Dimensionality Reduction for Density Ratio Estimation in High-dimensional Spaces

Masashi Sugiyama (sugi@cs.titech.ac.jp)
http://sugiyama-www.cs.titech.ac.jp/∼sugi/
Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Mathematisches Forchungsinstitute Oberwolfach
Schwarzwaldstr. 9-11, 77709 Oberwolfach-Walke, Germany

Motoaki Kawanabe (motoaki.kawanabe@first.fraunhofer.de)
Fraunhofer Institute FIRST.IDA
Kekuléstr. 7, D-12489 Berlin, Germany

Mathematisches Forchungsinstitute Oberwolfach
Schwarzwaldstr. 9-11, 77709 Oberwolfach-Walke, Germany

Pui Ling Chui (pauline@sg.cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract

The ratio of two probability density functions is becoming a quantity of interest these days in the machine learning and data mining communities since it can be used for various data processing tasks such as *non-stationarity adaptation*, *outlier detection*, and *feature selection*. Recently, several methods have been developed for directly estimating the density ratio without going through density estimation and were shown to work well in various practical problems. However, these methods still perform rather poorly when the dimensionality of the data domain is high. In this paper, we propose to incorporate a dimensionality reduction scheme into a density-ratio estimation procedure and experimentally show that the estimation accuracy in high-dimensional cases can be improved.

Keywords

density ratio estimation, dimensionality reduction, local Fisher discriminant analysis, unconstrained least-squares importance fitting

1 Introduction

The ratio of two probability density functions (a.k.a. the *importance*; see Fishman, 1996) is attracting a great deal of attention these days in the machine learning and data mining communities since it can be used for various statistical data processing tasks such as *covariate shift adaptation* (Shimodaira, 2000; Zadrozny, 2004; Sugiyama et al., 2007), *transfer learning* (Storkey & Sugiyama, 2007), *multi-task learning* (Bickel et al., 2008), *outlier detection* (Hido et al., 2008), *conditional density estimation* (Sugiyama et al., 2009), *variable selection* (Suzuki et al., 2008; Suzuki et al., 2009), *independent component analysis* (Suzuki & Sugiyama, 2009a), and *supervised dimensionality reduction* (Suzuki & Sugiyama, 2009b).

A naive approach to learning the density ratio is to estimate the two densities separately using a flexible technique such as *kernel density estimation* (Härdle et al., 2004) and then take the ratio of the estimated densities. However, this two-step approach is not reliable in practice since kernel density estimation performs poorly in high-dimensional cases; furthermore, division by an estimated density tends to magnify the estimation error.

Thus it is important to avoid density estimation when learning the density ratio. Actually, estimating the densities is more general than estimating the density ratio since knowing the densities implies knowing the ratio but not vice versa. Such a statement is sometimes referred to as *Vapnik's principle* (Vapnik, 1998) and 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.

Following this spirit, various methods have been developed for directly estimating the density ratio without going through density estimation (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009a). These methods are shown to compare favorably with naive kernel density estimation through extensive experiments. However, these methods still perform rather poorly when the dimensionality of the data domain is high.

The purpose of this paper is to develop a new method that can mitigate this problem. Our basic assumption behind the proposed method is that the difference of the two distributions (i.e., the distributions corresponding to the denominator and numerator of the density ratio) does not spread over the entire data domain, but is confined in a *subspace*—which we refer to as the *hetero-distributional subspace*. Once the hetero-distributional subspace can be identified, the density ratio is estimated only within this subspace, which leads to more stable and reliable estimation of the density ratio. We experimentally show that the proposed method—which we refer to as *Direct Density-ratio estimation with Dimensionality reduction* (D^3; pronounced as 'D-cube')—improves the accuracy of density ratio estimation in high-dimensional cases, while the computational cost is still kept moderate.

The rest of this paper is organized as follows. In Section 2, we formulate the problem of density ratio estimation and illustrate how the density ratio could be utilized in various data processing tasks. In Section 3, the basic idea of the proposed method D^3 is explained;

the details of the method are explained in Sections 4–6. Numerical examples are presented in Section 7 and concluding remarks are given in Section 8.

2 Formulation of Density Ratio Estimation Problem

In this section, we formulate the problem of density ratio estimation and briefly summarize possible usage of the density ratio in various data processing tasks.

2.1 Problem Formulation

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 first 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 article is to estimate the density ratio (also called the *importance* depending on the context)

$$r(\boldsymbol{x}) := \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} \tag{1}$$

from samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$. The subscripts 'nu' and 'de' denote 'numerator' and 'denominator', respectively.

2.2 Usage of Density Ratio in Data Processing

We are interested in estimating the density ratio since it is useful in various data processing tasks. Here we briefly review possible usage of the density ratio.

2.2.1 Covariate Shift Adaptation

Covariate shift (Shimodaira, 2000) is a situation in supervised learning where the input distributions change between the training and test phases but the conditional distribution of outputs given inputs remains unchanged. Under covariate shift, standard learning techniques such as maximum likelihood estimation are biased; the bias caused by covariate shift can be asymptotically canceled by weighting the loss function according to the importance (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Sugiyama et al., 2007). The basic idea of covariate shift adaptation is summarized in the following importance sampling identity:

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

where $r(\boldsymbol{x})$ is defined by Eq.(1). That is, the expectation of a function $g(\boldsymbol{x})$ over $p_{\mathrm{nu}}(\boldsymbol{x})$ can be computed by the importance-weighted expectation over $p_{\mathrm{de}}(\boldsymbol{x})$. Similarly, standard model selection criteria such as cross-validation or Akaike's information criterion lose their unbiasedness due to covariate shift; proper unbiasedness can be recovered by modifying the methods based on importance weighting (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Sugiyama et al., 2007; Huang et al., 2007; Quiñonero-Candela et al., 2009). Furthermore, the performance of active learning or the experiment design—the training input distribution is designed by the user to enhance the generalization performance—could also be improved by the use of the importance (Wiens, 2000; Kanamori & Shimodaira, 2003; Sugiyama, 2006; Sugiyama & Nakajima, 2009).

Thus the importance plays a central role in covariate shift adaptation and density-ratio estimation methods could be utilized for reducing the estimation bias under covariate shift. Examples of successful real-world applications include brain-computer interface (Sugiyama et al., 2007), robot control (Hachiya et al., 2009), speaker identification (Yamada et al., 2009), and natural language processing (Tsuboi et al., 2009). A similar importance-weighting idea also plays a central role in domain adaptation (Storkey & Sugiyama, 2007) and multi-task learning (Bickel et al., 2008).

2.2.2 Inlier-based Outlier Detection

Let us consider an outlier detection problem (Breunig et al., 2000; Schölkopf et al., 2001) of finding irregular samples in a dataset ('evaluation dataset') based on another dataset ('model dataset') that only contains regular samples. Defining the density ratio over the two sets of samples, we can see that the density-ratio values for regular samples 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 (Hido et al., 2008). Since the evaluation dataset has a wider support than the model dataset, we regard the evaluation dataset as samples corresponding to $p_{\mathrm{de}}(\boldsymbol{x})$ and the model dataset as samples corresponding to $p_{\mathrm{nu}}(\boldsymbol{x})$. Then outliers tend to have smaller density-ratio values (i.e., close to zero). As such, density-ratio estimation methods could be employed in outlier detection scenarios.

A similar idea could be used for change-point detection in time-series (Brodsky & Darkhovsky, 1993; Kawahara & Sugiyama, 2009) and two-sample problems in hypothesis testing (Henkel, 1979).

2.2.3 Conditional Density 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 density $q(\boldsymbol{y}|\boldsymbol{x})$. When the domain of $\boldsymbol{x}$ is continuous, conditional density estimation is not straightforward since a naive empirical approximation cannot be used (Bishop, 2006; Takeuchi et al., 2009).

In the context of density ratio estimation, 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 by

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

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

2.2.4 Mutual Information 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 $q(\boldsymbol{x}, \boldsymbol{y})$. 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

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

which plays a central role in information theory (Cover & Thomas, 1991).

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.

Mutual information can be used for measuring independence between random variables (Kraskov et al., 2004; Hulle, 2005) since it vanishes if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. Thus we can use density-ratio estimation methods, e.g., for variable selection (Suzuki et al., 2008; Suzuki et al., 2009), independent component analysis (Suzuki & Sugiyama, 2009a), and supervised dimensionality reduction (Suzuki & Sugiyama, 2009b).

3 Density Ratio Estimation with Dimensionality Reduction

As shown above, the density ratio is a useful quantity in various data processing tasks. However, the density ratio is usually unknown and needs to be estimated from data. Although methods of estimating the density ratio have been studied actively these days (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009a), estimating the density ratio in high-dimensional spaces is still a challenging problem. The goal of this paper is to give a practical density ratio estimation procedure for high-dimensional data. In this section, we describe the basic idea of the proposed procedure; details of the components of the procedure are described in the following sections.

3.1 Density Difference in Hetero-distributional Subspace

The basic assumption behind the proposed method is that the densities $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$ are not different in the entire space, but they are different only in some *subspace*. This

assumption can be mathematically formulated with the following linear mixing model.

Let $\{\boldsymbol{u}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ be i.i.d. samples drawn from an m-dimensional distribution with density $p_{\mathrm{de}}(\boldsymbol{u})$, where m is an integer such that $1 \leq m \leq d$ and we assume $p_{\mathrm{de}}(\boldsymbol{u}) > 0$ for all $\boldsymbol{u}$. Let $\{\boldsymbol{u}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ be i.i.d. samples drawn from another m-dimensional distribution with density $p_{\mathrm{nu}}(\boldsymbol{u})$. Let $\{\boldsymbol{v}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{v}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ be i.i.d. samples drawn from a $(d-m)$-dimensional distribution with density $p(\boldsymbol{v})$; we assume $p(\boldsymbol{v}) > 0$ for all $\boldsymbol{v}$. Let $\boldsymbol{A}$ be a $d \times m$ matrix and $\boldsymbol{B}$ be a $d \times (d-m)$ matrix such that the column vectors of $\boldsymbol{A}$ and $\boldsymbol{B}$ span the entire space. Based on these quantities, we consider the case where the samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ are generated as

$$\boldsymbol{x}_i^{\mathrm{de}} = \boldsymbol{A}\boldsymbol{u}_i^{\mathrm{de}} + \boldsymbol{B}\boldsymbol{v}_i^{\mathrm{de}},$$
$$\boldsymbol{x}_j^{\mathrm{nu}} = \boldsymbol{A}\boldsymbol{u}_j^{\mathrm{nu}} + \boldsymbol{B}\boldsymbol{v}_j^{\mathrm{nu}}.$$

Thus, $p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$ are expressed as

$$p_{\mathrm{de}}(\boldsymbol{x}) = c\, p_{\mathrm{de}}(\boldsymbol{u})p(\boldsymbol{v}),$$
$$p_{\mathrm{nu}}(\boldsymbol{x}) = c\, p_{\mathrm{nu}}(\boldsymbol{u})p(\boldsymbol{v}),$$

where c is the Jacobian between the observation $\boldsymbol{x}$ and the equi-/hetero-distributional components $(\boldsymbol{u}, \boldsymbol{v})$. We call $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{B})$ the *hetero-distributional subspace* and the *equi-distributional subspace*, respectively, where $\mathcal{R}(\cdot)$ denotes the range of a matrix. Note that $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{B})$ are not generally orthogonal (see Figure 1).

Under the above decomposability assumption, the density ratio is simplified as

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} = \frac{c\, p_{\mathrm{nu}}(\boldsymbol{u})p(\boldsymbol{v})}{c\, p_{\mathrm{de}}(\boldsymbol{u})p(\boldsymbol{v})} = \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} = r(\boldsymbol{u}). \tag{2}$$

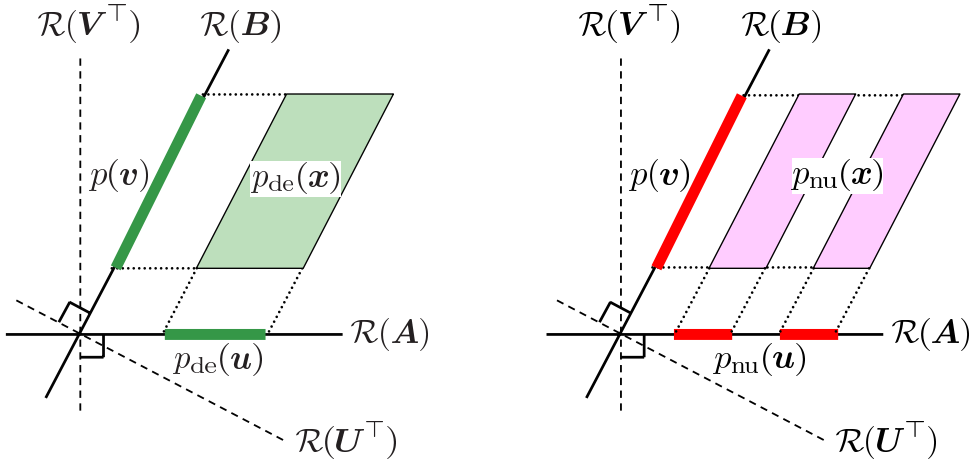
This means that the density ratio does not have to be estimated in the entire d-dimensional space, but only in the hetero-distributional subspace of dimension m ($\leq d$). Now we want to extract the hetero-distributional components $\boldsymbol{u}_i^{\mathrm{de}}$ and $\boldsymbol{u}_j^{\mathrm{nu}}$ from $\boldsymbol{x}_i^{\mathrm{de}}$ and $\boldsymbol{x}_j^{\mathrm{nu}}$, allowing estimation of the density ratio in $\mathcal{R}(\boldsymbol{A})$ via Eq.(2). As illustrated in Figure 1, the *oblique* projection of $\boldsymbol{x}_i^{\mathrm{de}}$ and $\boldsymbol{x}_j^{\mathrm{nu}}$ onto $\mathcal{R}(\boldsymbol{A})$ along $\mathcal{R}(\boldsymbol{B})$ allows us to obtain $\boldsymbol{u}_i^{\mathrm{de}}$ and $\boldsymbol{u}_j^{\mathrm{nu}}$.

3.2 Characterization of Hetero-distributional Subspace

Let us denote the oblique projection matrix onto $\mathcal{R}(\boldsymbol{A})$ along $\mathcal{R}(\boldsymbol{B})$ by $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$. In order to characterize the oblique projection matrix $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$, let us consider *dual* bases $\boldsymbol{U}$ and $\boldsymbol{V}$ for $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, i.e., $\boldsymbol{U}$ is an $m \times d$ matrix and $\boldsymbol{V}$ is a $(d-m) \times d$ matrix such that they are *bi-orthogonal* to each other:

$$\boldsymbol{U}\boldsymbol{B} = \boldsymbol{O}_{m\times(d-m)},$$
$$\boldsymbol{V}\boldsymbol{A} = \boldsymbol{O}_{(d-m)\times m},$$

where $\boldsymbol{O}_{m\times m'}$ denotes the $m \times m'$ matrix with all zeros. Thus $\mathcal{R}(\boldsymbol{B})$ and $\mathcal{R}(\boldsymbol{U}^\top)$ are orthogonal, and $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{V}^\top)$ are orthogonal where $^\top$ denotes the transpose. When



(a) $p_{\mathrm{de}}(\boldsymbol{x}) \propto p_{\mathrm{de}}(\boldsymbol{u})p(\boldsymbol{v})$ (b) $p_{\mathrm{nu}}(\boldsymbol{x}) \propto p_{\mathrm{nu}}(\boldsymbol{u})p(\boldsymbol{v})$

Figure 1: A schematic picture of the hetero-distributional subspace for $d = 2$ and $m = 1$. Let $\boldsymbol{A} \propto (1, 0)^\top$ and $\boldsymbol{B} \propto (1, 2)^\top$; then $\boldsymbol{U} \propto (2, -1)$ and $\boldsymbol{V} \propto (0, 1)$. $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{B})$ are called the hetero-distributional subspace and the equi-distributional subspace, respectively. If a data point $\boldsymbol{x}$ is projected onto $\mathcal{R}(\boldsymbol{A})$ along $\mathcal{R}(\boldsymbol{B})$, the equi-distributional component $\boldsymbol{v}$ can be eliminated and the hetero-distributional component $\boldsymbol{u}$ can be extracted.

$\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{B})$ are orthogonal, $\mathcal{R}(\boldsymbol{U}^\top)$ agrees with $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{V}^\top)$ agrees with $\mathcal{R}(\boldsymbol{B})$; however, they are different in general as illustrated in Figure 1.

The relation between $\boldsymbol{A}$ and $\boldsymbol{B}$ and the relation between $\boldsymbol{U}$ and $\boldsymbol{V}$ can be characterized in terms of the covariance matrix $\boldsymbol{\Sigma}$ (of either $p_{\mathrm{de}}(\boldsymbol{x})$ or $p_{\mathrm{nu}}(\boldsymbol{x})$) as

$$\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{B} = \boldsymbol{O}_{(d-m)\times m}, \tag{3}$$

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{O}_{m\times(d-m)}. \tag{4}$$

These orthogonarilies in terms of $\boldsymbol{\Sigma}$ follow from the statistical independence between the components in $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{B})$—more specifically, Eq.(3) follows from the fact that the *sphering* operation (transforming samples $\boldsymbol{x}$ by $\boldsymbol{\Sigma}^{-1/2}$ in advance) orthogonalizes independent components $\boldsymbol{u}$ and $\boldsymbol{v}$ (Hyvärinen et al., 2001) and Eq.(4) is its dual expression (Kawanabe et al., 2007). After sphering, the covariance matrix becomes identity and all the discussions become simpler. However, estimating the covariance matrix from samples is erroneous and taking its inverse further magnifies the estimation error. For this reason, we directly deal with non-orthogonal $\boldsymbol{A}$ and $\boldsymbol{B}$ below.

For normalization purposes, we further assume that

$$\boldsymbol{U}\boldsymbol{A} = \boldsymbol{I}_m,$$
$$\boldsymbol{V}\boldsymbol{B} = \boldsymbol{I}_{d-m},$$

where $\boldsymbol{I}_m$ denotes the m-dimensional identity matrix. Then the oblique projection ma-

trices $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$ and $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{B}),\mathcal{R}(\boldsymbol{A})}$ can be expressed as

$$\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})} = \boldsymbol{A}\boldsymbol{U},$$
$$\boldsymbol{P}_{\mathcal{R}(\boldsymbol{B}),\mathcal{R}(\boldsymbol{A})} = \boldsymbol{B}\boldsymbol{V},$$

which can be confirmed by the facts that $\boldsymbol{P}^2_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})} = \boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$ (*idempotence*), the null space of $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$ is $\mathcal{R}(\boldsymbol{B})$, and the range of $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$ is $\mathcal{R}(\boldsymbol{A})$; the same goes for $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{B}),\mathcal{R}(\boldsymbol{A})}$. The above expressions of $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{A}),\mathcal{R}(\boldsymbol{B})}$ and $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{B}),\mathcal{R}(\boldsymbol{A})}$ imply that $\boldsymbol{U}$ plays a role of expressing projected images in an m-dimensional coordinate system within $\mathcal{R}(\boldsymbol{A})$, and $\boldsymbol{V}$ plays a role of expressing projected images in a $(d-m)$-dimensional coordinate system within $\mathcal{R}(\boldsymbol{B})$. We call $\boldsymbol{U}$ and $\boldsymbol{V}$ *the hetero-distributional mapping* and *the equi-distributional mapping*, respectively.

Now $\boldsymbol{u}_i^{\mathrm{de}}$, $\boldsymbol{u}_j^{\mathrm{nu}}$, $\boldsymbol{v}_i^{\mathrm{de}}$, and $\boldsymbol{v}_j^{\mathrm{nu}}$ are expressed as

$$\boldsymbol{u}_i^{\mathrm{de}} = \boldsymbol{U}\boldsymbol{x}_i^{\mathrm{de}},$$
$$\boldsymbol{u}_j^{\mathrm{nu}} = \boldsymbol{U}\boldsymbol{x}_j^{\mathrm{nu}},$$
$$\boldsymbol{v}_i^{\mathrm{de}} = \boldsymbol{V}\boldsymbol{x}_i^{\mathrm{de}},$$
$$\boldsymbol{v}_j^{\mathrm{nu}} = \boldsymbol{V}\boldsymbol{x}_j^{\mathrm{nu}}.$$

Thus, if the hetero-distributional mapping $\boldsymbol{U}$ was estimated, estimation of the density ratio $r(\boldsymbol{x})$ could be carried out in a low-dimensional hetero-distributional subspace via Eq.(2).

We show how to estimate the hetero-distributional mapping $\boldsymbol{U}$ in Section 4, and then we give a method of estimating the density ratio within the hetero-distributional subspace in Section 5. For a while, we assume that the dimension m of the hetero-distributional subspace is known; we show how m is estimated in Section 6. Below, we refer to the proposed method as *Direct Density-ratio estimation with Dimensionality reduction* (D^3).

4 Identifying the Hetero-distributional Subspace by Supervised Dimensionality Reduction

In this section, we show how to estimate the hetero-distributional subspace.

4.1 Basic Idea

In order to estimate the hetero-distributional subspace, we need a criterion that reflects the degree of distributional difference in a subspace. A key observation in this context is that the existence of distributional difference can be checked whether samples from the two distributions can be separated from each other. That is, if we can distinguish samples of one distribution from the samples of the other distribution, we may conclude that two distributions are different; otherwise distributions may be similar. We employ this idea

for finding the hetero-distributional subspace. Let us denote the samples projected onto the hetero-distributional subspace by

$$\{\boldsymbol{u}_i^{\mathrm{de}} \mid \boldsymbol{u}_i^{\mathrm{de}} = \boldsymbol{U}\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}},$$
$$\{\boldsymbol{u}_j^{\mathrm{nu}} \mid \boldsymbol{u}_j^{\mathrm{nu}} = \boldsymbol{U}\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}.$$

Then our goal is to find the matrix $\boldsymbol{U}$ such that $\{\boldsymbol{u}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{u}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ are maximally separated from each other. For that purpose, we may use *any* supervised dimensionality reduction methods.

Among various supervised dimensionality reduction methods (e.g., Hastie & Tibshirani, 1996a; Hastie & Tibshirani, 1996b; Fukumizu et al., 2004; Goldberger et al., 2005; Globerson & Roweis, 2006), we decided to use *local Fisher discriminant analysis* (LFDA; Sugiyama, 2007) which is an extension of classical *Fisher discriminant analysis* (FDA; Fisher, 1936). LFDA has various practically useful properties, e.g., there is no limitation on the dimension of the reduced subspace, it works well even when data has multimodal structure (such as separate clusters), it is robust against outliers, its solution can be analytically computed using eigenvalue decomposition in a stable and efficient manner, and its experimental performance is shown to be better than other methods.

The rest of this section is devoted to reviewing technical details of LFDA—showing how to use it in search for the hetero-distributional subspace and discussing the validity of our choice. Below, we consider a set of binary-labeled training samples

$$\{(\boldsymbol{x}_k, y_k) \mid \boldsymbol{x}_k \in \mathbb{R}^d, y_k \in \{+1, -1\}\}_{k=1}^{n},$$

and reduce the dimensionality of $\boldsymbol{x}_k$ using an $m \times d$ transformation matrix $\boldsymbol{T}$ matrix as

$$\boldsymbol{T}\boldsymbol{x}_k.$$

Effectively, the training samples $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{n}$ correspond to the following setup: for $n = n_{\mathrm{de}} + n_{\mathrm{nu}}$,

$$\{\boldsymbol{x}_k\}_{k=1}^{n} = \{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}} \cup \{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}},$$
$$y_k = \begin{cases} +1 & \text{if } \boldsymbol{x}_k \in \{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}, \\ -1 & \text{if } \boldsymbol{x}_k \in \{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}. \end{cases}$$

4.2 Fisher Discriminant Analysis (FDA)

Since LFDA is an extention of FDA (Fisher, 1936), we first briefly review original FDA.

Let n_+ and n_- be the number of samples in class $+1$ and -1, respectively. Let $\boldsymbol{\mu}$, $\boldsymbol{\mu}_+$,

and $\boldsymbol{\mu}_-$ be the mean of $\{\boldsymbol{x}_k\}_{k=1}^n$, $\{\boldsymbol{x}_k|y_k=+1\}_{k=1}^n$, and $\{\boldsymbol{x}_k|y_k=-1\}_{k=1}^n$, respectively:

$$\boldsymbol{\mu} := \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k,$$
$$\boldsymbol{\mu}_+ := \frac{1}{n_+}\sum_{k:y_k=+1}\boldsymbol{x}_k,$$
$$\boldsymbol{\mu}_- := \frac{1}{n_-}\sum_{k:y_k=-1}\boldsymbol{x}_k.$$

Let $\boldsymbol{S}^{\mathrm{b}}$ and $\boldsymbol{S}^{\mathrm{w}}$ be the *between-class scatter matrix* and the *within-class scatter matrix* defined as

$$\boldsymbol{S}^{\mathrm{b}} := n_+(\boldsymbol{\mu}_+ - \boldsymbol{\mu})(\boldsymbol{\mu}_+ - \boldsymbol{\mu})^\top + n_-(\boldsymbol{\mu}_- - \boldsymbol{\mu})(\boldsymbol{\mu}_- - \boldsymbol{\mu})^\top,$$
$$\boldsymbol{S}^{\mathrm{w}} := \sum_{k:y_k=+1}(\boldsymbol{x}_k - \boldsymbol{\mu}_+)(\boldsymbol{x}_k - \boldsymbol{\mu}_+)^\top + \sum_{k:y_k=-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_-)(\boldsymbol{x}_k - \boldsymbol{\mu}_-)^\top.$$

The FDA transformation matrix $\boldsymbol{T}_{\mathrm{FDA}}$ is defined as

$$\boldsymbol{T}_{\mathrm{FDA}} := \operatorname*{argmax}_{\boldsymbol{T}\in\mathbb{R}^{m\times d}}\left[\mathrm{tr}(\boldsymbol{T}\boldsymbol{S}^{\mathrm{b}}\boldsymbol{T}^\top(\boldsymbol{T}\boldsymbol{S}^{\mathrm{w}}\boldsymbol{T}^\top)^{-1})\right].$$

That is, FDA seeks a transformation matrix $\boldsymbol{T}$ such that between-class scatter is maximized and within-class scatter is minimized in the embedding space $\mathbb{R}^m$.

Let $\{\boldsymbol{\phi}_l\}_{l=1}^d$ be the generalized eigenvectors associated with the generalized eigenvalues $\{\lambda_l\}_{l=1}^d$ of the following generalized eigenvalue problem:

$$\boldsymbol{S}^{\mathrm{b}}\boldsymbol{\phi} = \lambda\boldsymbol{S}^{\mathrm{w}}\boldsymbol{\phi}.$$

We assume that the generalized eigenvalues are sorted as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d.$$

Then a solution $\boldsymbol{T}_{\mathrm{FDA}}$ is analytically given as follows (e.g., Duda et al., 2001):

$$\boldsymbol{T}_{\mathrm{FDA}} = (\boldsymbol{\phi}_1|\boldsymbol{\phi}_2|\cdots|\boldsymbol{\phi}_m)^\top.$$

Thus FDA is computationally efficient.

FDA works very well if samples in each class are Gaussian with common covariance structure. However, it tends to give undesired results if samples in a class form several separate clusters or there exist outliers. Furthermore, the between-class scatter matrix $\boldsymbol{S}^{\mathrm{b}}$ is known to have rank one in the current setup (see e.g., Fukunaga, 1990), implying that we can obtain only one meaningful feature $\boldsymbol{\phi}_1$ through the FDA criterion; the remaining features $\{\boldsymbol{\phi}_l\}_{l=2}^d$ found by FDA are arbitrary in the null space of $\boldsymbol{S}^{\mathrm{b}}$. This is an essential limitation of FDA in dimensionality reduction.

4.3 Local Fisher Discriminant Analysis (LFDA)

In order to overcome the weaknesses of FDA explained above, LFDA has been introduced (Sugiyama, 2007). Here, we explain the main idea of LFDA briefly.

The scatter matrices $\boldsymbol{S}^{\mathrm{b}}$ and $\boldsymbol{S}^{\mathrm{w}}$ in original FDA can be expressed in the pairwise form as follows.

$$\boldsymbol{S}^{\mathrm{b}} = \frac{1}{2}\sum_{k,k'=1}^{n} W^{\mathrm{b}}_{k,k'}(\boldsymbol{x}_k - \boldsymbol{x}_{k'})(\boldsymbol{x}_k - \boldsymbol{x}_{k'})^\top,$$
$$\boldsymbol{S}^{\mathrm{w}} = \frac{1}{2}\sum_{k,k'=1}^{n} W^{\mathrm{w}}_{k,k'}(\boldsymbol{x}_k - \boldsymbol{x}_{k'})(\boldsymbol{x}_k - \boldsymbol{x}_{k'})^\top,$$

where

$$W^{\mathrm{b}}_{k,k'} := \begin{cases} 1/n - 1/n_+ & \text{if } y_k = y_{k'} = +1, \\ 1/n - 1/n_- & \text{if } y_k = y_{k'} = -1, \\ 1/n & \text{if } y_k \neq y_{k'}, \end{cases}$$
$$W^{\mathrm{w}}_{k,k'} := \begin{cases} 1/n_+ & \text{if } y_k = y_{k'} = +1, \\ 1/n_- & \text{if } y_k = y_{k'} = -1, \\ 0 & \text{if } y_k \neq y_{k'}. \end{cases}$$

Based on the above pairwise expression, let us define the *local* between-class scatter matrix $\boldsymbol{S}^{\mathrm{lb}}$ and the *local* within-class scatter matrix $\boldsymbol{S}^{\mathrm{lw}}$ as

$$\boldsymbol{S}^{\mathrm{lb}} := \frac{1}{2}\sum_{k,k'=1}^{n} W^{\mathrm{lb}}_{k,k'}(\boldsymbol{x}_k - \boldsymbol{x}_{k'})(\boldsymbol{x}_k - \boldsymbol{x}_{k'})^\top,$$
$$\boldsymbol{S}^{\mathrm{lw}} := \frac{1}{2}\sum_{k,k'=1}^{n} W^{\mathrm{lw}}_{k,k'}(\boldsymbol{x}_k - \boldsymbol{x}_{k'})(\boldsymbol{x}_k - \boldsymbol{x}_{k'})^\top,$$

where

$$W^{\mathrm{lb}}_{k,k'} := \begin{cases} A_{k,k'}(1/n - 1/n_+) & \text{if } y_k = y_{k'} = +1, \\ A_{k,k'}(1/n - 1/n_-) & \text{if } y_k = y_{k'} = -1, \\ 1/n & \text{if } y_k \neq y_{k'}, \end{cases}$$
$$W^{\mathrm{lw}}_{k,k'} := \begin{cases} A_{k,k'}/n_+ & \text{if } y_k = y_{k'} = +1, \\ A_{k,k'}/n_- & \text{if } y_k = y_{k'} = -1, \\ 0 & \text{if } y_k \neq y_{k'}. \end{cases}$$

$A_{k,k'}$ is the affinity value between $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k'}$ defined based on the *local scaling heuristic* (Zelnik-Manor & Perona, 2005):

$$A_{k,k'} := \exp\left(-\frac{\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\|^2}{\eta_k \eta_{k'}}\right).$$

η_k is the local scaling factor around $\boldsymbol{x}_k$ defined by

$$\eta_k := \|\boldsymbol{x}_k - \boldsymbol{x}_k^{(K)}\|,$$

where $\boldsymbol{x}_k^{(K)}$ denotes the K-th nearest neighbor of $\boldsymbol{x}_k$. A heuristic choice of $K = 7$ was shown to be useful through extensive simulations (Zelnik-Manor & Perona, 2005; Sugiyama, 2007). Note that the local scaling factors are computed in a classwise manner in LFDA.

Based on the local scatter matrices $\boldsymbol{S}^{\mathrm{lb}}$ and $\boldsymbol{S}^{\mathrm{lw}}$, the LFDA transformation matrix $\boldsymbol{T}_{\mathrm{LFDA}}$ is defined as

$$\boldsymbol{T}_{\mathrm{LFDA}} := \operatorname*{argmax}_{\boldsymbol{T}\in\mathbb{R}^{m\times d}} \left[\mathrm{tr}(\boldsymbol{T}\boldsymbol{S}^{\mathrm{lb}}\boldsymbol{T}^\top(\boldsymbol{T}\boldsymbol{S}^{\mathrm{lw}}\boldsymbol{T}^\top)^{-1})\right].$$

The definition of $\boldsymbol{S}^{\mathrm{lb}}$ and $\boldsymbol{S}^{\mathrm{lw}}$ implies that LFDA seeks a transformation matrix $\boldsymbol{T}$ such that *nearby* data pairs in the same class are made close and the data pairs in different classes are made apart; *far apart* data pairs in the same class are not imposed to be close.

By this localization effect, LFDA can overcome the weakness of original FDA against clustered data and outliers. When $A_{k,k'} = 1$ for all k, k' (i.e., no locality), $\boldsymbol{S}^{\mathrm{lw}}$ and $\boldsymbol{S}^{\mathrm{lb}}$ are reduced to $\boldsymbol{S}^{\mathrm{w}}$ and $\boldsymbol{S}^{\mathrm{b}}$. Thus, LFDA could be regarded as a natural localized variant of FDA. The between-class scatter matrix $\boldsymbol{S}^{\mathrm{b}}$ in original FDA had only rank one, while its local counterpart $\boldsymbol{S}^{\mathrm{lb}}$ in LFDA usually has full rank with no multiplicity in eigenvalues (given $n \geq d$). Therefore, LFDA can be applied to dimensionality reduction into *any* dimensional spaces, which is a significant advantage over original FDA.

A solution $\boldsymbol{T}_{\mathrm{LFDA}}$ can be computed in the same way as original FDA; namely, the LFDA solution is given as

$$\boldsymbol{T}_{\mathrm{LFDA}} = (\boldsymbol{\varphi}_1|\boldsymbol{\varphi}_2|\cdots|\boldsymbol{\varphi}_m)^\top,$$

where $\{\boldsymbol{\varphi}_l\}_{l=1}^d$ are the generalized eigenvectors associated with the generalized eigenvalues $\{\gamma_l\}_{l=1}^d$ of the following generalized eigenvalue problem:

$$\boldsymbol{S}^{\mathrm{lb}}\boldsymbol{\varphi} = \gamma\boldsymbol{S}^{\mathrm{lw}}\boldsymbol{\varphi}. \tag{5}$$

We assume that the generalized eigenvalues are sorted as

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d.$$

Thus LFDA is computationally as efficient as original FDA. A pseudo code of LFDA is summarized in Figure 2.

4.4 Use of LFDA for Finding Hetero-distributional Subspace

Finally, we show how to obtain an estimate of the transformation matrix $\boldsymbol{U}$ needed in the density-ratio estimation procedure (see Section 4.1 again) from the LFDA transformation matrix $\boldsymbol{T}_{\mathrm{LFDA}}$.

Input: Two sets of samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ on $\mathbb{R}^d$
Dimensionality of embedding space m $(1 \leq m \leq d)$
Output: $m \times d$ transformation matrix $\widehat{\boldsymbol{U}}$

$\widetilde{\boldsymbol{x}}_i^{\mathrm{de}} \longleftarrow$ 7th nearest neighbor of $\boldsymbol{x}_i^{\mathrm{de}}$ among $\{\boldsymbol{x}_{i'}^{\mathrm{de}}\}_{i'=1}^{n_{\mathrm{de}}}$ for $i = 1, 2, \ldots, n_{\mathrm{de}}$;
$\widetilde{\boldsymbol{x}}_j^{\mathrm{nu}} \longleftarrow$ 7th nearest neighbor of $\boldsymbol{x}_j^{\mathrm{nu}}$ among $\{\boldsymbol{x}_{j'}^{\mathrm{nu}}\}_{j'=1}^{n_{\mathrm{nu}}}$ for $j = 1, 2, \ldots, n_{\mathrm{nu}}$;
$\eta_i^{\mathrm{de}} \longleftarrow \|\boldsymbol{x}_i^{\mathrm{de}} - \widetilde{\boldsymbol{x}}_i^{\mathrm{de}}\|$ for $i = 1, 2, \ldots, n_{\mathrm{de}}$;
$\eta_j^{\mathrm{nu}} \longleftarrow \|\boldsymbol{x}_j^{\mathrm{nu}} - \widetilde{\boldsymbol{x}}_j^{\mathrm{nu}}\|$ for $j = 1, 2, \ldots, n_{\mathrm{nu}}$;
$A_{i,i'}^{\mathrm{de}} \longleftarrow \exp\left(-\dfrac{\|\boldsymbol{x}_i^{\mathrm{de}} - \boldsymbol{x}_{i'}^{\mathrm{de}}\|^2}{\eta_i^{\mathrm{de}}\eta_{i'}^{\mathrm{de}}}\right)$ for $i, i' = 1, 2, \ldots, n_{\mathrm{de}}$;
$A_{j,j'}^{\mathrm{nu}} \longleftarrow \exp\left(-\dfrac{\|\boldsymbol{x}_j^{\mathrm{nu}} - \boldsymbol{x}_{j'}^{\mathrm{nu}}\|^2}{\eta_j^{\mathrm{nu}}\eta_{j'}^{\mathrm{nu}}}\right)$ for $j, j' = 1, 2, \ldots, n_{\mathrm{nu}}$;
$\boldsymbol{X}_{\mathrm{de}} \longleftarrow (\boldsymbol{x}_1^{\mathrm{de}}|\boldsymbol{x}_2^{\mathrm{de}}|\cdots|\boldsymbol{x}_{n_{\mathrm{de}}}^{\mathrm{de}})$;
$\boldsymbol{X}_{\mathrm{nu}} \longleftarrow (\boldsymbol{x}_1^{\mathrm{nu}}|\boldsymbol{x}_2^{\mathrm{nu}}|\cdots|\boldsymbol{x}_{n_{\mathrm{nu}}}^{\mathrm{nu}})$;
$\boldsymbol{G}_{\mathrm{de}} \longleftarrow \boldsymbol{X}_{\mathrm{de}}\mathrm{diag}(\boldsymbol{A}_{\mathrm{de}}\boldsymbol{1}_{n_{\mathrm{de}}})\boldsymbol{X}_{\mathrm{de}}^\top - \boldsymbol{X}_{\mathrm{de}}\boldsymbol{A}_{\mathrm{de}}\boldsymbol{X}_{\mathrm{de}}^\top$;
$\boldsymbol{G}_{\mathrm{nu}} \longleftarrow \boldsymbol{X}_{\mathrm{nu}}\mathrm{diag}(\boldsymbol{A}_{\mathrm{nu}}\boldsymbol{1}_{n_{\mathrm{nu}}})\boldsymbol{X}_{\mathrm{nu}}^\top - \boldsymbol{X}_{\mathrm{nu}}\boldsymbol{A}_{\mathrm{nu}}\boldsymbol{X}_{\mathrm{nu}}^\top$;
$\boldsymbol{S}^{\mathrm{lw}} \longleftarrow \frac{1}{n_{\mathrm{de}}}\boldsymbol{G}_{\mathrm{de}} + \frac{1}{n_{\mathrm{nu}}}\boldsymbol{G}_{\mathrm{nu}}$;
$n \longleftarrow n_{\mathrm{de}} + n_{\mathrm{nu}}$;
$\boldsymbol{S}^{\mathrm{lb}} \longleftarrow (\frac{1}{n} - \frac{1}{n_{\mathrm{de}}})\boldsymbol{G}_{\mathrm{de}} + (\frac{1}{n} - \frac{1}{n_{\mathrm{nu}}})\boldsymbol{G}_{\mathrm{nu}} + \frac{n_{\mathrm{nu}}}{n}\boldsymbol{X}_{\mathrm{de}}\boldsymbol{X}_{\mathrm{de}}^\top + \frac{n_{\mathrm{de}}}{n}\boldsymbol{X}_{\mathrm{nu}}\boldsymbol{X}_{\mathrm{nu}}^\top$
$\quad -\frac{1}{n}\boldsymbol{X}_{\mathrm{de}}\boldsymbol{1}_{n_{\mathrm{de}}}(\boldsymbol{X}_{\mathrm{nu}}\boldsymbol{1}_{n_{\mathrm{nu}}})^\top - \frac{1}{n}\boldsymbol{X}_{\mathrm{nu}}\boldsymbol{1}_{n_{\mathrm{nu}}}(\boldsymbol{X}_{\mathrm{de}}\boldsymbol{1}_{n_{\mathrm{de}}})^\top$;
$\{\gamma_l, \boldsymbol{\varphi}_l\}_{l=1}^m \longleftarrow$ generalized eigenvalues and eigenvectors of $\boldsymbol{S}^{\mathrm{lb}}\boldsymbol{\varphi} = \gamma\boldsymbol{S}^{\mathrm{lw}}\boldsymbol{\varphi}$;
$\quad$ % $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d$
$\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^m \longleftarrow$ orthonormal basis of $\{\boldsymbol{\varphi}_l\}_{l=1}^m$;
$\quad$ % $\mathrm{span}(\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^{m'}) = \mathrm{span}(\{\boldsymbol{\varphi}_l\}_{l=1}^{m'})$ for $m' = 1, 2, \ldots, m$
$\widehat{\boldsymbol{U}} \longleftarrow (\widetilde{\boldsymbol{\varphi}}_1|\widetilde{\boldsymbol{\varphi}}_2|\cdots|\widetilde{\boldsymbol{\varphi}}_m)^\top$;

Figure 2: Pseudo code of LFDA. $\boldsymbol{1}_n$ denotes the n-dimensional vectors with all ones, and $\mathrm{diag}(\boldsymbol{b})$ denotes the diagonal matrix with diagonal elements specified by a vector $\boldsymbol{b}$. A MATLAB implementation of LFDA is available from '`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA/`'.

First, an orthonormal basis $\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^m$ of the LFDA subspace is computed from the generalized eigenvectors $\{\boldsymbol{\varphi}_l\}_{l=1}^m$ so that the span of $\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^{m'}$ agrees with the span of $\{\boldsymbol{\varphi}_l\}_{l=1}^{m'}$ for all m' $(1 \leq m' \leq m)$. This can be carried out e.g., by the Gram-Schmidt orthonormalization (see e.g., Albert, 1972). Then an estimate $\widehat{\boldsymbol{U}}$ is given as

$$\widehat{\boldsymbol{U}} := (\widetilde{\boldsymbol{\varphi}}_1|\widetilde{\boldsymbol{\varphi}}_2|\cdots|\widetilde{\boldsymbol{\varphi}}_m)^\top,$$

and the samples are transformed as

$$\widehat{\boldsymbol{u}}_i^{\mathrm{de}} := \widehat{\boldsymbol{U}}\boldsymbol{x}_i^{\mathrm{de}} \quad \text{for } i = 1, 2, \ldots, n_{\mathrm{de}},$$
$$\widehat{\boldsymbol{u}}_j^{\mathrm{nu}} := \widehat{\boldsymbol{U}}\boldsymbol{x}_j^{\mathrm{nu}} \quad \text{for } j = 1, 2, \ldots, n_{\mathrm{nu}}.$$

The above expression of $\widehat{\boldsymbol{U}}$ implies another useful advantage of LFDA. In the proposed density-ratio estimation procedure, we need the LFDA solution for each reduced dimensionality $m' = 1, 2, \ldots, d$ (see Section 6). However, we do not actually have to compute the LFDA solution for each m', but only to solve the generalized eigenvalue problem (5) once for $m' = d$ and compute the orthonormal basis $\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^{d}$; the solution for $m' < d$ can be obtained by simply taking the first m' basis vectors $\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^{m'}$.

4.5 Discussion

The hetero-distributional subspace can be identified by measuring the degree of distributional difference in a subspace and finding the subspace that has the smallest distributional difference.

The problem of measuring the degree of distributional difference is also called the *two-sample problem* in the context of significance tests, and the most fundamental approach would be the *t-test* and its multi-variate extension (Hotelling, 1951). However, the t-test is based on the parametric assumption that the target distribution is Gaussian and this model assumption is too restrictive in practical data analysis.

The *Kolmogorov-Smirnov test* and the *Wald-Wolfowitz runs test* are classical non-parametric methods of the two-sample problem; their multi-dimensional variants have also been developed (Bickel, 1969; Friedman & Rafsky, 1979). More recently, different types of non-parametric tests have been proposed, e.g., based on the permutation test (Hall & Tajvidi, 2002), the distance between the densities (Anderson et al., 1994; Biau & Gyorfi, 2005), and moment matching with the kernel trick (Borgwardt et al., 2006; Gretton et al., 2007). However, these methods tend to suffer from serious computational problems and therefore they are not applicable to large-scale problems.

Compared with other supervised dimensionality reduction methods such as the method based on discriminant adaptive nearest neighbor classifiers (Hastie & Tibshirani, 1996a), mixture discriminant analysis (Hastie & Tibshirani, 1996b), kernel dimensionality reduction (Fukumizu et al., 2004), neighborhood component analysis (Goldberger et al., 2005), and maximally collapsing metric learning (Globerson & Roweis, 2006), LFDA is advantageous in that it works well even when data has multimodal structure (such as separate clusters), it is robust against outliers, its solution can be analytically computed using eigenvalue decomposition in a stable and efficient manner, and its experimental performance is shown to be better than other methods.

Thus, for developing a density-ratio estimation method that is flexible and computationally efficient, the use of LFDA would be suitable.

5 Directly Estimating Density Ratio without Density Estimation

Given that the hetero-distributional subspace has been successfully identified, the next step is to estimate the density ratio within the subspace. In this section, we describe our

approach.

5.1 Basic Idea

A naive approach to estimating the density ratio $r(\boldsymbol{u})$ is to estimate the two densities $p_{\mathrm{de}}(\boldsymbol{u})$ and $p_{\mathrm{nu}}(\boldsymbol{u})$ separately from samples $\{\boldsymbol{u}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{u}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ using a flexible technique such as *kernel density estimation* (Härdle et al., 2004), and then take the ratio of the estimated densities. However, this naive approach is not reliable since division by an estimated density tends to magnify the estimation error, which can heavily degrade the estimation accuracy in practice[1].

To cope with this problem, several methods have been proposed for directly estimating the density ratio without going through density estimation (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009a). Among them, we decided to use the method called *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009a) since it has several advantages over the other methods, e.g., its solution can be analytically computed by solving a system of linear equations in a stable and efficient manner, model selection is possible via cross-validation (CV), the leave-one-out CV score can be computed analytically without repeating hold-out loops, and it is reported to perform well in experiments.

In the rest of this section, we review technical details of uLSIF and discuss the validity of our choice.

5.2 Linear Least-squares Estimation of Density Ratio

Let us model the density ratio $r(\boldsymbol{u})$ by the following linear model:

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

where

$$\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$$

are parameters to be learned from data samples and $\{\psi_\ell(\boldsymbol{u})\}_{\ell=1}^{b}$ are basis functions such that

$$\psi_\ell(\boldsymbol{u}) \geq 0 \text{ for all } \boldsymbol{u} \text{ and for } \ell = 1, 2, \ldots, b.$$

Note that b and $\{\psi_\ell(\boldsymbol{u})\}_{\ell=1}^{b}$ could be dependent on the samples $\{\boldsymbol{u}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{u}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ so kernel models are also allowed. We explain how the basis functions $\{\psi_\ell(\boldsymbol{u})\}_{\ell=1}^{b}$ are designed in Section 5.3.

[1] Furthermore, kernel density estimation does not allow us to choose the dimensionality of the hetero-distributional subspace (see Section 6).

The parameters $\{\alpha_\ell\}_{\ell=1}^b$ in the model $\widehat{r}(\boldsymbol{u})$ are determined so that the following squared error J_0 is minimized:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &:= \frac{1}{2}\int \left(\widehat{r}(\boldsymbol{u}) - r(\boldsymbol{u})\right)^2 p_{\mathrm{de}}(\boldsymbol{u})d\boldsymbol{u} \\ &= \frac{1}{2}\int \widehat{r}(\boldsymbol{u})^2 p_{\mathrm{de}}(\boldsymbol{u})d\boldsymbol{u} - \int \widehat{r}(\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u})d\boldsymbol{u} + \frac{1}{2}\int r(\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u})d\boldsymbol{u}, \end{aligned}$$

where the last term is a constant and therefore can be safely ignored. Let us denote the first two terms by J:

$$J(\boldsymbol{\alpha}) := \frac{1}{2}\int \widehat{r}(\boldsymbol{u})^2 p_{\mathrm{de}}(\boldsymbol{u})d\boldsymbol{u} - \int \widehat{r}(\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u})d\boldsymbol{u}. \tag{6}$$

Note that the same objective function can be obtained via the *Legendre-Fenchel duality* of a divergence (Nguyen et al., 2008).

Approximating the expectations in J by empirical averages and replacing $\boldsymbol{U}$ by its estimate $\widehat{\boldsymbol{U}}$ (see Section 4), we obtain

$$\begin{aligned} \widehat{J}(\boldsymbol{\alpha}) &:= \frac{1}{2n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}} \widehat{r}(\widehat{\boldsymbol{u}}_i^{\mathrm{de}})^2 - \frac{1}{n_{\mathrm{nu}}}\sum_{j=1}^{n_{\mathrm{nu}}} \widehat{r}(\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}) \\ &= \frac{1}{2}\sum_{\ell,\ell'=1}^{b} \alpha_\ell \alpha_{\ell'} \left(\frac{1}{n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}} \psi_\ell(\widehat{\boldsymbol{u}}_i^{\mathrm{de}})\psi_{\ell'}(\widehat{\boldsymbol{u}}_i^{\mathrm{de}})\right) - \sum_{\ell=1}^{b}\alpha_\ell \left(\frac{1}{n_{\mathrm{nu}}}\sum_{j=1}^{n_{\mathrm{nu}}} \psi_\ell(\widehat{\boldsymbol{u}}_j^{\mathrm{nu}})\right) \\ &= \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha}, \end{aligned}$$

where $\widehat{\boldsymbol{H}}$ is the $b \times b$ matrix with the (ℓ, ℓ')-th element

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

and $\widehat{\boldsymbol{h}}$ is the b-dimensional vector with the ℓ-th element

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

Now the optimization problem is formulated as follows.

$$\widetilde{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right], \tag{7}$$

where a penalty term $\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}/2$ is included for regularization purposes and λ (≥ 0) is a regularization parameter that controls strength of regularization. It is easy to confirm that the solution of Eq.(7) can be analytically computed as

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

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix.

The density ratio is always non-negative by definition, but some of the parameters $\{\widehat{\alpha}_\ell\}_{\ell=1}^b$ obtained through the above optimization problem could be negative. To cope with this problem, we modify the solution as

$$\widehat{\boldsymbol{\alpha}} := \max(\boldsymbol{0}_b, \widetilde{\boldsymbol{\alpha}}),$$

where $\boldsymbol{0}_b$ is the b-dimensional vectors with all zeros, and the 'max' operation for a pair of vectors is applied in the element-wise manner. Thanks to this analytic-form expression, the computation of the uLSIF solution is fast and stable.

It was theoretically shown that uLSIF is superior in statistical convergence and numerical stability (Kanamori et al., 2008; Kanamori et al., 2009b).

5.3 Basis Function Design

The performance of uLSIF depends on the choice of the basis functions $\{\psi_\ell(\boldsymbol{u})\}_{\ell=1}^b$. As explained below, the use of Gaussian basis functions would be reasonable:

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

where $K_\sigma(\boldsymbol{u}, \boldsymbol{u}')$ is the Gaussian kernel with kernel width σ:

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

By definition, the density ratio $r(\boldsymbol{u})$ tends to take large values if $p_{\mathrm{de}}(\boldsymbol{u})$ is small and $p_{\mathrm{nu}}(\boldsymbol{u})$ is large; conversely, $r(\boldsymbol{u})$ tends to be small (i.e., close to zero) if $p_{\mathrm{de}}(\boldsymbol{u})$ is large and $p_{\mathrm{nu}}(\boldsymbol{u})$ is small. When a function is approximated by a Gaussian kernel model, many kernels may be needed in the region where the output of the target function is large; on the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero. Following this heuristic, many kernels are allocated in the region where $p_{\mathrm{nu}}(\boldsymbol{u})$ has large values, which may be approximately achieved by setting the Gaussian centers at $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$.

Alternatively, we may locate $(n_{\mathrm{de}} + n_{\mathrm{nu}})$ Gaussian kernels at both $\{\widehat{\boldsymbol{u}}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$. However, in our preliminary experiments, this did not further improve the performance, but slightly increased the computational cost. When n_{nu} is very large, just using all the test input points $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ as Gaussian centers is already computationally rather demanding. To ease this problem, we practically use a subset of $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ as Gaussian centers for computational efficiency, i.e.,

$$\widehat{r}(\boldsymbol{u}) = \sum_{\ell=1}^{b} \alpha_\ell K_\sigma(\boldsymbol{u}, \boldsymbol{c}_\ell),$$

where $\{\boldsymbol{c}_\ell\}_{\ell=1}^b$ are template points randomly chosen from $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ without replacement and b $(\leq n_{\mathrm{nu}})$ is a prefixed number; in the experiments, we set

$$b = \min(100, n_{\mathrm{nu}}).$$

5.4 Model Selection

The performance of uLSIF depends on the kernel width σ and the regularization parameter λ. Model selection of uLSIF is possible based on cross-validation (CV) with respect to the error criterion (6).

A significant advantage of uLSIF is that the score of leave-one-out CV (LOOCV) can be computed analytically—thanks to this property, the computational complexity for performing LOOCV is reduced to the same order of magnitude as just computing a single solution, which is explained below.

In the current setting, two sets of samples $\{\widehat{\boldsymbol{u}}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ are given, which generally have different sample size. For

$$n := \min(n_{\mathrm{de}}, n_{\mathrm{nu}}),$$

we hold out $\widehat{\boldsymbol{u}}_k^{\mathrm{de}}$ and $\widehat{\boldsymbol{u}}_k^{\mathrm{nu}}$ $(k = 1, 2, \ldots, n)$ at the same time in the LOOCV procedure. Note that this is only for the sake of simplicity—the combination of the samples can be arbitrarily permuted without sacrificing the computational advantages.

Let $\widehat{r}^{(k)}(\boldsymbol{u})$ be an estimate of the density ratio obtained without $\widehat{\boldsymbol{u}}_k^{\mathrm{de}}$ and $\widehat{\boldsymbol{u}}_k^{\mathrm{nu}}$. Then the LOOCV score is expressed as

$$\widehat{J}^{\mathrm{LOOCV}} = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1}{2}(\widehat{r}^{(k)}(\widehat{\boldsymbol{u}}_k^{\mathrm{de}}))^2 - \widehat{r}^{(k)}(\widehat{\boldsymbol{u}}_k^{\mathrm{nu}})\right].$$

Our approach to efficiently computing the LOOCV score is to use the *Sherman-Woodbury-Morrison* formula (Golub & Loan, 1996) for computing matrix inverses: for an invertible square matrix $\boldsymbol{A}$ and vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ such that $\boldsymbol{\eta}^\top \boldsymbol{A}^{-1}\boldsymbol{\xi} \neq -1$, the Sherman-Woodbury-Morrison formula states that

$$(\boldsymbol{A} + \boldsymbol{\xi}\boldsymbol{\eta}^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{\xi}\boldsymbol{\eta}^\top\boldsymbol{A}^{-1}}{1 + \boldsymbol{\eta}^\top\boldsymbol{A}^{-1}\boldsymbol{\xi}}.$$

A pseudo code of uLSIF with LOOCV-based model selection is summarized in Figure 3.

5.5 Discussion

Kernel density estimation (KDE) is efficient in computation since no optimization is involved, and model selection is possible by CV (Härdle et al., 2004). However, the use of KDE in density ratio estimation can be inaccurate since division by an estimated density tends to magnify the estimation error.

Kernel mean matching (KMM) (Huang et al., 2007) overcomes this problem by directly estimating the density ratio. The basic idea of KMM is to find $\widehat{r}(\boldsymbol{u})$ such that the mean discrepancy between nonlinearly transformed samples drawn from $p_{\mathrm{de}}(\boldsymbol{u})$ and $p_{\mathrm{nu}}(\boldsymbol{u})$ is minimized in a *universal reproducing kernel Hilbert space* (Steinwart, 2001); the Gaussian kernel is an example of kernels that induce a universal reproducing kernel Hilbert space. Within this formulation, density ratio estimates at the given data samples

Input: Two sets of samples $\{\widehat{\boldsymbol{u}}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$
Output: Density ratio estimate $\widehat{r}(\boldsymbol{u})$

$b \longleftarrow \min(100, n_{\mathrm{nu}})$; $\quad n \longleftarrow \min(n_{\mathrm{de}}, n_{\mathrm{nu}})$;
Randomly choose b centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^{b}$ from $\{\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ without replacement;
For each candidate of Gaussian width σ

$$\widehat{H}_{\ell,\ell'} \longleftarrow \frac{1}{n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}} \exp\left(-\frac{\|\widehat{\boldsymbol{u}}_i^{\mathrm{de}}-\boldsymbol{c}_\ell\|^2+\|\widehat{\boldsymbol{u}}_i^{\mathrm{de}}-\boldsymbol{c}_{\ell'}\|^2}{2\sigma^2}\right) \text{ for } \ell,\ell'=1,2,\ldots,b;$$

$$\widehat{h}_\ell \longleftarrow \frac{1}{n_{\mathrm{nu}}}\sum_{j=1}^{n_{\mathrm{nu}}} \exp\left(-\frac{\|\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}-\boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } \ell=1,2,\ldots,b;$$

$$X_{\ell,k}^{\mathrm{de}} \longleftarrow \exp\left(-\frac{||\widehat{\boldsymbol{u}}_k^{\mathrm{de}}-\boldsymbol{c}_\ell||^2}{2\sigma^2}\right) \text{ for } k=1,2,\ldots,n \text{ and } \ell=1,2,\ldots,b;$$

$$X_{\ell,k}^{\mathrm{nu}} \longleftarrow \exp\left(-\frac{||\widehat{\boldsymbol{u}}_k^{\mathrm{nu}}-\boldsymbol{c}_\ell||^2}{2\sigma^2}\right) \text{ for } k=1,2,\ldots,n \text{ and } \ell=1,2,\ldots,b;$$

For each candidate of regularization parameter λ

$$\widehat{\boldsymbol{G}} \longleftarrow \widehat{\boldsymbol{H}} + \frac{\lambda(n_{\mathrm{de}}-1)}{n_{\mathrm{de}}}\boldsymbol{I}_b;$$

$$\boldsymbol{F} \longleftarrow \widehat{\boldsymbol{G}}^{-1}\boldsymbol{X}^{\mathrm{de}};$$

$$\boldsymbol{C} \longleftarrow \max\left(\boldsymbol{O}_{b\times n}, \frac{n_{\mathrm{de}}-1}{n_{\mathrm{de}}(n_{\mathrm{nu}}-1)}\left[n_{\mathrm{nu}}\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}}\boldsymbol{1}_n^\top - \widehat{\boldsymbol{G}}^{-1}\boldsymbol{X}^{\mathrm{nu}} + \boldsymbol{F}\,\mathrm{diag}\left(\frac{n_{\mathrm{nu}}\widehat{\boldsymbol{h}}^\top\boldsymbol{F}-\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{nu}}*\boldsymbol{F})}{n_{\mathrm{de}}\boldsymbol{1}_n^\top-\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{de}}*\boldsymbol{F})}\right)\right]\right);$$

$$\widehat{J}^{\mathrm{LOOCV}}(\sigma,\lambda) \longleftarrow \frac{\|(\boldsymbol{X}^{\mathrm{de}}*\boldsymbol{C})^\top\boldsymbol{1}_b\|^2}{2n} - \frac{\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{nu}}*\boldsymbol{C})\boldsymbol{1}_n}{n};$$

end
end

$$(\widehat{\sigma},\widehat{\lambda}) \longleftarrow \mathrm{argmin}_{(\sigma,\lambda)}\,\widehat{J}^{\mathrm{LOOCV}}(\sigma,\lambda);$$

$$\widetilde{H}_{\ell,\ell'} \longleftarrow \frac{1}{n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}} \exp\left(-\frac{\|\widehat{\boldsymbol{u}}_i^{\mathrm{de}}-\boldsymbol{c}_\ell\|^2+\|\widehat{\boldsymbol{u}}_i^{\mathrm{de}}-\boldsymbol{c}_{\ell'}\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell,\ell'=1,2,\ldots,b;$$

$$\widetilde{h}_\ell \longleftarrow \frac{1}{n_{\mathrm{nu}}}\sum_{j=1}^{n_{\mathrm{nu}}} \exp\left(-\frac{\|\widehat{\boldsymbol{u}}_j^{\mathrm{nu}}-\boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell=1,2,\ldots,b;$$

$$\widehat{\boldsymbol{\alpha}} \longleftarrow \max(\boldsymbol{0}_b, (\widetilde{\boldsymbol{H}}+\widehat{\lambda}\boldsymbol{I}_b)^{-1}\widetilde{\boldsymbol{h}});$$

$$\widehat{r}(\boldsymbol{u}) \longleftarrow \sum_{\ell=1}^{b}\widehat{\alpha}_\ell \exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right);$$

Figure 3: Pseudo code of uLSIF with LOOCV. $\boldsymbol{R}*\boldsymbol{R}'$ denotes the element-wise multiplication of matrices $\boldsymbol{R}$ and $\boldsymbol{R}'$ of the same size. For n-dimensional vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$, $\mathrm{diag}\left(\frac{\boldsymbol{r}}{\boldsymbol{r}'}\right)$ denotes the $n\times n$ diagonal matrix with i-th diagonal element r_i/r_i'. $\boldsymbol{1}_b$ denotes the b-dimensional vectors with all ones. $\boldsymbol{O}_{b\times n}$ denotes the $b\times n$ matrix with all zero. A MATLAB implementation of uLSIF is available from '`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/uLSIF/`'.

can be directly obtained by solving a convex quadratic programming problem. However, there is no objective model selection method and therefore model parameters such as the Gaussian width and the regularization parameter need to be chosen by hand, which is highly unreliable unless we have strong prior knowledge. Furthermore, the computation of KMM is rather expensive since a quadratic programming problem has to be solved.

Another approach to directly estimating the density ratio is to use a probabilistic classifier discriminating samples drawn from $p_{\mathrm{de}}(\boldsymbol{u})$ and samples drawn from $p_{\mathrm{nu}}(\boldsymbol{u})$—a kernelized variant of *logistic regression* (LogReg) is a suitable classifier for this purpose. It is known that the likelihood function of the kernel LogReg classifier is concave, so the regularized maximum likelihood solution with a convex regularizer can be uniquely obtained (Koh et al., 2007; Minka, 2007). An advantage of the LogReg method is that model selection (i.e., the choice of the kernel bandwidth as well as the regularization parameter) is possible by standard CV since the learning problem involved above is a standard supervised classification problem. However, LogReg is computationally rather expensive since a non-linear optimization problem has to be solved.

The *Kullback-Leibler importance estimation procedure* (KLIEP) (Sugiyama et al., 2008) also directly gives an estimate of the density-ratio function without going through density estimation by matching the two distributions in terms of the Kullback-Leibler divergence (Kullback & Leibler, 1951). The optimization criterion involved in KLIEP is a constrained non-linear optimization problem, which is convex. Thus, the global solution—which tends to be sparse—can be obtained, e.g., by simply performing gradient ascent and feasibility satisfaction iteratively. CV is also available for KLIEP and therefore model selection is possible, which is an advantage over KMM. However, the optimization problem involved in KLIEP is still computationally rather expensive.

Least-squares importance fitting (LSIF) (Kanamori et al., 2009a) is similar to KLIEP, but the squared loss is used instead of the log loss. The LSIF optimization problem is a convex quadratic programming problem and therefore a standard optimization software can be used for obtaining the global solution. CV is available, similar to LogReg and KLIEP. Furthermore, a notable advantage of LSIF is that a regularization-path tracking algorithm is available (cf. Best, 1982; Efron et al., 2004; Hastie et al., 2004). Thus, solutions for all regularization parameter values can be computed efficiently. In fact, the regularization-path tracking algorithm shows that a quadratic programming solver is no longer needed for obtaining the LSIF solution—just computing matrix inverses is enough. This highly contributes to saving the computation time. However, regularization-path tracking sometimes suffers from a numerical problem and therefore is not practically reliable.

uLSIF inherits good properties of the above methods, e.g., no density estimation is involved and a built-in model selection method is available. In addition to these preferable properties, the solution of uLSIF can be computed analytically through matrix inversion and therefore uLSIF is computationally very efficient and numerically stable. Furthermore, thanks to the availability of the closed-form solution of uLSIF, the LOOCV score can be analytically computed without repeating hold-out loops, which highly contributes to reducing the computation time in the model selection phase.

Thus, for developing a density-ratio estimation method that is flexible and computationally efficient, the use of uLSIF would be suitable.

6 Estimating the Dimension of the Hetero-distributional Subspace

So far, we explained how the dimensionality reduction idea could be incorporated into density ratio estimation, given that the dimension m of the hetero-distributional subspace is known in advance. In this section, we address how the dimension m of could be estimated from samples, which results in a practical procedure.

One possibility would be to compute the normalized LFDA criterion

$$\frac{1}{m'}\mathrm{tr}(\widehat{\boldsymbol{U}}\boldsymbol{S}^{\mathrm{lb}}\widehat{\boldsymbol{U}}^{\top}(\widehat{\boldsymbol{U}}\boldsymbol{S}^{\mathrm{lw}}\widehat{\boldsymbol{U}}^{\top})^{-1})$$

as a function of reduced dimensionality m' and choose the value of $\widehat{m}$ that maximizes the above criterion. However, in our preliminary experiments, this did not perform well since the above criterion measures the estimation error of the hetero-distributional subspace, not the estimation error of the density ratio.

A more sensible approach would be to use the LOOCV score of the uLSIF algorithm

$$\widehat{J}^{\mathrm{LOOCV}} = \frac{1}{\min(n_{\mathrm{de}}, n_{\mathrm{nu}})} \sum_{k=1}^{\min(n_{\mathrm{de}}, n_{\mathrm{nu}})} \left[\frac{1}{2}(\widehat{r}^{(k)}(\widehat{\boldsymbol{u}}_k^{\mathrm{de}}))^2 - \widehat{r}^{(k)}(\widehat{\boldsymbol{u}}_k^{\mathrm{nu}})\right]$$

as a function of the reduced dimensionality m' and choose the value of $\widehat{m}$ that minimizes the LOOCV score.

Thus our density ratio estimation procedure effectively combines LFDA and uLSIF. We refer to the proposed procedure as *Direct Density-ratio estimation with Dimensionality reduction* (D^3). The pseudo code of the entire algorithm of D^3 is summarized in Figure 4.

7 Numerical Examples

In this section, we investigate the experimental performance of the proposed method.

7.1 Illustrative Example

First, we illustrate how the proposed D^3 algorithm behaves.

Let the input domain be $\mathcal{D} = \mathbb{R}^2$ (i.e., $d = 2$) and the denominator and numerator densities are set as

$$\boldsymbol{x}^{\mathrm{de}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 1\end{pmatrix}\right),$$
$$\boldsymbol{x}^{\mathrm{nu}} \sim \frac{1}{2}N\left(\begin{pmatrix}-3\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) + \frac{1}{2}N\left(\begin{pmatrix}3\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),$$

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

Obtain orthonormal basis $\{\widetilde{\boldsymbol{\varphi}}_l\}_{l=1}^{d}$ using LFDA with $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$;
For each reduced dimension $m = 1, 2, \ldots, d$
  Form projection matrix: $\widehat{\boldsymbol{U}}_m = (\widetilde{\boldsymbol{\varphi}}_1 | \widetilde{\boldsymbol{\varphi}}_2 | \cdots | \widetilde{\boldsymbol{\varphi}}_m)^\top$;
  Project samples: $\{\widehat{\boldsymbol{u}}_{i,m}^{\mathrm{de}} \mid \widehat{\boldsymbol{u}}_{i,m}^{\mathrm{de}} = \widehat{\boldsymbol{U}}_m \boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_{j,m}^{\mathrm{nu}} \mid \widehat{\boldsymbol{u}}_{j,m}^{\mathrm{nu}} = \widehat{\boldsymbol{U}}_m \boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$;
  For each candidate of Gaussian width σ
    For each candidate of regularization parameter λ
      Compute LOOCV score $\widehat{J}^{\mathrm{LOOCV}}(m, \sigma, \lambda)$ using $\{\widehat{\boldsymbol{u}}_{i,m}^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_{j,m}^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$;
    end
  end
end
Choose the best model: $(\widehat{m}, \widehat{\sigma}, \widehat{\lambda}) \longleftarrow \mathrm{argmin}_{(m,\sigma,\lambda)} \widehat{J}^{\mathrm{LOOCV}}(m, \sigma, \lambda)$;
Estimate density ratio from $\{\widehat{\boldsymbol{u}}_{i,\widehat{m}}^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\widehat{\boldsymbol{u}}_{j,\widehat{m}}^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ using uLSIF with $(\widehat{\sigma}, \widehat{\lambda})$;

Figure 4: Pseudo code of Direct Density-ratio estimation with Dimensionality reduction (D^3).

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multi-variate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The profile of the densities and their ratio are illustrated in Figure 5 and Figure 8(a). We sample $n_{\mathrm{de}} = 100$ points from $p_{\mathrm{de}}(\boldsymbol{x})$ and $n_{\mathrm{nu}} = 100$ points from $p_{\mathrm{nu}}(\boldsymbol{x})$, respectively; the samples are illustrated in Figure 6. In this dataset, the distributions are different only in the one-dimensional subspace spanned by $(1, 0)^\top$, i.e., $m = 1$. The true hetero-distributional subspace is depicted by the solid line in Figure 6.

The dotted line in Figure 6 depicts the hetero-distributional subspace estimated by LFDA with reduced dimensionality $m' = 1$; when $m' = 2$, LFDA merely gives the entire space. This shows that for $m' = 1$, LFDA gives a very good estimate of the true hetero-distributional subspace.

Next, we chose reduced dimensionality m as well as the Gaussian width σ and the regularization parameter λ in uLSIF. Figure 7 depicts the LOOCV score of uLSIF, showing that

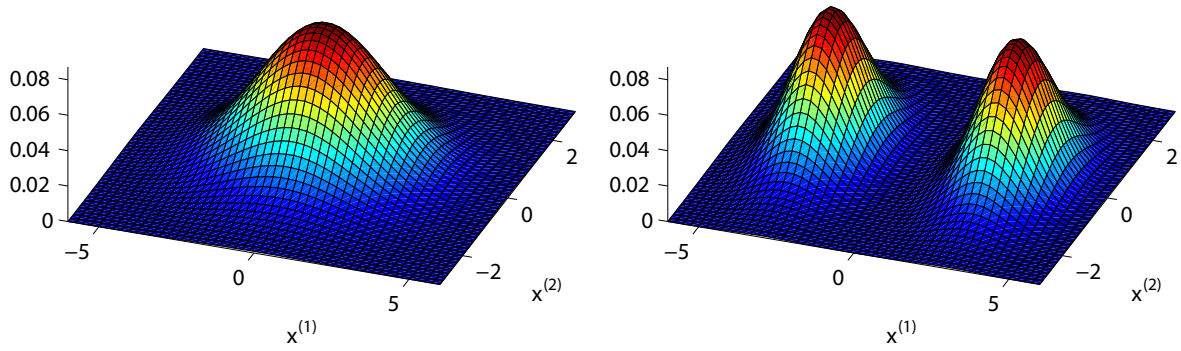
$$(\widehat{m}, \widehat{\sigma}, \widehat{\lambda}) = (1, 1, 10^{-0.5})$$

is the minimizer.

Finally, the density ratio is estimated by uLSIF. Figure 8 depicts the true density ratio, its estimate by uLSIF without dimensionality reduction, and its estimate by uLSIF with dimensionality reduction by LFDA. For uLSIF without dimensionality reduction,

$$(\widehat{\sigma}, \widehat{\lambda}) = (1, 10^{-0.5})$$

is chosen by LOOCV (see Figure 7 with $m' = 2$). This shows that when dimensionality reduction is not performed, independence between $r(\boldsymbol{x})$ and the second element $x^{(2)}$ (Figure 8(a)) is not captured and the estimated density ratio has Gaussian-structure along



(a) Profile of $p_{\mathrm{de}}(\boldsymbol{x})$ (b) Profile of $p_{\mathrm{nu}}(\boldsymbol{x})$

Figure 5: Two-dimensional toy dataset.

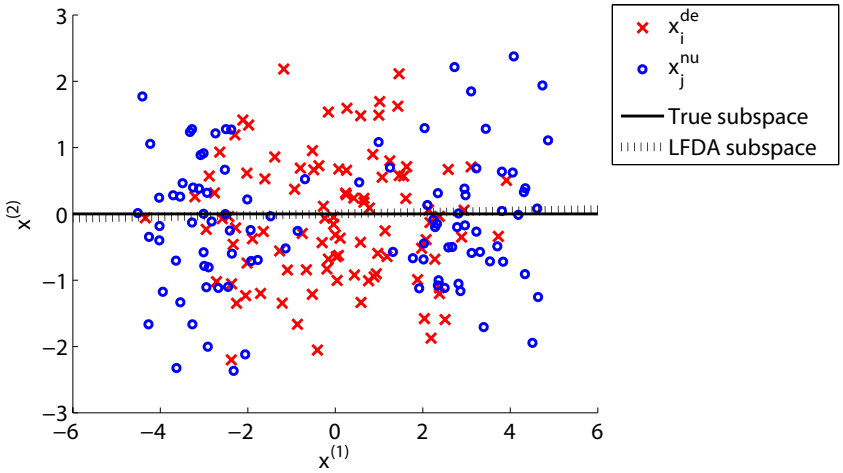


Figure 6: Samples and hetero-distributional subspace of two-dimensional toy dataset. The LFDA estimate of the hetero-distributional subspace is spanned by $(1.00, 0.01)^\top$, which is very close to the true hetero-distributional subspace spanned by $(1, 0)^\top$.

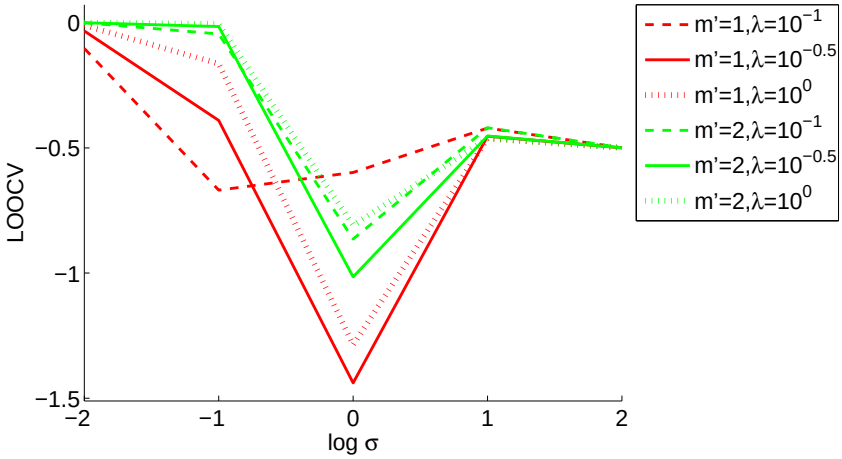
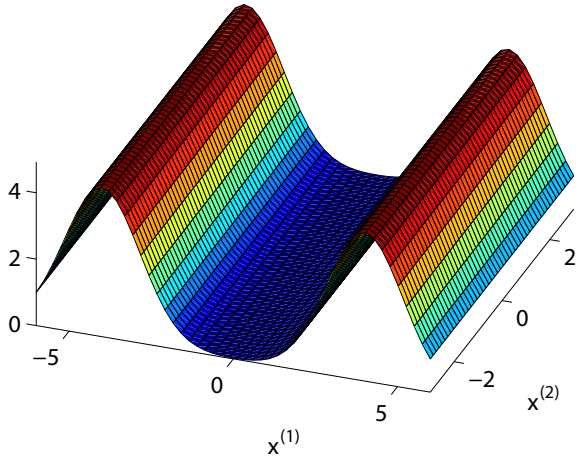
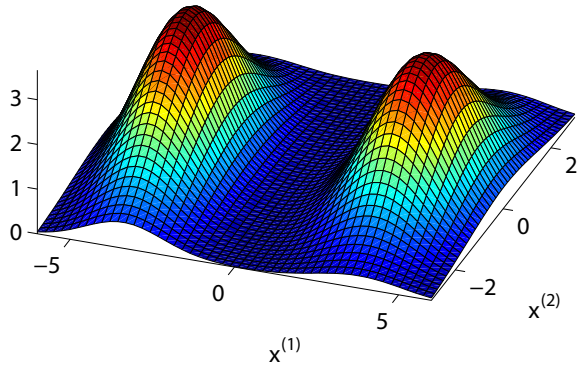


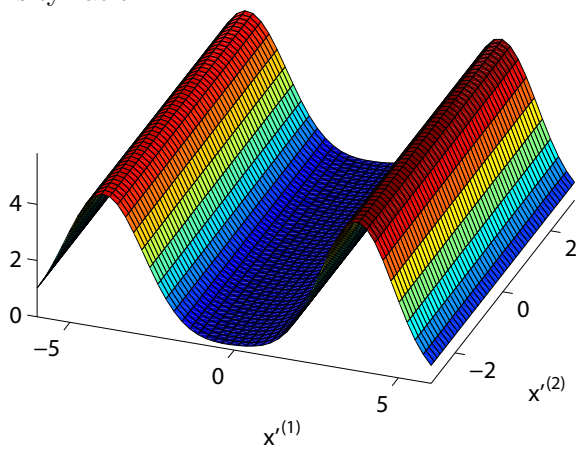
Figure 7: LOOCV score of uLSIF for two-dimensional toy dataset.



(a) True density-ratio



(b) Density-ratio estimation without dimensionality reduction. NMSE = 1.52×10^{-5}.



(c) Density-ratio estimation with dimensionality reduction. NMSE = 0.89×10^{-5}.

Figure 8: True and estimated density ratio functions. $x'^{(1)}$ and $x'^{(2)}$ in (c) denotes the LFDA solution, i.e., $x'^{(1)} = 1.00x^{(1)} + 0.01x^{(2)}$ and $x'^{(2)} = -0.01x^{(1)} + 1.00x^{(2)}$, respectively.

$x^{(2)}$ (Figure 8(b)). On the other hand, when dimensionality reduction is carried out, independence between $r(\boldsymbol{x})$ and $x^{(2)}$ can be successfully captured and a good result is obtained (Figure 8(c)).

The accuracy of an estimated density ratio is measured by the *normalized mean squared error* (NMSE):

$$\mathrm{NMSE} = \sum_{i=1}^{n_{\mathrm{de}}} \left(\frac{\widehat{r}(\boldsymbol{x}_i^{\mathrm{de}})}{\sum_{i'=1}^{n_{\mathrm{de}}} \widehat{r}(\boldsymbol{x}_{i'}^{\mathrm{de}})} - \frac{r(\boldsymbol{x}_i^{\mathrm{de}})}{\sum_{i'=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}_{i'}^{\mathrm{de}})} \right)^2. \tag{8}$$

By dimensionality reduction, NMSE is reduced from 1.52×10^{-5} to 0.89×10^{-5}; thus we gain 41.5% reduction in NMSE.

7.2 Performance Comparison using Artificial Datasets

Here, we investigate the performance of the D^3 algorithm using 6 artificial datasets. The input domain of the datasets is d-dimensional ($d \geq 2$) and the true dimensionality of the hetero-distributional subspace is $m = 1$ or 2. The equi-distributional component of the datasets is the $(d-m)$-dimensional Gaussian distribution with mean zero and covariance identity. The hetero-distributional component of each dataset is given as follows:

(a) Dataset 1 ('shifting', $m = 1$):

$$\boldsymbol{u}^{\mathrm{de}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),$$
$$\boldsymbol{u}^{\mathrm{nu}} \sim N\left(\begin{pmatrix}3\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right).$$

(b) Dataset 2 ('shrinking', $m = 2$):

$$\boldsymbol{u}^{\mathrm{de}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 4\end{pmatrix}\right),$$
$$\boldsymbol{u}^{\mathrm{nu}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right).$$

(c) Dataset 3 ('magnifying', $m = 2$):

$$\boldsymbol{u}^{\mathrm{de}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right),$$
$$\boldsymbol{u}^{\mathrm{nu}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 4\end{pmatrix}\right).$$

(d) Dataset 4 ('rotating', $m = 2$):

$$\boldsymbol{u}^{\mathrm{de}} \sim N\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 1\end{pmatrix}\right),$$
$$\boldsymbol{u}^{\mathrm{nu}} \sim N\left(\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}, \begin{pmatrix}5/2 & 3/2\\3/2 & 5/2\end{pmatrix}\right).$$

(e) Dataset 5 ('1-dimensional splitting', $m = 1$):

$$\boldsymbol{u}^{\mathrm{de}} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),$$
$$\boldsymbol{u}^{\mathrm{nu}} \sim \frac{1}{2}N\left(\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right) + \frac{1}{2}N\left(\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right).$$

(f) Dataset 6 ('2-dimensional splitting', $m = 2$):

$$\begin{aligned}
\boldsymbol{u}^{\mathrm{de}} &\sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),\\
\boldsymbol{u}^{\mathrm{nu}} &\sim \frac{1}{4}N\left(\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right) + \frac{1}{4}N\left(\begin{pmatrix}-2\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)\\
&\quad + \frac{1}{4}N\left(\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right) + \frac{1}{4}N\left(\begin{pmatrix}0\\-2\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right).
\end{aligned}$$

The number of samples is set as $n_{\mathrm{de}} = 200$ and $n_{\mathrm{nu}} = 1000$ for all the datasets. Examples of realized samples are illustrated in Figure 9. For each dimensionality $d = 2, 3, \ldots, 10$, the density ratio is estimated using the proposed method and the baseline method (uLSIF without dimensionality reduction). This experiment is repeated 100 times for each d with different random seed.

Figure 10 depicts choice of the dimensionality of the hetero-distributional subspace by LOOCV for each d. This shows that for the datasets 1, 2, 4, and 5, dimensionality choice by LOOCV works well. For the dataset 3, $\widehat{m} = 1$ is always chosen although the true dimensionality is $m = 2$. For the dataset 6, dimensionality choice is rather unstable, but still it works reasonably well.

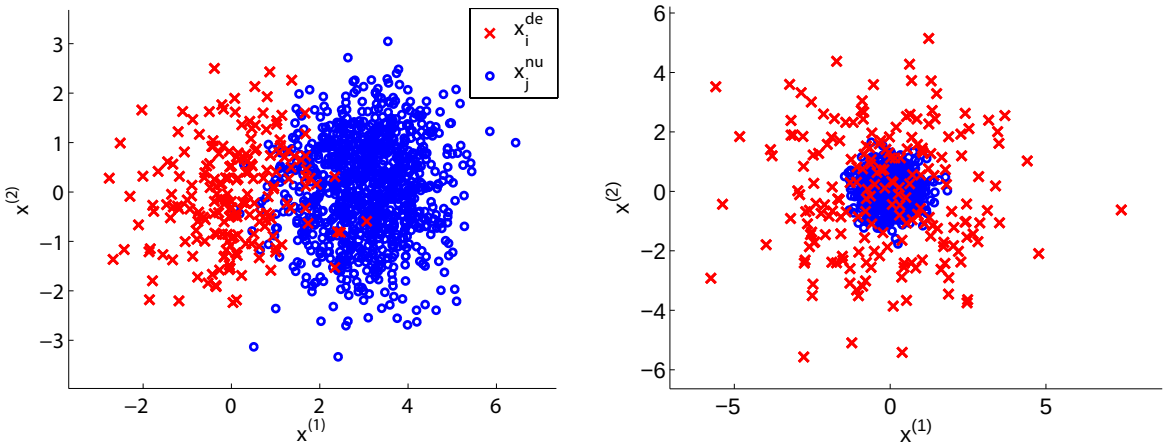
Figure 11 depicts the value of NMSE (see Eq.(8)) averaged over 100 trials. For each d, the *t-test* (see e.g., Henkel, 1979) at the significance level 5% is performed and the best method as well as the comparable method in terms of mean NMSE are indicated by '$\times$' (in other words, the method without the symbol '$\times$' is significantly worse than the other method). This shows that mean NMSE of the baseline method (no dimensionality reduction) tends to grow rapidly as the dimensionality d increases. On the other hand, increase of mean NMSE of the proposed D^3 algorithm is much smaller than that of the baseline method. Consequently, mean NMSE of D^3 is much smaller than that of the baseline method when the input dimensionality d is large. The difference of mean NMSE is statistically significant for the datasets 1, 2, 5, and 6.

From the above experiments, we experimentally confirmed that the proposed dimensionality scheme is useful in density ratio estimation.

7.3 Application to Inlier-based Outlier Detection

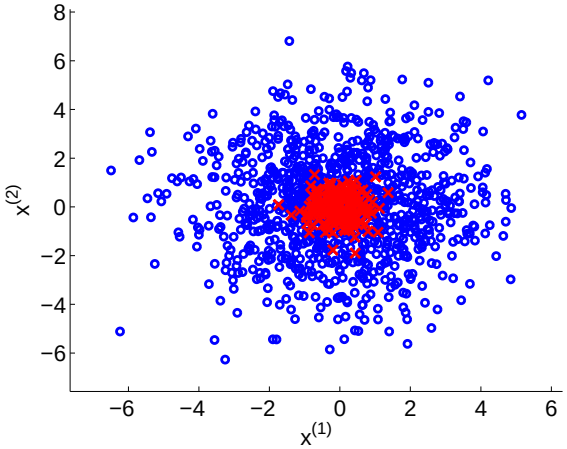
Finally, we apply the proposed method to inlier-based outlier detection (see Section 2.2.2).

We compare three schemes here—uLSIF with no dimensionality reduction, uLSIF with dimensionality reduction by LFDA (i.e., the proposed D^3 algorithm), and uLSIF with dimensionality reduction by *principal component analysis* (PCA) (Jolliffe, 1986). The datasets provided by IDA (Rätsch et al., 2001) are used for performance evaluation; we exclude the 'splice' dataset since it is discrete. The datasets are binary classification and each dataset consists of positive/negative and training/test samples for 20 or 100 trials. We use all positive training samples as inliers in the 'model' set, while we use all positive test samples as inliers and the first 5% of negative test samples as outliers in the

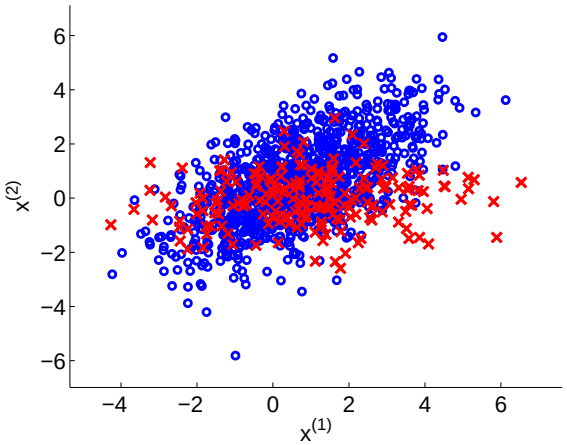


(a) Dataset 1 ($m = 1$)

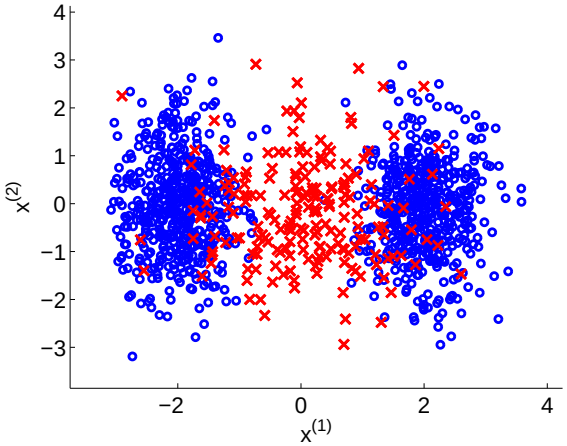
(b) Dataset 2 ($m = 2$)



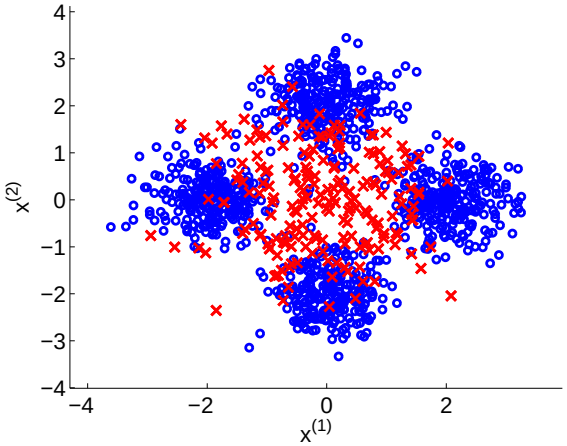
(c) Dataset 3 ($m = 2$)



(d) Dataset 4 ($m = 2$)

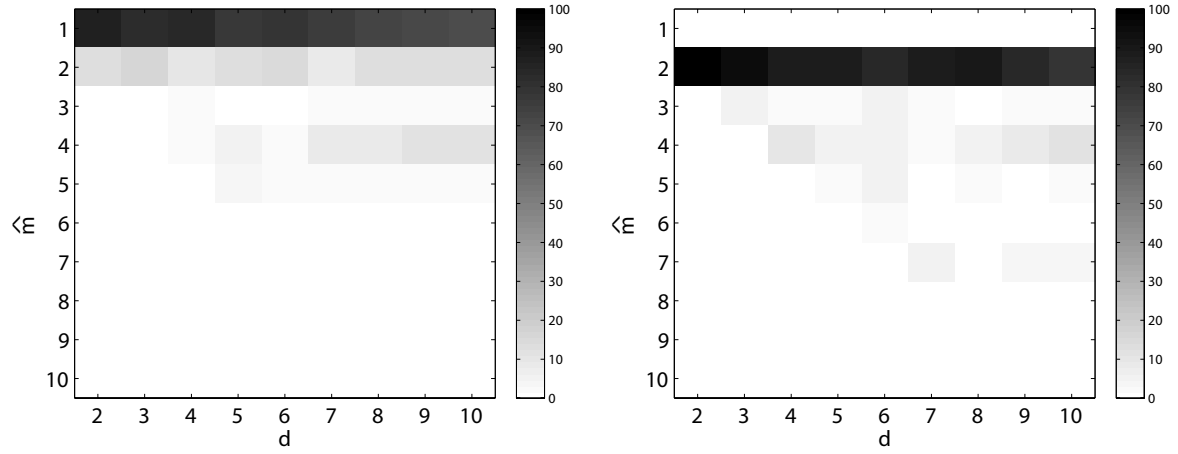


(e) Dataset 5 ($m = 1$)



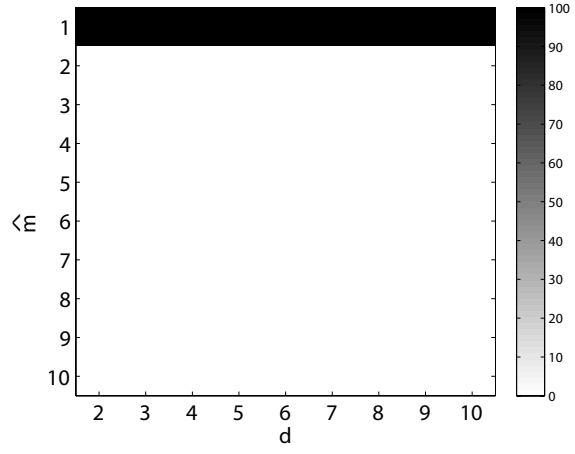
(f) Dataset 6 ($m = 2$)

Figure 9: Artificial datasets.

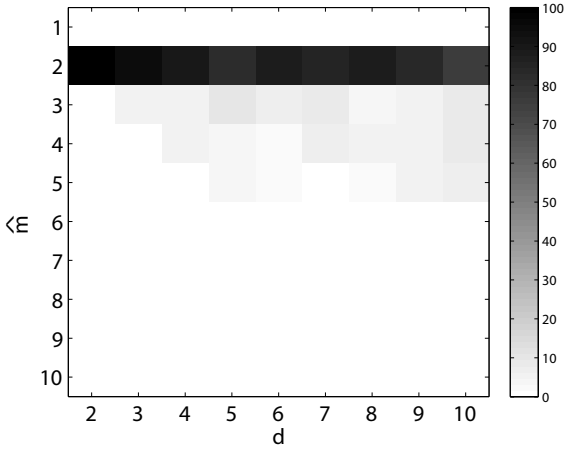


(a) Dataset 1 ($m = 1$)

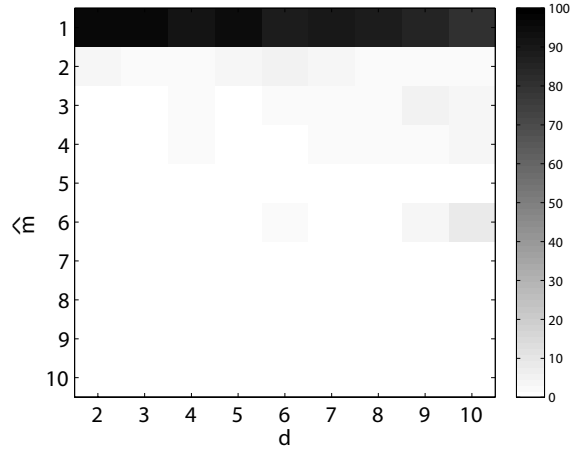
(b) Dataset 2 ($m = 2$)



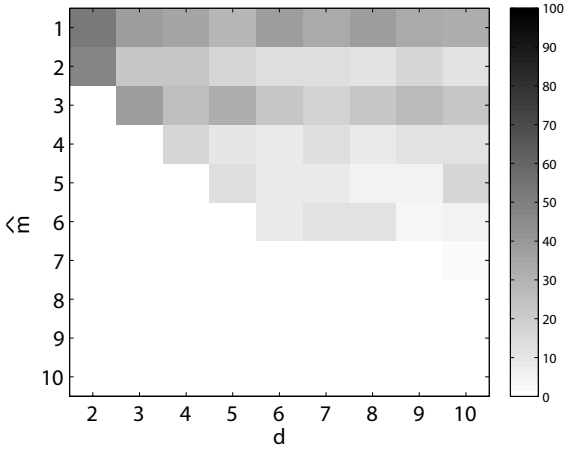
(c) Dataset 3 ($m = 2$)



(d) Dataset 4 ($m = 2$)

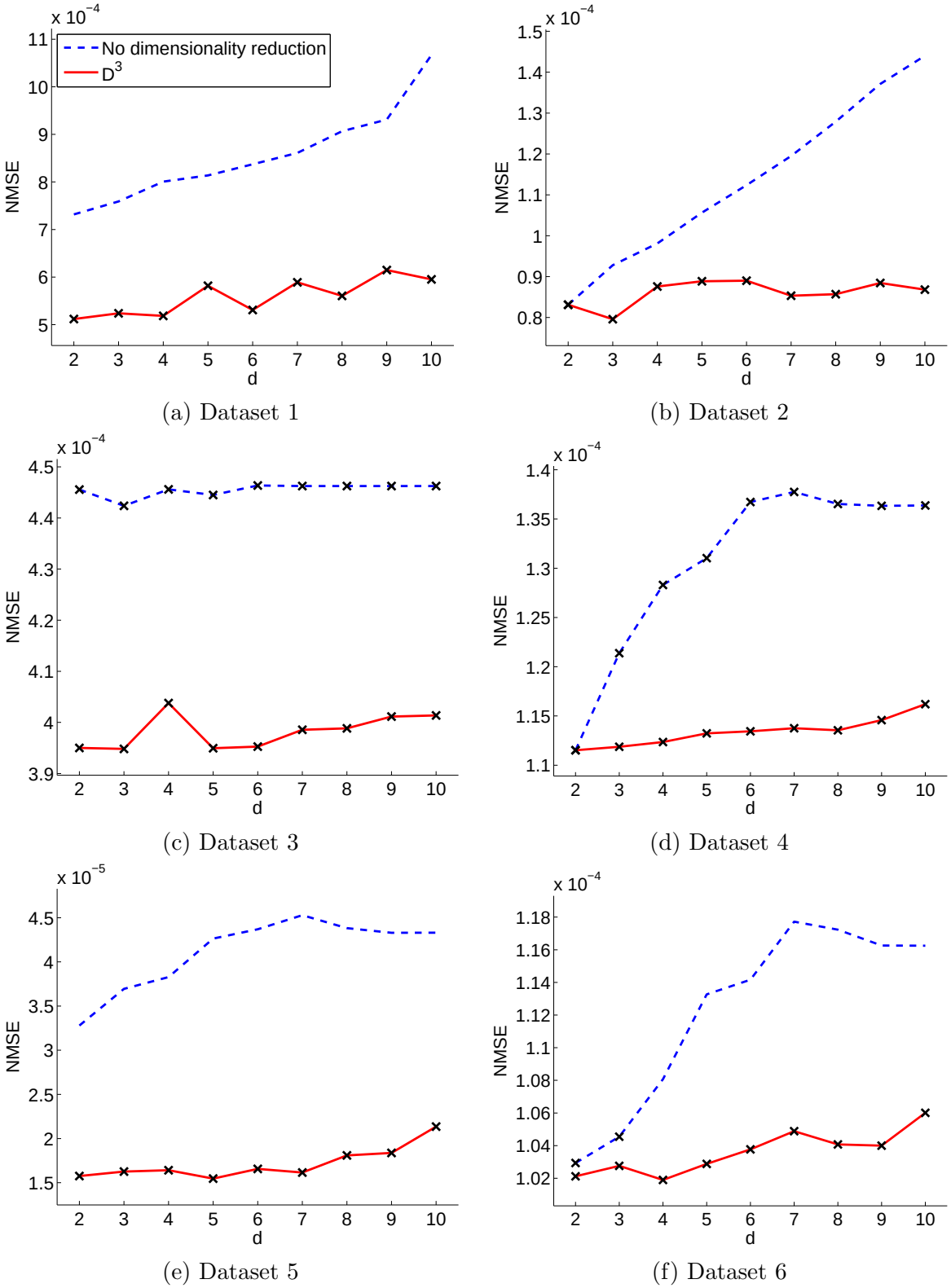


(e) Dataset 5 ($m = 1$)



(f) Dataset 6 ($m = 2$)

Figure 10: Dimension choice of hetero-distributional subspace by LOOCV.



(a) Dataset 1 (b) Dataset 2

(c) Dataset 3 (d) Dataset 4

(e) Dataset 5 (f) Dataset 6

Figure 11: Mean NMSE of the estimated density ratio functions. For each d, the t-test at the significance level 5% is performed and the best method as well as the comparable method in terms of mean NMSE are indicated by '×'.

Table 1: Mean AUC values and chosen dimensionality for outlier detection over 20 trials for the IDA datasets. The numbers in brackets are standard deviations. The best method in terms of the mean AUC value and comparable methods according to the t-test at the significance level 5% are specified by '$\circ$'.

		Plain uLSIF		LFDA+uLSIF (D^3)		PCA+uLSIF	
Dataset	d	AUC	$\widehat{m}$	AUC	$\widehat{m}$	AUC	$\widehat{m}$
banana	2	$\circ$0.686(0.108)	2(0)	$\circ$0.699(0.091)	1.9(0.3)	$\circ$0.644(0.136)	1.7(0.5)
b-cancer	9	0.665(0.099)	9(0)	$\circ$0.743(0.083)	3.1(2.2)	0.628(0.117)	3.5(2.7)
diabetes	8	0.621(0.079)	8(0)	$\circ$0.698(0.036)	3.6(1.7)	0.648(0.069)	2.9(2.2)
f-solar	9	0.393(0.044)	9(0)	$\circ$0.636(0.076)	1.9(1.4)	0.492(0.076)	3.0(2.1)
german	20	$\circ$0.629(0.060)	20(0)	$\circ$0.660(0.034)	8.4(3.8)	0.586(0.064)	8.9(7.6)
heart	13	0.839(0.063)	13(0)	$\circ$0.886(0.043)	2.9(2.3)	0.804(0.114)	3.2(3.8)
image	18	$\circ$0.621(0.058)	18(0)	$\circ$0.624(0.072)	6.1(2.1)	$\circ$0.614(0.095)	3.5(6.0)
thyroid	5	0.407(0.269)	5(0)	$\circ$0.824(0.077)	1.9(1.3)	0.708(0.202)	2.3(1.3)
titanic	3	0.595(0.102)	3(0)	$\circ$0.686(0.025)	1.5(0.8)	0.570(0.138)	1.5(0.7)
twonorm	20	0.934(0.013)	20(0)	0.913(0.044)	16.4(7.5)	$\circ$0.970(0.022)	8.5(6.7)
waveform	21	0.907(0.031)	21(0)	$\circ$0.930(0.013)	13.9(9.0)	$\circ$0.895(0.077)	5.6(3.9)
Average		0.663(0.084)	11.6(0)	0.754(0.054)	5.6(3.0)	0.687(0.101)	4.0(3.4)

'evaluation' set. Thus, we regard the positive samples as inliers and the negative samples as outliers.

In the evaluation of the performance of outlier detection methods, it is important to take into account both the detection rate (the amount of true outliers an outlier detection algorithm can find) and the detection accuracy (the amount of true inliers that an outlier detection algorithm misjudges as outliers). Since there is a trade-off between the detection rate and the detection accuracy, we adopt the area under the ROC curve (AUC) as our error metric (Bradley, 1997).

Note that a similar experiment has been carried out in Hido et al. (2008), where outlier samples are chosen *randomly* from the negative test samples. Since the choice of a small number of outliers significantly affect the AUC values in outlier detection experiments, we decided to choose outlier samples *deterministically* for obtaining reproducible experimental results. For this reason, our results given here are not necessarily the same as the numbers provided in Hido et al. (2008).

The mean and standard deviation of AUC values and chosen dimensionalities over 20 trials are summarized in Table 1, where the best method in terms of the mean AUC value and comparable methods according to the t-test at the significance level 5% are specified by '$\circ$'. The table shows that the proposed D^3 algorithm tends to outperform the baseline method (no dimensionality reduction). As a result, the average AUC value over all datasets is improved from 0.663 to 0.754 (we gain 13.7% increase in the mean AUC value). Dimensionality reduction by PCA also performs well for some datasets, but it tends to be outperformed by the baseline method for the other datasets. Consequently, the average AUC value over all datasets is slightly improved from 0.663 to 0.687, although this is

behind the proposed D^3 algorithm. Furthermore, the standard deviation of 'uLSIF+PCA' tends to be larger than the other methods, which would be due to the unsupervised nature of PCA. Thus PCA may not be reliable in practice.

Overall, the above experimental results show that the proposed D^3 algorithm is useful in inlier-based outlier detection scenarios.

8 Conclusions

The ratio of two probability density functions could be used for various data processing tasks, so it is important to estimate the density ratio accurately. In this paper, we proposed a new algorithm called *Direct Density-ratio estimation with Dimensionality reduction* (D^3) for improving density-ratio estimation accuracy in high-dimensional cases. The proposed method combines the supervised dimensionality reduction method called *local Fisher discriminant analysis* (LFDA) with the direct density-ratio estimation method called *unconstrained least-squares importance fitting* (uLSIF). The experiments showed that the proposed method compares favorably with the baseline method.

We chose LFDA in this paper since it was shown to be the most suitable in accuracy and computational efficiency (Sugiyama, 2007). On the other hand, supervised dimensionality reduction is one of the most active research topics and better methods will be developed in the future. Our framework allows us to use *any* supervised dimensionality reduction method for density ratio estimation. Thus, provided better methods of supervised dimensionality reduction, it is interesting to incorporate the new methods in the density ratio estimation framework and evaluate how accuracy and computational efficiency are improved.

In our experiments, choosing the dimensionality of hetero-distributional subspaces was slightly unstable in some cases, although the proposed method was no worse than the baseline method. Thus there is still room for further improvement in the choice of the dimensionality of hetero-distributional subspace, which needs to be further investigated in the future work.

Our formulation assumed that the components inside and outside the hetero-distributional subspace are statistically independent. A possible generalization of the proposed approach would be to weaken this condition, for example, following the line of Fukumizu et al. (2004) or Suzuki and Sugiyama (2009b). We focused on a linear hetero-distributional subspace, but we may consider a non-linear manifold and use a kernelized version of LFDA (Sugiyama, 2007). This would be a possible future direction to pursue.

Acknowledgments

We would like to thank Klaus-Robert Müller and Paul von Bünau for their fruitful comments, and Shohei Hido for sharing his simulation code with us. This work has been supported by MEXT Grant-in-Aid for Young Scientists (A) 20680007, SCAT, and AOARD.

References

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse.* New York and London: Academic Press.

Anderson, N., Hall, P., & Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.

Best, M. J. (1982). *An algorithm for the solution of the parametric quadratic programming problem* (Technical Report 82-24). Faculty of Mathematics, University of Waterloo.

Biau, G., & Gyorfi, L. (2005). On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, *51*, 3965–3973.

Bickel, P. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, *40*, 1–23.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for HIV therapy screening. *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)* (pp. 56–63). Helsinki, Finland: Omnipress.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning* (pp. 81–88).

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, *22*, e49–e57.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data.*

Brodsky, B., & Darkhovsky, B. (1993). *Nonparametric methods in change-point problems.* Dordrecht: Kluwer Academic Publishers.

Cheng, K. F., & Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*, 583–604.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York, NY, USA: John Wiley & Sons, Inc.

Duda, R. O., Hart, P. E., & Stor, D. G. (2001). *Pattern classification*. New York: Wiley.

Efron, B., Hastie, T., Tibshirani, R., & Johnstone, I. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Fishman, G. S. (1996). *Monte Carlo: Concepts, algorithms, and applications*. Berlin: Springer-Verlag.

Friedman, J., & Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, *7*, 697–717.

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

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Boston: Academic Press, Inc. Second edition.

Globerson, A., & Roweis, S. (2006). Metric learning by collapsing classes. *Advances in Neural Information Processing Systems 18* (pp. 451–458). Cambridge, MA: MIT Press.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems 17* (pp. 513–520). Cambridge, MA: MIT Press.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Gretton, A., Borgwardt, K. M., M.Rasch, Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19*, 513–520. Cambridge, MA: MIT Press.

Hachiya, H., Akiyama, T., Sugiyama, M., & Peters, J. (2009). Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*. to appear.

Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in highdimensional settings. *Biometrika*, *89*, 359–374.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, *5*, 1391–1415.

Hastie, T., & Tibshirani, R. (1996a). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*, 607–615.

Hastie, T., & Tibshirani, R. (1996b). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, *58*, 155–176.

Henkel, R. E. (1979). *Tests of significance*. Beverly Hills: SAGE Publication.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2008). Inlier-based outlier detection via direct density ratio estimation. *Proceedings of IEEE International Conference on Data Mining (ICDM2008)* (pp. 223–232). Pisa, Italy.

Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability* (pp. 23–41). Berkeley, CA., USA: University of California Press.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19*, 601–608. Cambridge, MA: MIT Press.

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

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Kanamori, T., Hido, S., & Sugiyama, M. (2008). *A least-squares approach to direct importance estimation* (Technical Report TR08-0003). Department of Computer Science, Tokyo Institute of Technology.

Kanamori, T., Hido, S., & Sugiyama, M. (2009a). Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. *Advances in Neural Information Processing Systems 21* (pp. 809–816). Cambridge, MA: MIT Press.

Kanamori, T., & Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, *116*, 149–162.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2009b). *Condition number analysis of kernel-based density ratio estimation* (Technical Report TR09-0006). Department of Computer Science, Tokyo Institute of Technology.

Kawahara, Y., & Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)* (pp. 389–400). Sparks, Nevada, USA.

Kawanabe, M., Sugiyama, M., Blanchard, G., & Müller, K.-R. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, *59*, 57–75.

Koh, K., Kim, S.-J., & Boyd, S. P. (2007). An interior-point method for large-scale l_1-regularized logistic regression. *Journal of Machine Learning Research*, *8*, 1519–1555.

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

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

Minka, T. P. (2007). *A comparison of numerical optimizers for logistic regression* (Technical Report). Microsoft Research.

Nguyen, X., Wainwright, M., & Jordan, M. (2008). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. C. Platt, D. Koller, Y. Singer and S. Roweis (Eds.), *Advances in neural information processing systems 20*, 1089–1096. Cambridge, MA: MIT Press.

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

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (Eds.). (2009). *Dataset shift in machine learning*. Cambridge, MA: MIT Press.

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning*, *42*, 287–320.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*, 1443–1471.

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

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

Storkey, A., & Sugiyama, M. (2007). Mixture regression for covariate shift. *Advances in Neural Information Processing Systems 19* (pp. 1337–1344). Cambridge, MA: MIT Press.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, *7*, 141–166.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, *8*, 1027–1061.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

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

Sugiyama, M., & Nakajima, S. (2009). Pool-based active learning in approximate linear regression. *Machine Learning*, *75*, 249–274.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., & Hachiya, H. (2009). *Least-squares conditional density estimation* (Technical Report TR09-0004). Department of Computer Science, Tokyo Institute of Technology.

Suzuki, T., & Sugiyama, M. (2009a). Estimating squared-loss mutual information for independent component analysis. *Independent Component Analysis and Signal Separation* (pp. 130–137). Berlin: Springer.

Suzuki, T., & Sugiyama, M. (2009b). *Sufficient dimension reduction via squared-loss mutual information estimation* (Technical Report TR09-0005). Department of Computer Science, Tokyo Institute of Technology.

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

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

Takeuchi, I., Nomura, K., & Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, *21*, 533–559.

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

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wiens, D. P. (2000). Robust weights and designs for biased regression models: Least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, *83*, 395–412.

Yamada, M., Sugiyama, M., & Matsui, T. (2009). Covariate shift adaptation for semi-supervised speaker identification. *Proceedings of 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)* (pp. 1661–1664). Taipei, Taiwan.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 903–910). New York, NY: ACM Press.

Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. *Advances in Neural Information Processing Systems 17* (pp. 1601–1608). Cambridge, MA: MIT Press.