Direct Density-ratio Estimation with Dimensionality Reduction via Least-squares Hetero-distributional Subspace Search

Masashi Sugiyama
Tokyo Institute of Technology
and PRESTO, Japan Science and Technology Agency,
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi

Makoto Yamada
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
yamada@sg.cs.titech.ac.jp

Paul von Bünau
Technical University of Berlin
Franklinstr. 28/29, 10587 Berlin, Germany.
buenau@cs.tu-berlin.de

Taiji Suzuki
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
s-taiji@stat.t.u-tokyo.ac.jp

Takafumi Kanamori
Nagoya University
Furocho, Chikusaku, Nagoya 464-8603, Japan.
kanamori@is.nagoya-u.ac.jp

Motoaki Kawanabe
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany.
motoaki.kawanabe@first.fraunhofer.de

Abstract

Methods for directly estimating the ratio of two probability density functions have been actively explored recently since they can be used for various data processing tasks such as *non-stationarity adaptation*, *outlier detection*, and *feature selection*. In this paper, we develop a new method which incorporates dimensionality reduction into a direct density-ratio estimation procedure. Our key idea is to find a low-dimensional subspace in which densities are significantly different and perform density ratio estimation only in this subspace. The proposed method, D^3-LHSS (Direct Density-ratio estimation with Dimensionality reduction via Least-squares Hetero-distributional Subspace Search), is shown to overcome the limitation of baseline methods.

Keywords

density ratio estimation, dimensionality reduction, unconstrained least-squares importance fitting

1 Introduction

Recently, it has been demonstrated that various machine learning and data mining tasks can be formulated in terms of the ratio of two probability density functions (Sugiyama et al., 2009; Sugiyama et al., 2011). Examples of such tasks include *covariate shift adaptation* (Shimodaira, 2000; Zadrozny, 2004; Sugiyama et al., 2007; Sugiyama & Kawanabe, 2010), *transfer learning* (Storkey & Sugiyama, 2007), *multi-task learning* (Bickel et al., 2008), *outlier detection* (Hido et al., 2008; Smola et al., 2009; Hido et al., 2010), *conditional density estimation* (Sugiyama et al., 2010c), *probabilistic classification* (Sugiyama, 2010), *variable selection* (Suzuki et al., 2009a), *independent component analysis* (Suzuki & Sugiyama, 2009), *supervised dimensionality reduction* (Suzuki & Sugiyama, 2010), and *causal inference* (Yamada & Sugiyama, 2010), For this reason, estimating the density ratio has been attracting a great deal of attention, and various approaches have been explored (Silverman, 1978; Ćwik & Mielniczuk, 1989; Gijbels & Mielniczuk, 1995; Sun & Woodroofe, 1997; Jacob & Oliveira, 1997; Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bensaid & Fabre, 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009a; Chen et al., 2009; Sugiyama et al., 2010b; Nguyen et al., 2010).

A naive approach to density ratio estimation is to approximate the two densities in the ratio (i.e., the numerator and the denominator) separately using a flexible technique such as non-parametric *kernel density estimation* (Silverman, 1986; Härdle et al., 2004), and then take the ratio of the estimated densities. However, this naive two-step approach is not reliable in practical situations since kernel density estimation performs poorly in high-dimensional cases; furthermore, division by an estimated density tends to magnify the estimation error. To improve the estimation accuracy, various methods have been developed for directly estimating the density ratio without going through density estimation, e.g., the moment matching method using reproducing kernels (Aronszajn, 1950;

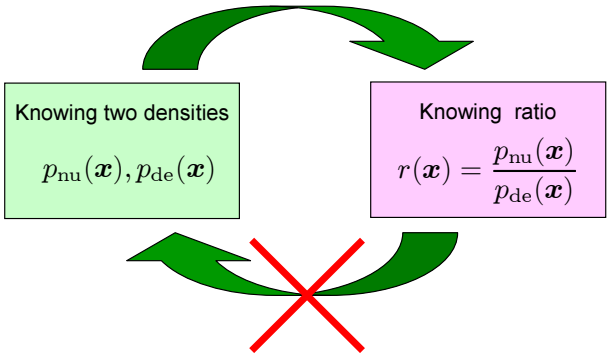


Figure 1: Density ratio estimation is substantially easier than density estimation. The density ratio $r(\boldsymbol{x})$ can be computed if two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are known. However, even if the density ratio is known, the two densities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

Steinwart, 2001) called *kernel mean matching* (KMM) (Huang et al., 2007; Quiñonero-Candela et al., 2009), the method based on *logistic regression* (LR) (Qin, 1998; Cheng & Chu, 2004; Bickel et al., 2007), the distribution matching method under the *Kullback-Leibler (KL) divergence* (Kullback & Leibler, 1951) called the *KL importance estimation procedure* (KLIEP) (Sugiyama et al., 2008; Nguyen et al., 2010), and the density-ratio matching methods under the squared-loss called *least-squares importance fitting* (LSIF) and *unconstrained LSIF* (uLSIF) (Kanamori et al., 2009a). These methods have been shown to compare favorably with naive kernel density estimation through extensive experiments.

The success of these direct density-ratio estimation methods could be intuitively understood through *Vapnik's principle* (Vapnik, 1998): "When solving a problem of interest, one should not solve a more general problem as an intermediate step". The *support vector machine* would be a successful example following this principle—instead of estimating the data generation model, it directly models the decision boundary which is simpler and sufficient for pattern recognition. In the current context, estimating the densities is more general than estimating the density ratio since knowing the two densities implies knowing the ratio, but not vice versa (Figure 1). Thus directly estimating the density ratio would be more promising than density ratio estimation via density estimation.

However, density ratio estimation in high-dimensional cases is still challenging even when the ratio is estimated directly without going through density estimation. Recently, an approach called *Direct Density-ratio estimation with Dimensionality reduction* (D^3) has been proposed (Sugiyama et al., 2010a). The basic idea of D^3 is the following two-step procedure: First a subspace in which the numerator and denominator densities are significantly different (called the *hetero-distributional subspace*) are identified, and then density ratio estimation is performed in this subspace. The rationale behind this approach is that, in practice, the distribution change does not occur in the entire space, but is often confined in a subspace. For example, in non-stationarity adaptation scenarios, the distribution change often occurs only for some attributes and other variables are stable; in

outlier detection scenarios, only a small number of attributes would cause a data sample to be an outlier.

In the D^3 algorithm, the hetero-distributional subspace is identified by searching a subspace in which samples drawn from the two distributions (i.e., the numerator and the denominator of the ratio) are separated from each other—this search is carried out in a computationally efficient manner using a supervised dimensionality reduction method called *local Fisher discriminant analysis* (LFDA) (Sugiyama, 2007). Then, within the identified hetero-distributional subspace, a direct density-ratio estimation method called *unconstrained least-squares importance Fitting* (uLSIF)—which was shown to be computationally efficient (Kanamori et al., 2009a) and numerically stable (Kanamori et al., 2009b)—is employed for obtaining the final density-ratio estimator. Through experiments, this D^3 procedure (which we refer to as D^3-LFDA/uLSIF) was shown to improve the performance in high-dimensional cases.

Although the framework of D^3 is promising, the above D^3-LFDA/uLSIF method possesses two fundamental weaknesses: the restrictive definition of the hetero-distributional subspace and the limiting ability of its search method. More specifically, the component inside the hetero-distributional subspace and its complementary component are assumed to be statistically independent in the original formulation (Sugiyama et al., 2010a). However, this assumption is rather restrictive and may not be fulfilled in practice. Also, in the above D^3 procedure, the hetero-distributional subspace is identified by searching a subspace in which samples drawn from the numerator and denominator distributions are separated from each other. If samples from the two distributions are separable, the two distributions would be significantly different. However, the opposite may not be always true, i.e., non-separability does not necessarily imply that the two distributions are different (consider two similar distributions with the common support). Thus LFDA (and any other supervised dimensionality reduction methods) does not necessarily identify the correct hetero-distributional subspace.

The goal of this paper is to give a new procedure of D^3 that can overcome the above weaknesses. First, we adopt a more general definition of the hetero-distributional subspace. More precisely, we remove the independence assumption between the component inside the hetero-distributional subspace and its complementary component. This allows us to apply the concept of D^3 to a wider class of problems. However, this general definition in turn makes the problem of searching the hetero-distributional subspace more challenging—supervised dimensionality reduction methods for separating samples drawn from the two distributions cannot be used anymore, but we need an alternative method that identifies the largest subspace such that the two *conditional* distributions are equivalent in its complementary subspace.

We prove that the hetero-distributional subspace can be identified by finding a subspace in which two *marginal* distributions are maximally different under the *Pearson divergence*, which is a squared-loss variant of the *Kullback-Leibler divergence* and is an instance of the *f-divergences* (Ali & Silvey, 1966; Csiszár, 1967). Then we propose a new method, which we call *Least-squares Hetero-distributional Subspace Search* (LHSS), for searching a subspace such that the Pearson divergence between two marginal distri-

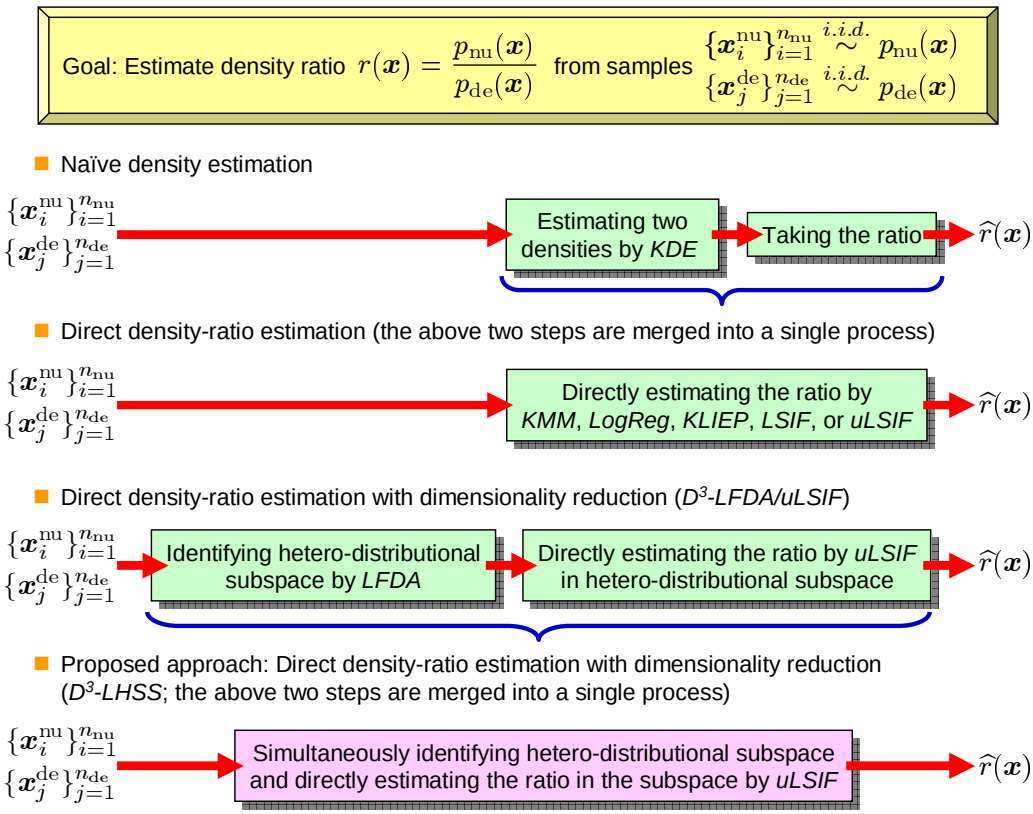


Figure 2: Existing and proposed density-ratio estimation approaches.

butions are maximized. An advantage of the LHSS method is that the subspace search (divergence estimation within a subspace) is carried out also using the density-ratio estimation method uLSIF. Thus the two steps in the D^3 procedure (first identifying the hetero-distributional subspace and then estimating the density ratio within the subspace) are merged into a single step. Thanks to this, the final density-ratio estimator can be automatically obtained without additional computation. We call the combined single-shot density-ratio estimation procedure *D^3 via LHSS* (D^3-LHSS). Through experiments, we show that the weaknesses of the existing approach can be successfully overcome by the D^3-LHSS approach.

Relation among the existing and proposed density-ratio estimation methods is summarized in Figure 2.

2 Formulation of Density-ratio Estimation Problem

In this section, we formulate the problem of density ratio estimation and review a relevant density-ratio estimation method. We briefly summarize possible usage of density ratios in various data processing tasks in Appendix A.

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 $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ from a distribution with density $p_{\mathrm{nu}}(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ from another distribution with density $p_{\mathrm{de}}(\boldsymbol{x})$. We assume that the latter density $p_{\mathrm{de}}(\boldsymbol{x})$ is strictly positive, i.e.,

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

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

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

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

2.2 Directly Estimating Density Ratios by Unconstrained Least-squares Importance Fitting (uLSIF)

As described in Appendix A, density ratios are useful in various data processing tasks. Since the density ratio is usually unknown and needs to be estimated from data, methods of estimating the density ratio have been actively explored recently (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009a). Here, we briefly review a direct density-ratio estimation method called *unconstrained least-squares importance fitting* (uLSIF) proposed by Kanamori et al. (2009a). For convenience in later sections, we replace the symbol $\boldsymbol{x}$ with $\boldsymbol{u}$, i.e., let us consider the problem of estimating the density ratio

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

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

2.2.1 Linear Least-squares Estimation of Density Ratios

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, b denotes the number of parameters, $^\top$ denotes the transpose of a matrix or a vector, 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 $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{u}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, meaning that kernel models are also allowed. We explain how the basis functions $\{\psi_\ell(\boldsymbol{u})\}_{\ell=1}^b$ are designed in Section 2.2.2.

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:

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

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

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

Note that the same objective function can be obtained via the *Legendre-Fenchel duality* of a divergence (Nguyen et al., 2010).

Approximating the expectations in J by empirical averages, we obtain

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

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

$$\widehat{H}_{\ell,\ell'} := \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} \psi_\ell(\boldsymbol{u}_j^{\mathrm{de}})\psi_{\ell'}(\boldsymbol{u}_j^{\mathrm{de}}), \tag{2}$$

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

$$\widehat{h}_\ell := \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \psi_\ell(\boldsymbol{u}_i^{\mathrm{nu}}). \tag{3}$$

Now the optimization problem is formulated as follows.

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right], \tag{4}$$

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 the strength of regularization. It is easy to confirm that the solution $\widehat{\boldsymbol{\alpha}}$ can be analytically computed as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}, \tag{5}$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix. Thanks to this analytic-form expression, uLSIF is computationally efficient compared with other density-ratio estimators which involve non-linear optimization (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Nguyen et al., 2010).

In the original uLSIF paper (Kanamori et al., 2009a), the above solution is further modified as

$$\widehat{\alpha}_\ell \longleftarrow \max(0, \widehat{\alpha}_\ell).$$

This modification may improve the estimation accuracy in finite sample cases since the true density ratio is non-negative. Even so, we still use Eq.(5) as it is since it is differentiable with respect to $\boldsymbol{U}$, where $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{x}$. This differentiability will play a crucial role in the next section. Note that, even without the above round-up modification, the solution is guaranteed to converge to the optimal vector asymptotically both in parametric and non-parametric cases (Kanamori et al., 2009a; Kanamori et al., 2009b). Thus omitting the above modification step may not have a strong effect.

It was theoretically shown that uLSIF possesses superior theoretical properties in statistical convergence and numerical stability (Kanamori et al., 2009a; Kanamori et al., 2009b).

2.2.2 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:

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

where $K(\boldsymbol{u}, \boldsymbol{u}')$ is the Gaussian kernel with kernel width σ (> 0):

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

By definition, the density ratio $r(\boldsymbol{u})$ tends to take large values if $p_{\mathrm{nu}}(\boldsymbol{u})$ is large and $p_{\mathrm{de}}(\boldsymbol{u})$ is small; conversely, $r(\boldsymbol{u})$ tends to be small (i.e., close to zero) if $p_{\mathrm{nu}}(\boldsymbol{u})$ is small and $p_{\mathrm{de}}(\boldsymbol{u})$ is large. When a non-negative function is approximated by a Gaussian kernel model, many kernels may be needed in the region where the output of the target function is large; on the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero (see Figure 3). Following this

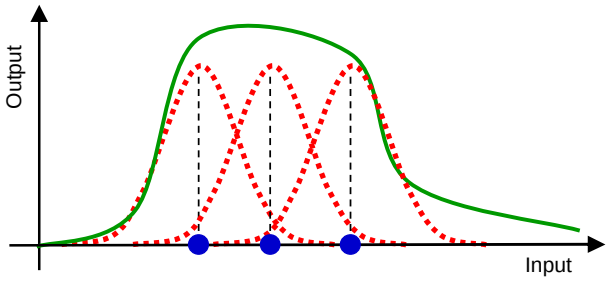


Figure 3: Heuristic of Gaussian kernel allocation.

heuristic, we allocate many kernels in the region where $p_{\mathrm{nu}}(\boldsymbol{u})$ takes large values, which may be approximately achieved by setting the Gaussian centers at $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$.

Alternatively, we may locate $(n_{\mathrm{nu}} + n_{\mathrm{de}})$ Gaussian kernels at both $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{u}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$. 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 $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ as Gaussian centers is already computationally rather demanding. To ease this problem, a subset of $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ may be used as Gaussian centers for computational efficiency, i.e., for a prefixed b $(\in \{1, 2, \ldots, n_{\mathrm{nu}}\})$, we use

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

where $\{\boldsymbol{c}_\ell\}_{\ell=1}^{b}$ are template points randomly chosen from $\{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ without replacement.

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 (1) (Kanamori et al., 2009a).

3 Direct Density-ratio Estimation with Dimensionality Reduction

Although uLSIF was shown to be a useful density ratio estimation method (Kanamori et al., 2009a), estimating the density ratio in high-dimensional spaces is still challenging. In this section, we propose a new method of direct density-ratio estimation that involves dimensionality reduction.

3.1 Hetero-distributional Subspace

Our basic idea is to first find a low-dimensional subspace in which the two densities are significantly different from each other, and then perform density ratio estimation only in this subspace. Although a similar framework has been explored in Sugiyama et al. (2010a), the current formulation is substantially more general than the previous approach, as explained below.

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

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

where $\boldsymbol{U}$ is an $m \times d$ matrix and $\boldsymbol{V}$ is a $(d-m) \times d$ matrix. In order to ensure the uniqueness of the decomposition, we assume (without loss of generality) that the row vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ form an orthonormal basis, i.e., $\boldsymbol{U}$ and $\boldsymbol{V}$ correspond to "projection" matrices that are orthogonally complementary to each other (see Figure 4). Then the two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ can be decomposed as

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

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

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

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

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

Thus, the density ratio does not have to be estimated in the entire d-dimensional space, but it is sufficient to estimate the ratio only in the m-dimensional subspace specified by $\boldsymbol{U}$.

Below, we will use the term, the *hetero-distributional subspace*, for indicating the subspace specified by $\boldsymbol{U}$ in which $p_{\mathrm{nu}}(\boldsymbol{u})$ and $p_{\mathrm{de}}(\boldsymbol{u})$ are different. More precisely, let $\mathcal{S}$ be a subspace specified by $\boldsymbol{U}$ and $\boldsymbol{V}$ such that

$$\mathcal{S} = \{\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{x} \mid p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}),\ \ \boldsymbol{u} = \boldsymbol{U}\boldsymbol{x},\ \ \boldsymbol{v} = \boldsymbol{V}\boldsymbol{x}\}.$$

Then the hetero-distributional subspace is defined as the *intersection* of all subspaces $\mathcal{S}$. Intuitively, the hetero-distributional subspace is the 'smallest' subspace specified by $\boldsymbol{U}$ such that $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u})$ and $p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})$ agree with each other. We refer to the orthogonal complement of the hetero-distributional subspace as the *homo-distributional subspace* (see Figure 4).

This formulation is a generalization of the one proposed in Sugiyama et al. (2010a) in which the components in the hetero-distributional subspace and its complimentary subspace are assumed to be independent of each other. On the other hand, we do not impose

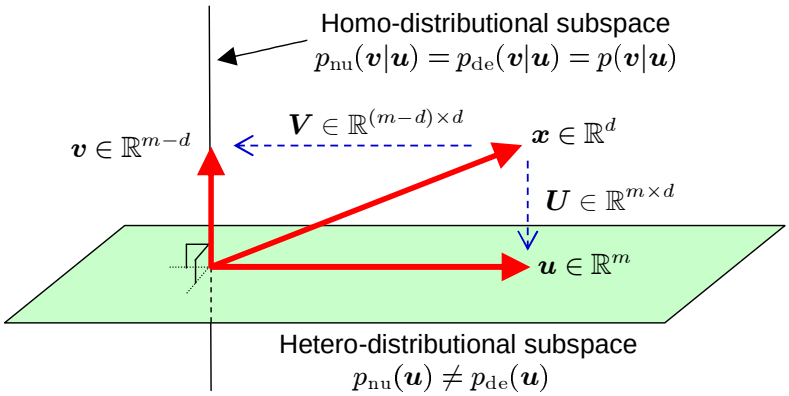


Figure 4: Hetero-distributional subspace.

such an independence assumption in the current paper. As will be demonstrated in Section 4.1, this generalization has a remarkable effect in extending the range of applications of direct density-ratio estimation with dimensionality reduction.

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

3.2 Estimating Pearson Divergence Using uLSIF

Here, we introduce a criterion for hetero-distributional subspace search and how it is estimated from data.

We use the *Pearson divergence* (PD) as our criterion for evaluating the discrepancy between two distributions. PD is a squared-loss variant of the *Kullback-Leibler divergence* (Kullback & Leibler, 1951), and is an instance of the *f-divergences*, which are also known as the Csiszár f-divergences (Csiszár, 1967) or the Ali-Silvey distances (Ali & Silvey, 1966). PD from $p_{\mathrm{nu}}(\boldsymbol{x})$ to $p_{\mathrm{de}}(\boldsymbol{x})$ is defined and expressed as

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

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

The following lemma (called the "*data processing*" inequality) characterizes the hetero-distributional subspace in terms of PD.

Lemma 1 *Let*

$$\begin{aligned}\mathrm{PD}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] &= \frac{1}{2}\int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1\right)^2 p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\ &= \frac{1}{2}\int \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \frac{1}{2}. \qquad (6)\end{aligned}$$

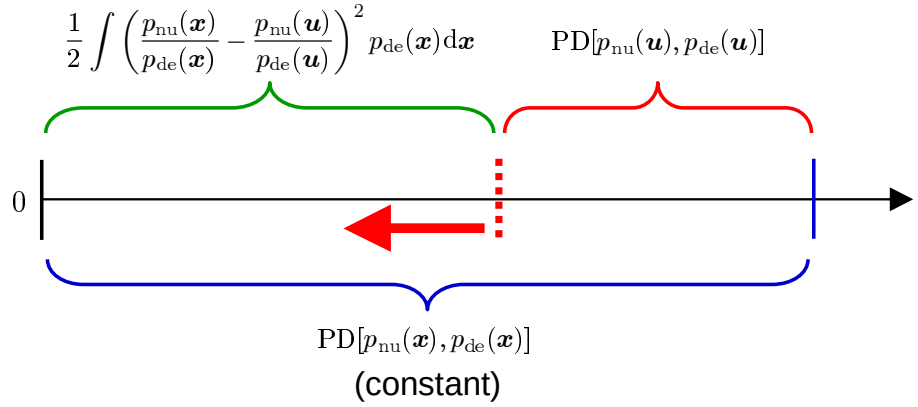


Figure 5: Since $\mathrm{PD}[p_{\rm nu}(\boldsymbol{x}),p_{\rm de}(\boldsymbol{x})]$ is constant, minimizing $\frac{1}{2}\int\left(\frac{p_{\rm nu}(\boldsymbol{x})}{p_{\rm de}(\boldsymbol{x})}-\frac{p_{\rm nu}(\boldsymbol{u})}{p_{\rm de}(\boldsymbol{u})}\right)^2 p_{\rm de}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is equivalent to maximizing $\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]$.

Then we have

$$\begin{aligned}\mathrm{PD}[p_{\rm nu}(\boldsymbol{x}),p_{\rm de}(\boldsymbol{x})]-\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]&=\frac{1}{2}\int\left(\frac{p_{\rm nu}(\boldsymbol{x})}{p_{\rm de}(\boldsymbol{x})}-\frac{p_{\rm nu}(\boldsymbol{u})}{p_{\rm de}(\boldsymbol{u})}\right)^2 p_{\rm de}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\&\geq 0.\end{aligned}\qquad(7)$$

A proof of the above lemma (for a class of f-divergences) is provided in Appendix B. The right-hand side of Eq.(7) is non-negative, and it vanishes if and only if $p_{\rm nu}(\boldsymbol{v}|\boldsymbol{u})=p_{\rm de}(\boldsymbol{v}|\boldsymbol{u})$. Since $\mathrm{PD}[p_{\rm nu}(\boldsymbol{x}),p_{\rm de}(\boldsymbol{x})]$ is a constant with respect to $\boldsymbol{U}$, maximizing $\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]$ with respect to $\boldsymbol{U}$ leads to $p_{\rm nu}(\boldsymbol{v}|\boldsymbol{u})=p_{\rm de}(\boldsymbol{v}|\boldsymbol{u})$ (Figure 5). That is, the hetero-distributional subspace can be characterized as the maximizer[1] of $\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]$.

Although the hetero-distributional subspace can be characterized as the maximizer of $\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]$, we cannot directly find the maximizer since $p_{\rm nu}(\boldsymbol{u})$ and $p_{\rm de}(\boldsymbol{u})$ are unknown. Here, we utilize a direct density-ratio estimator uLSIF (see Section 2.2) for approximating $\mathrm{PD}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]$ from samples. Let us replace the density ratio $p_{\rm nu}(\boldsymbol{u})/p_{\rm de}(\boldsymbol{u})$ in Eq.(6) by a density ratio estimator $\widehat{r}(\boldsymbol{u})$. Approximating the expectation over $p_{\rm nu}(\boldsymbol{u})$ by an empirical average over $\{\boldsymbol{u}_i^{\rm nu}\}_{i=1}^{n_{\rm nu}}$, we have the following PD estimator.

$$\widehat{\mathrm{PD}}[p_{\rm nu}(\boldsymbol{u}),p_{\rm de}(\boldsymbol{u})]:=\frac{1}{2n_{\rm nu}}\sum_{i=1}^{n_{\rm nu}}\widehat{r}(\boldsymbol{u}_i^{\rm nu})-\frac{1}{2}.$$

Since uLSIF was shown to be *consistent* (i.e., the solution converges to the optimal value) both in parametric and non-parametric cases (Kanamori et al., 2009a; Kanamori et al., 2009b), $\widehat{\mathrm{PD}}$ would be a consistent estimator of the true PD.

[1] As shown in Appendix B, the data processing inequality holds not only for PD, but also for *any* f-divergences. Thus the characterization of the hetero-distributional subspace is not limited to PD, but is applicable to all f-divergences.

3.3 Least-squares Hetero-distributional Subspace Search (LHSS)

Given the uLSIF-based PD estimator $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$, our next task is to find a maximizer of $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ with respect to $\boldsymbol{U}$, and identify the hetero-distributional subspace (cf. the data processing inequality given in Lemma 1). We call this procedure *Least-squares Hetero-distributional Subspace Search* (LHSS).

We may employ various optimization techniques to find a maximizer of $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$. Here we describe several possibilities.

3.3.1 Plain Gradient Algorithm

A gradient ascent algorithm would be a fundamental approach to non-linear smooth optimization. We utilize the following lemma.

Lemma 2 *The gradient of* $\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ *with respect to* $\boldsymbol{U}$ *is expressed as*

$$\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}} = \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \frac{\partial \widehat{h}_\ell}{\partial \boldsymbol{U}} - \frac{1}{2} \sum_{\ell,\ell'=1}^{b} \widehat{\alpha}_\ell \widehat{\alpha}_{\ell'} \frac{\partial \widehat{H}_{\ell,\ell'}}{\partial \boldsymbol{U}}, \tag{8}$$

where $\widehat{\boldsymbol{\alpha}}$ *is given by Eq.*(5) *and*

$$\frac{\partial \widehat{h}_\ell}{\partial \boldsymbol{U}} = \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \frac{\partial \psi_\ell(\boldsymbol{u}_i^{\mathrm{nu}})}{\partial \boldsymbol{U}}, \tag{9}$$

$$\frac{\partial \widehat{H}_{\ell,\ell'}}{\partial \boldsymbol{U}} = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \left(\frac{\partial \psi_\ell(\boldsymbol{u}_j^{\mathrm{de}})}{\partial \boldsymbol{U}} \psi_{\ell'}(\boldsymbol{u}_j^{\mathrm{de}}) + \psi_\ell(\boldsymbol{u}_j^{\mathrm{de}}) \frac{\partial \psi_{\ell'}(\boldsymbol{u}_j^{\mathrm{de}})}{\partial \boldsymbol{U}} \right), \tag{10}$$

$$\frac{\partial \psi_\ell(\boldsymbol{u})}{\partial \boldsymbol{U}} = -\frac{1}{\sigma^2} (\boldsymbol{u} - \boldsymbol{c}_\ell)(\boldsymbol{x} - \boldsymbol{c}'_\ell)^\top \psi_\ell(\boldsymbol{u}). \tag{11}$$

$\boldsymbol{c}'_\ell$ $(\in \mathbb{R}^d)$ *is a pre-image of* $\boldsymbol{c}_\ell$ $(\in \mathbb{R}^m)$*:*

$$\boldsymbol{c}_\ell = \boldsymbol{U} \boldsymbol{c}'_\ell.$$

A proof of the above lemma is provided in Appendix C. Note that $\{\widehat{\alpha}_\ell\}_{\ell=1}^b$ in Eq.(8) depend on $\widehat{\boldsymbol{U}}$ through $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ in Eq.(5), which was taken into account when deriving the gradient (see Appendix C). A plain gradient update rule is then given as

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

where t (> 0) is a learning rate. t may be chosen in practice by some approximate line search method such as *Armijo's rule* (Patriksson, 1999) or *backtracking line search* (Boyd & Vandenberghe, 2004).

A naive gradient update does not necessarily fulfill the orthonormality $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the m-dimensional identity matrix. Thus, after every gradient step, we need to orthonormalize $\boldsymbol{U}$ by, e.g., the *Gram-Schmidt process* (Golub & Loan, 1996) to guarantee its orthonormality. However, this may be rather time-consuming.

3.3.2 Natural Gradient Algorithm

In the Euclidean space, the ordinary gradient $\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}$ gives the steepest direction. On the other hand, in the current setup, the matrix $\boldsymbol{U}$ is restricted to be a member of the *Stiefel manifold* $\mathbb{S}_m^d(\mathbb{R})$:

$$\mathbb{S}_m^d(\mathbb{R}) := \{\boldsymbol{U} \in \mathbb{R}^{m\times d} \mid \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_m\}.$$

On a manifold, it is known that, not the ordinary gradient, but the *natural gradient* (Amari, 1998) gives the steepest direction. The natural gradient $\nabla\widehat{\mathrm{PD}}(\boldsymbol{U})$ at $\boldsymbol{U}$ is the projection of the ordinary gradient $\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}$ onto the tangent space of $\mathbb{S}_m^d(\mathbb{R})$ at $\boldsymbol{U}$.

If the tangent space is equipped with the canonical metric, i.e., for any $\boldsymbol{G}$ and $\boldsymbol{G}'$ in the tangent space,

$$\langle \boldsymbol{G}, \boldsymbol{G}' \rangle = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{G}^\top \boldsymbol{G}'\right), \tag{12}$$

the natural gradient is given by

$$\nabla\widehat{\mathrm{PD}}(\boldsymbol{U}) = \frac{1}{2}\left(\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}} - \boldsymbol{U}\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}^\top \boldsymbol{U}\right).$$

Then the *geodesic* from $\boldsymbol{U}$ to the direction of the natural gradient $\nabla\widehat{\mathrm{PD}}(\boldsymbol{U})$ over $\mathbb{S}_m^d(\mathbb{R})$ can be expressed using $t \in \mathbb{R}$ as

$$\boldsymbol{U}_t := \boldsymbol{U}\exp\left\{t\left(\boldsymbol{U}^\top \frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}} - \frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}^\top \boldsymbol{U}\right)\right\},$$

where 'exp' for a matrix denotes the *matrix exponential*, i.e., for a square matrix $\boldsymbol{T}$,

$$\exp(\boldsymbol{T}) := \sum_{k=0}^{\infty} \frac{1}{k!}\boldsymbol{T}^k. \tag{13}$$

Thus, line search along the geodesic in the natural gradient direction is equivalent to finding a maximizer from

$$\{\boldsymbol{U}_t \mid t \geq 0\}.$$

More details of geometric structure of the Stiefel manifold can be found in Nishimori and Akaho (2005).

A natural gradient update rule is then given as

$$\boldsymbol{U} \longleftarrow \boldsymbol{U}_t,$$

where t (> 0) is the learning rate. Since the orthonormality of $\boldsymbol{U}$ is automatically satisfied in the natural gradient method, it would be computationally more efficient than the plain gradient method. However, optimizing the $m \times d$ matrix $\boldsymbol{U}$ is still computationally expensive.

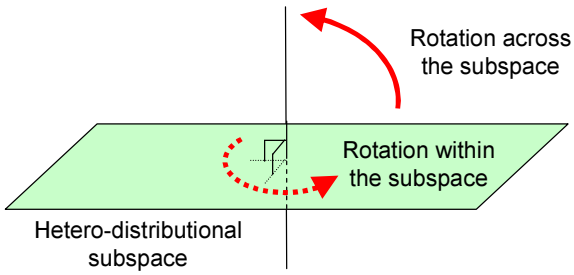


Figure 6: In the hetero-distributional subspace search, rotation which changes the subspace only matters (the solid arrow); rotation within the subspace (dotted arrow) can be ignored since this does not change the subspace. Similarly, rotation within the orthogonal complement of the hetero-distributional subspace can also be ignored (not depicted in the figure).

3.3.3 Givens Rotation

Another simple strategy for optimizing $\boldsymbol{U}$ is to rotate the matrix in the plane spanned by two coordinate axes (which is called the *Givens rotations*; see Golub & Loan, 1996). That is, we randomly choose a two-dimensional subspace spanned by the i-th and j-th variables, and rotate the matrix $\boldsymbol{U}$ within this subspace:

$$\boldsymbol{U} \longleftarrow \boldsymbol{R}_\theta^{(i,j)} \boldsymbol{U},$$

where $\boldsymbol{R}_\theta^{(i,j)}$ is the rotation matrix by angle θ within the subspace spanned by the i-th and j-th variables. $\boldsymbol{R}_\theta^{(i,j)}$ is equal to the identity matrix except that its elements (i,i), (i,j), (j,i), and (j,j) form a two-dimensional rotation matrix:

$$\begin{bmatrix} [\boldsymbol{R}_\theta^{(i,j)}]_{i,i} & [\boldsymbol{R}_\theta^{(i,j)}]_{i,j} \\ [\boldsymbol{R}_\theta^{(i,j)}]_{j,i} & [\boldsymbol{R}_\theta^{(i,j)}]_{j,j} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

The rotation angle θ ($0 \leq \theta \leq \pi$) may be optimized by some secant method (Press et al., 1992).

As shown above, the update rule of the Givens rotations is computationally very efficient. However, since the update direction is not optimized as in the plain/natural gradient methods, the Givens-rotation method could be potentially less efficient as an optimization strategy.

3.3.4 Subspace Rotation

Since we are searching for a subspace, rotation *within* the subspace does not have any influence on the objective value $\widehat{\mathrm{PD}}$ (see Figure 6). This implies that the number of parameters to be optimized in the gradient algorithm can be reduced.

For a *skew-symmetric* matrix $\boldsymbol{M}$ ($\in \mathbb{R}^{d\times d}$), i.e., $\boldsymbol{M}^\top = -\boldsymbol{M}$, rotation of $\boldsymbol{U}$ can be expressed as follows (Plumbley, 2005):

$$\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{O}_{m,(d-m)}\end{bmatrix} \exp(\boldsymbol{M}) \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix},$$

where $\boldsymbol{O}_{d,d'}$ is the $d \times d'$ matrix with all zeros, and $\exp(\boldsymbol{M})$ is the matrix exponential of $\boldsymbol{M}$ (see Eq.(13)). $\boldsymbol{M} = \boldsymbol{O}_{d,d}$ (i.e., $\exp(\boldsymbol{O}_{d,d}) = \boldsymbol{I}_d$) corresponds to no rotation. Here we update $\boldsymbol{U}$ through the matrix $\boldsymbol{M}$.

Let us adopt Eq.(12) as the inner product in the space of skew-symmetric matrices. Then we have the following lemma.

Lemma 3 *The derivative of* $\widehat{\mathrm{PD}}$ *with respect to* $\boldsymbol{M}$ *at* $\boldsymbol{M} = \boldsymbol{O}_{d,d}$ *is given by*

$$\left.\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{O}_{d,d}} = \begin{bmatrix}\boldsymbol{O}_{m,m} & \frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top\\ -(\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top)^\top & \boldsymbol{O}_{(d-m),(d-m)}\end{bmatrix}. \tag{14}$$

A proof of the above lemma is provided in Appendix D. The block structure of Eq.(14) has an intuitive explanation: the non-zero off-diagonal blocks correspond to the rotation angles *between* the hetero-distributional subspace and its orthogonal complement which do affect the objective function $\widehat{\mathrm{PD}}$. On the other hand, the derivative of rotation *within* the two subspaces vanishes because this does not change the objective value. Thus the variables to be optimized are only the angles corresponding to the non-zero off-diagonal blocks $\frac{\partial \widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top$, which includes only $m(d-m)$ variables. In contrast, the plain/natural gradient algorithms optimize the matrix $\boldsymbol{U}$, which contains md variables. Thus, when m is large, the subspace rotation approach may be computationally more efficient than the plain/natural gradient algorithms.

The gradient ascent update rule of $\boldsymbol{M}$ is given by

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

where t is a step-size. Then $\boldsymbol{U}$ is updated as

$$\boldsymbol{U} \longleftarrow \begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{O}_{m,(d-m)}\end{bmatrix} \exp(\boldsymbol{M}) \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}.$$

The *conjugate gradient* method (Golub & Loan, 1996) may be used for the update of $\boldsymbol{M}$.

Following the update of $\boldsymbol{U}$, its counterpart $\boldsymbol{V}$ also needs to be updated accordingly since the hetero-distributional subspace and its complement specified by $\boldsymbol{U}$ and $\boldsymbol{V}$ should be orthogonal to each other (see Figure 4). This can be achieved by setting

$$\boldsymbol{V} \longleftarrow \begin{bmatrix}\boldsymbol{\varphi}_1 & \boldsymbol{\varphi}_2 & \cdots & \boldsymbol{\varphi}_{d-m}\end{bmatrix}^\top,$$

where $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_{d-m}$ are orthonormal basis vectors in the orthogonal complement of the hetero-distributional subspace.

3.4 Proposed Algorithm: D^3-LHSS

Finally, we estimate the density ratio in the hetero-distributional subspace detected by the above LHSS method.

A notable fact of the LHSS algorithm is that the density ratio estimator in the hetero-distributional subspace has already been obtained during the hetero-distributional subspace search procedure. Thus, we do not need an additional estimation procedure—our final solution is simply given by

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

where $\widehat{\boldsymbol{U}}$ is a projection matrix obtained by the LHSS algorithm. $\{\widehat{\alpha}_\ell\}_{\ell=1}^b$ are the learned parameters for $\widehat{\boldsymbol{U}}$, which have been obtained and used when computing the gradient (see Lemma 2).

This expression implies that if the dimensionality is not reduced (i.e., $m = d$), the proposed method agrees with the original uLSIF (see Section 2.2). Thus, the proposed method could be regarded as a natural extension of uLSIF to high-dimensional data.

Given the true dimensionality m of the hetero-distributional subspace, we can estimate the hetero-distributional subspace by the LHSS algorithm. When m is unknown, we may choose the best dimensionality based on the CV score of the uLSIF estimator. We refer to our proposed procedure *D^3-LHSS* (D-cube LHSS; Direct Density-ratio estimation with Dimensionality reduction via Least-squares Hetero-distributional Subspace Search).

The complete procedure of D^3-LHSS is summarized in Figure 7. A MATLAB® implementation of D^3-LHSS is available from

'`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/D3LHSS/`'.

4 Experiments

In this section, we investigate the experimental performance of the proposed method. We employ the subspace rotation algorithm explained in Section 3.3.4 in our D^3-LHSS implementation. In uLSIF, the number of parameters is fixed to $b = 100$; the Gaussian width σ and the regularization parameter λ are chosen based on cross-validation.

4.1 Illustrative Examples

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

As explained in Section 1, the previous D^3 method, D^3-LFDA/uLSIF (Sugiyama et al., 2010a), has two potential weaknesses:

- The component $\boldsymbol{u}$ inside the hetero-distributional subspace and its complementary component $\boldsymbol{v}$ are assumed to be statistically independent (cf. Section 3.1).

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

For each reduced dimension $m = 1, 2, \ldots, d$
  Initialize embedding matrix $\boldsymbol{U}_m$ ($\in \mathbb{R}^{m \times d}$);
  Repeat until $\boldsymbol{U}_m$ converges
    Choose Gaussian width σ and regularization parameter λ by CV;
    Update $\boldsymbol{U}$ by some optimization method (see Section 3.3);
  end
  Obtain embedding matrix $\widehat{\boldsymbol{U}}_m$ and corresponding density-ratio estimator $\widehat{r}_m(\boldsymbol{x})$;
  Compute its CV value as a function of m;
end
Choose the best reduced dimensionality $\widehat{m}$ that minimizes the CV score;
Set $\widehat{r}(\boldsymbol{x}) = \widehat{r}_{\widehat{m}}(\boldsymbol{x})$;

Figure 7: Pseudo code of D^3-LHSS.

- Separability of samples drawn from two distributions implies that the two distributions are different, but non-separability does not necessarily imply that the two distributions are equivalent. Thus, D^3-LFDA/uLSIF may not be able to detect the subspace in which the two distributions are different, but samples are not really separable.

Here, through numerical examples, we illustrate these weaknesses of D^3-LFDA/uLSIF, and show these problems can be overcome by D^3-LHSS. Let us consider two-dimensional examples (i.e., $d = 2$), and suppose that the two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are different only in the one-dimensional subspace (i.e., $m = 1$) spanned by $(1, 0)^\top$:

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

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

"Rather-separate" dataset (Figure 8):

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

where $N(u; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2 with respect to u. This is an easy and simple dataset for the purpose of illustrating the usefulness of dimensionality reduction in density ratio estimation.

"Highly-overlapped" dataset (Figure 9):

$$p(v|u) = p(v) = N(v; 0, 1^2),$$
$$p_{\mathrm{nu}}(u) = N(u; 0, 0.6^2),$$
$$p_{\mathrm{de}}(u) = N(u; 0, 1.2^2).$$

Since v is independent of u, D^3-LFDA/uLSIF is still applicable in principle. However, u^{nu} and u^{de} are highly overlapped and are not clearly separable. Thus this dataset would be hard for D^3-LFDA/uLSIF.

"Dependent" dataset (Figure 10):

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

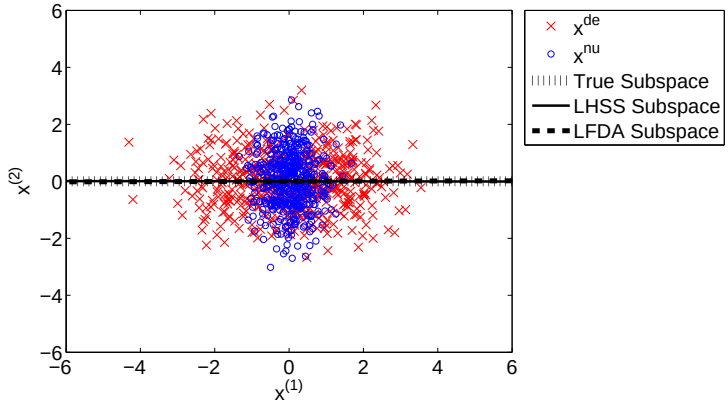
In this dataset, the *conditional* distribution $p(v|u)$ is common, but the *marginal* distributions $p_{\mathrm{nu}}(v)$ and $p_{\mathrm{de}}(v)$ are different. Since v is not independent of u, this dataset would be out of scope for D^3-LFDA/uLSIF.

The true hetero-distributional subspace for the "rather-separate" dataset is depicted by the dotted line in Figure 8(a); the solid line and the dashed line depict the hetero-distributional subspace found by LHSS and LFDA with reduced dimensionality $m = 1$, respectively. This graph shows that LHSS and LFDA both give very good estimates of the true hetero-distributional subspace. In Figure 8(c), Figure 8(d), and Figure 8(e), density ratio functions estimated by the plain uLSIF without dimensionality reduction, D^3-LFDA/uLSIF, and D^3-LHSS for the "rather-separate" dataset are depicted. These graphs show that both D^3-LHSS and D^3-LFDA/uLSIF give much better estimates of the density ratio function (see Figure 8(b) for the profile of the true density ratio function) than the plain uLSIF without dimensionality reduction. Thus, the usefulness of dimensionality reduction in density ratio estimation was illustrated.

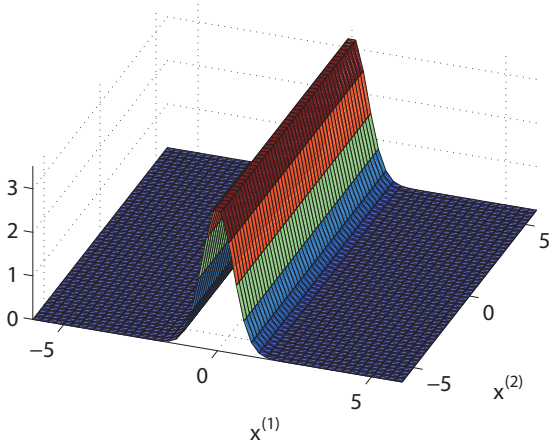
For the "highly-overlapped" dataset (Figure 9), LHSS gives a reasonable estimate of the hetero-distributional subspace, while LFDA is highly erroneous due to less separability. As a result, the density ratio function obtained by D^3-LFDA/uLSIF does not reflect the true redundant structure appropriately. On the other hand, D^3-LHSS still works well.

Finally, for the "dependent" dataset (Figure 10), LHSS gives an accurate estimate of the hetero-distributional subspace. However, LFDA gives a highly biased solution since the marginal distributions $p_{\mathrm{nu}}(v)$ and $p_{\mathrm{de}}(v)$ are no longer common in the "dependent" dataset. Consequently, the density ratio function obtained by D^3-LFDA/uLSIF is highly erroneous. In contrast, D^3-LHSS still works very well for the "dependent" dataset.

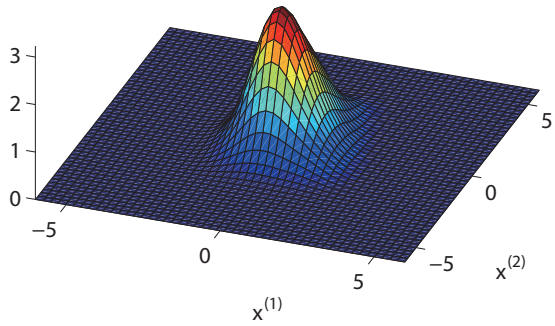
The experimental results for the "highly-overlapped" and "dependent" datasets illustrated typical failure modes of LFDA, and LHSS was shown to be able to successfully overcome these weaknesses of LFDA.



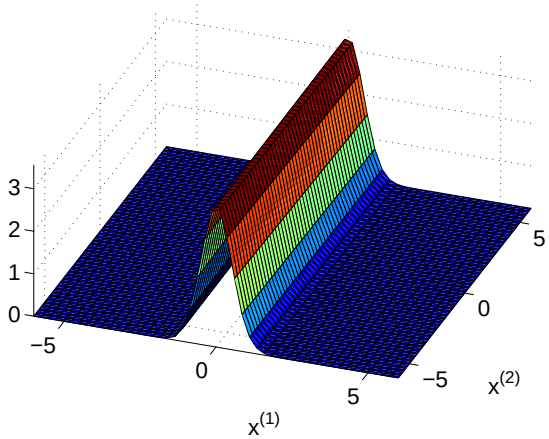
(a) Hetero-distributional subspace



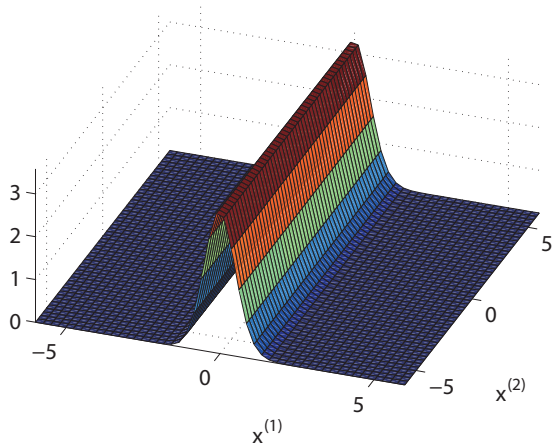
(b) $r(\boldsymbol{x})$



(c) $\widehat{r}(\boldsymbol{x})$ by plain uLSIF.

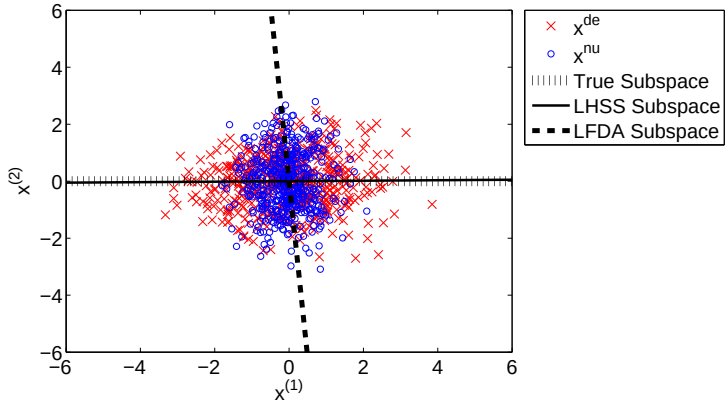


(d) $\widehat{r}(\boldsymbol{x})$ by D^3-LFDA/uLSIF.

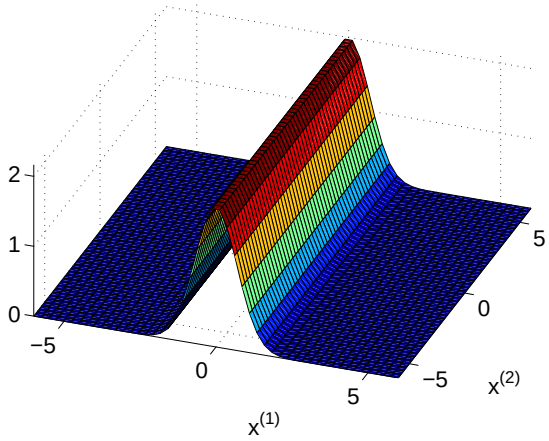


(e) $\widehat{r}(\boldsymbol{x})$ by D^3-LHSS.

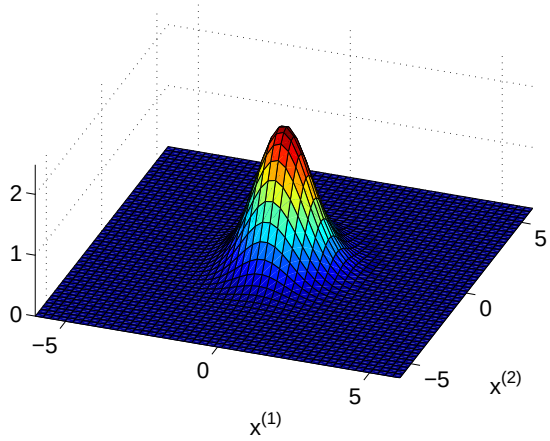
Figure 8: "Rather-separate" dataset.



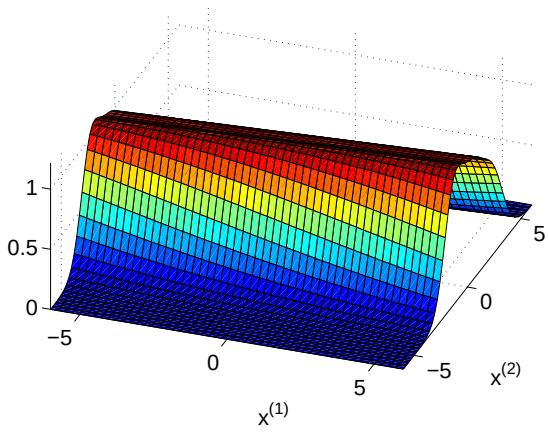
(a) Hetero-distributional subspace



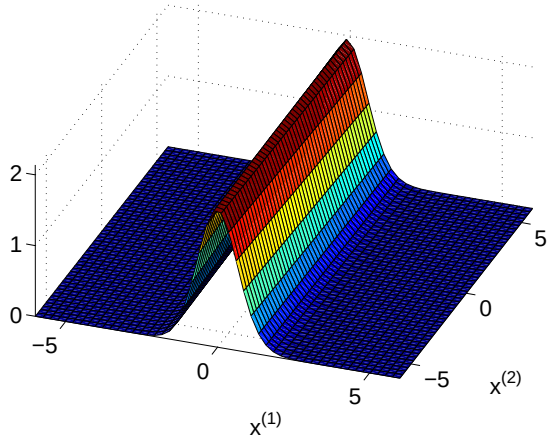
(b) $r(\boldsymbol{x})$



(c) $\widehat{r}(\boldsymbol{x})$ by plain uLSIF.

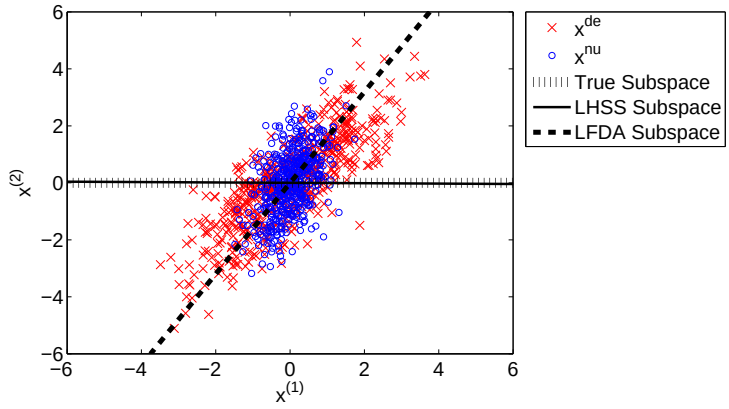


(d) $\widehat{r}(\boldsymbol{x})$ by D^3-LFDA/uLSIF.

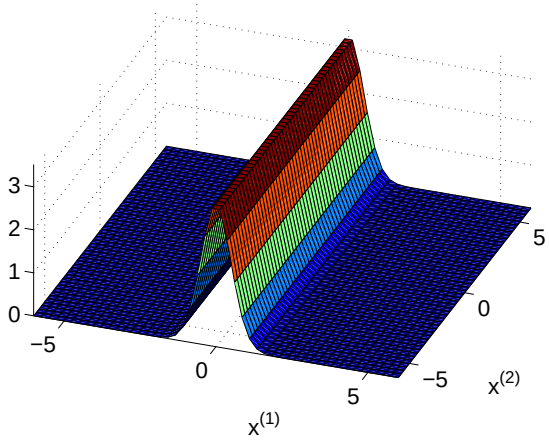


(e) $\widehat{r}(\boldsymbol{x})$ by D^3-LHSS.

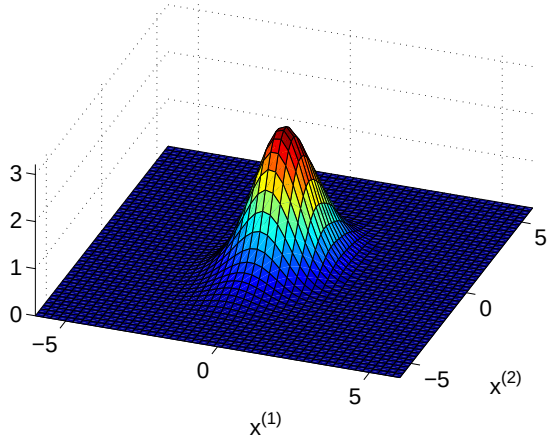
Figure 9: "Highly-overlapped" dataset.



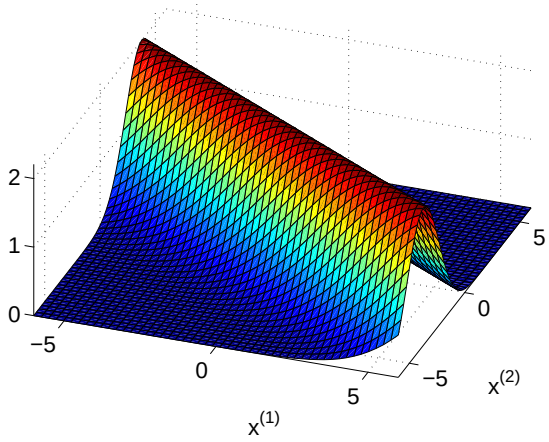
(a) Hetero-distributional subspace



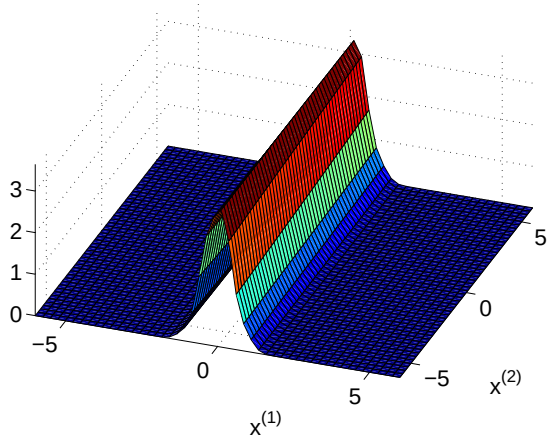
(b) $r(\boldsymbol{x})$



(c) $\widehat{r}(\boldsymbol{x})$ by plain uLSIF.



(d) $\widehat{r}(\boldsymbol{x})$ by D^3-LFDA/uLSIF.



(e) $\widehat{r}(\boldsymbol{x})$ by D^3-LHSS.

Figure 10: "Dependent" dataset.

4.2 Evaluation on Artificial Data

Next, we systematically compare the performance of the proposed D^3-LHSS with that of the plain uLSIF and D^3-LFDA/uLSIF for high-dimensional artificial data.

For the three datasets used in the previous experiments, we increase the entire dimensionality as $d = 2, 3, \ldots, 10$ by adding dimensions consisting of standard normal noise. The dimensionality of the hetero-distributional subspace is estimated based on the CV score of uLSIF. We evaluate the error of a density ratio estimator $\widehat{r}(\boldsymbol{x})$ by

$$\text{Error} := \frac{1}{2}\int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{\text{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{15}$$

which uLSIF tries to minimize (see Section 2.2).

The left graphs in Figure 11 show the density-ratio estimation error averaged over 100 runs as functions of the entire input dimensionality d. The best method in terms of the mean error and comparable methods according to the *t-test* (Henkel, 1979) at the significance level 1% are specified by '$\circ$'; otherwise methods are specified by '$\times$'.

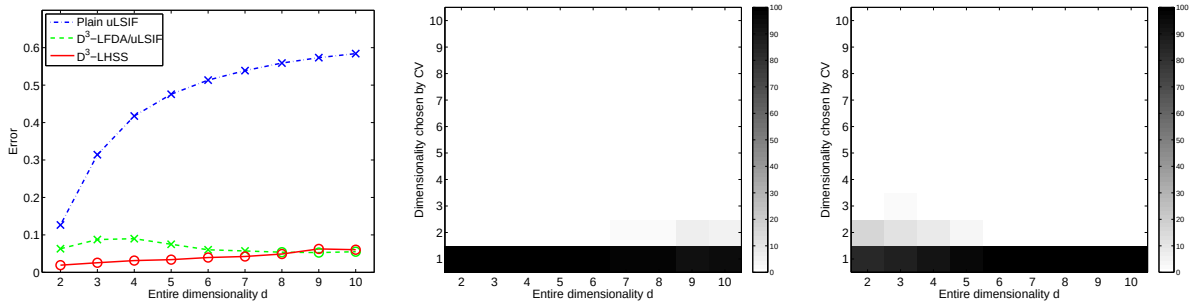
These plots show that, while the error of the plain uLSIF increases rapidly as the entire dimensionality d increases, that of the proposed D^3-LHSS is kept moderate. Consequently, the proposed method consistently outperforms the plain uLSIF. D^3-LHSS is comparable to D^3-LFDA/uLSIF for the "rather-separate" dataset, and D^3-LHSS significantly outperforms D^3-LFDA/uLSIF for the "highly-overlapped" and "dependent" datasets. Thus, D^3-LHSS was overall shown to compare favorably with the other approaches.

The choice of the dimensionality of the hetero-distributional subspace in D^3-LHSS and D^3-LFDA/uLSIF is illustrated in the middle and right columns of Figure 11; the darker the color is, the more frequently the corresponding dimensionality is chosen. The plots show that D^3-LHSS reasonably identifies the true dimensionality ($m = 1$ in the current setup) for all the three datasets, while D^3-LFDA/uLSIF performs well only for the "rather-separate" dataset. This happened because D^3-LFDA/uLSIF cannot find appropriate low-dimensional subspaces for the "highly-overlapped" and "dependent" datasets, and therefore the CV scores misled the choice of subspace dimensionality.

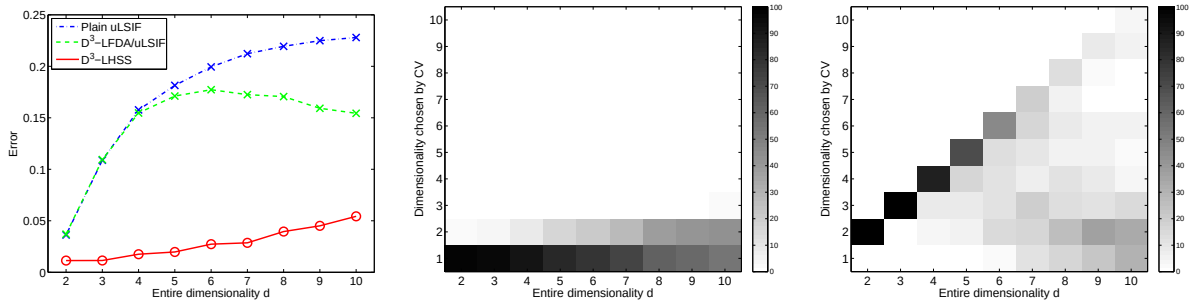
4.3 Inlier-based Outlier Detection for Benchmark Data

Finally, we apply the proposed method to inlier-based outlier detection, i.e., finding outliers in an evaluation dataset based on another "model" dataset that only contains inliers (see Section A.2 for details).

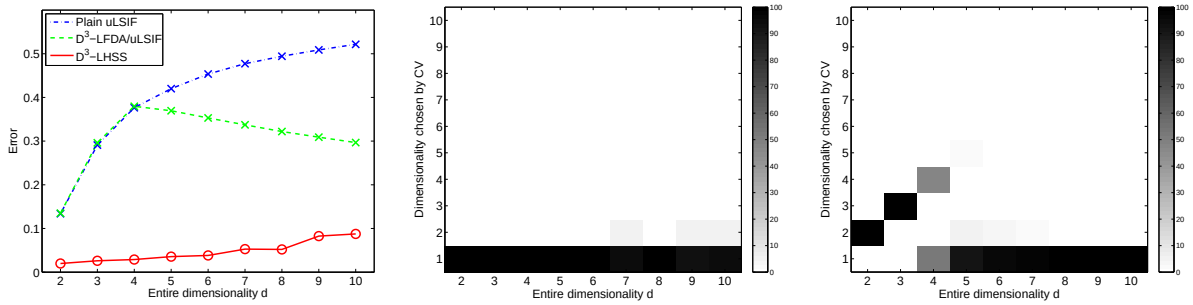
We use the *USPS hand-written digit dataset* taken from the *UCI Machine Learning Repository* (Asuncion & Newman, 2007). We regard samples in the class '1' as inliers and samples in other classes as outliers. We randomly take 500 samples from the class '1', and assign them to the model dataset. Then we randomly take 500 samples from the class '1' without overlap, and 25 samples from one of the other classes. From these samples, density ratio estimation is performed and the outlier score is computed. Since the USPS hand-written digit dataset contains 10 classes (i.e., from '0' to '9'), we have 9



(a) Density-ratio estimation error (b) Choice of Dimensionality by D^3-LHSS (c) Choice of Dimensionality by D^3-LFDA/uLSIF



(d) Density-ratio estimation error (e) Choice of Dimensionality by D^3-LHSS (f) Choice of Dimensionality by D^3-LFDA/uLSIF



(g) Density-ratio estimation error (h) Choice of Dimensionality by D^3-LHSS (i) Choice of Dimensionality by D^3-LFDA/uLSIF

Figure 11: Top: "Rather-separate" dataset. Middle: "Highly-overlapped" dataset. Bottom: "Dependent" dataset. Left: Density-ratio estimation error (15) averaged over 100 runs as a function of the entire data dimensionality d. The best method in terms of the mean error and comparable methods according to the *t-test* at the significance level 1% are specified by '$\circ$'; otherwise methods are specified by '$\times$'. Center: The dimensionality of the hetero-distributional subspace chosen by CV in LHSS. Right: The dimensionality of the hetero-distributional subspace chosen by CV in LFDA.

different tasks in total. The dimensionality of the samples is $d = 256$. For the D^3-LHSS and D^3-LFDA/uLSIF methods, we choose the dimensionality of the hetero-distributional subspace from $m = 1, 2, \ldots, 5$ by cross-validation.

When evaluating the performance of outlier detection methods, it is important to take into account both the *detection rate* (i.e., the amount of true outliers an outlier detection algorithm can find) and the *detection accuracy* (i.e., the amount of true inliers 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).

The mean and standard deviation of AUC scores over 100 runs with different random seeds are summarized in Table 1, where the best method in terms of the mean AUC and comparable methods according to the *t-test* at the significance level 1% are specified by '$\circ$'. The table shows that the proposed D^3-LHSS tends to outperform the plain uLSIF and D^3-LFDA/uLSIF. It is also note worthy that D^3-LFDA/uLSIF is actually outperformed by the plain uLSIF—the baseline method. This is perhaps because the numerator and denominator datasets are highly overlapped in outlier detection scenarios, so D^3-LFDA/uLSIF performs rather poorly (cf. Figure 9)

We also evaluate the performance of each method for an additional test dataset which is not used for density ratio estimation. The test dataset consists of 100 randomly chosen samples from the class '1' and 5 randomly chosen samples from the outlier class (which is the same as the evaluation dataset). The results are summarized in Table 2, showing that the advantage of the proposed method is still valid in this more challenging scenario.

5 Conclusions

Density ratios are becoming quantities of interest in the machine learning and data mining communities since it can be used for solving various important data processing tasks such as non-stationarity adaptation, outlier detection, and feature selection (Sugiyama et al., 2009; Sugiyama et al., 2011). In this paper, we tackled a challenging problem of estimating density ratios in high-dimensional spaces, and gave a new procedure in the framework of *Direct Density-ratio estimation with Dimensionality reduction* (D^3; D-cube). The basic idea of D^3 is to identify a subspace called the *hetero-distributional subspace*, in which two distributions (corresponding to the numerator and denominator of the density ratio) are different.

In the existing approach of D^3 (Sugiyama et al., 2010a), the hetero-distributional subspace is identified by finding a subspace in which samples drawn from the two distributions are maximally separated from each other. To this end, supervised dimensionality reduction methods such as *local Fisher discriminant analysis* (LFDA) (Sugiyama, 2007) are utilized. This approach was shown to work well when the components inside and outside the hetero-distributional subspace are statistically independent, and samples drawn from the two distributions are highly separable from each other in the hetero-distributional subspace.

Table 1: Outlier detection for the USPS hand-written digit dataset ($d = 256$). The means (and standard deviations in the bracket) of AUC scores over 100 runs for the evaluation dataset are summarized. The best method in terms of the mean AUC value and comparable methods according to the t-test at the significance level 1% are specified by '°'. The means (and standard deviations in the bracket) of the chosen dimensionality by cross-validation are also included in the table.

	D^3-LHSS		D^3-LFDA/uLSIF		Plain uLSIF
Data	AUC	$\widehat{m}$	AUC	$\widehat{m}$	AUC
Digit 2	°0.956 (0.035)	4.3 (0.8)	0.889 (0.104)	1.7 (1.1)	0.902 (0.038)
Digit 3	°0.967 (0.032)	4.4 (0.8)	0.868 (0.136)	1.8 (1.1)	0.921 (0.039)
Digit 4	°0.907 (0.061)	4.4 (0.9)	0.825 (0.104)	1.4 (0.6)	0.870 (0.036)
Digit 5	°0.965 (0.037)	4.3 (0.9)	0.882 (0.109)	1.6 (0.9)	0.906 (0.037)
Digit 6	°0.974 (0.022)	4.4 (0.8)	0.891 (0.090)	1.7 (1.1)	0.941 (0.029)
Digit 7	°0.924 (0.072)	4.4 (0.9)	0.642 (0.139)	2.3 (1.4)	0.878 (0.035)
Digit 8	°0.929 (0.051)	4.2 (1.0)	0.804 (0.147)	1.8 (1.1)	0.860 (0.033)
Digit 9	°0.942 (0.048)	4.6 (0.7)	0.790 (0.136)	1.8 (1.1)	0.892 (0.035)
Digit 0	°0.986 (0.019)	4.2 (0.9)	0.920 (0.071)	1.9 (0.8)	°0.979 (0.019)
Average	0.950 (0.051)	4.4 (0.9)	0.835 (0.142)	1.8 (1.1)	0.905 (0.049)

Table 2: Outlier detection for the USPS hand-written digit dataset ($d = 256$). The means (and standard deviations in the bracket) of AUC scores over 100 runs for unlearned test dataset are summarized.

	D^3-LHSS		D^3-LFDA/uLSIF		Plain uLSIF
Data	AUC	$\widehat{m}$	AUC	$\widehat{m}$	AUC
Digit 2	°0.946 (0.047)	4.3 (0.8)	0.817 (0.132)	1.7 (1.1)	0.905 (0.044)
Digit 3	°0.953 (0.061)	4.4 (0.8)	0.780 (0.161)	1.8 (1.1)	0.924 (0.045)
Digit 4	°0.880 (0.094)	4.4 (0.9)	0.767 (0.121)	1.4 (0.6)	°0.870 (0.063)
Digit 5	°0.954 (0.057)	4.3 (0,9)	0.813 (0.142)	1.6 (0.9)	0.906 (0.047)
Digit 6	°0.959 (0.052)	4.4 (0.8)	0.806 (0.141)	1.7 (1.1)	0.939 (0.040)
Digit 7	°0.909 (0.079)	4.4 (0.9)	0.689 (0.173)	2.3 (1.4)	0.877 (0.056)
Digit 8	°0.903 (0.078)	4.2 (1.0)	0.741 (0.173)	1.8 (1.1)	0.861 (0.049)
Digit 9	°0.932 (0.072)	4.6 (0.7)	0.793 (0.128)	1.8 (1.1)	0.894 (0.054)
Digit 0	°0.982 (0.039)	4.2 (0.9)	0.859 (0.098)	1.9 (0.8)	°0.982 (0.022)
Average	0.935 (0.073)	4.4 (0.9)	0.785 (0.150)	1.8 (1.1)	0.906 (0.060)

However, as illustrated in Section 4.1, violation of these conditions can cause significant performance degradation. This problem can be overcome in principle by finding a subspace such that two *conditional* distributions are similar to each other in its complementary subspace. However, comparing conditional distributions is a cumbersome task. To cope with this problem, we first proved that the hetero-distributional subspace can be characterized as the subspace in which two *marginal* distributions are maximally different under the *Pearson divergence* (Lemma 1). Based on this lemma, we proposed a new algorithm for finding the hetero-distributional subspace called *Least-squares Hetero-distributional Subspace Search* (LHSS). Since a density-ratio estimation method is utilized during hetero-distributional subspace search in the LHSS procedure, an additional density-ratio estimation step is not needed after hetero-distributional subspace search. Thus, two steps in the previous method (hetero-distributional subspace search followed by density ratio estimation in the identified subspace) were merged into a single step (see Figure 2). The proposed single-shot procedure, *D*^{3}*-LHSS* (D-cube LHSS), was shown to be able to overcome the limitations of the D^3-LFDA/uLSIF approach through experiments.

In the experiments in Section 4, we employed the subspace rotation algorithm explained in Section 3.3.4 in our D^3-LHSS implementation. Although we experimentally found that the subspace rotation algorithm is useful, this does not necessarily mean that subspace rotation is always the best performing algorithm. Other approaches explained in Section 3.3 may also be useful in some situations. Further investigating the optimization issue is an important future work.

We gave a general proof of the data processing inequality (Lemma 1) for a class of f-divergences (Ali & Silvey, 1966; Csiszár, 1967). Thus, the hetero-distributional subspace is characterized not only by the Pearson divergence, but also by *any* f-divergences. Since a framework of density ratio estimation for f-divergences has been provided in Nguyen et al. (2010), an interesting future direction is to develop hetero-distributional subspace search methods for general f-divergences.

Acknowledgments

MS was supported by SCAT, AOARD, and the JST PRESTO program. MY was supported by the JST PRESTO program. We thank Satoshi Hara for having performed preliminary experiments using an earlier version of the proposed method. Our special thanks also go to anonymous reviewers for their comments.

References

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

Akiyama, T., Hachiya, H., & Sugiyama, M. (2010). Efficient exploration through active

learning for value function approximation in reinforcement learning. *Neural Networks*, *23*, 639–648.

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

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

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

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Bensaid, N., & Fabre, J. P. (2007). Optimal asymptotic quadratic error of kernel estimators of Radon-Nikodym derivatives for strong mixing data. *Journal of Nonparametric Statistics*, *19*, 77–88.

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).

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning (ICML2007)* (pp. 81–88).

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

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

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.

Chen, S.-M., Hsu, Y.-S., & Liaw, J.-T. (2009). On kernel estimators of density ratio. *Statistics*, *43*, 463–479.

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. (2006). *Elements of information theory*. Hoboken, NJ, USA: John Wiley & Sons, Inc. 2nd edition.

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

Ćwik, J., & Mielniczuk, J. (1989). Estimating density ratio with application to discriminant analysis. *Communications in Statistics: Theory and Methods*, *18*, 3057–3069.

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

Gijbels, I., & Mielniczuk, J. (1995). Asymptotic properties of kernel estimators of the Radon-Nikodym derivative with applications to discriminant analysis. *Statistica Sinica*, *5*, 261–278.

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

Hachiya, H., Akiyama, T., Sugiyama, M., & Peters, J. (2009a). Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, *22*, 1399–1410.

Hachiya, H., Peters, J., & Sugiyama, M. (2009b). Efficient sample reuse in EM-based policy search. *Machine Learning and Knowledge Discovery in Databases* (pp. 469–484). Berlin: Springer.

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

Henkel, R. E. (1979). *Tests of significance*. Beverly Hills, CA, USA.: 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.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2010). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*. to appear.

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, USA: MIT Press.

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

Jacob, P., & Oliveira, P. E. (1997). Kernel estimators of general Radon-Nikodym derivatives. *Statistics*, *30*, 25–46.

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

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). arXiv.

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.

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.

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

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

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

Petersen, K. B., & Pedersen, M. S. (2008). *The matrix cookbook* (Technical Report). Technical University of Denmark.

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

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C*. Cambridge, UK: Cambridge University Press. 2nd edition.

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, Massachusetts, USA: MIT Press.

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

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *Journal of the Royal Statistical Society, Series C*, *27*, 26–33.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall.

Smola, A., Song, L., & Teo, C. H. (2009). Relative novelty detection. *Proceedings of Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)* (pp. 536–543). Clearwater Beach, FL, USA.

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, *36*, 111–147.

Storkey, A., & Sugiyama, M. (2007). Mixture regression for covariate shift. *Advances in Neural Information Processing Systems 19* (pp. 1337–1344). Cambridge, Massachusetts, USA: 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. (2010). Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, *E93-D*, 2690–2701.

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

Sugiyama, M., & Kawanabe, M. (2010). *Covariate shift adaptation: Towards machine learning in non-stationary environment*. Cambridge, Massachusetts, USA: MIT Press. to appear.

Sugiyama, M., Kawanabe, M., & Chui, P. L. (2010a). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

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., & Kanamori, T. (2010b). Density ratio estimation: A comprehensive review. *RIMS Kokyuroku*, 10–31.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2011). *Density ratio estimation in machine learning: A versatile tool for statistical data processing*. Cambridge, UK: Cambridge University Press.

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., & Okanohara, D. (2010c). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, *E93-D*, 583–594.

Sun, J., & Woodroofe, M. (1997). Semi-parametric estimates under biased sampling. *Statistica Sinica*, *7*, 545–575.

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

Suzuki, T., & Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)* (pp. 804–811). Sardinia, Italy.

Suzuki, T., Sugiyama, M., Kanamori, T., & Sese, J. (2009a). 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. *Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008)* (pp. 5–20). Antwerp, Belgium.

Suzuki, T., Sugiyama, M., & Tanaka, T. (2009b). Mutual information approximation via maximum likelihood estimation of density ratio. *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)* (pp. 463–467). Seoul, Korea.

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.

Ueki, K., Sugiyama, M., & Ihara, Y. (2010). Perceived age estimation under lighting condition change by covariate shift adaptation. *20th International Conference on Pattern Recognition (ICPR2010)* (pp. 3400–3403). Istanbul, Turkey.

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

Wahba, G. (1990). *Spline models for observational data*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

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. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)* (pp. 643–648). Atlanta, Georgia, USA: The AAAI Press.

Yamada, M., Sugiyama, M., & Matsui, T. (2010). Semi-supervised speaker identification under covariate shift. *Signal Processing*, *90*, 2353–2361.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML2004)* (pp. 903–910). New York, NY, USA: ACM Press.

A Usage of Density Ratios in Data Processing

We are interested in estimating density ratios since they are useful in various data processing tasks. Here, we briefly review possible usage of density ratios (Sugiyama et al., 2009; Sugiyama et al., 2011).

A.1 Covariate Shift Adaptation

Covariate shift (Shimodaira, 2000) is a situation in supervised learning where input distributions change between the training and test phases, but the conditional distribution of outputs given inputs remains unchanged. *Extrapolation* (i.e., prediction is made outside the training region) would be a typical example of covariate shift. Standard learning techniques such as maximum likelihood estimation are biased under covariate shift; the bias caused by covariate shift can be asymptotically canceled by weighting the loss function according to the *importance*[2] (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Sugiyama et al., 2007; Quiñonero-Candela et al., 2009; Sugiyama & Kawanabe, 2010).

[2]The test input density over the training input density is referred to as the importance in the context of *importance sampling* (Fishman, 1996).

The basic idea of covariate shift adaptation is summarized in the following importance sampling identity:

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

That is, the expectation of a function $\mathrm{loss}(\boldsymbol{x})$ over $p_{\mathrm{nu}}(\boldsymbol{x})$ can be computed by the importance-weighted expectation of $\mathrm{loss}(\boldsymbol{x})$ over $p_{\mathrm{de}}(\boldsymbol{x})$. Similarly, standard model selection criteria such as *cross-validation* (Stone, 1974; Wahba, 1990) or *Akaike's information criterion* (Akaike, 1974) lose their unbiasedness under 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). Furthermore, the performance of *active learning* or the *experiment design*, i.e., 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 used 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., 2009a; Akiyama et al., 2010; Hachiya et al., 2009b), speaker identification (Yamada et al., 2010), age prediction from face images (Ueki et al., 2010), wafer alignment in semiconductor exposure apparatus (Sugiyama & Nakajima, 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).

A.2 Inlier-based Outlier Detection

Let us consider an outlier detection problem 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; Smola et al., 2009; Hido et al., 2010). Since the evaluation dataset usually 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 (Kawahara & Sugiyama, 2009).

A.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 $\boldsymbol{x}$. Then a density-ratio estimation method directly gives an estimate of the conditional density (Sugiyama et al., 2010c).

When $\boldsymbol{y}$ is categorical, the same method can be used for probabilistic classification (Sugiyama, 2010).

A.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* $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y})$ between random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined by

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

which plays a central role in *information theory* (Cover & Thomas, 2006).

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, i.e.,

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

Then mutual information can be directly estimated using a density-ratio estimation method (Suzuki et al., 2008; Suzuki et al., 2009b). General divergence functionals can also be estimated in a similar way (Nguyen et al., 2010).

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 density-ratio estimation methods are applicable, e.g., to variable selection (Suzuki et al., 2009a), independent component analysis (Suzuki & Sugiyama, 2009), supervised dimensionality reduction (Suzuki & Sugiyama, 2010), and causal inference (Yamada & Sugiyama, 2010).

B Proof of Lemma 1

Here, let us consider the *f-divergences* (Ali & Silvey, 1966; Csiszár, 1967) and prove a similar inequality for a broader class of divergences. An f-divergence is defined using a convex function f such that $f(1) = 0$ as

$$I_f[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] := \int p_{\mathrm{de}}(\boldsymbol{x}) f\left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}.$$

The f-divergence is reduced to the Kullback-Leibler divergence if

$$f(t) = -\log t,$$

and the Pearson divergence if

$$f(t) = \frac{1}{2}(t-1)^2.$$

Using *Jensen's inequality* (Bishop, 2006), we have

$$\begin{aligned}
I_f[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] &= \iint p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}) p_{\mathrm{de}}(\boldsymbol{u}) f\left(\frac{p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}) p_{\mathrm{de}}(\boldsymbol{u})}\right) \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v} \\
&\geq \int p_{\mathrm{de}}(\boldsymbol{u}) f\left(\int p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}) \frac{p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}) p_{\mathrm{de}}(\boldsymbol{u})} \mathrm{d}\boldsymbol{v}\right) \mathrm{d}\boldsymbol{u} \\
&= \int p_{\mathrm{de}}(\boldsymbol{u}) f\left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} \int p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) \mathrm{d}\boldsymbol{v}\right) \mathrm{d}\boldsymbol{u} \\
&= \int p_{\mathrm{de}}(\boldsymbol{u}) f\left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}\right) \mathrm{d}\boldsymbol{u} \\
&= I_f[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})].
\end{aligned}$$

Thus, we have

$$I_f[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] - I_f[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] \geq 0,$$

and the equality holds if and only if $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u})$. ■

C Proof of Lemma 2

For

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

$\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ can be expressed as

$$\begin{aligned}
\widehat{\mathrm{PD}}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] &= \frac{1}{2} \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \widehat{h}_\ell - \frac{1}{2} \\
&= \frac{1}{2} \sum_{\ell,\ell'=1}^{b} \widehat{h}_\ell \widehat{h}_{\ell'} F_{\ell,\ell'} - \frac{1}{2}.
\end{aligned}$$

Thus, its partial derivative with respect to $\boldsymbol{U}$ is given by

$$\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{U}} = \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \frac{\partial\widehat{h}_\ell}{\partial\boldsymbol{U}} + \frac{1}{2}\sum_{\ell,\ell'=1}^{b} \widehat{h}_\ell \widehat{h}_{\ell'} \frac{\partial F_{\ell,\ell'}}{\partial\boldsymbol{U}}. \tag{16}$$

Since

$$\frac{\partial\boldsymbol{B}^{-1}}{\partial t} = -\boldsymbol{B}^{-1}\frac{\partial\boldsymbol{B}}{\partial t}\boldsymbol{B}^{-1}$$

holds for a square invertible matrix $\boldsymbol{B}$ (Petersen & Pedersen, 2008), it holds that

$$\frac{\partial\boldsymbol{F}}{\partial U_{k,k'}} = -(\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\frac{\partial\widehat{\boldsymbol{H}}}{\partial U_{k,k'}}(\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}.$$

Then we have

$$\begin{aligned}\sum_{\ell,\ell'=1}^{b} \widehat{h}_\ell\widehat{h}_{\ell'}\left[\frac{\partial\boldsymbol{F}}{\partial U_{k,k'}}\right]_{\ell,\ell'} &= -\widehat{\boldsymbol{h}}^\top(\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\frac{\partial\widehat{\boldsymbol{H}}}{\partial U_{k,k'}}(\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}} \\ &= -\sum_{\ell,\ell'=1}^{b}\widehat{\alpha}_\ell\widehat{\alpha}_{\ell'}\left[\frac{\partial\widehat{\boldsymbol{H}}}{\partial U_{k,k'}}\right]_{\ell,\ell'}.\end{aligned}$$

Substituting this into Eq.(16), we obtain Eq.(8). Eqs.(9) and (10) are clear from Eqs.(3) and (2). Finally, we prove Eq.(11). The basis function $\psi_\ell(\boldsymbol{u})$ can be expressed as

$$\psi_\ell(\boldsymbol{u}) = \psi_\ell(\boldsymbol{U}\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{U}(\boldsymbol{x}-\boldsymbol{c}'_\ell)\|^2}{2\sigma^2}\right).$$

Since $\frac{\partial\boldsymbol{a}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{a}}{\partial\boldsymbol{A}} = 2\boldsymbol{A}\boldsymbol{a}^\top\boldsymbol{a}$ (Petersen & Pedersen, 2008), we have

$$\frac{\partial\psi_\ell(\boldsymbol{u})}{\partial\boldsymbol{U}} = -\frac{1}{\sigma^2}\boldsymbol{U}(\boldsymbol{x}-\boldsymbol{c}'_\ell)(\boldsymbol{x}-\boldsymbol{c}'_\ell)^\top\exp\left(-\frac{\|\boldsymbol{U}(\boldsymbol{x}-\boldsymbol{c}'_\ell)\|^2}{2\sigma^2}\right),$$

from which we obtain Eq.(11). ■

D Proof of Lemma 3

The proof we provide here essentially follows the argument in Plumbley (2005).

For

$$\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}, \quad \boldsymbol{W}_0 = \begin{bmatrix}\boldsymbol{U}_0\\ \boldsymbol{V}_0\end{bmatrix},$$

rotation $\boldsymbol{W}$ from some $\boldsymbol{W}_0$ can be expressed as follows (Plumbley, 2005):

$$\boldsymbol{W} = \exp(\boldsymbol{M})\boldsymbol{W}_0, \tag{17}$$

where $\boldsymbol{M}$ is some skew-symmetric matrix. Let us consider the space of skew-symmetric matrices, and let $\boldsymbol{E}$ be an element in that space with unit length. Then the gradient of a function $\widehat{\mathrm{PD}}(\boldsymbol{M})$ with respect to $\boldsymbol{M}$, $\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{M}}$, in this space is given as an element whose inner product $\left\langle \frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{M}}, \boldsymbol{E} \right\rangle$ is equal to the derivative of $\widehat{\mathrm{PD}}(\boldsymbol{M})$ in the direction $\boldsymbol{E}$ (Plumbley, 2005). Thus, for $\boldsymbol{M} = t\boldsymbol{E}$ with t being a scalar, we have

$$\frac{\partial\widehat{\mathrm{PD}}(t\boldsymbol{E})}{\partial t} = \left\langle \frac{\partial\widehat{\mathrm{PD}}(\boldsymbol{M})}{\partial\boldsymbol{M}}, \boldsymbol{E} \right\rangle.$$

If we adopt Eq.(12) as the inner product of the space of skew-symmetric matrices, we have

$$\frac{\partial\widehat{\mathrm{PD}}(t\boldsymbol{E})}{\partial t} = \frac{1}{2}\mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}(\boldsymbol{M})}{\partial\boldsymbol{M}}\boldsymbol{E}^\top\right). \tag{18}$$

On the other hand, from Eq.(17) with $\boldsymbol{M} = t\boldsymbol{E}$, $\frac{\partial\boldsymbol{W}}{\partial t}$ can be expressed as follows (Petersen & Pedersen, 2008):

$$\frac{\partial\boldsymbol{W}}{\partial t} = \boldsymbol{E}\exp(t\boldsymbol{E})\boldsymbol{W}_0 = \boldsymbol{E}\boldsymbol{W}.$$

Then $\frac{\partial\widehat{\mathrm{PD}}(t\boldsymbol{E})}{\partial t}$ can be expressed as

$$\frac{\partial\widehat{\mathrm{PD}}(t\boldsymbol{E})}{\partial t} = \mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\frac{\partial\boldsymbol{W}}{\partial t}^\top\right) = \mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\boldsymbol{W}^\top\boldsymbol{E}^\top\right). \tag{19}$$

Since $\boldsymbol{E}$ is skew-symmetric, it can be expressed as

$$\boldsymbol{E} = \frac{1}{2}\boldsymbol{E} + \frac{1}{2}\boldsymbol{E} = \frac{1}{2}\boldsymbol{E} - \frac{1}{2}\boldsymbol{E}^\top.$$

Substituting this into Eq.(19), we have

$$\begin{aligned}
\frac{\partial\widehat{\mathrm{PD}}(t\boldsymbol{E})}{\partial t} &= \frac{1}{2}\mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\boldsymbol{W}^\top\boldsymbol{E}^\top\right) - \frac{1}{2}\mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\boldsymbol{W}^\top\boldsymbol{E}\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\boldsymbol{W}^\top\boldsymbol{E}^\top\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{W}\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}^\top\boldsymbol{E}^\top\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}\boldsymbol{W}^\top - \boldsymbol{W}\frac{\partial\widehat{\mathrm{PD}}}{\partial\boldsymbol{W}}^\top\right)\boldsymbol{E}^\top\right).
\end{aligned} \tag{20}$$

Combining Eqs.(18) and (20), we have

$$\begin{aligned}\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{M}} &= \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{W}}\boldsymbol{W}^\top - \boldsymbol{W}\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{W}}^\top \\ &= \begin{bmatrix} \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{U}^\top - \boldsymbol{U}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\right)^\top & \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top - \boldsymbol{U}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{V}}\right)^\top \\ \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{V}}\boldsymbol{U}^\top - \boldsymbol{V}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\right)^\top & \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{V}}\boldsymbol{V}^\top - \boldsymbol{V}\left(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{V}}\right)^\top \end{bmatrix}. \end{aligned} \tag{21}$$

Eq.(11) implies that $\frac{\partial\psi_\ell(\boldsymbol{u})}{\partial \boldsymbol{U}}\boldsymbol{U}^\top$ is symmetric. Then Eqs.(8) and (9) imply that $\frac{\partial\widehat{h}_\ell}{\partial \boldsymbol{U}}\boldsymbol{U}^\top$ and $\frac{\partial\widehat{H}_{\ell,\ell'}}{\partial \boldsymbol{U}}\boldsymbol{U}^\top$ are also symmetric. Consequently, Eq.(10) imply that $\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{U}^\top$ is symmetric:

$$\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{U}^\top = \left(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{U}^\top\right)^\top = \boldsymbol{U}\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}^\top.$$

Since the range of $\boldsymbol{V}$ is assumed to be orthogonal to the range of $\boldsymbol{U}$ (see Section 3.1), $\widehat{\mathrm{PD}}$ is independent of $\boldsymbol{V}$, and thus we have

$$\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{V}} = \boldsymbol{O}_{(d-m),d},$$

where $\boldsymbol{O}_{d,d'}$ is the $d \times d'$ matrix with all zeros. Then Eq.(21) yields

$$\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{M}} = \begin{bmatrix} \boldsymbol{O}_{m,m} & \frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top \\ -(\frac{\partial\widehat{\mathrm{PD}}}{\partial \boldsymbol{U}}\boldsymbol{V}^\top)^\top & \boldsymbol{O}_{(d-m),(d-m)} \end{bmatrix},$$

which concludes the proof. ■