Direct Density-Ratio Estimation with Dimensionality Reduction via Hetero-Distributional Subspace Analysis *

Makoto Yamada† and **Masashi Sugiyama**†‡
†Department of Computer Science, Tokyo Institute of Technology
‡Japan Science and Technology Agency
{yamada@sg. sugi@}cs.titech.ac.jp

Abstract

Methods for 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, feature selection, and conditional probability estimation. In this paper, we propose a new density-ratio estimator which incorporates *dimensionality reduction* into the density-ratio estimation procedure. Through experiments, the proposed method is shown to compare favorably with existing density-ratio estimators in terms of both accuracy and computational costs.

1 Introduction

During recent years, it has been shown that several machine learning and data mining tasks can be formulated with use of the ratio of two probability density functions (Sugiyama et al. 2009). Examples of such tasks are covariate shift adaptation (Shimodaira 2000), transfer learning (Storkey and Sugiyama 2007), multi-task learning (Bickel et al. 2008), outlier detection (Smola, Song, and Teo 2009; Hido et al. 2011), privacy-preserving data mining (Elkan 2010), feature selection (Suzuki et al. 2009), supervised dimensionality reduction (Suzuki and Sugiyama 2010), and causal inference (Yamada and Sugiyama 2010). For this reason, density ratio estimation has attracted a great deal of attention from machine learning and data mining communities and various approaches have been explored (Silverman 1978; Qin 1998; Huang et al. 2007; Sugiyama et al. 2008; Kanamori, Hido, and Sugiyama 2009; Nguyen, Wainwright, and Jordan 2010).

A naive way of density ratio estimation is to first estimate the two densities in the ratio (i.e., the numerator and the denominator) separately using a density estimator such as *kernel density estimation* (Silverman 1986; Lee and Gray 2006; Raykar, Duraiswami, and Zhao 2010), and then take the ratio of the estimated densities. However, this two-step approach is not reasonable since division by an estimated density tends to increase the estimation error of the dividend. To improve the estimation accuracy, direct density-ratio estimation methods (i.e., the density ratio is estimated without going through density estimation) were proposed recently such as the moment matching method using reproducing kernels called *kernel mean matching* (KMM) (Huang et al. 2007), the method based on *logistic regression* (LR) (Qin 1998), the distribution matching method under the *Kullback-Leibler (KL) divergence* (Kullback and Leibler 1951) called the *KL importance estimation procedure* (KLIEP) (Sugiyama et al. 2008; Nguyen, Wainwright, and Jordan 2010), and the density-ratio matching methods under the squared-loss called *least-squares importance fitting* (LSIF) and *unconstrained LSIF* (uLSIF) (Kanamori, Hido, and Sugiyama 2009). Through extensive experiments, direct density-ratio estimation methods have been shown to compare favorably with a naive two-step approach based on kernel density estimation.

Although the density ratio is estimated directly without going through density estimation, density ratio estimation in high-dimensional cases is still challenging. To deal with this issue, an approach called *Direct Density-ratio estimation with Dimensionality reduction* (D^3; D-cube) has been proposed (Sugiyama, Kawanabe, and Chui 2010). The key idea of D^3 is to find a subspace in which the numerator and denominator densities are significantly different (which is called the *hetero-distributional subspace*); then density ratio estimation is performed in this subspace.

The hetero-distributional subspace can be identified by the subspace in which two distributions are maximally separated. Based on this idea, a D^3 method called D^3-LFDA/uLSIF was proposed (Sugiyama, Kawanabe, and Chui 2010), which employs a supervised dimensionality reduction method called *local Fisher discriminant analysis* (LFDA) (Sugiyama 2007) for hetero-distributional subspace search; then the density ratio is estimated in the subspace by uLSIF. This method is computationally very efficient since LFDA and uLSIF both provide analytic-form solutions. However, maximum separability between two distributions does not necessarily imply that the two distributions are different. Thus, there exist cases in which D^3-LFDA/uLSIF cannot identify the correct hetero-distributional subspace.

To overcome this weakness, a new method called D^3-*least-squares hetero-distributional subspace search* (D^3-LHSS) was proposed (Sugiyama et al. 2011). D^3-LHSS searches the hetero-distributional subspace more directly so

*This work was partially supported by SCAT, AOARD, and the JST PRESTO program.

that the difference between two distributions in the subspace is maximized. Thanks to this direct formulation, D^3-LHSS can find *any* hetero-distributional subspace in principle. However, D^3-LHSS resorts to a gradient-based optimization scheme for subspace search, and thus it is computationally demanding.

The purpose of this paper is to improve the computational efficiency of D^3-LHSS. Our new method, which we call D^3-*hetero-distributional subspace analysis* (D^3-HSA), gives an analytic-form solution in each iteration of hetero-distributional subspace search, and thus is computationally more efficient than D^3-LHSS. Moreover, based on the above analytic-form solution, we develop a method to design a good initial value for optimization, which further contributes to reducing the computational cost and helps improving the estimation accuracy. Through experiments, we show that the proposed D^3-HSA approach is promising.

2 Problem Formulation

In this section, we describe the D^3 framework (Sugiyama, Kawanabe, and Chui 2010; Sugiyama et al. 2011).

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 $p_{\mathrm{de}}(\boldsymbol{x})$. Here, the subscripts 'nu' and 'de' denote 'numerator' and 'denominator', respectively. We assume that the latter density $p_{\mathrm{de}}(\boldsymbol{x})$ is strictly positive. The goal 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 $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$.

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

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

where $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{(d-m)\times d}$ are transformation matrices; the row vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ are assumed to form an orthonormal basis, i.e., $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonally complementary to each other. Then the two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ can always be decomposed as

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

A key assumption of the D^3 framework 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., $p_{\mathrm{nu}}(\boldsymbol{v}|\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{v}|\boldsymbol{u}) = p(\boldsymbol{v}|\boldsymbol{u})$. Then, the density-ratio can be simplified as

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u})}{p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{de}}(\boldsymbol{u})} = \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} =: r(\boldsymbol{u}). \tag{1}$$

This expression implies that the density ratio $r(\boldsymbol{x})$ 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 $\boldsymbol{U}$. For the moment, we assume that the true dimensionality m of the hetero-distributional subspace is known. How to estimate m from data is explained in Section 3.5.

3 Proposed Method: D^3-HSA

In this section, we describe our proposed method called the direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis (D^3-HSA).

3.1 PE Estimation-Maximization Framework

It was shown (Sugiyama et al. 2011) that the optimal transformation matrix that fulfills Eq.(1) can be characterized as

$$\boldsymbol{U}^* = \underset{\boldsymbol{U}\in\mathbb{R}^{m\times d}}{\mathrm{argmax}} \; \mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] \;\text{ s.t. } \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_m, \tag{2}$$

where $\top$ denotes the transpose of a matrix or a vector and $\boldsymbol{I}_m$ is the m-dimensional identity matrix. In the above, $\mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ is the *Pearson divergence* (PE) from $p_{\mathrm{nu}}(\boldsymbol{u})$ to $p_{\mathrm{de}}(\boldsymbol{u})$:

$$\begin{aligned}\mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] &:= \frac{1}{2}\int\left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}-1\right)^2 p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\\ &= \frac{1}{2}\int\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \frac{1}{2}. \end{aligned} \tag{3}$$

Note that $\mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})]$ vanishes if and only if $p_{\mathrm{nu}}(\boldsymbol{u}) = p_{\mathrm{de}}(\boldsymbol{u})$.

Based on Eq.(2), we develop the following iterative algorithm for learning $r(\boldsymbol{u})$:

(i) Initialization: Initialize the transformation matrix $\boldsymbol{U}$ (see Section 3.4).

(ii) PE estimation: For current $\boldsymbol{U}$, a PE estimator $\widehat{\mathrm{PE}}$ is obtained (see Section 3.2).

(iii) PE maximization: Given a PE estimator $\widehat{\mathrm{PE}}$, its maximizer with respect to $\boldsymbol{U}$ is obtained (see Section 3.3).

(iv) Convergence check: The above (ii) and (iii) are repeated until $\boldsymbol{U}$ fulfills some convergence criterion.

(v) Final density-ratio estimation: Obtain $\widehat{r}(\boldsymbol{u})$ under the learned transformation matrix $\boldsymbol{U}$ (see Section 3.5).

3.2 PE Estimation

In HSA, we employ a non-parametric PE estimator derived in Sugiyama et al. (2011), which is based on a density-ratio estimator called *unconstrained Least-squares Importance Fitting* (uLSIF) (Kanamori, Hido, and Sugiyama 2009). uLSIF was shown to achieve the optimal non-parametric convergence rate and the optimal numerical stability (Kanamori, Suzuki, and Sugiyama 2009). Below, we briefly describe the PE estimator. Let $\boldsymbol{u}_i^{\mathrm{nu}} = \boldsymbol{U}\boldsymbol{x}_i^{\mathrm{nu}}$ and $\boldsymbol{u}_j^{\mathrm{de}} = \boldsymbol{U}\boldsymbol{x}_j^{\mathrm{de}}$.

We model the density-ratio function $r(\boldsymbol{u})$ by

$$\sum_{i=1}^{n_{\mathrm{nu}}}\alpha_i K(\boldsymbol{u}, \boldsymbol{u}_i) = \boldsymbol{\alpha}^\top\boldsymbol{k}(\boldsymbol{u}), \tag{4}$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{n_{\mathrm{nu}}})^\top$ are parameters to be learned from data samples, $\boldsymbol{k}(\boldsymbol{u}) = (K(\boldsymbol{u}, \boldsymbol{u}_1), \ldots, K(\boldsymbol{u}, \boldsymbol{u}_{n_{\mathrm{nu}}}))^\top$ are the basis functions, and $K(\boldsymbol{u}, \boldsymbol{u}')$ is a kernel function.

The parameter $\boldsymbol{\alpha}$ is learned so that the following squared error is minimized (Kanamori, Hido, and Sugiyama 2009):

$$J_0(\boldsymbol{\alpha}) = \frac{1}{2}\int \left(\boldsymbol{\alpha}^\top \boldsymbol{k}(\boldsymbol{u}) - r(\boldsymbol{u})\right)^2 p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} = J(\boldsymbol{\alpha}) + C,$$

where C is a constant, and

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha}, \tag{5}$$

$$\boldsymbol{H} = \int \boldsymbol{k}(\boldsymbol{u})\boldsymbol{k}(\boldsymbol{u})^\top p_{\mathrm{de}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \;\; \boldsymbol{h} = \int \boldsymbol{k}(\boldsymbol{u}) p_{\mathrm{nu}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}.$$

Approximating the expectations in $\boldsymbol{H}$ and $\boldsymbol{h}$ included in J by empirical averages, we arrive at the following optimization problem:

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right],$$

where a regularization term $\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$ is included for avoiding overfitting, λ (≥ 0) is a regularization parameter, and

$$\widehat{\boldsymbol{H}} = \frac{1}{n_{\mathrm{de}}^2}\sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{k}(\boldsymbol{u}_j^{\mathrm{de}})\boldsymbol{k}(\boldsymbol{u}_j^{\mathrm{de}})^\top, \;\; \widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{k}(\boldsymbol{u}_i^{\mathrm{nu}}).$$

Differentiating the above objective function with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution as $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_{n_{\mathrm{nu}}})^{-1}\widehat{\boldsymbol{h}}$. Finally, a PE estimator is given as follows (cf. Eq.(3)):

$$\widehat{\mathrm{PE}} = \frac{1}{2}\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \frac{1}{2}. \tag{6}$$

Hyper-parameters included in the kernel function $K(\boldsymbol{u}, \boldsymbol{u}')$ and the regularization parameter λ can be optimized by cross-validation with respect to J (see Eq.(5)) as follows. First, samples $\mathcal{X}^{\mathrm{nu}} = \{\boldsymbol{u}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\mathcal{X}^{\mathrm{de}} = \{\boldsymbol{u}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ are divided into K disjoint subsets $\{\mathcal{X}_k^{\mathrm{nu}}\}_{k=1}^K$ and $\{\mathcal{X}_k^{\mathrm{de}}\}_{k=1}^K$, respectively. Then a density-ratio estimator $\widehat{r}_k(\boldsymbol{u})$ is obtained using $\mathcal{X}^{\mathrm{nu}} \backslash \mathcal{X}_k^{\mathrm{nu}}$ and $\mathcal{X}^{\mathrm{de}} \backslash \mathcal{X}_k^{\mathrm{de}}$, and the cost J is approximated using the hold-out samples $\mathcal{X}_k^{\mathrm{nu}}$ and $\mathcal{X}_k^{\mathrm{de}}$ as

$$J_k^{(K\text{-CV})} = \sum_{\boldsymbol{x}^{\mathrm{de}} \in \mathcal{X}_k^{\mathrm{de}}} \frac{\widehat{r}_k(\boldsymbol{u}^{\mathrm{de}})^2}{2|\mathcal{X}_k^{\mathrm{de}}|} - \sum_{\boldsymbol{x}^{\mathrm{nu}} \in \mathcal{X}_k^{\mathrm{nu}}} \frac{\widehat{r}_k(\boldsymbol{u}^{\mathrm{nu}})}{|\mathcal{X}_k^{\mathrm{nu}}|},$$

where $|\mathcal{X}|$ denotes the number of samples in the set $\mathcal{X}$. This hold-out procedure is repeated for $k = 1, \ldots, K$, and its average $J^{(K\text{-CV})}$ is outputted. We compute $J^{(K\text{-CV})}$ for all model candidates (i.e., the kernel parameter and the regularization parameter in the current case), and choose the model that minimizes $J^{(K\text{-CV})}$.

3.3 PE Maximization

Given the PE estimator $\widehat{\mathrm{PE}}$ (6), we next show how $\widehat{\mathrm{PE}}$ can be efficiently maximized with respect to $\boldsymbol{U}$:

$$\max_{\boldsymbol{U} \in \mathbb{R}^{m \times d}} \widehat{\mathrm{PE}} \;\; \text{s.t.}\; \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_m.$$

We propose to use a truncated negative quadratic function called the *Epanechnikov kernel* (Silverman 1986) as a kernel for $\boldsymbol{u}$:

$$K(\boldsymbol{u}, \boldsymbol{u}') = \max\left(0, 1 - \frac{\|\boldsymbol{u} - \boldsymbol{u}'\|^2}{2\sigma_{\mathrm{u}}^2}\right),$$

where σ_{u} is the kernel width.

Let $I(c)$ be the indicator function, i.e., $I(c) = 1$ if c is true and zero otherwise. Then, for the above kernel, $\widehat{\mathrm{PE}}$ can be expressed as

$$\widehat{\mathrm{PE}} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top\right) - \frac{1}{2}, \tag{7}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, and

$$\boldsymbol{D} = \frac{1}{n_{\mathrm{nu}}}\sum_{i,i'=1}^{n_{\mathrm{nu}}} \widehat{\alpha}_i(\boldsymbol{U}) I\left(\frac{\|\boldsymbol{U}\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{U}\boldsymbol{x}_{i'}^{\mathrm{nu}}\|^2}{2\sigma_{\mathrm{u}}^2} < 1\right) \times \left[\frac{1}{m}\boldsymbol{I}_d - \frac{1}{2\sigma_{\mathrm{u}}^2}(\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{x}_{i'}^{\mathrm{nu}})(\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{x}_{i'}^{\mathrm{nu}})^\top\right].$$

Here, by $\widehat{\alpha}_i(\boldsymbol{U})$, we explicitly indicated the fact that $\widehat{\alpha}_i$ depends on $\boldsymbol{U}$.

Let $\boldsymbol{D}'$ be $\boldsymbol{D}$ with $\boldsymbol{U}$ replaced by $\boldsymbol{U}'$, where $\boldsymbol{U}'$ is a transformation matrix obtained in the previous iteration. Thus, $\boldsymbol{D}'$ no longer depends on $\boldsymbol{U}$. Here we replace $\boldsymbol{D}$ in $\widehat{\mathrm{PE}}$ (see Eq.(7)) by $\boldsymbol{D}'$, which gives the following simplified PE estimate:

$$\widehat{\mathrm{PE}}' = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{U}\boldsymbol{D}'\boldsymbol{U}^\top\right) - \frac{1}{2}. \tag{8}$$

A maximizer of Eq.(8) can be obtained analytically by $(\boldsymbol{\varphi}_1|\cdots|\boldsymbol{\varphi}_m)^\top$, where $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m$ are the m principal components of $\boldsymbol{D}'$.

3.4 Initialization of $\boldsymbol{U}$

In the PE estimation-maximization framework described in Section 3.1, initialization of the transformation matrix $\boldsymbol{U}$ is important. Here we propose to initialize it based on PE maximization without dimensionality reduction.

More specifically, we determine the initial transformation matrix as $(\boldsymbol{\varphi}_1^{(0)}|\cdots|\boldsymbol{\varphi}_m^{(0)})^\top$, where $\boldsymbol{\varphi}_1^{(0)}, \ldots, \boldsymbol{\varphi}_m^{(0)}$ are the m principal components of $\boldsymbol{D}^{(0)}$:

$$\boldsymbol{D}^{(0)} = \frac{1}{n_{\mathrm{nu}}}\sum_{i,i'=1}^{n_{\mathrm{nu}}} \widehat{\alpha}_i^{(0)} I\left(\frac{\|\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{x}_{i'}^{\mathrm{nu}}\|^2}{2\sigma_{\mathrm{x}}^2} < 1\right) \times \left[\frac{1}{m}\boldsymbol{I}_d - \frac{1}{2\sigma_{\mathrm{x}}^2}(\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{x}_{i'}^{\mathrm{nu}})(\boldsymbol{x}_i^{\mathrm{nu}} - \boldsymbol{x}_{i'}^{\mathrm{nu}})^\top\right],$$

$$\widehat{\boldsymbol{\alpha}}^{(0)} = (\widehat{\boldsymbol{H}}^{(0)} + \lambda \boldsymbol{I}_{n_{\mathrm{nu}}})^{-1}\widehat{\boldsymbol{h}}^{(0)},$$

$$\widehat{\boldsymbol{H}}^{(0)} = \frac{1}{n_{\mathrm{de}}^2}\sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{k}'(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{k}'(\boldsymbol{x}_j^{\mathrm{de}})^\top, \;\; \widehat{\boldsymbol{h}}^{(0)} = \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{k}'(\boldsymbol{x}_i^{\mathrm{nu}}),$$

$$\boldsymbol{k}'(\boldsymbol{x}) = (K'(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K'(\boldsymbol{x}, \boldsymbol{x}_{n_{\mathrm{nu}}}))^\top,$$

$$K'(\boldsymbol{x}, \boldsymbol{x}') = \max\left(0, 1 - \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma_{\mathrm{x}}^2}\right).$$

σ_{x} is the kernel width and is chosen by cross-validation.

3.5 Density-Ratio Estimation in the Hetero-Distributional Subspace

Finally, a method of estimating the density ratio in the hetero-distributional subspace detected by the above HSA procedure is described.

A notable fact of the above HSA procedure is that the density-ratio estimator in the hetero-distributional subspace has already been obtained during the execution of the HSA algorithm—thus, an additional estimation procedure is not necessary. More specifically, the final solution is simply given by

$$\widehat{r}(\boldsymbol{x}) = \sum_{i=1}^{n_{\mathrm{nu}}} \widehat{\alpha}_i(\widehat{\boldsymbol{U}}) K(\widehat{\boldsymbol{U}}\boldsymbol{x}, \widehat{\boldsymbol{U}}\boldsymbol{x}_i),$$

where $\widehat{\boldsymbol{U}}$ is the transformation matrix obtained by the HSA algorithm. $\{\widehat{\alpha}_i(\widehat{\boldsymbol{U}})\}_{i=1}^{n_{\mathrm{nu}}}$ are the learned parameters under $\widehat{\boldsymbol{U}}$.

So far, we assumed that the true dimensionality m of the hetero-distributional subspace is known. When it is unknown, the best dimensionality based on the cross-validation score of the uLSIF estimator may be used in practice.

4 Experiments

In this section, we experimentally investigate the performance of the proposed and existing density-ration estimation methods using artificial and real-world datasets.

In all the experiments, we limit the number of basis kernels in HSA to 100 (see Eq.(4)), which were randomly chosen from all n_{nu} kernels. All the model parameters σ_{u}, σ_{x}, and λ are chosen by 5-fold cross-validation.

4.1 Illustration

Here, the performance of D^3-HSA is compared with that of the plain uLSIF (Kanamori, Hido, and Sugiyama 2009), D^3-LFDA/uLSIF (Sugiyama, Kawanabe, and Chui 2010), and D^3-LHSS (Sugiyama et al. 2011) using artificial datasets.

Suppose that the two densities $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are different only in a one-dimensional subspace (i.e., $m = 1$):

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

Let $n_{\mathrm{nu}} = n_{\mathrm{de}} = 1000$. The following datasets are used:

"Rather-separate" dataset:

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

where $N(u; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2 with respect to u, $N(\boldsymbol{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{v}$, and $\mathbf{0}_d$ denotes the d-dimensional vector with all zeros.

"Highly-overlapped" dataset:

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

"Dependent" dataset:

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

The error of a density-ratio estimator $\widehat{r}(\boldsymbol{x})$ is evaluated by

$$\mathrm{Error} := \frac{1}{2}\int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad (9)$$

which uLSIF tries to minimize. For the D^3-HSA, D^3-LHSS, and D^3-LFDA/uLSIF methods, we choose the dimensionality of the hetero-distributional subspace from $m = 1, \ldots, 5$ by cross-validation. In D^3-LHSS, the initialization matrix is chosen randomly.

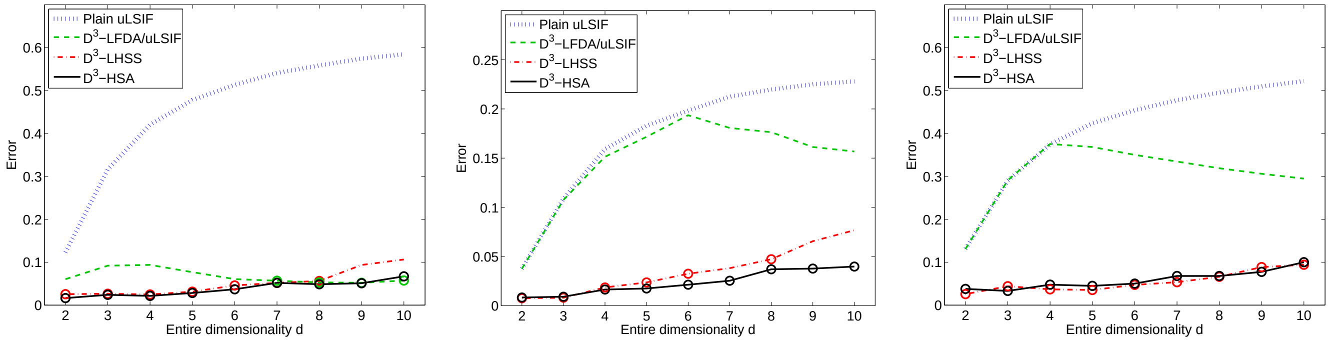
Figure 1 shows the density-ratio estimation error averaged over 50 runs as functions of the entire input dimensionality d, and Figure 2 shows the average computation time. These plots show that, while the error of the plain uLSIF increases rapidly as the entire dimensionality d increases. D^3-LFDA/uLSIF works reasonably well for the "rather-separate" dataset, but it performs poorly for the other two datasets. D^3-HSA and D^3-LHSS perform excellently for all the three datasets. Among D^3-HSA and D^3-LHSS, the computational cost of D^3-HSA is much smaller than D^3-LHSS. Thus, D^3-HSA overall compares favorably with the other approaches.

4.2 Application to Inlier-based Outlier Detection

Finally, we apply D^3-HSA to inlier-based outlier detection.

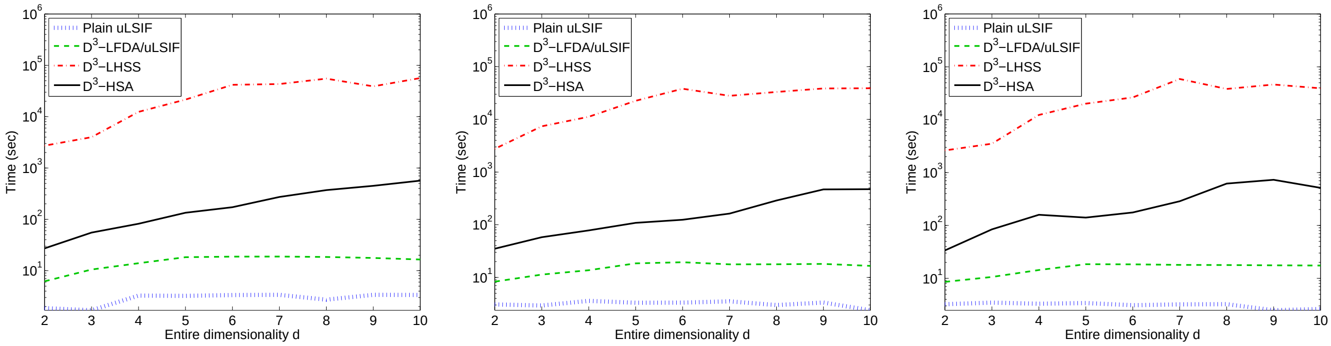
Let us consider an outlier detection problem of finding irregular samples in a dataset ("evaluation dataset") based on another dataset ("model dataset") that only contains regular samples. Defining the density ratio over the two sets of samples, we can see that the density-ratio values for regular samples are close to one, while those for outliers tend to be significantly deviated from one. Thus, density-ratio values could be used as an index of the degree of outlyingness (Smola, Song, and Teo 2009; Hido et al. 2011). 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.

We use the *USPS hand-written digit dataset* (Asuncion and Newman 2007). Each image consists of 256 (= 16 × 16) pixels and each pixel takes an integer value between 0 and 255 as the intensity level. 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. We applied principal component analysis to the 1025 samples, and extracted 50-dimensional feature vectors. 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 different tasks in total.



(a) “Rather-separate” dataset (b) “Highly-overlapped” dataset (c) “Dependent” dataset

Figure 1: Experimental results for artificial datasets. Density-ratio estimation error (9) averaged over 50 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 ‘$\circ$’.



(a) “Rather-separate” dataset (b) “Highly-overlapped” dataset (c) “Dependent” dataset

Figure 2: Experimental results for artificial datasets. The average computation time for each method. The average computation time includes cross-validation for choosing the dimensionality of the hetero-distributional subspace.

For the D^3-LFDA/uLSIF, D^3-LHSS, and D^3-HSA methods, we choose the dimensionality of the hetero-distributional subspace from $m = 5, 10, 15, \ldots, 50$ by cross-validation. In D^3-LHSS, the initialization matrix of LHSS is chosen randomly.

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 50 runs with different random seeds are summarized in Table 1. The table shows that the proposed D^3-HSA tends to outperform the plain uLSIF, D^3-LFDA/uLSIF, and D^3-LHSS with reasonable computation time.

5 Conclusion

In this paper, we proposed a novel density-ratio estimation method called *direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis* (D^3-HSA), which is more accurate and computationally efficient than existing methods. In D^3-HSA, a transformation matrix is estimated by iteratively performing *Pearson divergence* (PE) estimation and maximization, both of which are *analytically* carried out. Moreover, we gave a systematic method to design an initial transformation matrix. We applied the proposed D^3-HSA to density-ratio estimation and outlier detection tasks and experimentally showed that the proposed method is promising.

References

Asuncion, A., and Newman, D. 2007. UCI machine learning repository.

Bickel, S.; Bogojeska, J.; Lengauer, T.; and Scheffer, T. 2008. Multi-task learning for HIV therapy screening. In McCallum, A., and Roweis, S., eds., *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)*, 56–63.

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.

Elkan, C. 2010. Privacy-preserving data mining via importance

Table 1: Outlier detection for the USPS hand-written digit dataset ($d = 50$). The means (and standard deviations in the bracket) of AUC scores over 50 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 bold face. The means (and standard deviations in the bracket) of the dimensionality chosen by cross-validation are also included in the table. At the bottom, the average computation time (and standard deviations in the bracket) for each method over all experiments is shown. The computation time includes cross-validation for choosing the dimensionality of the hetero-distributional subspace.

	D^3-HSA		D^3-LHSS		D^3-LFDA/uLSIF		Plain uLSIF
Data	AUC	$\widehat{m}$	AUC	$\widehat{m}$	AUC	$\widehat{m}$	AUC
Digit 2	**0.987**(0.010)	11.8(7.4)	0.964(0.035)	19.8(14.1)	0.888(0.043)	44.6(14.8)	0.900(0.041)
Digit 3	**0.989**(0.011)	10.9(7.0)	0.975(0.030)	13.8(11.1)	0.917(0.044)	45.7(13.1)	0.923(0.039)
Digit 4	**0.984**(0.013)	13.3(6.8)	0.925(0.063)	15.3(12.9)	0.846(0.046)	44.6(14.8)	0.857(0.040)
Digit 5	**0.987**(0.011)	13.3(7.8)	0.976(0.025)	13.8(12.3)	0.895(0.045)	45.5(13.6)	0.905(0.040)
Digit 6	**0.991**(0.010)	13.3(7.7)	**0.983**(0.020)	14.5(11.1)	0.929(0.041)	44.1(14.8)	0.943(0.027)
Digit 7	**0.974**(0.018)	11.6(5.9)	**0.956**(0.050)	14.5(11.7)	0.836(0.078)	33.6(21.2)	0.882(0.036)
Digit 8	**0.980**(0.012)	12.7(7.6)	0.929(0.074)	14.9(11.7)	0.839(0.063)	48.2 (8.9)	0.847(0.039)
Digit 9	**0.988**(0.010)	12.9(8.3)	0.956(0.040)	13.9(11.1)	0.872(0.074)	47.3(10.8)	0.887(0.034)
Digit 0	**0.985**(0.014)	9.2(4.6)	**0.992**(0.012)	13.4(10.4)	0.969(0.032)	24.9(21.9)	**0.983**(0.016)
Time [sec]	129.2 (49.6)		4028.7 (2945.7)		22.3 (7.6)		4.4 (1.9)

weighting. In *ECML/PKDD Workshop on Privacy and Security Issues in Data Mining and Machine Learning (PSDML2010)*.

Hido, S.; Tsuboi, Y.; Kashima, H.; Sugiyama, M.; and Kanamori, T. 2011. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems* 26(2):309–336.

Huang, J.; Smola, A.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2007. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, 601–608. Cambridge, MA: MIT Press.

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

Kanamori, T.; Suzuki, T.; and Sugiyama, M. 2009. Condition number analysis of kernel-based density ratio estimation. Technical report, arXiv. http://www.citebase.org/abstract?id=oai:arXiv.org:0912.2800.

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

Lee, D., and Gray, A. 2006. Faster Gaussian summation: Theory and empirical experiments. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, 281–288.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56(11):5847–5861.

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

Raykar, V. C.; Duraiswami, R.; and Zhao, L. H. 2010. Fast computation of kernel estimators. *Journal of Computational and Graphical Statistics* 19(1):205–220.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):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*. New York: Chapman and Hall.

Smola, A.; Song, L.; and Teo, C. H. 2009. Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)*, 536–543.

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

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

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

Sugiyama, M.; Yamada, M.; von Bünau, P.; Suzuki, T.; Kanamori, T.; and Kawanabe, M. 2011. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks* 24(2):183–198.

Sugiyama, M.; Kawanabe, M.; and Chui, P. L. 2010. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks* 23(1):44–59.

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

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

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

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