Density-Ratio Matching under the Bregman Divergence: A Unified Framework of Density-Ratio Estimation

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

Taiji Suzuki
The University of Tokyo, Japan.
s-taiji@stat.t.u-tokyo.ac.jp

Takafumi Kanamori
Nagoya University, Japan.
kanamori@is.nagoya-u.ac.jp

Abstract

Estimation of the ratio of probability densities has attracted a great deal of attention since it can be used for addressing various statistical paradigms. A naive approach to density-ratio approximation is to first estimate numerator and denominator densities separately and then take their ratio. However, this two-step approach does not perform well in practice, and methods for directly estimating density ratios without density estimation have been explored. In this paper, we first give a comprehensive review of existing density-ratio estimation methods and discuss their pros and cons. Then we propose a new framework of density-ratio estimation in which a density-ratio model is fitted to the true density-ratio under the Bregman divergence. Our new framework includes existing approaches as special cases, and is substantially more general. Finally, we develop a robust density-ratio estimation method under the power divergence, which is a novel instance in our framework.

Keywords

Density ratio, Bregman divergence, Logistic regression, Kernel mean matching, Kullback-Leibler importance estimation procedure, Least-squares importance fitting

1 Introduction

The *ratio* of probability densities can be used for various statistical data processing purposes (Sugiyama et al., 2009, 2012) such as discriminant analysis (Silverman, 1978), non-stationarity adaptation (Shimodaira, 2000; Sugiyama and Müller, 2005; Sugiyama et al., 2007; Quiñonero-Candela et al., 2009; Sugiyama and Kawanabe, 2011), multi-task learning (Bickel et al., 2008), outlier detection (Hido et al., 2008; Smola et al., 2009; Hido et al., 2011), two-sample test (Keziou and Leoni-Aubin, 2005; Sugiyama et al., 2011a) change detection in time series (Kawahara and Sugiyama, 2009), conditional density estimation (Sugiyama et al., 2010), and probabilistic classification (Sugiyama, 2010).

Furthermore, *mutual information*—which plays a central role in information theory (Cover and Thomas, 2006)—can be estimated via density-ratio estimation (Suzuki et al., 2008, 2009b). Since mutual information is a measure of statistical independence between random variables, density-ratio estimation can be used also for variable selection (Suzuki et al., 2009a), dimensionality reduction (Suzuki and Sugiyama, 2010), independent component analysis (Suzuki and Sugiyama, 2009), causal inference (Yamada and Sugiyama, 2010), clustering (Kimura and Sugiyama, 2011), and cross-domain object matching (Yamada and Sugiyama, 2011) Thus, density-ratio estimation is a versatile tool for statistical data processing.

A naive approach to approximating a density-ratio is to separately estimate the two densities corresponding to the numerator and denominator of the ratio, and then take the ratio of the estimated densities. However, this naive approach is not reliable in high-dimensional problems since division by an estimated quantity can magnify the estimation error of the dividend. To overcome this drawback, various approaches to directly estimating density-ratios without going through density estimation have been explored recently, including the *moment matching approach* (Gretton et al., 2009), the *probabilistic classification approach* (Qin, 1998; Cheng and Chu, 2004), the *density matching approach* (Sugiyama et al., 2008; Tsuboi et al., 2009; Yamada and Sugiyama, 2009; Nguyen et al., 2010; Yamada et al., 2010), and the *density-ratio fitting approach* (Kanamori et al., 2009).

The purpose of this paper is to provide a general framework of density-ratio estimation that accommodates the above methods. More specifically, we propose a new density-ratio estimation approach called *density-ratio matching*—a density-ratio model is fitted to the true density-ratio function under the *Bregman divergence* (Bregman, 1967). We further develop a robust density-ratio estimation method under the *power divergence* (Basu et al., 1998), which is a novel instance in our general framework. Note that the Bregman divergence has been widely used in machine learning literature so far (Collins et al., 2002; Murata et al., 2004; Tsuda et al., 2005; Dhillon and Sra, 2006; Cayton, 2008; Wu et al., 2009), and the current paper explores a new application of the Bregman divergence in the framework of density-ratio estimation.

The rest of this paper is organized as follows. After the problem formulation below, we give a comprehensive review of density-ratio estimation methods in Section 2. In Section 3, we describe our new framework for density-ratio estimation. Finally, we conclude in Section 4.

Problem Formulation: The problem of density-ratio estimation addressed in this paper is formulated as follows. Let $\mathcal{X}$ ($\subset \mathbb{R}^d$) 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 $p_{\mathrm{nu}}^*(\boldsymbol{x})$ defined on $\mathcal{X}$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ from another distribution with density $p_{\mathrm{de}}^*(\boldsymbol{x})$ defined on $\mathcal{X}$.

$$\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}^*(\boldsymbol{x}) \quad \text{and} \quad \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}^*(\boldsymbol{x}).$$

We assume that $p_{\mathrm{de}}^*(\boldsymbol{x})$ is strictly positive over the domain $\mathcal{X}$. 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 $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$. 'nu' and 'de' indicate 'numerator' and 'denominator', respectively.

2 Existing Density-Ratio Estimation Methods

In this section, we give a comprehensive review of existing density-ratio estimation methods.

2.1 Moment Matching

Here, we describe a framework of density-ratio estimation based on *moment matching*.

2.1.1 Finite-Order Approach

First, we describe methods of finite-oder moment-matching for density-ratio estimation.

The simplest implementation of moment matching would be to match the first-order moment (i.e., the mean):

$$\operatorname*{argmin}_r \left\| \int \boldsymbol{x} r(\boldsymbol{x}) p_{\mathrm{de}}^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{x} p_{\mathrm{nu}}^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Its non-linear variant can be obtained using some non-linear function $\boldsymbol{\phi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^t$ as

$$\operatorname*{argmin}_r \mathrm{MM}'(r),$$

where

$$\mathrm{MM}'(r) := \left\| \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p_{\mathrm{de}}^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{\phi}(\boldsymbol{x}) p_{\mathrm{nu}}^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2.$$

'MM' stands for 'moment matching'. Let us ignore the irrelevant constant in $\mathrm{MM}'(r)$ and define the rest as $\mathrm{MM}(r)$:

$$\mathrm{MM}(r) := \left\| \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2 - 2 \left\langle \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \int \boldsymbol{\phi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

In practice, the expectations over $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ in $\mathrm{MM}(r)$ are replaced by sample averages. That is, for an n_{de}-dimensional vector

$$\boldsymbol{r}^*_{\mathrm{de}} := (r^*(\boldsymbol{x}^{\mathrm{de}}_1), \ldots, r^*(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}}))^\top,$$

where $^\top$ denotes the transpose, an estimator $\widehat{\boldsymbol{r}}_{\mathrm{de}}$ of $\boldsymbol{r}^*_{\mathrm{de}}$ can be obtained by solving the following optimization problem.

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} := \operatorname*{argmin}_{\boldsymbol{r} \in \mathbb{R}^{n_{\mathrm{de}}}} \widehat{\mathrm{MM}}(\boldsymbol{r}), \tag{2}$$

where

$$\widehat{\mathrm{MM}}(\boldsymbol{r}) := \frac{1}{n^2_{\mathrm{de}}} \boldsymbol{r}^\top \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{r} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{r}^\top \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}}. \tag{3}$$

$\mathbf{1}_n$ denotes the n-dimensional vector with all ones. $\boldsymbol{\Phi}_{\mathrm{nu}}$ and $\boldsymbol{\Phi}_{\mathrm{de}}$ are the $t \times n_{\mathrm{nu}}$ and $t \times n_{\mathrm{de}}$ *design matrices* defined by

$$\boldsymbol{\Phi}_{\mathrm{nu}} := (\boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{nu}}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{nu}}_{n_{\mathrm{nu}}})) \ \text{ and } \ \boldsymbol{\Phi}_{\mathrm{de}} := (\boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}})),$$

respectively. Taking the derivative of the objective function (3) with respect to $\boldsymbol{r}$ and setting it to zero, we have

$$\frac{2}{n^2_{\mathrm{de}}} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{r} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}} = \mathbf{0}_t,$$

where $\mathbf{0}_t$ denotes the t-dimensional vector with all zeros. Solving this equation with respect to $\boldsymbol{r}$, one can obtain the solution analytically as

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}})^{-1} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}}.$$

One may add a normalization constraint

$$\frac{1}{n_{\mathrm{de}}} \mathbf{1}^\top_{n_{\mathrm{de}}} \boldsymbol{r} = 1$$

to the optimization problem (2). Then the optimization problem becomes a *convex linearly-constrained quadratic program*. Since there is no known method for obtaining the

analytic-form solution for convex linearly-constrained quadratic programs, a numerical solver may be needed to compute the solution. Furthermore, a non-negativity constraint

$$\boldsymbol{r} \geq \boldsymbol{0}_{n_{\mathrm{de}}}$$

and/or an upper bound for a positive constant B, i.e.,

$$\boldsymbol{r} \leq B\boldsymbol{1}_{n_{\mathrm{de}}}$$

may also be incorporated in the optimization problem (2), where inequalities for vectors are applied in the element-wise manner. Even with these modifications, the optimization problem is still a convex linearly-constrained quadratic program, so its solution can be numerically computed by standard optimization software.

The above *fixed-design* method gives estimates of the density-ratio values only at the denominator sample points $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$. Below, we consider the *induction* setup, where the entire density-ratio function $r^*(\boldsymbol{x})$ is estimated (Qin, 1998; Kanamori et al., 2012).

We use the following linear density-ratio model for density-ratio function learning:

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell \psi_\ell(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta}, \tag{4}$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a basis function vector and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. We assume that the basis functions are non-negative.

$$\boldsymbol{\psi}(\boldsymbol{x}) \geq \boldsymbol{0}_b.$$

Then model outputs at $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ are expressed in terms of the parameter vector $\boldsymbol{\theta}$ as

$$(r(\boldsymbol{x}_1^{\mathrm{de}}), \ldots, r(\boldsymbol{x}_{n_{\mathrm{de}}}^{\mathrm{de}}))^\top = \boldsymbol{\Psi}_{\mathrm{de}}^\top \boldsymbol{\theta},$$

where $\boldsymbol{\Psi}_{\mathrm{de}}$ is the $b \times n_{\mathrm{de}}$ *design matrix* defined by

$$\boldsymbol{\Psi}_{\mathrm{de}} := (\boldsymbol{\psi}(\boldsymbol{x}_1^{\mathrm{de}}), \ldots, \boldsymbol{\psi}(\boldsymbol{x}_{n_{\mathrm{de}}}^{\mathrm{de}})). \tag{5}$$

Then, following Eq.(2), the parameter $\boldsymbol{\theta}$ is learned as follows.

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\frac{1}{n_{\mathrm{de}}^2} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{\Psi}_{\mathrm{de}}^\top \boldsymbol{\theta} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{nu}} \boldsymbol{1}_{n_{\mathrm{nu}}} \right]. \tag{6}$$

Taking the derivative of the above objective function with respect to $\boldsymbol{\theta}$ and setting it to zero, we have the solution $\widehat{\boldsymbol{\theta}}$ analytically as

$$\widehat{\boldsymbol{\theta}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{\Psi}_{\mathrm{de}}^\top)^{-1} \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{nu}} \boldsymbol{1}_{n_{\mathrm{nu}}}.$$

One may include a normalization constraint, a non-negativity constraint (given that the basis functions are non-negative), and a regularization constraint to the optimization problem (6):

$$\frac{1}{n_{\mathrm{de}}}\mathbf{1}_{n_{\mathrm{de}}}^{\top}\boldsymbol{\Psi}_{\mathrm{de}}^{\top}\boldsymbol{\theta}=1, \quad \boldsymbol{\theta}\geq\mathbf{0}_b, \text{ and } \boldsymbol{\theta}\leq B\mathbf{1}_b.$$

Then the optimization problem becomes a convex linearly-constrained quadratic program, whose solution can be obtained by a standard numerical solver.

The upper-bound parameter B, which works as a regularizer, may be optimized by *cross-validation* (CV) with respect to the moment-matching error MM defined by Eq.(1). Availability of CV would be one of the advantages of the inductive method (i.e., learning the entire density-ratio function).

2.1.2 Infinite-Order Approach: KMM

Matching a finite number of moments does not necessarily lead to the true density-ratio function $r^*(\boldsymbol{x})$, even if infinitely many samples are available. In order to guarantee that the true density-ratio function can always be obtained in the large-sample limit, all moments up to the infinite order need to be matched. Here we describe a method of infinite-oder moment-matching called *kernel mean matching* (KMM), which allows one to efficiently match all the moments using kernel functions (Huang et al., 2007; Gretton et al., 2009).

The basic idea of KMM is essentially the same as the finite-order approach, but a *universal reproducing kernel* $K(\boldsymbol{x},\boldsymbol{x}')$ (Steinwart, 2001) is used as a non-linear transformation. The *Gaussian kernel*

$$K(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right) \tag{7}$$

is an example of universal reproducing kernels. It has been shown that the solution of the following optimization problem agrees with the true density-ratio (Huang et al., 2007; Gretton et al., 2009):

$$\min_{r\in\mathcal{H}} \left\|\int K(\boldsymbol{x},\cdot)p_{\mathrm{nu}}^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int K(\boldsymbol{x},\cdot)r(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{\mathcal{H}}^2,$$

where $\mathcal{H}$ denotes a universal reproducing kernel Hilbert space and $\|\cdot\|_{\mathcal{H}}$ denotes its norm.

An empirical version of the above problem is expressed as

$$\min_{\boldsymbol{r}\in\mathbb{R}^{n_{\mathrm{de}}}}\left[\frac{1}{n_{\mathrm{de}}^2}\boldsymbol{r}^{\top}\boldsymbol{K}_{\mathrm{de,de}}\boldsymbol{r}-\frac{2}{n_{\mathrm{de}}n_{\mathrm{nu}}}\boldsymbol{r}^{\top}\boldsymbol{K}_{\mathrm{de,nu}}\mathbf{1}_{n_{\mathrm{nu}}}\right],$$

where $\boldsymbol{K}_{\mathrm{de,de}}$ and $\boldsymbol{K}_{\mathrm{de,nu}}$ denote the kernel Gram matrices defined by

$$[\boldsymbol{K}_{\mathrm{de,de}}]_{j,j'}=K(\boldsymbol{x}_j^{\mathrm{de}},\boldsymbol{x}_{j'}^{\mathrm{de}}) \text{ and } [\boldsymbol{K}_{\mathrm{de,nu}}]_{j,i}=K(\boldsymbol{x}_j^{\mathrm{de}},\boldsymbol{x}_i^{\mathrm{nu}}), \tag{8}$$

respectively. In the same way as the finite-order case, the solution can be obtained analytically as

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \boldsymbol{K}_{\mathrm{de,de}}^{-1} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}}. \tag{9}$$

If necessary, one may include a non-negativity constraint, a normalization constraint, and an upper bound in the same way as the finite-order case. Then the solution can be numerically obtained by solving a convex linearly-constrained quadratic programming problem.

For a linear density-ratio model (4), an inductive variant of KMM is formulated as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\frac{1}{n_{\mathrm{de}}^2} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,de}} \boldsymbol{\Psi}_{\mathrm{de}}^\top \boldsymbol{\theta} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}} \right],$$

and the solution $\widehat{\boldsymbol{\theta}}$ is given by

$$\widehat{\boldsymbol{\theta}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,de}} \boldsymbol{\Psi}_{\mathrm{de}})^{-1} \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}}.$$

2.1.3 Remarks

The infinite-order moment matching method, *kernel mean matching* (KMM), can efficiently match all the moments by making use of universal reproducing kernels. Indeed, KMM has an excellent theoretical property that it is consistent (Huang et al., 2007; Gretton et al., 2009). However, KMM has a limitation in model selection—there is no known method for determining the kernel parameter (i.e., the Gaussian kernel width). A popular heuristic of setting the Gaussian width to the median distance between samples (Schölkopf and Smola, 2002) would be useful in some cases, but this may not always be reasonable.

In the above, moment matching was performed in terms of the squared norm, which led to an analytic-form solution (if no constraint is imposed). As shown in Kanamori et al. (2012), moment matching can be systematically generalized to various divergences.

2.2 Probabilistic Classification

Here, we describe a framework of density-ratio estimation through *probabilistic classification*.

2.2.1 Basic Framework

The basic idea of the probabilistic classification approach is to obtain a probabilistic classifier that separates numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and denominator samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$.

Let us assign a label $y = +1$ to $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $y = -1$ to $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, respectively. Then the two densities $p_{\mathrm{nu}}^*(\boldsymbol{x})$ and $p_{\mathrm{de}}^*(\boldsymbol{x})$ are written as

$$p_{\mathrm{nu}}^*(\boldsymbol{x}) = p^*(\boldsymbol{x}|y=+1) \quad \text{and} \quad p_{\mathrm{de}}^*(\boldsymbol{x}) = p^*(\boldsymbol{x}|y=-1),$$

respectively. Note that y is regarded as a random variable here. An application of Bayes' theorem,

$$p^*(\boldsymbol{x}|y) = \frac{p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y)},$$

yields that the density-ratio $r^*(\boldsymbol{x})$ can be expressed in terms of y as follows:

$$\begin{aligned} r^*(\boldsymbol{x}) &= \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})} = \left(\frac{p^*(y=+1|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y=+1)}\right)\left(\frac{p^*(y=-1|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y=-1)}\right)^{-1} \\ &= \frac{p^*(y=-1)}{p^*(y=+1)}\frac{p^*(y=+1|\boldsymbol{x})}{p^*(y=-1|\boldsymbol{x})}. \end{aligned}$$

The ratio $p^*(y=-1)/p^*(y=+1)$ may be simply approximated by the ratio of the sample size:

$$\frac{p^*(y=-1)}{p^*(y=+1)} \approx \frac{n_{\mathrm{de}}/(n_{\mathrm{de}}+n_{\mathrm{nu}})}{n_{\mathrm{nu}}/(n_{\mathrm{de}}+n_{\mathrm{nu}})} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}.$$

The 'class'-posterior probability $p^*(y|\boldsymbol{x})$ may be approximated by separating $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ using a probabilistic classifier. Thus, given an estimator of the class-posterior probability, $\widehat{p}(y|\boldsymbol{x})$, a density-ratio estimator $\widehat{r}(\boldsymbol{x})$ can be constructed as

$$\widehat{r}(\boldsymbol{x}) = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}\frac{\widehat{p}(y=+1|\boldsymbol{x})}{\widehat{p}(y=-1|\boldsymbol{x})}. \tag{10}$$

A practical advantage of the probabilistic classification approach would be its easy implementability. Indeed, one can directly use standard probabilistic classification algorithms for density-ratio estimation. Another, more important advantage of the probabilistic classification approach is that model selection (i.e., tuning the basis functions and the regularization parameter) is possible by standard *cross-validation* since the estimation problem involved in this framework is a standard supervised classification problem.

Below, two probabilistic classification algorithms are described. For making the explanation simple, we consider a set of paired samples $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{n}$, where, for $n = n_{\mathrm{nu}} + n_{\mathrm{de}}$,

$$\begin{aligned} (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &:= (\boldsymbol{x}^{\mathrm{nu}}_1, \ldots, \boldsymbol{x}^{\mathrm{nu}}_{n_{\mathrm{nu}}}, \boldsymbol{x}^{\mathrm{de}}_1, \ldots, \boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}}), \\ (y_1, \ldots, y_n) &:= (\underbrace{+1, \ldots, +1}_{n_{\mathrm{nu}}}, \underbrace{-1, \ldots, -1}_{n_{\mathrm{de}}}). \end{aligned}$$

2.2.2 Logistic Regression

Here, a popular probabilistic classification algorithm called *logistic regression* (Hastie et al., 2001) is explained.

A logistic regression classifier employs a parametric model of the following form for expressing the class-posterior probability $p^*(y|\boldsymbol{x})$,

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1+\exp\left(-y\boldsymbol{\psi}(\boldsymbol{x})^\top\boldsymbol{\theta}\right)},$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a basis function vector and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. The parameter vector $\boldsymbol{\theta}$ is determined so that the *penalized log-likelihood* is maximized, which can be expressed as the following minimization problem:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\sum_{k=1}^{n} \log \left(1 + \exp\left(-y_k \boldsymbol{\psi}(\boldsymbol{x}_k)^\top \boldsymbol{\theta}\right)\right) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right], \tag{11}$$

where $\lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}$ is a penalty term included for regularization purposes.

Since the objective function in Eq.(11) is convex, the global optimal solution can be obtained by a standard non-linear optimization technique such as the *gradient descent method* or *(quasi-)Newton methods* (Hastie et al., 2001; Minka, 2007). Finally, a density-ratio estimator $\widehat{r}_{\mathrm{LR}}(\boldsymbol{x})$ is given by

$$\widehat{r}_{\mathrm{LR}}(\boldsymbol{x}) = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \frac{1 + \exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\theta}}\right)}{1 + \exp\left(-\boldsymbol{\psi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\theta}}\right)} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\theta}}\right),$$

where 'LR' stands for 'logistic regression'.

Suppose that the logistic regression model $p(y|\boldsymbol{x};\boldsymbol{\theta})$ satisfies the following two conditions:

- The constant function is included in the basis functions, i.e., there exists $\boldsymbol{\theta}^\circ$ such that

$$\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta}^\circ = 1.$$

- The model is *correctly specified*, i.e., there exists $\boldsymbol{\theta}^*$ such that

$$p(y|\boldsymbol{x};\boldsymbol{\theta}^*) = p^*(y|\boldsymbol{x}).$$

Then it was proved that the logistic regression approach is optimal among a class of semi-parametric estimators in the sense that the asymptotic variance is minimized (Qin, 1998). However, when the model is misspecified (which would be the case in practice), the density matching approach explained in Section 2.3 would be more preferable (Kanamori et al., 2010).

When *multi-class* logistic regression classifiers are used, density-ratios among multiple densities can be estimated simultaneously (Bickel et al., 2008). This is useful, e.g., for solving *multi-task learning* problems (Caruana et al., 1997).

2.2.3 Least-Squares Probabilistic Classifier

Although the performance of these general-purpose non-linear optimization techniques has been improved together with the evolution of computer environment in the last decade, training logistic regression classifiers is still computationally expensive. Here,

an alternative probabilistic classification algorithm called *least-squares probabilistic classifier* (LSPC; Sugiyama, 2010) is described. LSPC is computationally more efficient than logistic regression, with comparable accuracy in practice.

In LSPC, the class-posterior probability $p^*(y|\boldsymbol{x})$ is modeled as

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^{b} \theta_\ell \psi(\boldsymbol{x},y) \ = \boldsymbol{\psi}(\boldsymbol{x},y)^\top \boldsymbol{\theta},$$

where $\boldsymbol{\psi}(\boldsymbol{x},y)$ ($\in \mathbb{R}^b$) is a non-negative basis function vector, and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. The class label y takes a value in $\{1,\ldots,c\}$, where c is the number of classes.

The basic idea of LSPC is to express the class-posterior probability $p^*(y|\boldsymbol{x})$ in terms of the equivalent density-ratio expression: $p^*(\boldsymbol{x},y)/p^*(\boldsymbol{x})$. Then the density-ratio estimation method called *unconstrained least-squares importance fitting* (uLSIF; Kanamori et al., 2009) is used for estimating this density-ratio. Since uLSIF will be reviewed in detail in Section 2.4.3, we only describe the final solution here.

Let

$$\widehat{\boldsymbol{H}} := \frac{1}{n}\sum_{k=1}^{n}\sum_{y=1}^{c} \boldsymbol{\psi}(\boldsymbol{x}_k,y)\boldsymbol{\psi}(\boldsymbol{x}_k,y)^\top \quad \text{and} \quad \widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\psi}(\boldsymbol{x}_k,y_k).$$

Then the uLSIF solution is given analytically as $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}}$, where λ (≥ 0) is the regularization parameter and $\boldsymbol{I}_b$ is the b-dimensional identity matrix. In order to assure that the output of LSPC is a probability, the outputs are normalized and negative outputs are rounded up to zero (Yamada et al., 2011):

$$\widehat{p}(y|\boldsymbol{x}) = \frac{\max(0, \boldsymbol{\psi}(\boldsymbol{x},y)^\top \widehat{\boldsymbol{\theta}})}{\sum_{y'=1}^{c} \max(0, \boldsymbol{\psi}(\boldsymbol{x},y')^\top \widehat{\boldsymbol{\theta}})}.$$

A standard choice of basis functions $\boldsymbol{\psi}(\boldsymbol{x},y)$ would be a *kernel* model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\ell=1}^{n} \theta_\ell^{(y)} K(\boldsymbol{x},\boldsymbol{x}_\ell), \tag{12}$$

where $K(\boldsymbol{x},\boldsymbol{x}')$ is some kernel function such as the *Gaussian kernel* (7). Then the matrix $\widehat{\boldsymbol{H}}$ becomes block-diagonal. Thus, we only need to train a model with n parameters separately c times for each class $y = 1,\ldots,c$. Since all the diagonal block matrices are the same, the computational complexity for computing the solution is $\mathcal{O}(n^3 + cn^2)$.

Let us further reduce the number of kernels in model (12). To this end, we focus on a kernel function $K(\boldsymbol{x},\boldsymbol{x}')$ that is "localized". Examples of such localized kernels include the popular Gaussian kernel. The idea is to reduce the number of kernels by locating the kernels only at samples belonging to the *target* class:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\ell=1}^{n_y} \theta_\ell^{(y)} K(\boldsymbol{x},\boldsymbol{x}_\ell^{(y)}), \tag{13}$$

where n_y is the number of training samples in class y and $\{\boldsymbol{x}_k^{(y)}\}_{k=1}^{n_y}$ is the training input samples in class y. The rationale behind this model simplification is as follows. By definition, the class-posterior probability $p^*(y|\boldsymbol{x})$ takes large values in the regions where samples in class y are dense; conversely, $p^*(y|\boldsymbol{x})$ takes smaller values (i.e., close to zero) in the regions where samples in class y are sparse. When a non-negative function is approximated by a localized kernel model, many kernels may be needed in the region where the output of the target function is large; on the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero. Following this heuristic, many kernels are allocated in the region where $p^*(y|\boldsymbol{x})$ takes large values, which can be achieved by Eq.(13).

This model simplification allows one to further reduce the computational cost since the size of the target blocks in matrix $\widehat{\boldsymbol{H}}$ is further reduced. In order to determine the n_y-dimensional parameter vector $\boldsymbol{\theta}^{(y)} = (\theta_1^{(y)}, \ldots, \theta_{n_y}^{(y)})^\top$ for each class y, we only need to solve the following system of n_y linear equations:

$$(\widehat{\boldsymbol{H}}^{(y)} + \lambda \boldsymbol{I}_{n_y})\boldsymbol{\theta}^{(y)} = \widehat{\boldsymbol{h}}^{(y)}, \tag{14}$$

where $\widehat{\boldsymbol{H}}^{(y)}$ is the $n_y \times n_y$ matrix, and $\widehat{\boldsymbol{h}}^{(y)}$ is the n_y-dimensional vector defined as

$$\widehat{H}_{\ell,\ell'}^{(y)} := \frac{1}{n_y}\sum_{k=1}^{n_y} K(\boldsymbol{x}_k^{(y)}, \boldsymbol{x}_\ell^{(y)})K(\boldsymbol{x}_k^{(y)}, \boldsymbol{x}_{\ell'}^{(y)}) \quad \text{and} \quad \widehat{h}_\ell^{(y)} := \frac{1}{n_y}\sum_{k=1}^{n_y} K(\boldsymbol{x}_k^{(y)}, \boldsymbol{x}_\ell^{(y)}).$$

Let $\widehat{\boldsymbol{\theta}}^{(y)}$ be the solution of Eq.(14). Then the final solution is given by

$$\widehat{p}(y|\boldsymbol{x}) = \frac{\max\left(0, \sum_{\ell=1}^{n_y} \widehat{\theta}_\ell^{(y)} K(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)})\right)}{\sum_{y'=1}^{c} \max\left(0, \sum_{\ell=1}^{n_{y'}} \widehat{\theta}_\ell^{(y')} K(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y')})\right)}. \tag{15}$$

For the simplified model (13), the computational complexity for computing the solution is $\mathcal{O}(cn_y^3)$—when $n_y = n/c$ for all y, this is equal to $\mathcal{O}(c^{-2}n^3)$. Thus, this approach is computationally highly efficient for multi-class problems with large c.

A MATLAB® implementation of LSPC is available from

http://sugiyama-www.cs.titech.ac.jp/˜sugi/software/LSPC/

2.2.4 Remarks

Density-ratio estimation by probabilistic classification can successfully avoid density estimation by casting the problem of density-ratio estimation as the problem of estimating the 'class'-posterior probability. An advantage of the probabilistic classification approach over the moment matching approach explained in Section 2.1 is that cross-validation can

be used for model selection. Furthermore, existing software packages of probabilistic classification algorithms can be directly used for density-ratio estimation.

The probabilistic classification approach with logistic regression was shown to have a suitable theoretical property (Qin, 1998): if the logistic regression model is *correctly specified*, the probabilistic classification approach is optimal among a broad class of semi-parametric estimators. However, this strong theoretical property is not true when the correct model assumption is not fulfilled.

An advantage of the probabilistic classification approach is that it can be used for estimating density-ratios among multiple densities by multi-class probabilistic classifiers. In this context, the *least-squares probabilistic classifier* (LSPC) would be practically useful due to its computational efficiency.

2.3 Density Matching

Here, we describe a framework of density-ratio estimation by *density matching* under the KL divergence.

2.3.1 Basic Framework

Let $r(\boldsymbol{x})$ be a model of the true density-ratio $r^*(\boldsymbol{x}) = p^*_{\mathrm{nu}}(\boldsymbol{x})/p^*_{\mathrm{de}}(\boldsymbol{x})$. Then the numerator density $p^*_{\mathrm{nu}}(\boldsymbol{x})$ may be modeled by $p_{\mathrm{nu}}(\boldsymbol{x}) = r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$. Now let us consider the KL divergence from $p^*_{\mathrm{nu}}(\boldsymbol{x})$ to $p_{\mathrm{nu}}(\boldsymbol{x})$:

$$\mathrm{KL}'(p^*_{\mathrm{nu}} \| p_{\mathrm{nu}}) := \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{nu}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = C - \mathrm{KL}(r),$$

where $C := \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$ is a constant irrelevant to r, and $\mathrm{KL}(r)$ is the relevant part:

$$\mathrm{KL}(r) := \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i).$$

Since $p_{\mathrm{nu}}(\boldsymbol{x})$ is a probability density function, its integral should be one:

$$1 = \int p_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j).$$

Furthermore, the density $p_{\mathrm{nu}}(\boldsymbol{x})$ should be non-negative, which can be achieved by $r(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$. Combining these equations together, we have the following optimization problem.

$$\begin{aligned} \max_{r} \quad & \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i) \\ \text{s.t.} \quad & \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) = 1 \text{ and } r(\boldsymbol{x}) \geq 0 \text{ for all } \boldsymbol{x}. \end{aligned}$$

This formulation is called the *KL importance estimation procedure* (KLIEP; Sugiyama et al., 2008).

Possible hyper-parameters in KLIEP (such as basis parameters and regularization parameters) can be optimized using *cross-validation* with respect to the KL divergence, where the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ appearing in the objective function may only be cross-validated (Sugiyama et al., 2008).

Below, practical implementations of KLIEP for various density-ratio models are described.

2.3.2 Linear and Kernel Models

Let us employ a linear model for density-ratio estimation.

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell \psi_\ell(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta}, \tag{16}$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a non-negative basis function vector, and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. Then the KLIEP optimization problem for the linear model is expressed as follows (Sugiyama et al., 2008).

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^b} \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log(\boldsymbol{\psi}(\boldsymbol{x}_i^{\mathrm{nu}})^\top \boldsymbol{\theta}) \quad \text{s.t.} \quad \overline{\boldsymbol{\psi}}_{\mathrm{de}}^\top \boldsymbol{\theta} = 1 \text{ and } \boldsymbol{\theta} \geq \mathbf{0}_b,$$

where $\overline{\boldsymbol{\psi}}_{\mathrm{de}} := \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\psi}(\boldsymbol{x}_j^{\mathrm{de}})$.

Since the above optimization problem is *convex*, there exists the unique global optimum solution. Furthermore, the KLIEP solution tends to be *sparse*, i.e., many parameters take exactly zero, because of the non-negativity constraint. Such sparsity would contribute to reducing the computation time when computing estimated density-ratio values. As can be confirmed from the above optimization problem, the denominator samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ appear only in terms of the basis-transformed mean $\overline{\boldsymbol{\psi}}_{\mathrm{de}}$. Thus, KLIEP for linear models is computationally efficient even when the number n_{de} of denominator samples is very large.

The performance of KLIEP depends on the choice of the basis functions $\boldsymbol{\psi}(\boldsymbol{x})$. As explained below, the use of the following Gaussian kernel model would be reasonable:

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{n_{\mathrm{nu}}} \theta_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell^{\mathrm{nu}}), \tag{17}$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel (7). The reason why the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$, not the denominator samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, are chosen as the Gaussian centers is as follows. By definition, the density-ratio $r^*(\boldsymbol{x})$ tends to take large values if $p_{\mathrm{de}}^*(\boldsymbol{x})$ is small and $p_{\mathrm{nu}}^*(\boldsymbol{x})$ is large. Conversely, $r^*(\boldsymbol{x})$ tends to be small (i.e., close to zero) if $p_{\mathrm{de}}^*(\boldsymbol{x})$ is large and $p_{\mathrm{nu}}^*(\boldsymbol{x})$ is small. 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. Following this heuristic, many kernels are allocated in the region where $p^*_{\mathrm{nu}}(\boldsymbol{x})$ takes large values, which can be achieved by setting the Gaussian centers at $\{\boldsymbol{x}^{\mathrm{nu}}_i\}^{n_{\mathrm{nu}}}_{i=1}$.

The KLIEP methods for linear/kernel models are referred to as *linear KLIEP* (L-KLIEP) and *kernel KLIEP* (K-KLIEP), respectively. A MATLAB® implementation of the K-KLIEP algorithm is available from

http://sugiyama-www.cs.titech.ac.jp/~sugi/software/KLIEP/

2.3.3 Log-Linear Models

Another popular model choice would be the *log-linear model* (Tsuboi et al., 2009; Kanamori et al., 2010):

$$r(\boldsymbol{x};\boldsymbol{\theta},\theta_0) = \exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top\boldsymbol{\theta} + \theta_0\right), \tag{18}$$

where θ_0 is a normalization parameter. From the normalization constraint

$$\frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j;\boldsymbol{\theta},\theta_0) = 1,$$

θ_0 is determined as

$$\widehat{\theta}_0 = -\log\left(\frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\exp\left(\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)^\top\boldsymbol{\theta}\right)\right).$$

Then the density-ratio model is expressed as

$$r(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top\boldsymbol{\theta}\right)}{\frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\exp\left(\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)^\top\boldsymbol{\theta}\right)}.$$

By definition, outputs of the log-linear model $r(\boldsymbol{x};\boldsymbol{\theta})$ are non-negative for all $\boldsymbol{x}$. Thus, we do not need the non-negativity constraint on the parameter. Then the KLIEP optimization criterion is expressed as

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^b}\left[\overline{\boldsymbol{\psi}}_{\mathrm{nu}}^\top\boldsymbol{\theta} - \log\left(\frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\exp(\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)^\top\boldsymbol{\theta})\right)\right],$$

where $\overline{\boldsymbol{\psi}}_{\mathrm{nu}} := \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{nu}}_i)$. This is an unconstrained convex optimization problem, so the global optimal solution can be obtained by, e.g., the gradient method and (quasi-)Newton methods. Since the numerator samples $\{\boldsymbol{x}^{\mathrm{nu}}_i\}^{n_{\mathrm{nu}}}_{i=1}$ appear only in terms of the basis-transformed mean $\overline{\boldsymbol{\psi}}_{\mathrm{nu}}$, KLIEP for log-linear models is computationally efficient even when the number n_{nu} of numerator samples is very large (cf. KLIEP for linear/kernel models is computationally efficient when n_{de} is very large; see Section 2.3.2).

The KLIEP method for log-linear models is called *log-linear KLIEP* (LL-KLIEP).

2.3.4 Gaussian Mixture Models

In the Gaussian kernel model (17), the Gaussian shape is spherical and its width is controlled by a single width parameter σ. It is possible to use correlated Gaussian kernels, but choosing the covariance matrix via cross-validation would be computationally intractable.

Another option is to also estimate the covariance matrix directly from data. For this purpose, the *Gaussian mixture model* comes in handy (Yamada and Sugiyama, 2009):

$$r(\boldsymbol{x}; \{\theta_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^c) = \sum_{k=1}^{c} \theta_k K(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{19}$$

where c is the number of mixing components, $\{\theta_k\}_{k=1}^c$ are mixing coefficients, $\{\boldsymbol{\mu}_k\}_{k=1}^c$ are means of Gaussian functions, $\{\boldsymbol{\Sigma}_k\}_{k=1}^c$ are covariance matrices of Gaussian functions, and $K(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian kernel with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$K(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{20}$$

Note that $\boldsymbol{\Sigma}$ should be *positive definite*, i.e., all the eigenvectors of $\boldsymbol{\Sigma}$ should be strictly positive.

For the Gaussian mixture model (19), the KLIEP optimization problem is expressed as

$$\begin{aligned} \max_{\{\theta_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^c} \quad & \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log\left(\sum_{k=1}^{c} \theta_k K(\boldsymbol{x}_i^{\mathrm{nu}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\ \text{s.t.} \quad & \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \sum_{k=1}^{c} \theta_k K(\boldsymbol{x}_j^{\mathrm{de}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = 1, \\ & \theta_k \geq 0 \text{ and } \boldsymbol{\Sigma}_k \succ \boldsymbol{O} \text{ for } k = 1, \ldots, c, \end{aligned}$$

where $\boldsymbol{\Sigma}_k \succ \boldsymbol{O}$ means that $\boldsymbol{\Sigma}_k$ is positive definite.

The above optimization problem is *non-convex*, and there is no known method for obtaining the global optimal solution. In practice, a local optimal solution may be numerically obtained by, e.g., a fixed-point method.

The KLIEP method for Gaussian mixture models is called *Gaussian-mixture KLIEP* (GM-KLIEP).

2.3.5 Probabilistic PCA Mixture Models

The Gaussian mixture model explained above would be more flexible than linear/kernel/log-linear models and suitable for approximating correlated density-ratio functions. However, when the target density-ratio function is (locally) rank-deficient, its behavior could be unstable since inverse covariance matrices are included in the Gaussian function (see Eq.(20)). To cope with this problem, the use of *a mixture of probabilistic principal component analyzers* (PPCA; Tipping and Bishop, 1999) was proposed for density-ratio estimation (Yamada et al., 2010).

The PPCA mixture model is defined as

$$r(\boldsymbol{x}; \{\theta_k, \boldsymbol{\mu}_k, \sigma_k^2, \boldsymbol{W}_k\}_{k=1}^c) = \sum_{k=1}^c \theta_k K(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2, \boldsymbol{W}_k),$$

where c is the number of mixing components and $\{\theta_k\}_{k=1}^c$ are mixing coefficients. $K(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2, \boldsymbol{W})$ is a PPCA model defined by

$$K(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2, \boldsymbol{W}) = (2\pi\sigma^2)^{-\frac{d}{2}} \det(\boldsymbol{C})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where 'det' denotes the determinant, $\boldsymbol{\mu}$ is the mean of the Gaussian function, σ^2 is the variance of the Gaussian function, $\boldsymbol{W}$ is a $d \times m$ 'projection' matrix onto a m-dimensional *latent* space (where $m \leq d$), and $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_d$.

Then the KLIEP optimization criterion is expressed as

$$\begin{aligned} \max_{\{\theta_k, \boldsymbol{\mu}_k, \sigma_k^2, \boldsymbol{W}_k\}_{k=1}^c} \quad & \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log\left(\sum_{k=1}^c \theta_k K(\boldsymbol{x}_i^{\mathrm{nu}}; \boldsymbol{\mu}_k, \sigma_k^2, \boldsymbol{W}_k)\right) \\ \text{s.t.} \quad & \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \sum_{k=1}^c \theta_k K(\boldsymbol{x}_j^{\mathrm{de}}; \boldsymbol{\mu}_k, \sigma_k^2, \boldsymbol{W}_k) = 1, \\ & \theta_k \geq 0 \text{ for } k = 1, \ldots, c. \end{aligned}$$

The above optimization is non-convex, so a local optimal solution may be found by some algorithm in practice. When the dimensionality of the latent space, m, is equal to the entire dimensionality d, PPCA models are reduced to ordinary Gaussian models. Thus, PPCA models can be regarded as an extension of Gaussian models to (locally) rank-deficient data.

The KLIEP method for PPCA mixture models is called *PPCA-mixture KLIEP* (PM-KLIEP).

2.3.6 Remarks

Density-ratio estimation by density matching under the KL divergence allows one to avoid density estimation when estimating density-ratios (Section 2.3.1). Furthermore, cross-validation with respect to the KL divergence is available for model selection.

The method, called the *KL importance estimation procedure* (KLIEP), is applicable to a variety of models such as linear models, kernel models, log-linear models, Gaussian mixture models, and probabilistic principal-component-analyzer mixture models.

2.4 Density-Ratio Fitting

Here, we describe a framework of density-ratio estimation by *least-squares density-ratio fitting* (Kanamori et al., 2009).

2.4.1 Basic Framework

The model $r(\boldsymbol{x})$ of the true density-ratio function $r^*(\boldsymbol{x}) = p^*_{\mathrm{nu}}(\boldsymbol{x})/p^*_{\mathrm{de}}(\boldsymbol{x})$ is learned so that the following squared error SQ' is minimized:

$$\begin{aligned}\mathrm{SQ}'(r) &:= \frac{1}{2}\int \left(r(\boldsymbol{x}) - r^*(\boldsymbol{x})\right)^2 p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \\ &= \frac{1}{2}\int r(\boldsymbol{x})^2 p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int r(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int r^*(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned}$$

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

$$\mathrm{SQ}(r) := \frac{1}{2}\int r(\boldsymbol{x})^2 p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int r(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Approximating the expectations in SQ by empirical averages, we obtain the following optimization problem:

$$\min_r \left[\sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j)^2 - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} r(\boldsymbol{x}^{\mathrm{nu}}_i)\right]. \tag{21}$$

We refer to this formulation as *least-squares importance fitting* (LSIF). Possible hyper-parameters (such as basis parameters and regularization parameters) can be optimized by *cross-validation* with respect to the SQ criterion (Kanamori et al., 2009).

Below, two implementations of LSIF for the following linear/kernel models are described:

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell \psi_\ell(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta},$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a non-negative basis function vector, and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. Since this model is the same form as that used in KLIEP for linear/kernel models (Section 2.3.2), we may use the same basis design idea described there.

For the above linear/kernel models, Eq.(21) is expressed as

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^b}\left[\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta}\right],$$

where

$$\widehat{\boldsymbol{H}} := \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)^\top \quad \text{and} \quad \widehat{\boldsymbol{h}} := \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{nu}}_i). \tag{22}$$

2.4.2 Implementation with Non-Negativity Constraint

Here, we describe an implementation of LSIF *with* non-negativity constraint.

Let us impose non-negativity constraint $\boldsymbol{\theta} \geq \mathbf{0}_b$ since the density-ratio function is non-negative by definition. Let us further add the following regularization term to the objective function:

$$\mathbf{1}_b^\top \boldsymbol{\theta} = \|\boldsymbol{\theta}\|_1 := \sum_{\ell=1}^{b} |\theta_\ell|.$$

The term $\mathbf{1}_b^\top \boldsymbol{\theta}$ works as the ℓ_1-regularizer if it is combined with the non-negativity constraint. Then the optimization problem is expressed as follows.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^b} \quad \left[\frac{1}{2} \boldsymbol{\theta}^\top \widehat{\boldsymbol{H}} \boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \lambda \mathbf{1}_b^\top \boldsymbol{\theta} \right] \quad \text{s.t.} \quad \boldsymbol{\theta} \geq \mathbf{0}_b,$$

where λ (≥ 0) is the regularization parameter. We refer to this method as *constrained LSIF* (cLSIF; Kanamori et al., 2009). The cLSIF optimization problem is a convex quadratic program, so the unique global optimal solution may be computed by a standard optimization software.

We can also use the ℓ_2-regularizer $\boldsymbol{\theta}^\top \boldsymbol{\theta}$, instead of the ℓ_1-regularizer $\mathbf{1}_b^\top \boldsymbol{\theta}$, without changing the computational property (i.e., the optimization problem is still a convex quadratic program). However, using the ℓ_1-regularizer would be more advantageous since the solution tends to be *sparse*, i.e., many parameters take exactly zero (Williams, 1995; Tibshirani, 1996; Chen et al., 1998). Furthermore, as shown in Kanamori et al. (2009), the use of the ℓ_1-regularizer allows one to compute the entire *regularization path* efficiently (Best, 1982; Efron et al., 2004; Hastie et al., 2004), which highly improves the computational cost in the model selection phase.

An R implementation of cLSIF is available from

http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/

2.4.3 Implementation without Non-Negativity Constraint

Here, we describe another implementation of LSIF *without* the non-negativity constraint called *unconstrained LSIF* (uLSIF).

Without the non-negativity constraint, the linear regularizer $\mathbf{1}_b^\top \boldsymbol{\theta}$ used in cLSIF does not work as a regularizer. For this reason, a quadratic regularizer $\boldsymbol{\theta}^\top \boldsymbol{\theta}$ is adopted here. Then we have the following optimization problem.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\frac{1}{2} \boldsymbol{\theta}^\top \widehat{\boldsymbol{H}} \boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \frac{\lambda}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]. \tag{23}$$

Eq.(23) is an unconstrained convex quadratic program, and the solution can be computed *analytically* by solving the following system of linear equations:

$$(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b) \boldsymbol{\theta} = \widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix. The solution $\widehat{\boldsymbol{\theta}}$ of the above equation is given by

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}.$$

Since the non-negativity constraint $\boldsymbol{\theta} \geq \boldsymbol{0}_b$ was dropped, some of the obtained parameters could be negative. To compensate for this approximation error, the solution may be modified as follows (Kanamori et al., 2012):

$$\max(0, \boldsymbol{\psi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\theta}}).$$

This is the solution of the approximation method called *unconstrained LSIF* (uLSIF; Kanamori et al., 2009). An advantage of uLSIF is that the solution can be analytically computed just by solving a system of linear equations. Therefore, its computation is stable when λ is not too small.

A practically important advantage of uLSIF over cLSIF is that the score of *leave-one-out cross-validation* (LOOCV) can be computed analytically (Kanamori et al., 2009)—thanks to this property, the computational complexity for performing LOOCV is the same order as just computing a single solution.

A MATLAB® implementation of uLSIF is available from

http://sugiyama-www.cs.titech.ac.jp/~sugi/software/uLSIF/

and an R implementation of uLSIF is available from

http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/

2.4.4 Remarks

One can successfully avoid density estimation by least-squared density-ratio fitting. The least-squares methods for linear/kernel models are computationally more advantageous than alternative approaches such as moment matching (Section 2.1), probabilistic classification (Section 2.2), and density matching (Section 2.3). Indeed, the constrained method (cLSIF) for the ℓ_1-regularizer is equipped with a *regularization path tracking* algorithm. Furthermore, the unconstrained method (uLSIF) allows one to compute the density-ratio estimator analytically; the leave-one-out cross-validation score can also be computed in a closed form. Thus, the overall computation of uLSIF including model selection is highly efficient.

The fact that uLSIF has an analytic-form solution is actually very useful beyond its computational efficiency. When one wants to optimize some criterion defined using a density-ratio estimate (e.g., *mutual information*, see Cover and Thomas, 2006), the analytic-form solution of uLSIF allows one to compute the *derivative* of the target criterion analytically. Then one can develop, e.g., gradient-based and (quasi-)Newton algorithms for optimization. This property can be successfully utilized, e.g., in identifying the central subspace in *sufficient dimension reduction* (Suzuki and Sugiyama, 2010), finding independent components in *independent component analysis* (Suzuki and Sugiyama,

2011), performing dependence-minimizing regression in *causality learning* (Yamada and Sugiyama, 2010), and identifying the hetero-distributional subspace in *direct density-ratio estimation with dimensionality reduction* (Sugiyama et al., 2011b).

3 Unified Framework by Density-Ratio Matching

As reviewed in the previous section, various density-ratio estimation methods have been developed so far. In this section, we propose a new framework of density-ratio estimation by *density-ratio matching* under the *Bregman divergence* (Bregman, 1967), which includes various useful divergences (Banerjee et al., 2005; Stummer, 2007). This framework is a natural extension of the least-squares approach described in Section 2.4, and includes the existing approaches reviewed in the previous section as special cases (Section 3.2). Then we provide interpretation of density-ratio matching from two different views in Section 3.3. Finally, we give a new instance of density-ratio matching based on the *power divergence* in Section 3.4.

3.1 Basic Framework

A basic idea of density-ratio matching is to directly fit a density-ratio model $r(\boldsymbol{x})$ to the true density-ratio function $r^*(\boldsymbol{x})$ under some divergence. At a glance, this density-ratio matching problem is equivalent to the *regression problem*, which is aimed at estimating a real-valued function. However, density-ratio matching is essentially different from regression since samples of the true density-ratio function are not available. Here, we employ the *Bregman* (BR) divergence for measuring the discrepancy between the true density-ratio function and the density-ratio model.

The BR divergence is an extension of the *Euclidean distance* to a class of divergences that share similar properties. Let f be a differentiable and *strictly convex* function. Then the BR divergence associated with f from t^* to t is defined as

$$\mathrm{BR}'_f(t^* \| t) := f(t^*) - f(t) - \partial f(t)(t^* - t),$$

where ∂f is the derivative of f. Note that

$$f(t) + \partial f(t)(t^* - t)$$

is the value of the first-order *Taylor expansion* of f around t evaluated at t^*. Thus, the BR divergence evaluates the difference between the value of f at point t^* and its linear extrapolation from t (see Figure 1). $\mathrm{BR}'_f(t^* \| t)$ is a convex function with respect to t^*, but not necessarily convex with respect to t.

Here the discrepancy from the true density-ratio function r^* to a density-ratio model r is measured using the BR divergence as

$$\mathrm{BR}'_f\left(r^* \| r\right) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) \Big(f(r^*(\boldsymbol{x})) - f(r(\boldsymbol{x})) - \partial f(r(\boldsymbol{x}))(r^*(\boldsymbol{x}) - r(\boldsymbol{x})) \Big) \mathrm{d}\boldsymbol{x}. \tag{24}$$

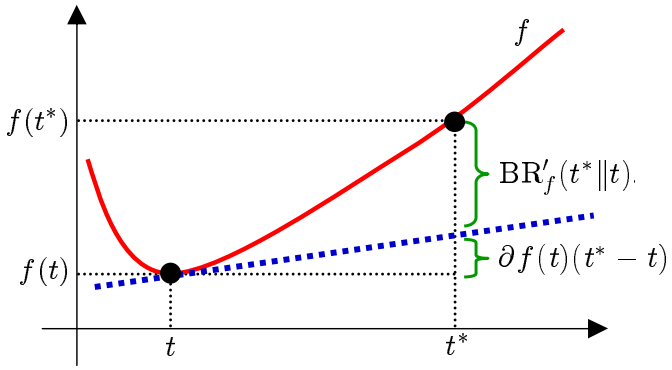


Figure 1: Bregman divergence $\mathrm{BR}'_f(t^*\|t)$.

A motivation for this choice is that the BR divergence allows one to directly obtain an *empirical approximation* for any f. Indeed, let us first extract a relevant part of $\mathrm{BR}'_f(r^*\|r)$ as

$$\mathrm{BR}'_f(r^*\|r) = \mathrm{BR}_f(r) + C,$$

where $C := \int p^*_{\mathrm{de}}(\boldsymbol{x}) f(r^*(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ is a constant independent of r, and

$$\mathrm{BR}_f(r) := \int p^*_{\mathrm{de}}(\boldsymbol{x})\Big(\partial f(r(\boldsymbol{x}))r(\boldsymbol{x}) - f(r(\boldsymbol{x}))\Big)\mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x})\partial f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{25}$$

Then an empirical approximation $\widehat{\mathrm{BR}}_f(r)$ of $\mathrm{BR}_f(r)$ is given by

$$\widehat{\mathrm{BR}}_f(r) := \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\Big(\partial f(r(\boldsymbol{x}^{\mathrm{de}}_j))r(\boldsymbol{x}^{\mathrm{de}}_j) - f(r(\boldsymbol{x}^{\mathrm{de}}_j))\Big) - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\partial f(r(\boldsymbol{x}^{\mathrm{nu}}_i)). \tag{26}$$

This immediately gives the following optimization criterion.

$$\min_r \widehat{\mathrm{BR}}_f(r),$$

where r is searched within some class of functions.

3.2 Existing Methods as Density-Ratio Matching

Here, we show that various density-ratio estimation methods reviewed in the previous section can be accommodated in the density-ratio matching framework (see Table 1).

3.2.1 Least-Squares Importance Fitting

Here, we show that the *least-squares importance fitting* (LSIF) approach introduced in Section 2.4.1 is an instance of density-ratio matching. More specifically, there exists a

Table 1: Summary of density-ratio estimation methods. In the table, 'LSIF', 'KMM', 'LR', and 'KLIEP' stand for 'least-squares importance fitting', 'kernel mean matching', 'logistic regression', and 'Kullback-Leibler Importance Estimation Procedure', respectively.

Method (Section)	$f(t)$	Model selection	Optimization
LSIF (3.2.1)	$(t-1)^2/2$	Available	Analytic
KMM (3.2.2)	$(t-1)^2/2$	Partially unavailable	Analytic
LR (3.2.3)	$t\log t-(1+t)\log(1+t)$	Available	Convex
KLIEP (3.2.4)	$t\log t-t$	Available	Convex
Robust (3.4)	$(t^{1+\alpha}-t)/\alpha$, $\alpha>0$	Available	Convex ($0<\alpha\leq 1$) Non-convex ($\alpha>1$)

BR divergence such that the optimization problem of density-ratio matching is reduced to that of LSIF.

When

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

BR (24) is reduced to the squared (SQ) distance:

$$\mathrm{SQ}'(t^*\|t):=\frac{1}{2}(t^*-t)^2.$$

Following Eqs.(25) and (26), let us denote SQ without an irrelevant constant term by $\mathrm{SQ}(r)$ and its empirical approximation by $\widehat{\mathrm{SQ}}(r)$, respectively:

$$\mathrm{SQ}(r):=\frac{1}{2}\int p^*_{\mathrm{de}}(\boldsymbol{x})r(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x}-\int p^*_{\mathrm{nu}}(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$
$$\widehat{\mathrm{SQ}}(r):=\frac{1}{2n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}r(\boldsymbol{x}^{\mathrm{de}}_j)^2-\frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}r(\boldsymbol{x}^{\mathrm{nu}}_i).$$

This agrees with the LSIF formulation given in Section 2.4.1.

3.2.2 Kernel Mean Matching

Here, we show that the solution of the moment matching method, *kernel mean matching* (KMM) introduced in Section 2.1, actually agrees with that of *unconstrained LSIF* (uLSIF; see Section 2.4.3) for specific kernel models. Since uLSIF was shown to be an instance of density-ratio matching in Section 3.2.1, the KMM solution can also be obtained in the density-ratio matching framework.

Let us consider the following kernel density-ratio model:

$$r(\boldsymbol{x})=\sum_{\ell=1}^{n_{\mathrm{de}}}\theta_\ell K(\boldsymbol{x},\boldsymbol{x}^{\mathrm{de}}_\ell),\tag{27}$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ is a *universal reproducing kernel* (Steinwart, 2001) such as the Gaussian kernel (7). Note that uLSIF and KLIEP use the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ as Gaussian centers, while the model (27) adopts the denominator samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ as Gaussian centers. For the density-ratio model (27), the matrix $\widehat{\boldsymbol{H}}$ and the vector $\widehat{\boldsymbol{h}}$ defined by Eq.(22) are expressed as

$$\widehat{\boldsymbol{H}} = \frac{1}{n_{\mathrm{de}}} \boldsymbol{K}_{\mathrm{de,de}}^2 \quad \text{and} \quad \widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}},$$

where $\boldsymbol{K}_{\mathrm{de,de}}$ and $\boldsymbol{K}_{\mathrm{de,nu}}$ are defined in Eq.(8). Then the (unregularized) uLSIF solution (see Section 2.4.3 for details) is expressed as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{uLSIF}} = \widehat{\boldsymbol{H}}^{-1} \widehat{\boldsymbol{h}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \boldsymbol{K}_{\mathrm{de,de}}^{-2} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}}. \tag{28}$$

On the other hand, let us consider an inductive variant of KMM for the kernel model (27) (see Section 2.1.2). For the density-ratio model (27), the design matrix $\boldsymbol{\Psi}_{\mathrm{de}}$ defined by Eq.(5) agrees with $\boldsymbol{K}_{\mathrm{de,de}}$. Then the KMM solution is given as follows (see Section 2.1.2):

$$\widehat{\boldsymbol{\theta}}_{\mathrm{KMM}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,de}} \boldsymbol{\Psi}_{\mathrm{de}})^{-1} \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}} = \widehat{\boldsymbol{\theta}}_{\mathrm{uLSIF}}.$$

3.2.3 Logistic Regression

Here, we show that the *logistic regression* approach introduced in Section 2.2.2 is an instance of density-ratio matching. More specifically, there exists a BR divergence such that the optimization problem of density-ratio matching is reduced to that of the logistic regression approach.

When

$$f(t) = t \log t - (1+t) \log(1+t),$$

BR (24) is reduced to the *binary Kullback-Leibler* (BKL) divergence:

$$\mathrm{BKL}'(t^* \| t) := (1+t^*) \log \frac{1+t}{1+t^*} + t^* \log \frac{t}{t^*}.$$

The name 'BKL' comes from the fact that $\mathrm{BKL}'(t^* \| t)$ is expressed as

$$\mathrm{BKL}'(t^* \| t) = (1+t^*) \mathrm{KL}_{\mathrm{bin}} \left(\frac{1}{1+t^*} \middle\| \frac{1}{1+t} \right),$$

where $\mathrm{KL}_{\mathrm{bin}}$ is the KL divergence for binary random variables defined as

$$\mathrm{KL}_{\mathrm{bin}}(p, q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

for $0 < p, q < 1$. Thus, BKL' agrees with $\mathrm{KL}_{\mathrm{bin}}$ up to the constant factor $(1+t^*)$.

Following Eqs.(25) and (26), let us denote BKL without an irrelevant constant term by $\mathrm{BKL}(r)$ and its empirical approximation by $\widehat{\mathrm{BKL}}(r)$, respectively:

$$\begin{aligned}\mathrm{BKL}(r) &:= -\int p^*_{\mathrm{de}}(\boldsymbol{x}) \log \frac{1}{1+r(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{r(\boldsymbol{x})}{1+r(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}, \\ \widehat{\mathrm{BKL}}(r) &:= -\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \log \frac{1}{1+r(\boldsymbol{x}^{\mathrm{de}}_j)} - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log \frac{r(\boldsymbol{x}^{\mathrm{nu}}_i)}{1+r(\boldsymbol{x}^{\mathrm{nu}}_i)}. \end{aligned} \tag{29}$$

Eq.(29) is a generalized expression of logistic regression (Qin, 1998). Indeed, when $n_{\mathrm{de}} = n_{\mathrm{nu}}$, the ordinary logistic regression formulation (11) can be obtained from Eq.(29) (up to a regularizer) if the *log-linear* density-ratio model (18) without the constant term θ_0 is used.

3.2.4 Kullback-Leibler Importance Estimation Procedure

Here, we show that the *KL importance estimation procedure* (KLIEP) introduced in Section 2.3.1 is an instance of density-ratio matching. More specifically, there exists a BR divergence such that the optimization problem of density-ratio matching is reduced to that of the KLIEP approach.

When

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

BR (24) is reduced to the *unnormalized Kullback-Leibler* (UKL) divergence:

$$\mathrm{UKL}'(t^* \| t) := t^* \log \frac{t^*}{t} - t^* + t.$$

Following Eqs.(25) and (26), let us denote UKL without an irrelevant constant term by $\mathrm{UKL}(r)$ and its empirical approximation by $\widehat{\mathrm{UKL}}(r)$, respectively:

$$\mathrm{UKL}(r) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{30}$$

$$\widehat{\mathrm{UKL}}(r) := \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i). \tag{31}$$

Let us further impose that the ratio model $r(\boldsymbol{x})$ is non-negative for all $\boldsymbol{x}$ and is normalized with respect to $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$:

$$\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) = 1.$$

Then the optimization criterion is reduced to as follows.

$$\begin{aligned}&\max_{r} \ \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\log r(\boldsymbol{x}_i^{\mathrm{nu}})\\ &\text{s.t.} \ \ \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}_j^{\mathrm{de}}) = 1 \text{ and } r(\boldsymbol{x}) \geq 0 \text{ for all } \boldsymbol{x}.\end{aligned}$$

This agrees with the KLIEP formulation reviewed in Section 2.3.1.

3.3 Interpretation of Density-Ratio Matching

Here, we show the correspondence between the density-ratio matching approach and a divergence estimation method, and the correspondence between the density-ratio matching approach and a moment-matching approach.

3.3.1 Divergence Estimation View

We first show that our density-ratio matching formulation can be interpreted as *divergence estimation* based on the *Ali-Silvey-Csiszár* (ASC) divergence (Ali and Silvey, 1966; Csiszár, 1967), which is also known as the *f-divergence*.

Let us consider the ASC divergence for measuring the discrepancy between two probability density functions. An ASC divergence is defined using a *convex function* f such that $f(1) = 0$ as follows:

$$\mathrm{ASC}_f(p^*_{\mathrm{nu}}\|p^*_{\mathrm{de}}) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) f\left(\frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}. \tag{32}$$

The ASC divergence is reduced to the *Kullback-Leibler* (KL) divergence (Kullback and Leibler, 1951) if $f(t) = t\log t$, and the *Pearson* (PE) divergence (Pearson, 1900) if $f(t) = \frac{1}{2}(t-1)^2$.

Let $\partial f(t)$ be the *sub-differential* of f at a point t ($\in \mathbb{R}$), which is a set defined as follows (Rockafellar, 1970):

$$\partial f(t) := \{z \in \mathbb{R} \mid f(s) \geq f(t) + z(s-t), \ \forall s \in \mathbb{R}\}.$$

If f is differentiable at t, then the sub-differential is reduced to the ordinary derivative. Although the sub-differential is a set in general, for simplicity, we treat $\partial f(r)$ as a single element if there is no confusion. Below, we assume that f is *closed*, i.e., its *epigraph* is a closed set (Rockafellar, 1970).

Let f^* be the *conjugate dual function* associated with f defined as

$$f^*(u) := \sup_t[tu - f(t)] = -\inf_t[f(t) - tu].$$

Since f is a closed convex function, we also have

$$f(t) = -\inf_u [f^*(u) - tu]. \tag{33}$$

For the KL divergence where $f(t) = t \log t$, the conjugate dual function is given by $f^*(u) = \exp(u-1)$. For the PE divergence where $f(t) = (t-1)^2/2$, the conjugate dual function is given by $f^*(u) = u^2/2 + u$.

Substituting Eq.(33) into Eq.(32), we have the following lower bound (Keziou, 2003):

$$\mathrm{ASC}_f(p^*_{\mathrm{nu}} \| p^*_{\mathrm{de}}) = -\inf_g \mathrm{ASC}'_f(g),$$

where

$$\mathrm{ASC}'_f(g) := \int f^*(g(\boldsymbol{x})) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int g(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{34}$$

By taking the derivative of the integrand for each $\boldsymbol{x}$ and equating it to zero, we can show that the infimum of ASC'_f is attained at g such that

$$\partial f^*(g(\boldsymbol{x})) = \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})} = r^*(\boldsymbol{x}).$$

Thus, minimizing $\mathrm{ASC}'_f(g)$ yields the true density-ratio function $r^*(\boldsymbol{x})$.

For some g, there exists r such that

$$g = \partial f(r).$$

Then $f^*(g)$ is expressed as

$$f^*(g) = \sup_s \Big[s \partial f(r) - f(s) \Big].$$

According to the *variational principle* (Jordan et al., 1999), the supremum in the right-hand side of the above equation is attained at $s = r$. Thus, we have

$$f^*(g) = r \partial f(r) - f(r).$$

Then the lower bound $\mathrm{ASC}'_f(g)$ defined by Eq.(34) can be expressed as

$$\mathrm{ASC}'_f(g) = \int p^*_{\mathrm{de}}(\boldsymbol{x}) \Big(r(\boldsymbol{x}) \partial f(r(\boldsymbol{x})) - f(r(\boldsymbol{x})) \Big) \mathrm{d}\boldsymbol{x} - \int \partial f(r(\boldsymbol{x})) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

This is equivalent to the criterion BR_f defined by Eq.(25). Thus, density-ratio matching under the BR divergence can be interpreted as divergence estimation under the ASC divergence.

3.3.2 Moment Matching View

Next, we investigate the correspondence between the density-ratio matching approach and a moment-matching approach. To this end, we focus on the ideal situation where the true density-ratio function r^* is included in the density-ratio model r.

The non-linear version of finite-order moment matching (see Section 2.1.1) learns the density-ratio model r so that the following criterion is minimized:

$$\left\| \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{\phi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2,$$

where $\boldsymbol{\phi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^m$ is some vector-valued function. Under the assumption that the density-ratio model r can represent the true density-ratio r^*, we have the following estimation equation:

$$\int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{\phi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbf{0}_m, \tag{35}$$

where $\mathbf{0}_m$ denotes the m-dimensional vector with all zeros.

On the other hand, the density-ratio matching approach described in Section 3.1 learns the density-ratio model r so that the following criterion is minimized:

$$\int p^*_{\mathrm{de}}(\boldsymbol{x}) \partial f(r(\boldsymbol{x})) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{de}}(\boldsymbol{x}) f(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \partial f(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}.$$

Taking the derivative of the above criterion with respect to parameters in the density-ratio model r and equate it to zero, we have the following estimation equation:

$$\int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x}) \nabla r(\boldsymbol{x}) \partial^2 f(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \nabla r(\boldsymbol{x}) \partial^2 f(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} = \mathbf{0}_b,$$

where ∇ denotes the differential operator with respect to parameters in the density-ratio model r, and b is the number of parameters. This implies that putting

$$\boldsymbol{\phi}(\boldsymbol{x}) = \nabla r(\boldsymbol{x}) \partial^2 f(r(\boldsymbol{x}))$$

in Eq.(35) gives the same estimation equation as density-ratio matching, resulting in the same optimal solution.

3.4 Basu's Power Divergence for Robust Density-Ratio Estimation

Finally, we introduce a new instance of density-ratio matching based on Basu's *power* divergence (BA divergence; Basu et al., 1998).

3.4.1 Derivation

For $\alpha > 0$, let

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

Then BR (24) is reduced to the BA divergence:

$$\mathrm{BA}'_\alpha(t^* \| t) := t^\alpha (t - t^*) - \frac{t^*(t^\alpha - (t^*)^\alpha)}{\alpha}.$$

Following Eqs.(25) and (26), let us denote BA'_α without an irrelevant constant term by $\mathrm{BA}_\alpha(r)$ and its empirical approximation by $\widehat{\mathrm{BA}}_\alpha(r)$, respectively:

$$\mathrm{BA}_\alpha(r) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x})^{\alpha+1} \mathrm{d}\boldsymbol{x} - \left(1 + \frac{1}{\alpha}\right) \int p^*_{\mathrm{nu}}(\boldsymbol{x}) r(\boldsymbol{x})^\alpha \mathrm{d}\boldsymbol{x} + \frac{1}{\alpha},$$

$$\widehat{\mathrm{BA}}_\alpha(r) := \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j)^{\alpha+1} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} r(\boldsymbol{x}^{\mathrm{nu}}_i)^\alpha + \frac{1}{\alpha}.$$

The density-ratio model r is determined so that $\widehat{\mathrm{BA}}_\alpha(r)$ is minimized.

When $\alpha = 1$, the BA divergence is reduced to the twice SQ divergence (see Section 2.4):

$$\widehat{\mathrm{BA}}_1 = 2\widehat{\mathrm{SQ}}.$$

Similarly, the fact

$$\lim_{\alpha \to 0} \frac{t^\alpha - 1}{\alpha} = \log t$$

implies that the BA divergence tends to the UKL divergence as $\alpha \to 0$ (see Section 3.2.4):

$$\lim_{\alpha \to 0} \widehat{\mathrm{BA}}_\alpha(r) = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i) = \widehat{\mathrm{UKL}}(r).$$

Thus, the BA divergence essentially includes the SQ and UKL divergences as special cases, and is substantially more general.

3.4.2 Robustness

Let us take the derivative of $\widehat{\mathrm{BA}}_\alpha(r)$ with respect to parameters included in the density-ratio model r, and equate it to zero. Then we have the following estimation equation:

$$\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j)^\alpha \nabla r(\boldsymbol{x}^{\mathrm{de}}_j) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} r(\boldsymbol{x}^{\mathrm{nu}}_i)^{\alpha-1} \nabla r(\boldsymbol{x}^{\mathrm{nu}}_i) = \boldsymbol{0}_b, \tag{36}$$

where ∇ is the differential operator with respect to parameters in the density-ratio model r, b denotes the number of parameters, and $\mathbf{0}_b$ denotes the b-dimensional vector with all zeros.

As explained in Section 3.4.1, the BA method with $\alpha \to 0$ corresponds to KLIEP (using the UKL divergence). According to Eq.(31), the estimation equation of KLIEP is given as follows (this also agrees with Eq.(36) with $\alpha = 0$):

$$\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \nabla r(\boldsymbol{x}_j^{\mathrm{de}}) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} r(\boldsymbol{x}_i^{\mathrm{nu}})^{-1} \nabla r(\boldsymbol{x}_i^{\mathrm{nu}}) = \mathbf{0}_b.$$

Comparing this with Eq.(36), we see that the BA method can be regarded as a weighted version of KLIEP according to $r(\boldsymbol{x}_j^{\mathrm{de}})^\alpha$ and $r(\boldsymbol{x}_i^{\mathrm{nu}})^\alpha$. When $r(\boldsymbol{x}_j^{\mathrm{de}})$ and $r(\boldsymbol{x}_i^{\mathrm{nu}})$ are less than 1, the BA method down-weights the effect of those samples. Thus, 'outlying' samples relative to the density-ratio model r tend to have less influence on parameter estimation, which will lead to *robust* estimators (Basu et al., 1998).

Since LSIF corresponds to $\alpha = 1$, LSIF is more robust against outliers than KLIEP (which corresponds to $\alpha \to 0$) in the above sense, and BA with $\alpha > 1$ would be even more robust.

3.4.3 Numerical Examples

Here we illustrate the behavior of the robust density-ratio estimation method based on the BA divergence using artificial data sets.

Let the numerator and denominator densities be defined as follows (Figure 2(a)):

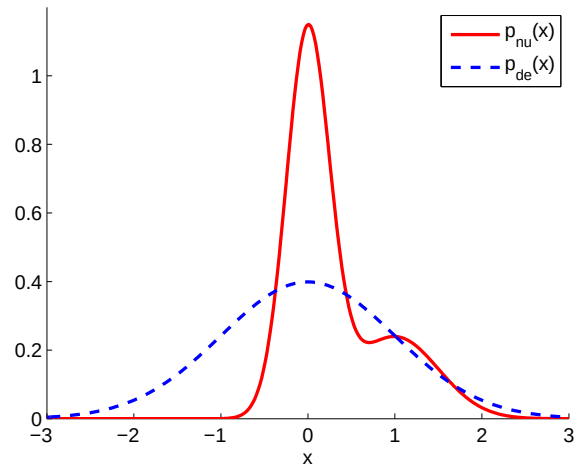
$$p_{\mathrm{nu}}^*(x) = 0.7N\left(x; 0, 0.25^2\right) + 0.3N\left(x; 1, 0.5^2\right) \quad \text{and} \quad p_{\mathrm{de}}^*(x) = N(x; 0, 1^2),$$

where $N(x; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2,. We draw $n_{\mathrm{nu}} = n_{\mathrm{de}} = 300$ samples from each density, which are illustrated in Figure 2(b).

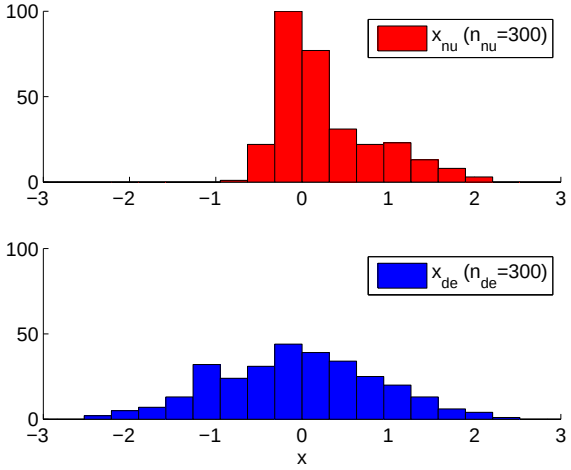
Here, we employ the Gaussian-kernel density-ratio model (17), and determine the model parameters so that $\widehat{\mathrm{BA}}_\alpha(r)$ with a quadratic regularizer is minimized under the non-negativity constraint:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathbb{R}^b} \; & \left[\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \left(\sum_{\ell=1}^{n_{\mathrm{nu}}} \theta_\ell K(\boldsymbol{x}_j^{\mathrm{nu}}, \boldsymbol{x}_\ell^{\mathrm{nu}}) \right)^{\alpha+1} \right. \\ & \left. - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \left(\sum_{\ell=1}^{n_{\mathrm{nu}}} \theta_\ell K(\boldsymbol{x}_i^{\mathrm{de}}, \boldsymbol{x}_\ell^{\mathrm{nu}}) \right)^{\alpha} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right] \\ \text{s.t.} \quad & \boldsymbol{\theta} \geq \mathbf{0}_b. \end{aligned} \tag{37}$$

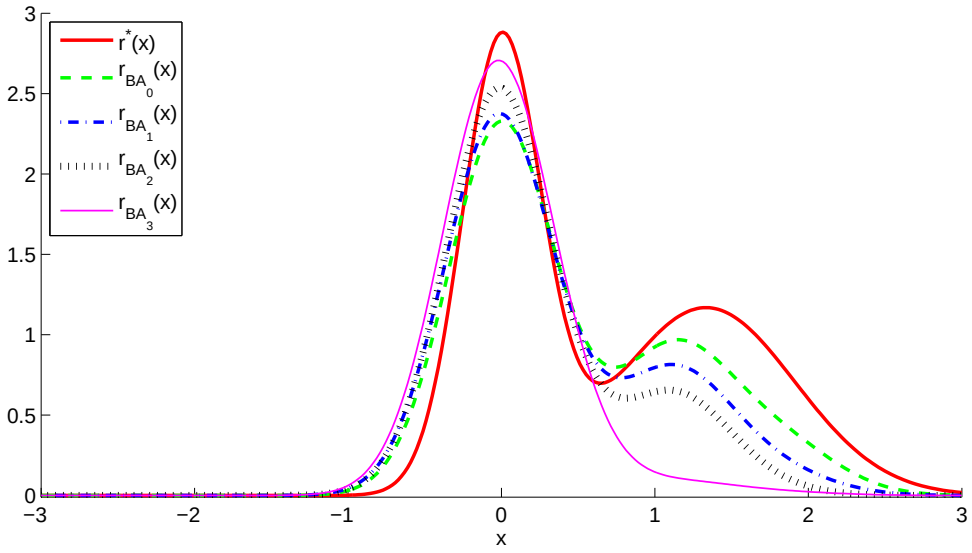
Note that this optimization problem is convex for $0 < \alpha \leq 1$. In our implementation, we solve the above optimization problem by gradient-projection, i.e., the parameters are iteratively updated by gradient descent with respect to the objective function, and the



(a) Numerator and denominator density functions.



(b) Numerator and denominator sample points



(c) True and estimated density-ratio functions

Figure 2: Numerical examples.

solution is projected back to the feasible region by rounding-up negative parameters to zero. Before solving the optimization problem (37), we run uLSIF (see Section 2.4.3) and obtain cross-validation estimates of the Gaussian width σ and the regularization parameter λ. We then fix the Gaussian width and the regularization parameter in the BA method to these values, and solve the optimization problem (37) by gradient-projection with $\boldsymbol{\theta} = \mathbf{1}_b/b$ as the initial solution.

Figure 2(c) shows the true and estimated density-ratio functions by the BA methods for $\alpha = 0, 1, 2, 3$. The true density-ratio function has two peaks—higher one at $x = 0$ and lower one at around $x = 1.2$. The graph shows that, as α increases, estimated density-

ratio functions tend to focus on approximating the higher peak and ignore the lower peak. Thus, if numerator samples drawn from the right-hand Gaussian (i.e., $N\left(x;1,0.5^2\right)$) are regarded as outliers, the BA methods with larger α are more robust against these outliers.

We further investigate the issue of robustness against outliers more systematically. Let

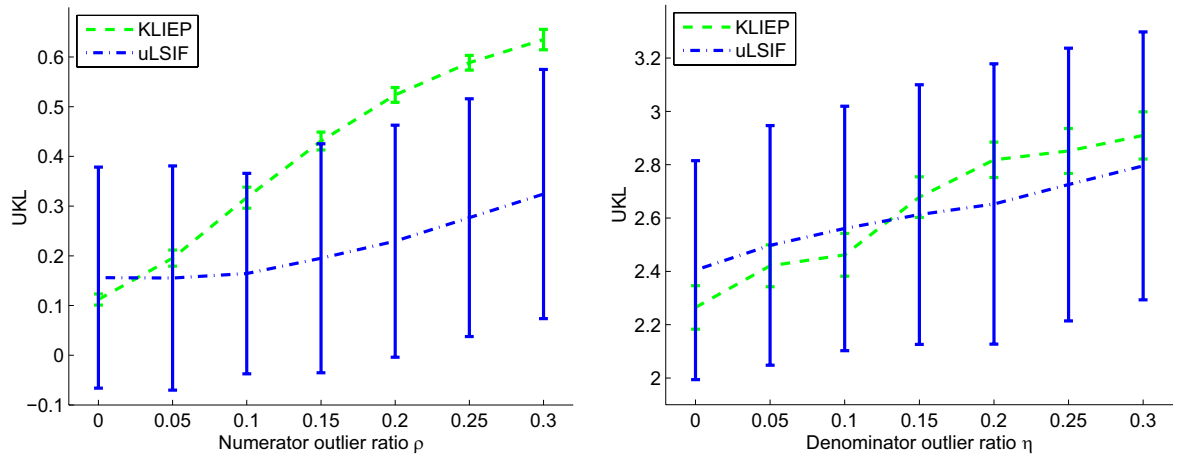
$$p^*_{\mathrm{nu}}(x) = (1-\rho)N\left(x;0,0.25^2\right) + \rho N\left(x;1,0.5^2\right),$$
$$p^*_{\mathrm{de}}(x) = (1-\eta)N(x;0,1^2) + \eta N(x;0,0.5^2),$$

where ρ and η are the numerator and denominator outlier ratio, respectively; samples drawn from the second densities $N\left(x;1,0.5^2\right)$ and $N(x;0,0.5^2)$ are regarded as outliers. Let $n_{\mathrm{nu}} = n_{\mathrm{de}} = 300$, and we evaluate how the accuracy of density-ratio estimation is influenced by outliers. In the first set of experiments, we fix the denominator outlier ratio to $\eta = 0$ (i.e., no outlier) and change the numerator outlier ratio as $\rho = 0, 0.05, 0.1, \ldots, 0.3$. In the second set of experiments, we fix the numerator outlier ratio to $\rho = 0$ (i.e., no outlier) and change the denominator outlier ratio as $\eta = 0, 0.05, 0.1, \ldots, 0.3$. The approximation error of a density-ratio estimator $\widehat{r}$ is evaluated by $\mathrm{UKL}\left(\widehat{r}\right)$ defined by Eq.(30), which correspond to the BA divergence with $\alpha \to 0$ as explained in Section 3.4.1. Here, $\mathrm{UKL}\left(\widehat{r}\right)$ is numerically approximated using 1000 samples independently taken following $p^*_{\mathrm{nu}}(x)$ with $\rho = 0$ (i.e., no outliers) and 1000 samples independently taken following $p^*_{\mathrm{de}}(x)$ with $\eta = 0$ (i.e., no outliers). Note that these samples are not used for obtaining a density-ratio estimator $\widehat{r}$. For obtaining density-ratio estimators, we use off-the-shelf MATLAB implementation of KLIEP (which corresponds to the BA method with $\alpha \to 0$) and uLSIF (which corresponds to the BA method with $\alpha = 1$) available from the web (see Section 2.3 and Section 2.4). This renders a more practical setup of density-ratio estimation.

The median and standard deviation of UKL values for KLIEP and uLSIF over 100 runs are plotted in Figure 3. Note that the standard deviation is divided by 5 in the plots for clear visibility. The graphs show that KLIEP works better than uLSIF when the outlier ratio is small. This is natural consequences since KLIEP tries to minimizes UKL (see Section 3.2.4). However, as the outlier ratio increases, the approximation error of KLIEP grows rapidly. On the other hand, the approximation error of uLSIF grows rather mildly, showing the robustness of uLSIF against outliers. This phenomenon well agrees with the argument in Section 3.4.2.

However, the error bars of uLSIF are much larger than KLIEP. This would be caused by the fact that the 'effective' number of samples used in uLSIF is smaller than that of KLIEP due to the down-weighting effect explained in Section 3.4.2. Thus, the statistical efficiency of uLSIF would be lower than KLIEP, which is a common trade-off in robust statistical methods (Huber, 1981).

Another observation from these experimental results is that numerator outliers more strongly degrade the accuracy of KLIEP than denominator outliers.



(a) The numerator outlier ratio ρ is changed while the denominator outlier ratio is fixed to $\eta = 0$ (i.e., no outliers).

(b) The denominator outlier ratio η is changed while the numerator outlier ratio is fixed to $\rho = 0$ (i.e., no outliers).

Figure 3: The median and standard deviation of UKL values for KLIEP and uLSIF over 100 runs when the number of outlier samples is changed. For clear visibility, the standard deviation is divided by 5 in the plots.

4 Conclusions

In this paper, we addressed the problem of density-ratio estimation. We first provided a comprehensive review of density-ratio estimation methods, including the *moment matching approach* (Section 2.1), the *probabilistic classification approach* (Section 2.2), the *density matching approach* (Section 2.3), and the *density-ratio fitting approach* (Section 2.4). Then we proposed a novel framework of density-ratio estimation by density-ratio fitting under the *Bregman divergence* (Section 3.1). We showed that our novel framework accommodates the existing approaches reviewed above, and is substantially more general. Within this novel framework, we developed a robust density-ratio estimation method based on Basu's *power* divergence.

The power divergence method allows us to systematically compare the robustness of the density matching approach based on the KL divergence (KLIEP) and the density-ratio fitting approach based on the Pearson divergence (uLSIF). However, the robustness of the probabilistic classification approach is still unknown, which needs to be investigated in our future work.

Experimentally, we observed that numerator outliers tend to more significantly degrade the accuracy of KLIEP than denominator samples, while uLSIF is reasonably stable for both cases. It is interesting to investigate this experimental tendency theoretically, together with convergence properties of the robust method.

In the power divergence method, the choice of robustness parameter α is an open issue. Although there seems to be no universal way for this (Basu et al., 1998; Jones et al., 2001; Fujisawa and Eguchi, 2008), a practical approach would be to use cross-validation over a

fixed divergence such as the squared distance.

Acknowledgements

Masashi Sugiyama was supported by SCAT, AOARD, and the JST PRESTO program. Taiji Suzuki was supported by MEXT Grant-in-Aid for Young Scientists (B) 22700289. Takafumi Kanamori was supported by MEXT Grant-in-Aid for Young Scientists (B) 20700251.

References

Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society, Series B 28(1):131–142

Banerjee A, Merugu S, Dhillon IS, Ghosh J (2005) Clustering with Bregman divergences. Journal of Machine Learning Research 6:1705–1749

Basu A, Harris IR, Hjort NL, Jones MC (1998) Robust and efficient estimation by minimising a density power divergence. Biometrika 85(3):549–559

Best MJ (1982) An algorithm for the solution of the parametric quadratic programming problem. Tech. Rep. 82-24, Faculty of Mathematics, University of Waterloo

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

Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7:200–217

Caruana R, Pratt L, Thrun S (1997) Multitask learning. Machine Learning 28:41–75

Cayton L (2008) Fast nearest neighbor retrieval for Bregman divergences. In: McCallum A, Roweis S (eds) Proceedings of the 25th Annual International Conference on Machine Learning (ICML2008), Omnipress, pp 112–119

Chen SS, Donoho DL, Saunders MA (1998) Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20(1):33–61

Cheng KF, Chu CK (2004) Semiparametric density estimation under a two-sample density ratio model. Bernoulli 10(4):583–604

Collins M, Schapire RE, Singer Y (2002) Logistic regression, adaboost and Bregman distances. Machine Learning 48(1-3):253–285

Cover TM, Thomas JA (2006) Elements of Information Theory, 2nd edn. John Wiley & Sons, Inc., Hoboken, NJ, USA

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

Dhillon I, Sra S (2006) Generalized nonnegative matrix approximations with Bregman divergences. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA, pp 283–290

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. The Annals of Statistics 32(2):407–499

Fujisawa H, Eguchi S (2008) Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis 99(9):2053–2081

Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B (2009) Covariate shift by kernel mean matching. In: Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (eds) Dataset Shift in Machine Learning, MIT Press, Cambridge, MA, USA, chap 8, pp 131–160

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, USA

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

Hido S, Tsuboi Y, Kashima H, Sugiyama M, Kanamori T (2008) Inlier-based outlier detection via direct density ratio estimation. In: Giannotti F, Gunopulos D, Turini F, Zaniolo C, Ramakrishnan N, Wu X (eds) Proceedings of IEEE International Conference on Data Mining (ICDM2008), Pisa, Italy, pp 223–232

Hido S, Tsuboi Y, Kashima H, Sugiyama M, 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 KM, Schölkopf B (2007) Correcting sample selection bias by unlabeled data. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA, USA, pp 601–608

Huber PJ (1981) Robust Statistics. Wiley, New York, NY, USA

Jones MC, Hjort NL, Harris IR, Basu A (2001) A comparison of related density-based minimum divergence estimators. Biometrika 88:865–873

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Machine Learning 37(2):183

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

Kanamori T, Suzuki T, Sugiyama M (2010) Theoretical analysis of density ratio estimation. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E93-A(4):787–798

Kanamori T, Suzuki T, Sugiyama M (2012) Kernel-based least-squares density-ratio estimation I: Statistical analysis. Machine Learning To appear

Kawahara Y, Sugiyama M (2009) Change-point detection in time-series data by direct density-ratio estimation. In: Park H, Parthasarathy S, Liu H, Obradovic Z (eds) Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009), Sparks, Nevada, USA, pp 389–400

Keziou A (2003) Dual representation of ϕ-divergences and applications. Comptes Rendus Mathématique 336(10):857–862

Keziou A, Leoni-Aubin S (2005) Test of homogeneity in semiparametric two-sample density ratio models. Comptes Rendus Mathématique 340(12):905–910

Kimura M, Sugiyama M (2011) Dependence-maximization clustering with least-squares mutual information. Journal of Advanced Computational Intelligence and Intelligent Informatics 15(7):800–805

Kullback S, Leibler RA (1951) On information and sufficiency. Annals of Mathematical Statistics 22:79–86

Minka TP (2007) A comparison of numerical optimizers for logistic regression. Tech. rep., Microsoft Research, URL http://research.microsoft.com/~minka/papers/logreg/minka-logreg.pdf

Murata N, Takenouchi T, Kanamori T, Eguchi S (2004) Information geometry of U-boost and Bregman divergence. Neural Computation 16(7):1437–1481

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

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 50(302):157–175

Qin J (1998) Inferences for case-control and semiparametric two-sample density ratio models. Biometrika 85(3):619–630

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (eds) (2009) Dataset Shift in Machine Learning. MIT Press, Cambridge, MA, USA

Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton, NJ, USA

Schölkopf B, Smola AJ (2002) Learning with Kernels. MIT Press, Cambridge, MA, USA

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 BW (1978) Density ratios, empirical likelihood and cot death. Journal of the Royal Statistical Society, Series C 27(1):26–33

Smola A, Song L, Teo CH (2009) Relative novelty detection. In: van Dyk D, Welling M (eds) Proceedings of Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009), Clearwater Beach, FL, USA, JMLR Workshop and Conference Proceedings, vol 5, pp 536–543

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

Stummer W (2007) Some Bregman distances between financial diffusion processes. Proceedings in Applied Mathematics and Mechanics 7:1050,503–1050,504

Sugiyama M (2010) Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. IEICE Transactions on Information and Systems E93-D(10):2690–2701

Sugiyama M, Kawanabe M (2011) Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. MIT Press, Cambridge, MA, USA, to appear

Sugiyama M, Müller KR (2005) Input-dependent estimation of generalization error under covariate shift. Statistics & Decisions 23(4):249–279

Sugiyama M, Krauledat M, Müller KR (2007) Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research 8:985–1005

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(4):699–746

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, Takeuchi I, Suzuki T, Kanamori T, Hachiya H, Okanohara D (2010) Least-squares conditional density estimation. IEICE Transactions on Information and Systems E93-D(3):583–594

Sugiyama M, Suzuki T, Itoh Y, Kanamori T, Kimura M (2011a) Least-squares two-sample test. Neural Networks 24(7):735–751

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

Sugiyama M, Suzuki T, Kanamori T (2012) Density Ratio Estimation in Machine Learning. Cambridge University Press, Cambridge, UK, to appear

Suzuki T, Sugiyama M (2009) Estimating squared-loss mutual information for independent component analysis. In: Adali T, Jutten C, Romano JMT, Barros AK (eds) Independent Component Analysis and Signal Separation, Springer, Berlin, Germany, Lecture Notes in Computer Science, vol 5441, pp 130–137

Suzuki T, Sugiyama M (2010) Sufficient dimension reduction via squared-loss mutual information estimation. In: Teh YW, Tiggerington M (eds) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010), Sardinia, Italy, JMLR Workshop and Conference Proceedings, vol 9, pp 804–811

Suzuki T, Sugiyama M (2011) Least-squares independent component analysis. Neural Computation 23(1):284–301

Suzuki T, Sugiyama M, Sese J, Kanamori T (2008) Approximating mutual information by maximum likelihood density ratio estimation. In: Saeys Y, Liu H, Inza I, Wehenkel L, de Peer YV (eds) Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008), Antwerp, Belgium, JMLR Workshop and Conference Proceedings, vol 4, pp 5–20

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

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

Tibshirani R (1996) Regression shrinkage and subset selection with the lasso. Journal of the Royal Statistical Society, Series B 58(1):267–288

Tipping ME, Bishop CM (1999) Mixtures of probabilistic principal component analyzers. Neural Computation 11(2):443–482

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

Tsuda K, Rätsch G, Warmuth M (2005) Matrix exponential gradient updates for on-line learning and Bregman projection. In: Saul LK, Weiss Y, Bottou L (eds) Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, pp 1425–1432

Williams PM (1995) Bayesian regularization and pruning using a Laplace prior. Neural Computation 7(1):117–143

Wu L, Jin R, Hoi SCH, Zhu J, Yu N (2009) Learning Bregman distance functions and its application for semi-supervised clustering. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (eds) Advances in Neural Information Processing Systems 22, pp 2089–2097

Yamada M, Sugiyama M (2009) Direct importance estimation with Gaussian mixture models. IEICE Transactions on Information and Systems E92-D(10):2159–2162

Yamada M, 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), The AAAI Press, Atlanta, Georgia, USA, pp 643–648

Yamada M, Sugiyama M (2011) Cross-domain object matching with model selection. In: Gordon G, Dunson D, Dudík M (eds) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011), Fort Lauderdale, Florida, USA, pp 807–815

Yamada M, Sugiyama M, Wichern G, Simm J (2010) Direct importance estimation with a mixture of probabilistic principal component analyzers. IEICE Transactions on Information and Systems E93-D(10):2846–2849

Yamada M, Sugiyama M, Wichern G, Simm J (2011) Improving the accuracy of least-squares probabilistic classifiers. IEICE Transactions on Information and Systems E94-D(6):1337–1340