IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E93-A, no.4, pp.787–798, 2010.

Theoretical Analysis of Density Ratio Estimation

Takafumi Kanamori (kanamori@is.nagoya-u.ac.jp)
Nagoya University

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

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Tokyo Institute of Technology
and
Japan Science and Technology Agency

Abstract

Density ratio estimation has gathered a great deal of attention recently since it can be used for various data processing tasks. In this paper, we consider three methods of density ratio estimation: (A) the numerator and denominator densities are separately estimated and then the ratio of the estimated densities is computed, (B) a logistic regression classifier discriminating denominator samples from numerator samples is learned and then the ratio of the posterior probabilities is computed, and (C) the density ratio function is directly modeled and learned by minimizing the empirical Kullback-Leibler divergence. We first prove that when the numerator and denominator densities are known to be members of the exponential family, (A) is better than (B) and (B) is better than (C). Then we show that once the model assumption is violated, (C) is better than (A) and (B). Thus in practical situations where no exact model is available, (C) would be the most promising approach to density ratio estimation.

Keywords

density ratio estimation, density estimation, logistic regression, asymptotic analysis, Gaussian assumption.

1 Introduction

The ratio of two probability density functions has been demonstrated to be useful in various data processing tasks [21], such as non-stationarity adaptation [18, 35, 23, 22, 17, 27], outlier detection [7, 19], conditional density estimation [26], feature selection [31, 30], feature extraction [29], and independent component analysis [28]. Thus accurately estimating the density ratio is an important and challenging research topic in the machine learning and data mining communities.

A naive approach to density ratio estimation is (A) density ratio estimation by separate maximum likelihood density estimation—first the numerator and denominator densities are separately estimated and then the ratio of the estimated densities is computed. However, density estimation is substantially more difficult than density ratio estimation and the above two-shot process of first estimating the densities and then taking their ratio is thought to be less accurate. To cope with this problem, various alternative methods have been developed recently, which allow one to estimate the density ratio without going through density estimation [16, 8, 15, 25, 10].

In this paper, we consider the following two methods in addition to the method (A): (B) density ratio estimation by logistic regression [16, 3, 1]—a logistic regression classifier discriminating numerator samples from denominator samples is used for density ratio estimation, and (C) direct density ratio estimation by empirical Kullback-Leibler divergence minimization [15, 25]—the density ratio function is directly modeled and learned. The goal of this paper is to theoretically compare the accuracy of these three density ratio estimation schemes.

We first prove that when the numerator and denominator densities are known to be members of the exponential family, (A) is better than (B) and (B) is better than (C). The fact that (A) is better than (B) could be regarded as an extension of the existing result for binary classification [5]—estimating data generating densities by maximum likelihood estimation has higher statistical efficiency than logistic regression in classification scenarios. On the other hand, the fact that (B) is better than (C) follows from the fact that (B) has the smallest asymptotic variance in a class of semi-parametric estimators [16].

We then show that when the model assumption is violated, (C) is better than (A) and (B). Our statement is that the estimator obtained by (C) converges to the projection of the true density ratio function onto the target parametric model (i.e., the optimal approximation in the model), while the estimators obtained by (A) and (B) do not generally converge to the projection.

Since model misspecification would be a usual situation in practice, (C) is the most promising approach in density ratio estimation. In a regression framework, an asymptotic analysis with a similar spirit exists [14].

2 Density Ratio Estimation

In this section, we formulate the problem of density ratio estimation and review three density ratio estimators.

2.1 Problem Formulation

Let $\mathcal{X}$ $(\subset \mathbb{R}^d)$ be the data domain and suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^n$ drawn from a distribution with density $p_{\mathrm{nu}}^*(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^n$ drawn from another distribution with density $p_{\mathrm{de}}^*(\boldsymbol{x})$:

$$\begin{aligned}\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^n &\overset{i.i.d.}{\sim} p_{\mathrm{nu}}^*(\boldsymbol{x}),\\ \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^n &\overset{i.i.d.}{\sim} p_{\mathrm{de}}^*(\boldsymbol{x}).\end{aligned}$$

The subscripts 'nu' and 'de' denote 'numerator' and 'denominator', respectively. We assume that the latter density $p_{\mathrm{de}}^*(\boldsymbol{x})$ is strictly positive, i.e.,

$$p_{\mathrm{de}}^*(\boldsymbol{x}) > 0, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

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

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

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

The goal of this paper is to theoretically compare the performance of the following three density ratio estimators:

(A) Density ratio estimation by separate maximum likelihood density estimation (see Section 2.3 for details),

(B) Density ratio estimation by logistic regression [16, 3, 1] (see Section 2.4 for details),

(C) Direct density ratio estimation by empirical Kullback-Leibler divergence minimization [15, 25] (see Section 2.5 for details).

2.2 Measure of Accuracy

Let us consider the *unnormalized Kullback-Leibler divergence* [2] from the true density $p_{\mathrm{nu}}^*(\boldsymbol{x})$ to its estimator $\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})$:

$$\mathrm{UKL}(p_{\mathrm{nu}}^* \| \widehat{r} \cdot p_{\mathrm{de}}^*) := \int p_{\mathrm{nu}}^*(\boldsymbol{x}) \log \frac{p_{\mathrm{nu}}^*(\boldsymbol{x})}{\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} - 1 + \int \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1}$$

$\mathrm{UKL}(p_{\mathrm{nu}}^*(\boldsymbol{x}) \| \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x}))$ is non-negative for all $\widehat{r}$ and vanishes if and only if $\widehat{r} = r^*$. If $\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})$ is normalized to be a probability density function, i.e.,

$$\int \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1,$$

then the unnormalized Kullback-Leibler divergence is reduced to the ordinary Kullback-Leibler divergence [12]:

$$\mathrm{KL}(p^*_{\mathrm{nu}} \| \widehat{r} \cdot p^*_{\mathrm{de}}) := \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{\widehat{r}(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}. \tag{2}$$

In our theoretical analysis, we use the expectation of $\mathrm{UKL}(p^*_{\mathrm{nu}} \| \widehat{r} \cdot p^*_{\mathrm{de}})$ over $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^n$ and $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^n$ as the measure of accuracy of a density ratio estimator $\widehat{r}(\boldsymbol{x})$:

$$J(\widehat{r}) := \mathbb{E}\left[\mathrm{UKL}(p^*_{\mathrm{nu}} \| \widehat{r} \cdot p^*_{\mathrm{de}})\right], \tag{3}$$

where $\mathbb{E}$ denotes the expectation over $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^n$ and $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^n$.

In the rest of this section, the three methods of density ratio estimation we are dealing with are described in detail.

2.3 Method (A): Density Ratio Estimation by Separate Maximum Likelihood Density Estimation

For $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$, two parametric models $p_{\mathrm{nu}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{nu}})$ and $p_{\mathrm{de}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{de}})$ such that

$$\begin{aligned}
&\int p_{\mathrm{nu}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{nu}}) \mathrm{d}\boldsymbol{x} = 1, \quad \forall \boldsymbol{\theta}_{\mathrm{nu}} \in \Theta_{\mathrm{nu}}, \\
&\quad p_{\mathrm{nu}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{nu}}) \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \ \forall \boldsymbol{\theta}_{\mathrm{nu}} \in \Theta_{\mathrm{nu}}, \\
&\int p_{\mathrm{de}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{de}}) \mathrm{d}\boldsymbol{x} = 1 \quad \forall \boldsymbol{\theta}_{\mathrm{de}} \in \Theta_{\mathrm{de}}, \\
&\quad p_{\mathrm{de}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{de}}) \geq 0 \quad \forall \boldsymbol{x} \in \mathcal{X}, \ \forall \boldsymbol{\theta}_{\mathrm{de}} \in \Theta_{\mathrm{de}},
\end{aligned}$$

are prepared. Then the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{nu}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{de}}$ are computed separately from $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^n$ and $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^n$:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} &:= \underset{\boldsymbol{\theta}_{\mathrm{nu}} \in \Theta_{\mathrm{nu}}}{\mathrm{argmax}} \left[\sum_{i=1}^n \log p_{\mathrm{nu}}(\boldsymbol{x}^{\mathrm{nu}}_i; \boldsymbol{\theta}_{\mathrm{nu}}) \right], \\
\widehat{\boldsymbol{\theta}}_{\mathrm{de}} &:= \underset{\boldsymbol{\theta}_{\mathrm{de}} \in \Theta_{\mathrm{de}}}{\mathrm{argmax}} \left[\sum_{j=1}^n \log p_{\mathrm{de}}(\boldsymbol{x}^{\mathrm{de}}_j; \boldsymbol{\theta}_{\mathrm{de}}) \right].
\end{aligned}$$

Note that the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{nu}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{de}}$ minimize the empirical Kullback-Leibler divergences from the true densities $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ to their models $p_{\mathrm{nu}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{nu}})$ and $p_{\mathrm{de}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{de}})$, respectively:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} &= \underset{\boldsymbol{\theta}_{\mathrm{nu}} \in \Theta_{\mathrm{nu}}}{\mathrm{argmin}} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x}^{\mathrm{nu}}_i)}{p_{\mathrm{nu}}(\boldsymbol{x}^{\mathrm{nu}}_i; \boldsymbol{\theta}_{\mathrm{nu}})} \right], \\
\widehat{\boldsymbol{\theta}}_{\mathrm{de}} &= \underset{\boldsymbol{\theta}_{\mathrm{de}} \in \Theta_{\mathrm{de}}}{\mathrm{argmin}} \left[\frac{1}{n} \sum_{j=1}^n \log \frac{p^*_{\mathrm{de}}(\boldsymbol{x}^{\mathrm{de}}_j)}{p_{\mathrm{nu}}(\boldsymbol{x}^{\mathrm{de}}_j; \boldsymbol{\theta}_{\mathrm{de}})} \right].
\end{aligned}$$

Finally, a density ratio estimator is constructed by taking the ratio of the estimated densities:

$$\widehat{r}_{\mathrm{A}}(\boldsymbol{x}) := \frac{p_{\mathrm{nu}}(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_{\mathrm{nu}})}{p_{\mathrm{de}}(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_{\mathrm{de}})} \left(\frac{1}{n} \sum_{j=1}^{n} \frac{p_{\mathrm{nu}}(\boldsymbol{x}_j^{\mathrm{de}}; \widehat{\boldsymbol{\theta}}_{\mathrm{nu}})}{p_{\mathrm{de}}(\boldsymbol{x}_j^{\mathrm{de}}; \widehat{\boldsymbol{\theta}}_{\mathrm{de}})} \right)^{-1},$$

where the estimator is normalized so that

$$\frac{1}{n} \sum_{j=1}^{n} \widehat{r}_{\mathrm{A}}(\boldsymbol{x}_j^{\mathrm{de}}) = 1.$$

2.4 Method (B): Density Ratio Estimation by Logistic Regression

Let us assign a selector variable $y =$ 'nu' to samples drawn from $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $y =$ 'de' to samples drawn from $p^*_{\mathrm{de}}(\boldsymbol{x})$, i.e., the two densities are written as

$$\begin{aligned} p^*_{\mathrm{nu}}(\boldsymbol{x}) &= q^*(\boldsymbol{x} | y = \text{`nu'}), \\ p^*_{\mathrm{de}}(\boldsymbol{x}) &= q^*(\boldsymbol{x} | y = \text{`de'}). \end{aligned}$$

Since

$$\begin{aligned} q^*(\boldsymbol{x} | y = \text{`nu'}) &= \frac{q^*(y = \text{`nu'} | \boldsymbol{x}) q^*(\boldsymbol{x})}{q^*(y = \text{`nu'})}, \\ q^*(\boldsymbol{x} | y = \text{`de'}) &= \frac{q^*(y = \text{`de'} | \boldsymbol{x}) q^*(\boldsymbol{x})}{q^*(y = \text{`de'})}, \end{aligned}$$

the density ratio can be expressed in terms of y as

$$\begin{aligned} r^*(\boldsymbol{x}) &= \frac{q^*(y = \text{`nu'} | \boldsymbol{x})}{q^*(y = \text{`nu'})} \frac{q^*(y = \text{`de'})}{q^*(y = \text{`de'} | \boldsymbol{x})} \\ &= \frac{q^*(y = \text{`nu'} | \boldsymbol{x})}{q^*(y = \text{`de'} | \boldsymbol{x})}, \end{aligned}$$

where we used the fact that

$$q^*(y = \text{`nu'}) = q^*(y = \text{`de'}) = \frac{1}{2}$$

in the current setup.

The conditional probability $q^*(y|\boldsymbol{x})$ could be approximated by discriminating $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^n$ from $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^n$ using a *logistic regression* classifier, i.e., for a parametric function $r(\boldsymbol{x}; \boldsymbol{\theta})$ such that

$$r(\boldsymbol{x}; \boldsymbol{\theta}) \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \ \forall \boldsymbol{\theta} \in \Theta, \tag{4}$$

the conditional probabilities $q^*(y = \text{`nu'}|\boldsymbol{x})$ and $q^*(y = \text{`de'}|\boldsymbol{x})$ are modeled by

$$\begin{aligned} q(y = \text{`nu'}|\boldsymbol{x};\boldsymbol{\theta}) &= \frac{r(\boldsymbol{x};\boldsymbol{\theta})}{1 + r(\boldsymbol{x};\boldsymbol{\theta})}, \\ q(y = \text{`de'}|\boldsymbol{x};\boldsymbol{\theta}) &= \frac{1}{1 + r(\boldsymbol{x};\boldsymbol{\theta})}. \end{aligned}$$

Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{B}}$ is computed from $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^n$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^n$:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{B}} := \mathop{\mathrm{argmax}}_{\boldsymbol{\theta}\in\Theta} \left[\sum_{i=1}^n \log \frac{r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta})}{1 + r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta})} + \sum_{j=1}^n \log \frac{1}{1 + r(\boldsymbol{x}_j^{\mathrm{de}};\boldsymbol{\theta})} \right]. \tag{5}$$

Note that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{B}}$ minimizes the empirical Kullback-Leibler divergences from the true density $q^*(\boldsymbol{x}, y)$ to its estimator $q(y|\boldsymbol{x};\boldsymbol{\theta})q^*(\boldsymbol{x})$:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\mathrm{B}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta} \Bigg[& \frac{1}{2n} \sum_{i=1}^n \log \frac{q^*(\boldsymbol{x}_i^{\mathrm{nu}}, y = \text{`nu'})}{q(y = \text{`nu'}|\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta})q^*(\boldsymbol{x}_i^{\mathrm{nu}})} \\ & + \frac{1}{2n} \sum_{j=1}^n \log \frac{q^*(\boldsymbol{x}_j^{\mathrm{de}}, y = \text{`de'})}{q(y = \text{`de'}|\boldsymbol{x}_j^{\mathrm{de}};\boldsymbol{\theta})q^*(\boldsymbol{x}_j^{\mathrm{de}})} \Bigg]. \end{aligned}$$

Finally, a density ratio estimator is constructed by taking the ratio of $q(y = \text{`nu'}|\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})$ and $q(y = \text{`de'}|\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})$ with proper normalization:

$$\begin{aligned} \widehat{r}_{\mathrm{B}}(\boldsymbol{x}) &:= \frac{q(y = \text{`nu'}|\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})}{q(y = \text{`de'}|\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})} \left(\frac{1}{n} \sum_{j=1}^n \frac{q(y = \text{`nu'}|\boldsymbol{x}_j^{\mathrm{de}};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})}{q(y = \text{`de'}|\boldsymbol{x}_j^{\mathrm{de}};\widehat{\boldsymbol{\theta}}_{\mathrm{B}})} \right)^{-1} \\ &= r(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{B}}) \left(\frac{1}{n} \sum_{j=1}^n r(\boldsymbol{x}_j^{\mathrm{de}};\widehat{\boldsymbol{\theta}}_{\mathrm{B}}) \right)^{-1}. \end{aligned}$$

2.5 Method (C): Direct Density Ratio Estimation by Empirical Unnormalized Kullback-Leibler Divergence Minimization

For the density ratio function $r^*(\boldsymbol{x})$, a parametric model $r(\boldsymbol{x};\boldsymbol{\theta})$ such that Eq.(4) is fulfilled is prepared. Then the following estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ is computed from $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^n$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^n$:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{C}} := \mathop{\mathrm{argmax}}_{\boldsymbol{\theta}\in\Theta} \left[\sum_{i=1}^n \log r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta}) - \sum_{j=1}^n r(\boldsymbol{x}_j^{\mathrm{de}};\boldsymbol{\theta}) \right]. \tag{6}$$

Note that $\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ minimizes the empirical unnormalized Kullback-Leibler divergence from the true density $p_{\mathrm{nu}}^*(\boldsymbol{x})$ to its estimator $\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})$:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{C}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{p_{\mathrm{nu}}^*(\boldsymbol{x}_i^{\mathrm{nu}})}{\widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})p_{\mathrm{de}}^*(\boldsymbol{x}_i^{\mathrm{nu}})} - 1 + \frac{1}{n} \sum_{j=1}^n \widehat{r}(\boldsymbol{x}_j^{\mathrm{de}}) \right].$$

Finally, a density ratio estimator is obtained by

$$\widehat{r}_{\mathrm{C}}(\boldsymbol{x}) := r(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_{\mathrm{C}}) \left(\frac{1}{n} \sum_{j=1}^{n} r(\boldsymbol{x}_j^{\mathrm{de}}; \widehat{\boldsymbol{\theta}}_{\mathrm{C}}) \right)^{-1} .$$

3 Exponential Models

In our theoretical analysis, we employ the *exponential model*, which is explained in this section.

3.1 Exponential Models for Densities and Ratios

We use the following exponential model for the densities $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$.

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}) - \varphi(\boldsymbol{\theta}) \right\}, \quad \theta \in \Theta, \tag{7}$$

where $h(\boldsymbol{x})$ is a *base measure*, $\boldsymbol{\xi}(\boldsymbol{x})$ is a *sufficient statistic*, $\varphi(\boldsymbol{\theta})$ is a *normalization factor*, and $^\top$ denotes the transpose of a vector [13]. The exponential model includes various popular models as special cases, e.g., the normal, exponential, gamma, chi-square, and beta distributions.

Correspondingly, we use the following exponential model for the ratio $r^*(\boldsymbol{x})$.

$$r(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0) = \exp\left\{ \theta_0 + \boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}) \right\}, \quad \theta \in \Theta, \quad \theta_0 \in \mathbb{R}. \tag{8}$$

3.2 Method (A)

For the exponential model (7), the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{nu}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{de}}$ are given by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} = \underset{\boldsymbol{\theta}_{\mathrm{nu}} \in \Theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) - n\varphi(\boldsymbol{\theta}) \right],$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{de}} = \underset{\boldsymbol{\theta}_{\mathrm{de}} \in \Theta}{\operatorname{argmax}} \left[\sum_{j=1}^{n} \boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) - n\varphi(\boldsymbol{\theta}) \right],$$

where irrelevant constants are ignored. The density ratio estimator $\widehat{r}_{\mathrm{A}}(\boldsymbol{x})$ for the exponential density model is expressed as

$$\widehat{r}_{\mathrm{A}}(\boldsymbol{x}) = \exp\left\{ \widehat{\boldsymbol{\theta}}_{\mathrm{A}}^\top \boldsymbol{\xi}(\boldsymbol{x}) \right\} \left(\frac{1}{n} \sum_{j=1}^{n} \exp\left\{ \widehat{\boldsymbol{\theta}}_{\mathrm{A}}^\top \boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) \right\} \right)^{-1},$$

where

$$\widehat{\boldsymbol{\theta}}_{\mathrm{A}} := \widehat{\boldsymbol{\theta}}_{\mathrm{nu}} - \widehat{\boldsymbol{\theta}}_{\mathrm{de}}.$$

One can use the other estimator such as

$$\widetilde{r}_{\mathrm{A}}(\boldsymbol{x}) = \exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{A}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}) - \varphi(\widehat{\boldsymbol{\theta}}_{\mathrm{nu}}) + \varphi(\widehat{\boldsymbol{\theta}}_{\mathrm{de}})\right\}$$

instead of $\widehat{r}_{\mathrm{A}}(\boldsymbol{x})$. We compare $\widetilde{r}_{\mathrm{A}}(\boldsymbol{x})$ to Method (B) and Method (C), since the same normalization factor as $\widetilde{r}_{\mathrm{A}}(\boldsymbol{x})$ appears in the other methods as shown below. This fact facilitates the theoretical analysis.

3.3 Method (B)

For the exponential model (8), the optimization problem (5) is expressed as

$$\begin{aligned}(\widehat{\boldsymbol{\theta}}_{\mathrm{B}}, \widehat{\theta}_{\mathrm{B},0}) &= \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmax}} \left[\sum_{i=1}^{n}\log\frac{r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta},\theta_0)}{1+r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta},\theta_0)} + \sum_{j=1}^{n}\log\frac{1}{1+r(\boldsymbol{x}_j^{\mathrm{de}};\boldsymbol{\theta},\theta_0)}\right] \\ &= \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmax}} \left[\sum_{i=1}^{n}\log\frac{\exp\left\{\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}})\right\}}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}})\right\}}\right. \\ &\qquad\qquad\left. + \sum_{j=1}^{n}\log\frac{1}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}}\right].\end{aligned}$$

The density ratio estimator $\widehat{r}_{\mathrm{B}}(\boldsymbol{x})$ for the exponential ratio model is expressed as

$$\widehat{r}_{\mathrm{B}}(\boldsymbol{x}) = \exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{B}}^{\top}\boldsymbol{\xi}(\boldsymbol{x})\right\}\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{B}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\right)^{-1}.$$

3.4 Method (C)

For the exponential model (8), the optimization problem (6) is expressed as

$$\begin{aligned}(\widehat{\boldsymbol{\theta}}_{\mathrm{C}}, \widehat{\theta}_{\mathrm{C},0}) &= \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmax}} \left[\frac{1}{n}\sum_{i=1}^{n}\log r(\boldsymbol{x}_i^{\mathrm{nu}};\boldsymbol{\theta},\theta_0) - \frac{1}{n}\sum_{j=1}^{n} r(\boldsymbol{x}_i^{\mathrm{de}};\boldsymbol{\theta},\theta_0)\right] \\ &= \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmax}} \left[\frac{1}{n}\sum_{i=1}^{n}(\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}})) - \frac{1}{n}\sum_{j=1}^{n}\exp\left\{\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\right]. \qquad (9)\end{aligned}$$

The density ratio estimator $\widehat{r}_{\mathrm{C}}(\boldsymbol{x})$ for the exponential ratio model is expressed as

$$\widehat{r}_{\mathrm{C}}(\boldsymbol{x}) = \exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x})\right\}\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\right)^{-1}.$$

4 Accuracy Analysis for Correctly Specified Exponential Models

In this section, we theoretically analyze the accuracy of the above three density ratio estimators under the assumption that the true densities $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ both belong to the exponential family, i.e., there exist $\boldsymbol{\theta}^*_{\mathrm{nu}} \in \Theta$ and $\boldsymbol{\theta}^*_{\mathrm{de}} \in \Theta$ such that

$$p^*_{\mathrm{nu}}(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}^*_{\mathrm{nu}}),$$
$$p^*_{\mathrm{de}}(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}^*_{\mathrm{de}}).$$

Since the ratio of two exponential densities also belongs to the exponential model, the above assumption implies that there exist $\boldsymbol{\theta}^* \in \Theta$ and $\theta^*_0 \in \mathbb{R}$ such that

$$r^*(\boldsymbol{x}) = r(\boldsymbol{x}; \boldsymbol{\theta}^*, \theta^*_0). \tag{10}$$

It is straightforward to extended the results in this section to general parametric models, since we focus on the first-order asymptotics of the estimators. An arbitrary parametric model $p(\boldsymbol{x}; \boldsymbol{\theta})$ has the same first-order asymptotics as the exponential model of the form

$$p_{\mathrm{exp}}(\boldsymbol{x}; \boldsymbol{\theta}) \propto \exp\{\log p(\boldsymbol{x}; \boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \log p(\boldsymbol{x}; \boldsymbol{\theta}^*)\}$$

around the parameter $\boldsymbol{\theta}^*$. Thus the same theoretical property holds.

First, we analyze the asymptotic behavior of $J(\widehat{r}_{\mathrm{A}})$. Then we have the following lemma (proofs of all lemmas, theorems, and corollaries are provided in Appendix).

Lemma 1 $J(\widehat{r}_{\mathrm{A}})$ *can be asymptotically expressed as*

$$J(\widehat{r}_{\mathrm{A}}) = \frac{1}{2n}\left[\dim\Theta + \mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{de}})^{-1}\right) + \mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}})\right] + \mathcal{O}(n^{-3/2}),$$

where $\mathcal{O}(\cdot)$ denotes the asymptotic order. $\boldsymbol{F}(\boldsymbol{\theta})$ denote the Fisher information matrix of the exponential model $p(\boldsymbol{x}; \boldsymbol{\theta})$:

$$\boldsymbol{F}(\boldsymbol{\theta}) := \int \nabla \log p(\boldsymbol{x}; \boldsymbol{\theta}) \nabla \log p(\boldsymbol{x}; \boldsymbol{\theta})^\top p(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x},$$

where ∇ denotes the partial differential operator with respect to $\boldsymbol{\theta}$. $\mathrm{PE}(p\|q)$ *denotes the Pearson divergence of two densities p and q defined as*

$$\mathrm{PE}(p\|q) := \frac{1}{2}\int \frac{(p(\boldsymbol{x}) - q(\boldsymbol{x}))^2}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}. \tag{11}$$

Next, we investigate the asymptotic behavior of $J(\widehat{r}_{\mathrm{B}})$ and $J(\widehat{r}_{\mathrm{C}})$. Let y be the selector variable taking 'nu' or 'de' as defined in Section 2.4. The statistical model of the joint probability for $\boldsymbol{z} = (\boldsymbol{x}, y)$ is defined as

$$q(\boldsymbol{z}; \boldsymbol{\theta}, \theta_0) = q(y|\boldsymbol{x}; \boldsymbol{\theta}, \theta_0) \times \frac{p^*_{\mathrm{nu}}(\boldsymbol{x}) + p^*_{\mathrm{de}}(\boldsymbol{x})}{2}, \tag{12}$$

where $q(y|\boldsymbol{x};\boldsymbol{\theta},\theta_0)$ is the conditional probability of y such that

$$\begin{aligned} q(y=\text{`nu'}|\boldsymbol{x};\boldsymbol{\theta},\theta_0) &= \frac{r(\boldsymbol{x};\boldsymbol{\theta},\theta_0)}{1+r(\boldsymbol{x};\boldsymbol{\theta},\theta_0)} \\ &= \frac{\exp\left\{\theta_0+\boldsymbol{\theta}^\top\boldsymbol{\xi}(\boldsymbol{x})\right\}}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^\top\boldsymbol{\xi}(\boldsymbol{x})\right\}}, \\ q(y=\text{`de'}|\boldsymbol{x};\boldsymbol{\theta},\theta_0) &= \frac{1}{1+r(\boldsymbol{x};\boldsymbol{\theta},\theta_0)} \\ &= \frac{1}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^\top\boldsymbol{\xi}(\boldsymbol{x})\right\}}. \end{aligned}$$

The Fisher information matrix of the model (12) is denoted as

$$\widetilde{\boldsymbol{F}}(\boldsymbol{\theta},\theta_0)\in\mathbb{R}^{(\dim\Theta+1)\times(\dim\Theta+1)}.$$

The submatrix of $\widetilde{\boldsymbol{F}}(\boldsymbol{\theta},\theta_0)$ formed by the first $(\dim\Theta)$ rows and the first $(\dim\Theta)$ columns is defined as

$$\int\nabla\log q(\boldsymbol{z};\boldsymbol{\theta},\theta_0)\nabla\log q(\boldsymbol{z};\boldsymbol{\theta},\theta_0)^\top q(\boldsymbol{z};\boldsymbol{\theta},\theta_0)\mathrm{d}\boldsymbol{z}.$$

The inverse matrix of $\widetilde{\boldsymbol{F}}(\boldsymbol{\theta},\theta_0)$ is expressed as

$$\widetilde{\boldsymbol{F}}(\boldsymbol{\theta},\theta_0)^{-1}=\begin{pmatrix}\boldsymbol{H}_{11}(\boldsymbol{\theta},\theta_0) & \boldsymbol{h}_{12}(\boldsymbol{\theta},\theta_0)\\ \boldsymbol{h}_{12}(\boldsymbol{\theta},\theta_0)^\top & h_{22}(\boldsymbol{\theta},\theta_0)\end{pmatrix}, \tag{13}$$

where $\boldsymbol{H}_{11}(\boldsymbol{\theta},\theta_0)$ is a $(\dim\Theta)\times(\dim\Theta)$ matrix. Then we have the following lemmas.

Lemma 2 *$J(\widehat{r}_{\mathrm{B}})$ can be asymptotically expressed as*

$$J(\widehat{r}_{\mathrm{B}})=\frac{1}{2n}\left[\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})\boldsymbol{H}_{11}(\boldsymbol{\theta}^*,\theta_0^*)\right)+\mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}})\right]+\mathcal{O}(n^{-3/2}),$$

where $(\boldsymbol{\theta}^*,\theta_0^*)$ is defined in Eq.(10).

Lemma 3 *$J(\widehat{r}_{\mathrm{C}})$ can be asymptotically expressed as*

$$J(\widehat{r}_{\mathrm{C}})=\frac{1}{2n}\left[\dim\Theta+\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})^{-1}\boldsymbol{G}\right)+\mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}})\right]+\mathcal{O}(n^{-3/2}),$$

where

$$\boldsymbol{G}:=\int r^*(\boldsymbol{x})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})^\top p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Based on the above lemmas, we compare the accuracy of the three methods. For the accuracy of (A) and (B), we have the following theorem.

Theorem 4 *Asymptotically, the inequality*

$$J(\widehat{r}_{\mathrm{A}}) \leq J(\widehat{r}_{\mathrm{B}})$$

holds.

Thus the method (A) is more accurate than the method (B) in terms of the expected unnormalized Kullback-Leibler divergence (3). Theorem 4 may be regarded as an extension of the result for binary classification [5]: estimating data generating Gaussian densities by maximum likelihood estimation has high statistical efficiency than logistic regression in the sense of classification error rate.

Next, we compare the accuracy of (B) and (C).

Theorem 5 *Asymptotically, the inequality*

$$J(\widehat{r}_{\mathrm{B}}) \leq J(\widehat{r}_{\mathrm{C}})$$

holds.

Thus the method (B) is more accurate than the method (C) in terms of the expected unnormalized Kullback-Leibler divergence (3). This inequality is a direct consequence of the paper by Qin [16]. In that paper, it was shown that the method (B) has the smallest asymptotic variance in a class of semi-parametric estimators. It is easy to see the method (C) is included in the class.

Finally, we compare the accuracy of (A) and (C). From Theorem 4 and Theorem 5, we immediately have the following corollary.

Corollary 6 *The inequality*

$$J(\widehat{r}_{\mathrm{A}}) \leq J(\widehat{r}_{\mathrm{C}})$$

holds.

It was advocated that one should avoid solving more difficult intermediate problems when solving a target problem [33]. This statement is sometimes referred to as “Vapnik's principle”, and the *support vector machine* [4] would be a successful example of this principle—instead of estimating a data generation model, it directly models the decision boundary which is sufficient for pattern recognition.

If we followed Vapnik's principle, directly estimating the ratio $r^*(\boldsymbol{x})$ would be more promising than estimating the two densities $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ since knowing $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ implies knowing $r^*(\boldsymbol{x})$ but not vice versa; indeed, $r^*(\boldsymbol{x})$ cannot be uniquely decomposed into $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$. Thus Corollary 6 is at a glance counter-intuitive. However, Corollary 6 would be reasonable since the method (C) does not make use of the knowledge that *each* density is exponential, but only the knowledge that the ratio is exponential. Thus the method (A) can utilize the a priori model information more effectively. Thanks to the additional knowledge that the both densities belong to the exponential model, the intermediate problems (i.e., density estimation) were actually made easier in terms of Vapnik's principle.

5 Accuracy Analysis for Misspecified Exponential Models

In this section, we theoretically analyze the approximation error of the three density ratio estimators for misspecified exponential models, i.e., the true densities and ratio are not necessarily included in the exponential models. The unnormalized Kullback-Leibler divergence is employed to measure the approximation error.

First, we study the convergence of the method (A). Let $\overline{p}_{\mathrm{nu}}(\boldsymbol{x})$ and $\overline{p}_{\mathrm{de}}(\boldsymbol{x})$ be the projections of the true densities $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ onto the model $p(\boldsymbol{x};\boldsymbol{\theta})$ in terms of the Kullback-Leibler divergence (2):

$$\begin{aligned}\overline{p}_{\mathrm{nu}}(\boldsymbol{x}) &:= p(\boldsymbol{x};\overline{\boldsymbol{\theta}}_{\mathrm{nu}}),\\ \overline{p}_{\mathrm{de}}(\boldsymbol{x}) &:= p(\boldsymbol{x};\overline{\boldsymbol{\theta}}_{\mathrm{de}}),\end{aligned}$$

where

$$\begin{aligned}\overline{\boldsymbol{\theta}}_{\mathrm{nu}} &:= \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta}\left[\int p^*_{\mathrm{nu}}(\boldsymbol{x})\log\frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p(\boldsymbol{x};\boldsymbol{\theta})}\mathrm{d}\boldsymbol{x}\right],\\ \overline{\boldsymbol{\theta}}_{\mathrm{de}} &:= \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta}\left[\int p^*_{\mathrm{de}}(\boldsymbol{x})\log\frac{p^*_{\mathrm{de}}(\boldsymbol{x})}{p(\boldsymbol{x};\boldsymbol{\theta})}\mathrm{d}\boldsymbol{x}\right].\end{aligned}$$

This means that $\overline{p}_{\mathrm{nu}}(\boldsymbol{x})$ and $\overline{p}_{\mathrm{de}}(\boldsymbol{x})$ are the optimal approximations to $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ in the model $p(\boldsymbol{x};\boldsymbol{\theta})$ in terms of the Kullback-Leibler divergence. Let

$$\overline{r}_{\mathrm{A}}(\boldsymbol{x}) := \frac{\overline{p}_{\mathrm{nu}}(\boldsymbol{x})}{\overline{p}_{\mathrm{de}}(\boldsymbol{x})}.$$

Since the ratio of two exponential densities also belongs to the exponential model, there exists $\overline{\boldsymbol{\theta}}_{\mathrm{A}}\in\Theta$ such that

$$\overline{r}_{\mathrm{A}}(\boldsymbol{x}) = r(\boldsymbol{x};\overline{\boldsymbol{\theta}}_{\mathrm{A}},\bar{\theta}_{\mathrm{A},0}).$$

Then we have the following lemma.

Lemma 7 $\widehat{r}_{\mathrm{A}}$ *converges in probability to* $\overline{r}_{\mathrm{A}}$ *as* $n\to\infty$.

Next, we investigate the convergence of the method (B). Let $q^*(\boldsymbol{x},y)$ be the joint probability defined as

$$q^*(\boldsymbol{x},y) = q^*(y|\boldsymbol{x})\times\frac{p^*_{\mathrm{nu}}(\boldsymbol{x})+p^*_{\mathrm{de}}(\boldsymbol{x})}{2}, \tag{14}$$

where $q^*(y|\boldsymbol{x})$ is the conditional probability of y such that

$$\begin{aligned}q^*(y=\text{`nu'}|\boldsymbol{x}) &= \frac{r^*(\boldsymbol{x})}{1+r^*(\boldsymbol{x})},\\ q^*(y=\text{`de'}|\boldsymbol{x}) &= \frac{1}{1+r^*(\boldsymbol{x})}.\end{aligned}$$

The model (12) is used to estimate $q^*(\boldsymbol{x}, y)$, and let $\overline{q}(\boldsymbol{x}, y)$ be the projection of the true density $q^*(\boldsymbol{x}, y)$ onto the model (12) in terms of the Kullback-Leibler divergence (2):

$$\overline{q}(\boldsymbol{x}, y) := q(\boldsymbol{x}, y; \overline{\boldsymbol{\theta}}_{\mathrm{B}}, \bar{\theta}_{\mathrm{B},0}), \tag{15}$$

where

$$(\overline{\boldsymbol{\theta}}_{\mathrm{B}}, \bar{\theta}_{\mathrm{B},0}) := \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmin}} \left[\int \sum_{y\in\{\text{`nu'},\text{`de'}\}} q^*(\boldsymbol{x}, y) \log \frac{q^*(y|\boldsymbol{x})}{q(y|\boldsymbol{x}; \boldsymbol{\theta}, \theta_0)} \mathrm{d}\boldsymbol{x} \right].$$

This means that $\overline{q}(\boldsymbol{x}, y)$ is the optimal approximation to $q^*(\boldsymbol{x}, y)$ in the model

$$q(y|\boldsymbol{x}; \boldsymbol{\theta}, \theta_0) \frac{p^*_{\mathrm{nu}}(\boldsymbol{x}) + p^*_{\mathrm{de}}(\boldsymbol{x})}{2}$$

in terms of the Kullback-Leibler divergence. Let

$$\overline{r}_{\mathrm{B}}(\boldsymbol{x}) := r(\boldsymbol{x}; \overline{\boldsymbol{\theta}}_{\mathrm{B}}, \bar{\theta}_{\mathrm{B},0}).$$

Then we have the following lemma.

Lemma 8 $\widehat{r}_{\mathrm{B}}$ *converges in probability to* $\overline{r}_{\mathrm{B}}$ *as* $n \to \infty$.

Finally, we study the convergence of the method (C). Suppose that the model $r(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0)$ in Eq.(8) is employed. Let $\overline{r}_{\mathrm{C}}(\boldsymbol{x})$ be the projection of the true ratio function $r^*(\boldsymbol{x})$ onto the model $r(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0)$ in terms of the unnormalized Kullback-Leibler divergence (1):

$$\overline{r}_{\mathrm{C}}(\boldsymbol{x}) := r(\boldsymbol{x}; \overline{\boldsymbol{\theta}}_{\mathrm{C}}, \bar{\theta}_{\mathrm{C},0}),$$

where

$$(\overline{\boldsymbol{\theta}}_{\mathrm{C}}, \bar{\theta}_{\mathrm{C},0}) := \underset{(\boldsymbol{\theta},\theta_0)\in\Theta\times\mathbb{R}}{\mathrm{argmin}} \left[\int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{r^*(\boldsymbol{x})}{r(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0)} \mathrm{d}\boldsymbol{x} - 1 + \int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0) \mathrm{d}\boldsymbol{x} \right]. \tag{16}$$

This means that $\overline{r}_{\mathrm{C}}(\boldsymbol{x})$ is the optimal approximation to $r^*(\boldsymbol{x})$ in the model $r(\boldsymbol{x}; \boldsymbol{\theta})$ in terms of the unnormalized Kullback-Leibler divergence. Then we have the following lemma.

Lemma 9 $\widehat{r}_{\mathrm{C}}$ *converges in probability to* $\overline{r}_{\mathrm{C}}$ *as* $n \to \infty$.

Based on the above lemmas, we investigate the relation among the three methods. Lemma 9 implies that the method (C) is consistent to the optimal approximation $\overline{r}_{\mathrm{C}}$. However, as we will show below, the methods (A) and (B) are not consistent to the optimal approximation $\overline{r}_{\mathrm{C}}$ in general. Let us measure the deviation of a density ratio function $\overline{r}'$ from $\overline{r}$ by

$$\mathrm{D}(\overline{r}', \overline{r}) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) \left(\overline{r}'(\boldsymbol{x}) - \overline{r}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}.$$

Then we have the following theorem.

Theorem 10 *The inequalities*

$$\mathrm{D}(\overline{r}_{\mathrm{A}}, \overline{r}_{\mathrm{C}}) \geq \left| \int p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}_{\mathrm{A}}(\boldsymbol{x}) d\boldsymbol{x} - 1 \right|^2,$$
$$\mathrm{D}(\overline{r}_{\mathrm{B}}, \overline{r}_{\mathrm{C}}) \geq \left| \int p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}_{\mathrm{B}}(\boldsymbol{x}) d\boldsymbol{x} - 1 \right|^2$$

hold. More generally, for any $\overline{r}$ in the exponential model,

$$\mathrm{D}(\overline{r}, \overline{r}_{\mathrm{C}}) \geq \left| \int p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}(\boldsymbol{x}) d\boldsymbol{x} - 1 \right|^2 \tag{17}$$

holds.

When the model is misspecified, $p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}_{\mathrm{A}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}_{\mathrm{B}}(\boldsymbol{x})$ are not probability densities in general. Then Theorem 10 implies that the method (A) and the method (B) are not consistent to the optimal approximation $\overline{r}_{\mathrm{C}}$.

Since model misspecification would be a usual situation in practice, the method (C) is the most promising approach in density ratio estimation.

Finally, for the consistency of the method (A), we also have the following additional result.

Corollary 11 *If $p^*_{\mathrm{de}}(\boldsymbol{x})$ belongs to the exponential model* (7)*, i.e., there exists $\overline{\boldsymbol{\theta}}_{\mathrm{de}} \in \Theta$ such that*

$$p^*_{\mathrm{de}}(\boldsymbol{x}) = p(\boldsymbol{x}; \overline{\boldsymbol{\theta}}_{\mathrm{de}}),$$

then

$$\overline{r}_{\mathrm{A}} = \overline{r}_{\mathrm{C}}$$

holds even when $p^*_{\mathrm{nu}}(\boldsymbol{x})$ does not belong to the exponential model (7)*.*

This corollary means that, as long as $p^*_{\mathrm{de}}(\boldsymbol{x})$ is correctly specified, the method (A) is still consistent.

6 Conclusions

In this paper, we theoretically investigated the accuracy of three density ratio estimation approaches: (A) density ratio estimation by separate maximum likelihood density estimation, (B) density ratio estimation by logistic regression, and (C) direct density ratio estimation by empirical Kullback-Leibler divergence minimization. Intuitively, the method (C) seems to be better than the other approaches due to "Vapnik's principle"—one should not solve more difficult intermediate problems (density estimation in the current context) when solving a target problem (density ratio estimation in the current context).

However, as we proved in Section 4, the method (A) is more accurate than the other approaches when the numerator and denominator densities are known to be members of the exponential family. This result is at first sight counter-intuitive, but it would be reasonable because the methods (B) and (C) do not make use of the knowledge that *each* density is exponential, but only the knowledge that their ratio is exponential. Thus the method (A) can utilize the a priori model information more effectively than the other methods. We note that this result is not contradictory to Vapnik's principle since the additional knowledge that the densities belong to the exponential model is utilized to make the intermediate problems (density estimation) substantially easier.

On the other hand, once the correct model assumption is not fulfilled, the method (C) was shown to be consistent to the optimal approximation in the model, while the methods (A) and (B) are not consistent in general (see Section 5). The fact that the direct method outperforms the other approaches in the absence of the additional knowledge would follow Vapnik's principle.

It seems to be a common phenomenon in various situations that a method which works optimally for correctly specified models performs poorly for misspecified models and conversely a method which works well for misspecified models performs poorly for correctly specified models. For example, in active learning (or the experiment design), the traditional variance-only approach works optimally for correctly specified models [6]. However, it was shown that the traditional method works poorly once the correct model assumption is slightly violated [20]. To cope with this problem, various active learning methods which do not require the correct model assumption have been developed and shown to work better than the traditional method for misspecified models [34, 11, 20, 9, 24]. However, these methods cannot outperform the traditional method when the model is correctly specified. Thus the performance loss for correctly specified models would be the price one has to pay for acquiring robustness against model misspecification.

Model misspecification would almost always occur in practice, so developing methods for misspecified models is crucial. Based on these observations, we conclude that the direct density ratio approach (C) is the most promising density ratio estimation method.

Acknowledgments

MS was supported by SCAT, AOARD, and the JST PRESTO program.

A Asymptotic Expansion of Measure of Accuracy

First, we show some fundamental results used for proving Lemma 1, Lemma 2, and Lemma 3.

Using the Taylor expansion

$$\log(1+t) = t - \frac{t^2}{2} + \mathcal{O}(t^3),$$

we have the following expansion:

$$\begin{aligned}\log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{\widehat{r}(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})} &= \log \frac{r^*(\boldsymbol{x})}{\widehat{r}(\boldsymbol{x})} \\ &= -\log \frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} \\ &= -\left(\frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} - 1\right) + \frac{1}{2}\left(\frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} - 1\right)^2 + \mathcal{O}_p\left(\left|\frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} - 1\right|^3\right),\end{aligned}$$

where $\mathcal{O}_p(\cdot)$ denotes the stochastic order. Substituting this expansion into the unnormalized Kullback-Leibler divergence $\mathrm{UKL}(p^*_{\mathrm{nu}}\|\widehat{r}\cdot p^*_{\mathrm{de}})$, we obtain

$$\mathrm{UKL}(p^*_{\mathrm{nu}}\|\widehat{r}\cdot p^*_{\mathrm{de}}) = \mathrm{PE}(p^*_{\mathrm{nu}}\|\widehat{r}\cdot p^*_{\mathrm{de}}) + \mathcal{O}(\|\widehat{r}/r^* - 1\|^3), \tag{18}$$

where 'PE' denotes the Pearson divergence defined by Eq.(11) and $\|\widehat{r}/r^* - 1\|$ is defined as

$$\|\widehat{r}/r^* - 1\| := \left(\int p^*_{\mathrm{nu}}(\boldsymbol{x})|\widehat{r}(\boldsymbol{x})/r^*(\boldsymbol{x}) - 1|^2 \mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

Under a regularity condition of asymptotic statistics, the expectation $\mathbb{E}\left[\|\widehat{r}/r^* - 1\|^3\right]$ is of order $\mathcal{O}(n^{-3/2})$:

$$\mathbb{E}\left[\|\widehat{r}/r^* - 1\|^3\right] = \mathcal{O}(n^{-3/2}).$$

See Theorem 5.23 in [32] for the details of the regularity condition on general M-estimators. Hence, the measure of accuracy $J(\widehat{r})$ can be represented as

$$J(\widehat{r}) = \mathbb{E}\left[\mathrm{PE}(p^*_{\mathrm{nu}}\|\widehat{r}\cdot p^*_{\mathrm{de}})\right] + \mathcal{O}(n^{-3/2}). \tag{19}$$

Then we have the following lemma.

Lemma 12 (Asymptotics of measure of accuracy) *Let $\widehat{\boldsymbol{\theta}}$ be an estimator of the parameter $\boldsymbol{\theta}^*$ in r^*, and $\widehat{r}(\boldsymbol{x})$ be the estimator defined as*

$$\widehat{r}(\boldsymbol{x}) := \exp\left\{\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}(\boldsymbol{x})\right\}\left(\frac{1}{n}\sum_{j=1}^{n} \exp\left\{\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}(\boldsymbol{x}^{\mathrm{de}}_j)\right\}\right)^{-1}.$$

Then, the measure of accuracy of $\widehat{r}$ is asymptotically given as

$$J(\widehat{r}) = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}}) \cdot \mathbb{E}\left[\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^\top\right]\right) + \frac{1}{2n}\mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}}) + \mathcal{O}(n^{-3/2}), \tag{20}$$

where $\delta\boldsymbol{\theta}$ denotes the deviation of $\widehat{\boldsymbol{\theta}}$ from the parameter $\boldsymbol{\theta}^*$:

$$\delta\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*.$$

Proof: The probabilistic order of $\delta\boldsymbol{\theta}$ is $\mathcal{O}_p(n^{-1/2})$. Let $\boldsymbol{\eta}_{\mathrm{nu}}$ be

$$\boldsymbol{\eta}_{\mathrm{nu}} := \int \boldsymbol{\xi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Using the Taylor expansion

$$\begin{aligned} \exp(t) &= 1 + t + \mathcal{O}(t^2), \\ \log(1+t) &= t + \mathcal{O}(t^2), \end{aligned}$$

we have the following the asymptotic expansion of $\widehat{r}$:

$$\begin{aligned} \log \widehat{r}(\boldsymbol{x}) &= \log \frac{r^*(\boldsymbol{x}) \exp\left\{\delta\boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x})\right\}}{\frac{1}{n}\sum_{j=1}^n r^*(\boldsymbol{x}_j^{\mathrm{de}}) \exp\left\{\delta\boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}} \\ &= \log r^*(\boldsymbol{x}) + \delta\boldsymbol{\theta}^\top \boldsymbol{\xi}(\boldsymbol{x}) \\ &\quad - \log\left\{1 + \left(\frac{1}{n}\sum_{j=1}^n r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1\right) + \delta\boldsymbol{\theta}^\top \cdot \frac{1}{n}\sum_{j=1}^n r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) + \mathcal{O}_p(n^{-1})\right\} \\ &= \log r^*(\boldsymbol{x}) + \delta\boldsymbol{\theta}^\top(\boldsymbol{\xi}(\boldsymbol{x}) - \boldsymbol{\eta}_{\mathrm{nu}}) - \left(\frac{1}{n}\sum_{j=1}^n r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1\right) + \mathcal{O}_p(n^{-1}). \end{aligned}$$

Therefore, we have

$$\frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} - 1 = \delta\boldsymbol{\theta}^\top(\boldsymbol{\xi}(\boldsymbol{x}) - \boldsymbol{\eta}_{\mathrm{nu}}) - \left(\frac{1}{n}\sum_{j=1}^n r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1\right) + \mathcal{O}_p(n^{-1}).$$

Substituting the above expression into the Pearson divergence in Eq.(19), we obtain

$$\begin{aligned} \mathrm{PE}(p^*_{\mathrm{nu}} \| \widehat{r} \cdot p^*_{\mathrm{de}}) &= \frac{1}{2}\int p^*_{\mathrm{nu}}(\boldsymbol{x}) \left(\frac{\widehat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^\top\right) + \frac{1}{2}\left(\frac{1}{n}\sum_{j=1}^n (r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1)\right)^2 + \mathcal{O}_p(n^{-3/2}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\mathrm{PE}(p^*_{\mathrm{nu}} \| \widehat{r} \cdot p^*_{\mathrm{de}})\right] &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}}) \cdot \mathbb{E}\left[\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^\top\right]\right) + \frac{1}{2n}\int p^*_{\mathrm{de}}(\boldsymbol{x})(r^*(\boldsymbol{x}) - 1)^2 \mathrm{d}\boldsymbol{x} \\ &\quad + \mathcal{O}(n^{-3/2}) \\ &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}}) \cdot \mathbb{E}\left[\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^\top\right]\right) + \frac{1}{2n}\mathrm{PE}(p^*_{\mathrm{de}} \| p^*_{\mathrm{nu}}) + \mathcal{O}(n^{-3/2}). \end{aligned}$$

Applying Eq (18) to the above equation, we obtain Eq.(20). ■

B Proof of Lemma 1

According to Lemma 12, we need to compute the asymptotic variance of estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{A}}$ in order to compute the measure of accuracy of $\widehat{r}_{\mathrm{A}}$. Based on the standard asymptotic statistics, the asymptotic variance of the maximum likelihood estimator for the exponential family is given as

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} - \boldsymbol{\theta}^*_{\mathrm{nu}}) &\sim N(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})^{-1}),\\ \sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{de}} - \boldsymbol{\theta}^*_{\mathrm{de}}) &\sim N(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{de}})^{-1}),\end{aligned}$$

when the sample size n goes to infinity. Under the regularity condition of parametric estimation, the bias of estimator is given as

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} - \boldsymbol{\theta}^*_{\mathrm{nu}}\right] &= \mathcal{O}(n^{-1}),\\ \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathrm{de}} - \boldsymbol{\theta}^*_{\mathrm{de}}\right] &= \mathcal{O}(n^{-1}).\end{aligned}$$

Then, for

$$\begin{aligned}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{A}} &:= \widehat{\boldsymbol{\theta}}_{\mathrm{A}} - (\boldsymbol{\theta}^*_{\mathrm{nu}} - \boldsymbol{\theta}^*_{\mathrm{de}})\\ &= (\widehat{\boldsymbol{\theta}}_{\mathrm{nu}} - \boldsymbol{\theta}^*_{\mathrm{nu}}) - (\widehat{\boldsymbol{\theta}}_{\mathrm{de}} - \boldsymbol{\theta}^*_{\mathrm{de}}),\end{aligned}$$

we have

$$\mathbb{E}\left[\delta\widehat{\boldsymbol{\theta}}_{\mathrm{A}}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{A}}^{\top}\right] = \frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})^{-1} + \frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{de}})^{-1} + \mathcal{O}(n^{-3/2}),$$

where we used the fact that $\widehat{\boldsymbol{\theta}}_{\mathrm{nu}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{de}}$ are independent. Substituting the above asymptotic variance of $\delta\widehat{\boldsymbol{\theta}}_{\mathrm{A}}$ into the first term of Eq.(20), we obtain

$$\begin{aligned}J(\widehat{r}_{\mathrm{A}}) &= \frac{1}{2n}\left[\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})^{-1} + \boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{de}})^{-1})\right) + \mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}})\right] + \mathcal{O}(n^{-3/2}),\\ &= \frac{1}{2n}\left[\dim\Theta + \mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{nu}})\boldsymbol{F}(\boldsymbol{\theta}^*_{\mathrm{de}})^{-1}\right) + \mathrm{PE}(p^*_{\mathrm{de}}\|p^*_{\mathrm{nu}})\right] + \mathcal{O}(n^{-3/2}),\end{aligned}$$

which concludes the proof. ■

C Proof of Lemma 2

Let $(\widehat{\boldsymbol{\theta}}_{\mathrm{B}}, \widehat{\theta}_{\mathrm{B},0})$ be the maximum likelihood estimator with the model (12). Let

$$\begin{aligned}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{B}} &:= \widehat{\boldsymbol{\theta}}_{\mathrm{B}} - \boldsymbol{\theta}^*\\ &= \widehat{\boldsymbol{\theta}}_{\mathrm{B}} - (\boldsymbol{\theta}^*_{\mathrm{nu}} - \boldsymbol{\theta}^*_{\mathrm{de}}).\end{aligned}$$

Based on the standard asymptotic statistics, the asymptotic variance of the maximum likelihood estimator for the exponential family is given as

$$\sqrt{n}\delta\boldsymbol{\theta}_{\mathrm{B}} \sim N\left(\mathbf{0}, \boldsymbol{H}_{11}(\boldsymbol{\theta}^*, \theta_0^*)\right),$$

when the sample size n goes to infinity. $\boldsymbol{H}_{11}(\boldsymbol{\theta}, \theta_0)$ is the submatrix of the inverse matrix of the Fisher information matrix as defined in Eq.(13) and $(\boldsymbol{\theta}^*, \theta_0^*)$ is the parameter corresponding to the density ratio $r^*(\boldsymbol{x})$. Hence, the asymptotic variance of $\delta\boldsymbol{\theta}_{\mathrm{B}}$ is given as

$$\mathbb{E}\left[\delta\boldsymbol{\theta}_{\mathrm{B}}\delta\boldsymbol{\theta}_{\mathrm{B}}^{\top}\right] = \frac{1}{n}\boldsymbol{H}_{11}(\boldsymbol{\theta}^*, \theta_0^*) + \mathcal{O}(n^{-3/2}).$$

Substituting the asymptotic variance of $\delta\widehat{\boldsymbol{\theta}}_{\mathrm{B}}$ into the first term of Eq.(20), we establish the lemma. ■

D Proof of Lemma 3

By simple calculation, we find that the optimal solution $(\widehat{\boldsymbol{\theta}}_{\mathrm{C}}, \widehat{\theta}_{\mathrm{C},0})$ satisfies

$$\widehat{\theta}_{\mathrm{C},0} = -\log\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\right).$$

The extremal condition for Eq.(9) with the above expression provides the following equation:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) = \frac{\sum_{j=1}^{n}\exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})}{\sum_{j=1}^{n}\exp\left\{\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}}. \tag{21}$$

Let $\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ be

$$\begin{aligned}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}} &:= \widehat{\boldsymbol{\theta}}_{\mathrm{C}} - \boldsymbol{\theta}^* \\ &= \widehat{\boldsymbol{\theta}}_{\mathrm{C}} - (\boldsymbol{\theta}_{\mathrm{nu}}^* - \boldsymbol{\theta}_{\mathrm{de}}^*).\end{aligned}$$

Then, Eq.(21) is represented as

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) = \frac{\sum_{j=1}^{n}r^*(\boldsymbol{x}_j^{\mathrm{de}})\exp\left\{\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})}{\sum_{j=1}^{n}r^*(\boldsymbol{x}_j^{\mathrm{de}})\exp\left\{\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^{\top}\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\right\}}.$$

Using the Taylor expansion

$$\begin{aligned}\exp(t) &= 1 + t + \mathcal{O}(t^2), \\ \frac{1}{1-t} &= 1 + t + \mathcal{O}(t^2),\end{aligned}$$

the asymptotic expansion of the right-hand side of the above equation yields

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) &= \left\{\frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) + \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})^\top \delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}\right\} \\
&\quad \times \left\{1 - \left(\frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1\right) - \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})^\top \delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}\right\} + \mathcal{O}_p(n^{-1}) \\
&= \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) - \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\left(\frac{1}{n}\sum_{j'=1}^{n} r^*(\boldsymbol{x}_{j'}^{\mathrm{de}}) - 1\right) \\
&\quad + \left\{\frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})^\top\right. \\
&\qquad \left. - \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}})\frac{1}{n}\sum_{j'=1}^{n} r^*(\boldsymbol{x}_{j'}^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_{j'}^{\mathrm{de}})^\top\right\}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}} + \mathcal{O}_p(n^{-1}) \\
&= \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) - (\boldsymbol{\eta}_{\mathrm{nu}} + \mathcal{O}_p(n^{-1/2}))\left(\frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}}) - 1\right) \\
&\quad + \boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}} + \mathcal{O}_p(n^{-1}) \\
&= \boldsymbol{\eta}_{\mathrm{nu}} + \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})\left(\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) - \boldsymbol{\eta}_{\mathrm{nu}}\right) + \boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}} + \mathcal{O}_p(n^{-1}).
\end{aligned}
$$

Hence, we obtain

$$
\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) - \boldsymbol{\eta}_{\mathrm{nu}}) - \frac{1}{n}\sum_{j=1}^{n} r^*(\boldsymbol{x}_i^{\mathrm{de}})(\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{de}}) - \boldsymbol{\eta}_{\mathrm{nu}}) + \mathcal{O}_p(n^{-1}). \tag{22}
$$

When the sample size goes to infinity, the central limit theorem provides

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{\xi}(\boldsymbol{x}_i^{\mathrm{nu}}) - \boldsymbol{\eta}_{\mathrm{nu}}) &\sim N(\boldsymbol{0}, \boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)), \\
\frac{1}{\sqrt{n}}\sum_{j=1}^{n} r^*(\boldsymbol{x}_j^{\mathrm{de}})(\boldsymbol{\xi}(\boldsymbol{x}_j^{\mathrm{de}}) - \boldsymbol{\eta}_{\mathrm{nu}}) &\sim N(\boldsymbol{0}, \boldsymbol{G}),
\end{aligned} \tag{23}
$$

where $\boldsymbol{G}$ is the matrix defined in Lemma 3. Combining Eqs.(22) and (23), we obtain the following expression of the asymptotic variance of $\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$:

$$
\mathbb{E}\left[\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}^\top\right] = \frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)^{-1} + \frac{1}{n}\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)^{-1}\boldsymbol{G}\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)^{-1} + \mathcal{O}(n^{-3/2}).
$$

Substituting the asymptotic variance of $\delta\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ into the first term of Eq.(20), we establish the lemma. ■

E Proof of Theorem 4

We compare the coefficients of order $\mathcal{O}(n^{-1})$ in $J(\widehat{r}_{\mathrm{A}})$ and $J(\widehat{r}_{\mathrm{B}})$, and prove the following inequality:

$$\dim\Theta + \mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{de}}^*)^{-1}\right) \le \mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)\boldsymbol{H}_{11}(\boldsymbol{\theta}^*,\theta_0^*)\right). \tag{24}$$

Let $\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\theta_0)$ be the Fisher information matrix of the logistic model

$$\begin{aligned} q_{\boldsymbol{\eta}}(y=\text{`nu'}|\boldsymbol{x};\boldsymbol{\theta},\theta_0) &= \frac{\exp\left\{\theta_0+\boldsymbol{\theta}^\top(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta})\right\}}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^\top(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta})\right\}},\\ q_{\boldsymbol{\eta}}(y=\text{`de'}|\boldsymbol{x};\boldsymbol{\theta},\theta_0) &= \frac{1}{1+\exp\left\{\theta_0+\boldsymbol{\theta}^\top(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta})\right\}}, \end{aligned} \tag{25}$$

where $\boldsymbol{\eta}$ is a fixed vector. Let us represent $\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\theta_0)^{-1}$ in a block form as

$$\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}(\boldsymbol{\theta},\theta_0)^{-1} = \begin{pmatrix} \boldsymbol{H}_{\boldsymbol{\eta},11}(\boldsymbol{\theta},\theta_0) & \boldsymbol{h}_{\boldsymbol{\eta},12}(\boldsymbol{\theta},\theta_0) \\ \boldsymbol{h}_{\boldsymbol{\eta},12}(\boldsymbol{\theta},\theta_0)^\top & h_{\boldsymbol{\eta},22}(\boldsymbol{\theta},\theta_0) \end{pmatrix}.$$

When the functions $1,\xi_1(\boldsymbol{x}),\ldots,\xi_k(\boldsymbol{x})$ are linearly independent, the maximum likelihood estimator (mle) of $\boldsymbol{\theta}$ for model (25) is given by $\widehat{\boldsymbol{\theta}}_{\mathrm{B}}$. The equality

$$\theta_0+\boldsymbol{\theta}^\top\boldsymbol{\xi}(\boldsymbol{x}) = \widetilde{\theta}_0+\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta})$$

implies $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$ and $\theta_0=\widetilde{\theta}_0-\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\eta}=\widetilde{\theta}_0-\boldsymbol{\theta}^\top\boldsymbol{\eta}$. Due to the equality $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$, we see that the mle of $\boldsymbol{\theta}$ is equal to that of $\widetilde{\boldsymbol{\theta}}$, and hence, the variance is unchanged under the parameter transformation, that is,

$$\boldsymbol{H}_{\boldsymbol{\eta},11}(\widetilde{\boldsymbol{\theta}},\widetilde{\theta}_0) = \boldsymbol{H}_{\boldsymbol{\eta},11}(\boldsymbol{\theta},\theta_0+\boldsymbol{\theta}^\top\boldsymbol{\eta}) = \boldsymbol{H}_{11}(\boldsymbol{\theta},\theta_0)$$

holds for any $\boldsymbol{\eta}$. The Fisher information matrix $\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}$ can be represented as

$$\begin{aligned} \widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}^*,\theta_0^*+\boldsymbol{\theta}^{*\top}\boldsymbol{\eta}) &= \frac{1}{2}\int\frac{p_{\mathrm{nu}}^*(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})}{p_{\mathrm{nu}}^*(\boldsymbol{x})+p_{\mathrm{de}}^*(\boldsymbol{x})}\begin{pmatrix}\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}\\ 1\end{pmatrix}\begin{pmatrix}(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta})^\top & 1\end{pmatrix}\mathrm{d}\boldsymbol{x}\\ &= \begin{pmatrix}\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11} & \widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}\\ \widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}^\top & \widetilde{f}_{\boldsymbol{\eta},22}\end{pmatrix}. \end{aligned}$$

The first equality is obtained by the straightforward calculation of the Fisher information matrix. Applying the matrix inversion formula to the block form, we obtain

$$\begin{aligned} \boldsymbol{H}_{11}(\boldsymbol{\theta}^*,\theta_0^*) &= \boldsymbol{H}_{\boldsymbol{\eta},11}(\boldsymbol{\theta}^*,\theta_0^*+\boldsymbol{\theta}^{*\top}\boldsymbol{\eta})\\ &= \widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1}+\frac{\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1}\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}^\top\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1}}{\widetilde{f}_{\boldsymbol{\eta},22}-\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}^\top\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1}\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}}. \end{aligned}$$

Since $\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}}$ is positive definite, we have

$$\widetilde{f}_{\boldsymbol{\eta},22}-\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}^{\top}\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1}\widetilde{\boldsymbol{f}}_{\boldsymbol{\eta},12}>0,$$

and hence, we obtain the inequality

$$\boldsymbol{H}_{11}(\boldsymbol{\theta}^*,\theta_0^*)\succeq\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta},11}^{-1},$$

for any $\boldsymbol{\eta}$. In the above formula, $\boldsymbol{A}\succeq\boldsymbol{B}$ indicates the fact that the matrix $\boldsymbol{A}-\boldsymbol{B}$ is positive semidefinite. On the other hand, the inequalities

$$\begin{aligned}
\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}_{\mathrm{nu}},11}&=\frac{1}{2}\int\frac{p_{\mathrm{nu}}^*(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})}{p_{\mathrm{nu}}^*(\boldsymbol{x})+p_{\mathrm{de}}^*(\boldsymbol{x})}(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})^{\top}\mathrm{d}\boldsymbol{x}\\
&\preceq\frac{1}{2}\int p_{\mathrm{nu}}^*(\boldsymbol{x})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{nu}})^{\top}\mathrm{d}\boldsymbol{x}\\
&=\frac{1}{2}\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*),\\
\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}_{\mathrm{de}},11}&=\frac{1}{2}\int\frac{p_{\mathrm{nu}}^*(\boldsymbol{x})p_{\mathrm{de}}^*(\boldsymbol{x})}{p_{\mathrm{nu}}^*(\boldsymbol{x})+p_{\mathrm{de}}^*(\boldsymbol{x})}(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{de}})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{de}})^{\top}\mathrm{d}\boldsymbol{x}\\
&\preceq\frac{1}{2}\int p_{\mathrm{de}}^*(\boldsymbol{x})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{de}})(\boldsymbol{\xi}(\boldsymbol{x})-\boldsymbol{\eta}_{\mathrm{de}})^{\top}\mathrm{d}\boldsymbol{x}\\
&=\frac{1}{2}\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{de}}^*)
\end{aligned}$$

hold. Therefore, we obtain

$$\begin{aligned}
\boldsymbol{H}_{11}(\boldsymbol{\theta}^*,\theta_0^*)&\succeq\frac{1}{2}\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}_{\mathrm{nu}},11}^{-1}+\frac{1}{2}\widetilde{\boldsymbol{F}}_{\boldsymbol{\eta}_{\mathrm{de}},11}^{-1}\\
&\succeq\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)^{-1}+\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{de}}^*)^{-1}.
\end{aligned}$$

By multiplying $\boldsymbol{F}(\boldsymbol{\theta}_{\mathrm{nu}}^*)$ from the left-hand side and taking the trace of both sides, we obtain the inequality (24). ■

F Proof of Theorem 10

We prove the general expression (17) for any $\overline{r}$ in the exponential model. The optimality condition of Eq.(16) provides the equality

$$\int p_{\mathrm{de}}^*(\boldsymbol{x})\overline{r}_{\mathrm{C}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int p_{\mathrm{nu}}^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1.$$

Hence, we have

$$\int p_{\mathrm{de}}^*(\boldsymbol{x})(\overline{r}(\boldsymbol{x})-\overline{r}_{\mathrm{C}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}=\int p_{\mathrm{de}}^*(\boldsymbol{x})\overline{r}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-1.$$

Applying the Schwarz inequality to the above equality, we obtain

$$\mathrm{D}(\overline{r}, \overline{r}_{\mathrm{C}}) \geq \left| \int p^*_{\mathrm{de}}(\boldsymbol{x})\, \overline{r}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - 1 \right|^2 .$$

Thus, $\overline{r}$ is different from $\overline{r}_{\mathrm{C}}$ unless $p^*_{\mathrm{de}} \cdot \overline{r}$ is a probability density. ■

G Proof of Corollary 11

The optimality condition of the method (A) provides the equality

$$\int p^*_{\mathrm{nu}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \overline{p}_{\mathrm{nu}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Substituting the equality $\overline{p}_{\mathrm{nu}}(\boldsymbol{x}) = \overline{p}_{\mathrm{de}}(\boldsymbol{x}) \overline{r}_{\mathrm{A}}(\boldsymbol{x})$ into the above expression, we have

$$\int p^*_{\mathrm{nu}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \overline{p}_{\mathrm{de}}(\boldsymbol{x}) \overline{r}_{\mathrm{A}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

When p^*_{de} belongs to the exponential model, we have $p^*_{\mathrm{de}} = \overline{p}_{\mathrm{de}}$ and thus, the equality

$$\int p^*_{\mathrm{nu}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int p^*_{\mathrm{de}}(\boldsymbol{x}) \overline{r}_{\mathrm{A}}(\boldsymbol{x}) \boldsymbol{\xi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

holds. The above equation is exactly the same as the optimality condition of Eq.(16) for the method (C). Thus, $\overline{r}_{\mathrm{A}} = \overline{r}_{\mathrm{C}}$ holds. ■

References

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

[2] N. Cesa-Bianchi and G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, Cambridge, NY, 2006.

[3] K.F. Cheng and C.K. Chu, "Semiparametric density estimation under a two-sample density ratio model," Bernoulli, vol.10, no.4, pp.583–604, 2004.

[4] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol.20, pp.273–297, 1995.

[5] B. Efron, "The efficiency of logistic regression compared to normal discriminant analysis," Journal of the American Statistical Association, vol.70, no.352, pp.892–898, 1975.

[6] V.V. Fedorov, Theory of Optimal Experiments, Academic Press, New York, 1972.

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

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

[9] T. Kanamori, "Pool-based active learning with optimal sampling distribution and its information geometrical interpretation," Neurocomputing, vol.71, no.1–3, pp.353–362, 2007.

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

[11] T. Kanamori and H. Shimodaira, "Active learning algorithm using the maximum weighted log-likelihood estimator," Journal of Statistical Planning and Inference, vol.116, no.1, pp.149–162, 2003.

[12] S. Kullback and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol.22, pp.79–86, 1951.

[13] E.L. Lehmann and G. Casella, Theory of Point Estimation, second ed., Springer, New York, 1998.

[14] P. Liang and M. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008), ed. A. McCallum and S. Roweis, pp.584–591, Omnipress, 2008.

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

[16] J. Qin, "Inferences for case-control and semiparametric two-sample density ratio models," Biometrika, vol.85, no.3, pp.619–639, 1998.

[17] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, eds., Dataset Shift in Machine Learning, MIT Press, Cambridge, MA, 2009.

[18] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," Journal of Statistical Planning and Inference, vol.90, no.2, pp.227–244, 2000.

[19] A. Smola, L. Song, and C.H. Teo, "Relative novelty detection," JMLR Workshop and Conference Proceedings, ed. D. van Dyk and M. Welling, Twelfth International Conference on Artificial Intelligence and Statistics, vol.5, pp.536–543, 2009.

[20] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization error," Journal of Machine Learning Research, vol.7, pp.141–166, Jan. 2006.

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

[22] M. Sugiyama, M. Krauledat, and K.R. Müller, "Covariate shift adaptation by importance weighted cross validation," Journal of Machine Learning Research, vol.8, pp.985–1005, May 2007.

[23] M. Sugiyama and K.R. Müller, "Input-dependent estimation of generalization error under covariate shift," Statistics & Decisions, vol.23, no.4, pp.249–279, 2005.

[24] M. Sugiyama and S. Nakajima, "Pool-based active learning in approximate linear regression," Machine Learning, vol.75, no.3, pp.249–274, 2009.

[25] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," Annals of the Institute of Statistical Mathematics, vol.60, no.4, pp.699–746, 2008.

[26] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, "Least-squares conditional density estimation," IEICE Transactions on Information and Systems, vol.E93-D, no.3, 2010. to appear.

[27] M. Sugiyama, P. von Bünau, M. Kawanabe, and K.R. Müller, Covariate Shift Adaptation: Towards Machine Learning in Non-Stationary Environment, MIT Press, Cambridge, MA. to appear.

[28] T. Suzuki and M. Sugiyama, "Estimating squared-loss mutual information for independent component analysis.," Independent Component Analysis and Signal Separation, ed. T. Adali, C. Jutten, J.M.T. Romano, and A.K. Barros, Lecture Notes in Computer Science, vol.5441, Berlin, pp.130–137, Springer, 2009.

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

[30] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes," BMC Bioinformatics, vol.10, no.1, p.S52, 2009.

[31] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori, "Approximating mutual information by maximum likelihood density ratio estimation," JMLR Workshop and Conference Proceedings, ed. Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y.V. de Peer, New Challenges for Feature Selection in Data Mining and Knowledge Discovery, vol.4, pp.5–20, 2008.

[32] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.

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

[34] D.P. Wiens, "Robust weights and designs for biased regression models: Least squares and generalized M-estimation," Journal of Statistical Planning and Inference, vol.83, no.2, pp.395–412, 2000.

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