f-Divergence Estimation and Two-Sample Homogeneity Test under Semiparametric Density-Ratio Models

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

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

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

Abstract

A density ratio is defined by the ratio of two probability densities. We study the inference problem of density ratios and apply a semi-parametric density-ratio estimator to the two-sample homogeneity test. In the proposed test procedure, the f-divergence between two probability densities is estimated using a density-ratio estimator. The f-divergence estimator is then exploited for the two-sample homogeneity test. We derive an optimal estimator of f-divergence in the sense of the asymptotic variance in a semiparametric setting, and provide a statistic for two-sample homogeneity test based on the optimal estimator. We prove that the proposed test dominates the existing empirical likelihood score test. Through numerical studies, we illustrate the adequacy of the asymptotic theory for finite-sample inference.

Keywords

Density ratio, divergence, two-sample test, semiparametric model, asymptotic expansion.

1 Introduction

In this paper, we study the two-sample homogeneity test under semiparametric density-ratio models. An estimator of density ratios is exploited to obtain a test statistic. For two probability densities $p_{\mathrm{n}}(x)$ and $p_{\mathrm{d}}(x)$ over a probability space $\mathcal{X}$, the density ratio $r(x)$ is defined as the ratio of these densities, that is,

$$r(x) = \frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)},$$

in which p_{n} (p_{d}) denotes the "numerator" ("denominator") of the density ratio. For statistical examples and motivations of the density ratio model, see [20, 5, 11] and the references therein. Qin [20] has studied the inference problem of density ratios under retrospective sampling plans, and proved that in the sense of Godambe [8], the estimating function obtained from the prospective likelihood is optimal in a class of unbiased estimating functions for semiparametric density ratio models. As a similar approach, a semiparametric density ratio estimator based on logistic regression is studied in [4].

The density ratio is closely related to the inference of divergences. A divergence is a discrepancy measure between pairs of multivariate probability densities, and the f-divergence [1, 6] is a class of divergences based on the ratio of two probability densities. For a strictly convex function f satisfying $f(1) = 0$, the f-divergence between two probability densities $p_{\mathrm{d}}(x)$ and $p_{\mathrm{n}}(x)$ is defined by

$$D_f(p_{\mathrm{d}}, p_{\mathrm{n}}) = \int_{\mathcal{X}} p_{\mathrm{d}}(x) f\left(\frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)}\right) dx. \tag{1}$$

Since f is strictly convex, the f-divergence is non-negative and takes zero if and only if $p_{\mathrm{n}} = p_{\mathrm{d}}$ holds. Popular divergences such as the Kullback-Leibler (KL) divergence [16], the Hellinger distance, and the Pearson divergence are included in the f-divergence class. In statistics, machine learning, and information theory, the f-divergence is often exploited as a metric between probability distributions, even though the divergence does not necessarily satisfy the definition of the metric.

A central topic in this line of research is to estimate the divergence based on samples from each probability distribution. A typical approach is to exploit non-parametric estimators of probability densities for the estimation of divergence [25, 26].

The conjugate expression of f can be exploited for the estimation of the f-divergence in the context of one-sample problems [3, 12] and two-sample problems [14, 15]. A kernel-based estimator of the f-divergence has been developed by using a non-parametric density-ratio model [19].

Once the divergence between two probability densities is estimated, the homogeneity test can be conducted. In the homogeneity test, the null hypothesis is represented as $H_0 : p_{\mathrm{n}} = p_{\mathrm{d}}$ against the complementary alternative $H_1 : p_{\mathrm{n}} \neq p_{\mathrm{d}}$. If an estimate of $D_f(p_{\mathrm{d}}, p_{\mathrm{n}})$ is beyond some positive value, the null hypothesis is rejected and the alternative is accepted. Keziou [13] and Keziou and Leoni-Aubin [14, 15] have studied the homogeneity test using a f-divergence estimator for semiparametric density-ratio models. On the other hand,

Fokianos et al. [7] adopted a more direct approach. They have proposed the Wald-type score test derived from the empirical likelihood estimator of density ratios. In our paper, we consider the optimality of f-divergence estimators, and investigate the relation between the test statistic using the f-divergence estimator and the Wald-type score test derived from the empirical likelihood estimator.

The rest of this paper is organized as follows: In Section 2 we introduce a class of estimators of density ratios for semiparametric density-ratio models. In Section 3, we consider the asymptotic property of the f-divergence estimator. The main results of this paper are presented in Section 4 and Section 5. Among a class of estimators, we present the optimal estimator of the f-divergence, which is then exploited for two-sample homogeneity test. In one-sample problems, Broniatowski and Keziou [3] proposed the estimator using the conjugate expression of the f-divergence, while they argued neither its optimality nor its efficiency. A main contribution of this paper is to present the optimal estimator of the f-divergence in the sense of asymptotic variance under the semiparametric density-ratio models. Then, we propose a test statistic based on the optimal f-divergence estimator, and investigate its power function. Numerical studies are provided in Section 6, illustrating the adequacy of our asymptotic theory for finite-sample inference. Section 7 is devoted to concluding remarks. Some calculations are deferred to Appendix.

We summarize some notations to be used throughout the paper. For a vector (matrix) a, $\|a\|$ is the Euclidean (Frobenius) norm of a, and a^T denotes the transposition of a. The first and the second derivative of $f : \mathbb{R} \to \mathbb{R}$ are denoted as f' and f'', respectively. The gradient column vector of the function $g(\theta)$ with respect to the parameter θ is represented as ∇g, i.e., $\nabla g = (\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_d})^T$. For a vector valued function $g(\theta) = (g_1(\theta), \ldots, g_d(\theta))$, ∇g denotes the Jacobian matrix $(\nabla g)_{ij} = \frac{\partial g_i}{\partial \theta_j}$. For a vector-valued function $\eta(x;\theta) = (\eta_1(x;\theta), \ldots, \eta_d(x;\theta))^T$, let $\mathcal{L}[\eta(x;\theta)]$ be the linear space $\mathcal{L}[\eta(x;\theta)] = \left\{ \sum_{k=1}^d a_k \, \eta_k(x;\theta) \;\middle|\; a_1, \ldots, a_d \in \mathbb{R} \right\}$. Let $N_d(\mu, \Sigma)$ be the d-dimensional normal distribution with the mean vector μ and the variance-covariance matrix Σ. The dimension d may be dropped if there is no confusion. For a sequence of random variables X_n, $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{p} X$ denote the convergence in law to X and the convergence in probability to X, respectively. We also use the probabilistic orders, $O_p(\cdot)$ and $o_p(\cdot)$, which are defined in [24].

2 Estimation of Density Ratios

We introduce the method of estimating density ratios according to Qin [20]. Suppose that two sets of samples are independently generated from each probability:

$$x_1^{(\mathrm{n})}, \ldots, x_{m_\mathrm{n}}^{(\mathrm{n})} \sim_{i.i.d.} p_\mathrm{n}, \qquad x_1^{(\mathrm{d})}, \ldots, x_{m_\mathrm{d}}^{(\mathrm{d})} \sim_{i.i.d.} p_\mathrm{d}.$$

The model for the density ratio is defined by $r(x;\theta)$ with the parameter $\theta \in \Theta \subset \mathbb{R}^d$. We assume that the true density ratio is represented as

$$r(x) = \frac{p_\mathrm{n}(x)}{p_\mathrm{d}(x)} = r(x;\theta^*)$$

with some $\theta^* \in \Theta$. The model for the density ratio $r(x;\theta)$ is regarded as a semiparametric model for probability densities. That is, even if $r(x;\theta^*) = p_{\mathrm{n}}(x)/p_{\mathrm{d}}(x)$ is specified, there are yet infinite degrees of freedom for the probability densities p_{n} and p_{d}.

The moment matching estimator for the density ratio has been proposed by Qin [20]. Let $\eta(x;\theta) \in \mathbb{R}^d$ be a vector-valued function from $\mathcal{X} \times \Theta$ to $\mathbb{R}^d$, and the estimation function Q_η is defined as

$$Q_\eta(\theta) = \frac{1}{m_{\mathrm{d}}} \sum_{i=1}^{m_{\mathrm{d}}} r(x_i^{(\mathrm{d})};\theta)\eta(x_i^{(\mathrm{d})};\theta) - \frac{1}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} \eta(x_j^{(\mathrm{n})};\theta).$$

Since $p_{\mathrm{n}}(x) = r(x;\theta^*)p_{\mathrm{d}}(x)$ holds, the expectation of $Q_\eta(\theta)$ over the observed samples vanishes at $\theta = \theta^*$. In addition, the estimation function $Q_\eta(\theta)$ converges to its expectation in the large sample limit. Thus, the estimator $\widehat{\theta}$ defined as a solution of the estimating equation,

$$Q_\eta(\widehat{\theta}) = 0,$$

has the statistical consistency under some mild assumption, see [20] for details. Below, we show a sufficient condition for the consistency and the asymptotic normality of $\widehat{\theta}$.

The moment matching estimation of the density ratio contains a wide range of estimators. Several authors have proposed various density-ratio estimators [13, 14, 15, 19, 23, 10]. These estimators with a finite-dimensional model $r(x;\theta)$ can all be represented as a moment matching estimator. These existing methods, however, are intended to be applied with kernel methods which have been developed in machine learning [22]. The kernel density estimators for probability densities are also exploited as another approach to density ratio estimation [17, 9, 2].

Before presenting the asymptotic results, we prepare some notations. $\mathrm{E}_{\mathrm{n}}[\,\cdot\,]$ and $\mathrm{V}_{\mathrm{n}}[\,\cdot\,]$ denote the expectation and the variance (or the variance-covariance matrix for multi-dimensional random variables) under the probability p_{n}, and $\mathrm{E}_{\mathrm{d}}[\,\cdot\,]$ and $\mathrm{V}_{\mathrm{d}}[\,\cdot\,]$ are defined in the same way for the probability p_{d}. The expectation and the variance by the joint probability of all samples, $x_i^{(\mathrm{n})}\,(i = 1,\ldots,m_{\mathrm{n}})$, $x_j^{(\mathrm{d})}\,(j = 1,\ldots,m_{\mathrm{d}})$ are denoted as $\mathbb{E}[\,\cdot\,]$ and $\mathbb{V}[\,\cdot\,]$, respectively. The covariance matrix between two random variables by the joint probability of all samples is also denoted as $\mathrm{Cov}[\,\cdot\,,\,\cdot\,]$.

We introduce the asymptotic theory of density ratio estimation. Let ρ and m be $\rho = m_{\mathrm{n}}/m_{\mathrm{d}}$, $m = (\frac{1}{m_{\mathrm{n}}} + \frac{1}{m_{\mathrm{d}}})^{-1} = \frac{m_{\mathrm{n}}m_{\mathrm{d}}}{m_{\mathrm{n}}+m_{\mathrm{d}}}$, and let U_η be the d by d matrix defined by

$$U_\eta(\theta) = \mathrm{E}_{\mathrm{n}}[\eta(x;\theta)\nabla \log r(x;\theta)^T],$$

where $\eta(x;\theta)$ is a d-dimensional vector-valued function. Suppose that $U_\eta(\theta)$ is non-degenerate in the vicinity of $\theta = \theta^*$. Below, the notation ρ is also used as the large sample limit of $m_{\mathrm{n}}/m_{\mathrm{d}}$, and we assume that $0 < \rho < \infty$ holds even in the limit.

We introduce the asymptotic property of the density ratio estimator. Assumptions for asymptotic expansion are explicitly presented below. Since the details are shown in [24, Section 5], we skip the proof of the consistency and the asymptotic normality of density ratio estimators. A similar assumption is studied in [3] for one-sample problems.

Assumption 1 (consistency of $\widehat{\theta}$)

1. *The estimator $\widehat{\theta} \in \Theta$ such that $Q_\eta(\widehat{\theta}) = 0$ exists.*
2. *Both*

$$\sup_{\theta\in\Theta} \Big\| \frac{1}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} \eta(x_j^{(\mathrm{n})};\theta) - E_n[\eta(x;\theta)] \Big\|, \text{ and,}$$

$$\sup_{\theta\in\Theta} \Big\| \frac{1}{m_{\mathrm{d}}} \sum_{i=1}^{m_{\mathrm{d}}} \eta(x_i^{(\mathrm{d})};\theta) r(x_i^{(\mathrm{d})};\theta) - E_d[\eta(x;\theta)r(x;\theta)] \Big\|$$

   converge in probability to zero in the large sample limit.
3. $\inf_{\theta:\|\theta-\theta^*\|\geq\varepsilon} \big\| E_n[\eta(x;\theta)] - E_d[\eta(x;\theta)r(x;\theta)] \big\| > 0$ *holds for any* $\varepsilon > 0$.

Note that 2) in Assumption 1 and the triangle inequality lead to the uniform convergence,

$$\sup_{\theta\in\Theta} \big| \|Q_\eta(\theta)\| - \|\mathbb{E}[Q_\eta(\theta)]\| \big| \xrightarrow{p} 0.$$

Along the argument in [24, Section 5], we can prove the asymptotic consistency, $\widehat{\theta} \xrightarrow{p} \theta^*$ in the large sample limit.

For the asymptotic normality, we assume the following conditions.

Assumption 2 (asymptotic normality of $\widehat{\theta}$)

1. *The estimator of the density ratio, $\widehat{\theta}$, exists and is consistent.*
2. *The expectations, $E_n[\|\eta(x;\theta^*)\|^2]$, $E_n[\|\nabla\eta(x;\theta^*)\|]$, $E_d[\|\eta(x;\theta^*)r(x;\theta^*)\|^2]$ and $E_d[\|\nabla(\eta(x;\theta^*)r(x;\theta^*))\|]$ are finite. In the vicinity of θ^*, each element of the second derivatives of $\eta(x;\theta)$ and $\eta(x;\theta)r(x;\theta)$ with respect to θ are dominated by a p_n-integrable function and a p_d-integrable function, respectively.*
3. *The matrix $U_\eta(\theta)$ is non-singular in the vicinity of $\theta = \theta^*$.*

Under Assumption 2, the asymptotic expansion of the estimating equation $Q_\eta(\widehat{\theta}) = 0$ around $\theta = \theta^*$ yields the following convergence in law,

$$\begin{aligned} \sqrt{m}(\widehat{\theta} - \theta^*) &= -\sqrt{m} U_\eta^{-1} Q_\eta + o_p(1) \\ &\xrightarrow{d} N_d\left(0,\, U_\eta^{-1} \frac{\rho \mathrm{V_d}[r\eta] + \mathrm{V_n}[\eta]}{\rho+1} (U_\eta^T)^{-1}\right), \end{aligned} \tag{2}$$

where functions are evaluated at $\theta = \theta^*$. The asymptotic variance above is derived from the equalities,

$$\mathbb{E}[Q_\eta] = 0 \quad \text{and} \quad m \cdot \mathbb{E}[Q_\eta Q_\eta^T] = \frac{\rho \mathrm{V_d}[r\eta] + \mathrm{V_n}[\eta]}{\rho+1}.$$

Qin [20] has shown that the prospective likelihood minimizes the asymptotic variance in the class of moment matching estimators. More precisely, for the density ratio model $r(x;\theta) = \exp\{\alpha + \phi(x;\beta)\}$, $\theta = (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^{d-1}$, the vector-valued function η_{opt} defined by

$$\eta_{\mathrm{opt}}(x;\theta) = \frac{1}{1 + \rho r(x;\theta)} \nabla \log r(x;\theta) \tag{3}$$

minimizes the asymptotic variance (2).

3 Estimation of f-divergence

We consider the estimation of the f-divergence. As shown in (1), the f-divergence is represented as the expectation $f(r(x))$, i.e.,

$$D_f(p_{\mathrm{d}}, p_{\mathrm{n}}) = \int p_{\mathrm{d}}(x) f\left(\frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)}\right) dx = \int p_{\mathrm{d}}(x) f\big(r(x)\big) dx,$$

where $r(x) = p_{\mathrm{n}}(x)/p_{\mathrm{d}}(x)$.

The conjugate representation of the function f is available for the f-divergence estimation [3, 12, 13, 14, 15, 19]. For a convex function f, the conjugate function of f is defined as

$$f^*(w) = \sup_{r \in \mathbb{R}} \{rw - f(r)\},$$

and satisfies

$$f(r) = \sup_{w \in \mathbb{R}} \{rw - f^*(w)\} = rf'(r) - f^*(f'(r)) \tag{4}$$

under a mild assumption on f [21]. Substituting the above expression into the f-divergence, we have

$$D_f(p_{\mathrm{d}}, p_{\mathrm{n}}) = \sup_{w} \left\{ \int p_{\mathrm{n}}(x) w(x) dx - \int p_{\mathrm{d}}(x) f^*(w(x)) dx \right\}, \tag{5}$$

where the supremum is taken over all measurable functions and the supremum is attained at $w(x) = f'(r(x))$. Based on (5), one can consider the f-divergence estimator $\widehat{D}_f$ by replacing the true distributions with their empirical versions:

$$\widehat{D}_f = \sup_{\theta \in \Theta} \left\{ \frac{1}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} f'(r(x_j^{(\mathrm{n})};\theta)) - \frac{1}{m_{\mathrm{d}}} \sum_{i=1}^{m_{\mathrm{d}}} f^*(f'(r(x_i^{(\mathrm{d})};\theta))) \right\}. \tag{6}$$

The above estimator has been considered in some works [3, 19]. Suppose that the maximum value of (6) is attained in the interior of Θ. Then, the extremal condition

$$\frac{1}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} \nabla(f'(r(x_j^{(\mathrm{n})};\theta))) - \frac{1}{m_{\mathrm{d}}} \sum_{i=1}^{m_{\mathrm{d}}} r(x_i^{(\mathrm{d})};\theta) \nabla(f'(r(x_i^{(\mathrm{d})};\theta))) = 0,$$

holds at the optimal solution, where the identity $(f^*)'(f'(r)) = r$ is used. This is the moment matching estimator of the density ratio with

$$\eta(x;\theta) = \nabla(f'(r(x;\theta))) = f''(r(x;\theta))\nabla r(x;\theta). \tag{7}$$

We consider an extension of the f-divergence estimator. In the above estimator, the density ratio is estimated by the moment matching estimator with $\eta(x;\theta) = f''(r(x;\theta))\nabla r(x;\theta)$. Then, the estimated density ratio $r(x;\widehat{\theta})$ is substituted into the expression of the f-divergence derived from the decomposition $f(r) = rf'(r) - f^*(f'(r))$. As an extension, we consider arbitrary moment matching estimators using $\eta(x;\theta)$, and any decomposition of the function f such that

$$f(r) = rf_{\mathrm{n}}(r) + f_{\mathrm{d}}(r). \tag{8}$$

The decomposition (4) corresponds to

$$f_{\mathrm{n}}(r) = f'(r), \quad f_{\mathrm{d}}(r) = -f^*(f'(r)). \tag{9}$$

Then, the f-divergence is represented as

$$\int p_{\mathrm{d}}(x) f\big(r(x)\big)dx = \int p_{\mathrm{n}}(x) f_{\mathrm{n}}(r(x))dx + \int p_{\mathrm{d}}(x) f_{\mathrm{d}}(r(x))dx \tag{10}$$

since $r(x) = p_{\mathrm{n}}(x)/p_{\mathrm{d}}(x)$ holds. The empirical version of (10) provides an estimate of the f-divergence,

$$\widehat{D}_f = \frac{1}{m_{\mathrm{d}}}\sum_{i=1}^{m_{\mathrm{d}}} f_{\mathrm{d}}(r(x_i^{(\mathrm{d})};\widehat{\theta})) + \frac{1}{m_{\mathrm{n}}}\sum_{j=1}^{m_{\mathrm{n}}} f_{\mathrm{n}}(r(x_j^{(\mathrm{n})};\widehat{\theta})), \tag{11}$$

where the parameter $\widehat{\theta}$ is estimated by the estimation function Q_η. In the next section, we study the optimal choice of η and the decomposition $f_{\mathrm{n}}, f_{\mathrm{d}}$.

We may consider a wider class of f-divergence estimators than the estimator of the form (11). For example, one may exploit non-parametric estimators of probability densities with semi-parametric density ratio models to estimate the f-divergence. The estimator (11), however, has the advantage of being simple to analyze statistical properties, since they are the plug-in type estimator. Hence, in this paper, we focus on the estimator (11).

Using the estimator $\widehat{D}_f$, we can conduct the homogeneity test with hypotheses

$$H_0 : p_{\mathrm{n}} = p_{\mathrm{d}}, \qquad H_1 : p_{\mathrm{n}} \neq p_{\mathrm{d}}. \tag{12}$$

When the null hypothesis is true, the f-divergence $D_f(p_{\mathrm{d}}, p_{\mathrm{n}})$ is equal to zero and otherwise $D_f(p_{\mathrm{d}}, p_{\mathrm{n}})$ takes a positive real value. Thus, the null hypothesis is rejected when $\widehat{D}_f > t$ holds, where t is a positive constant determined from the significance level of the test.

4 Optimal Estimator of f-divergence

In a semiparametric setting, we consider the optimal estimator of the f-divergence. The asymptotic variance is used as the criterion for the comparison of estimators [8]. For the model $r(x;\theta)$ and the function $f(r)$, we assume the following conditions.

Assumption 3

1. *The model $r(x;\theta)$ includes the constant function 1.*
2. *For any $\theta \in \Theta$, $1 \in \mathcal{L}[\nabla \log r(x;\theta)]$ holds.*
3. *f is third-order differentiable, and a strictly convex function satisfying $f(1) = f'(1) = 0$.*

Standard models of density ratios satisfy 1) and 2) of Assumption 3. Later, we show some examples.

In addition, we assume the following conditions to justify the asymptotic expansion of the estimator $\widehat{D}_f$.

Assumption 4 (asymptotic expansion of $\widehat{D}_f$)

1. *For the estimator $\widehat{\theta}$, $\sqrt{m}(\widehat{\theta} - \theta^*)$ converges in distribution to a centered multivariate normal distribution.*
2. *For the decomposition $f(r) = f_d(r) + r f_n(r)$, suppose that $E_d[|f_d(r(x;\theta^*))|^2]$, $E_d[\|\nabla f_d(r(x;\theta^*))\|]$, $E_n[|f_n(r(x;\theta^*))|^2]$, and $E_n[\|\nabla f_n(r(x;\theta^*))\|]$ are finite. In the vicinity of θ^*, the second derivatives of $f_{\mathrm{n}}(r(x;\theta))$ and $f_{\mathrm{d}}(r(x;\theta))$ with respect to θ are dominated by a p_{n}-integrable function and a p_{d}-integrable function, respectively.*
3. *$E_n[(f'(r;\theta^*)) - f_n(r(x;\theta^*))\nabla \log r(x;\theta^*)]$ exists.*

Under Assumption 4, the delta method is available. See [24, Section 3] for details.

We compare the asymptotic variance of two estimators for the f-divergence; one is the estimator $\widehat{D}_f$ derived from the moment matching estimator using $\eta(x;\theta)$ and the decomposition $f(r) = f_{\mathrm{d}}(r) + r f_{\mathrm{n}}(r)$, and the other is the estimator $\bar{D}_f$ defined by the density ratio estimator using $\bar{\eta}(x;\theta)$ and the decomposition $f(r) = \bar{f}_{\mathrm{d}}(r) + r\bar{f}_{\mathrm{n}}(r)$. In order to compare the variances of these estimators, we consider the following formula,

$$0 \leq \mathbb{V}[\widehat{D}_f - \bar{D}_f] = \mathbb{V}[\widehat{D}_f] - \mathbb{V}[\bar{D}_f] - 2\,\mathrm{Cov}[\,\widehat{D}_f - \bar{D}_f,\, \bar{D}_f\,].$$

Suppose that the third term vanishes for any $\widehat{D}_f$, then we have the inequality $\mathbb{V}[\bar{D}_f] \leq \mathbb{V}[\widehat{D}_f]$ for any $\widehat{D}_f$. This implies that the estimator $\bar{D}_f$ is the asymptotically optimal estimator for the f-divergence.

We compute the covariance $\mathrm{Cov}[\,\widehat{D}_f - \bar{D}_f, \bar{D}_f\,]$. Let the column vectors $c(\theta)$ and $\bar{c}(\theta) \in \mathbb{R}^d$ be

$$c(\theta) = \mathrm{E_n}[\{f'(r(x;\theta)) - f_\mathrm{n}(r(x;\theta))\}\nabla \log r(x;\theta)],$$
$$\bar{c}(\theta) = \mathrm{E_n}[\{f'(r(x;\theta)) - \bar{f}_\mathrm{n}(r(x;\theta))\}\nabla \log r(x;\theta)].$$

Then, under Assumption 3 and Assumption 4, some calculation of the covariance gives the equality

$$m(1+\rho^{-1}) \cdot \mathrm{Cov}[\,\widehat{D}_f - \bar{D}_f, \bar{D}_f\,] = \mathrm{E_n}\big[\big\{\bar{f}_\mathrm{n}(r) - f_\mathrm{n}(r) + \bar{c}^T U_{\bar{\eta}}^{-1}\bar{\eta} - c^T U_\eta^{-1}\eta\big\} \cdot \{f(r) - (r+\rho^{-1})(\bar{f}_\mathrm{n}(r) + \bar{c}^T U_{\bar{\eta}}^{-1}\bar{\eta})\}\big] + o(1), \quad (13)$$

in which r denotes the density ratio $r(x) = r(x;\theta^*)$, and the functions in (13) are evaluated at $\theta = \theta^*$. See Appendix A for the computation of (13). Then, we study the sufficient condition that the above covariance vanishes.

Theorem 1 *Suppose Assumption 3 and Assumption 4 for the decomposition of f, and suppose that $\bar{f}_\mathrm{d}(r(x;\theta))$, $\bar{f}_\mathrm{n}(r(x;\theta))$ and $\bar{\eta}(x;\theta)$ satisfy*

$$f(r(x;\theta)) - (r(x;\theta)+\rho^{-1})(\bar{f}_\mathrm{n}(r(x;\theta)) + \bar{c}^T U_{\bar{\eta}}^{-1}\bar{\eta}(x;\theta)) \in \mathcal{L}[\nabla \log r(x;\theta)] \quad (14)$$

for all $\theta \in \Theta$. Then the estimator $\bar{D}_f$ using $\bar{\eta}(x;\theta)$ and the decomposition $f(r) = \bar{f}_\mathrm{d}(r) + r\bar{f}_\mathrm{n}(r)$ satisfies

$$\lim_{m\to\infty} m\mathbb{V}[\bar{D}_f] \le \lim_{m\to\infty} m\mathbb{V}[\widehat{D}_f],$$

that is, $\bar{D}_f$ uniformly attains the minimum asymptotic variance in terms of the f-divergence estimation.

Proof. Remember that $U_{\bar{\eta}} = \mathrm{E_n}[\bar{\eta}(x;\theta)\nabla \log r(x;\theta)^T]$. For any p_n and p_d such that $p_\mathrm{n}(x)/p_\mathrm{d}(x) = r(x;\theta)$, we have

$$\begin{aligned}
&\mathrm{E_n}\big[\big\{\bar{f}_\mathrm{n}(r) - f_\mathrm{n}(r) + \bar{c}^T U_{\bar{\eta}}^{-1}\bar{\eta} - c^T U_\eta^{-1}\eta\big\}\nabla \log r(x;\theta)^T\big] \\
&= \mathrm{E_n}\big[\{\bar{f}_\mathrm{n}(r(x;\theta)) - f_\mathrm{n}(r(x;\theta))\}\nabla \log r(x;\theta)^T\big] + \bar{c}^T U_{\bar{\eta}}^{-1}U_{\bar{\eta}} - c^T U_\eta^{-1}U_\eta \\
&= \mathrm{E_n}\big[\{\bar{f}_\mathrm{n}(r(x;\theta)) - f_\mathrm{n}(r(x;\theta))\}\nabla \log r(x;\theta)^T\big] + \bar{c}^T - c^T \\
&= 0.
\end{aligned}$$

Hence, when (14) holds, we have

$$m(1+\rho^{-1}) \cdot \mathrm{Cov}[\,\widehat{D}_f - \bar{D}_f, \bar{D}_f\,] = o(1)$$

for any $\widehat{D}_f$.

Intuitively, the meaning of Eq. (14) is related to the inference of probability distributions. As shown in [24, Section 5], for the probabilistic model $p(x;\theta)$, the Z-estimator using $\eta(x;\theta)$ is available to estimate the parameter θ, i.e., the estimator is given by

zero of the empirical mean of $\eta(x;\theta)$. In this case, the minimum asymptotic variance of the parameter estimation is achieved when $\eta(x;\theta) \in \mathcal{L}[\nabla \log p(x;\theta)]$ holds. Note that in the exponential family, $\nabla \log p(x;\theta)$ leads to the sufficient statistics. In other words, $\nabla \log p(x;\theta)$ implies the most informative direction for the parameter estimation. In the density ratio estimation, $\nabla \log r(x;\theta)$ corresponds to the score function $\nabla \log p(x;\theta)$. The equality

$$f(r) - (r + \rho^{-1})(\bar{f}_{\mathrm{n}}(r) + \bar{c}U_{\bar{\eta}}^{-1}\bar{\eta}) = \{f_{\mathrm{d}}(r) - r\bar{c}U_{\bar{\eta}}^{-1}\bar{\eta}\} - \rho^{-1}\{f_{\mathrm{n}}(r) + \bar{c}U_{\bar{\eta}}^{-1}\bar{\eta}\}$$

implies that an estimator of D_f is asymptotically given as $\widehat{D}_f = \widehat{E}_{\mathrm{de}}[f_{\mathrm{d}}(r) - r\bar{c}U_{\bar{\eta}}^{-1}\bar{\eta}] + \widehat{E}_{\mathrm{nu}}[f_{\mathrm{n}}(r) - \bar{c}U_{\bar{\eta}}^{-1}\bar{\eta}]$, where $\widehat{E}_{\mathrm{de}}$ (resp. $\widehat{E}_{\mathrm{de}}$) is the empirical expectation over the samples $x_1^{(\mathrm{d})}, \ldots, x_{m_{\mathrm{d}}}^{(\mathrm{d})}$ ($x_1^{(\mathrm{n})}, \ldots, x_{m_{\mathrm{n}}}^{(\mathrm{n})}$)[1]. Similarly to the estimation of probabilities, the minimum asymptotic variance of $\widehat{D}_f$ is achieved when the estimation function is parallel to the most informative direction $\mathcal{L}[\nabla \log r(x;\theta)]$.

In the following corollaries, we present some sufficient conditions for (14).

Corollary 1 *Under Assumption 3 and Assumption 4, suppose that, for $r = r(x;\theta)$,*

$$f(r) - (r + \rho^{-1})\bar{f}_{\mathrm{n}}(r) \in \mathcal{L}[\nabla \log r] \tag{15}$$

holds for all $\theta \in \Theta$. Then, the function $\bar{\eta} = \eta_{\mathrm{opt}}$ defined in (3) *and the decomposition $f(r) = \bar{f}_{\mathrm{d}}(r) + r\bar{f}_{\mathrm{n}}(r)$ satisfy the condition* (14)*.*

Proof. We see that the condition (15) and the equality $(r + \rho^{-1})\eta_{\mathrm{opt}} = \rho^{-1}\nabla \log r$ assure the condition (14).
Based on Corollary 1 we see that the estimator defined from

$$f_{\mathrm{d}}(r) = \frac{f(r)}{1 + \rho r},\ f_{\mathrm{n}}(r) = \frac{\rho f(r)}{1 + \rho r},\ \eta(x;\theta) = \eta_{\mathrm{opt}}(x;\theta) \tag{16}$$

leads to an optimal estimator of the f-divergence. In the optimal estimator, the function f is decomposed according to the ratio of the logistic model, $1/(1 + \rho r)$ and $\rho r/(1 + \rho r)$.

We show another sufficient condition.

Corollary 2 *Under Assumption 3 and Assumption 4, suppose that for $r = r(x;\theta)$ and $\bar{\eta} = \bar{\eta}(x;\theta)$,*

$$f(r) - (r + \rho^{-1})f'(r) \in \mathcal{L}[\nabla \log r]$$

and

$$f'(r) - \bar{f}_{\mathrm{n}}(r) \in \mathcal{L}[\bar{\eta}]$$

hold for all $\theta \in \Theta$. Then, the decomposition $f(r) = \bar{f}_{\mathrm{d}}(r) + r\bar{f}_{\mathrm{n}}(r)$ and the vector-valued function $\bar{\eta}(x;\theta)$ satisfy (14)*.*

[1]The equality $m(1 + \rho^{-1}) \cdot \mathrm{Cov}(\widehat{D}_f - \bar{D}_f,\ \widehat{E}_{\mathrm{nu}}[g_{\mathrm{nu}}] + \widehat{E}_{\mathrm{de}}[g_{\mathrm{de}}]) = -\mathrm{E}_{\mathrm{n}}[Z \cdot (g_{\mathrm{de}} - \rho^{-1}g_{\mathrm{nu}})]$ holds for any g_{de} and g_{nu}. We omit the definition of the random variable Z.

Table 1: Mean square errors of KL-divergence estimators are shown. $p_{\rm n}$ ($p_{\rm d}$) is the probability density of $N(0,1)$ (resp. $N(\mu,1)$).

	mean square error: $m\mathbb{E}[(\widehat{D}_f - D_f)^2]$ ($m_{\rm d} = m_{\rm n} = 50$)									
μ	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
optimal estimator	0.038	0.143	0.299	0.685	1.119	1.808	2.502	3.642	5.276	8.354
conjugate representation	0.040	0.171	0.381	0.893	1.480	2.723	4.125	6.329	9.553	16.830

Proof. When $f'(r(x;\theta)) - \bar{f}_{\rm n}(r(x;\theta)) \in \mathcal{L}[\bar{\eta}(x;\theta)]$ holds, there exists a vector $b \in \mathbb{R}^d$ such that $f'(r(x;\theta)) - \bar{f}_{\rm n}(r(x;\theta)) = b^T\bar{\eta}(x;\theta)$. Remember that $\bar{c}(\theta) = {\rm E}_{\rm n}[\{f'(r) - \bar{f}_{\rm n}(r)\}\nabla \log r]$ and $U_{\bar{\eta}}(\theta) = {\rm E}_{\rm n}[\bar{\eta}\nabla \log r^T]$. Thus, $\bar{c}^T U_{\bar{\eta}}^{-1} = {\rm E}_{\rm n}[b^T\bar{\eta}(\nabla \log r)^T]{\rm E}_{\rm n}[\bar{\eta}(\nabla \log r)^T]^{-1} = b^T$ holds. Hence, we have $\bar{c}^T U_{\bar{\eta}}^{-1}\bar{\eta}(x;\theta) = b^T\bar{\eta}(x;\theta) = f'(r(x;\theta)) - \bar{f}_{\rm n}(r(x;\theta))$, and we can confirm that (14) holds under the assumption.

We consider the decomposition derived from the conjugate representation, $f(r) = -f^*(f'(r)) + rf'(r)$, that is, $f_{\rm d}(r) = -f^*(f'(r))$ and $f_{\rm n}(r) = f'(r)$, where f^* is the conjugate function of f. For the conjugate representation, the second condition in Corollary 2 is always satisfied, since $f'(r) - f_{\rm n}(r) = 0$ holds. Then, the decomposition based on the conjugate representation leads to an optimal estimator when the model $r = r(x;\theta)$ and the function f satisfy

$$f(r) - (r + \rho^{-1})f'(r) \in \mathcal{L}[\nabla \log r]. \tag{17}$$

Later, we show some examples.

We compare the decomposition (16) and that defined from the conjugate representation (9). As shown above, the conjugate representation leads to an optimal estimator if (17) holds. However, there exists a pair of function f and model $r(x;\theta)$ which does not meet (17) as shown in Example 1 below. In this case, the optimality of the estimator based on the conjugate representation is not guaranteed. On the other hand, the decomposition (16) always leads to an optimal estimator without specific conditions on $f(r)$ and $r(x;\theta)$, as long as the argument on the asymptotic expansion is valid.

We show some examples in which Corollary 1 or Corollary 2 is applicable to construct the optimal estimator.

Example 1 *Let the model be* $r(x;\theta) = \exp\{\theta^T\phi(x)\}$, $\theta \in \mathbb{R}^d$ *with* $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))^T$ *and* $\phi_1(x) = 1$. *Then* $\mathcal{L}[\nabla \log r(x;\theta)]$ *is spanned by* $1, \phi_2(x), \ldots, \phi_d(x)$. *The f-divergence with* $f(r) = -\log r + r - 1$ *leads to the KL-divergence. Let* $f_{\rm d}(r) = -\log r - 1$ *and* $f_{\rm n}(r) = 1$, *then we can confirm that* (15) *is satisfied. Hence, the function* $\eta = \eta_{\rm opt}$ *and the decomposition* $f_{\rm d}(r) = -\log r - 1$ *and* $f_{\rm n}(r) = 1$ *lead to an optimal estimator of the KL-divergence. We see that there is redundancy for the decomposition of f. Indeed, for any constants* $c_0, c_1 \in \mathbb{R}$, *the function* $c_0 + c_1 \log r(x;\theta)$ *is included in* $\mathcal{L}[\nabla \log r(x;\theta)]$. *Hence the decomposition*

$$f_{\rm n}(r) = \frac{r + c_1 \log r + c_0}{r + \rho^{-1}}, \;\; f_{\rm d}(r) = r - \log r - 1 - rf_{\rm n}(r)$$

with $\bar{\eta} = \eta_{\mathrm{opt}}$ *also leads to an optimal estimator. The decomposition in* (16) *is realized by setting* $c_0 = -1$, $c_1 = -1$. *Next, we consider the conjugate expression of the KL-divergence. For* $f(r) = -\log r + r - 1$ *and* $r(x;\theta) = \exp\{\theta^T\phi(x)\}$, *we have* $f(r(x;\theta)) - (r(x;\theta) + \rho^{-1})f'(r(x;\theta)) = -\theta^T\phi(x) - \rho^{-1} + \rho^{-1}\exp\{-\theta^T\phi(x)\}$. *In general, the function* $\exp\{-\theta^T\phi(x)\}$ *is not represented by the linear combination of* $\phi_1(x), \ldots, \phi_d(x)$, *and thus, the condition in Corollary 2 will not hold. Therefore, the conjugate expression of the KL-divergence is not optimal in general.*

Table 1 shows numerical results of the estimation of the KL-divergence between $N(\mu, 1)$ and $N(0, 1)$. Under the model $r(x;\theta) = \exp\{\alpha + \beta x\}$, $\theta = (\alpha, \beta) \in \mathbb{R}^2$, we compare the optimal estimator using (16) and the estimator defined from the conjugate representation of the KL-divergence. The sample size is set to $m_{\mathrm{d}} = m_{\mathrm{n}} = 50$, and the averaged values of the square error $m(\widehat{D}_f - D_f)^2$ over 1000 runs are presented for several μ. We see that the optimal estimator outperforms the estimator using the conjugate representation.

Example 2 *Let the model be* $r(x;\theta) = \exp\{\theta^T\phi(x)\}$, $\theta \in \mathbb{R}^d$ *with* $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))^T$ *and* $\phi_1(x) = 1$. *Then, the linear space* $\mathcal{L}[\nabla \log r(x;\theta)]$ *is spanned by* $\{\phi_1(x), \ldots, \phi_d(x)\}$ *and thus* $\mathcal{L}[\nabla \log r(x;\theta)]$ *includes the function of the form* $c_0 + c_1 \log r(x;\theta)$ *for* $c_0, c_1 \in \mathbb{R}$. *Let the convex function* $f(r)$ *be*

$$f(r) = \frac{1}{1+\rho}\log\frac{1+\rho}{1+\rho r} + r\,\frac{\rho}{1+\rho}\log\frac{r(1+\rho)}{1+\rho r} \tag{18}$$

for $\rho > 0$. *Then the corresponding* f*-divergence is reduced to mutual information:*

$$\int p_{\mathrm{d}}(x) f\left(\frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)}\right) dx = \int \sum_{y=\mathrm{n},\mathrm{d}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dx,$$

in which y *is the binary random variable taking "n" or "d"; the joint probability of* x *and* y *is defined as* $p(x,\mathrm{n}) = p_{\mathrm{n}}(x)\frac{\rho}{1+\rho}$ *and* $p(x,\mathrm{d}) = p_{\mathrm{d}}(x)\frac{1}{1+\rho}$. *The equality* $p_{\mathrm{d}} = p_{\mathrm{n}}$ *implies that the conditional probability* $p(x|y)$ *is independent of* y. *Thus, mutual information becomes zero if and only if* $p_{\mathrm{d}} = p_{\mathrm{n}}$ *holds. For any moment matching estimator, we can confirm that the following decomposition satisfies the condition in Corollary 2:*

$$f_{\mathrm{d}}(r) = \frac{1}{1+\rho}\log\frac{1+\rho}{1+\rho r},\ f_{\mathrm{n}}(r) = \frac{\rho}{1+\rho}\log\frac{r(1+\rho)}{1+\rho r}. \tag{19}$$

Note that the above decomposition with the model $r(x;\theta) = \exp\{\theta^T\phi(x)\}$ *also satisfies the condition in Corollary 1. As pointed out in [14, 15], the decomposition above is derived from the conjugate expression of* (18). *In this example, we present another characterization, that is, an optimal estimator for mutual information.*

Example 3 *Let* $r(x;\theta) = 1 + \theta^T\phi(x)$ *and* $\phi_1(x) = 1$. *The subspace* $\mathcal{L}[\nabla \log r(x;\theta)]$ *is spanned by* $\{\phi_1/r, \ldots, \phi_d/r\}$, *and thus* $\mathcal{L}[\nabla \log r(x;\theta)]$ *includes the function of the form* $c_0 + c_1/r(x;\theta)$ *for* $c_0, c_1 \in \mathbb{R}$. *Let the convex function* f *be* $f(r) =$

$\frac{1}{\rho+1}\left(r-1+(1+\rho r)\log\frac{1+\rho r}{r(1+\rho)}\right)$ *for* $\rho>0$. *Then the corresponding* f*-divergence is expressed as*

$$\int p_{\mathrm{d}}(x) f\left(\frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)}\right) dx = \mathrm{KL}\left(\frac{p_{\mathrm{d}}+\rho\, p_{\mathrm{n}}}{1+\rho}, p_{\mathrm{n}}\right),$$

where KL *is the Kullback-Leibler divergence. Corollary 1 assures that the decomposition* $f_{\mathrm{d}}(r)=\frac{1}{\rho+1}\left(\frac{r-1}{1+\rho r}+\log\frac{1+\rho r}{r(1+\rho)}\right)$, $f_{\mathrm{n}}(r)=\frac{\rho}{\rho+1}\left(\frac{r-1}{1+\rho r}+\log\frac{1+\rho r}{r(1+\rho)}\right)$ *and the moment matching estimator using* $\eta=\eta_{\mathrm{opt}}$ *lead to an optimal estimator for the above* f*-divergence. On the other hand, due to Corollary 2, we can confirm that the decomposition derived from the conjugate expression,* $f_{\mathrm{d}}(r)=\frac{1}{1+\rho}\log\frac{1+\rho r}{r(1+\rho)}$, $f_{\mathrm{n}}(r)=f'(r)=\frac{1}{r(1+\rho)}\left(r-1+\rho r\log\frac{1+\rho r}{r(1+\rho)}\right)$ *leads to another optimal estimator.*

5 Homogeneity Test Exploiting f-divergence Estimators

For the homogeneity test of p_{n} and p_{d}, we need to know the asymptotic distribution of $\widehat{D}_f$ under the null hypothesis, $H_0: p_{\mathrm{n}}=p_{\mathrm{d}}$ in (12). In this section, we assume $\frac{p_{\mathrm{n}}(x)}{p_{\mathrm{d}}(x)}=r(x;\theta^*)=1$. We consider the optimal estimator $\widehat{D}_f$ defined from (16). The asymptotic distribution of the optimal estimator is given by the following theorem.

Theorem 2 *Let* $p_{\mathrm{n}}(x)/p_{\mathrm{d}}(x)=r(x;\theta^*)=1$. *Suppose that Assumption 3 and Assumption 4 hold, and that in the vicinity of* θ^*, *the third-order derivatives of* $f_{\mathrm{n}}(r(x;\theta))$ *and* $f_{\mathrm{d}}(r(x;\theta))$ *with respect to* θ *are dominated by a* p_{d}*-integrable function. We assume that the* d *by* d *symmetric matrix* $U_\eta=\mathrm{E}_{\mathrm{n}}[\eta(\nabla\log r)^T]$ *with* $\eta=\eta_{\mathrm{opt}}=\frac{1}{1+\rho r}\nabla\log r$ *is non-degenerate in the vicinity of* $\theta=\theta^*$. *Let* $\widehat{D}_f$ *be the estimator defined from* (16). *Then, in terms of the asymptotic distribution of* $\widehat{D}_f$, *we obtain* $\frac{2m}{f''(1)}\widehat{D}_f \xrightarrow{d} \chi^2_{d-1}$, *where* χ^2_k *is the chi-square distribution with* k *degrees of freedom.*

The proof is deferred to Appendix B. For the homogeneity test of p_{n} and p_{d}, the null hypothesis $p_{\mathrm{n}}=p_{\mathrm{d}}$ is rejected if

$$\widehat{D}_f \geq \frac{f''(1)}{2m}\chi^2_{d-1}(1-\alpha) \tag{20}$$

is satisfied, where $\chi^2_{d-1}(1-\alpha)$ is the chi-square $100(1-\alpha)$ percent point function with $d-1$ degrees of freedom. The homogeneity test based on (20) with the optimal choice (16) is referred to as $\widehat{D}_f$*-based test*.

We consider the power function of the homogeneity test, and compare the proposed method to the other method. A standard approach for the homogeneity test is exploiting the asymptotic distribution of the empirical likelihood estimator $\widehat{\theta}$. Under the model

$$r(x;\theta)=\exp\{\alpha+\phi(x;\beta)\},\ \theta=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{d-1}, \tag{21}$$

Fokianos et al. [7] pointed out that the asymptotic distribution of the empirical likelihood estimator $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta}) \in \mathbb{R} \times \mathbb{R}^{d-1}$ under the null hypothesis $p_{\mathrm{n}} = p_{\mathrm{d}}$ is given as

$$\sqrt{m}(\widehat{\beta} - \beta^*) \xrightarrow{d} N_{d-1}(0, \mathrm{V}_{\mathrm{n}}[\nabla_\beta \phi]^{-1}),$$

where $\theta^* = (\alpha^*, \beta^*)$ and $\nabla_\beta \phi$ is the $d-1$ dimensional gradient vector of $\phi(x;\beta)$ at $\beta = \beta^*$ with respect to the parameter β. Then the null hypothesis is rejected if the Wald-type test statistic

$$S = m(\widehat{\beta} - \beta^*)^T \widehat{V}_{\mathrm{n}}[\nabla_\beta \phi](\widehat{\beta} - \beta^*) \tag{22}$$

is larger than $\chi^2_{d-1}(1-\alpha)$, where $\widehat{V}_{\mathrm{n}}[\nabla_\beta \phi]$ is a consistent estimator of $\mathrm{V}_{\mathrm{n}}[\nabla_\beta \phi]$. In this paper, the homogeneity test based on the statistic S is referred to as the *empirical likelihood score test*. Fokianos et al. [7] studied statistical properties of the empirical likelihood score test through numerical experiments, and reported that the power of the empirical likelihood score test is comparable to the standard t-test and F-test which are available under parametric models.

Alternative test statistic is given by the empirical likelihood ratio test. Let $\ell(\theta)$ be

$$\ell(\theta) = \sum_{i=1}^{m_{\mathrm{d}}} \log \frac{1}{\rho r(x_i^{(\mathrm{d})};\theta) + 1} + \sum_{j=1}^{m_{\mathrm{n}}} \log \frac{\rho r(x_j^{(\mathrm{n})};\theta)}{\rho r(x_j^{(\mathrm{n})};\theta) + 1}.$$

Then, the empirical likelihood ratio test uses the statistic,

$$R = 2 \max_{\theta \in \Theta} \{\ell(\theta) - \ell(\theta^*)\}. \tag{23}$$

It is shown that under the null hypothesis $r(x) = 1$, the statistic, R, converges in distribution to the chi-square distribution with $d-1$ degrees of freedom [15]. The empirical likelihood ratio test is closely related to the mutual information shown in Example 2. Indeed, the estimator of the mutual information derived from (6) and (18) satisfies $R = 2(m_{\mathrm{n}} + m_{\mathrm{d}})\widehat{D}_f$. This formula has been pointed out in [15]. Our argument in Example 2 guarantees that this estimator attains the minimum asymptotic variance for the estimation of mutual information.

We show that the power of the $\widehat{D}_f$-based test is not less than that of the empirical likelihood score test under the setup of local alternative, where the distributions p_{n} and p_{d} vary according to the sample size. To compute the power of the test, we assume the following conditions.

Assumption 5

1. *f is third-order differentiable, and a strictly convex function satisfying $f(1) = f'(1) = 0$.*
2. *The density ratio model $r(x;\theta)$ is represented as (21), and suppose the true density ratio is given as $r(x;\theta^*) = 1$. Note that for any $\theta \in \Theta$, $1 \in \mathcal{L}[\nabla \log r(x;\theta)]$ holds.*

3. *In the vicinity of θ^*, the third-order derivatives of $f_{\mathrm{n}}(r(x;\theta))$ and $f_{\mathrm{d}}(r(x;\theta))$ with respect to θ are dominated by a p_{d}-integrable function.*

4. *For a fixed probability density $p(x)$, we assume $p_{\mathrm{d}}(x) = p(x)$. The probability density $p_{\mathrm{n}}^{(m)}$ is represented as $p_{\mathrm{n}}^{(m)}(x) = p_{\mathrm{d}}(x)r(x;\theta_m)$, where the parameter θ_m is defined as $\theta_m = \theta^* + h_m/\sqrt{m}$ for $h_m \in \mathbb{R}^d$ satisfying $\lim_{m\to\infty} h_m = h \in \mathbb{R}^d$.*

5. *The matrix-valued functions $M(\theta)$ and $U(\theta)$ are defined as*

$$M(\theta) = \mathrm{E}_{\mathrm{d}}[\nabla \log r(x;\theta) \nabla \log r(x;\theta)^T],$$
$$U(\theta) = \mathrm{E}_{\mathrm{d}}[\frac{1}{1+\rho r(x;\theta)} \nabla \log r(x;\theta) \nabla \log r(x;\theta)^T].$$

   We assume that $M(\theta)$ and $U(\theta)$ are continuous and non-degenerate in the vicinity of θ^*.

6. *Let $V[\nabla r]$ be the variance-covariance matrix of $\nabla \log r(x;\theta^*) = \nabla r(x;\theta^*)$ under $p_{\mathrm{d}}(x)(= p(x))$. We assume*

$$\frac{1}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} \nabla r(x_j^{(\mathrm{n})};\theta^*) \nabla r(x_j^{(\mathrm{n})};\theta^*)^T \xrightarrow{p} M(\theta^*), \tag{24}$$

$$\sqrt{m}\, U(\theta_m)(\widehat{\theta} - \theta_m) \xrightarrow{d} N\big(0, \frac{1}{(1+\rho)^2} V[\nabla r]\big), \tag{25}$$

   when m tends to infinity.

Note that (25) is reduced to (2) with $\eta = \eta_{\mathrm{opt}}$, when $h_m = 0 \in \mathbb{R}^d$ holds. See [24, Section 14] and [18, Section 11.4.2] for details of the asymptotic theory when the probability distribution depends on the sample size.

In the above, one can make the assumption weaker such that the probability p_{d} also varies according to the sample size. We adopt the simplified assumption to avoid technical difficulties.

Theorem 3 *Under Assumption 5, the power function of the $\widehat{D}_f$-based test with significance level α is asymptotically given as $\Pr\{Y \geq \chi^2_{d-1}(1-\alpha)\}$, where Y is the random variable whose distribution function is the non-central chi-square distribution with $d-1$ degrees of freedom and non-centrality parameter $h^T M(\theta^*) h$. Moreover, the asymptotic power function of the empirical likelihood score test* (22) *is the same.*

The proof is given in Appendix C. Theorem 3 implies that, under the local alternative, the power function of the $\widehat{D}_f$-based test does not depend on choice of the f-divergence, and that the empirical likelihood score test has the same power as the $\widehat{D}_f$-based test.

Next, we consider the power function under the misspecification case.

Theorem 4 *We assume that the density ratio $p_{\mathrm{n}}^{(m)}/p_{\mathrm{d}}$ is not realized by the model $r(x;\theta)$, and that $p_{\mathrm{n}}^{(m)}$ is represented as $p_{\mathrm{n}}^{(m)}(x) = p_{\mathrm{d}}(x)\left(r(x;\theta_m) + \frac{s_m(x)+\varepsilon_m}{\sqrt{m}}\right)$, where $s_m(x)$ satisfies $\mathrm{E}_{\mathrm{d}}[s_m(x)] = 0$. Suppose that $\lim_{m\to\infty}\varepsilon_m = \varepsilon$ holds. Suppose Assumption 5 except the definition of $p_{\mathrm{n}}^{(m)}(x)$. Then, under the setup of the local alternative, the power function of the $\widehat{D}_f$-based test is larger than or equal to that of the empirical likelihood score test.*

The proof is given in Appendix D. Even in the misspecification case, the assumptions (24) and (25) is valid, since eventually the limit of $p_{\mathrm{n}}^{(m)}/p_{\mathrm{d}}$ is realized by the model $r(x;\theta^*) = 1$. In [18, Section 11.4.2], detailed explanation on the asymptotic theory under local alternative is presented. Theorem 3 and Theorem 4 indicate that the $\widehat{D}_f$-based test is more powerful than the empirical likelihood score test regardless of whether the model $r(x;\theta)$ is correct or slightly misspecified.

6 Numerical Studies

In this section, we report numerical results for illustrating the adequacy of the asymptotic theory for finite-sample inference.

We examine two f-divergences for the homogeneity test. One is the KL-divergence defined by $f(r) = r - 1 - \log(r)$ as shown in Example 1, and the test statistic is derived from (16). This is referred to as the *KL-based test*. The other is mutual information defined by (18), and the estimator $\widehat{D}_f$ is derived from the optimal decomposition (16) and the moment matching estimator $\eta = \eta_{\mathrm{opt}}$. This is referred to as the *MI-based test*. The *empirical likelihood ratio test* (23) is also applied. As we mentioned, the statistic of the empirical likelihood ratio test is equivalent to the estimator of mutual information using the conjugate representation (19) and the moment matching estimator $\eta = \eta_{\mathrm{opt}}$. This fact has been investigated by Keziou and Leoni-Aubin [14, 15]. The MI-based test and empirical likelihood ratio test share the same moment matching estimator using η_{opt}, and the difference comes from the way of decomposing the function f. In the following, we observe that the MI-based test and empirical likelihood ratio test provide almost the same results. We compare these methods to the empirical likelihood score test (22) proposed in [7], and the Hotelling T^2-test. The null hypothesis of the test is $H_0 : p_{\mathrm{n}} = p_{\mathrm{d}}$ and the alternative is $H_1 : p_{\mathrm{n}} \neq p_{\mathrm{d}}$. The type-I error and the power function of these tests are computed.

First we assume that the null hypothesis $p_{\mathrm{n}} = p_{\mathrm{d}}$ is correct, and we compute the type-I error. We consider three cases: in the first case, the distributions of p_{n} and p_{d} are given as the one-dimensional exponential distribution with rate parameter $\lambda = 0.1, 1$ or 5; in the second case, the distributions of p_{n} and p_{d} are given as the 10-dimensional normal distribution $N_{10}(0, I_{10})$; and in the third case, each element of the 10-dimensional vector $x \in \mathbb{R}^{10}$ is independent and identically distributed from the t-distribution with 10 or 5 degrees of freedom. For the k-dimensional vector $x = (x_1, \ldots, x_k)$, the semiparametric

model for density ratio is defined as

$$r(x;\theta) = \exp\left\{\alpha + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{k} \beta_{k+j}\, x_j^2\right\} \tag{26}$$

with the $(2k+1)$-dimensional parameter $\theta = (\alpha, \beta_1, \ldots, \beta_{2k})$. The sample size is set to $m_{\mathrm{n}} = m_{\mathrm{d}}$ and varies from 10 to 100 for one-dimensional random variables and from 100 to 1000 for 10-dimensional random variables. The significance level of the test is set to 0.05, and the type-I errors are averaged over 1000 runs. For each case, the averaged type-I errors of the KL-based test, MI-based test, empirical likelihood ratio test, and empirical likelihood score test are shown in Table 2. In the exponential distribution, the type-I error of empirical likelihood score test is larger than the significance level even with large sample size. On the other hand, the type-I errors of the KL-based test, MI-based test and empirical likelihood ratio test are close to the significance level for large sample size. In addition, the type-I error comes closer to the significance level when the rate parameter λ becomes bigger. This is because large λ corresponds to small variance, and hence, the estimation accuracy of the f-divergence is high for large λ. In the normal case, all of the type-I errors converge to the significance level with modest sample size. In the case of the t-distribution, the type-I error of the empirical likelihood score test is larger than the significance level even with large sample size. On the other hand, the type-I errors of the other tests are close to the significance level with moderate sample size even for the t-distribution.

Next, we compute the power function of the KL-based test, MI-based test, empirical likelihood ratio test, empirical likelihood score test, and Hotelling T^2-test. In the numerical simulations, $p_{\mathrm{n}}(x)$ is fixed and $p_{\mathrm{d}}(x)$ is varied by changing the parameters in the distribution such as the rate parameter, the mean parameter or the scale parameter. We consider the following three setups:

1. $p_{\mathrm{n}}(x)$ is given as the one-dimensional exponential distribution with rate parameter $\lambda_{\mathrm{nu}} = 0.1, 1$ or 5, and $p_{\mathrm{d}}(x)$ is also the one-dimensional exponential distribution with rate parameter λ_{de}. The parameter λ_{de} varies from $0.6 \cdot \lambda_{\mathrm{nu}}$ to $1.4 \cdot \lambda_{\mathrm{nu}}$. We use two density ratio models: one is given as $r(x;\theta) = \exp\{\alpha + \beta_1 x\}$, $\theta = (\alpha, \beta_1) \in \mathbb{R}^2$, and the other is given as $r(x;\theta) = \exp\{\alpha + \beta_1 x + \beta_2 x^2\}$, $\theta = (\alpha, \beta_1, \beta_2) \in \mathbb{R}^3$. The sample size is set to $m_{\mathrm{n}} = m_{\mathrm{d}} = 100$.

2. $p_{\mathrm{n}}(x)$ is defined as the 10-dimensional standard normal distribution, or the 10-dimensional t-distribution with 10 or 5 degrees of freedom. The sample $x^{(\mathrm{d})} = (x_1^{(\mathrm{d})}, \ldots, x_{10}^{(\mathrm{d})})$ from p_{d} is computed such that

$$x_\ell^{(\mathrm{d})} = x_\ell + \mu, \quad \ell = 1, \ldots, 10, \tag{27}$$

   where $x = (x_1, \ldots, x_{10}) \sim p_{\mathrm{n}}$, that is, the mean parameter $\mu \in \mathbb{R}$ is added to each element of x. Hence, $p_{\mathrm{n}} = p_{\mathrm{d}}$ holds for $\mu = 0$. The sample size is set to $m_{\mathrm{n}} = m_{\mathrm{d}} = 500$ or 1000, and the density ratio models (26) with $k = 10$ is used.

Table 2: Averaged Type-I errors over 1000 runs are shown as functions of the number of samples. The exponential distribution, normal distribution, t-distribution with 10 degrees of freedom (df), and t-distribution with 5 degrees of freedom are examined as p_{n} and p_{d}. In the table, "KL", "MI", "ratio", and "score" denote the KL-based test, MI-based test, empirical likelihood ratio test, and empirical likelihood score test, respectively.

	1-dim exp. dist. ($\lambda = 0.1$)				1-dim exp. dist. ($\lambda = 1$)				1-dim exp. dist. ($\lambda = 5$)			
$m_{\mathrm{n}}(= m_{\mathrm{d}})$	KL	MI	ratio	score	KL	MI	ratio	score	KL	MI	ratio	score
10	0.177	0.084	0.084	0.206	0.170	0.083	0.083	0.203	0.178	0.067	0.079	0.218
50	0.095	0.067	0.067	0.124	0.099	0.068	0.068	0.134	0.109	0.078	0.080	0.139
100	0.085	0.059	0.059	0.104	0.092	0.067	0.067	0.107	0.079	0.054	0.057	0.102
	10-dim standard normal				10-dim. t-dist. (df=10)				10-dim. t-dist. (df=5)			
$m_{\mathrm{n}}(= m_{\mathrm{d}})$	KL	MI	ratio	score	KL	MI	ratio	score	KL	MI	ratio	score
100	0.152	0.102	0.102	0.200	0.179	0.092	0.092	0.244	0.254	0.104	0.104	0.384
500	0.058	0.051	0.051	0.059	0.084	0.073	0.073	0.114	0.099	0.056	0.056	0.189
1000	0.062	0.058	0.058	0.069	0.063	0.056	0.056	0.078	0.089	0.067	0.067	0.159

3. $p_{\mathrm{n}}(x)$ is given as the same distribution as the second setup, and the sample $x^{(\mathrm{d})} = (x_1^{(\mathrm{d})}, \ldots, x_{10}^{(\mathrm{d})})$ from p_{d} is computed such that

$$x_\ell^{(\mathrm{d})} = \sigma \times x_\ell, \quad \ell = 1, \ldots, 10, \tag{28}$$

where $x = (x_1, \ldots, x_{10}) \sim p_{\mathrm{n}}$, that is, the scale parameter $\sigma > 0$ is multiplied to each element of $x \in \mathbb{R}^{10}$. Hence, the null hypothesis $p_{\mathrm{n}} = p_{\mathrm{d}}$ corresponds to $\sigma = 1$. The sample size is set to $m_{\mathrm{n}} = m_{\mathrm{d}} = 500$ or 1000, and the density ratio models (26) with $k = 10$ is used.

In the first setup for the exponential distribution, both of the two density ratio models include the true density ratio. In the second and the third setups, when both p_{n} and p_{d} are the 10-dimensional normal distribution, the density ratio model (26) includes the true density ratio. For the t-distribution, however, the true ratio $r(x)$ resides outside of the model (26). In all simulations, the significance level is 0.05, and the power functions are averaged over 1000 runs.

Table 3 shows the averaged power functions for the first setup. Under the model $r(x;\theta) = \exp\{\alpha + \beta_1 x\}$, the power functions of all tests behave in a similar way, and are almost independent of the value of λ_{nu}. Under the larger model $r(x;\theta) = \exp\{\alpha + \beta_1 x + \beta_2 x^2\}$, however, except the Hotelling T^2-test, the power is slightly smaller than that under the smaller density ratio model.

Table 4 shows the averaged power functions for the setup (27). The mean parameter μ varies from -0.1 to 0.1. When both p_{n} and p_{d} are the normal distribution, the power functions of the KL-based test, MI-based test, empirical likelihood ratio test, and empirical likelihood score test almost coincide with each other. The power of the Hotelling T^2-test is slightly larger than the others. This result is obvious, since the Hotelling T^2-test works well under the normal distribution. Under the t-distribution with 5 degree of freedom, the power of empirical likelihood score test around $\mu = 0$ is much larger than the significance level, 0.05. That is, the empirical likelihood score test is not conservative, and will lead false positive with high probability. In the MI-based test and empirical likelihood ratio test, the power around $\mu = 0$ is close to the significance level and the power is comparable to the Hotelling T^2-test outside of the vicinity of $\mu = 0$.

Table 5 shows the averaged power functions when the scale parameter σ in (28) varies from 0.9 to 1.1. In this case, the means of p_{n} and p_{d} are the same, and hence the Hotelling T^2-test fails to detect the difference between p_{n} and p_{d}. In addition, we see that the power function of the empirical likelihood score test is biased, that is, the power function takes the minimum value at σ less than 1. This is because the estimated variance, $\widehat{V}_{\mathrm{n}}$, based on the empirical likelihood score estimator tends to take slightly small values than the true variance. In the MI-based test and empirical likelihood ratio test, the power around $\sigma = 1$ is close to the significance level, while the power of the KL-based test is slightly larger than the significance level around $\sigma = 1$.

In the first numerical study on the exponential distribution, a test with the smaller density ratio model performs better than that using the larger model. We see that the power of the Hotelling T^2-test is high, since the Hotelling T^2-test detects the difference

in the mean value which is directly connected to the rate parameter of the exponential distribution. As shown in the numerical results of the second and the third setups, when the model $r(x;\theta)$ is correct, the powers of the KL-based test, MI-based test, empirical likelihood ratio test, and empirical likelihood score test are almost the same. Thus, the numerical simulations meet the theoretical results in Theorem 3. The empirical likelihood score test has large type-I error and the power is slightly biased especially when the samples are generated from the t-distribution. Throughout the simulations, the MI-based test and empirical likelihood ratio test have comparable power to the other methods, while the type-I error is well controlled. In the simulations, we see that the null distribution of the MI-based test and that of the empirical likelihood ratio test are approximated by the asymptotic distribution more accurately than that of the KL-based test, although the first-order asymptotic theory provided in Section 5 does not explain the difference between the MI-based test and KL-based test. We expect that higher order asymptotic theory is needed to better understand the difference among f-divergences for the homogeneity test.

7 Conclusion

We have addressed inference methods of density ratios and their application to homogeneity test under the semiparametric models. We showed that the estimator introduced in [20] provides an optimal estimator of the f-divergence with appropriate decomposition of the function f, and proposed a test statistic for homogeneity test using the optimal f-divergence estimator. It is revealed that the power function of the $\widehat{D}_f$-based test does not depend on the choice of the f-divergence up to the first order under the local alternative setup. Additionally, the $\widehat{D}_f$-based test and empirical likelihood score test [7] were shown to have asymptotically the same power. For misspecified density-ratio models, we showed that the $\widehat{D}_f$-based test usually has greater power than the empirical likelihood score test. In numerical studies, the MI-based test and empirical likelihood ratio test provided the most reliable results than the others, that is, the null distribution was well approximated by the asymptotic distribution with moderate samples size, and the power was comparable to the Hotelling T^2-test even under the normal case.

The choice of the f-divergence is an important open problem for the homogeneity test. In our first-order asymptotic theory, the choice of the f-divergence does not affect the power function of the $\widehat{D}_f$-based test. Hence, higher order asymptotic theory may be necessary to make clear the difference among f-divergences for the homogeneity test.

In this paper, we considered the estimators of the form (11), while we can use a wider class of estimators for the inference of divergences. In terms of the class of f-divergence estimators, we have two challenging future works: one is to study the optimal estimator among *all* estimators of the f-divergence, and another is to specify how large the class of estimators (11) is among all estimators.

Table 3: Averaged power functions over 1000 runs are shown. The probability $p_{\mathrm{n}}(x)$ is given as the one-dimensional exponential distribution with rate $\lambda_{\mathrm{nu}} = 0.1$, 1, or 5, and the probability $p_{\mathrm{d}}(x)$ is the one-dimensional exponential distribution with rate λ_{de}. The parameter λ_{de} varies from $0.6 \cdot \lambda_{\mathrm{nu}}$ to $1.4 \cdot \lambda_{\mathrm{nu}}$. In the table, "KL", "MI","ratio","score", and "Hote." denote the KL-based test, MI-based test, empirical likelihood ratio test, empirical likelihood score test and Hotelling T^2-test, respectively.

$m_{\mathrm{n}} = m_{\mathrm{d}} = 100,\ r(x;\theta) = \exp\{\alpha + \beta_1 x\}$

	1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 0.1$)					1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 1$)					1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 5$)				
$\lambda_{\mathrm{de}}/\lambda_{\mathrm{nu}}$	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
0.6	0.961	0.954	0.954	0.971	0.950	0.952	0.937	0.937	0.959	0.929	0.958	0.949	0.948	0.969	0.947
0.7	0.770	0.733	0.733	0.815	0.726	0.769	0.731	0.731	0.806	0.720	0.741	0.702	0.702	0.790	0.694
0.8	0.366	0.331	0.331	0.423	0.321	0.405	0.358	0.358	0.457	0.345	0.377	0.339	0.339	0.422	0.323
0.9	0.145	0.126	0.125	0.167	0.118	0.146	0.126	0.126	0.175	0.123	0.131	0.119	0.118	0.166	0.111
1.0	0.050	0.048	0.048	0.048	0.047	0.049	0.048	0.048	0.050	0.045	0.058	0.055	0.054	0.056	0.051
1.1	0.096	0.107	0.107	0.062	0.103	0.093	0.107	0.107	0.067	0.100	0.104	0.111	0.111	0.073	0.110
1.2	0.242	0.264	0.264	0.179	0.257	0.241	0.259	0.259	0.185	0.253	0.227	0.249	0.244	0.166	0.237
1.3	0.431	0.455	0.455	0.334	0.442	0.442	0.466	0.466	0.351	0.458	0.406	0.432	0.432	0.316	0.424
1.4	0.617	0.636	0.636	0.536	0.628	0.634	0.662	0.662	0.528	0.654	0.620	0.649	0.649	0.542	0.634

$m_{\mathrm{n}} = m_{\mathrm{d}} = 100,\ r(x;\theta) = \exp\{\alpha + \beta_1 x + \beta_2 x^2\}$

	1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 0.1$)					1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 1$)					1-dim exp. dist. ($\lambda_{\mathrm{nu}} = 5$)				
$\lambda_{\mathrm{de}}/\lambda_{\mathrm{nu}}$	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
0.6	0.936	0.907	0.906	0.952	0.945	0.943	0.916	0.916	0.958	0.944	0.942	0.913	0.914	0.954	0.942
0.7	0.686	0.616	0.615	0.749	0.685	0.682	0.590	0.591	0.755	0.684	0.696	0.619	0.621	0.764	0.683
0.8	0.393	0.302	0.300	0.467	0.347	0.381	0.290	0.290	0.456	0.336	0.383	0.294	0.296	0.458	0.342
0.9	0.159	0.105	0.105	0.224	0.119	0.149	0.088	0.089	0.216	0.103	0.186	0.118	0.119	0.261	0.124
1.0	0.092	0.069	0.069	0.113	0.047	0.094	0.062	0.062	0.105	0.054	0.082	0.059	0.061	0.105	0.059
1.1	0.090	0.093	0.092	0.080	0.101	0.078	0.092	0.092	0.060	0.091	0.083	0.086	0.090	0.063	0.108
1.2	0.169	0.197	0.197	0.101	0.250	0.178	0.209	0.209	0.106	0.259	0.179	0.204	0.209	0.094	0.256
1.3	0.330	0.376	0.376	0.183	0.439	0.337	0.384	0.384	0.168	0.442	0.312	0.355	0.355	0.156	0.430
1.4	0.482	0.529	0.529	0.286	0.623	0.477	0.526	0.526	0.277	0.597	0.488	0.551	0.551	0.275	0.657

Table 4: Averaged power functions over 1000 runs are shown as functions of the mean parameter of the probability $p_{\mathrm{d}}(x)$, where $p_{\mathrm{d}}(x)$ is defined by (27) through the probability p_{n}. The normal distribution, t-distribution with 10 degrees of freedom (df), and t-distribution with 5 degrees of freedom are examined as p_{n}. In the table, "KL", "MI","ratio","score", and "Hote." denote the KL-based test, MI-based test, empirical likelihood ratio test, empirical likelihood score test and Hotelling T^2-test, respectively.

$m_{\mathrm{n}} = m_{\mathrm{d}} = 500$

	10-dim. normal dist.					10-dim. t-dist. (df=10)					10-dim. t-dist. (df=5)				
μ	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
-0.10	0.905	0.900	0.900	0.908	0.955	0.818	0.794	0.794	0.826	0.903	0.707	0.671	0.671	0.740	0.726
-0.08	0.716	0.696	0.696	0.720	0.814	0.569	0.535	0.535	0.597	0.673	0.509	0.439	0.439	0.554	0.532
-0.06	0.386	0.358	0.358	0.389	0.480	0.344	0.311	0.311	0.357	0.369	0.348	0.273	0.273	0.413	0.298
-0.04	0.168	0.155	0.155	0.185	0.210	0.176	0.150	0.150	0.206	0.179	0.186	0.135	0.135	0.272	0.143
-0.02	0.112	0.096	0.096	0.123	0.105	0.088	0.066	0.066	0.111	0.056	0.126	0.083	0.083	0.224	0.085
0.00	0.071	0.061	0.061	0.080	0.046	0.074	0.058	0.058	0.102	0.048	0.118	0.076	0.076	0.203	0.056
0.02	0.089	0.081	0.081	0.106	0.080	0.098	0.079	0.079	0.117	0.082	0.132	0.087	0.087	0.233	0.062
0.04	0.187	0.161	0.161	0.203	0.206	0.170	0.147	0.147	0.200	0.173	0.183	0.130	0.130	0.257	0.137
0.06	0.421	0.395	0.395	0.429	0.508	0.346	0.311	0.312	0.370	0.390	0.335	0.274	0.274	0.391	0.294
0.08	0.709	0.686	0.686	0.706	0.812	0.586	0.553	0.553	0.582	0.664	0.511	0.449	0.449	0.559	0.523
0.10	0.875	0.861	0.861	0.873	0.945	0.820	0.806	0.806	0.831	0.891	0.724	0.676	0.676	0.733	0.781

$m_{\mathrm{n}} = m_{\mathrm{d}} = 1000$

	10-dim. normal dist.					10-dim. t-dist. (df=10)					10-dim. t-dist. (df=5)				
μ	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
-0.10	0.998	0.998	0.998	0.998	1.000	0.993	0.993	0.993	0.993	0.998	0.961	0.957	0.957	0.962	0.977
-0.08	0.971	0.969	0.969	0.969	0.990	0.901	0.890	0.890	0.900	0.956	0.814	0.796	0.796	0.833	0.873
-0.06	0.758	0.748	0.748	0.762	0.857	0.602	0.594	0.594	0.617	0.734	0.515	0.467	0.467	0.551	0.573
-0.04	0.334	0.311	0.312	0.341	0.412	0.255	0.240	0.240	0.275	0.324	0.263	0.214	0.214	0.328	0.283
-0.02	0.113	0.108	0.108	0.116	0.121	0.128	0.112	0.112	0.144	0.102	0.127	0.092	0.092	0.178	0.085
0.00	0.058	0.052	0.052	0.068	0.042	0.054	0.047	0.047	0.077	0.050	0.085	0.050	0.050	0.145	0.035
0.02	0.107	0.103	0.103	0.120	0.115	0.106	0.093	0.093	0.123	0.102	0.131	0.102	0.102	0.205	0.086
0.04	0.380	0.370	0.370	0.389	0.474	0.280	0.265	0.264	0.300	0.352	0.282	0.229	0.229	0.348	0.279
0.06	0.741	0.725	0.725	0.746	0.848	0.609	0.588	0.588	0.621	0.728	0.513	0.478	0.478	0.566	0.592
0.08	0.973	0.971	0.971	0.972	0.991	0.902	0.898	0.898	0.900	0.957	0.805	0.782	0.782	0.825	0.883
0.10	0.999	0.999	0.999	0.999	1.000	0.987	0.987	0.987	0.989	0.998	0.961	0.954	0.954	0.961	0.982

Table 5: Averaged power functions over 1000 runs are shown as functions of the scale parameter of the probability $p_{\mathrm{d}}(x)$, where $p_{\mathrm{d}}(x)$ is defined by (28) through the probability p_{n}. The normal distribution, t-distribution with 10 degrees of freedom (df), and t-distribution with 5 degrees of freedom are examined as p_{n}. In the table, "KL", "MI","ratio","score", and "Hote." denote the KL-based test, MI-based test, empirical likelihood ratio test, empirical likelihood score test and Hotelling T^2-test, respectively.

$m_{\mathrm{n}} = m_{\mathrm{d}} = 500$

	10-dim. normal dist.					10-dim. t-dist. (df=10)					10-dim. t-dist. (df=5)				
σ	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
0.90	0.998	0.999	0.999	0.995	0.051	0.981	0.993	0.993	0.924	0.053	0.819	0.870	0.870	0.484	0.054
0.92	0.960	0.967	0.967	0.905	0.069	0.814	0.848	0.848	0.642	0.046	0.560	0.638	0.638	0.250	0.061
0.94	0.725	0.761	0.761	0.578	0.037	0.508	0.562	0.563	0.277	0.062	0.265	0.329	0.329	0.117	0.048
0.96	0.319	0.356	0.356	0.206	0.053	0.217	0.255	0.255	0.111	0.056	0.159	0.188	0.188	0.103	0.054
0.98	0.088	0.093	0.093	0.063	0.047	0.088	0.090	0.090	0.066	0.056	0.113	0.109	0.109	0.121	0.051
1.00	0.068	0.057	0.057	0.069	0.054	0.067	0.055	0.055	0.100	0.031	0.129	0.076	0.076	0.208	0.052
1.02	0.156	0.118	0.118	0.217	0.057	0.132	0.091	0.091	0.214	0.051	0.213	0.097	0.097	0.374	0.051
1.04	0.404	0.315	0.315	0.512	0.060	0.307	0.213	0.213	0.463	0.051	0.331	0.177	0.177	0.544	0.052
1.06	0.762	0.687	0.687	0.837	0.059	0.596	0.462	0.462	0.746	0.047	0.544	0.314	0.314	0.749	0.053
1.08	0.975	0.955	0.955	0.991	0.051	0.882	0.789	0.789	0.949	0.053	0.740	0.531	0.531	0.898	0.039
1.10	0.999	0.997	0.997	0.999	0.044	0.988	0.967	0.967	0.997	0.047	0.903	0.762	0.762	0.964	0.044

$m_{\mathrm{n}} = m_{\mathrm{d}} = 1000$

	10-dim. normal dist.					10-dim. t-dist. (df=10)					10-dim. t-dist. (df=5)				
σ	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.	KL	MI	ratio	score	Hote.
0.90	1.000	1.000	1.000	1.000	0.049	1.000	1.000	1.000	1.000	0.047	0.989	0.990	0.990	0.929	0.060
0.92	1.000	1.000	1.000	1.000	0.052	0.996	0.997	0.997	0.993	0.033	0.885	0.914	0.914	0.657	0.036
0.94	0.975	0.980	0.980	0.963	0.056	0.856	0.889	0.889	0.748	0.060	0.546	0.625	0.625	0.289	0.051
0.96	0.652	0.685	0.686	0.527	0.046	0.398	0.467	0.467	0.265	0.054	0.249	0.309	0.309	0.134	0.048
0.98	0.147	0.170	0.170	0.096	0.049	0.128	0.136	0.136	0.083	0.065	0.081	0.091	0.091	0.079	0.051
1.00	0.050	0.049	0.049	0.063	0.047	0.075	0.065	0.065	0.088	0.053	0.098	0.067	0.067	0.153	0.048
1.02	0.186	0.144	0.144	0.243	0.037	0.181	0.124	0.124	0.270	0.061	0.172	0.086	0.086	0.310	0.056
1.04	0.670	0.604	0.605	0.735	0.049	0.525	0.419	0.419	0.666	0.047	0.428	0.268	0.268	0.628	0.054
1.06	0.979	0.971	0.971	0.987	0.061	0.890	0.858	0.858	0.950	0.059	0.739	0.574	0.574	0.864	0.049
1.08	1.000	1.000	1.000	1.000	0.058	0.998	0.997	0.997	1.000	0.051	0.932	0.850	0.850	0.979	0.054
1.10	1.000	1.000	1.000	1.000	0.047	1.000	1.000	1.000	1.000	0.053	0.990	0.975	0.975	0.997	0.053

8 Acknowledgements

The authors are grateful to Dr. Hironori Fujisawa and Dr. Masayuki Henmi of Institute of Statistical Mathematics, Dr. Fumiyasu Komaki of University of Tokyo, and the anonymous reviewers for their helpful comments. T. Kanamori was partially supported by Grant-in-Aid for Young Scientists (20700251), T. Suzuki was partially supported by Grant-in-Aid for Young Scientists (22700289), and M. Sugiyama was supported by SCAT, AOARD, and the JST PRESTO program.

A Asymptotic expansion of $\widehat{D}_f$

We make a supplementary statement on the asymptotic expansions of $\widehat{D}_f$ and $\bar{D}_f$ in (13). We consider the asymptotic expansion of the estimator (11). For $f(r) = f_{\mathrm{d}}(r) + r f_{\mathrm{n}}(r)$, let us define $\mathbb{G}f$ be

$$\mathbb{G}f = \frac{\sqrt{m}}{m_{\mathrm{d}}} \sum_{i=1}^{m_{\mathrm{d}}} \left[f_{\mathrm{d}}(r(x_i^{(\mathrm{d})}; \theta^*)) - \mathrm{E}_{\mathrm{d}}[f_{\mathrm{d}}(r(x; \theta^*))] \right] + \frac{\sqrt{m}}{m_{\mathrm{n}}} \sum_{j=1}^{m_{\mathrm{n}}} \left[f_{\mathrm{n}}(r(x_j^{(\mathrm{n})}; \theta^*)) - \mathrm{E}_{\mathrm{n}}[f_{\mathrm{n}}(r(x; \theta^*))] \right],$$

and $\bar{\mathbb{G}}f$ is also defined in the same way for the other decomposition $f(r) = \bar{f}_{\mathrm{d}}(r) + r\bar{f}_{\mathrm{n}}(r)$. Remember that

$$c = \mathrm{E}_{\mathrm{n}}\left[\{ f'(r(x; \theta^*)) - f_{\mathrm{n}}(r(x; \theta^*)) \} \nabla \log r(x; \theta^*) \right],$$
$$\bar{c} = \mathrm{E}_{\mathrm{n}}\left[\{ f'(r(x; \theta^*)) - \bar{f}_{\mathrm{n}}(r(x; \theta^*)) \} \nabla \log r(x; \theta^*) \right].$$

The Taylor expansion around $\theta = \theta^*$ yields $f_{\mathrm{d}}(r(x; \widehat{\theta})) = f_{\mathrm{d}}(r(x; \theta^*)) + {f_{\mathrm{d}}}'(r(x; \theta^*)) r(x; \theta^*) \nabla \log r(x; \theta^*)^T (\widehat{\theta} - \theta^*) + O(\|\widehat{\theta} - \theta^*\|^2)$. We have the same expansion for $f_{\mathrm{n}}(r(x; \widehat{\theta}))$. By using the above expansions with Assumption 3 and Assumption 4, we have

$$\sqrt{m}(\widehat{D}_f - D_f) = \mathbb{G}f - \sqrt{m}\, c^T U_\eta^{-1} Q_\eta(\theta^*) + o_p(1),$$
$$\sqrt{m}(\bar{D}_f - D_f) = \bar{\mathbb{G}}f - \sqrt{m}\, \bar{c}^T U_{\bar{\eta}}^{-1} Q_{\bar{\eta}}(\theta^*) + o_p(1).$$

The first and the second terms of $\sqrt{m}(\widehat{D}_f - D_f)$ and $\sqrt{m}(\bar{D}_f - D_f)$ converge in distribution to a centered normal distribution. For $p_{\mathrm{n}} = p_{\mathrm{d}}$, however, these terms may vanish and $\sqrt{m}\widehat{D}_f$ becomes of the order $o_p(1)$, as shown in Appendix B. Substituting the above expression into $m \cdot \mathrm{Cov}[\widehat{D}_f - \bar{D}_f, \bar{D}_f] = \mathbb{E}[\{\sqrt{m}(\widehat{D}_f - D_f) - \sqrt{m}(\bar{D}_f - D_f)\}\sqrt{m}(\bar{D}_f - D_f)] + o(1)$, we obtain (13).

B Proof of Theorem 2

Proof. Let $\delta\widehat{\theta} = \widehat{\theta} - \theta^*$. Then, due to (2), we have $\sqrt{m}\,\delta\widehat{\theta} = -\sqrt{m}\,U_\eta^{-1}Q_\eta + o_p(1)$, where $\eta = \eta_{\mathrm{opt}}$ defined in (3). Let $f_{\mathrm{d}}(r) = f(r)/(1+\rho r)$ and $f_{\mathrm{n}}(r) = \rho f(r)/(1+\rho r)$. Then we have $f_{\mathrm{d}}(1) = {f_{\mathrm{d}}}'(1) = f_{\mathrm{n}}(1) = {f_{\mathrm{n}}}'(1) = 0$ and ${f_{\mathrm{d}}}''(1) + {f_{\mathrm{n}}}''(1) = f''(1)$, since $f(1) = f'(1) = 0$ is assumed. The asymptotic expansion of $m\widehat{D}_f$ around $\theta = \theta^*$ leads to

$$\begin{aligned} m\widehat{D}_f &= \frac{{f_{\mathrm{d}}}''(1)}{2}\sqrt{m}\delta\widehat{\theta}^T \mathrm{E}_{\mathrm{d}}[\nabla r(x;\theta^*)\nabla r(x;\theta^*)^T]\sqrt{m}\delta\widehat{\theta}, \\ &\quad + \frac{{f_{\mathrm{n}}}''(1)}{2}\sqrt{m}\delta\widehat{\theta}^T \mathrm{E}_{\mathrm{n}}[\nabla r(x;\theta^*)\nabla r(x;\theta^*)^T]\sqrt{m}\delta\widehat{\theta} + o_p(1) \\ &= \frac{(1+\rho)^2 f''(1)}{2}\sqrt{m}Q_\eta^T(\mathrm{E}_{\mathrm{n}}[\nabla r(x;\theta^*)\nabla r(x;\theta^*)^T])^{-1}\sqrt{m}Q_\eta + o_p(1), \end{aligned}$$

since $p_{\mathrm{n}} = p_{\mathrm{d}}$ and $r(x;\theta^*) = 1$ hold. The asymptotic distribution of $\sqrt{m}Q_\eta$ is the Gaussian distribution with mean zero and variance-covariance matrix $\mathrm{V}_{\mathrm{n}}[\nabla r]/(1+\rho)^2$, since the equality $\eta_{\mathrm{opt}}(x;\theta^*) = \nabla \log r(x;\theta^*)/(1+\rho) = \nabla r(x;\theta^*)/(1+\rho)$ holds. Let M be the d by d matrix defined as $M = \mathrm{E}_{\mathrm{n}}[\nabla r(x;\theta^*)\nabla r(x;\theta^*)^T]$, and $\sqrt{V}$ be a d by d matrix such that $\sqrt{V}\sqrt{V}^T = \mathrm{V}_{\mathrm{n}}[\nabla r]$. Then asymptotically

$$\frac{2m}{f''(1)}\widehat{D}_f \xrightarrow{d} Z_d^T\sqrt{V}^T M^{-1}\sqrt{V}Z_d$$

holds, where Z_d is the d-dimensional random vector whose distribution is the d-dimensional standard Gaussian distribution, that is, $Z_d \sim N_d(0, I_d)$. Let $\sqrt{M}$ be the symmetric positive definite matrix such that $M = \sqrt{M}\sqrt{M}$, and the vector μ be $\mu = \mathrm{E}_{\mathrm{n}}[\nabla r(x;\theta^*)]$. Note that $\sqrt{M}$ is well-defined, since M is a positive definite matrix. Let P be the d by d matrix $P = I - \sqrt{M}^{-1}\mu\mu^T\sqrt{M}^{-1}$, then P is the projection matrix along the vector $\sqrt{M}^{-1}\mu$. Indeed, for the vector $b \in \mathbb{R}^d$ such that $b^T\nabla \log r(x;\theta^*) = \nabla r(x;\theta^*)^T b = 1$, we have $\mathrm{E}_{\mathrm{n}}[\nabla r] = \mathrm{E}_{\mathrm{n}}[\nabla r(\nabla r)^T b] = \mathrm{E}_{\mathrm{n}}[\nabla r(\nabla r)^T]b$, and thus, $\|\sqrt{M}^{-1}\mu\|^2 = \mathrm{E}_{\mathrm{n}}[\nabla r]^T\mathrm{E}_{\mathrm{n}}[\nabla r\nabla r^T]^{-1}\mathrm{E}_{\mathrm{n}}[\nabla r] = \mathrm{E}_{\mathrm{n}}[\nabla r]^T b = 1$ holds. We can choose $\sqrt{V} = \sqrt{M}P$, since $\sqrt{V}\sqrt{V}^T = M - \mu\mu^T$ holds. As a result, we have $Z_d^T\sqrt{V}^T M^{-1}\sqrt{V}Z_d = Z_d^T P Z_d$, and the distribution of $Z_d^T P Z_d$ is the chi-square distribution with $d-1$ degrees of freedom.

C Proof of Theorem 3

First, we calculate the power function of $\widehat{D}_f$-based test. Proof. The equality $p_{\mathrm{n}}^{(m)}(x) = p_{\mathrm{d}}(x)r(x;\theta_m)$ leads to $\mathrm{E}_{\mathrm{d}}[\nabla r(x;\theta^*)]^T h = 0$. Indeed

$$\begin{aligned} &\int p_{\mathrm{n}}^{(m)}(x)dx = \int p_{\mathrm{d}}(x)r(x;\theta_m)dx \\ &\Longrightarrow 1 = 1 + \mathrm{E}_{\mathrm{d}}[\nabla r(x;\theta^*)]^T\frac{h_m}{\sqrt{m}} + o(1/\sqrt{m}) \end{aligned} \tag{29}$$

holds, since $\theta_m = \theta^* + h_m/\sqrt{m}$. Thus, we have $E[\nabla r(x;\theta^*)]^T h = 0$ when m tends to infinity. Let the matrix M be $M(\theta^*) = E[\nabla r(x;\theta^*)\nabla r(x;\theta^*)^T]$, the vector μ be $E[\nabla r(x;\theta^*)]$, and $\sqrt{V}$ be a matrix such that $\sqrt{V}\sqrt{V}^T = V[\nabla r]$. Let $\delta\widehat{\theta}_m$ be $\widehat{\theta} - \theta_m$. Under Assumption 5, the asymptotic expansion gives

$$\begin{aligned}\frac{2m}{f''(1)}\widehat{D}_f &= (\sqrt{m}\delta\theta_m + h_m)^T M(\sqrt{m}\delta\theta_m + h_m) + o_p(1)\\ &= (\sqrt{m}U(\theta_m)\delta\theta_m + U(\theta_m)h)^T U(\theta_m)^{-1} M U(\theta_m)^{-1}\\ &\quad \times (\sqrt{m}U(\theta_m)\delta\theta_m + U(\theta_m)h) + o_p(1)\\ &\xrightarrow{d} \left\|\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h\right\|^2, \quad Z_d \sim N_d(0, I_d), \end{aligned} \tag{30}$$

Since $\lim_{m\to\infty} U(\theta_m) = M/(1+\rho)$ and $\sqrt{m}U(\theta_m)\delta\theta_m \xrightarrow{d} \sqrt{V}Z_d/(1+\rho)$ hold. In the same way as the proof of Theorem 2, we see that $\sqrt{M}^{-1}\sqrt{V}$ is the projection matrix along the vector $\sqrt{M}^{-1}\mu$. Moreover, $\sqrt{M}h$ is orthogonal to the vector $\sqrt{M}^{-1}\mu$ since $\mu^T h = 0$ holds. As a result, we see that the distribution function of $\left\|\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h\right\|^2$ is the non-central chi-square distribution with $d-1$ degrees of freedom and non-centrality parameter $h^T M(\theta^*)h$.

Next, we calculate the power function of empirical likelihood score test. The notations M and μ are the same as the proof above. Proof. From the definition of the statistic S, we have

$$\begin{aligned} S &= m(\widehat{\beta} - \beta^*)^T \widehat{V}_{\mathrm{n}}[\nabla_\beta \phi](\widehat{\beta} - \beta^*)\\ &= m(\widehat{\theta} - \theta^*)^T V(\widehat{\theta} - \theta^*) + o_p(1), \end{aligned}$$

where $V = V[\nabla r]$. In the same way as (30), we have

$$m(\widehat{\theta} - \theta^*)^T V(\widehat{\theta} - \theta^*) + o_p(1) \xrightarrow{d} \left\|\sqrt{V}^T\sqrt{M}^{-1}(\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h)\right\|^2.$$

The matrix $\sqrt{V}^T\sqrt{M}^{-1}$ is the projection matrix along the vector $\sqrt{M}^{-1}\mu$ and $\mu^T h = 0$ holds. Then we see that the vector $\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h$ is orthogonal to $\sqrt{M}^{-1}\mu$. This implies $\|\sqrt{V}^T\sqrt{M}^{-1}(\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h)\|^2 = \|\sqrt{M}^{-1}\sqrt{V}Z_d + \sqrt{M}h\|^2$. Thus, under the local alternative setup, the limit distribution of the test statistic S is the non-central chi-square distribution with the same parameter as $\widehat{D}_f$-based test.

D Proof of Theorem 4

Below, the notations $M = E[\nabla r(x;\theta^*)\nabla r(x;\theta^*)]$ and $\mu = E[\nabla r(x;\theta^*)]$ are used. Proof. In the same way as (29), we have $\mu^T h + \varepsilon = 0$. Let the random vector W be $W = PZ_d + \sqrt{M}h$, $Z_d \sim N_d(0, I_d)$, where P is the projection matrix along the vector $\sqrt{M}^{-1}\mu$ as defined in the proof of Theorem 2. According to the proof of Theorem 3 in Appendix C, the power of $\widehat{D}_f$-based test is asymptotically equal to $\Pr\left\{\|W\|^2 \geq \chi^2_{d-1}(1-\alpha)\right\}$, and

that of empirical likelihood score test is equal to $\Pr\left\{\|PW\|^2 \geq \chi^2_{d-1}(1-\alpha)\right\}$. We have the equality $W = PW + c\sqrt{M}^{-1}h$ with some $c \in \mathbb{R}$. Note that generally $\sqrt{M}h$ is not orthogonal to $\sqrt{M}^{-1}\mu$ in the misspecified case, since $(\sqrt{M}^{-1}\mu)^T\sqrt{M}h = \mu^T h = -\varepsilon$ holds. For $\varepsilon \neq 0$, we have $c \neq 0$ and then the inequality $\|W\|^2 > \|PW\|^2$ holds. As a result, the power of $\widehat{D}_f$-based test is larger than or equal to that of empirical likelihood score test under the misspecified setup.

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

[2] N. Bensaid and J. P. Fabre. Optimal asymptotic quadratic error of kernel estimators of Radon-Nikodym derivatives for strong mixing data. *Journal of Nonparametric Statistics*, 19(2):77–88, 2007.

[3] M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100:16–26, 2009.

[4] K. F. Cheng and C. K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

[5] T. F. Cox and G. Ferry. Robust logistic discrimination. *Biometrika*, 78(4):841–849, 1991.

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

[7] K. Fokianos, B. Kedem, J. Qin, and D. A. Short. A semiparametric approach to the one-way layout. *Technometrics*, 43:56–64, 2001.

[8] V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31:1208–1211, 1960.

[9] P. Jacoba and P. E. Oliveirab. Kernel estimators of general Radon-Nikodym derivatives. *Statistics*, 30:25–46, 1997.

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

[11] R. Kay and S. Little. Transformation of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74(3):495–501, 1987.

[12] A. Keziou. Dual representation of ϕ-divergences and applications. *C. R. Acad. Sci. Paris, Ser. I*, 336(10):857–862, 2003.

[13] A. Keziou. *Utilisation des divergences entre mesures en statistique inferentielle*. PhD thesis, UPMC University, 2003.

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

[15] A. Keziou and S. Leoni-Aubin. On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138(4):915–928, 2008.

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

[17] J. Kwik and J. Mielniczuk. Estimating density ratio with application to discriminant analysis. *Commun. Statist. –Theory Meth.*, 18(8):3057–3069, 1989.

[18] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

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

[20] J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–639, 1998.

[21] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[22] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[23] 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*, 60(4):699–746, 2008.

[24] A. W. Van der Vaart. *Asymptotic statistics*. Cambridge Ser. Stat. Probab. Math. Cambridge Univ. Press, Cambridge, 1998.

[25] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55:2392–2405, 2009.

[26] Q. Wang, S. R. Kulkarni, and S. Verdú. Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5:265–353, 2009.