

Article

Statistical Analysis of Distance Estimators with Density Differences and Density Ratios

Takafumi Kanamori ^{1,*}, Masashi Sugiyama ²

¹ Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan

² Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan;

E-Mail: sugi@cs.titech.ac.jp

* Author to whom correspondence should be addressed; E-Mail: kanamori@is.nagoya-u.ac.jp;
Tel.: +81-52-789-4598.

Received: 21 October 2013; in revised form: 27 January 2014 / Accepted: 7 February 2014 /

Published: 17 February 2014

Abstract: Estimating a discrepancy between two probability distributions from samples is an important task in statistics and machine learning. There are mainly two classes of discrepancy measures: distance measures based on the density difference, such as the L_p -distances, and divergence measures based on the density ratio, such as the ϕ -divergences. The intersection of these two classes is the L_1 -distance measure, and thus, it can be estimated either based on the density difference or the density ratio. In this paper, we first show that the Bregman scores, which are widely employed for the estimation of probability densities in statistical data analysis, allows us to estimate the density difference and the density ratio directly without separately estimating each probability distribution. We then theoretically elucidate the robustness of these estimators and present numerical experiments.

Keywords: density difference; density ratio; L_1 -distance; Bregman score; robustness

1. Introduction

In statistics and machine learning, estimating a discrepancy between two probability distributions from samples has been extensively studied [1], because discrepancy estimation is useful in solving various real-world data analysis tasks, including covariate shift adaptation [2,3], conditional probability estimation [4], outlier detection [5] and divergence-based two-sample testing [6].

There are mainly two classes of discrepancy measures for probability densities. One is genuine distances on function spaces, such as the L_s -distance for $s \geq 1$, and the other is divergence measures, such as the Kullback–Leibler divergence and the Pearson divergence. Typically, distance measures in the former class can be represented using the difference of two probability densities, while those in the later class are represented using the ratio of two probability densities. Therefore, it is important to establish statistical methods to estimate the density difference and the density ratio.

A naive way to estimate the density difference and the density ratio consists of two steps: two probability densities are separately estimated in the first step, and then, their difference or ratio is computed in the second step. However, such a two-step approach is not favorable in practice, because the density estimation in the first step is carried out without regard to the second step of taking the difference or ratio. To overcome this problem, the authors in [7–11] studied the estimation of the density difference and the density ratio in a semi-parametric manner without separately modeling each probability distribution.

The intersection of the density difference-based distances and the density ratio-based divergences is the L_1 -distance, and thus, it can be estimated either based on the density difference or the density ratio. In this paper, we first propose a novel direct method to estimate the density difference and the density ratio based on the Bregman scores [12]. We then show that the density-difference approach to L_1 -distance estimation is more robust than the density-ratio approach. This fact has already been pointed out in [10] based on a somewhat intuitive argument: the density difference is always bounded, while the density ratio can be unbounded. In this paper, we theoretically support this claim by providing detailed theoretical analysis of the robustness properties.

There are some related works to our study. Density ratio estimation was intensively investigated in the machine learning community [4,7,8]. As shown in [6], the density-ratio is available to estimate the ϕ -divergence [13,14]. However, estimation of the L_1 -distance, which is a member of the ϕ -divergence, was not studied, since it does not satisfy the regularity condition, which is required to investigate the statistical asymptotic property. On the other hand, the least mean squares estimator of the density-difference was proposed in [10], and the robustness property was numerically investigated. In the present paper, not only the least squares estimator, but also general score estimators for density differences are considered, and their robustness properties are theoretically investigated.

The rest of the paper is structured as follows. In Section 2, we describe two approaches to L_1 -distance estimation based on the density difference and the density ratio. In Section 3, we introduce the Bregman scores, which are widely employed for the estimation of probability densities in statistical data analysis. In Section 4, we apply the Bregman score to the estimation of the density difference and the density ratio. In Section 5, we introduce a robustness measure in terms of which the proposed estimators are analyzed in the following sections. In Section 6, we consider statistical models without the scale parameter (called the non-scale models) and investigate the robustness of the density difference and density ratio estimators. In Section 7, we consider statistical models with the scale parameter (called the scale models) and show that the estimation using the scale models is reduced to the estimation using the non-scale models. Then, we apply the theoretical results on the non-scale models to the scale models and elucidate the robustness of the scale models. In Section 8, numerical examples on L_1 -distance estimation are presented. Finally, we conclude in Section 9.

2. Estimation of L_1 -Distance

Let $p(x)$ and $q(x)$ be two probability densities. In this section, we introduce two approaches to estimating discrepancy measures: an approach based on the density difference, $p - q$, and an approach based on the density ratio, p/q .

2.1. L_1 -Distance As the Density Difference and Density Ratio

The density difference, $p - q$, is directly used to compute the L_s -distance between two probability densities:

$$d_s(p, q) = \left(\int |p(x) - q(x)|^s dx \right)^{1/s}, \quad (1)$$

where $s \geq 1$. On the other hand, the density ratio, p/q , appears in the ϕ -divergence [13,14] defined as:

$$\int \phi \left(\frac{p(x)}{q(x)} \right) q(x) dx,$$

where ϕ is a strictly convex function, such that $\phi(1) = 0$. The ϕ -divergence is non-negative and vanishes only when $p = q$ holds. Hence, it can be regarded as an extension of the distance between p and q . The class of ϕ -divergences includes many important discrepancy measures, such as the Kullback–Leibler divergence ($\phi(z) = z \log z$), the Pearson distance ($\phi(z) = (1 - z)^2$) and the L_1 -distance ($\phi(z) = |1 - z|$). The intersection of the ϕ -divergence and the L_s -distance is the L_1 -distance:

$$d_1(p, q) = \int |p(x) - q(x)| dx.$$

The purpose of our work is to compare the statistical properties of the density-difference approach and the density-ratio approach to the estimation of the L_1 -distance between probability densities p and q defined on \mathbb{R}^d . For the estimation of the L_1 -distance, we use two sets of identically and independently distributed (i.i.d.) samples:

$$x_1, \dots, x_n \sim p, \quad y_1, \dots, y_m \sim q. \quad (2)$$

In both density-difference and density-ratio approaches, semi-parametric statistical models are used, which will be explained below.

2.2. Density-Difference Approach

The difference of two probability densities, $f(x) = p(x) - q(x)$, is widely applied to statistical inference [10]. A parametric statistical model for the density difference $f(x)$ is denoted as:

$$\mathcal{M}_{\text{diff}} = \{f(x; \theta) = f_\theta(x) \mid \theta \in \Theta_k\}, \quad (3)$$

where Θ_k is the k -dimensional parameter space. The density-difference model, $f(x; \theta)$, can take both positive and negative values, and its integral should vanish. Note that there are infinitely many

degrees of freedom to specify the probability densities, p and q , even when the density difference $f = p - q$ is specified. Hence, the density-difference model is regarded as a semi-parametric model for probability densities.

Recently, a density-difference estimator from Samples (2) that does not involve separate estimation of two probability densities has been proposed [10]. Once a density-difference estimator, $\hat{f} \in \mathcal{M}_{\text{diff}}$, is obtained, the L_1 -distance can be immediately estimated as:

$$d_1(p, q) = \int |f(x)|dx \approx \int |\hat{f}(x)|dx.$$

The L_1 -distance has the invariance property under the variable change. More specifically, let $x = \psi(z)$ be a one-to-one mapping on \mathbb{R}^d and $f_\psi(z)$ be $f(\psi(z))|J_\psi(z)|$, where J_ψ is the Jacobian determinant of ψ . For $f(x) = p(x) - q(x)$, the function, $f_\psi(z)$, is the density difference between p and q in the z -coordinate. Then, we have:

$$\int |f(x)|dx = \int |f_\psi(z)|dz, \tag{4}$$

due to the formula of variable change for probability densities. When the transformed data, z , with the model, $f_\psi(z)$, is used instead of the model, $f(x)$, the L_1 -distance in the z -coordinate can be found in the same way as the L_1 -distance in the x -coordinate.

Note that this invariance property does not hold for general distance measures. Indeed, we have:

$$(d_s(p, q))^s = \int |f(x)|^s dx = \int |f_\psi(z)|^s |J_\psi(z)|^{1-s} dz$$

for general distance measures.

2.3. Density-Ratio Approach

The density ratio of two probability densities, $p(x)$ and $q(x)$, is defined as $r(x) = p(x)/q(x)$, which is widely applied in statistical inference as the density difference [4]. Let:

$$\mathcal{M}_{\text{ratio}} = \{r(x; \theta) = r_\theta(x) \mid \theta \in \Theta_k\} \tag{5}$$

be the k -dimensional parametric statistical model of the density ratio, $r(x)$. From the definition of the density ratio, the function, $r(x; \theta)$, should be non-negative. Various estimators of the density ratio based on the Samples (2) that do not involve separate estimation of two probability densities have been developed so far [7,8,11]. Once the density ratio estimator, $\hat{r} \in \mathcal{M}_{\text{ratio}}$, is obtained, the L_1 -distance between p and q can be immediately estimated as:

$$d_1(p, q) = \int |1 - r(x)|q(y)dy \approx \int |1 - \hat{r}(x)|q(y)dy \approx \frac{1}{m} \sum_{j=1}^m |1 - \hat{r}(y_j)|.$$

Differently from the L_1 -distance estimator using the density difference, the numerical integral should be replaced with the sample mean, since the density, $q(y)$, is unknown. In the density-ratio approach, the variable transformation maintains the L_1 -distance, as well as the estimation of the density difference

itself. For the one-to-one mapping $y = \psi(z)$, let $r_\psi(z)$ be $r(\psi(z))$ and the probability density, $q_\psi(z)$, be $q(\psi(z))|J_\psi(z)|$. Then, we have:

$$d_1(p, q) = \int |1 - r(y)|q(y)dy = \int |1 - r_\psi(z)|q_\psi(z)dz \approx \frac{1}{m} \sum_{j=1}^m |1 - r_\psi(z_j)|,$$

where z_j is the transformed sample, such that $y_j = \psi(z_j)$. In the density-ratio approach, the L_1 -distance for the transformed data does not require the computation of the Jacobian determinant, J_ψ .

3. Bregman Scores

The Bregman score is an extension of the log-likelihood function, and it is widely applied in statistical inference [12,15–18]. In this section, we briefly review the Bregman score. See [12] for details.

For functions f and g on \mathbb{R}^d , the Bregman score, $S(f, g)$, is a class of real-valued functions that satisfy the inequality:

$$S(f, g) \geq S(f, f). \tag{6}$$

Clearly, the inequality becomes the equality for $f = g$. If the equality $S(f, g) = S(f, f)$ leads to $f = g$, $S(f, g)$ is called the strict Bregman score. The minimization problem of the strict Bregman score, $S(f, g)$, i.e., $\min_g S(f, g)$, has the uniquely optimal solution $g = f$.

Let us introduce the definition of Bregman scores. For a function, f , defined on the Euclidean space, \mathbb{R}^d , let $G(f)$ be a real-valued convex functional. The functional, $G(f)$, is called the potential below. The functional derivative of $G(f)$ is denoted as $G'(x; f)$, which is defined as the function satisfying the equality:

$$\lim_{\varepsilon \rightarrow 0} \frac{G(f + \varepsilon h) - G(f)}{\varepsilon} = \int G'(x; f)h(x)\lambda(dx) \tag{7}$$

for any function, $h(x)$, with a regularity condition, where $\lambda(\cdot)$ is the base measure. Then, the Bregman score, $S(f, g)$, for functions f, g is defined as:

$$S(f, g) = -G(g) - \int G'(x, g)(f(x) - g(x))\lambda(dx). \tag{8}$$

Due to the convexity of $G(f)$, we have:

$$G(f) - G(g) - \int G'(x, g)(f(x) - g(x))\lambda(dx) \geq 0,$$

which is equivalent to Inequality (6). Let \mathcal{F} be a set of functions defined on \mathbb{R}^d . If \mathcal{F} is a convex set and the potential $G(f)$ is strictly convex on \mathcal{F} , the associated Bregman score is strict.

When $G(f)$ is expressed as:

$$G(f) = \int U(f(x))\lambda(dx),$$

with a convex differentiable function $U : \mathbb{R} \rightarrow \mathbb{R}$, the corresponding Bregman score is referred to as the separable Bregman score, which is given as:

$$S(f, g) = - \int \{U(g(x)) + U'(g(x))(f(x) - g(x))\} \lambda(dx).$$

Due to the computational tractability, the separable Bregman scores are often used in real-world data analysis.

If f is a probability density, the Bregman score is expressed as:

$$S(f, g) = \int f(x)\ell(x, g)\lambda(dx), \tag{9}$$

where $\ell(x, g)$ is given by

$$\ell(x, g) = -G'(x, g) - G(g) + \int G'(y, g)g(y)\lambda(dy). \tag{10}$$

The function, $\ell(x, g)$, is regarded as the loss of the forecast using $g \in \mathcal{F}$ for an outcome, $x \in \mathbb{R}^d$. The function of the form of Equation (9) is called a proper scoring rule, and its relation to the Bregman score has been extensively investigated [12,17,19]. When the i.i.d. samples, x_1, \dots, x_n , are observed from the probability density, f , the minimization problem of the empirical mean over the probability model, g , i.e.,

$$\min_g \frac{1}{n} \sum_{i=1}^n \ell(x_i, g),$$

is expected to provide a good estimate of the probability density, f .

Below, let us introduce exemplary Bregman scores:

Example 1 (Kullback–Leibler (KL) score). *The Kullback–Leibler (KL) score for the probability densities, $p(x)$ and $q(x)$, is defined as:*

$$S(p, q) = - \int p(x) \log q(x) dx,$$

which is the separable Bregman score with the potential function:

$$G(p) = \int p(x) \log p(x) dx,$$

i.e., the negative entropy. The difference $S(p, q) - S(p, p)$ is called the KL divergence [20]. The KL score is usually defined for probability densities, but an extension to non-negative functions is also available. Hence, the KL score is applicable to the estimation of the density ratio [8,11]. However, it is not possible to directly use the KL score to estimate the density difference, because it can take negative values.

Example 2 (Density-power score). *Let α be a positive number and f and g be functions that can take both positive and negative values. Then, the density-power score with the base measure $\lambda(\cdot)$ is defined as:*

$$S(f, g) = \alpha \int |g(x)|^{1+\alpha} \lambda(dx) - (1 + \alpha) \int f(x)|g(x)|^{\alpha-1} g(x) \lambda(dx).$$

See [21,22] for the details of the density-power score for probability densities. The potential of the density power score is given as:

$$G(f) = \int |f(x)|^{1+\alpha} dx.$$

Hence, the density-power score is the separable Bregman score. Letting α be zero, then the difference of the density-power scores $(S(p, q) - S(p, p))/\alpha$ for probability densities p and q tends to the KL divergence.

Example 3 (Pseudo-spherical score; γ -score). For $\alpha > 0$ and $g \neq 0$, the pseudo-spherical score [16] is defined as:

$$S(f, g) = - \frac{\int f(x) |g(x)|^{\alpha-1} g(x) \lambda(dx)}{\left(\int |g(x)|^{1+\alpha} \lambda(dx) \right)^{\alpha/(1+\alpha)}}.$$

This is the Bregman score derived from the potential function:

$$G(f) = \left(\int |f(x)|^{1+\alpha} \lambda(dx) \right)^{1/(1+\alpha)},$$

implying that the pseudo-spherical score is a non-separable Bregman score. For probability densities, p and q , the monotone transformation of the pseudo-spherical score, $-\log(-S(p, q))$, is called the γ -score [23], which is used for robust parameter estimation of probability densities. In the limiting case of $\alpha \rightarrow 0$, the difference of the γ -scores, $-\log(-S(p, q)) + \log(-S(p, p))$, recovers the KL-divergence. Note that the corresponding potential is not strictly convex on a set of functions, but is strictly convex on the set of probability densities. As a result, the equality $S(p, q) = S(p, p)$ for the probability densities, p, q , leads to $p = q$, while the equality $S(f, g) = S(f, f)$ for functions f and g yields that f and g are linearly dependent. The last assertion comes from the equality condition of the Hölder's inequality.

When the model, $f(x; \theta)$, includes the scale parameter, c , i.e., $f(x; \theta) = cg(x; \bar{\theta})$ with the parameter $\theta = (c, \bar{\theta}) \in \Theta_k$ for $c \in \mathbb{R}$ and $\bar{\theta} \in \Theta_{k-1}$, the pseudo-spherical score does not work. This is because the potential is not strictly convex on the statistical model with the scale parameter, and hence, the scale parameter, c , is not estimable when the pseudo-spherical score is used.

The density-power score and the pseudo-spherical score in the above examples include the non-negative parameter, α . When α is an odd integer, the absolute-value operator in the scores can be removed, which is computationally advantageous. For this reason, we set the parameter, α , to a positive odd integer when the Bregman score is used for the estimation of the density difference.

4. Direct Estimation of Density Differences and Density Ratios Using Bregman Scores

The Bregman scores are applicable not only to the estimation of probability densities, but also to the estimation of density differences and density ratios. In this section, we propose estimators for density differences and density ratios, and show their theoretical properties.

4.1. Estimators for Density Differences and Density Ratios

First of all, let us introduce a way to directly estimate the density difference based on the Bregman scores. Let $\mathcal{M}_{\text{diff}}$ be the statistical Model (3) to estimate the true density-difference $f(x) = p(x) - q(x)$ defined on the Euclidean space, \mathbb{R}^d . Let the base measure, $\lambda(\cdot)$, be the Lebesgue measure. Then, for the density-difference model, $f_\theta \in \mathcal{M}_{\text{diff}}$, the Bregman score Equation (8) is given as:

$$S_{\text{diff}}(f, f_\theta) = \int p(x) \ell(x, f_\theta) dx - \int q(x) \ell(x, f_\theta) dx,$$

where $\ell(x, f_\theta)$ is defined in Equation (10). This can be approximated by the empirical mean based on the Samples (2) as follows. Let δ be the Dirac delta function and \tilde{f} be the difference between two empirical densities,

$$\tilde{f}(z) = \tilde{p}(z) - \tilde{q}(z) = \frac{1}{n} \sum_{i=1}^n \delta(z - x_i) - \frac{1}{m} \sum_{j=1}^m \delta(z - y_j).$$

Then, we have:

$$S_{\text{diff}}(f, f_\theta) \approx S_{\text{diff}}(\tilde{f}, f_\theta) = \frac{1}{n} \sum_{i=1}^n \ell(x_i, f_\theta) - \frac{1}{m} \sum_{j=1}^m \ell(y_j, f_\theta).$$

If the target density difference, f , is included in the model M_{diff} , the minimizer of the strict Bregman score, $S_{\text{diff}}(\tilde{f}, f_\theta)$, with respect to $f_\theta \in M_{\text{diff}}$ is expected to produce a good estimator of f .

Next, we use the Bregman scores to estimate the density ratio $r(x) = p(x)/q(x)$. Let us define $q(x)$ as the base measure of the Bregman score. Given the density-ratio model, M_{ratio} , of Equation (5), the Bregman score of the model, $r_\theta \in M_{\text{ratio}}$, is given as:

$$\begin{aligned} S_{\text{ratio}}(r, r_\theta) &= -G(r_\theta) - \int G'(x, r_\theta)(r(x) - r_\theta(x))q(x)dx \\ &= -G(r_\theta) - \int G'(x, r_\theta)p(x)dx + \int G'(x, r_\theta)r_\theta(x)q(x)dx. \end{aligned}$$

Using the Samples (2), we can approximate the score, $S_{\text{ratio}}(r, r_\theta)$, with the base measure, q , as:

$$S_{\text{ratio}}(r, r_\theta) \approx -G(r_\theta) - \frac{1}{n} \sum_{i=1}^n G'(x_i, r_\theta) + \frac{1}{m} \sum_{j=1}^m G'(y_j, r_\theta)r_\theta(y_j).$$

For example, the density-power score for the density ratio is given as:

$$\begin{aligned} S_{\text{ratio}}(r, r_\theta) &= -(1 + \alpha) \int r_\theta(x)^\alpha r(x)q(x)dx + \alpha \int r_\theta(x)^{1+\alpha}q(x)dx \\ &= -(1 + \alpha) \int r_\theta(x)^\alpha p(x)dx + \alpha \int r_\theta(x)^{1+\alpha}q(x)dx, \end{aligned}$$

in which the equality $r(x)q(x) = p(x)$ is used. We can also obtain similar approximation for the pseudo-spherical score.

4.2. Invariance of Estimators

We show that the estimators obtained by the density-power score and the pseudo-spherical score have the affine invariance property. Suppose that the Samples (2) are distributed on the d -dimensional Euclidean space, and let us consider the affine transformation of samples, such that $x_i = Ax'_i + b$ and $y_j = Ay'_j + b$, where A is an invertible matrix and b is a vector. Let $f_{A,b}(x)$ be the transformed density-difference $|\det A|f(Ax' + b)$ and $\tilde{f}_{A,b}$ be the difference of the empirical distributions defined from the samples, x'_i, y'_j . Let $S_{\text{diff}}(f, g)$ be the density-power score or the pseudo-spherical score with a positive odd integer, α , for the estimation of the density difference. Then, we have:

$$S_{\text{diff}}(\tilde{f}_{A,b}, f_{A,b}) = |\det A|^\alpha S_{\text{diff}}(\tilde{f}, f).$$

Let \widehat{f} ($\widehat{f}_{A,b}$) be the estimator based on the samples, $\{x_i\}$ and $\{y_j\}$ ($\{x'_i\}$ and $\{y'_j\}$). Then, the above equality leads to $(\widehat{f})_{A,b} = \widehat{f}_{A,b}$, implying that the estimator is invariant under the affine transformation of the data. In addition, Equality (4) leads to:

$$\int |\widehat{f}(x)|dx = \int |(\widehat{f})_{A,b}(x)|dx = \int |\widehat{f}_{A,b}(x)|dx.$$

This implies that the affine transformation of the data does not affect the estimated L_1 -distance. The same invariance property holds for the density-ratio estimators based on the density-power score and the pseudo-spherical score.

5. Robustness Measure

The robustness of the estimator is an important feature in practice, since typically real-world data includes outliers that may undermine the reliability of the estimator. In this section, we introduce robustness measures of estimators against outliers.

In order to define robustness measures, let us briefly introduce the influence function in the setup of the density-difference estimation. Let $p(x)$ and $q(x)$ be the true probability densities of each dataset in Samples (2). Suppose that these probabilities are shifted to:

$$\begin{aligned} p_\varepsilon(x) &= (1 - \varepsilon)p(x) + \varepsilon\delta(x - z_p), \\ q_\varepsilon(x) &= (1 - \varepsilon)q(x) + \varepsilon\delta(x - z_q), \end{aligned} \tag{11}$$

by the outliers, z_p and z_q , respectively. A small positive number, ε , denotes the ratio of outliers. Let θ^* be the true model parameter of the density difference $f(x) = p(x) - q(x)$, i.e., $f(x) = f(x; \theta^*) \in \mathcal{M}_{\text{diff}}$. Let us define the parameter, θ_ε , as the minimum solution of the problem,

$$\min_{\theta \in \Theta} S_{\text{diff}}(p_\varepsilon - q_\varepsilon, f_\theta).$$

Clearly, $\theta_0 = \theta^*$ holds. For the density-difference estimator using the Bregman score, $S_{\text{diff}}(f, f_\theta)$, with the model, $\mathcal{M}_{\text{diff}}$, the influence function is defined as:

$$\text{IF}_{\text{diff}}(\theta^*; z_p, z_q) = \lim_{\varepsilon \searrow 0} \frac{\theta_\varepsilon - \theta^*}{\varepsilon}.$$

Intuitively, the estimated parameter is distributed around $\theta^* + \varepsilon \cdot \text{IF}_{\text{diff}}(\theta^*; z_p, z_q)$ under the existence of outliers, z_p and z_q , with the small contamination ratio, ε . The influence function for the density ratio is defined in the same manner, and it is denoted as $\text{IF}_{\text{ratio}}(\theta^*; z_p, z_q)$.

The influence function provides several robustness measures of estimators. An example is the gross error sensitivity defined as $\sup_{z_p, z_q} \|\text{IF}_{\text{diff}}(\theta^*; z_p, z_q)\|$, where $\|\cdot\|$ is the Euclidean norm. The estimator that uniformly minimizes the gross error sensitivity over the parameter, θ , is called the most B(bias)-robust estimator. The most B-robust estimator minimizes the worst-case influence of outliers. For the one-dimensional normal distribution, the median estimator is the most B-robust for the estimation of the mean parameter [24].

In this paper, we consider another robustness measure, called the *redescending property*. The estimator satisfying the following redescending property,

$$\lim_{\|z_p\|, \|z_q\| \rightarrow \infty} \|\text{IF}_{\text{diff}}(\theta; z_p, z_q)\| = 0 \quad \text{for all } \theta \in \Theta,$$

is called the redescending estimator [23–26]. The redescending property is preferable to stable inference, since the influence of extreme outliers can be ignored. Furthermore, in the machine learning literature, learning algorithms with the redescending property for classification problems have been proposed, under the name of the robust support vector machines [27–29]. Note that the most B-robust estimator is not necessarily a redescending estimator, and *vice versa*. It is known that, for the estimation of probability densities, the pseudo-spherical score has the redescending property, while the density-power score does not necessarily provide the redescending estimator [23,30].

In the next sections, we apply the density-power score and the pseudo-spherical score to estimate the density difference or the density ratio and investigate their robustness.

6. Robustness under Non-Scale Models

In this section, we consider statistical models without the scaling parameter, and investigate the robustness of the density-difference and density-ratio estimators based on the density-power score and the pseudo-spherical score.

6.1. Non-Scale Models

The model satisfying the following assumption is called the *non-scale model*:

Assumption 1. Let \mathcal{M} be the model of density differences or density ratios. For $c \in \mathbb{R}$ and $f \in \mathcal{M}$, such that $c \neq 0$ and $f \neq 0$, $cf \in \mathcal{M}$ holds only when $c = 1$.

The density-power score and the pseudo-spherical score are the strict Bregman score on the non-scale models. Indeed, the density-power score is the strict Bregman score, as pointed out in Example 2. For the pseudo-spherical score, suppose that the equality $S(f, g) = S(f, f)$ holds for the non-zero functions, f and g . Then, g is proportional to f . When f and g are both included in a non-scale model, we have $f = g$. Thus, the pseudo-spherical score on the non-scale model is also the strict Bregman score.

6.2. Density-Difference Approach

Here, we consider the robustness of density-difference estimation using the non-scale models. Assumption 1 implies that the model, $f_\theta(x)$, does not include the scale parameter. An example is the model consisting of two probability models,

$$f_\theta(x) = p_{\theta_1}(x) - p_{\theta_2}(x), \quad \theta = (\theta_1, \theta_2),$$

such that $\theta_1 \neq \theta_2$, where p_{θ_1} and p_{θ_2} are parametric models of the normal distributions. The above model, f_θ , is still a semi-parametric model, because even when $f_\theta(x)$ is specified, the pair of probability densities, p and q , such that $f_\theta = p - q$, have infinitely many degrees of freedom.

The following theorem shows the robustness of the density-difference estimator. The proof is found in Appendix A.

Theorem 1. Suppose that Assumption 1 holds for the density-difference model, $\mathcal{M}_{\text{diff}}$. We assume that the true density-difference, f , is included in $\mathcal{M}_{\text{diff}}$ and that $f = f_{\theta^*} \in \mathcal{M}_{\text{diff}}$ holds. For the Bregman score, $S_{\text{diff}}(f, g)$, of the density difference, let J be the matrix, each element of which is given as:

$$J_{ij} = \frac{\partial}{\partial \theta_i \partial \theta_j} S_{\text{diff}}(f, f_{\theta}) \Big|_{\theta = \theta^*}.$$

Suppose that J is invertible. Then, under the model, $\mathcal{M}_{\text{diff}}$, the influence function of the density-power score with a positive odd parameter, α , is given as:

$$\begin{aligned} & \text{IF}_{\text{diff}}(\theta^*; z_p, z_q) \\ &= -\alpha(1 + \alpha)J^{-1} \left(f(z_p)^{\alpha-1} \frac{\partial f}{\partial \theta}(z_p; \theta^*) - f(z_q)^{\alpha-1} \frac{\partial f}{\partial \theta}(z_q; \theta^*) - \int f(x)^{\alpha} \frac{\partial f}{\partial \theta}(x; \theta^*) dx \right), \end{aligned} \quad (12)$$

where $\frac{\partial f}{\partial \theta}$ is the k -dimensional gradient vector of the function, f , with respect to the parameter, θ . In addition, we suppose that f is not the zero function. Then, under the model, $\mathcal{M}_{\text{diff}}$, the influence function of the pseudo-spherical score with a positive odd parameter, α , is given as:

$$\begin{aligned} & \text{IF}_{\text{diff}}(\theta^*; z_p, z_q) \\ &= \alpha(1 + \alpha)J^{-1} \\ & \times \left((f(z_p)^{\alpha} - f(z_q)^{\alpha}) \frac{\int f(x)^{\alpha} \frac{\partial f}{\partial \theta}(x; \theta^*) dx}{\int f(x)^{1+\alpha} dx} - f(z_p)^{\alpha-1} \frac{\partial f}{\partial \theta}(z_p; \theta^*) + f(z_q)^{\alpha-1} \frac{\partial f}{\partial \theta}(z_q; \theta^*) \right). \end{aligned} \quad (13)$$

Theorem 1 implies that the density-difference estimation with the pseudo-spherical score under non-scale models has the redescending property. For the density difference $f = p - q$, the limiting condition,

$$\lim_{\|x\| \rightarrow \infty} f(x) = 0,$$

will hold in many practical situations. Hence, for $\alpha > 1$, the assumption:

$$\lim_{\|x\| \rightarrow \infty} f_{\theta}(x)^{\alpha-1} \frac{\partial f}{\partial \theta}(x; \theta) = 0$$

for all $\theta \in \Theta_k$ will not be a strong condition for the density-difference model. Under the above limiting conditions, the influence Function (13) tends to zero, as z_p and z_q go to the infinite point. As a result, the pseudo-spherical score produces a redescending estimator. On the other hand, the density-power score does not have the redescending property, since the last term in Equation (12) does not vanish, when z_p and z_q tend to the infinite point.

Let us consider the L_1 -distance estimation using the density-difference estimator. The L_1 -distance estimator under the Contamination (11) is distributed around:

$$\int |f(x; \theta_{\varepsilon})| dx \approx \int \left| f(x; \theta^*) + \varepsilon \text{IF}_{\text{diff}}(\theta^*; z_p, z_q) \cdot \frac{\partial f}{\partial \theta}(x; \theta^*) \right| dx,$$

which implies that the bias term is expressed as the inner product of the influence function and the gradient of the density-difference model. Let $b_{\text{diff},\varepsilon}$ be:

$$b_{\text{diff},\varepsilon} = \varepsilon \int |\text{IF}_{\text{diff}}(\theta^*; z_p, z_q) \cdot \frac{\partial f}{\partial \theta}(x; \theta^*)| dx.$$

Then, the bias of the L_1 -distance estimator induced by outliers is approximately bounded above by $b_{\text{diff},\varepsilon}$. Since the pseudo-spherical score with the non-scale model provides the redescending estimator for the density difference, the L_1 -distance estimator based on the pseudo-spherical score also has the redescending property against the outliers.

6.3. Density-Ratio Approach

The following theorem provides the influence function of the density-ratio estimators. Since the proof is almost the same as that of Theorem 1, we omit the detailed calculation.

Theorem 2. *Suppose that Assumption 1 holds for the density-ratio model, $\mathcal{M}_{\text{ratio}}$. We assume that the true density-ratio $r(x) = p(x)/q(x)$ is included in:*

$$\mathcal{M}_{\text{ratio}} = \{r(x; \theta) = r_\theta(x) \mid \theta \in \Theta_k\},$$

and that $r = r_{\theta^*} \in \mathcal{M}_{\text{ratio}}$ holds. For the Bregman score, $S_{\text{ratio}}(r, r_\theta)$, with the base measure, $q(x)$, let J be the matrix, each element of which is given as:

$$J_{ij} = \frac{\partial}{\partial \theta_i \partial \theta_j} S_{\text{ratio}}(r, r_\theta) \Big|_{\theta = \theta^*}.$$

Suppose that J is invertible. Then, the influence function of the density-power score with a positive real parameter, α , is given as:

$$\text{IF}_{\text{ratio}}(\theta^*; z_p, z_q) = -\alpha(1 + \alpha)J^{-1} \left(r(z_p)^{\alpha-1} \frac{\partial r}{\partial \theta}(z_p; \theta^*) - r(z_q)^\alpha \frac{\partial r}{\partial \theta}(z_q; \theta^*) \right).$$

The influence function of the pseudo-spherical score with a positive real parameter, α , is given as:

$$\begin{aligned} & \text{IF}_{\text{ratio}}(\theta^*; z_p, z_q) \\ &= -\alpha(1 + \alpha)J^{-1} \\ & \times \left((r(z_p)^\alpha - r(z_q)^{\alpha+1}) \frac{\int r(x)^\alpha \frac{\partial r}{\partial \theta}(x; \theta^*) q(x) dx}{\int r(x)^{\alpha+1} q(x) dx} + r(z_p)^{\alpha-1} \frac{\partial r}{\partial \theta}(z_p; \theta^*) - r(z_q)^\alpha \frac{\partial r}{\partial \theta}(z_q; \theta^*) \right). \end{aligned}$$

The density ratio is a non-negative function. Hence, we do not need to care about the absolute value in the density-power score and the pseudo-spherical score. As a result, the parameter, α , in these scores is allowed to take any positive real number in the above theorem.

For the density ratio $r(x) = p(x)/q(x)$, a typical limiting condition is:

$$\lim_{\|x\| \rightarrow \infty} r(x) = \infty.$$

For example, the density ratio of two Gaussian distributions with the same variance and different means leads to an unbounded density ratio. Hence, the influence function can tend to infinity. As a result, the density-ratio estimator is sensitive to the shift in probability distributions.

Let us consider the L_1 -distance estimation using the density ratio. The L_1 -distance estimator under the Contamination (11) is distributed around:

$$\int |1 - r(y; \theta_\varepsilon)|q(y)dy \approx \int \left| 1 - r(y; \theta^*) - \varepsilon \text{IF}_{\text{ratio}}(\theta^*; z_p, z_q) \cdot \frac{\partial r}{\partial \theta}(y; \theta^*) \right| q(y)dy.$$

Thus, the bias of the L_1 -distance estimator induced by the outliers is approximately bounded above by:

$$b_{\text{ratio},\varepsilon} = \varepsilon \int \left| \text{IF}_{\text{diff}}(\theta^*; z_p, z_q) \cdot \frac{\partial r}{\partial \theta}(y; \theta^*) \right| q(y)dy.$$

The influence function for the density-ratio estimator can take an arbitrarily large value. In addition, the empirical approximation of the integral is also affected by outliers. Hence, the density-ratio estimator does not necessarily provide a robust estimator for the L_1 -distance measure.

7. Robustness under Scale Models

In this section, we consider the estimation of density differences using the model with the scale parameter. For such a model, the pseudo-spherical score does not work as shown in Example 3. Furthermore, in the previous section, we presented the instability of the density-ratio estimation against the gross outliers. Hence, in this section, we focus on the density-difference estimation using the density-power score with the scale models.

7.1. Decomposition of Density-Difference Estimation Procedure

We show that the estimation procedure of the density difference using the density-power score is decomposed into two steps: estimation using the pseudo-spherical score with the non-scale model and estimation of the scale parameter. Note that the estimation in the first step has already been investigated in the last section.

Let us consider the statistical model satisfying the following assumption:

Assumption 2. Let $\mathcal{M}_{\text{diff}}$ be the model for the density difference. For all $f \in \mathcal{M}_{\text{diff}}$ and all $c \in \mathbb{R}$, $cf \in \mathcal{M}_{\text{diff}}$ holds.

The model satisfying the above assumption is referred to as the *scale model*. A typical example of the scale model is the linear model:

$$\mathcal{M}_{\text{diff}} = \left\{ \sum_{\ell=1}^k \theta_\ell \psi_\ell(x) \mid \theta_\ell \in \mathbb{R}, \ell = 1, \dots, k \right\},$$

where ψ_ℓ is the basis functions, such that $\int \psi_\ell(x)dx = 0$ holds for all $\ell = 1, \dots, k$.

Suppose that the k -dimensional scale model, $\mathcal{M}_{\text{diff}}$, is parametrized as:

$$\mathcal{M}_{\text{diff}} = \{ f(x; \theta) = c g_{\bar{\theta}}(x) \mid c \in \mathbb{R}, \bar{\theta} \in \Theta_{k-1}, \theta = (c, \bar{\theta}) \}. \tag{14}$$

The parameter, c , is the scale parameter, and $\bar{\theta}$ in Equation (14) is called the shape parameter. We assume that any $g_{\bar{\theta}}$ is not equal to the zero-function. The parametrization of the model, $\mathcal{M}_{\text{diff}}$, may

not provide one-to-one correspondence between the parameter $\theta = (c, \bar{\theta})$ and the function, $cg_{\bar{\theta}}$, e.g., $c = 0$. We assume that in the vicinity of the true density-difference, the parametrization, θ , and the function, $cg_{\bar{\theta}}$, has a one-to-one correspondence. Define the model, $\mathcal{M}_{\text{diff},c}$, to be the $(k - 1)$ -dimensional non-scale model:

$$\mathcal{M}_{\text{diff},c} = \{cg_{\bar{\theta}}(x) \mid \bar{\theta} \in \Theta_{k-1}\}.$$

For the pseudo-spherical score, the equality $S(f, g) = S(f, cg)$ holds for $c > 0$. Thus, the scale parameter is not estimable. Let us study the statistical property of the estimator based on the density-power score with the scale models:

Theorem 3. *Let us consider the density-difference estimation. Define $S_{\text{diff},\alpha}^{\text{pow}}(f, g)$ and $S_{\text{diff},\alpha}^{\text{ps}}(f, g)$ as the density-power score and the pseudo-spherical score with a positive odd number, α , and the base measure of these scores is given by the Lebesgue measure, respectively. Let f_0 be a function and $\bar{c}g_{\bar{\theta}} \in \mathcal{M}_{\text{diff}}$ be the optimal solution of the problem,*

$$\min_f S_{\text{diff},\alpha}^{\text{pow}}(f_0, f), \quad \text{s.t. } f \in \mathcal{M}_{\text{diff}}.$$

We assume that $\bar{c}g_{\bar{\theta}} \neq 0$. Then, $g_{\bar{\theta}}$ is given as the optimal solution of:

$$\min_g S_{\text{diff},\alpha}^{\text{ps}}(f_0, g), \quad \text{s.t. } g \in \mathcal{M}_{\text{diff},c} \cup \mathcal{M}_{\text{diff},-c}, \tag{15}$$

where c is any fixed non-zero constant. In addition, the optimal scale parameter is expressed as:

$$\bar{c} = \int f_0(x)g_{\bar{\theta}}(x)^\alpha dx \Big/ \int g_{\bar{\theta}}(x)^{1+\alpha} dx.$$

The empirical density-difference, \tilde{f} , is allowed as the function, f_0 , in the above theorem. The proof is found in Appendix B. The same theorem for the non-negative functions is shown in [25].

Theorem 3 indicates that the minimization of the density-power score on the scale model is decomposed into two stages. Suppose that the true density-difference is $f_0 = p - q = c^*g_{\bar{\theta}^*} \in \mathcal{M}_{\text{diff}}$. At the first stage of the estimation, the minimization problem (15) is solved on the non-scale model $\mathcal{M}_{\text{diff},\pm c^*}$. Then, at the second stage, the scale parameter is estimated. Though c^* is unknown, the estimation procedure can be virtually interpreted as the two-stage procedure using the non-scale model, $\mathcal{M}_{\text{diff},\pm c^*}$.

7.2. Statistical Properties of Density-Difference Estimation

Based on the two-stage procedure for the minimization of the density-power score, we investigate the statistical properties of the density-difference estimator.

As shown in Section 6.2, the estimator using the pseudo-spherical score over the non-scale model, $\mathcal{M}_{\text{diff},c^*}$, has the redescending property. Hence, the extreme outliers have little impact on the estimation of the shape parameter, $\bar{\theta}$. Under the Contamination (11), let us define $\bar{\theta}_\epsilon$ as the optimal solution of the problem,

$$\min_g S_{\text{ps},\alpha}(p_\epsilon - q_\epsilon, g), \quad g \in \mathcal{M}_{\text{diff},c^*}.$$

As the outliers, z_p and z_q , tend to the infinite point, we have:

$$\bar{\theta}_\varepsilon = \bar{\theta}^* + o(\varepsilon),$$

because the estimation of the shape parameter using the pseudo-spherical score with the non-scale model has the redescending property, as shown in the last section.

The scale parameter is given as:

$$c_\varepsilon = \frac{\int (p_\varepsilon(x) - q_\varepsilon(x))g_{\bar{\theta}_\varepsilon}(x)^\alpha dx}{\int g_{\bar{\theta}_\varepsilon}(x)^{1+\alpha} dx} = (1 - \varepsilon) \frac{\int f(x)g_{\bar{\theta}_\varepsilon}(x)^\alpha dx}{\int g_{\bar{\theta}_\varepsilon}(x)^{1+\alpha} dx} + \varepsilon \frac{g_{\bar{\theta}_\varepsilon}(z_p)^\alpha - g_{\bar{\theta}_\varepsilon}(z_q)^\alpha}{\int g_{\bar{\theta}_\varepsilon}(x)^{1+\alpha} dx}.$$

As the outliers, z_p and z_q , tend to the infinite point, the second term in the above expression converges to zero. Hence, the scale parameter is given as:

$$c_\varepsilon = (1 - \varepsilon)c^* + o(\varepsilon),$$

from which we have:

$$c_\varepsilon g_{\bar{\theta}_\varepsilon} = (1 - \varepsilon)(p - q) + o(\varepsilon).$$

The above analysis shows that the extreme outliers with the small contamination ratio, ε , will make the intensity of the estimated density-difference smaller by the factor, $1 - \varepsilon$, and the estimated density-difference is distributed around $(1 - \varepsilon)(p - q)$. Hence, the contamination with extreme outliers has little impact on the shape parameter in the density-difference estimator, when the density-power score is used.

Let us consider the L_1 -distance estimation using the density-difference estimator. Suppose that the true density-difference, $p - q$, is estimated by the density-power score with the scale model. Then, the L_1 -distance estimator under the Contamination (11) is distributed around:

$$\int |c_\varepsilon g_{\bar{\theta}_\varepsilon}(x)| dx = (1 - \varepsilon) \int |p(x) - q(x)| dx + o(\varepsilon). \tag{16}$$

When the contamination ratio, ε , is small, even extreme outliers do not significantly affect the L_1 -distance estimator. The bias induced by the extreme outliers depends only on the contamination ratio. If prior knowledge on the contamination ratio, ε , is available, one can approximately correct the bias of the L_1 -distance measure by multiplying the constant obtained by the prior knowledge to the L_1 -distance estimator.

8. Numerical Experiments

We conducted numerical experiments to evaluate the statistical properties of L_1 -distance estimators. We used synthetic datasets. Let $N(\mu, \sigma^2)$ be the one-dimensional normal distribution with mean μ and variance σ^2 . In the standard setup, let us assume that the samples are drawn from the normal distributions,

$$x_1, \dots, x_n \sim N(0, 1), \quad y_1, \dots, y_m \sim N(1, 1).$$

In addition, some outliers are observed from:

$$\tilde{x}_1, \dots, \tilde{x}_{n'} \sim N(0, \tau^2), \quad \tilde{y}_1, \dots, \tilde{y}_{m'} \sim N(0, \tau^2),$$

where the variance, τ , is much larger than one. Based on the two datasets, $\{x_1, \dots, x_n, \tilde{x}_1, \dots, \tilde{x}_{n'}\}$ and $\{y_1, \dots, y_m, \tilde{y}_1, \dots, \tilde{y}_{m'}\}$, the L_1 -distance between $N(0, 1)$ and $N(1, 1)$ is estimated.

Below, we show the models and estimators used in the L_1 -distance estimation. The non-scale model for the density-difference is defined as:

$$f_\theta(x) = \phi(x; \mu_1, \sigma_1) - \phi(x; \mu_2, \sigma_2), \quad \theta = (\mu_1, \mu_2, \sigma_1, \sigma_2),$$

where $\phi(x; \mu, \sigma)$ is the probability density of $N(\mu, \sigma^2)$. To estimate the parameters, the density-power score and the pseudo-spherical score are used. As the scale model, we employ $cf_\theta(x)$ with the parameter, θ , and $c > 0$. As shown in Equation (16), the estimator has the bias, when the samples are contaminated by outliers. Ideally, the multiplication of $(1 - \varepsilon)^{-1}$ to the L_1 -distance estimator with the scale model will improve the estimation accuracy, where ε is the contamination ratio $n'/n (= m'/m)$. In the numerical experiments, also, the bias corrected estimator was examined, though the bias correction requires prior knowledge on the contamination ratio. For the statistical model for density-ratio estimation, we used the scale model,

$$r(x; \theta) = \exp\{\theta_0 + \theta_1 x + \theta_2 x^2\}, \quad \theta \in \mathbb{R}^3,$$

and the density-power score as the loss function. Furthermore, we evaluated the two-step approach in which the L_1 -distance is estimated from the separately estimated probability densities. We employed the density-power score with the statistical model, $\phi(x; \mu, \sigma)$, to estimate the probability density of each dataset.

In numerical experiments, the error of the L_1 -distance estimator, $\hat{d}_1(p, q)$, was measured by the relative error, $|1 - \hat{d}_1(p, q)/d_1(p, q)|$. The number of training samples varied from 1,000 to 10,000, and that of the outliers varied from zero (no outlier) to 100. The parameter, α , in the score function was set to $\alpha = 1$ or 3 for the density-difference (DF)-based estimators and $\alpha = 0.1$ for the density-ratio (DR)-based estimators. For the density-ratio estimation, the score with large α easily yields numerical errors, since the power of the exponential model tends to become extremely large. For each setup, the averaged relative error of each estimator was computed over 100 iterations.

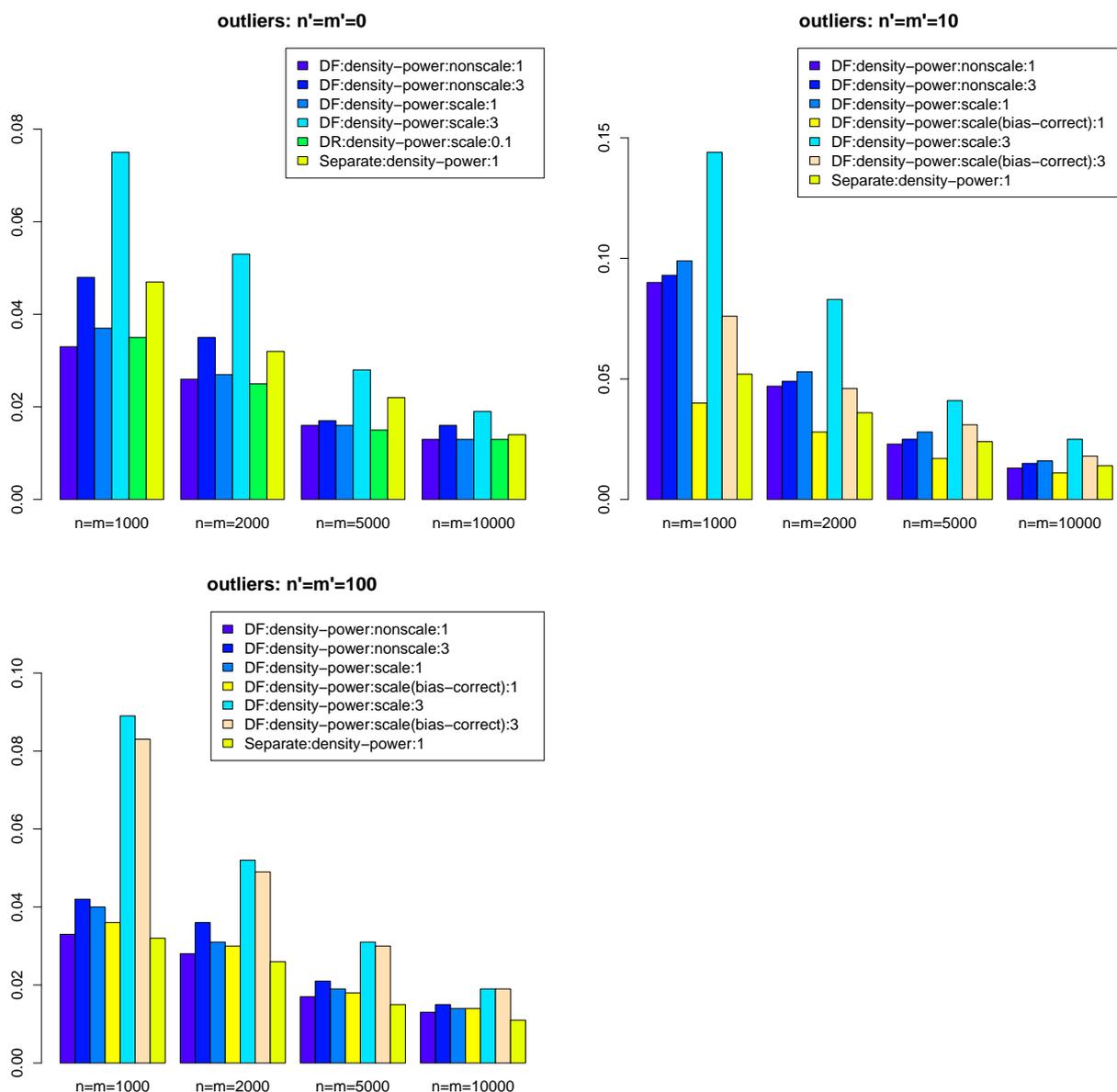
Table 1. Averaged relative error and standard deviation of L_1 -distance estimators over 100 iterations: DF (DR) denotes the estimator based on the density-difference (density-ratio), and “separate” denotes the L_1 -distance estimator using the separately estimated probability densities. The density-power score or pseudo-spherical score with parameter α is employed with the scale or non-scale model. The number of samples in the standard setup is $n = m = 1,000, 2,000, 5,000$, or $10,000$, and the number of outliers is set to $n' = m' = 0$ (i.e., no outliers), 10 or 100. When the samples are contaminated by outliers, the density-ratio-based estimator becomes extremely unstable and numerical error occurs.

outliers: $n' = m' = 0$ (no outlier)					
DF/DR estimator:model	α	$n = m = 1,000$	$n = m = 2,000$	$n = m = 5,000$	$n = m = 10,000$
DF density-power:nonscale	1	0.033 (0.028)	0.026 (0.024)	0.016 (0.014)	0.013 (0.010)
DF density-power:nonscale	3	0.048 (0.032)	0.035 (0.029)	0.017 (0.014)	0.016 (0.010)
DF density-power:scale	1	0.037 (0.030)	0.027 (0.025)	0.016 (0.013)	0.013 (0.010)
DF density-power:scale	3	0.075 (0.069)	0.053 (0.058)	0.028 (0.027)	0.019 (0.017)
DF pseudo-sphere:nonscale	1	0.610 (0.450)	0.604 (0.396)	0.451 (0.320)	0.452 (0.294)
DF pseudo-sphere:nonscale	3	0.782 (0.532)	0.739 (0.491)	0.604 (0.440)	0.500 (0.379)
DR density-power:scale	0.1	0.035 (0.026)	0.025 (0.022)	0.015 (0.013)	0.013 (0.009)
Separate:density-power	1	0.047 (0.038)	0.032 (0.024)	0.022 (0.017)	0.014 (0.010)

outliers: $n' = m' = 10, \tau = 100$					
DF/DR estimator:model	α	$n = m = 1,000$	$n = m = 2,000$	$n = m = 5,000$	$n = m = 10,000$
DF density-power:nonscale	1	0.033 (0.026)	0.028 (0.022)	0.017 (0.013)	0.013 (0.010)
DF density-power:nonscale	3	0.042 (0.033)	0.036 (0.029)	0.021 (0.016)	0.015 (0.012)
DF density-power:scale	1	0.040 (0.030)	0.031 (0.025)	0.019 (0.014)	0.014 (0.011)
DF density-power:scale (bias-correct)	1	0.036 (0.030)	0.030 (0.025)	0.018 (0.014)	0.014 (0.011)
DF density-power:scale	3	0.089 (0.077)	0.052 (0.047)	0.031 (0.024)	0.019 (0.016)
DF density-power:scale (bias-correct)	3	0.083 (0.075)	0.049 (0.046)	0.030 (0.023)	0.019 (0.016)
DF pseudo-sphere:nonscale	1	0.658 (0.474)	0.632 (0.424)	0.515 (0.370)	0.417 (0.297)
DF pseudo-sphere:nonscale	3	0.969 (0.494)	0.743 (0.487)	0.677 (0.483)	0.506 (0.421)
DR density-power:scale	0.1	–	–	–	–
Separate:density-power	1	0.032 (0.023)	0.026 (0.019)	0.015 (0.011)	0.011 (0.008)

outliers: $n' = m' = 100, \tau = 100$					
DF/DR estimator:model	α	$n = m = 1,000$	$n = m = 2,000$	$n = m = 5,000$	$n = m = 10,000$
DF density-power:nonscale	1	0.090 (0.042)	0.047 (0.028)	0.023 (0.014)	0.013 (0.010)
DF density-power:nonscale	3	0.093 (0.053)	0.049 (0.032)	0.025 (0.020)	0.015 (0.012)
DF density-power:scale	1	0.099 (0.043)	0.053 (0.029)	0.028 (0.017)	0.016 (0.011)
DF density-power:scale (bias-correct)	1	0.040 (0.031)	0.028 (0.022)	0.017 (0.013)	0.011 (0.009)
DF density-power:scale	3	0.144 (0.100)	0.083 (0.047)	0.041 (0.030)	0.025 (0.016)
DF density-power:scale (bias-correct)	3	0.076 (0.094)	0.046 (0.041)	0.031 (0.025)	0.018 (0.014)
DF pseudo-sphere:nonscale	1	0.557 (0.461)	0.511 (0.399)	0.501 (0.372)	0.465 (0.305)
DF pseudo-sphere:nonscale	3	0.807 (0.507)	0.739 (0.508)	0.581 (0.458)	0.534 (0.396)
DR density-power:scale	0.1	–	–	–	–
Separate:density-power	1	0.052 (0.036)	0.036 (0.031)	0.024 (0.017)	0.014 (0.009)

Figure 1. Averaged relative error of L_1 -distance estimators over 100 iterations are plotted. The estimators with extremely large relative errors are omitted. DF denotes the estimator based on the density-difference, and the rightmost number in the legend denotes α in the density-power score. The number of samples in the standard setup is $n = m = 1,000, 2,000, 5,000, \text{ or } 10,000$, and the number of outliers is set to $n' = m' = 0$ (*i.e.*, no outliers), 10 or 100.



The numerical results are depicted in Figure 1, and details are shown in Table 1. In the figure, estimators with extremely large relative errors are omitted. As shown in Table 1, the estimation accuracy of the DR-based estimator was severely degraded by the contaminated samples. On the other hand, DF-based estimators were robust against outliers. The DF-based estimator with the pseudo-spherical score is less accurate than that with the density-power score. In the statistical inference, there is the trade-off between the efficiency and robustness. Though the pseudo-spherical score provides a redescending estimator, the efficiency of the estimator is not high in practice. In the estimation of the probability

density, the pseudo-spherical score having the parameter, α , ranging from 0.1 to one provides a robust and efficient estimator, and the estimator with large α became inefficient [23]. This is because the estimator with large α tends to ignore most of the samples. In the density-difference estimation, the parameter, α , should be a positive odd number. Hence, in our setup, the estimator using the pseudo-spherical score became inefficient. In terms of the density-power score, the corresponding DF-based estimator has the bounded influence function. As a result, the estimator is efficient and rather robust against the outliers. Furthermore, we found that the bias correction by multiplying the constant factor, $(1 - \epsilon)^{-1}$, improves the estimation accuracy.

When there is no outlier, the two-step approach using the separately estimated probability densities has larger relative errors than the DF-based estimators using the density-power score. For the contaminated samples, the two-step approach is superior to the other methods, especially when the sample size is less than 2,000. In this case, the separate density estimation with the density-power score efficiently reduces the influence of the outliers. For the larger sample size, however, the DF-based estimators using the density-power score are comparable with the two-step approach. When the rate of the outliers is moderate, the DF-based approach works well, even though the statistical model is based on the semiparametric modeling, which has less information than the parametric modeling used in the two-step approach.

9. Conclusions

In this paper, we first proposed to use the Bregman score to estimate density differences and density ratios, and then, we studied the robustness property of the L_1 -distance estimator. We showed that the pseudo-spherical score provides a redescending estimator of the density difference under non-scale models. However, the estimator based on the density-power score does not have the redescending property against extreme outliers. In the scale models, the pseudo-spherical score does not work, since the corresponding potential is not strictly convex on the function space. We proved that the density-power score provides a redescending estimator for the shape parameter in the scale models. Under extreme outliers, the shift in the L_1 -distance estimator using the scale model is calculated. The density-power score provides a redescending estimator for the shape parameter in the scale models. Moreover, we proved that the L_1 -distance estimator is not significantly affected by extreme outliers. In addition, we showed that prior knowledge on the contamination ratio, ϵ , can be used to correct the bias of the L_1 -distance estimator. In numerical experiments, the density-power score provides an efficient and robust estimator in comparison to the pseudo-spherical score. This is because the pseudo-spherical score with large α tends to ignore most of the samples and, thus, becomes inefficient. In a practical setup, the density-power score will provide a satisfactory result. Furthermore, we illustrated that the bias correction by using the prior knowledge on the contamination ratio improves L_1 -distance estimators using scale models.

Besides the Bregman scores, there are other useful classes of estimators, such as local scoring rules [12,18,30]. It is therefore an interesting direction to pursue the possibility of applying another class of scoring rules to the estimation of density differences and density ratios.

A. Proof of Theorem 1

For the density-difference $f(x) = f_{\theta^*}(x) = p(x) - q(x)$, we define $f_\epsilon(x)$ as the contaminated density-difference,

$$\begin{aligned} f_\epsilon(x) &= (1 - \epsilon)p(x) + \epsilon\delta(x - z_p) - (1 - \epsilon)q(x) - \epsilon\delta(x - z_q) \\ &= f(x) + \epsilon \{ \delta(x - z_p) - \delta(x - z_q) - f(x) \}. \end{aligned}$$

Let g be the function $g(x) = \delta(x - z_p) - \delta(x - z_q) - f_{\theta^*}(x)$. By using the implicit function theorem to the \mathbb{R}^k -valued function:

$$(\theta, \epsilon) \mapsto \frac{\partial}{\partial \theta} S(f_\epsilon, f_\theta) = \frac{\partial}{\partial \theta} \int \ell(x, f_\theta) f_\epsilon(x) dx$$

around $(\theta, \epsilon) = (\theta^*, 0)$, we have:

$$\text{IF}_{\text{diff}}(\theta^*; z_p, z_q) = -J^{-1} \frac{\partial}{\partial \theta} S(g, f_\theta) \Big|_{\theta=\theta^*}.$$

The computation of the above derivative for each score yields the results.

B. Proof of Theorem 3

Let us consider the minimization of $S_{\text{diff},\alpha}^{\text{pow}}(f_0, f)$ subject to $f \in \mathcal{M}_{\text{diff}}$. For $c_g \in \mathcal{M}_{\text{diff}}$, we have:

$$S_{\text{diff},\alpha}^{\text{pow}}(f_0, c_g) = \alpha c^{1+\alpha} \int g(x)^{1+\alpha} dx - (1 + \alpha)c^\alpha \int f_0(x)g(x)^\alpha dx$$

For a fixed $g \in \mathcal{M}_{\text{diff},1}$, the minimizer of $S_{\text{diff},\alpha}^{\text{pow}}(f_0, c_g)$ with respect to $c \in \mathbb{R}$ is given as:

$$c_g = \int f_0(x)g(x)^\alpha dx \Big/ \int g(x)^{1+\alpha} dx, \tag{17}$$

since g is not the zero function. Substituting the optimal c_g into $S_{\text{diff},\alpha}^{\text{pow}}(f_0, c_g)$, we have:

$$S_{\text{diff},\alpha}^{\text{pow}}(f_0, c_g) = - \frac{\left(\int f_0(x)g(x)^\alpha dx \right)^{1+\alpha}}{\left(\int g(x)^{1+\alpha} dx \right)^\alpha} = -(S_{\text{diff},\alpha}^{\text{ps}}(f_0, g))^{1+\alpha}$$

for the positive odd number, α . Hence, the optimal solution of $S_{\text{diff},\alpha}^{\text{pow}}(f_0, c_g)$ subject to $c_g g \in \mathcal{M}_{\text{diff}}$ is obtained by solving:

$$\min_g S_{\text{diff},\alpha}^{\text{ps}}(f_0, g), \quad g \in \mathcal{M}_{\text{diff},c} \cup \mathcal{M}_{\text{diff},-c} \tag{18}$$

where c is any fixed non-zero number. Since $\alpha + 1$ is an even number, we need to take into account two sub-models, $\mathcal{M}_{\text{diff},c}$ and $\mathcal{M}_{\text{diff},-c}$, in order to reduce the optimization of $S_{\text{diff},\alpha}^{\text{pow}}$ to $S_{\text{diff},\alpha}^{\text{ps}}$.

Acknowledgments

TK was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant number 24500340, and MS was partially supported by JSPS KAKENHI Grant number 25700022 and Asian Office of Aerospace Research and Development (AOARD).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Sugiyama, M.; Liu, S.; du Plessis, M.C.; Yamanaka, M.; Yamada, M.; Suzuki, T.; Kanamori, T. Direct divergence approximation between probability distributions and its applications in machine learning. *J. Comput. Sci. Eng.* **2013**, *7*, 99–111.
2. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Infer.* **2000**, *90*, 227–244.
3. Sugiyama, M.; Kawanabe, M. *Machine Learning in Non-Stationary Environments : Introduction to Covariate Shift Adaptation (Adaptive Computation and Machine Learning)*; MIT Press: Cambridge, MA, USA, 2012.
4. Sugiyama, M.; Suzuki, T.; Kanamori, T. *Density Ratio Estimation in Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
5. Hido, S.; Tsuboi, Y.; Kashima, H.; Sugiyama, M.; Kanamori, T. Inlier-based Outlier Detection via Direct Density Ratio Estimation. In Proceedings of IEEE International Conference on Data Mining (ICDM2008), Pisa, Italy, 15–19 December 2008.
6. Kanamori, T.; Suzuki, T.; Sugiyama, M. f-Divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. Inform. Theor.* **2012**, *58*, 708–720.
7. Kanamori, T.; Hido, S.; Sugiyama, M. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In *Advances in Neural Information Processing Systems 21*; MIT Press: Cambridge, MA, USA, 2009.
8. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inform. Theor.* **2010**, *56*, 5847–5861.
9. Qin, J. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **1998**, *85*, 619–639.
10. Sugiyama, M.; Kanamori, T.; Suzuki, T.; du Plessis, M.C.; Liu, S.; Takeuchi, I. Density-difference estimation. *Neural. Comput.* **2013**, *25*, 2734–2775.
11. Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Büna, P.; Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Ann. Inst. Stat. Math.* **2008**, *60*, 699–746.
12. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
13. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from Another. *J. Roy. Stat. Soc. Series B* **1966**, *28*, 131–142.

14. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
15. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3.
16. Good, I.J. Comment on “Measuring Information and Uncertainty” by R. J. Buehler. In *Foundations of Statistical Inference*; Godambe, V.P., Sprott, D.A., Eds.; Dove: Mineola, NY, USA, 1971; p. 337339.
17. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U -Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
18. Parry, M.; Dawid, A.P.; Lauritzen, S. Proper local scoring rules. *Ann. Stat.* **2012**, *40*, 561–592.
19. Hendrickson, A.D.; Buehler, R.J. Proper scores for probability forecasters. *Ann. Mathe. Stat.* **1971**, *42*, 19161921.
20. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: Landon, UK, 2006.
21. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
22. Basu, A.; Shioya, H.; Park, C. Monographs on statistics and applied probability. In *Statistical Inference: The Minimum Distance Approach*; Taylor & Francis: Landon, UK, 2010.
23. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
24. Hampel, F.R.; Rousseeuw, P.J.; Ronchetti, E.M.; Stahel, W.A. *Robust Statistics. The Approach Based on Influence Functions*; John Wiley and Sons, Inc.: Landon, UK, 1986.
25. Eguchi, S.; Kato, S. Entropy and divergence associated with power function and the statistical application. *Entropy* **2010**, *12*, 262–274.
26. Maronna, R.; Martin, R.; Yohai, V. *Robust Statistics: Theory and Methods*; Wiley: Landon, UK, 2006.
27. Wu, Y.; Liu, Y. Robust truncated hinge loss support vector machines. *J. Am. Stat. Assoc.* **2007**, *102*, 974–983.
28. Xu, H.; Caramanis, C.; Mannor, S.; Yun, S. Risk Sensitive Robust Support Vector Machines. In Proceedings of the 48th IEEE Conference on Decision Control, Shanghai, China, 15–18 December 2009; pp. 4655–4661.
29. Xu, L.; Crammer, K.; Schuurmans, D. *Robust Support Vector Machine Training via Convex Outlier Ablation*; AAAI: Boston, MA, USA, 2006; pp. 536–542.
30. Kanamori, T.; Fujisawa, H. Affine invariant divergences associated with composite scores and its applications. *Bernoulli* **2014**, submitted.