Density Ratio Estimation with Doubly Strong Robustness

Ryosuke Nagumo [1 2] **Hironori Fujisawa** [3 1]

Abstract

We develop two density ratio estimation (DRE) methods with robustness to outliers. These are based on the divergence with a weight function to weaken the adverse effects of outliers. One is based on the Unnormalized Kullback-Leibler divergence, called Weighted DRE, and its optimization is a convex problem. The other is based on the γ-divergence, called γ-DRE, which improves a normalizing term problem of Weighted DRE. Its optimization is a DC (Difference of Convex functions) problem and needs more computation than a convex problem. These methods have doubly strong robustness, which means robustness to the heavy contamination of both the reference and target distributions. Numerical experiments show that our proposals are more robust than the previous methods.

1. Introduction

Density ratio estimation (DRE) is a statistical method to directly estimate the ratio of two probability density functions without estimating each function (Nguyen et al., 2007; Sugiyama et al., 2012b). DRE is used in many applications, including change detection (Kawahara & Sugiyama, 2009; Liu et al., 2013), outlier detection (Hido et al., 2011), covariate shift adaptation (Shimodaira, 2000; Zhang et al., 2023), and two-sample test (Wornowizki & Fried, 2016; Kim et al., 2021). Its parametric formulation is also called the differential graphical model (Liu et al., 2014; 2017a;b). Differential graphical models are used in protein and genetic interaction mapping (Ideker & Krogan, 2012) and brain imaging (Na et al., 2020).

Density ratio estimation is not robust when the data exist in a region where the density function values are small (Smola et al., 2009; Yamada et al., 2011). Consider a density ratio $r(x) = p(x)/q(x)$, where we call $p(x)$ a reference distribution and $q(x)$ a target distribution. When the data exist in a region where $p(x)$ or $q(x)$ is small, the density ratio $r(x)$ tends to be wrongly estimated. This situation occurs, for example, when $p(x)$ and $q(x)$ have no common support, which is a usual case in the high-dimensional setting (Rhodes et al., 2020; Kato & Teshima, 2021; Choi et al., 2021; 2022; Srivastava et al., 2023).

Our interest is in the case where outliers contaminate the main distributions (Maronna et al., 2006; Hampel et al., 2011). Because outliers exist in a region where the density functions of the main distribution are small, the outliers have adverse effects on the estimation of the density ratio.

For density estimation, many robust methods have been proposed, including the Huber loss (Huber, 1964), the trimming estimator (Hadi & Luceno, 1997), the density power divergence (Basu et al., 1998), and the γ-divergence (Fujisawa & Eguchi, 2008). These robust estimation methods realized real-world applications, including yeast gene expression (Yang & Lozano, 2015) and gene function regulation (Hirose et al., 2017).

For the outlier-robust density ratio estimation, only the limited research has been done as far as we know (Sugiyama et al., 2012a). The most promising method is Trimmed DRE (Liu et al., 2017c). This method assumes that outliers have larger density ratio values than inliers. It also assumes that only the reference dataset is contaminated while the target dataset remains clean. Under these assumptions, Trimmed DRE is shown to be robust experimentally and theoretically.

Although Trimmed DRE is shown to be robust, there are some limitations to its robustness. The density ratio values of outliers do not necessarily have larger values than inliers. That breaks the assumption that this method relies on. Besides, Trimmed DRE is not robust to the contamination of the target dataset, which restricts its usage in real-world applications. For example, in the change point detection in time series data, a sequence is divided into subsequences and assigned to the reference and target datasets sequentially (Liu et al., 2017a). Then, both the reference and target datasets can include outliers. The robustness to the reference contamination alone is insufficient for time series applications.

[1]The Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan [2]Panasonic Holdings Corporation, Osaka, Japan [3]Institute of Statistical Mathematics, Tokyo, Japan. Correspondence to: Ryosuke Nagumo <nagumo@ism.ac.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

To overcome these limitations, we develop density ratio estimation methods with doubly strong robustness, where *doubly* means two types of contamination of the reference and target datasets, and *strong* means independence from the contamination ratio. We propose Weighted DRE and γ-DRE, which realize doubly strong robustness. Both methods employ a weight function which weakens the adverse effects of outliers. Weighted DRE minimizes the Unnormalized Kullback-Leibler (UKL) divergence, and the minimization is a convex problem. γ-DRE minimizes the γ-divergence to overcome a drawback of Weighted DRE in estimating the normalizing term. Its minimization is a DC (Difference of Convex functions) problem but needs more computation than Weighted DRE. These methods firstly achieve doubly strong robustness as far as we know.

This paper is organized as follows. Section 2 introduces the conventional DRE method from the viewpoint of the density ratio function and the discrepancy measure. In Section 3, we propose two DRE methods with robustness, Weighted DRE and γ-DRE, and theoretically show that they have doubly strong robustness. In Section 4, numerical experiments illustrate that the proposed methods are more robust than the past ones.

2. Density Ratio Estimation

The density ratio is defined as the ratio of two density functions. Let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be the strictly positive density functions of the reference and target datasets, respectively, for $\boldsymbol{x} \in \mathbb{R}^d$. The true density ratio can be written as $r(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}$. To estimate the density ratio, we employ the density ratio function $r_{\boldsymbol{\beta}}(\boldsymbol{x})$, where $\boldsymbol{\beta}$ is a parameter, and then measure the discrepancy between the true density ratio $r(\boldsymbol{x})$ and the density ratio function $r_{\boldsymbol{\beta}}(\boldsymbol{x})$. The choice of the density ratio function and the discrepancy measure realizes the various DRE methods.

2.1. Density Ratio Function

The formulation of the density ratio function is categorized into three patterns: parametric models, non-parametric models, and deep models.

For the parametric models (Liu et al., 2014; 2017b), the density ratio function is defined as

$$r_{\boldsymbol{\theta},C}(\boldsymbol{x}) = C r_{\boldsymbol{\theta}}(\boldsymbol{x}) = C \exp\left(\boldsymbol{\theta}^T h(\boldsymbol{x})\right), \qquad (1)$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is the difference parameter, $C \in \mathbb{R}$ is the normalizing term, and $h(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^p$ is the feature transform function. A detailed discussion of this formulation will be given in Appendix A. This parametric function is useful for sparse estimation in high-dimensional settings (Liu et al., 2017b).

For the non-parametric models, the density ratio function can be defined by the linear model (Sugiyama et al., 2008; Kanamori et al., 2009; 2012) as

$$r^{\mathrm{lm}}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \psi(\boldsymbol{x}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^b$ is a weight parameter and $\psi(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^b$ is a basis function. Another formulation is the log-linear model (Tsuboi et al., 2009; Kanamori et al., 2010) defined as

$$r^{\mathrm{llm}}_{\boldsymbol{\theta},C}(\boldsymbol{x}) = C \exp\left(\boldsymbol{\theta}^T \psi(\boldsymbol{x})\right).$$

The standard choice of the basis function is $\boldsymbol{\theta}^T \psi(\boldsymbol{x}) = \sum_{i=1}^{n_p} \theta_i K(\boldsymbol{x}, \boldsymbol{x}_i^{(p)})$, where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel and $\boldsymbol{x}_i^{(p)}$ for $i = 1, ..., n_p$ are the data points in the reference dataset. These non-parametric models are useful for complex distributions (Sugiyama et al., 2012a).

For the deep models, some density ratio methods using deep neural nets have been proposed (Rhodes et al., 2020; Kato & Teshima, 2021; Choi et al., 2021; 2022; Srivastava et al., 2023). These methods are useful for high-dimensional and unstructured data such as images.

2.2. Discrepancy Measure

Statistical divergences are reasonable choices to measure the discrepancy between the true density ratio $r(\boldsymbol{x})$ and the density ratio function $r_{\boldsymbol{\beta}}(\boldsymbol{x})$. The most famous one is the Bregman (BR) divergence (Bregman, 1967; Sugiyama et al., 2012a; Kato & Teshima, 2021). Let f be a differentiable and strictly convex function with the derivative ∂f. Then, we quantify the discrepancy of $r(\boldsymbol{x})$ and $r_{\boldsymbol{\beta}}(\boldsymbol{x})$ as

$$\begin{aligned} & D_{\mathrm{BR}}(r, r_{\boldsymbol{\beta}}) \\ &= \int [f\left(r(\boldsymbol{x})\right) - f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right) \\ &\qquad - \partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right)\left(r(\boldsymbol{x}) - r_{\boldsymbol{\beta}}(\boldsymbol{x})\right)] q(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int [\partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right) r_{\boldsymbol{\beta}}(\boldsymbol{x}) - f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right)] q(\boldsymbol{x}) d\boldsymbol{x} \\ &\qquad - \int \partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right) p(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{const.} \end{aligned}$$

The Bregman divergence is a general expression, and the choice of f realizes various methods. For example, the UKL (Unnormalized Kullback-Leibler) divergence (Nguyen et al., 2007) and KLIEP (Kullback-Leibler Importance Estimation Procedure) (Sugiyama et al., 2008) adopt $f(t) = t \log t - t$, LSIF (Least-Squares Importance Fitting) (Kanamori et al., 2009) and KMM (Kernel Mean Matching) (Gretton et al., 2009) adopt $f(t) = (t-1)^2/2$, and the BKL (Binary Kullback-Leibler) divergence (Hastie et al., 2001) adopts $f(t) = t \log t - (1+t)\log(1+t)$.

Given two datasets, $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p} \sim_{i.i.d.} p(\boldsymbol{x})$ for the reference and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q} \sim_{i.i.d.} q(\boldsymbol{x})$ for the target, the Bregman

divergence without the constant term is approximated by

$$\hat{D}_{\mathrm{BR}}(r, r_{\boldsymbol{\beta}}) = \frac{1}{n_q}\sum_{i=1}^{n_q}\left\{\partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x}_i^{(q)})\right) r_{\boldsymbol{\beta}}(\boldsymbol{x}_i^{(q)}) - f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x}_i^{(q)})\right)\right\} - \frac{1}{n_p}\sum_{i=1}^{n_p}\partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x}_i^{(p)})\right).$$

3. Robust Density Ratio Estimation

Our idea to achieve a robust estimation is to introduce a weight function $w(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}_+$ to reduce the adverse effects of outliers in the estimator (Maronna et al., 2006). We propose Weighted DRE and γ-DRE to realize this idea.

3.1. Weighted DRE

3.1.1. Formulation

The Bregman divergence can include the weight function as the base measure:

$$D_{\mathrm{BR}}(r, r_{\boldsymbol{\beta}}; w) = \int [\partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right) r_{\boldsymbol{\beta}}(\boldsymbol{x}) - f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right)] q(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} - \int \partial f\left(r_{\boldsymbol{\beta}}(\boldsymbol{x})\right) p(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{const.}$$

With the base measure $w(\boldsymbol{x})d\boldsymbol{x}$, the Bregman divergence still has the following properties: (i) $D_{\mathrm{BR}}(r, r_{\boldsymbol{\beta}}; w) \geq 0$, (ii) $D_{\mathrm{BR}}(r, r_{\boldsymbol{\beta}}; w) = 0 \Leftrightarrow r = r_{\boldsymbol{\beta}}$.

Some combinations of the density ratio function $r_{\boldsymbol{\beta}}$ and the convex function f can realize the robust estimation. Here, we show an example of the parametric DRE with the UKL divergence. When adopting the parametric density ratio function (1) and $f(t) = t\log t - t$, the formulation of the UKL divergence can be given as

$$\begin{aligned} D_{\mathrm{UKL}}(r, r_{\boldsymbol{\theta},C}; w) &= \int r_{\boldsymbol{\theta},C} q(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} - \int \log r_{\boldsymbol{\theta},C} p(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} \\ &= C\int \exp\left(\boldsymbol{\theta}^T h(\boldsymbol{x})\right) q(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad - \int \left(\boldsymbol{\theta}^T h(\boldsymbol{x}) + \log C\right) p(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{const.} \end{aligned} \tag{2}$$

Because (2) is convex about C, the optimal normalizing term is

$$C^* = \frac{\int w(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}}{\int \exp(\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}}. \tag{3}$$

Then, the parameter $\boldsymbol{\theta}$ is estimated by minimizing

$$\begin{aligned} D_{\mathrm{UKL}}(r, r_{\boldsymbol{\theta}}; w) &= D_{\mathrm{UKL}}(r, r_{\boldsymbol{\theta},C^*}; w) \\ &= -\int \boldsymbol{\theta}^T h(\boldsymbol{x}) w(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad + \int w(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad \times \log \int \exp(\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{const.} \end{aligned} \tag{4}$$

This UKL divergence is empirically approximated without the constant term by two datasets $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}$ and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q}$:

$$\hat{D}_{\mathrm{UKL}}(r, r_{\boldsymbol{\theta}}; w) = -\frac{1}{n_p}\sum_{i=1}^{n_p}\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(p)}) w(\boldsymbol{x}_i^{(p)}) + \frac{1}{n_p}\sum_{i=1}^{n_p} w(\boldsymbol{x}_i^{(p)}) \times \log \frac{1}{n_q}\sum_{i=1}^{n_q}\exp(\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)})) w(\boldsymbol{x}_i^{(q)}). \tag{5}$$

This objective function is convex about $\boldsymbol{\theta}$ and can be minimized via gradient descent. For sparse estimation, we can add a regularization term, for example, $\lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2\|\boldsymbol{\theta}\|_2^2$ with the positive tuning parameters λ_1 and λ_2 (Zou & Hastie, 2005). The more efficient optimization method via the Lagrangian dual problem is discussed in Appendix B.1. We call this estimation procedure Weighted DRE.

3.1.2. Doubly Strong Robustness

Let us consider the robustness of the estimator given by minimizing the UKL divergence (4). Suppose that the reference and target datasets are contaminated by outliers, more precisely, drawn from the contaminated distributions (Huber, 2004; Maronna et al., 2006; Hampel et al., 2011) given by

$$p^\dagger(\boldsymbol{x}) = (1-\varepsilon_p)p^*(\boldsymbol{x}) + \varepsilon_p\delta_p(\boldsymbol{x}),$$
$$q^\dagger(\boldsymbol{x}) = (1-\varepsilon_q)q^*(\boldsymbol{x}) + \varepsilon_q\delta_q(\boldsymbol{x}),$$

respectively, where $p^*(\boldsymbol{x})$ and $q^*(\boldsymbol{x})$ are the true density functions, $\delta_p(\boldsymbol{x})$ and $\delta_q(\boldsymbol{x})$ are the density functions of outliers, and ε_p and ε_q $(0 \leq \varepsilon_p < 1, 0 \leq \varepsilon_q < 1)$ are the contamination ratios of outliers. The contaminated density ratio is $r^\dagger(\boldsymbol{x}) = p^\dagger(\boldsymbol{x})/q^\dagger(\boldsymbol{x})$, and the true density ratio is $r^*(\boldsymbol{x}) = p^*(\boldsymbol{x})/q^*(\boldsymbol{x})$.

The target parameter of the minimization problem (4) can have a latent bias in the contaminated setting. Because we only have the contaminated density functions $p^\dagger(\boldsymbol{x})$ and $q^\dagger(\boldsymbol{x})$, the target parameter in the contaminated setting is

$$\boldsymbol{\theta}^\dagger_{\mathrm{UKL}} = \operatorname*{argmin}_{\boldsymbol{\theta}} D_{\mathrm{UKL}}(r^\dagger, r_{\boldsymbol{\theta}}; w).$$

This is different from the true target parameter

$$\boldsymbol{\theta}^*_{\mathrm{UKL}} = \operatorname*{argmin}_{\boldsymbol{\theta}} D_{\mathrm{UKL}}(r^*, r_{\boldsymbol{\theta}}; w).$$

The robust statistics aims to make the latent bias $\boldsymbol{\theta}^{\dagger}_{\mathrm{UKL}} - \boldsymbol{\theta}^{*}_{\mathrm{UKL}}$ sufficiently small (Huber, 2004; Hampel et al., 2011).

To achieve the small bias, we assume that the weight function can ignore the outlier distributions.

Assumption 3.1. Suppose that $\boldsymbol{\theta}$ exists in a compact convex set Θ. Let

$$f_q(\boldsymbol{x}, \boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) \delta_q(\boldsymbol{x}),$$
$$\nu_1 = \int w(\boldsymbol{x}) \delta_p(\boldsymbol{x}) d\boldsymbol{x}, \ \ \nu_2 = \int \boldsymbol{\theta}^T h(\boldsymbol{x}) w(\boldsymbol{x}) \delta_p(\boldsymbol{x}) d\boldsymbol{x},$$
$$\nu_3 = \int f_q(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}, \ \ \nu_4 = \int h(\boldsymbol{x}) f_q(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x},$$
$$\nu_5 = \int h(\boldsymbol{x}) h(\boldsymbol{x})^T f_q(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}.$$

Let $\mathcal{A}_j$ be the index set of the elements of ν_j, and $\nu = \max\{\sup_{\boldsymbol{\theta} \in \Theta} |\nu_{ja}|\}_{j=1....,5; a \in \mathcal{A}_j}$. Then, ν is sufficiently small.

Let us consider an example where Assumption 3.1 is satisfied.

Example 3.2. Let $\delta_{\boldsymbol{a}}(\boldsymbol{x})$ be a Dirac delta function at $\boldsymbol{a}$. Consider the well-used outlier distributions: $\delta_p(\boldsymbol{x}) = \delta_{\boldsymbol{x}_o^{(p)}}(\boldsymbol{x})$ and $\delta_q(\boldsymbol{x}) = \delta_{\boldsymbol{x}_o^{(q)}}(\boldsymbol{x})$ with outliers $\boldsymbol{x}_o^{(p)}$ and $\boldsymbol{x}_o^{(q)}$. Let

$$h(\boldsymbol{x}) = (x_1 x_1, x_1 x_2, \ldots, x_d x_d)^T,$$

which is usually used in Gaussian distributions and Ising models (Liu et al., 2017b). Let

$$w(\boldsymbol{x}) = \exp(-\|x\|_4^4),$$

where $\|x\|_4^4 = x_1^4 + x_2^4 + ... + x_d^4$. If $\boldsymbol{x}_o^{(p)}$ and $\boldsymbol{x}_o^{(q)}$ are strong outliers, more precisely, if $\|\boldsymbol{x}_o^{(p)}\|$ and $\|\boldsymbol{x}_o^{(q)}\|$ are sufficiently large, this example satisfies Assumption 3.1 because

$$\nu_1 = w(\boldsymbol{x}_o^{(p)}), \quad \nu_2 = \boldsymbol{\theta}^T h(\boldsymbol{x}_o^{(p)}) w(\boldsymbol{x}_o^{(p)}),$$
$$\nu_3 = \exp(\boldsymbol{\theta}^T h(\boldsymbol{x}_o^{(q)})) w(\boldsymbol{x}_o^{(q)}),$$
$$\nu_4 = h(\boldsymbol{x}_o^{(q)}) \nu_3, \quad \nu_5 = h(\boldsymbol{x}_o^{(q)}) h(\boldsymbol{x}_o^{(q)})^T \nu_3.$$

Importantly, the weight function $w(\boldsymbol{x})$ should converge to zero more rapidly than the feature transform function $h(\boldsymbol{x})$ and the parametric density ratio function $r_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\boldsymbol{\theta}^T h(\boldsymbol{x}))$. An example of the practical choice of the weight function is discussed in Appendix C.

Under Assumption 3.1, the following theorem holds.

Theorem 3.3. *Under Assumption 3.1, we have*

$$D_{\mathrm{UKL}}(r^{\dagger}, r_{\boldsymbol{\theta}}; w) = (1 - \varepsilon_p) \left\{ D_{\mathrm{UKL}}(r^*, r_{\boldsymbol{\theta}}; w) + \mathrm{const} + O(\varepsilon_r \nu) \right\},$$

where

$$\varepsilon_r = \max\left\{ \frac{\varepsilon_p}{1 - \varepsilon_p}, -\frac{\varepsilon_p \log(1 - \varepsilon_q)}{1 - \varepsilon_p}, \frac{\varepsilon_q}{1 - \varepsilon_q} \right\}.$$

Furthermore, we assume that $\boldsymbol{\theta}^{\dagger}_{\mathrm{UKL}}$ and $\boldsymbol{\theta}^{*}_{\mathrm{UKL}}$ are unique interior points in Θ. Then, we have

$$\boldsymbol{\theta}^{\dagger}_{\mathrm{UKL}} - \boldsymbol{\theta}^{*}_{\mathrm{UKL}} = O(\varepsilon_r \nu).$$

The outline of the proof is owing to (Fujisawa & Eguchi, 2008). The proof is given in Appendix D.1.

This theorem implies that the latent bias $\boldsymbol{\theta}^{\dagger}_{\mathrm{UKL}} - \boldsymbol{\theta}^{*}_{\mathrm{UKL}}$ can be sufficiently small even when the contamination ratios ε_p and ε_q are not small. This property can be called strong robustness, which is considered for one contaminated distribution in the past papers (Fujisawa & Eguchi, 2008; Hirose et al., 2017; Hung et al., 2018; Kawashima & Fujisawa, 2023). The above theorem shows doubly strong robustness because there are two contaminated distributions.

3.2. γ-DRE

Although Weighted DRE seems reasonable, it has a slight drawback in estimating the normalizing term C in (3). This normalizing term consists of the reference and target dataset and may not be precise due to randomness. The normalizing term C is a nuisance parameter, and our interest is only in estimating the difference parameter $\boldsymbol{\theta}$. In this section, we propose another DRE method without estimating the normalizing term C.

3.2.1. Formulation

The γ-divergence (Fujisawa & Eguchi, 2008) is a popular method to estimate a density function robustly under heavily contaminated conditions. Let $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ be strictly positive functions, not necessarily density functions, and γ be a positive constant. The γ-divergence $D_\gamma(g, f)$ is written as $D_\gamma(g, f) = d_\gamma(g, f) - d_\gamma(g, g)$, where $d_\gamma(g, f)$ is the γ-cross entropy:

$$d_\gamma(g, f) = -\frac{1}{\gamma} \log \int g(\boldsymbol{x}) f(\boldsymbol{x})^\gamma d\boldsymbol{x} + \frac{1}{1+\gamma} \log \int f(\boldsymbol{x})^{1+\gamma} d\boldsymbol{x}.$$

We propose γ-DRE, which minimizes the γ-cross entropy (or the γ-divergence) between the true density ratio $r(\boldsymbol{x})$ and the parametric density ratio function $r_{\boldsymbol{\theta},C}(\boldsymbol{x}) = C r_{\boldsymbol{\theta}}(\boldsymbol{x}) =$

$C\exp(\boldsymbol{\theta}^T h(\boldsymbol{x}))$ with the base measure of $w(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}$:

$$\begin{aligned}
&d_\gamma(r, r_{\boldsymbol{\theta},C}; wq)\\
=&-\frac{1}{\gamma}\log\int r_{\boldsymbol{\theta},C}(\boldsymbol{x})^\gamma w(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\\
&+\frac{1}{1+\gamma}\log\int r_{\boldsymbol{\theta},C}(\boldsymbol{x})^{1+\gamma} w(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\\
=&-\frac{1}{\gamma}\log\int \exp(\gamma\boldsymbol{\theta}^T h(\boldsymbol{x}))w(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\\
&+\frac{1}{1+\gamma}\log\int \exp((1+\gamma)\boldsymbol{\theta}^T h(\boldsymbol{x}))w(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\\
=&d_\gamma(r, r_{\boldsymbol{\theta}}; wq).
\end{aligned} \tag{6}$$

The γ-divergence with the base measure $w(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}$, that is, $D_\gamma(r, r_{\boldsymbol{\theta}}; wq) = d_\gamma(r, r_{\boldsymbol{\theta}}; wq) - d_\gamma(r, r; wq)$, satisfies the following: (i) $D_\gamma(r, r_{\boldsymbol{\theta}}; wq) \geq D_\gamma(r, r; wq)$, (ii) $D_\gamma(r, r_{\boldsymbol{\theta}}; wq) = 0 \Leftrightarrow r_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha\, r(\boldsymbol{x})$, where $\alpha > 0$ is a constant. The property (ii) is slightly different from the conventional one, which implies that the normalizing term C vanishes in (6).

The above objective function is empirically approximated by two datasets $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}$ and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q}$:

$$\begin{aligned}
&\hat{d}_\gamma(r, r_{\boldsymbol{\theta}}; wq)\\
=&-\frac{1}{\gamma}\log\frac{1}{n_p}\sum_{i=1}^{n_p}\exp(\gamma\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(p)}))w(\boldsymbol{x}_i^{(p)})\\
&+\frac{1}{1+\gamma}\log\frac{1}{n_q}\sum_{i=1}^{n_q}\exp((1+\gamma)\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)}))w(\boldsymbol{x}_i^{(q)})\\
\triangleq&-g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta}).
\end{aligned} \tag{7}$$

The minimization about $\boldsymbol{\theta}$ in (7) is not a convex problem but a DC (Difference of Convex functions) problem because $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ are convex. This problem can be iteratively solved by the dual problem of Fenchel-Rockafellar (Dinh & Thi, 1997):

$$\boldsymbol{\theta}^{(k)} = \operatorname*{argmin}_{\boldsymbol{\theta}} g_2(\boldsymbol{\theta}) - \boldsymbol{\theta}^T \partial g_1(\boldsymbol{\theta}^{(k-1)}), \tag{8}$$

where k is an iteration number. At each iteration, the objective function in (8) is convex and can be minimized by optimization methods, including gradient descent. The more efficient algorithm in the high-dimensional setting will be discussed in Appendix B.2.

3.2.2. Doubly Strong Robustness

γ-DRE also has doubly strong robustness. We assume a similar property of the weight function to Assumption 3.1.

Assumption 3.4. Suppose that $\boldsymbol{\theta}$ exists in a compact convex set Θ. Let

$$\begin{aligned}
&f_{gm}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp((m+\gamma)\boldsymbol{\theta}^T h(\boldsymbol{x}))w(\boldsymbol{x})\delta_g(\boldsymbol{x}),\\
&\nu_1' = \int f_{p0}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x}, \quad \nu_2' = \int f_{q1}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x},\\
&\nu_3' = \int h(\boldsymbol{x}) f_{p0}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x}, \quad \nu_4' = \int h(\boldsymbol{x}) f_{q1}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x},\\
&\nu_5' = \int h(\boldsymbol{x})h(\boldsymbol{x})^T f_{p0}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x},\\
&\nu_6' = \int h(\boldsymbol{x})h(\boldsymbol{x})^T f_{q1}(\boldsymbol{x}, \boldsymbol{\theta})d\boldsymbol{x}.
\end{aligned}$$

Let $\mathcal{A}_j'$ be the index set of the elements of ν_j', and $\nu' = \max\{\sup_{\boldsymbol{\theta}\in\Theta}|\nu_{ja}'|\}_{j=1,\ldots,6;a\in\mathcal{A}_j'}$. Then, ν' is sufficiently small.

The same example as in Example 3.2 satisfies Assumption 3.4, if γ is set to a moderate value, for instance, $\gamma \in (0, 1]$ (Fujisawa & Eguchi, 2008).

Under Assumption 3.4, the following theorem holds.

Theorem 3.5. *Under Assumption 3.4, we have*

$$\begin{aligned}
&d_\gamma(r^\dagger, r_{\boldsymbol{\theta}}; wq^\dagger)\\
&= d_\gamma(r^*, r_{\boldsymbol{\theta}}; wq^*) + \text{const} + O(\varepsilon_r'\nu'),
\end{aligned}$$

where

$$\varepsilon_r' = \max\left\{\frac{\varepsilon_p}{1-\varepsilon_p}, \frac{\varepsilon_q}{1-\varepsilon_q}\right\}.$$

Let

$$\begin{aligned}
\boldsymbol{\theta}_\gamma^\dagger &= \operatorname*{argmin}_{\boldsymbol{\theta}} d_\gamma(r^\dagger, r_{\boldsymbol{\theta}}; wq^\dagger),\\
\boldsymbol{\theta}_\gamma^* &= \operatorname*{argmin}_{\boldsymbol{\theta}} d_\gamma(r^*, r_{\boldsymbol{\theta}}; wq^*).
\end{aligned}$$

Furthermore, we assume that $\boldsymbol{\theta}_\gamma^\dagger$ and $\boldsymbol{\theta}_\gamma^*$ are unique interior points in Θ and the Hessian matrix of $d_\gamma(r^*, r_{\boldsymbol{\theta}}; wq^*)$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_\gamma^*$ is positive definite. Then, we have

$$\boldsymbol{\theta}_\gamma^\dagger - \boldsymbol{\theta}_\gamma^* = O(\varepsilon_r'\nu').$$

The proof is given in Appendix D.2. γ-DRE also has doubly strong robustness because Theorem 3.5 holds even when the contamination ratio ε_p and ε_q are not small.

3.3. Related Works

The most promising robust DRE method is Trimmed DRE (Liu et al., 2017c). It trims outliers by assuming that they have larger values of the density ratio function than inliers. The estimator of Trimmed DRE is given by optimizing the weighted KL divergence, which reduces into a min-max problem of a convex function:

$$\max_{\boldsymbol{\theta}}\min_{\boldsymbol{w}\in[0,1]^{n_p}, <\mathbf{1},\boldsymbol{w}>=\nu n_p}\sum_{i=1}^{n_p} w_i \log r_{\boldsymbol{\theta},C^\circ}(\boldsymbol{x}_i^{(p)}),$$

where $\boldsymbol{w} \in \mathbb{R}^{n_p}$ is a weight parameter for the reference dataset, $\nu \in (0, 1]$ is a trimming quantile, and $C^\circ = n_q / \sum_{i=1}^{n_q} \exp(\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)}))$. In estimating $\boldsymbol{w}$ given $\boldsymbol{\theta}$, w_is with the top $100(1-\nu)$ % of $\log r_{\boldsymbol{\theta}, C^\circ}(\boldsymbol{x}_i^{(p)})$ values become 0, and the rest becomes $1/n_p$. In estimating $\boldsymbol{\theta}$ given $\boldsymbol{w}$, data with $w_i = 0$ is removed from the maximization of the empirical density ratio function. Thus, the parameter $\boldsymbol{\theta}$ is estimated with only inliers.

Although Trimmed DRE is shown to be robust, the assumption it relies on does not hold in some situations. The parametric density ratio function $r_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(\boldsymbol{\theta}^T h(\boldsymbol{x}))$ does not have large values for some outliers. For simplicity, let us consider the cases where $\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta}_p, \boldsymbol{\beta}_q \in \mathbb{R}$, $\boldsymbol{\theta}^T h(\boldsymbol{x}) = \theta x, \theta = \beta_p - \beta_q$, and the outlier is $x_o = c$ ($c > 0$). We have the following three cases when $c \to \infty$:

(i) When $\beta_p > \beta_q$, $r_\theta(x_o) \propto \exp(|\theta| c) \to \infty$.

(ii) When $\beta_p < \beta_q$, $r_\theta(x_o) \propto \exp(-|\theta| c) \to 0$.

(iii) When $\beta_p = \beta_q$, $r_\theta(x_o) \propto \exp(0) \to \text{const.}$

These cases indicate that Trimmed DRE is not robust when $r_\theta(x) \nrightarrow \infty$ ($c \to \infty$). A similar discussion can be given when $\boldsymbol{\theta}^T h(\boldsymbol{x}) = \sum_{u,v=1}^{d} \theta_{u,v} x_u x_v$, a typical setting for multivariate Gaussian distributions and Ising models (Liu et al., 2017b).

Another limitation is that Trimmed DRE is robust when outliers contaminate only the reference dataset while keeping the target dataset clean. The normalizing term C° is calculated by the target dataset without trimming. If the target dataset is contaminated, the calculation of the normalizing term will be unstable.

Our proposals, Weighted DRE and γ-DRE, overcome the above shortages. These methods assume that outlier distributions are negligible in the sense of Assumption 3.1 or Assumption 3.4 and are robust to the contamination of the reference and target datasets. They also need no hyperparameter tuning depending on the true contamination ratio, although Trimmed DRE needs to tune the trimming quantile ν.

4. Numerical Experiments

4.1. Difference between Precision Matrices

We estimated the difference between the precision matrices in normal distributions. Let $\boldsymbol{x} = (x_1, x_2)^T \in \mathbb{R}^2$, and the true density functions be denoted by $p^*(\boldsymbol{x}) = f_{\lambda_p}(\boldsymbol{x})$ and $q^*(\boldsymbol{x}) = f_{\lambda_q}(\boldsymbol{x})$, where $f_\lambda(\boldsymbol{x}) = N\left(\boldsymbol{0}, \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}^{-1}\right)$, for the reference and target distributions, respectively. This setting is similar to the previous research (Liu et al., 2014; 2017c). The parameters λ_p and λ_q ranged from -0.9 to 0.9. The outlier distribution was set to $\delta(\boldsymbol{x}) = N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$. We prepared four settings of the reference distribution $p^\dagger(\boldsymbol{x})$ and the target distribution $q^\dagger(\boldsymbol{x})$:

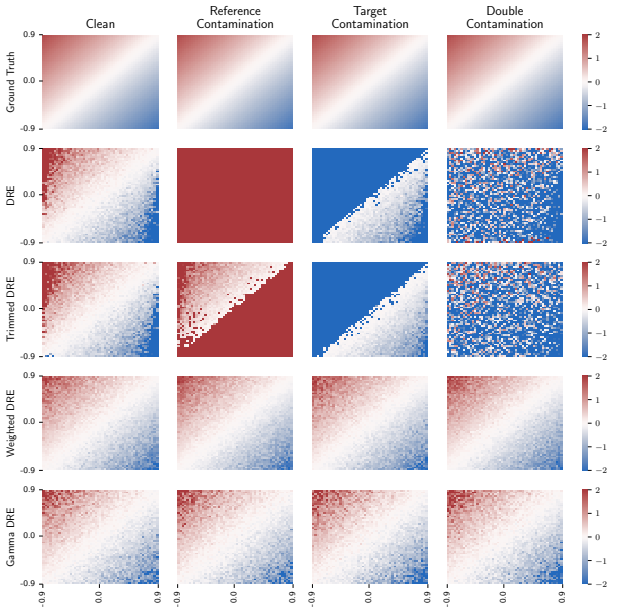

Figure 1. Estimation of the difference parameter $\theta_{1,2}$ between the off-diagonal elements of the precision matrices in normal distributions. Each column shows the settings of the contamination: "clean", "reference contamination", "target contamination", and "double contamination". Each row shows the ground truth and the estimation methods: DRE, Trimmed DRE, Weighted DRE, and γ-DRE. The x-axis and y-axis in each figure indicate the true values of the off-diagonal elements of the reference and target distributions, respectively.

- "clean": $p^\dagger(\boldsymbol{x}) = p^*(\boldsymbol{x})$, $q^\dagger(\boldsymbol{x}) = q^*(\boldsymbol{x})$.
- "reference contamination": $p^\dagger(\boldsymbol{x}) = 0.8p^*(\boldsymbol{x}) + 0.2\delta(\boldsymbol{x})$, $q^\dagger(\boldsymbol{x}) = q^*(\boldsymbol{x})$.
- "target contamination": $p^\dagger(x) = p^*(x)$, $q^\dagger(\boldsymbol{x}) = 0.8q^*(\boldsymbol{x}) + 0.2\delta(\boldsymbol{x})$.
- "double contamination": $p^\dagger(\boldsymbol{x}) = 0.8p^*(\boldsymbol{x}) + 0.2\delta(\boldsymbol{x})$, $q^\dagger(\boldsymbol{x}) = 0.8q^*(\boldsymbol{x}) + 0.2\delta(\boldsymbol{x})$.

The dataset sizes were set to $n_p = n_q = 100$.

We compared four methods (DRE, Trimmed DRE, Weighted DRE, and γ-DRE). The density ratio function was parametrized by $\boldsymbol{\theta}^T h(\boldsymbol{x}) = \sum_{u,v=1}^{2} \theta_{u,v} x_u x_v$. The ground truth is given by $\theta_{1,2} = \theta_{2,1} = \lambda_q - \lambda_p$ and $\theta_{1,1} = \theta_{2,2} = 0$

Table 1. MSE of the estimated parameters in the "clean" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
0.0	0.0	0.0	0.0123 (0.01645)	0.0196 (0.0393)	0.0240 (0.0491)	**0.0097 (0.0185)**
0.4	0.0	0.4	0.0565 (0.06743)	0.0551 (0.0548)	0.0883 (0.0765)	**0.0493 (0.0400)**
0.8	-0.4	0.4	0.1535 (0.10345)	0.1428 (0.0966)	0.2349 (0.1351)	**0.1387 (0.0850)**
1.2	-0.8	0.4	0.3035 (0.61568)	**0.2363 (0.2536)**	0.5086 (0.2216)	0.2631 (0.1855)
1.6	-0.8	0.8	1.5608 (2.75867)	1.4278 (2.6771)	0.7467 (0.3520)	**0.2931 (0.2971)**

Table 2. MSE of the estimated parameters in the "double contamination" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
0.0	0.0	0.0	5.3757 (5.5380)	5.5567 (4.8591)	0.0299 (0.0455)	**0.0131 (0.0165)**
0.4	0.0	0.4	5.5245 (5.5945)	6.0835 (6.4211)	0.0840 (0.0845)	**0.0689 (0.0704)**
0.8	-0.4	0.4	8.3192 (7.8744)	9.0380 (7.0335)	0.2950 (0.1739)	**0.1629 (0.0955)**
1.2	-0.8	0.4	8.2561 (7.7996)	10.2449 (8.2807)	0.5750 (0.2867)	**0.2175 (0.1636)**
1.6	-0.8	0.8	8.4907 (8.8919)	9.3233 (8.7860)	0.9466 (0.5066)	**0.3042 (0.2891)**

because the true density ratio $r^*(\boldsymbol{x}) = p^*(\boldsymbol{x})/q^*(\boldsymbol{x}) \propto \exp((\lambda_q - \lambda_p)x_1x_2)$. The weight function was set to $w(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}\|_4^4/50\right)$. The trimming quantile ν in Trimmed DRE was set to the true contamination ratio. The parameter γ in γ-DRE was set to 0.01. No regularization term was added to the objective function.

Figure 1 shows that Weighted DRE and γ-DRE successfully estimated the ground truth of $\theta_{1,2}$ in all the contaminated settings. This was because the experimental setting satisfies Assumption 3.1 and Assumption 3.4.

DRE failed to estimate the ground truth in the contaminated settings except for a part of the "target contamination". In the "target contamination" setting, DRE estimated the ground truth robustly where $\theta_{1,2} < 0$ in the lower right of the figure. This was because the influence of the outliers vanished from the second term of the objective function in (5), where $\exp(\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)})) = \exp(\sum_{u,v=1}^2 \theta_{u,v} x_{i,u}^{(q)} x_{i,v}^{(q)}) \approx 0$ with $\theta_{1,2} = \theta_{2,1} < 0$, $\theta_{1,1} = \theta_{2,2} = 0$, and $x_{i,1}^{(q)}, x_{i,2}^{(q)} \approx 100$. In the "reference contamination" and the "target contamination" settings, the estimated values were heavily biased in the opposite direction because the reference and target datasets had the opposite signs in the objective function (5). In the case of "double contamination", the color tendency was unclear because, for example, the estimated values tended to be positive/negative when the generated outliers were stronger on the reference/target than the other one.

Trimmed DRE successfully estimated the ground truth in the "reference contamination" setting where $\theta_{1,2} > 0$ in the upper left of the figure. In this region, the density ratio values of the outliers became large by $\exp(\sum_{u,v=1}^2 \theta_{u,v} x_u x_v)$, where $\theta_{1,2} = \theta_{2,1} > 0$, $\theta_{1,1} = \theta_{2,2} = 0$, and $x_1, x_2 \approx 100$. Because the estimator of Trimmed DRE was designed to be robust when outliers had larger density ratio values than inliers, the estimator was robust in that region. In other regions, the estimation performance was similar to that of DRE. These results agree with the theoretical property discussed in Section 3.3.

We calculated the mean squared error (MSE) values of the estimated values on some pairs of the true parameters λ_p and λ_q. The experiments were repeated 100 times. More details are given in Appendix E, including the extended cases of the parameter pairs and the contamination settings.

Tables 1 and 2 show that Weighted DRE and γ-DRE achieved robustness without compromising the precision of the estimation. In the range of $0.0 \leq \theta_{1,2} \leq 1.2$ in Table 1 (the "clean" setting), the MSE values of Weighted DRE and γ-DRE were comparable to those of DRE and Trimmed DRE. Besides, at $\theta_{1,2} = 1.6$, Weighted DRE and γ-DRE had the significantly smaller MSE values than DRE and Trimmed DRE. That might be because the weight function excluded "weak outliers" sampled from the main body of the density functions. In Table 2 (the "double contamination" setting), the MSE values of Weighted DRE and γ-DRE were significantly smaller than those of DRE and Trimmed DRE, which shows the robustness of our proposals.

Tables 1 and 2 also show that γ-DRE estimated the ground truth slightly better than Weighted DRE. The MSE values of γ-DRE were smaller than those of Weighted DRE in all the settings. γ-DRE did not estimate the normalizing term (3), so the estimated values might have small errors.

4.2. Change Detection in Time Series Data

As an experiment with real-world data, we performed a change detection in time series data. We used a human activity dataset provided by the Human Activity Sensing

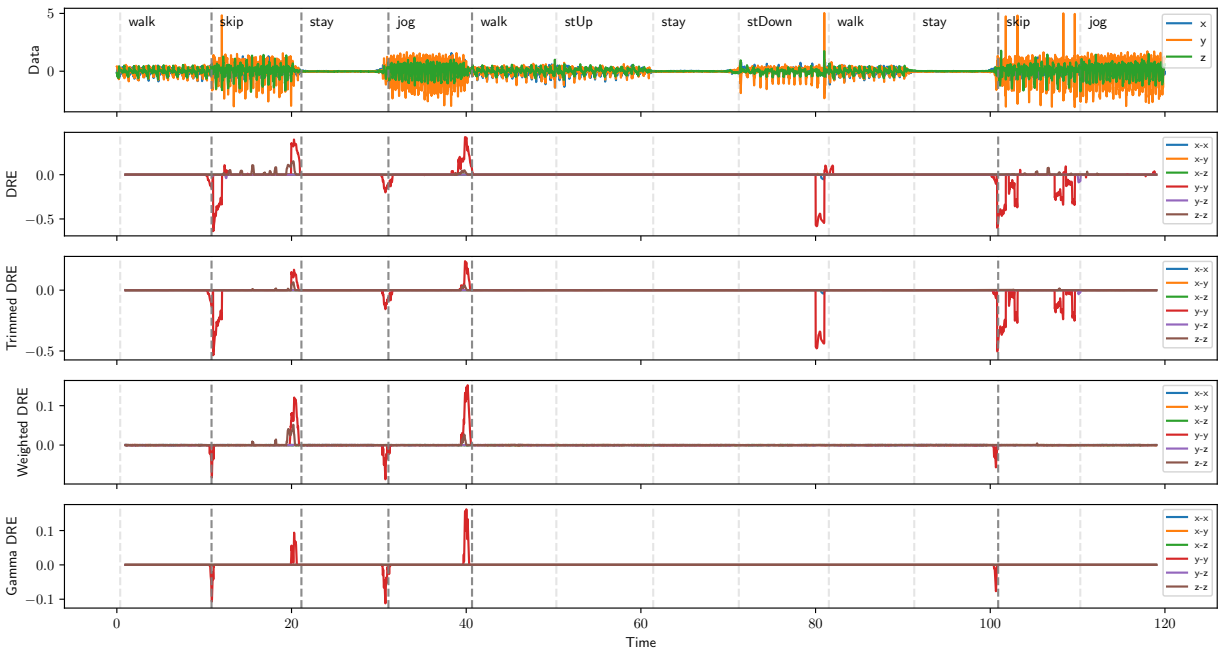

Figure 2. Change detection of the human activity sensor data. The top figure shows the sequence of the three-dimensional data (x-, y-, and z-axis). Each figure below the top shows all the estimated parameters of DRE, Trimmed DRE, Weighted DRE, and γ-DRE. The vertical dashed lines are the ground truth of the change points with the category labels of the human activities. Bold dashed lines indicate the changes from and to "skip" or "jog".

Consortium (HASC) Challenge 2011, which was used in the experiment of change detection with DRE (Liu et al., 2013). This is a task to segment time series data measured by acceleration sensors according to 6 categories: "stay", "walk", "jog", "skip", "stair up", and "stair down". The measured data has three dimensions (x-, y-, and z-axis).

In the change detection task in time series data, the original sequence was divided and assigned to the reference and target datasets sequentially. The dataset sizes of the reference and target datasets were set to $n_p = n_q = 100$.

We compared four methods (DRE, Trimmed DRE, Weighted DRE, and γ-DRE). The parametric function was designed as $\boldsymbol{\theta}^T h(\boldsymbol{x}) = \sum_{u,v=1}^{3} \theta_{u,v} x_u x_v$. The weight function was set to $w(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}\|_4^4/5\right)$ in Weighted DRE and γ-DRE. We added the elastic net regularization term $\lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2\|\boldsymbol{\theta}\|_2^2$ with $\lambda_1 = \lambda_2 = 0.5$ to the objective function. The trimming quantile ν was set to 0.9 in Trimmed DRE. The tuning parameter γ was set to 0.01 in γ-DRE. All the estimated parameters, $\theta_{u,v}$ for $1 \le u \le v \le 3$, were used as the anomaly levels instead of the density ratio values because $\theta_{u,v} = 0$ at no-change points.

Figure 2 shows that Weighted DRE and γ-DRE successfully detected the change points without disturbance by outliers. The top row of Figure 2 indicates that the changes from and to "skip" or "jog" are large and other changes are small. All the methods successfully detected these large change points. DRE falsely detected the outliers at the time around $[10, 20]$ and $[100, 110]$, and Trimmed DRE did so at the time around $[10, 12]$ and $[100, 110]$. Besides, they estimated larger anomaly levels at the outliers than those at the true change points. Weighted DRE and γ-DRE did not detect these outliers. At the time around 80, Weighted DRE and γ-DRE could not detect the change point, although DRE and Trimmed DRE could. This was the change point between "stair down" and "walk", but there was no apparent change in the original data. DRE and Trimmed DRE seemed to detect the outlier, not the change point.

5. Conclusions

We have presented two density ratio estimation methods, Weighted DRE and γ-DRE. Their estimators are robust even when many outliers contaminate both the reference and target datasets, which is called doubly strong robustness. Theoretical analysis reveals that the weight function should converge to zero more rapidly than the density ratio function. Numerical experiments show that our proposals are more robust than the previous methods for synthetic and real-world datasets.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

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

Choi, K., Liao, M., and Ermon, S. Featurized density ratio estimation. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161, pp. 172–182, 2021.

Choi, K., Meng, C., Song, Y., and Ermon, S. Density ratio estimation via infinitesimal classification. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022*, 2022.

Dinh, T. P. and Thi, H. A. L. Convex analysis approach to d.c. programming: Theory, algorithm and applications. 1997.

Fujisawa, H. and Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008. ISSN 0047-259X.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, pp. 131–160, 2009.

Hadi, A. S. and Luceno, A. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25(3):251–272, 1997.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. *Robust Statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2001.

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

Hirose, K., Fujisawa, H., and Sese, J. Robust sparse gaussian graphical modeling. *Journal of Multivariate Analysis*, 161:172–190, 2017. ISSN 0047-259X.

Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

Huber, P. J. *Robust Statistics*. John Wiley & Sons, 2004.

Hung, H., Jou, Z.-Y., and Huang, S.-Y. Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1), 2018.

Ideker, T. and Krogan, N. J. Differential network biology. *Molecular Systems Biology*, 8:565, 2012.

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009.

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

Kanamori, T., Suzuki, T., and Sugiyama, M. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86:335–367, 2012.

Kato, M. and Teshima, T. Non-negative bregman divergence minimization for deep direct density ratio estimation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5320–5333. PMLR, 18–24 Jul 2021.

Kawahara, Y. and Sugiyama, M. Change-point detection in time-series data by direct density-ratio estimation. pp. 389–400, 2009.

Kawashima, T. and Fujisawa, H. Robust regression against heavy heterogeneous contamination. *Metrika*, 86:421–442, 2023.

Kim, B., Liu, S., and Kolar, M. Two-Sample Inference for High-Dimensional Markov Networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):939–962, 09 2021. ISSN 1369-7412.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013. ISSN 0893-6080.

Liu, S., Quinn, J., Gutmann, M., Suzuki, T., and Sugiyama, M. Direct learning of sparse changes in markov networks by density ratio estimation. *Neural computation*, 26: 1169–1197, 03 2014.

Liu, S., Fukumizu, K., and Suzuki, T. Learning sparse structural changes in high-dimensional markov networks: A review on methodologies and theories. *Behaviormetrika*, 44:265–296, 2017a.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M., and Fukumizu, K. Support consistency of direct sparse-change learning in Markov networks. *The Annals of Statistics*, 45(3):959 – 990, 2017b.

Liu, S., Takeda, A., Suzuki, T., and Fukumizu, K. Trimmed density ratio estimation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017c.

Maronna, R. A., Martin, R. D., and Yohai, V. J. *Robust Statistics: Theory and Methods*. John Wiley & Sons, 2006.

Na, S., Kolar, M., and Koyejo, O. Estimating differential latent variable graphical models with applications to brain connectivity. *Biometrika*, 108(2):425–442, 09 2020. ISSN 0006-3444.

Nguyen, X., Wainwright, M. J., and Jordan, M. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Rhodes, B., Xu, K., and Gutmann, M. U. Telescoping density-ratio estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4905–4916. Curran Associates, Inc., 2020.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758.

Smola, A., Song, L., and Teo, C. H. Relative novelty detection. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 536–543, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

Srivastava, A., Han, S., Xu, K., Rhodes, B., and Gutmann, M. Estimating the density ratio between distributions with high discrepancy using multinomial logistic regression. *Transactions on Machine Learning Research*, 2023(3): 1–23, 2023. ISSN 2835-8856.

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

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012a.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012b.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

Wornowizki, M. and Fried, R. Two-sample homogeneity tests based on divergence measures. *Computational Statistics*, 31:291–313, 2016.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25:1324–1370, 2011.

Yang, E. and Lozano, A. C. Robust gaussian graphical modeling with the trimmed graphical lasso. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Zhang, Y.-J., Zhang, Z.-Y., Zhao, P., and Sugiyama, M. Adapting to continuous covariate shift via online density ratio estimation. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29074–29113. Curran Associates, Inc., 2023.

Zou, H. and Hastie, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412.

A. Discussion of Parametric Density Ratio Function

We justify the formulation of the parametric density ratio function (1) in the case of an exponential family. If two density functions belong to an exponential family, more precisely, $p(\boldsymbol{x}) = f_{\beta_p}(\boldsymbol{x})$ and $q(\boldsymbol{x}) = f_{\beta_q}(\boldsymbol{x})$, where $f_\beta(\boldsymbol{x}) = \exp\left\{\boldsymbol{\beta}^T h(\boldsymbol{x}) + b(\boldsymbol{x}) - \log A(\boldsymbol{\beta})\right\}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$, $h(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^p$, $b(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}$, and $A(\boldsymbol{\beta}) : \mathbb{R}^p \mapsto \mathbb{R}$, the true density ratio can be written as

$$r(\boldsymbol{x}) = \frac{A(\boldsymbol{\beta}_q)}{A(\boldsymbol{\beta}_p)} \exp\left\{(\boldsymbol{\beta}_p - \boldsymbol{\beta}_q)^T h(\boldsymbol{x})\right\}. \tag{9}$$

In the formulation of the parametric density ratio function, we estimate the difference $\boldsymbol{\beta}_p - \boldsymbol{\beta}_q$ directly without estimating each parameter $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_q$ (Liu et al., 2014; 2017a). To approximate the true density ratio $r(\boldsymbol{x})$, the parametric density ratio function $r_{\boldsymbol{\theta},C}(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ is introduced by

$$r_{\boldsymbol{\theta},C}(\boldsymbol{x}) = C \exp\left(\boldsymbol{\theta}^T h(\boldsymbol{x})\right). \tag{10}$$

By comparing (9) and (10), the optimal parameters $\boldsymbol{\theta}^\circ$ and C° are given by

$$\boldsymbol{\theta}^\circ = \boldsymbol{\beta}_p - \boldsymbol{\beta}_q, \quad C^\circ = \frac{A(\boldsymbol{\beta}_q)}{A(\boldsymbol{\beta}_p)}.$$

We can easily show that $\boldsymbol{\theta}^\circ$ and C° are the minimizer of (2).

Notably, the parametric density ratio function $r_{\boldsymbol{\theta},C}(\boldsymbol{x})$ can be applied to any density ratio, even when the density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ do not necessarily belong to an exponential family. In that case, $\boldsymbol{\theta}^\circ$ and C° cannot be written explicitly.

B. Efficient Optimization by Dual Problem

We show the primal-dual trick enables us the efficient calculation of the estimator of Weighted DRE and γ-DRE, which is scalable to the high-dimensional data (Boyd & Vandenberghe, 2004; Liu et al., 2014). The core of this trick is to avoid calculating the log-expectation term.

B.1. Dual Problem of Weighted DRE

By adding L1 and L2 norms to (5), the objective function of Weighted DRE, $\mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\theta})$, can be written as

$$\mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\theta}) = -\boldsymbol{\theta}^T \boldsymbol{\xi} + \kappa \log \sum_{i=1}^{n_q} \exp(\boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)})) w(\boldsymbol{x}_i^{(q)}) + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2, \tag{11}$$

where,

$$\boldsymbol{\xi} = \frac{1}{n_p} \sum_{i=1}^{n_p} h(\boldsymbol{x}_i^{(p)}) w(\boldsymbol{x}_i^{(p)}) \in \mathbb{R}^d, \quad \kappa = \frac{1}{n_p} \sum_{i=1}^{n_p} w(\boldsymbol{x}_i^{(p)}) \in \mathbb{R},$$

and λ_1 and λ_2 are regularization parameters for L1 and L2 norms, respectively.

The idea to convert this objective function (11) into the dual problem is to introduce the variable $z_i = \boldsymbol{\theta}^T h(\boldsymbol{x}_i^{(q)})$ for $i = 1, ..., n_q$ in the log-sum-exp term. This equation can be introduced into the original objective function (11) as the constraints in the form of Lagrange multiplier. Here, we can convert the minimization of (11) into min-max problem: $\min_{\boldsymbol{\theta},\boldsymbol{z}} \max_{\boldsymbol{\alpha}} \mathcal{L}_{UKL}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\alpha})$, where

$$\mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\alpha}) = -\boldsymbol{\theta}^T \boldsymbol{\xi} + \kappa \log \sum_{i=1}^{n_q} \exp(z_i) w(\boldsymbol{x}_i^{(q)}) + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2 - (\boldsymbol{z} - \Phi^T \boldsymbol{\theta})^T \boldsymbol{\alpha}, \tag{12}$$

where $\Phi = (h(\boldsymbol{x}_1^{(q)}), ..., h(\boldsymbol{x}_{n_q}^{(q)})) \in \mathbb{R}^{d \times n_q}$, and $\boldsymbol{\alpha} \in \mathbb{R}^{n_q}$ is a Lagrange multiplier. We can convert this Lagrange function into the dual function by minimizing $\boldsymbol{\theta}$ and $\boldsymbol{z}$ respectively:

$$\begin{aligned} \mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\alpha}) &:= \min_{\boldsymbol{\theta},\boldsymbol{z}} \mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\alpha}) \\ &= \min_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^T (\Phi \boldsymbol{\alpha} - \boldsymbol{\xi}) + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2\right] + \min_{\boldsymbol{z}} \left[\kappa \log \sum_{i=1}^{n_q} \exp(z_i) w(\boldsymbol{x}_i^{(q)}) - \boldsymbol{z}^T \boldsymbol{\alpha}\right]. \end{aligned} \tag{13}$$

Firstly, let us consider the first term:

$$\psi_1(\boldsymbol{\theta}) = -\boldsymbol{\theta}^T\boldsymbol{\eta} + \lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2\|\boldsymbol{\theta}\|_2^2, \tag{14}$$

where $\boldsymbol{\eta} = \boldsymbol{\xi} - \Phi\boldsymbol{\alpha}$. Since $\psi_1(\boldsymbol{\theta})$ is convex, equating the differential $\nabla\psi_1(\boldsymbol{\theta})$ to zero gives us the minimizer:

$$\hat{\theta}_j = \frac{1}{2\lambda_2} s_{\lambda_1}(\eta_j), \quad \text{for } j = 1, ..., d, \tag{15}$$

where $s_\lambda(x)$ is a soft threshold function: $s_\lambda(x) = \mathrm{sgn}(x)(|x| - \lambda)_+$. By substituting the optimal $\hat{\boldsymbol{\theta}}$ into (14),

$$\min_{\boldsymbol{\theta}} \psi_1(\boldsymbol{\theta}) = -\frac{1}{4\lambda_2}\sum_{j=1}^{d}(|\eta_j| - \lambda_1)_+^2. \tag{16}$$

Secondly, let us consider the second term:

$$\psi_2(\boldsymbol{z}) = \kappa \log \sum_{i=1}^{n_q} \exp(z_i) w(\boldsymbol{x}_i^{(q)}) - \boldsymbol{z}^T\boldsymbol{\alpha}. \tag{17}$$

Since $\psi_2(\boldsymbol{z})$ is also convex, equating the differential to zero gives us the minimizer $\hat{\boldsymbol{z}}$:

$$\alpha_i = \kappa \frac{\exp(\hat{z}_i) w(\boldsymbol{x}_i^{(q)})}{\sum_{i=1}^{n_q} \exp(\hat{z}_i) w(\boldsymbol{x}_i^{(q)})}, \quad \text{for } i = 1, ..., n_q.$$

This equation gives us the constraints which $\boldsymbol{\alpha}$ should satisfy: $\alpha_i > 0$ for $i = 1, ..., n_q$ and $\sum_{i=1}^{n_q}\alpha_i = \kappa$. By substituting the optimal $\hat{\boldsymbol{z}}$ into (17),

$$\min_{\boldsymbol{z}} \psi_2(\boldsymbol{z}) = -\sum_{i=1}^{n_q} \alpha_i \left(\log\alpha_i - \log w(\boldsymbol{x}_i^{(q)})\right) + \text{const.} \tag{18}$$

By substituting (16) and (18) into (13), we should solve the problem of $\max_{\boldsymbol{\alpha}} \mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\alpha})$:

$$\begin{aligned} &\mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\alpha}) = -\frac{1}{4\lambda_2}\sum_{j=1}^{d}(|\eta_j| - \lambda_1)_+^2 - \sum_{i=1}^{n_q}\alpha_i\left(\log\alpha_i - \log w(\boldsymbol{x}_i^{(q)})\right) \\ &\text{subject to} \quad \alpha_i > 0, \text{ for } i = 1, ..., n_q, \quad \sum_{i=1}^{n_q}\alpha_i = \kappa. \end{aligned}$$

This dual function $\mathcal{L}_{\mathrm{UKL}}(\boldsymbol{\alpha})$ is concave and can be solved by optimizing methods, including gradient descent. After convergence, we can convert the estimated dual parameter $\hat{\boldsymbol{\alpha}}$ to the primal parameter $\hat{\boldsymbol{\theta}}$ by (15). Because this formulation changes the estimation problem of $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ to that of $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{n_q}$, this primal-dual trick is useful in the high-dimensional setting where $p \geq n_q$.

B.2. Dual Problem of γ-DRE

The optimization problem of γ-DRE is an iterative minimization of (8). Given the previous estimated parameter $\hat{\boldsymbol{\theta}}^{(k-1)}$, the objective function can be written as

$$\mathcal{L}_\gamma(\boldsymbol{\theta}) = -\boldsymbol{\theta}^T\boldsymbol{\xi}^{(k-1)} + \frac{1}{1+\gamma}\log\sum_{i=1}^{n_q}\exp\left((1+\gamma)\boldsymbol{\theta}^T h\left(\boldsymbol{x}_i^{(q)}\right) w\left(\boldsymbol{x}_i^{(q)}\right)\right) + \lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2\|\boldsymbol{\theta}\|_2^2,$$

where

$$\boldsymbol{\xi}^{(k-1)} = \sum_{i=1}^{n_p} \frac{\exp\left(\gamma\hat{\boldsymbol{\theta}}^{(k-1)T} h\left(\boldsymbol{x}_i^{(p)}\right) w\left(\boldsymbol{x}_i^{(p)}\right)\right)}{\sum_{j=1}^{n_p}\exp\left(\gamma\hat{\boldsymbol{\theta}}^{(k-1)T} h\left(\boldsymbol{x}_j^{(p)}\right) w\left(\boldsymbol{x}_j^{(p)}\right)\right)} h\left(\boldsymbol{x}_i^{(p)}\right).$$

By replacing $\boldsymbol{\xi}$ by $\boldsymbol{\xi}^{(k-1)}$ and κ by $1/(1+\gamma)$ in (11), the same discussion in Appendix B.1 holds. Notably, this estimation procedure should be iterated until the estimated parameter $\hat{\boldsymbol{\theta}}^{(k)}$ converges.

C. Choice of Weight Function

The weight function is expected to satisfy Assumptions 3.1 or 3.4. In practice, the weight function can be set as $w(x) = \exp(-\|\frac{x-\mathrm{Med}}{\mathrm{MADN}}\|_4^4/\tau)$, where Med is the median and $\mathrm{MADN} = \mathrm{Med}(\{|\boldsymbol{x}_i - \mathrm{Med}(\mathcal{D})|\}_{i=1}^n)/0.675$ is the normalized median absolute deviation of the dataset $\mathcal{D}$. This transformation can be interpreted as the robust version of the standardization. The hyper-parameter τ should be selected such that usual samples have large weights and outliers have small weights. Because the value of $\|\frac{x-\mathrm{Med}}{\mathrm{MADN}}\|$ increases as the dimension size increases, the appropriate value of τ can change. We adopted $\tau = 50$ for the two-dimensional standard Gaussian in Section 4.1.

Here, we show how to choose the hyper-parameter τ for the one-dimensional standard Gaussian distribution. Figure 3 shows that $\tau = 10$ is appropriate compared with the probability density function of the standard Gaussian distribution. The weight function with $\tau = 1$ has too-narrow window, which will eliminate the usual samples. The weight function with $\tau = 100$ has too-wide window, which cannot eliminate the outlier samples.

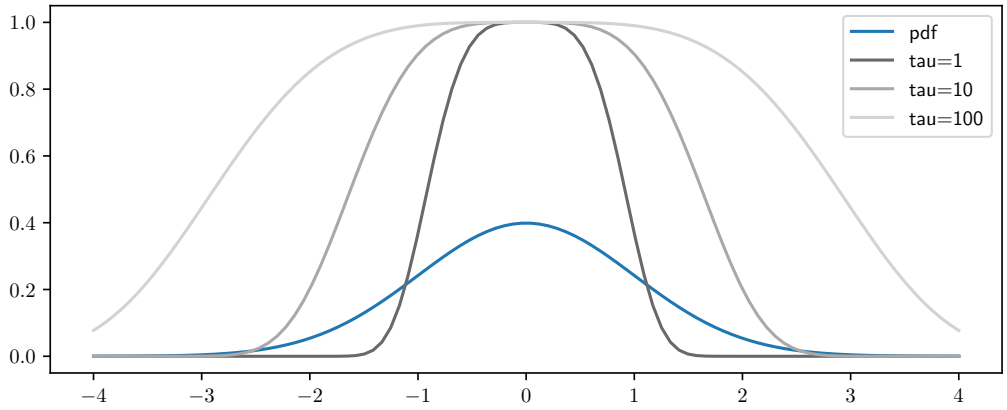

Figure 3. Comparison between the probability density function of the one-dimensional Gaussian distribution and the weight function with the hyper-parameters of $\tau = 1, 10, 100$.

D. Proof of Doubly Strong Robustness

D.1. Proof of Theorem 3.3

Lemma D.1. *Let $f(\boldsymbol{x})$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{x})$ be C^2-class real-valued functions on a compact convex set $\mathcal{X}$. Let the Hessian matrices of $f(\boldsymbol{x})$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{x})$ be denoted by $J(\boldsymbol{x})$ and $J_{\boldsymbol{\beta}}(\boldsymbol{x})$, respectively. Assume that $J(\boldsymbol{x})$ is positive definite and $J_{\boldsymbol{\beta}}(\boldsymbol{x})$ is continuous about $\boldsymbol{\beta}$. Suppose that $\|\boldsymbol{\alpha}\| = O(\nu)$ and $\|\boldsymbol{\beta}\| = O(\nu)$ for a sufficiently small value ν. Assume*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f_{\boldsymbol{\alpha}}(\boldsymbol{x}) - f(\boldsymbol{x})| = O(\nu), \tag{19}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|J_{\boldsymbol{\beta}}(\boldsymbol{x}) - J(\boldsymbol{x})\|_{\max} = O(\nu). \tag{20}$$

Let $\boldsymbol{x}^*$ and $\boldsymbol{x}^*_{\boldsymbol{\alpha}}$ be minimizers of $f(\boldsymbol{x})$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{x})$, respectively. Assume that $\boldsymbol{x}^*$ and $\boldsymbol{x}^*_{\boldsymbol{\alpha}}$ are unique interior points in $\mathcal{X}$. Then, we have $\boldsymbol{x}^*_{\boldsymbol{\alpha}} - \boldsymbol{x}^* = O(\nu)$.

Proof. Let the smallest eigenvalues of $J(\boldsymbol{x})$ and $J_{\boldsymbol{\beta}}(\boldsymbol{x})$ over the set $\mathcal{X}$ be denoted by a and $a_{\boldsymbol{\beta}}$, respectively. Let $c = \min\{a, \min_{0\le\|\boldsymbol{\beta}\|\le\nu_0} a_{\boldsymbol{\beta}}\}$ with a sufficiently small fixed value $\nu_0 > 0$. From the positive definiteness of $J(\boldsymbol{x})$ and (20), we have $c > 0$. By Taylor expansion,

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T J(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{x}^*) \ge f(\boldsymbol{x}^*) + \frac{c}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2,$$

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = f_{\boldsymbol{\alpha}}(\boldsymbol{x}^*_{\boldsymbol{\alpha}}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*_{\boldsymbol{\alpha}})^T J_{\boldsymbol{\beta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\alpha}})(\boldsymbol{x} - \boldsymbol{x}^*_{\boldsymbol{\alpha}}) \ge f_{\boldsymbol{\alpha}}(\boldsymbol{x}^*_{\boldsymbol{\alpha}}) + \frac{c}{2}\|\boldsymbol{x} - \boldsymbol{x}^*_{\boldsymbol{\alpha}}\|^2,$$

where $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}_{\boldsymbol{\alpha}}$ are appropriate values. Using these inequalities, we have

$$f(\boldsymbol{x}_{\boldsymbol{\alpha}}^*) - f(\boldsymbol{x}^*) \geq \frac{c}{2}\|\boldsymbol{x}_{\boldsymbol{\alpha}}^* - \boldsymbol{x}^*\|^2,$$
$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}^*) - f_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\boldsymbol{\alpha}}^*) \geq \frac{c}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_{\boldsymbol{\alpha}}^*\|^2,$$

and

$$\begin{aligned} c\|\boldsymbol{x}_{\boldsymbol{\alpha}}^* - \boldsymbol{x}^*\|^2 &\leq f(\boldsymbol{x}_{\boldsymbol{\alpha}}^*) - f(\boldsymbol{x}^*) + f_{\boldsymbol{\alpha}}(\boldsymbol{x}^*) - f_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\boldsymbol{\alpha}}^*) \\ &\leq |f(\boldsymbol{x}_{\boldsymbol{\alpha}}^*) - f_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\boldsymbol{\alpha}}^*)| + |f_{\boldsymbol{\alpha}}(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)| = O(\nu), \end{aligned}$$

where the last order holds from (19). The proof is complete. □

Proof of Theorem 3.3. We have

$$\begin{aligned} &D_{\mathrm{UKL}}(r^\dagger, r_{\boldsymbol{\theta}}; w) \\ &= -\int \boldsymbol{\theta}^T h(\boldsymbol{x}) w(\boldsymbol{x}) p^\dagger(\boldsymbol{x}) d\boldsymbol{x} + \int w(\boldsymbol{x}) p^\dagger(\boldsymbol{x}) d\boldsymbol{x} \times \log \int \exp(\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q^\dagger(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{const} \\ &= -(1-\varepsilon_p)\int \boldsymbol{\theta}^T h(\boldsymbol{x}) w(\boldsymbol{x}) p^*(\boldsymbol{x}) d\boldsymbol{x} - \varepsilon_p \nu_2 \\ &\quad + \left\{(1-\varepsilon_p)\int w(\boldsymbol{x}) p^*(\boldsymbol{x}) d\boldsymbol{x} + \varepsilon_p \nu_1\right\} \times \log\left\{(1-\varepsilon_q)\int \exp(\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q^*(\boldsymbol{x}) d\boldsymbol{x} + \varepsilon_q \nu_3\right\} + \mathrm{const} \\ &= (1-\varepsilon_p)\left\{D_{\mathrm{UKL}}(r^*, r_{\boldsymbol{\theta}}; w) + \mathrm{const} + O(\varepsilon_r \nu)\right\}. \end{aligned}$$

Let $f(\boldsymbol{\theta}) = D_{\mathrm{UKL}}(r^*, r_{\boldsymbol{\theta}}; w)$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = D_{\mathrm{UKL}}(r^\dagger, r_{\boldsymbol{\theta}}; w)/(1-\varepsilon_p) + \mathrm{const}$ with $\boldsymbol{\alpha} = (\nu_j)_{j=1}^3$. These are C^2-class real-valued functions on a compact convex set Θ. The above shows (19). Let the Hessian matrices of $f(\boldsymbol{\theta})$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ be denoted by $J(\boldsymbol{\theta})$ and $J_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ with $\boldsymbol{\beta} = (\nu_j)_{j=1}^5$, respectively. $J_{\boldsymbol{\beta}}(\boldsymbol{x})$ is continuous about $\boldsymbol{\beta}$. We can easily show that $J(\boldsymbol{\theta})$ is positive definite after simple calculation. The conditions $\|\boldsymbol{\alpha}\| = O(\nu)$ and $\|\boldsymbol{\beta}\| = O(\nu)$ are satisfied from the definition of ν. In a similar manner to the above, we can show (20). The minimizers $\boldsymbol{\theta}_{\mathrm{UKL}}^\dagger$ and $\boldsymbol{\theta}_{\mathrm{UKL}}^*$ are unique interior points in Θ from the assumption of the theorem. Therefore, all the conditions assumed in Lemma D.1 are satisfied, and we have $\boldsymbol{\theta}_{\mathrm{UKL}}^\dagger - \boldsymbol{\theta}_{\mathrm{UKL}}^* = O(\varepsilon_r \nu)$. □

D.2. Proof of Theorem 3.5

Proof. It follows that

$$\begin{aligned} &d_\gamma(r^\dagger, r_{\boldsymbol{\theta}}; w\, q^\dagger) \\ &= -\frac{1}{\gamma}\log\int \exp(\gamma \boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) p^\dagger(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{1+\gamma}\log\int \exp((1+\gamma)\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q^\dagger(\boldsymbol{x}) d\boldsymbol{x} \\ &= -\frac{1}{\gamma}\log\left\{(1-\varepsilon_p)\int \exp(\gamma \boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) p^*(\boldsymbol{x}) d\boldsymbol{x} + \varepsilon_p \nu_1'\right\} \\ &\quad + \frac{1}{1+\gamma}\log\left\{(1-\varepsilon_q)\int \exp((1+\gamma)\boldsymbol{\theta}^T h(\boldsymbol{x})) w(\boldsymbol{x}) q^*(\boldsymbol{x}) d\boldsymbol{x} + \varepsilon_q \nu_2'\right\} \\ &= d_\gamma(r^*, r_{\boldsymbol{\theta}}; w\, q^*) + \mathrm{const} + O(\varepsilon_r' \nu'). \end{aligned}$$

Let $f(\boldsymbol{\theta}) = d_\gamma(r^*, r_{\boldsymbol{\theta}}; w\, q^*)$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = d_\gamma(r^\dagger, r_{\boldsymbol{\theta}}; w\, q^\dagger)$ with $\boldsymbol{\alpha} = (\nu_j')_{j=1}^2$. These are C^2-class real-valued functions on a compact convex set Θ. The above shows (19). Let the Hessian matrices of $f(\boldsymbol{\theta})$ and $f_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ be denoted by $J(\boldsymbol{\theta})$ and $J_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ with $\boldsymbol{\beta} = (\nu_j')_{j=1}^6$, respectively. $J_{\boldsymbol{\beta}}(\boldsymbol{x})$ is continuous about $\boldsymbol{\beta}$. The conditions $\|\boldsymbol{\alpha}\| = O(\nu')$ and $\|\boldsymbol{\beta}\| = O(\nu')$ are satisfied from the definition of ν'. In a similar manner to the above, we can show (20). The minimizers $\boldsymbol{\theta}_\gamma^\dagger$ and $\boldsymbol{\theta}_\gamma^*$ are unique interior points in Θ from the assumption of the theorem. If $J(\boldsymbol{\theta})$ is positive definite, then all the conditions assumed in Lemma D.1 are satisfied and we have $\boldsymbol{\theta}_\gamma^\dagger - \boldsymbol{\theta}_\gamma^* = O(\varepsilon_r' \nu')$. However, from the assumption of the theorem, we only have the condition that $J(\boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_\gamma^*$, not over Θ. We need to bridge the gap. From the continuity of $J(\boldsymbol{\theta})$ and (19), we can take a restricted compact convex set $\tilde{\Theta} \subset \Theta$ such that $J(\boldsymbol{\theta})$ is positive definite over $\tilde{\Theta}$ and $\boldsymbol{\theta}_\gamma^\dagger, \boldsymbol{\theta}_\gamma^* \in \tilde{\Theta}$. Considering Lemma D.1 on the restricted set $\tilde{\Theta}$, we can have $\boldsymbol{\theta}_\gamma^\dagger - \boldsymbol{\theta}_\gamma^* = O(\varepsilon_r' \nu')$. □

E. MSE Values in Section 4.1

We calculated the MSE values between the estimated parameters and the ground truth. The experimental settings are described in Section 4.1. The initial parameters of the iterative calculation were randomly generated from a normal distribution for each experiment. The mean and the standard deviation of the squared errors were calculated.

Tables 3, 4, 5, and 6 show the MSE values in the "clean", "reference contamination", "target contamination", and "double contamination" settings, respectively. Table 3 shows that Weighted DRE and γ-DRE had comparable MSE values to DRE and Trimmed DRE in the range of $-1.2 \leq \theta_{1,2} \leq 1.2$. At $\theta_{1,2} = -1.6$ and $\theta_{1,2} = 1.6$, Weighted DRE and γ-DRE had significantly smaller MSE values than DRE and Trimmed DRE. This seems to be because Weighted DRE and γ-DRE eliminated the weak outliers. These tables also show that γ-DRE estimated the ground truth slightly better than Weighted DRE in all the settings. γ-DRE need not estimate the normalizing term, so it could have the lower MSE values.

Table 3. MSE of the estimated parameters in the "clean" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
-1.6	0.8	-0.8	1.0746 (1.5161)	1.2785 (3.0056)	0.7038 (0.3516)	**0.3288 (0.5279)**
-1.2	0.4	-0.8	0.4389 (0.7856)	0.4844 (0.9658)	0.4534 (0.2723)	**0.2797 (0.1731)**
-0.8	0.4	-0.4	0.1523 (0.1136)	0.1419 (0.0974)	0.2377 (0.1564)	**0.1373 (0.0755)**
-0.4	0.4	0.0	**0.0435 (0.0400)**	0.0438 (0.0417)	0.0834 (0.0883)	0.0513 (0.0400)
0.0	0.0	0.0	0.0123 (0.0164)	0.0196 (0.0393)	0.0240 (0.0491)	**0.0097 (0.0185)**
0.4	0.0	0.4	0.0565 (0.0674)	0.0551 (0.0548)	0.0883 (0.0765)	**0.0493 (0.0400)**
0.8	-0.4	0.4	0.1535 (0.1034)	0.1428 (0.0966)	0.2349 (0.1351)	**0.1387 (0.0850)**
1.2	-0.8	0.4	0.3035 (0.6157)	**0.2363 (0.2536)**	0.5086 (0.2216)	0.2631 (0.1855)
1.6	-0.8	0.8	1.5608 (2.7587)	1.4278 (2.6771)	0.7467 (0.3520)	**0.2931 (0.2971)**

Table 4. MSE of the estimated parameters in the "reference contamination" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
-1.6	0.8	-0.8	3950298.5757 (22926.6509)	35323.5783 (5264.2799)	0.9017 (0.3773)	**0.3218 (0.3032)**
-1.2	0.4	-0.8	3950542.0491 (25067.6394)	34954.4141 (4712.6414)	0.5781 (0.2530)	**0.2431 (0.1692)**
-0.8	0.4	-0.4	3958872.0875 (23063.6142)	34460.0170 (7649.7992)	0.2463 (0.1283)	**0.1531 (0.0870)**
-0.4	0.4	0.0	3956367.0423 (24930.7475)	25795.7112 (16823.4622)	0.0902 (0.0926)	**0.0436 (0.0340)**
0.0	0.0	0.0	3959068.1508 (25197.0345)	21001.3299 (19309.3519)	0.0258 (0.0527)	**0.0079 (0.0115)**
0.4	0.0	0.4	3960324.3266 (24236.2236)	19655.8422 (19797.1869)	0.1094 (0.1430)	**0.0548 (0.0427)**
0.8	-0.4	0.4	3957741.4148 (23384.0926)	20133.8894 (19859.9808)	0.2834 (0.1815)	**0.1421 (0.0773)**
1.2	-0.8	0.4	3961153.3083 (25523.0181)	21343.8024 (20202.3100)	0.5734 (0.2484)	**0.2712 (0.1711)**
1.6	-0.8	0.8	3968357.1323 (24595.6865)	21251.8250 (20522.5385)	0.9429 (0.3470)	**0.2951 (0.5538)**

Table 5. MSE of the estimated parameters in the "target contamination" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
-1.6	0.8	-0.8	549087.3957 (499298.7574)	351177.4384 (319271.9316)	0.7067 (0.3616)	**0.2708 (0.2885)**
-1.2	0.4	-0.8	637923.7106 (480891.8776)	357119.1806 (318166.7869)	0.4389 (0.2751)	**0.2971 (0.1992)**
-0.8	0.4	-0.4	507397.3182 (500213.5196)	363576.7793 (317394.2769)	0.2191 (0.1321)	**0.1538 (0.0952)**
-0.4	0.4	0.0	499076.9452 (501614.2931)	306914.5394 (320727.8412)	0.0790 (0.0667)	**0.0436 (0.0331)**
0.0	0.0	0.0	585877.3689 (481541.5629)	327077.0420 (312615.8724)	0.0186 (0.0274)	**0.0122 (0.0184)**
0.4	0.0	0.4	564275.2849 (462204.3313)	392242.2477 (307921.5425)	0.0829 (0.0748)	**0.0607 (0.0468)**
0.8	-0.4	0.4	650588.5428 (358591.2366)	430804.4411 (250226.4449)	0.2494 (0.1304)	**0.1332 (0.0845)**
1.2	-0.8	0.4	870574.0795 (138590.6119)	539042.4358 (108722.4856)	0.5089 (0.2182)	**0.2447 (0.1827)**
1.6	-0.8	0.8	891624.2286 (141870.1246)	563625.7989 (96903.0755)	0.7265 (0.3316)	**0.2620 (0.2594)**

Table 6. MSE of the estimated parameters in the "double contamination" setting (the standard deviations are in parentheses).

$\theta_{1,2}$	λ_p	λ_q	DRE	TRIMMED DRE	WEIGHTED DRE	γ-DRE
-1.6	0.8	-0.8	4.9189 (5.4462)	3.5911 (3.7100)	0.9394 (0.3565)	**0.2962 (0.4160)**
-1.2	0.4	-0.8	47.3175 (443.5423)	3.2073 (3.6291)	0.5737 (0.2540)	**0.2325 (0.1936)**
-0.8	0.4	-0.4	4.3060 (4.4582)	4.4132 (4.3617)	0.2862 (0.1567)	**0.1344 (0.0937)**
-0.4	0.4	0.0	5.1599 (5.0391)	4.5406 (4.6529)	0.0962 (0.1021)	**0.0463 (0.0391)**
0.0	0.0	0.0	5.3757 (5.5380)	5.5567 (4.8591)	0.0299 (0.0455)	**0.0131 (0.0165)**
0.4	0.0	0.4	5.5245 (5.5945)	6.0835 (6.4211)	0.0840 (0.0845)	**0.0689 (0.0704)**
0.8	-0.4	0.4	8.3192 (7.8744)	9.0380 (7.0335)	0.2950 (0.1739)	**0.1629 (0.0955)**
1.2	-0.8	0.4	8.2561 (7.7996)	10.2449 (8.2807)	0.5750 (0.2867)	**0.2175 (0.1636)**
1.6	-0.8	0.8	8.4907 (8.8919)	9.3233 (8.7860)	0.9466 (0.5066)	**0.3042 (0.2891)**

F. Comparison between *Stable* and *Robust* DRE

Our problem setting is different from the previously discussed problem setting in (Yamada et al., 2011), which we call *stable* estimation. The *stable* estimation tackles the situation where the two density functions have no common supports, which is the usual case in the high-dimensional setting (Rhodes et al., 2020; Kato & Teshima, 2021; Choi et al., 2021; 2022; Srivastava et al., 2023). The *robust* estimation aims to eliminate the disturbance effects by outliers. For example, when considering the density ratio of the main and outlier distributions mixture, the *stable* DRE estimates the density ratio of the mixtures, whereas the *robust* DRE only estimates that of the main distributions.

We compare the performance of RuLSIF (Relative unconstrained Least-Squares Importance Fitting) (Yamada et al., 2011), one of the famous *stable* non-parametric DRE methods, and Weighted DRE in the case where the different outliers contaminate the main distribution. We prepare three cases:

- Uncontaminated: $p(x) = q(x) = N(0, 1)$.
- Same outliers: $p(x) = q(x) = 0.8N(0, 1) + 0.2N(15, 1)$.
- Different outliers: $p(x) = 0.8N(0, 1) + 0.2N(15, 1)$ and $q(x) = 0.8N(0, 1) + 0.2N(10, 1)$.

In all the cases, the ground truth of the density ratio is $r(x) = 1$. The dataset sizes are $n_p = n_q = 100$. We used a Python code of RuLSIF provided on GitHub [1]. The hyper-parameter α in RuLSIF is set as 0.95, achieving the high stability.

Figure 4 shows that RuLSIF fails to estimate the ground truth in the "Different outliers" setting, although Weighted DRE can estimate it correctly. In the "Different outliers" setting, RuLSIF falsely detects the change of the outlier distributions, that is, $N(15, 1)$ in $p(x)$ and $N(10, 1)$ in $q(x)$. Because this misspecification affects the estimation of the main distributions $N(0, 1)$, the estimated density ratio is different from the ground truth $r(x) = 1$.

[1] https://github.com/hoxo-m/densratio_py/

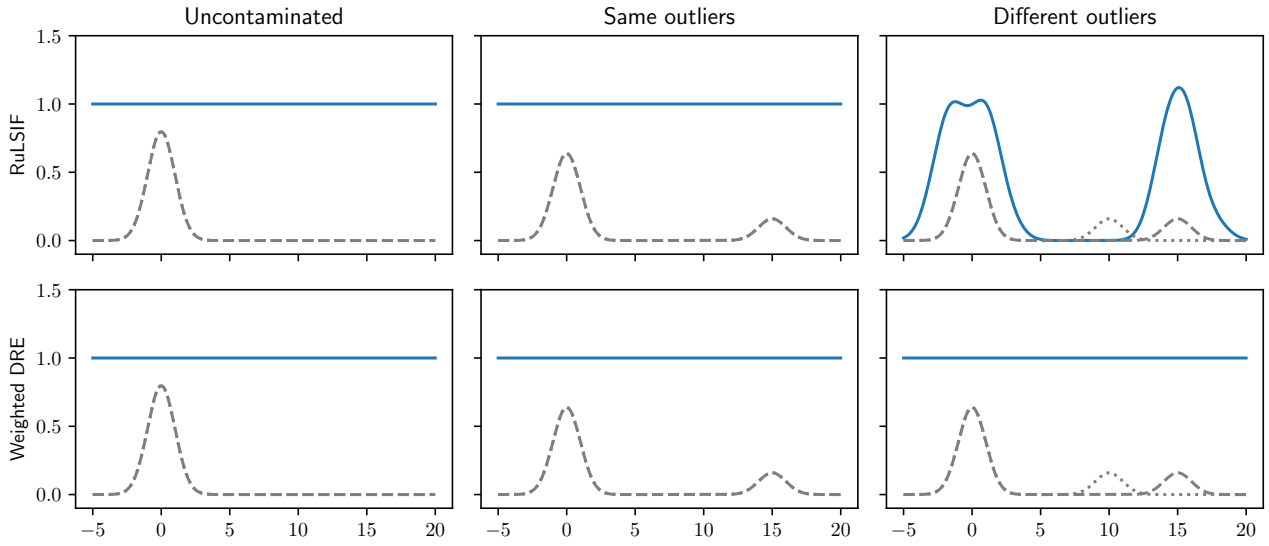

Figure 4. Comparison of the performance between RuLSIF and Weighted DRE. The x-axis shows the data values of x, and the y-axis shows the values of the density ratio or the density functions. The blue lines show the estimated density ratio values, and the gray dashed lines show the density functions. For all figures, the true density ratio is $r(x) = 1$.