Unbiased Estimating Equation and Latent Bias Under f-Separable Bregman Distortion Measures

Masahiro Kobayashi, *Member, IEEE*, and Kazuho Watanabe, *Member, IEEE*

***Abstract*— We discuss unbiased estimating equations in a class of objective functions using a monotonically increasing function f and Bregman divergence. The choice of the function f gives desirable properties, such as robustness against outliers. To obtain unbiased estimating equations, analytically intractable integrals are generally required as bias correction terms. In this study, we clarify the combination of Bregman divergence, statistical model, and function f in which the bias correction term vanishes. Focusing on Mahalanobis and Itakura–Saito distances, we generalize fundamental existing results and characterize a class of distributions of positive reals with a scale parameter, including the gamma distribution as a special case. We also generalized these results to general model classes characterized by one-dimensional Bregman divergence. Furthermore, we discuss the possibility of latent bias minimization when the proportion of outliers is large, which is induced by the extinction of the bias correction term. We conducted numerical experiments to show that the latent bias can approach zero under heavy contamination of outliers or very small inliers.**

***Index Terms*— f-separable distortion measures, Bregman divergence, Itakura–Saito distance, latent bias, M-estimators, unbiased estimating equations.**

I. Introduction

THE maximum likelihood estimation (MLE) for the statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$ estimates the parameter $\boldsymbol{\theta}$ by minimizing the negative log-likelihood. It is equivalent to empirical inference under the Kullback–Leibler (KL)-divergence. However, MLE is susceptible to outliers or mismatches of the assumed model. In robust statistics, estimation methods weakening adverse effects of outliers have been studied [1], [2]. M-estimation is a well-known method that changes the negative log-likelihood function $-\frac{1}{n}\sum_{i=1}^{n}\log p(\boldsymbol{x}_i|\boldsymbol{\theta})$ in MLE to a general robust objective function $\frac{1}{n}\sum_{i=1}^{n}\rho(\boldsymbol{x}_i,\boldsymbol{\theta})$ [1], [2]. This is equivalent to changing the assumed model to heavy-tailed distribution under the KL-divergence measure. Another definition is given by solving the following estimating equation with respect to the parameter $\boldsymbol{\theta}$,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\boldsymbol{x}_i,\boldsymbol{\theta})=\mathbf{0}, \tag{1}$$

where the function ψ is the generalized score function in MLE [1], [2]. Although the two definitions are not always the same, they concur if the condition $\psi(\boldsymbol{x},\boldsymbol{\theta})=\frac{\partial\rho(\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ holds.

Another well-known estimation method is the minimum divergence estimation, which is to change KL-divergence to another divergence [3], [4]. Well-known divergences are f- and Bregman divergences, which are defined by a strictly convex function [5]. A common part of these divergences is only KL-divergence [6]. The f-divergence includes the α-divergence as a subclass, which plays an important role in information geometry [7]. Particularly, estimators based on the minimization of Hellinger distance, which is an α-divergence, have robustness and completely efficient properties [8]. However, the continuous model estimation using f-divergence requires the use of non-parametric kernel density estimators instead of the true distribution. Non-parametric kernel density estimators have the disadvantages of a bandwidth selection problem and worse estimation efficiency in high dimensions. There are known methods to deal with this problem, such as using the same kernel to represent empirical and model distributions [9], or using a dual representation of divergence [10], [11], [12]. However, the estimation based on minimizing the Bregman divergence does not require kernel density estimator because the empirical distribution is available. Broniatowski et al. [13] called this type of divergence in which the empirical distribution can be used instead of the true distribution decomposable divergence. Jana and Basu [14] simply called such divergence non-kernel divergence and showed that single-integral and non-kernel divergences are limited to Bregman divergence. The β-divergence, also known as density power divergence, which belongs to the Bregman divergence is the first non-kernel divergence proposed as an extension of M-estimation (1) [15]. Since then, robust non-kernel divergences applicable to empirical inference have been developed.

The minimization of these divergences reduces to estimating equations by the weighted (negative) score function $s(\boldsymbol{x},\boldsymbol{\theta})=\frac{\partial l(\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, where $l(\boldsymbol{x},\boldsymbol{\theta})=-\log p(\boldsymbol{x}|\boldsymbol{\theta})$. Conversely, these divergences are constructed from estimating equations. The

Manuscript received 20 July 2022; revised 28 November 2023; accepted 6 February 2024. Date of publication 16 February 2024; date of current version 16 July 2024. This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP23K16849 and Grant JP19K11825. An earlier version of this paper was presented in part at the 2020 IEEE Information Theory Workshop [DOI: 10.1109/ITW46852.2021.9457678]. *(Corresponding author: Masahiro Kobayashi.)*

Masahiro Kobayashi is with the Information and Media Center, Toyohashi University of Technology, Toyohashi 441-8580, Japan (e-mail: kobayashi@imc.tut.ac.jp).

Kazuho Watanabe is with the Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi 441-8580, Japan (e-mail: wkazuho@cs.tut.ac.jp).

Communicated by P. Netrapalli, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3366538.

Digital Object Identifier 10.1109/TIT.2024.3366538

following two types of estimating equations are well known:

$$\frac{1}{n}\sum_{i=1}^{n}\xi(l(\boldsymbol{x}_i,\boldsymbol{\theta}))s(\boldsymbol{x}_i,\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))s(\boldsymbol{X},\boldsymbol{\theta})\right], \tag{2}$$

$$\frac{\sum_{i=1}^{n}\xi(l(\boldsymbol{x}_i,\boldsymbol{\theta}))s(\boldsymbol{x}_i,\boldsymbol{\theta})}{\sum_{j=1}^{n}\xi(l(\boldsymbol{x}_j,\boldsymbol{\theta}))} = \frac{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))s(\boldsymbol{X},\boldsymbol{\theta})\right]}{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))\right]}, \tag{3}$$

where $\xi:\mathbb{R}\to\mathbb{R}$ works as the weight function. These estimating equations are included in the M-estimation framework (1) by putting the function ψ as follows:

$$\psi(\boldsymbol{x},\boldsymbol{\theta}) = \xi(l(\boldsymbol{x},\boldsymbol{\theta}))s(\boldsymbol{x},\boldsymbol{\theta}) - \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))s(\boldsymbol{X},\boldsymbol{\theta})\right],$$
$$\psi(\boldsymbol{x},\boldsymbol{\theta}) = \xi(l(\boldsymbol{x},\boldsymbol{\theta}))s(\boldsymbol{x},\boldsymbol{\theta})\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))\right] - \xi(l(\boldsymbol{x},\boldsymbol{\theta}))\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{X},\boldsymbol{\theta}))s(\boldsymbol{X},\boldsymbol{\theta})\right],$$

respectively [16]. Equation (2) is called the *non-normalized estimating equation* because the summation of weights of score functions is not one. This estimating equation corresponds to the Bregman divergence, and its special cases (β-divergence [15], and its generalizations [17], [18], [19]) and variants (U-divergence [20], Ψ-divergence [21]). Equation (3) is the *normalized estimating equation* because the summation of weights of score functions is one. Windham [22] proposed an estimator using density power weight, which is equivalent to the solution to equation (3). Then, Jones et al. [23] constructed the corresponding divergence. This divergence, called γ-divergence, has the property that the latent bias can be minimized when the proportion of outliers is large and that the divergence with such a property is unique under some assumptions [24]. This property of the γ-divergence was extended to the normalized estimating equation (3) with the general weight ξ [16]. However, these approaches require bias correction terms, i.e., the right-hand sides of (2) and (3), which generally result in analytically intractable integrals. This is true for most practical models except for a few simple cases. For example, if the weight function ξ is a power function that corresponds to β- and γ-divergences, and the statistical model is a specific case within the exponential family, such as Gaussian, gamma, and inverse Gaussian distributions, the bias correction term can be calculated. However, if the statistical model is complex or for general weight functions, analytical computation becomes intractable.

Recently, following the success of divergences using density power weight such as β- and γ-divergences, the extension, unification, and relationship of these divergences have been investigated. There are two mainstream research directions. The first direction is to extend the existing divergences within non-kernel divergences. Kanamori and Fujisawa [25] proposed Hölder divergence, which establishes invariance to the affine transformation of random variables based on the composite score. In their approach, the proportion of outliers can be estimated by considering the unnormalized density [26]. Kuchibhotla et al. [27] proposed the bridge density power divergence (BDPD), which is constructed by a convex combination of estimating equations (2) and (3) using the density power weight. They tried to deal with the spurious global solution problem that γ-divergence produces. Both Hölder divergence and BDPD include β- and γ-divergences as special cases. The γ-divergence is generated by the logarithmic transformation of each term of the β-divergence. Ray et al. [28] proposed the functional density power divergence (FDPD) which is generated by the general functional transformation of each term of the β-divergence. It contains BDPD [27] and Jones et al.'s general divergence [23]. To provide better robustness versus efficiency trade-off, the expansion of β-divergence has been investigated within the Bregman divergence framework [17], [18], [19]. Notably, β-divergence, γ-divergence, Bregman divergence, and BDPD correspond to M-estimation (1), whereas Hölder divergence and FDPD do not necessarily correspond to it.

The second direction is to extend beyond the non-kernel divergence framework to the super family of divergences that include many existing divergences. Ghosh et al. [29] proposed super (S)-divergence, which generalizes α- and β-divergences and has two tuning parameters. The cases of continuous and discrete models have been investigated, and it has been reported that the regions outside α- and β-divergences show good performance [29], [30], [31]. This divergence has been further generalized [32]. Maji et al. [33] proposed a logarithm transformation divergence for each term of S-divergence, which includes Rényi- [34] and γ- [24] divergences and called it logarithmic S-divergence (LSD). S-divergence and LSD correspond to the non-normalized and normalized estimating equations, respectively. However, these estimating equations cannot be expressed by the summation of independent and identically distributed data points as in (2) and (3). In estimators based on both S-divergence and LSD, the asymptotic variance depends on only one of the two tuning parameters. Maji et al. [35] proposed C-divergence, which is a very wide divergence class and includes f- [5] and generalized S- [32] divergences. In fact, this divergence was previously proposed by Vonta et al. [36] and used for testing. Basak and Basu [37] considered a new divergence by giving the argument of the Bregman divergence a power of the density function. It is called generalized S–Bregman divergence, which includes S- [29] and Bregman exponential [17] divergences and has three tuning parameters.

In this paper, we consider the M-estimation under f-separable distortion measures, which were proposed to extend linear distortion, such as the average distortion to nonlinear distortion, and for which the rate-distortion function was studied [38]. It was also used to solve the estimation problem with Bregman divergence as the base distortion measure, and a simple clustering or vector quantization algorithm was constructed [39]. In this paper, this class of objective functions is called the f-separable Bregman distortion measure. Note that this distortion measure is defined by neither f- nor Bregman divergences between the aforementioned probability distributions. It is defined by a function f, not necessarily convex, and the Bregman divergence between vector or scalar variables. The M-estimation under this distortion measure, as discussed in Section III, can be viewed as a deviance-based estimation of the regular exponential family model. The unbiasedness of the estimating equation of deviance-based methods

has been studied, and some sufficient conditions for it have been obtained [40], [41]. However, these results only apply to the case where the data-generating distribution is included in the assumed model. On the other hand, the M-estimation of the location family is proved to have an unbiased estimating equation for general symmetric distributions [2]. It is unknown in what cases of f-separable Bregman distortion measures the estimating equation is unbiased for such a general class of distributions. If an estimating equation is unbiased, it can be regarded as normalized, and the estimator has the potential to minimize the latent bias, even if the proportion of outliers is large.

In this paper, we study the conditions for bias correction terms of f-separable Bregman distortion measures to vanish and characterize the combination of Bregman divergence, statistical model, and function f. Focusing on Mahalanobis and Itakura–Saito (IS) distances, we specify the conditions for the general model classes and the function f to achieve unbiased estimating equations. We also describe the properties of these general model classes and discuss the relationship between these models and regular exponential family. Furthermore, we discuss if the latent bias can be minimized when the proportion of outliers is large. We compare the M-estimation under the f-separable IS distortion measure with the estimation methods minimizing β- and γ-divergences theoretically in terms of asymptotic efficiency and numerically through experiments examining latent bias under heavy contamination.

II. f-Separable Bregman Distortion Measures

This section introduces the estimation method based on f-separable Bregman distortion measures [39]. We consider estimating the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$ of a statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$ when given the data $\boldsymbol{x}^n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i = (x_i^{(1)}, \cdots, x_i^{(d)})^{\mathrm{T}} \in \boldsymbol{\chi} \subseteq \mathbb{R}^d$. We assume that $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$ is the data-generating distribution, and the parameter $\boldsymbol{\theta}$ is the expected value of $\boldsymbol{x}$ under the model, i.e., $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{X}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$ if it exists. The objective function is defined by

$$L_f(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} f\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right), \tag{4}$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a differentiable and continuous monotonically increasing function, $d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\chi} \times \boldsymbol{\Theta} \to \mathbb{R}_+$ is the Bregman divergence, and $\mathbb{R}_+$ is the set of nonnegative real numbers. The Bregman divergence is defined by a differentiable strictly convex function $\phi : \boldsymbol{\chi} \to \mathbb{R}$ as

$$d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) \triangleq \phi(\boldsymbol{x}) - \phi(\boldsymbol{\theta}) - \langle \boldsymbol{x} - \boldsymbol{\theta}, \nabla\phi(\boldsymbol{\theta}) \rangle, \tag{5}$$

where $\nabla\phi$ is its gradient vector and $\langle \cdot, \cdot \rangle$ is the inner product. The corresponding estimating equation to the objective function (4) is given by

$$\frac{1}{n} \sum_{i=1}^{n} f'(d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta})) \frac{\partial}{\partial \boldsymbol{\theta}} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \tag{6}$$

where f' is the derivative of f. This is not generally unbiased. The estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}^*$ is given by the solution of the estimating equation (6). The property of the estimator depends on the function f. For example, if the function f is concave, the estimator is robust against outliers. Then, the update rule of the estimator $\hat{\boldsymbol{\theta}}$ is given by

$$\boldsymbol{\theta} = \frac{\sum_{i=1}^{n} f'\left(d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta})\right) \boldsymbol{x}_i}{\sum_{j=1}^{n} f'\left(d_\phi(\boldsymbol{x}_j, \boldsymbol{\theta})\right)} \tag{7}$$

which gives the iterative algorithm. When the function f satisfies $f(0) > -\infty$, the update rule converges with finite iteration [39]. In other words, the estimator is one of the local minima of the objective function (4).

The original f-separable distortion measures are defined by f-mean with respect to some base distortion d [38]. From the viewpoint of f-mean, representative examples are the log–sum–exp function and power mean, which are, respectively, given by the following functions:

$$f_\alpha(z) = \begin{cases} \frac{1-\exp(-\alpha z)}{\alpha} & (\alpha \neq 0), \\ z & (\alpha = 0), \end{cases} \tag{8}$$

$$f_\beta(z) = \begin{cases} \frac{(z+a)^\beta - 1}{\beta} & (\beta \neq 0), \\ \log(z + a) & (\beta = 0), \end{cases} \tag{9}$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $a \in \mathbb{R}_+$. When $\alpha = 0$ or $\beta = 0$, functions (8) and (9) are, respectively, given by the following continuous limits:

$$f_0(z) = \lim_{\alpha \to 0} f_\alpha(z) = z,$$
$$f_0(z) = \lim_{\beta \to 0} f_\beta(z) = \log(z + a).$$

If tuning parameters satisfy $\alpha > 0$ and $\beta < 1$, the estimators become robust. When $\alpha = 0$ and $\beta = 1$, functions (8) and (9) become linear functions. The derivative of functions (8) and (9) are, respectively, given by

$$f'_\alpha(z) = \exp(-\alpha z),$$
$$f'_\beta(z) = (z + a)^{\beta - 1}.$$

III. Relation to Robust Divergences

First, we show that the minimization of $L_f(\boldsymbol{\theta})$ is derived from deviance-based M-estimation of the expectation parameter under the regular exponential family,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = r_\phi(\boldsymbol{x}) \exp(-d_\phi(\boldsymbol{x}, \boldsymbol{\theta})), \tag{10}$$

where $r_\phi(\boldsymbol{x})$ is uniquely determined by the strictly convex function ϕ and $\boldsymbol{\theta}$ is the expectation parameter, i.e., $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{X}]$ [42]. The deviance function [40] is defined by

$$\Delta(\boldsymbol{x}, \boldsymbol{\theta}) = 2 \left[l(\boldsymbol{x}, \boldsymbol{\theta}) - \inf_{\boldsymbol{\tau}} \left(l(\boldsymbol{x}, \boldsymbol{\tau}) \right) \right]. \tag{11}$$

The objective function of the deviance-based M-estimation is defined by

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\sqrt{\Delta(\boldsymbol{x}_i, \boldsymbol{\theta})}),$$

where $\rho : \mathbb{R}_+ \to \mathbb{R}$. If the function ρ is differentiable, the corresponding estimating equation is given by

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\rho'(\sqrt{\Delta(\boldsymbol{x}_i, \boldsymbol{\theta})})}{\sqrt{\Delta(\boldsymbol{x}_i, \boldsymbol{\theta})}} s(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{0},$$

where ρ' is the derivative of ρ. From (11), the deviance function of the regular exponential family (10) is given by

$$\begin{aligned}&\Delta(\boldsymbol{x},\boldsymbol{\theta})\\&=2\left[-\log r_\phi(\boldsymbol{x})+d_\phi(\boldsymbol{x},\boldsymbol{\theta})-\inf_{\boldsymbol{\tau}}\left(-\log r_\phi(\boldsymbol{x})+d_\phi(\boldsymbol{x},\boldsymbol{\tau})\right)\right]\\&=2\left[-\log r_\phi(\boldsymbol{x})+d_\phi(\boldsymbol{x},\boldsymbol{\theta})+\log r_\phi(\boldsymbol{x})\right]=2d_\phi(\boldsymbol{x},\boldsymbol{\theta}).\end{aligned}$$

Thus, the objective function and estimating equation under the regular exponential family are, respectively, given by

$$\frac{1}{n}\sum_{i=1}^n\rho\left(\sqrt{2d_\phi(\boldsymbol{x}_i,\boldsymbol{\theta})}\right), \tag{12}$$

$$\frac{1}{n}\sum_{i=1}^n\frac{\rho'\left(\sqrt{2d_\phi(\boldsymbol{x}_i,\boldsymbol{\theta})}\right)}{\sqrt{2d_\phi(\boldsymbol{x}_i,\boldsymbol{\theta})}}\frac{\partial}{\partial\boldsymbol{\theta}}d_\phi(\boldsymbol{x}_i,\boldsymbol{\theta})=\mathbf{0}, \tag{13}$$

where $s(\boldsymbol{x},\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}d_\phi(\boldsymbol{x},\boldsymbol{\theta})$ is the negative score function. The objective functions of f-separable Bregman distortion measures (4) and deviance-based M-estimation (12) are related as follows $f(z)=\rho(\sqrt{2z})$. Similarly, the estimating equations (6) and (13) have the following relationship $f'(z)=\rho'(\sqrt{2z})/\sqrt{2z}$.

Next, we turn to empirical inference based on robust divergences under the regular exponential family (10). Suppose for a moment that the bias correction term can be ignored. In this case, the non-normalized estimating equation (2) is given by

$$\frac{1}{n}\sum_{i=1}^n\xi\left(l(\boldsymbol{x}_i,\boldsymbol{\theta})\right)\frac{\partial}{\partial\boldsymbol{\theta}}d_\phi(\boldsymbol{x}_i,\boldsymbol{\theta})=\mathbf{0}. \tag{14}$$

Compared with this equation, the estimating equation (6) of f-separable Bregman distortion measures can be interpreted as a weighted score function. We focus on the arguments of the weight functions of (6) and (14). The only difference is the term $\inf_{\boldsymbol{\theta}}l(\boldsymbol{x},\boldsymbol{\theta})=-\log r_\phi(\boldsymbol{x})$. Specifically, if the domain of the function f' is extended to $(-\infty,\infty)$, the function f' works identically to the weight function ξ. In view of this relation, function (8) associated with the log–sum–exp function yields the estimation methods that minimize β- and γ-divergences with the non-normalized and normalized estimating equations (2) and (3), respectively. In particular, if the function f is given by (8) and the squared distance is used to estimate the standard Gaussian location parameter, the estimating equation (6) reduces to

$$\frac{1}{n}\sum_{i=1}^n\exp\left(-\alpha\|\boldsymbol{x}_i-\boldsymbol{\theta}\|^2\right)(\boldsymbol{x}_i-\boldsymbol{\theta})=\mathbf{0},$$

which is the same estimating equation as β- and γ-divergences. In other words, when we assume the regular exponential family and the function (8), then it is related to the estimation based on the power of the statistical model.

In this section, we assumed the bias correction term is exactly $\mathbf{0}$; however, it is not generally true. With the combination of the model and Bregman divergence discussed in the next section, the estimating equation (6) becomes unbiased without any bias correction term for any function f satisfying the condition given in the main theorems.

IV. Conditions for Unbiased Estimating Equation

In general, the estimator based on f-separable Bregman distortion measures introduced in Section II does not satisfy consistency because its estimating equation is not necessarily unbiased. Thus, to satisfy an unbiased estimating equation, we must subtract the bias correction term $b_f(\boldsymbol{\theta})$ from the objective function (4) as follows:

$$\begin{aligned}L_f(\boldsymbol{\theta})&=\frac{1}{n}\sum_{i=1}^n f\left(d_\phi\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right)\right)-b_f(\boldsymbol{\theta}),\\b_f(\boldsymbol{\theta})&=-\int\nabla\nabla\phi(\boldsymbol{\theta})\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)(\boldsymbol{X}-\boldsymbol{\theta})\right]d\boldsymbol{\theta},\end{aligned} \tag{15}$$

where $\int\cdot d\boldsymbol{\theta}$ denotes the indefinite integral with respect to $\boldsymbol{\theta}$. Differentiating (15) with respect to $\boldsymbol{\theta}$ and setting it to $\mathbf{0}$, we obtain the following estimating equation,

$$\begin{aligned}&\nabla\nabla\phi(\boldsymbol{\theta})\frac{1}{n}\sum_{i=1}^n f'\left(d_\phi\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right)\right)(\boldsymbol{x}_i-\boldsymbol{\theta})\\&=\nabla\nabla\phi(\boldsymbol{\theta})\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)(\boldsymbol{X}-\boldsymbol{\theta})].\end{aligned}$$

Multiplying both sides by the inverse matrix $\left(\nabla\nabla\phi(\boldsymbol{\theta})\right)^{-1}$ yields the following non-normalized estimating equation:

$$\frac{1}{n}\sum_{i=1}^n f'(d_\phi\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right))(\boldsymbol{x}_i-\boldsymbol{\theta})=\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)(\boldsymbol{X}-\boldsymbol{\theta})].$$

We can also consider the normalized estimating equation as follows:

$$\frac{\sum_{i=1}^n f'(d_\phi\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right))(\boldsymbol{x}_i-\boldsymbol{\theta})}{\sum_{j=1}^n f'\left(d_\phi\left(\boldsymbol{x}_j,\boldsymbol{\theta}\right)\right)}=\frac{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right))(\boldsymbol{X}-\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)]}.$$

Fujisawa [16] elucidated that this estimating equation can possibly minimize the latent bias, even for a large proportion of outliers. In both cases, it is necessary to calculate the integral for bias correction for each combination of statistical model, Bregman divergence, and function f. However, the integral may not exist or be analytically intractable in many cases. In this paper, we discuss the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^n f'\left(d_\phi\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right)\right)(\boldsymbol{x}_i-\boldsymbol{\theta})=\mathbf{0}.$$

It means that the bias correction term is independent of the parameter $\boldsymbol{\theta}$. In other words, the following equation is satisfied,

$$\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)(\boldsymbol{X}-\boldsymbol{\theta})\right]=\mathbf{0}. \tag{16}$$

Then, this estimating equation is automatically normalized. Therefore, the estimator has the possibility to minimize the latent bias, even for many outliers. In the remaining section, we characterize the combination of the statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$, Bregman divergence $d_\phi(\boldsymbol{x},\boldsymbol{\theta})$, and function f, where the bias correction term vanishes. In what follows, the statistical model considered is generally not the regular exponential family.

In particular, we focus on Mahalanobis and IS distances. To estimate the location parameter of elliptical distribution,

it is known that the bias correction term vanishes, and the estimator is consistent under certain conditions on the function f [2]. Moreover, it is known that the bias correction term vanishes for a log-gamma regression model. This is equivalent to the case where IS distance is used, and the model is the gamma distribution [41]. In this paper, we derive a simple condition of the function f that induces an unbiased estimating equation. Even if the weight function f' is changed, the calculation of the bias correction term is unnecessary; only the simple conditions should be checked. In particular, for the IS distance, the class of the model is extended to a more general class.

A. Mahalanobis Distance

Suppose the strictly convex function is given by $\phi(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a positive definite matrix. Then, the corresponding Bregman divergence is given by

$$d_{\mathrm{Mah.}}(\boldsymbol{x},\boldsymbol{\theta}) \triangleq (\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\theta}),$$

which is called Mahalanobis distance. If the positive definite matrix $\boldsymbol{A}$ is identity, it reduces to the squared distance,

$$\|\boldsymbol{x}-\boldsymbol{\theta}\|^2 = \sum_{j=1}^{d}(x^{(j)}-\theta^{(j)})^2.$$

We assume that the statistical model is the elliptical distribution.

Definition 1 (Elliptical Distribution [43], [44, pp. 46–47]): For $\boldsymbol{x} \in \mathbb{R}^d$ and the location parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R}^d$ and a nonnegative function $g : \mathbb{R}_+ \to \mathbb{R}_+$, let $C_{\mathrm{Mah.}} = 2\pi^{\frac{d}{2}}/\Gamma(\frac{d}{2})\int_0^\infty t^{d-1}g(t^2)dt < \infty$ be the normalization constant, and the positive definite matrix $\boldsymbol{A}$ be the inverse of a fixed covariance matrix. Then, the elliptical distribution is defined by the following probability density function,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{|\boldsymbol{A}|^{\frac{1}{2}}}{C_{\mathrm{Mah.}}}g((\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\theta})). \tag{17}$$

This distribution includes Gaussian, Laplace, t distributions and so on [44].

Assumption 1: There exists the elliptical distribution (17) corresponding to the nonnegative function g, i.e., $C_{\mathrm{Mah.}} = 2\pi^{\frac{d}{2}}/\Gamma(\frac{d}{2})\int_0^\infty t^{d-1}g(t^2)dt < \infty$.

Theorem 1: Under Assumption 1, if and only if the following condition holds against the combination of the function f and the statistical model (17), the estimating equation without a bias correction term or equivalently (16) holds:

$$\int_0^\infty g(t)f'(t)t^{\frac{d-1}{2}}dt < \infty. \tag{18}$$

The proof of Theorem 1 is in Appendix A. Although in this case, the unbiased estimating equation is intuitively trivial because of the symmetry around $\boldsymbol{\theta}$ and has been pointed out in the literature [2], the explicit condition for unbiasedness has never been discussed.

B. IS Distance

Suppose the strictly convex function is given by $\phi(x) = -\log x$. Then, the corresponding Bregman divergence is given by

$$d_{\mathrm{IS}}(x,\theta) \triangleq \frac{x}{\theta} - \log\frac{x}{\theta} - 1, \tag{19}$$

which is called the IS distance.

Definition 2 (IS Distribution): For $x \in \mathbb{R}_+ \setminus \{0\}$ and the scale parameter $\theta \in \Theta = \mathbb{R}_+\setminus\{0\}$, and a nonnegative function $g : \mathbb{R}_+ \to \mathbb{R}_+$, we define the following probability density function with the normalization constant $C_{\mathrm{IS}} < \infty$,

$$p(x|\theta) = \frac{1}{C_{\mathrm{IS}}}\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta)). \tag{20}$$

The normalization constant C_{IS} independent of the parameter θ is given by

$$C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))dx = \int_0^\infty \frac{1}{t}g(d_{\mathrm{IS}}(t,1))dt. \tag{21}$$

We used integration by substitution $t = x/\theta$. When the expectation exists, the scale parameter coincides with the expectation. In particular, if $g(z) = \exp(-kz)$, the IS distribution reduces to the gamma distribution with the known shape parameter $k > 0$,

$$p(x|\theta) = \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)}x^{k-1}\exp\left(-\frac{k}{\theta}x\right),$$

where $\Gamma(\cdot)$ is the gamma function. Details of the IS distribution are described in Section V-A.

Assumption 2: There exists the IS distribution (20) corresponding to the nonnegative function g, i.e., $C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x}g(d_{\mathrm{IS}}(x,1))dx < \infty$.

Theorem 2: Under Assumption 2, if and only if the following condition holds against the combination of the function f and statistical model (20), the estimating equation without a bias correction term or equivalently (16) holds:

$$\int_0^\infty g(t)f'(t)dt < \infty. \tag{22}$$

Proof: From the left-hand side of (16), substituting the IS distance (19) and IS distribution (20), we have

$$\begin{aligned}
&\mathbb{E}_{p(x|\theta)}\left[f'\left(d_{\mathrm{IS}}\left(X,\theta\right)\right)\left(X-\theta\right)\right] \\
&= \int_0^\infty \frac{1}{C_{\mathrm{IS}}}\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))f'(d_{\mathrm{IS}}(x,\theta))(x-\theta)dx \\
&\propto \int_0^\theta \frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))f'(d_{\mathrm{IS}}(x,\theta))(x-\theta)dx \\
&\quad + \int_\theta^\infty \frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))f'(d_{\mathrm{IS}}(x,\theta))(x-\theta)dx \\
&= \theta\int_\infty^0 g(t)f'(t)dt + \theta\int_0^\infty g(t)f'(t)dt = 0.
\end{aligned}$$

We used integration by substitution, $t = d_{\mathrm{IS}}(x,\theta)$. Therefore, if the integral (22) exists, then (16) holds, i.e., the unbiased estimating equation holds without any bias correction term. Conversely, the above discussion also shows that $\mathbb{E}_{p(x|\theta)}\left[|f'(d_{\mathrm{IS}}(X,\theta))(X-\theta)|\right] \propto 2\theta\int_0^\infty g(t)f'(t)dt$. This means that the condition (22) is also necessary. □

1) Example: Gamma Distribution: In the case of the function (8) and gamma distribution with the known shape parameter $k > 0$, i.e., $g(z) = \exp(-kz)$, then the integral in equation (22) is given by

$$\int_0^\infty \exp(-kz)\exp(-\alpha z)dz = \int_0^\infty \exp(-(k+\alpha)z)dz.$$

Therefore, the condition $\alpha > -k$ must be satisfied for the integral to be bounded. In other words, the lower limit of α that satisfies the unbiased estimating equation differs for each shape parameter k. Since $k > 0$, we can see that the condition of Theorem 2 is satisfied if $\alpha > 0$, for which the estimator is robust against outliers.

For the function (9) and gamma distribution with the known shape parameter $k > 0$, the integral in equation (22) becomes

$$\int_0^\infty \exp(-kz)(z+a)^{\beta-1}dz.$$

When $a > 0$, the condition of Theorem (22) holds for $\beta < \infty$. Moreover, when $a = 0$, the condition of Theorem (22) holds for $0 < \beta < \infty$. However, it does not hold for $\beta \leq 0$.

C. Other Bregman Divergence

The conditions of Theorems 1 and 2 are the same in one-dimensional case. A common point is that the statistical model is expressed by the Bregman divergence used for estimation. Hence, the results of Theorems 1 and 2 can be extended to a wider class of continuous distributions written by Bregman divergence. We assume the following statistical model, which is defined by one-dimensional Bregman divergence.

Definition 3 (Continuous Bregman Distribution): For $x \in (a,b) \subseteq \mathbb{R}$, the parameter $\theta \in \Theta = (a,b) \subseteq \mathbb{R}$, and the function $g : \mathbb{R}_+ \to \mathbb{R}_+$, we define the following probability density function with the normalization constant satisfying $C_\phi(\theta) < \infty$,

$$p(x|\theta) = \frac{1}{C_\phi(\theta)} \frac{\phi'(x) - \phi'(\theta)}{x - \theta} g(d_\phi(x,\theta)). \tag{23}$$

Here, $a \in \mathbb{R} \cup \{-\infty\}$ and $b \in \mathbb{R} \cup \{\infty\}$ express the left and right edges of the support of the probability density function which depend on the strictly convex function ϕ. For example, if $\phi(x) = -\log x$, $a = 0$ and $b = \infty$, and if $\phi(x) = x^2$, $a = -\infty$ and $b = \infty$. In general, the normalization constant $C_\phi(\theta)$ depends on the parameter θ. This distribution includes one-dimensional elliptical and IS distribution. Specifically, if (24) holds, and the expected value exists, $\mathbb{E}[X] < \infty$, $\mathbb{E}[X] = \theta$ holds from the estimating equation (16) with $f(z) = z$, and the condition for it is given by (25). Details of the continuous Bregman distribution is discussed in Section V-B.

Assumption 3:

1) Bregman divergence satisfies the following for any θ and a positive constant ζ (including ∞) with respect to the support (a,b) of (23):

$$\lim_{x \to a} d_\phi(x,\theta) = \lim_{x \to b} d_\phi(x,\theta) = \zeta. \tag{24}$$

2) Bregman divergence used for estimation corresponds to that of the model (23).

Assumption 4: There exists the continuous Bregman distribution (23) corresponding to the nonnegative function g, i.e., $C_\phi(\theta) < \infty$.

Theorem 3: Under Assumption 3 and Assumption 4, if and only if the following condition holds against the combination of the function f and statistical model (23), the estimating equation without a bias correction term or equivalently (16) holds:

$$\int_0^\zeta g(t)f'(t)dt < \infty. \tag{25}$$

Proof: From the left-hand side of equation (16), substituting the one-dimensional Bregman divergence (5) and the continuous Bregman distribution (23), we have

$$\begin{aligned}
&\mathbb{E}_{p(x|\theta)}\left[f'\left(d_\phi\left(X,\theta\right)\right)\left(X-\theta\right)\right] \\
&= \int_a^b \frac{1}{C_\phi(\theta)} \frac{\phi'(x) - \phi'(\theta)}{x-\theta} g(d_\phi(x,\theta)) f'(d_\phi(x,\theta))(x-\theta)dx \\
&\propto \int_a^\theta (\phi'(x) - \phi'(\theta)) g(d_\phi(x,\theta)) f'(d_\phi(x,\theta))dx \\
&\quad + \int_\theta^b (\phi'(x) - \phi'(\theta)) g(d_\phi(x,\theta)) f'(d_\phi(x,\theta))dx \\
&= \int_\zeta^0 g(t)f'(t)dt + \int_0^\zeta g(t)f'(t)dt = 0.
\end{aligned}$$

We used integration by substitution as $t = d_\phi(x,\theta)$ and (24). Therefore, if integral (25) exists, then (16) holds, i.e., the unbiased estimating equation holds without any bias correction term. Conversely, the above discussion also shows that $\mathbb{E}_{p(x|\theta)}\left[|f'(d_\phi(X,\theta))(X-\theta)|\right] \propto 2\int_0^\zeta g(t)f'(t)dt$. This means that the condition (25) is also necessary. □

Note that Assumption 3 must be satisfied for the integration by substitution to imply the unbiased estimating equation.

The elliptical and IS distributions are rare examples with unbiased estimating equations for the corresponding f-separable Bregman distortion measures and include the corresponding regular exponential family models.

Remark 1: In this section, we derived the condition under which the estimating equation holds without the bias correction term. Generally, the bias correction term does not vanish. We showed rare examples, Mahalanobis, IS, and one-dimensional Bregman divergences, for which the bias correction term vanishes. We emphasize that conditions (18), (22), and (25) of the theorems are easy to check. For example, when the statistical model is the gamma distribution, i.e., the function g is $\exp(-kz)$, we immediately see that the condition is satisfied with respect to the function f' which is a polynomial. Thus, the conditions of the theorem can narrow the range within which the function f or f' can be chosen.

Remark 2: This paper focuses on unbiasedness of the estimating equations, which is the necessary condition for the consistency of estimators. On the other hand, even when the unbiasedness of estimating equations does not hold, the generalization performance may be good due to the trade-off between bias and variance. For example, the Lq-likelihood estimator is the case where the bias correction term is truncated from the β-divergence [45]. However, under small samples, the exchange of bias and variance has been shown to improve

the generalization performance. For f-separable Bregman distortion measures, the generalization performance when the estimating equations are not unbiased is unknown and is a subject for future work.

V. Details of Statistical Models

In Section IV, we clarified conditions that consist of the Bregman divergence, statistical model, and the class of function f for the unbiased estimating equation to hold without the bias correction term. We have newly defined the IS distribution and its generalized continuous Bregman distribution. In this section, we describe the properties of these distributions. We also discuss the relationship between the continuous Bregman distribution and regular exponential family.

A. IS Distribution

1) Relationship Between Function g, Expected Value $\mathbb{E}[X]$, and Normalization Constant C_{IS}:

Lemma 1: Under Assumption 2, the following relation holds with respect to the expected value $\mathbb{E}[X]$ and the function g which composes the IS distribution,

$$\int_0^\infty g(t)dt < \infty \iff \mathbb{E}_{p(x|\theta)}[X] = \theta < \infty.$$

Proof: This relation immediately holds from Theorem 2 by substituting $f'(z) = 1$. □

Lemma 2: Under Assumption 2, the following relation holds with respect to the expected value $\mathbb{E}[X]$ and the normalization constant C_{IS} of the IS distribution,

$$\mathbb{E}_{p(x|\theta)}[X] = \theta < \infty \iff C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x} g(d_{\mathrm{IS}}(x,1))dx = \int_0^\infty g(d_{\mathrm{IS}}(x,1))dx < \infty. \tag{26}$$

Proof: We assume that the expected value $\mathbb{E}[X]$ exists and is θ. From the definition and finiteness of expected value $\mathbb{E}[X]$, we have

$$\infty > \theta = \mathbb{E}_{p(x|\theta)}[X] = \int_0^\infty \frac{1}{C_{\mathrm{IS}}}\frac{1}{x} g(d_{\mathrm{IS}}(x,\theta))x dx = \frac{1}{C_{\mathrm{IS}}}\int_0^\infty g(d_{\mathrm{IS}}(x,\theta))dx = \theta\frac{1}{C_{\mathrm{IS}}}\int_0^\infty g(d_{\mathrm{IS}}(x,1))dx.$$

Here, the normalization constant C_{IS} must satisfy

$$C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x} g(d_{\mathrm{IS}}(x,1))dx = \int_0^\infty g(d_{\mathrm{IS}}(x,1))dx < \infty. \tag{27}$$

Therefore, we have

$$\mathbb{E}_{p(x|\theta)}[X] = \theta < \infty \Rightarrow (27).$$

Conversely, when we assume (27), we have

$$(27) \Rightarrow \mathbb{E}_{p(x|\theta)}[X] = \theta < \infty.$$

Therefore, (26) holds. □

Theorem 4: Under Assumption 2, the following relation holds with respect to the expected value $\mathbb{E}[X]$, the normalization constant C_{IS}, and the function g which composes the IS distribution,

$$\int_0^\infty g(t)dt < \infty \iff \mathbb{E}_{p(x|\theta)}[X] = \theta < \infty \iff C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x} g(d_{\mathrm{IS}}(x,1))dx = \int_0^\infty g(d_{\mathrm{IS}}(x,1))dx < \infty.$$

Proof: Theorem 4 immediately holds from Lemma 1 and Lemma 2. □

Lemma 1 shows that when the function g satisfy $\int_0^\infty g(t)dt < \infty$, meaning that $g \in L^1(\mathbb{R}_+)$, if the IS distribution exists, its expected value $\mathbb{E}[X]$ is θ. In other words, the existence of the expected value depends only on the function g. This property holds in the general continuous Bregman distribution described in Section V-B (Corollary 1).

Lemma 2 means the normalization constant C_{IS} is expressed in another form than (21). Then, it holds that

$$C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x} g(d_{\mathrm{IS}}(x,1))dx = \int_0^\infty g(d_{\mathrm{IS}}(x,1))dx.$$

The integrand of the normalization constant (27) does not have factor $\frac{1}{x}$ which diverges infinity when x goes to 0. This fact has an advantage in calculating the normalization constant numerically. However, the relation (26) between the normalization constant and expected value does not hold on the continuous Bregman distribution. It is a property of the scale family as discussed in Appendix B (Theorem 7).

Remark 3: Lemma 1 is obtained from the property of the continuous Bregman distribution. Similarly, Lemma 2 is obtained from the property of the scale family. The IS distribution belongs to the continuous Bregman distribution and scale family. Therefore, Theorem 4 is obtained.

2) Examples of the IS Distribution:

a) Gamma distribution: When we choose the function $g(z) = \exp(-kz)$, the IS distribution becomes the gamma distribution with the known shape parameter $k > 0$. Then, $\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))$ is expressed as follows:

$$\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta)) = \frac{1}{x}\exp\left(-k d_{\mathrm{IS}}(x,\theta)\right) = \left(\frac{e}{\theta}\right)^k x^{k-1}\exp\left(-\frac{k}{\theta}x\right).$$

The normalization constant C_{IS} is given by

$$C_{\mathrm{IS}} = \int_0^\infty \frac{1}{x}\exp\left(-k d_{\mathrm{IS}}(x,\theta)\right)dx = \left(\frac{e}{k}\right)^k \Gamma(k).$$

Therefore, the gamma distribution is obtained

$$p(x|\theta) = \frac{1}{C_{\mathrm{IS}}}\frac{1}{x}\exp(-k d_{\mathrm{IS}}(x,\theta)) = \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)}x^{k-1}\exp\left(-\frac{k}{\theta}x\right). \tag{28}$$

The gamma distribution is also expressed as

$$p(x|\beta,k) = \frac{x^{k-1}}{\Gamma(k)\beta^k}\exp\left(-\frac{x}{\beta}\right).$$

The parameters β and k are called scale and shape parameters, respectively. This model corresponds to (28) by the transformation with respect to parameter $\theta = k\beta$ or the change of random variable $Y = kX$. It is worth nothing that the parameter θ is also the scale parameter and expected value.

b) Mixture distribution: Let $g_1(z) : \mathbb{R}_+ \to \mathbb{R}_+$ and $g_2(z) : \mathbb{R}_+ \to \mathbb{R}_+$ be nonnegative functions. We define the function g by a convex combination, i.e., $g(z) = bg_1(z) + (1-b)g_2(z)$, where the coefficient satisfies $0 \le b \le 1$. Then, the IS distribution with respect to the function g is given by the convex combination of IS distributions with the same parameter θ:

$$p(x|\theta) = w\frac{1}{C_1}\frac{1}{x}g_1\left(d_{\mathrm{IS}}(x,\theta)\right) + (1-w)\frac{1}{C_2}\frac{1}{x}g_2\left(d_{\mathrm{IS}}(x,\theta)\right),$$
$$w = \frac{bC_1}{bC_1 + (1-b)C_2},$$

where $C_1 = \int_0^\infty \frac{1}{x}g_1(d_{\mathrm{IS}}(x,1))dx$ and $C_2 = \int_0^\infty \frac{1}{x}g_2(d_{\mathrm{IS}}(x,1))dx$ are the normalization constants of the component distributions, respectively, and w is the mixing coefficient. Similarly, if g is a convex combination of three or more functions, the IS mixture is generated. If $g_1(z)$ and $g_2(z)$ are given by $\exp(-k_1 z)$ and $\exp(-k_2 z)$ respectively, the IS mixture reduces to the gamma mixture with the same parameter θ, where k_1 and k_2 are positive.

B. Continuous Bregman Distribution

1) Relationship Between Function g and Expected Value $\mathbb{E}[X]$:

Corollary 1: Under (24) of Assumption 3 and Assumption 4, the following relation holds with respect to the expected value $\mathbb{E}[X]$ and the function g which composes the continuous Bregman distribution,

$$\int_0^\zeta g(t)dt < \infty \iff \mathbb{E}_{p(x|\theta)}[X] = \theta < \infty.$$

Proof: Corollary 1 follows immediately from Theorem 3 as $f'(z) = 1$. □

2) Examples of the Continuous Bregman Distribution: We show common examples of the continuous Bregman distribution. Note that the normalization constants of the following examples and those of the corresponding continuous Bregman distributions are different.

a) One-dimensional elliptical distribution: We set $\phi(x) = x^2$. Then, (23) becomes the one-dimensional elliptical distribution as follows:

$$p(x|\theta) = \frac{1}{C_{\mathrm{Mah.}}}g((x-\theta)^2).$$

b) IS distribution: We set $\phi(x) = -\log x$. Then, (23) becomes the IS distribution as follows:

$$p(x|\theta) = \frac{1}{C_{\mathrm{IS}}}\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta)).$$

We explained this distribution in Section V-A.

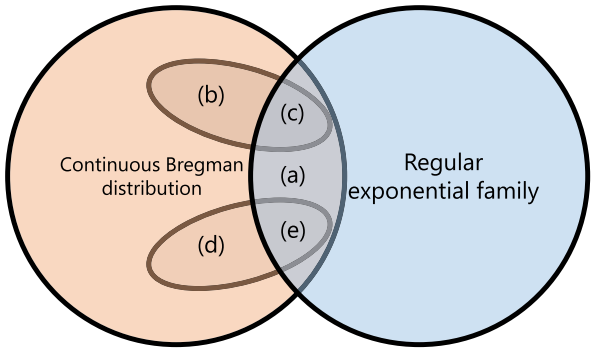


Fig. 1. Comparison between the continuous Bregman distribution and regular exponential family. (a) The intersection between the continuous Bregman distribution and regular exponential family. (b) One-dimensional elliptical distribution; (c) Gaussian distribution; (d) IS distribution; (e) gamma distribution.

c) Mixture distribution: As in the case of the IS mixture, we define the function g by a convex combination, i.e., $g(z) = bg_1(z)+(1-b)g_2(z)$. Then, we obtain the following continuous Bregman mixture:

$$p(x|\theta) = w\frac{1}{C_{\phi,1}(\theta)}\frac{\phi'(x)-\phi'(\theta)}{x-\theta}g_1(d_\phi(x,\theta)) + (1-w)\frac{1}{C_{\phi,2}(\theta)}\frac{\phi'(x)-\phi'(\theta)}{x-\theta}g_2(d_\phi(x,\theta)),$$
$$w = \frac{bC_{\phi,1}(\theta)}{bC_{\phi,1}(\theta) + (1-b)C_{\phi,2}(\theta)},$$

where all component distributions have the same parameter θ and depend on the same strictly convex function ϕ. Thus, the supports of the component distributions are all same. Here, $C_{\phi,1}(\theta)$ and $C_{\phi,2}(\theta)$ are the normalization constants of the component distributions. Similarly, for a convex combination of three or more functions g, the continuous Bregman mixture can be generated.

C. Relation to Regular Exponential Family

We consider the relationship between the continuous Bregman distribution (23) and the regular exponential family (10).

Assumption 5:

1) Let $g(z) = \exp(-kz)$ with $k > 0$.
2) For all x, the factor
$$\frac{1}{C_\phi(\theta)}\frac{\phi'(x)-\phi'(\theta)}{x-\theta}$$
does not depend on the parameter θ.

Proposition 1: Under Assumption 5, the continuous Bregman distribution becomes the regular exponential family as follows:

$$p(x|\theta) = \frac{1}{C_\phi(\theta)}\frac{\phi'(x)-\phi'(\theta)}{x-\theta}\exp(-d_\phi(x,\theta)) = r_\phi(x)\exp(-d_\phi(x,\theta)),$$

where $r_\phi(x)$ is uniquely determined by the strictly convex function ϕ [42], i.e.,

$$r_\phi(x) = \frac{1}{C_\phi(\theta)}\frac{\phi'(x)-\phi'(\theta)}{x-\theta}.$$

Figure 1 shows the relationship between the continuous Bregman distribution and regular exponential family. The intersection between the continuous Bregman distribution and the regular exponential family include Gaussian and gamma distributions. The condition 2 of Assumption 5 is a tight condition. Even if the condition 1 of Assumption 5 holds and the continuous Bregman distribution corresponding to the strictly convex function ϕ exists, it is not necessarily the regular exponential family.

VI. Latent Bias

In this section, we discuss the possibility of the latent bias minimization for many outliers, which is induced by the vanishing bias correction term. The estimating equation of the f-separable Bregman distortion measures reduces to the normalized estimating equation. From the viewpoint of the normalized estimating equation, the condition of latent bias minimization was shown as a theorem [16]. We can apply this theorem. From the viewpoint of the objective function, we see that the definitions of outliers are different for f-separable distortion measures and γ-divergence. As a by-product, we obtain a solution to a drawback of γ-divergence in the case of the exponential model. In what follows, we assume that the function f is twice differentiable.

A. Contaminated Distribution

We assume that the data-generating distribution is given as follows:

$$\tilde{p}(\boldsymbol{x}) = (1-\varepsilon)p(\boldsymbol{x}|\boldsymbol{\theta}^*) + \varepsilon c(\boldsymbol{x}), \tag{29}$$

where $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$ is the target distribution, $c(\boldsymbol{x})$ is the contamination distribution that generates outliers, and ε is the proportion of outliers. Suppose the parameter $\hat{\boldsymbol{\theta}}$ estimated from the data generated from this distribution is expressed asymptotically as $\tilde{\boldsymbol{\theta}}$, i.e., $\hat{\boldsymbol{\theta}} \xrightarrow{P} \tilde{\boldsymbol{\theta}}$. Here, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ is called the latent bias, which expresses the bias caused by the contamination distribution [16].

B. Definition of Outliers: γ-Divergence

In the estimation based on γ-divergence, it is assumed that the following quantity can be made arbitrarily small for an appropriately large $\gamma_0 > 0$ as an assumption regarding outliers,

$$\nu_p = \left[\mathbb{E}_{c(\boldsymbol{x})}\left[p(\boldsymbol{X}|\boldsymbol{\theta}^*)^{\gamma_0}\right]\right]^{\frac{1}{\gamma_0}}. \tag{30}$$

This assumption means that outliers are distributed over the region where the likelihood is small in the target distribution $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$. Since nothing about the outlier proportion is assumed, it is also possible to deal with the case where the proportion of outlier is large.

However, Kuchibhotla et al. [27] reported the following two disadvantages. First, the γ-divergence is adversely affected by data at the edge of the support of the target model, like location–scale family. For example, in estimating of the scale parameter of the exponential distribution, a wrong global solution is generated when a very small inlier around $x = 0$, such as $x = 10^{-4}$, is mixed [23]. Here, an inlier means a data point near zero [1, p. 140]. Secondly, the estimator, which can achieve the latent bias minimization, is a local solution. Nevertheless, the solution selection criteria have yet to be established. A solution to these problems has been invented by Kuchibhotla et al. [27]; however, they are not fully resolved.

C. Definition of Outliers: f-Separable Bregman Distortion Measures

In the estimation based on f-separable Bregman distortion measures, we consider that the following quantity can be made arbitrarily small for an appropriate function f as an assumption regarding outliers,

$$\nu_{d_\phi} = \mathbb{E}_{c(\boldsymbol{x})}\left[f\left(d_\phi\left(\boldsymbol{X}, \boldsymbol{\theta}^*\right)\right)\right]. \tag{31}$$

In other words, if the function f is parameterized by a parameter α, (31) can be arbitrarily reduced for an appropriately large parameter α. This assumption corresponds to the assumption (30) of γ-divergence. It means that when the random variable follows the contamination distribution, i.e., $\boldsymbol{X} \sim c(\boldsymbol{x})$, an outlier is in the region where $d_\phi(\boldsymbol{X}, \boldsymbol{\theta}^*) \to \infty$ is satisfied. When estimating the location parameter of the elliptical distribution using Mahalanobis distance, the definition of outlier is the same as (30), i.e., $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \to \infty$ is regarded as the outlier. However, when estimating the scale parameter of the IS distribution using the IS distance, the definition of outlier is not the same as (30). In this case, from

$$\lim_{x\to 0} d_{\mathrm{IS}}(x,\theta) = \lim_{x\to\infty} d_{\mathrm{IS}}(x,\theta) = \infty,$$

the data near 0 or ∞ are regarded as outliers. In other words, the estimator based on f-separable IS distortion measures is robust against large outliers and very small inliers to which γ-divergence is vulnerable.

D. Necessary Condition for Latent Bias Minimization

We express the estimating equation with the data-generating distribution $\tilde{p}(\boldsymbol{x})$ as follows:

$$\psi_{\tilde{p}}(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{p}(\boldsymbol{x})}\left[f'\left(d_\phi(\boldsymbol{X}, \boldsymbol{\theta})\right)(\boldsymbol{\theta} - \boldsymbol{X})\right] = \boldsymbol{0}.$$

The solution to this estimating equation is given by $\tilde{\boldsymbol{\theta}}$.

Assumption 6: We denote the smallest eigenvalue of

$$\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta}^*)}\left[f''\left(d_\phi(\boldsymbol{X}, \boldsymbol{\theta}^*)\right)(\boldsymbol{X} - \boldsymbol{\theta}^*)(\boldsymbol{X} - \boldsymbol{\theta}^*)^{\mathrm{T}}\right] \boldsymbol{\nabla}\boldsymbol{\nabla}\phi(\boldsymbol{\theta}^*)$$

by $\lambda_{\min}$. We assume the following,

$$\lambda_{\min} > -\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta}^*)}\left[f'\left(d_\phi(\boldsymbol{X}, \boldsymbol{\theta}^*)\right)\right].$$

Theorem 5: Assume Assumption 6 holds and that for any sufficiently small $\eta > 0$,

$$\left\|\psi_c\left(\psi^{-1}_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{\tau})\right)\right\| < \eta\frac{1-\varepsilon}{\varepsilon} \tag{32}$$

for $\|\boldsymbol{\tau}\| < \eta$. Then, there exists a solution $\tilde{\boldsymbol{\theta}}$ of $\psi_{\tilde{p}}(\boldsymbol{\theta}) = \boldsymbol{0}$ such that $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| < \eta$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ for $\tilde{p}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}^*)$.

Proof: Under Assumption 6, the following matrix is positive definite,

$$\left.\frac{\partial \psi_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta}^*)}\left[\left.\frac{\partial f'\left(d_\phi(\boldsymbol{X}, \boldsymbol{\theta})\right)(\boldsymbol{\theta} - \boldsymbol{X})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right],$$

because

$$\left.\frac{\partial \psi_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta}^*)}\left[f''(d_\phi(\boldsymbol{X},\boldsymbol{\theta}^*))(\boldsymbol{X}-\boldsymbol{\theta}^*)(\boldsymbol{X}-\boldsymbol{\theta}^*)^{\mathrm{T}}\right]\boldsymbol{\nabla}\boldsymbol{\nabla}\phi(\boldsymbol{\theta}^*) + \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta}^*)}\left[f'\left(d_\phi(\boldsymbol{X},\boldsymbol{\theta}^*)\right)\right]\boldsymbol{E},$$

where $\boldsymbol{E}$ is the identity matrix. The positive definiteness of this matrix and the condition (32) ensure that all the assumptions of [16, Theorem 3.1] are satisfied. Hence, the assertions of the theorem directly follow from those of [16, Theorem 3.1]. □

These assertions of the theorem mean that the latent bias can be arbitrarily small and that Fisher consistency holds. The positive definiteness of $\partial\psi_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ in the proof is necessary for the implicit function theorem. It is difficult to actually confirm the condition (32) of Theorem 5. Therefore, we assume condition (3.1) in the literature [16] corresponds to the condition (32) of Theorem 5. In our case, condition (3.1) [16] reduces to the following:

$$\|\psi_c(\boldsymbol{\theta}^*)\| = \left\|\mathbb{E}_{c(\boldsymbol{x})}\left[f'\left(d_\phi(\boldsymbol{X},\boldsymbol{\theta}^*)\right)(\boldsymbol{X}-\boldsymbol{\theta}^*)\right]\right\| \tag{33}$$

can be made arbitrarily small. The fact that $\|\psi_c(\boldsymbol{\theta}^*)\|$ can be made arbitrarily small implies that $\|\psi_c(\boldsymbol{\theta})\|$ can also be made arbitrarily small in the vicinity of $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ by the continuity. Since $\psi^{-1}_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{\tau})$ is contained in the vicinity of $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ for a sufficiently small $\eta>0$, (32) is implied by the fact that $\|\psi_c(\boldsymbol{\theta}^*)\|$ can be made arbitrarily small. If the quantity (31) can be made arbitrarily small under an appropriate function f, then (33) can also be made arbitrarily small. However, the converse is not generally true.

If the contamination distribution $c(\boldsymbol{x})$ has the point mass at $\|\boldsymbol{x}\|\to\infty$, (33) can be rewritten as

$$\lim_{\|\boldsymbol{x}\|\to\infty}\left\|f'\left(d_\phi(\boldsymbol{x},\boldsymbol{\theta}^*)\right)(\boldsymbol{x}-\boldsymbol{\theta}^*)\right\|, \tag{34}$$

then the following condition is required for the limit (34) to be arbitrarily small:

$$\lim_{z\to\infty} f'(z) = 0,$$

which is also a necessary condition for the influence function to be bounded [39]. When (34) equals to 0, the influence function has a desirable property called the redescending property. In other words, when the redescending property holds, the influence of the sufficiently large outliers is ignored. For functions (8) and (9), sufficient conditions for the redescending property were investigated [39].

Remark 4: The estimator that can achieve the latent bias minimization is one of the local solutions of (4). Thus, its solution selection problem occurs. In other words, this problem is the initial value selection of the iterative update rule (7).

E. Strategy for Initial Value Selection

The estimator that can achieve the latent bias minimization is a local minimum solution given by a fixed point of the iterative algorithm. Thus, we need the strategy for initial value selection. We have already assumed that (33) is sufficiently small. It means that the contamination distribution is far from the target distribution. Furthermore, we consider that the proportion of the target distribution $1-\varepsilon$ is larger than the proportion of the contamination distribution ε. In other words, we assume $\varepsilon<0.5$. We apply the K-means clustering with two clusters to the dataset and roughly separate it into the data generated from the target and contamination distributions. The initial values of the cluster centers are set at the minimum and maximum values of the data. If the target and contamination distributions are ideally separated, it can be expected that the initial value near the true value is obtained.

VII. Asymptotic Property

The estimation based on f-separable Bregman distortion measures, which satisfies the unbiasedness of the estimating equation, can be interpreted as an M-estimation (1), where

$$\psi(\boldsymbol{x},\boldsymbol{\theta}) = f'(d_\phi(\boldsymbol{x},\boldsymbol{\theta}))(\boldsymbol{x}-\boldsymbol{\theta}).$$

Therefore, under appropriate assumptions, the following consistency and asymptotic normality of the estimator follow from the asymptotic theory of M-estimation [1], [2], [46],

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^*,$$
$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) \xrightarrow{d} N(\boldsymbol{0},\Sigma(\boldsymbol{\theta}^*)),$$

where $\Sigma(\boldsymbol{\theta}^*) = \boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\boldsymbol{I}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)$,

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\left[f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)\right]^2(\boldsymbol{X}-\boldsymbol{\theta})(\boldsymbol{X}-\boldsymbol{\theta})^{\mathrm{T}}\right],$$
$$\boldsymbol{J}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\frac{\partial f'\left(d_\phi\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right)(\boldsymbol{\theta}-\boldsymbol{X})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right].$$

If the data are generated from the distribution (29), the asymptotic variance is given by $\Sigma(\boldsymbol{\theta}^*)/(1-\varepsilon)$ [16, Theorem 4.2].

A. Gamma Distribution

We assume that the statistical model is the gamma distribution $p(x|\theta) = \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)}x^{k-1}\exp\left(-\frac{k}{\theta}x\right)$, the function f is (8), and Bregman divergence is the IS distance (19). Then, the asymptotic variance of the estimator is given by

$$V[\hat{\theta}] = \Sigma(\theta^*) = \frac{\Gamma(2\alpha+k)\Gamma(k)}{[\Gamma(\alpha+k)]^2}\frac{(\alpha+k)^{2(\alpha+1+k)}}{(2\alpha+k)^{2\alpha+1+k}}\frac{1}{k^{2+k}}\theta^{*2},$$

the tuning parameter satisfies $\alpha>-0.5k$. For the exponential distribution ($k=1$), we can compare the asymptotic relative efficiencies (AREs) of the estimators based on minimizing the f-separable IS distortion measures and β- and γ-divergences. The ARE is given by $\frac{V[\hat{\theta}_{\mathrm{MLE}}]}{V[\hat{\theta}]}$, where $V[\hat{\theta}_{\mathrm{MLE}}]$ is the asymptotic variance of the maximum likelihood estimator ($\alpha=0$). The asymptotic variances of the estimators based on β- and γ-divergences were derived for the exponential distribution [15], [23]. Figure 2 shows their AREs, when the tuning parameter $\alpha=\beta=\gamma$. We observe that the range of tuning parameter $\alpha=\beta=\gamma>0$ induces the robustness against outliers. As shown in Figure 2, for the function (8) and IS distance, the ARE is generally greater than that of β-divergence in the range of tuning parameter $\alpha<2$. The ARE is also greater than that of γ-divergence in the entire range of the tuning parameter. However, in general, the ARE and robustness have

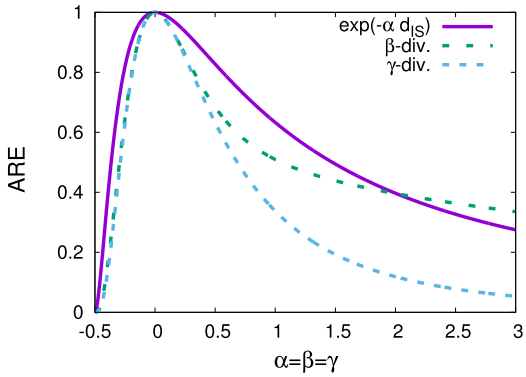


Fig. 2. Comparison of ARE under the exponential model ($k = 1$).

trade-off relationship. Therefore, it is essential to choose the tuning parameter appropriately, taking into account both of them.

We show the behavior of the estimator with respect to the number of data with numerical examples where the results were averaged over 10,000 trials. The number of data is given as $n = 30$, 50 and from 100 to 1000 in steps of 100. Figure 3 shows the bias and mean squared error (MSE) (log–log plot) of the estimator based on the f-separable IS distortion measure when the data are given in the exponential distribution ($\theta = 1$). In addition, Figure 4 shows the bias and MSE (log–log plot) of the estimator when the distribution is contaminated by a Gaussian distribution with a proportion of contamination $\varepsilon = 0.4$. Figure 3 shows that both bias and MSE converge to zero with the same order of convergence regardless of the value of the tuning parameter. However, looking at the values of bias and MSE with a fixed number of data, it can be seen that they reach a minimum at $\alpha = 0$ (MLE) and increase monotonically as α increases. Figure 4 shows that bias and MSE both converge to larger values for $\alpha = 0.25$ than for $\alpha = 0$, and converge to 0 for $\alpha = 0.5$ and above when the proportion of contamination is large. The phenomenon of bias and MSE taking larger values for small values of the tuning parameters than for $\alpha = 0$ is discussed in Section VIII-B1 with a fixed number of data. When the proportion of contamination is small or the contamination distribution is farther from the target distribution, the values of bias and MSE converge to zero even when the tuning parameters are small. This shows that if the target distribution contains a distribution of data, the estimator asymptotically approaches the true value of the target distribution, regardless of whether the data are contaminated or not.

VIII. Numerical Experiments

This section discusses the results of experiments conducted to demonstrate the latent bias minimization by f-separable Bregman distortion measures under heavy contamination. Generally, in the location parameter estimation, the bias correction term vanishes, and the estimating equation is normalized. In any case, the latent bias can be minimized under heavy contamination. However, in scale parameter estimation, the latent bias minimization is difficult under heavy contamination. Thus, we focus on scale parameter estimation using f-separable IS distortion measures.

A. Setup

We use the function (8). When the target is the exponential distribution and the Bregman divergence is the IS distance, Assumption 6 holds for $\alpha \geq 0$ in (8). We consider condition (32) of Theorem 5. If the contamination distribution is sufficiently far away from the target distribution, the integrand in (33) approaches 0 exponentially. Therefore, condition (32) of Theorem 5 holds for sufficiently large α.

Competitors are the estimation methods based on β- and γ-divergences, which include tuning parameters β and γ, respectively. These divergences have weight functions ξ that are power functions. Estimation based on the β-divergence is expected to have a non-zero latent bias because the β-divergence corresponds to the non-normalized estimating equation. If $\alpha = \beta = \gamma = 0$, the estimation methods reduce to the exact MLE under the assumed model. When the tuning parameters are significantly large, the estimation methods are robust against outliers. For each estimation method, the iterative method is used by giving the initial value of parameter θ. The fixed point of the iterative method is treated as an estimator of θ. In the cases of β- and γ-divergences, we set the true value to the initial value of parameter θ to investigate the behavior of the solution near the true value. For the estimation based on f-separable distortion measures, we obtain the initial value from the method of Section VI-E. Estimation based on β- and γ-divergences is advantageous because the initial value is not always close to the true value in f-separable Bregman distortion measure-based estimation. Parameter θ is estimated from 100 data samples. The proportion of outliers is $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$. The reported results are averaged over 100 trials. We considered the following situations.

1) Exponential Distribution With Outlier Contamination: In this experiment, we investigate the behavior of the latent bias under significant outlier contamination. The data-generating distribution is the following:

$$(1-\varepsilon)\mathrm{Exp}(\theta^* = 1) + \varepsilon N(\mu_{\mathrm{out}}, \sigma^2_{\mathrm{out}} = 1),$$

where the location parameter of the contamination distribution is $\mu_{\mathrm{out}} \in \{10, 20, 30\}$.

2) Exponential Distribution With Inlier Contamination: The aim of this experiment is to investigate the behavior under the small inlier contamination, for which it was reported the estimation based on minimizing γ-divergence generates a spurious global solution [23], [27]. The data-generating distribution is the following:

$$(1-\varepsilon)\mathrm{Exp}(\theta^* = 1) + \varepsilon\delta(x - 10^{-4}),$$

where δ is the Dirac delta function.

3) Gamma Distribution With Outlier Contamination: We investigate the behavior of the latent bias in the gamma distribution when the shape parameter k is greater than and less than one when outliers are mixed. The data-generating distribution is the following:

$$(1-\varepsilon)\mathrm{Gam}(\theta^* = 1|k) + \varepsilon N(\mu_{\mathrm{out}} = 20, \sigma^2_{\mathrm{out}} = 1),$$

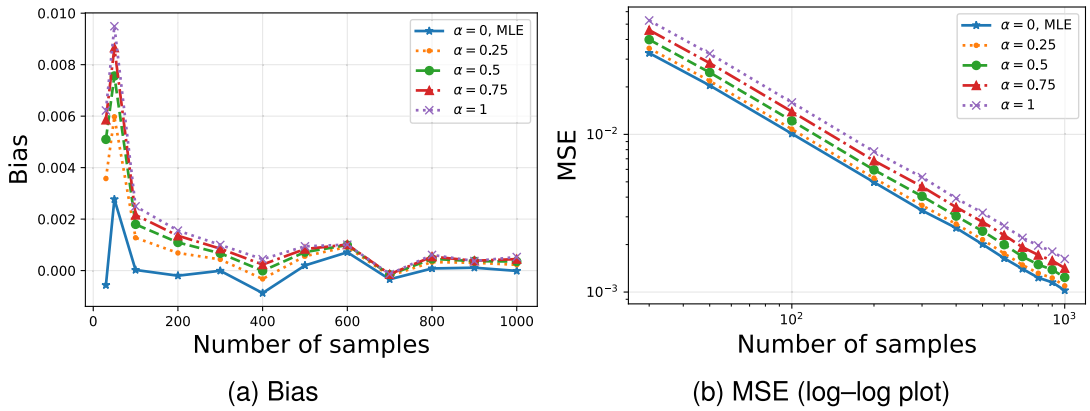


Fig. 3. Bias and MSE of the non-contaminated exponential distribution against the number of samples.

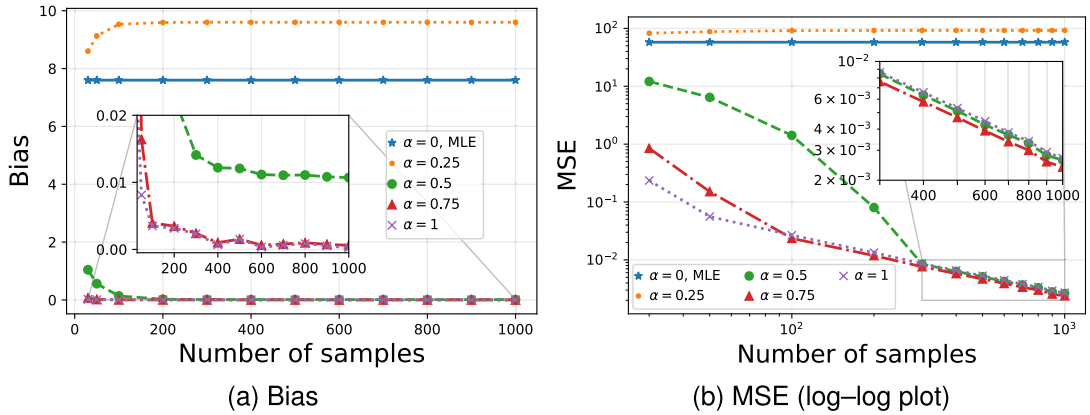


Fig. 4. Bias and MSE of the exponential distribution contaminated with outliers ($\varepsilon = 0.4$) against the number of samples. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 20, \sigma^2_{\text{out}} = 1)$.

where the shape parameter is $k \in \{0.5, 2\}$.

4) Gaussian Distribution Estimation Under Outlier Contamination: In this experiment, we apply the estimation based on minimizing the f-separable IS distortion measures to the variance estimation problem of the Gaussian distribution. However, we need the location parameter to estimate the variance.

Therefore, we estimate the location parameter μ by minimizing the f-separable squared distortion measures as the function f is (8). In other words, we estimate the location and variance parameters simultaneously. The update rules of location and variance are respectively given as follows:

$$\mu = \frac{\sum_{i=1}^n \exp\left(-\alpha \frac{(x_i-\mu)^2}{2\sigma^2}\right) x_i}{\sum_{j=1}^n \exp\left(-\alpha \frac{(x_j-\mu)^2}{2\sigma^2}\right)}, \tag{35}$$

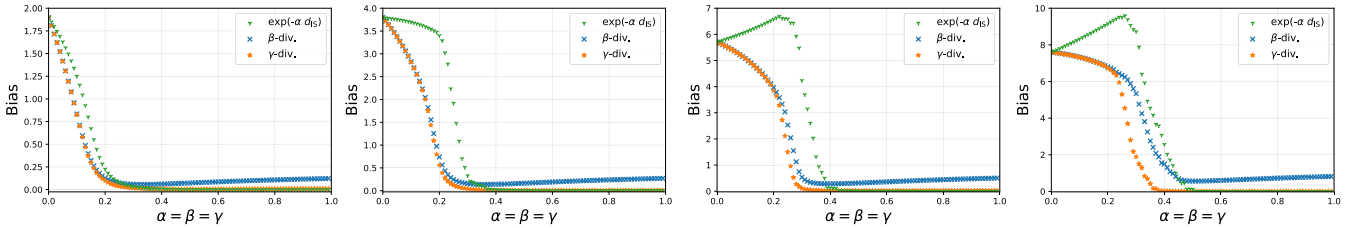
$$\sigma^2 = \frac{\sum_{i=1}^n \exp\left(-\frac{\alpha}{2} d_{\text{IS}}\left((x_i-\mu)^2, \sigma^2\right)\right)(x_i-\mu)^2}{\sum_{j=1}^n \exp\left(-\frac{\alpha}{2} d_{\text{IS}}\left((x_j-\mu)^2, \sigma^2\right)\right)}.$$

The update rule (35) of the location parameter is the same as those based on minimizing β- and γ-divergences. In other words, the result of the variance estimation causes the difference in estimation. The update rules and the objective function of the estimation based on minimizing β- and γ-divergences are given in Appendices C-A3 and C-B3. The data-generating distribution is given as follows:

$$(1-\varepsilon)N(\mu^* = 0, \sigma^{*2} = 1) + \varepsilon N(\mu_{\text{out}} = 5, \sigma^2_{\text{out}} = 1).$$

B. Results

1) Exponential Distribution With Outlier Contamination: First, we discuss the influence of the proportion of outliers when the location parameter of the contamination is 20 (Figure 5). For the f-separable IS distortion measure, the bias goes to zero regardless of the proportion of outliers. It is achieved when we set the tuning parameter α to a large value. However, when the proportion of outliers is greater than or equal to 0.3, the bias increases once and then approaches 0 as the tuning parameter α increases. To find out the cause of this, we investigated the shape of the objective function. When $\alpha = 0$, the estimation method is the MLE of the exponential distribution, and the solution of the objective function is unique. Since the objective function changes continuously as the parameter α increases, the solution of the objective function is unique when α is small. The unique solution moves to the direction of the target or contamination distributions as α increases. We observed that the moving direction of the unique solution depends on the proportion of outliers. For $\varepsilon = 0.1$ and 0.2, the unique solution moves to the direction of the target distribution. However, for $\varepsilon = 0.3$ and 0.4, the unique solution moves in the direction of the contamination distribution.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 5. Bias of the exponential distribution contaminated with outliers. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 20, \sigma^2_{\text{out}} = 1)$.

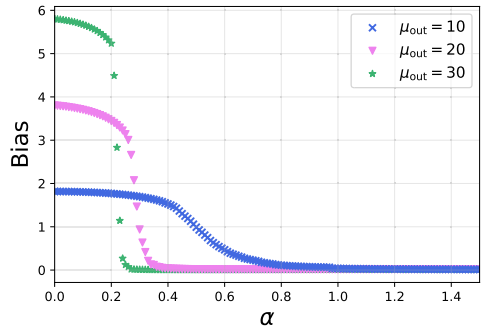


Fig. 6. Results for the different locations of the contamination distribution $N(\mu_{\text{out}}, \sigma^2 = 1)$ with $\varepsilon = 0.2$.

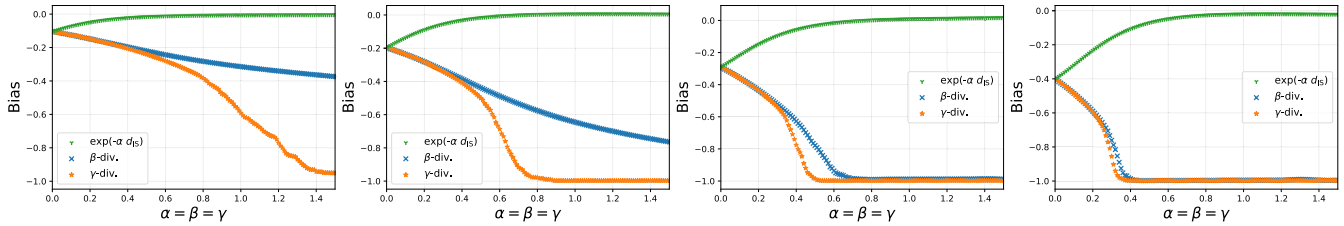
Consequently, the bias increased up to a certain value of α (Figures 5 (c) and (d)). However, a local solution is generated in the objective function as α increases. The selection of this generated local solution causes the phenomenon that the bias suddenly approaches 0, as illustrated in Figures 5 (b), (c), and (d).

Next, we discuss the influence of the location parameter of the contamination distribution. Figure 6 shows the results of the f-separable IS distortion measure when the location parameter μ of contamination distribution is 10, 20, and 30, and the proportion of outliers is 0.2. The farther the contamination distribution that generates the outliers is from the target distribution to be estimated, the smaller the minimum value α when the bias reaches zero. It is worth nothing that the curve in the figure changes continuously only when the location parameter of the contamination distribution is 10. This is because when the target and contamination distributions are close to each other, the unique solution for sufficiently small α moves toward the true parameter of the target distribution with respect to the increase in α.

2) Exponential Distribution With Inlier Contamination: Figure 7 shows the results of the inlier contamination experiment. Only the f-separable IS distortion measure has achieved bias going to zero, while neither β- nor γ-divergences has been achieved. This is because, in the objective function of β- or γ-divergence, the solution near the true value moves toward zero with respect to the increase in β or γ. Especially in the case of γ-divergence, there are suspicious solution near $\theta = 0$ [23], [27] and the solution near the true value. Additionally, when γ exceeds a certain value, the solution near the true value disappears. Therefore, above a certain value of γ, the estimator is given as a solution near $\theta = 0$, so the bias approaches -1. In other words, for β- or γ-divergence, the bias cannot approach zero no matter how the initial value of parameter θ is tuned.

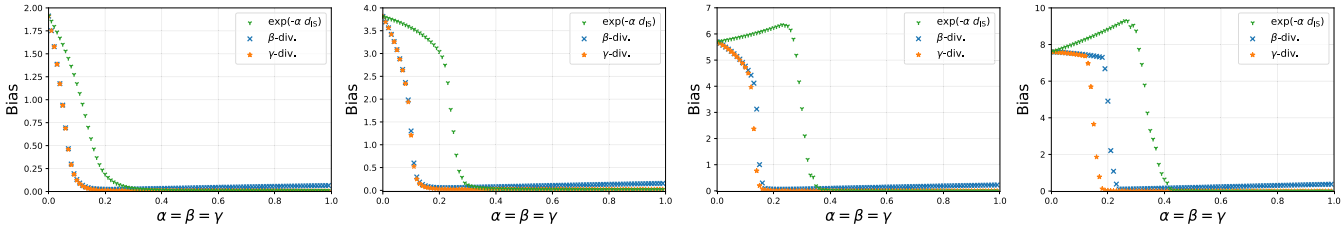
3) Gamma Distribution With Outlier Contamination: Figures 8 and 9 show the biases of the contaminated gamma distribution when the true shape parameter k is 2 and 0.5, respectively. The β-divergence-based-estimator was numerically unstable when using the iterative algorithm. Therefore, it was obtained through grid search. When the true shape parameter k is 2, the behavior of the bias is almost the same as in the exponential distribution (Figure 8). However, when the true shape parameter is 0.5, the behavior of the bias differs significantly from that of the exponential distribution (Figure 9). The β- and γ-divergences-based-estimators did not reduce the bias to zero; it increased as the proportion of outliers increased. The objective function and the estimating equations for the β- and γ-divergences with respect to the gamma distribution include the gamma function. Because of the constraint that the argument of the gamma function is positive, if the true shape parameter k is less than one, the tuning parameters β and γ that can be adjusted are constrained to be in the range $[0, k/(1-k))$. Here, the range of tuning parameters for β- and γ-divergences is $[0, 1)$. The case of f-separable IS distortion measure-based estimation, has achieved bias converging to zero. In estimating the contaminated gamma distribution based on the f-separable IS distortion measure, the bias can be reduced to zero independent of the known shape parameter.

4) Gaussian Distribution Estimation Under Outlier Contamination: Figures 10 and 11 show the biases of the mean and the variance of the contaminated Gaussian distribution estimation experiment, respectively. To estimate the mean, the bias achieved zero regardless of the proportion of outliers and estimation methods. However, as the proportion of outliers increases, the variance estimation results start to differ. Thus, the process of the bias of the mean estimation results going to zero starts to differ when the tuning parameter value increases. In particular, this difference becomes more prominent as the proportion of outliers increases. In the estimation based on the f-separable distortion measure and γ-divergence, the bias of the variance estimation can approach zero by increasing the tuning parameter regardless of the proportion of outliers. The tuning parameter value achieving bias near zero is smaller for γ-divergence than of the f-separable IS distortion measure. In the estimation based on β-divergence, the bias cannot approach 0, even when the proportion of outliers is 0.1, and it worsens as the proportion of outliers ratio increases.



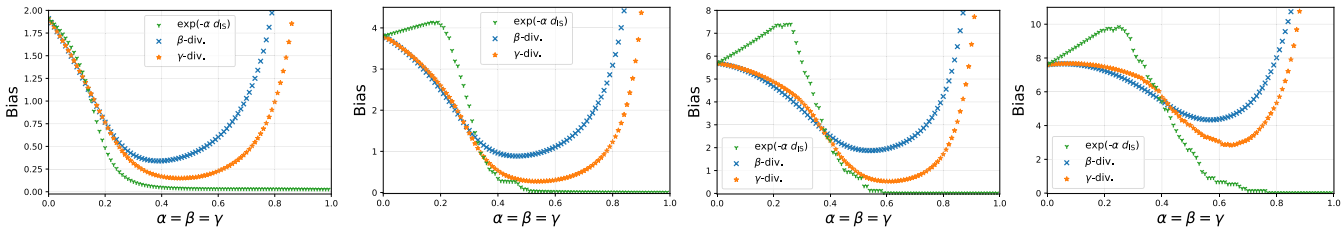
(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 7. Bias of exponential distribution contaminated with inliers. The contamination distribution is Dirac delta: $\delta(x - 10^{-4})$.



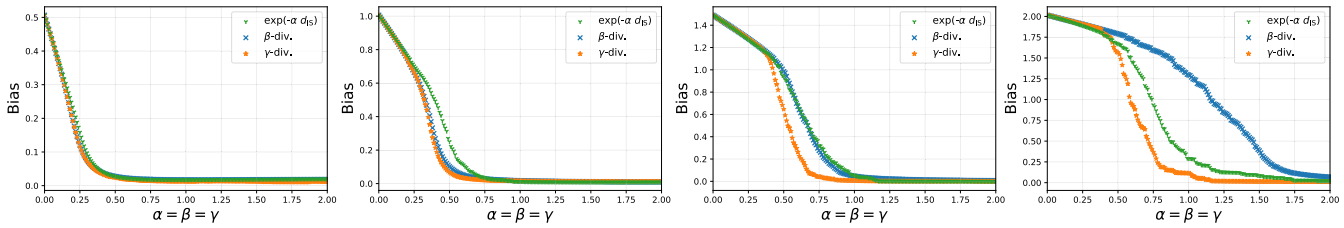
(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 8. Bias of the gamma distribution ($k = 2$) contaminated with outliers. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 20, \sigma^2_{\text{out}} = 1)$.



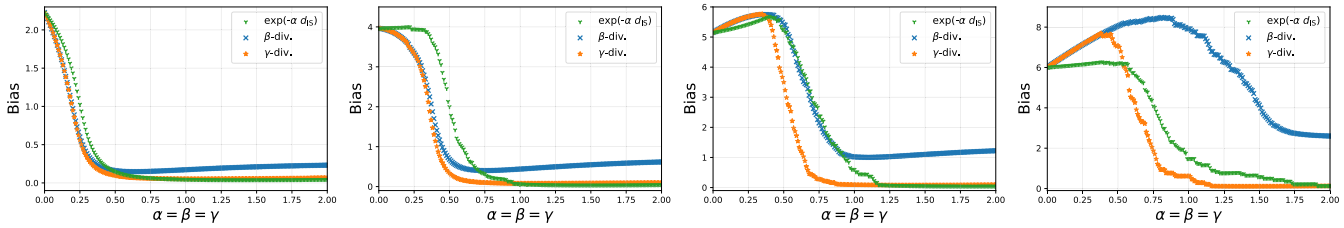
(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 9. Bias of the gamma distribution ($k = 0.5$) contaminated with outliers. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 20, \sigma^2_{\text{out}} = 1)$.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 10. Bias of the Gaussian mean parameter μ contaminated with outliers. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 5, \sigma^2_{\text{out}} = 1)$.

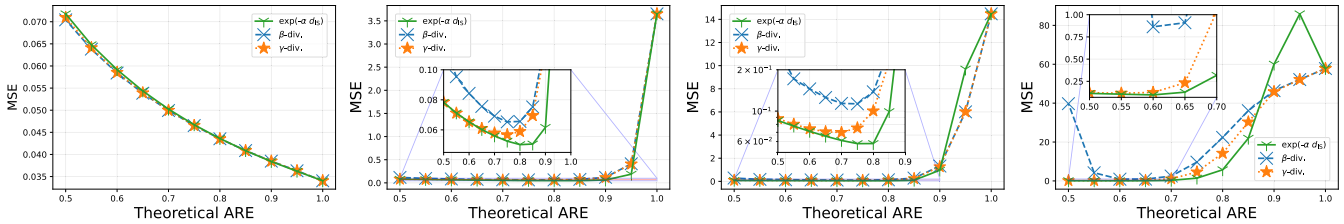


(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.3$ (d) $\varepsilon = 0.4$

Fig. 11. Bias of the Gaussian variance parameter σ^2 contaminated with outliers. The contamination distribution is the Gaussian distribution: $N(\mu_{\text{out}} = 5, \sigma^2_{\text{out}} = 1)$.

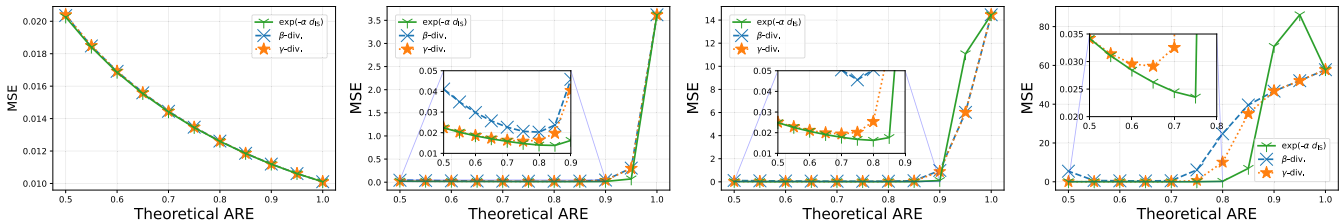
C. Trade-off Between Sample Efficiency and Robustness

We demonstrated that the bias can reach zero with the f-separable IS distortion measure-based estimation. However, the performance of the estimator should be measured under the trade-off between efficiency and robustness. We focus on the exponential distribution for which the AREs of the estimators are discussed in Section VII (Figure 2). The contamination distribution is the Gaussian distribution with $\mu_{\text{out}} = 20$ and



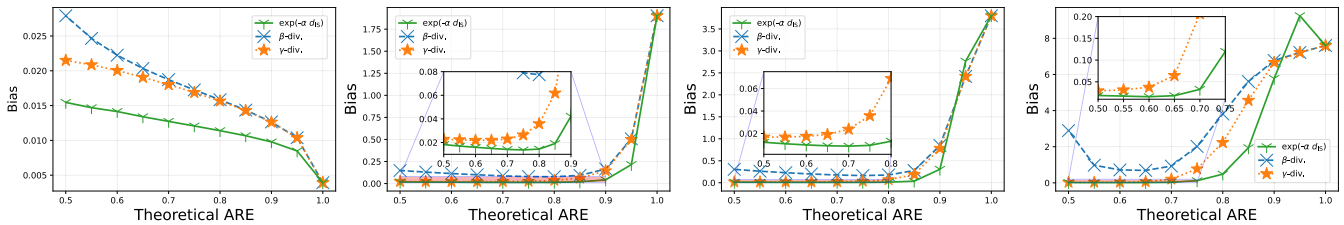
(a) $\varepsilon = 0$ (b) $\varepsilon = 0.1$ (c) $\varepsilon = 0.2$ (d) $\varepsilon = 0.4$

Fig. 12. MSE of the exponential distribution with the outliers under $n = 30$ samples.



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.1$ (c) $\varepsilon = 0.2$ (d) $\varepsilon = 0.4$

Fig. 13. MSE of the exponential distribution with the outliers under $n = 100$ samples.



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.1$ (c) $\varepsilon = 0.2$ (d) $\varepsilon = 0.4$

Fig. 14. Bias of the exponential distribution with the outliers under $n = 30$ samples.

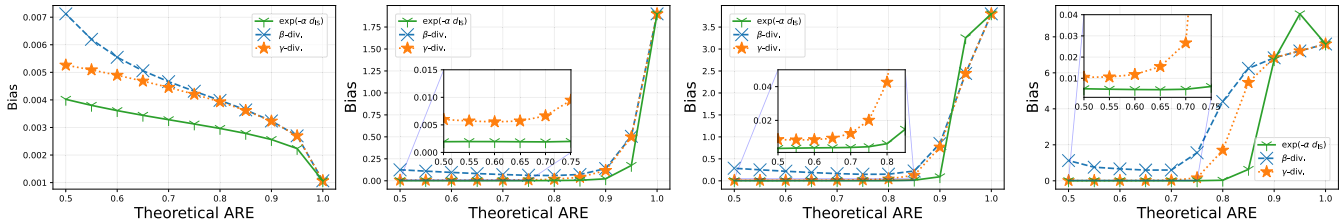
TABLE I

CORRESPONDENCE BETWEEN ARE AND TUNING PARAMETERS UNDER THE EXPONENTIAL DISTRIBUTION

ARE	α	β	γ
1	0	0	0
0.95	0.2140	0.1286	0.1275
0.9	0.3354	0.1972	0.1930
0.85	0.4488	0.2605	0.2508
0.8	0.5630	0.3249	0.3063
0.75	0.6823	0.3941	0.3617
0.7	0.8099	0.4720	0.4185
0.65	0.9489	0.5638	0.4779
0.6	1.1030	0.6779	0.5410
0.55	1.2763	0.8295	0.6091
0.5	1.4747	1.0488	0.6836

$\sigma^2_{\text{out}} = 1$. The proportion of outliers was set to $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4\}$. We changed the number of data samples as 30, 50, 100, and 1000. Since we observed similar tendencies in results as discussed below, we omit showing the results, for $n = 50$ and 1000, and for $\varepsilon = 0.3$ in this paper. We averaged the results over $10,000$ trials. For comparison, the tuning parameters corresponding to AREs between 0.5 and 1 were obtained from Figure 2 for f-separable IS distortion measure, β- and γ-divergences. This means that all three estimators have an equal ARE if there is no contamination by outliers. Note that ARE and tuning parameters are inversely proportional. Table I shows the correspondence between ARE and tuning parameters under the exponential distribution, where α, β, and γ are tuning parameters of f-separable IS distortion measure using (8), β-divergence, and γ-divergence, respectively. The full correspondence between ARE and tuning parameters is shown in Figure 2.

Figures 12–15 show the MSE and bias of the estimators for $n = 30$ and 100. Note that, the x-axis represents ARE, not tuning parameters. When ARE is close to one, it is weak against outliers but efficient. Conversely, when ARE is close to 0.5, it is outlier-resistant but less efficient. When outliers are not mixed ($\varepsilon = 0$), for all estimators, both MSE and bias monotonically increase as ARE decreases. This works as a sanity check for using ARE despite its asymptotic nature. For $\varepsilon = 0$, the MSEs of all three estimators are roughly same, whereas the bias of the f-separable IS distortion measure is smaller than those of β- and γ-divergences. When the proportion of outliers is 0.1, the MSEs of f-separable IS distortion measure, and β- and γ-divergences show similar trends. However, the f-separable IS distortion measure shows slightly better performance overall. In particular, it shows better results for both MSE and bias for AREs around 0.95. When the proportion of outliers is greater than 0.2, the MSE of the f-separable IS distortion measure is greather than those of β- and γ-divergences for AREs between 0.9 and 1. This is because the bias is then increased relative to the MLE (Figures 14 and 15). Considering the bias, when ARE is 0.9, the bias in the f-separable IS distortion measure is smaller than those in β- and γ-divergences, but the MSE is greater, and



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.1$ (c) $\varepsilon = 0.2$ (d) $\varepsilon = 0.4$

Fig. 15. Bias of the exponential distribution with the outliers under $n = 100$ samples.

hence the variance is larger. On the other hand, f-separable IS distortion measure shows better performance than β- and γ-divergeces in most regions where ARE is less than 0.85.

IX. Conclusion

In this paper, we discussed the condition for the unbiased estimating equation in the class of parameter estimation by minimizing the f-separable Bregman distortion measures. Its condition consists of the statistical model, Bregman divergence, and function f. We clarified that the condition the function f and statistical model should satisfy is characterized by a simple integral for Mahalanobis and IS distances. These results were extended to the case of one-dimensional Bregman divergence. In estimating the scale parameter of the gamma distribution, divergence-based estimation generally requires bias correction terms. Furthermore, we proved that the vanishing bias correction term implies the possibility of minimizing latent bias caused by the large proportion of outliers. We demonstrated that the latent bias could approach zero through experiments with outliers or very small inliers mixed. For the choice of function f, there is a trade-off between robustness against outliers and model efficiency. Methods for determining the tuning parameters of divergence have been studied [47], [48], [49]. These methods can be used to determine the appropriate function f and the strictly convex function ϕ.

Appendix A
Proof of Theorem 1

We decompose the positive definite matrix $\boldsymbol{A}$ as, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{V}^{-1} = \boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with positive eigenvalues. Then, Mahalanobis distance is rewritten as

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\theta})\\ =\ &(\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\theta})\\ =\ &\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{j=1}^{d}\lambda_j y_j^2 = \|\boldsymbol{s}\|^2,\end{aligned}$$

where $\boldsymbol{y} = \boldsymbol{V}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\theta})$, $\boldsymbol{s} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{y}$ and λ_j is the j-th element of the diagonal matrix $\boldsymbol{\Lambda}$. From (16), we have

$$\begin{aligned}&\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[f'(d_{\mathrm{Mah.}}(\boldsymbol{X},\boldsymbol{\theta}))(\boldsymbol{X}-\boldsymbol{\theta})\right]\\ =\ &\int_{\mathbb{R}^d}\frac{|\boldsymbol{A}|^{\frac{1}{2}}}{C_{\mathrm{Mah.}}}g\left(d_{\mathrm{Mah.}}(\boldsymbol{x},\boldsymbol{\theta})\right)f'\left(d_{\mathrm{Mah.}}(\boldsymbol{x},\boldsymbol{\theta})\right)(\boldsymbol{x}-\boldsymbol{\theta})d\boldsymbol{x}\\ =\ &\int_{\mathbb{R}^d}\frac{|\boldsymbol{V}||\boldsymbol{\Lambda}|^{\frac{1}{2}}}{C_{\mathrm{Mah.}}}g\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}\right)f'\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}\right)\boldsymbol{V}\boldsymbol{y}\left|\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right|d\boldsymbol{y}\\ =\ &|\boldsymbol{V}|^2|\boldsymbol{\Lambda}|^{\frac{1}{2}}\boldsymbol{V}\int_{\mathbb{R}^d}\frac{1}{C_{\mathrm{Mah.}}}g\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}\right)f'\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}\right)\boldsymbol{y}d\boldsymbol{y}\\ =\ &|\boldsymbol{\Lambda}|^{\frac{1}{2}}\boldsymbol{V}\int_{\mathbb{R}^d}\frac{1}{C_{\mathrm{Mah.}}}g\left(\|\boldsymbol{s}\|^2\right)f'\left(\|\boldsymbol{s}\|^2\right)\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\boldsymbol{s}\left|\frac{\partial(\boldsymbol{y})}{\partial(\boldsymbol{s})}\right|d\boldsymbol{s}\\ =\ &\boldsymbol{V}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\int_{\mathbb{R}^d}\frac{1}{C_{\mathrm{Mah.}}}g\left(\|\boldsymbol{s}\|^2\right)f'\left(\|\boldsymbol{s}\|^2\right)\boldsymbol{s}d\boldsymbol{s}\\ =\ &\boldsymbol{V}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\mathbb{E}_{p(\boldsymbol{s})}\left[f'\left(\|\boldsymbol{S}\|^2\right)\boldsymbol{S}\right],\end{aligned}\tag{36}$$

where Jacobians are given by

$$\left|\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right| = |\boldsymbol{V}|,$$
$$\left|\frac{\partial(\boldsymbol{y})}{\partial(\boldsymbol{s})}\right| = |\boldsymbol{\Lambda}|^{-\frac{1}{2}},$$

respectively. Notably, because the matrix $\boldsymbol{V}$ is an orthogonal matrix, $|\boldsymbol{V}|^2 = 1$. Here, the random vector $\boldsymbol{S}$ follows a spherical distribution $p(\boldsymbol{s}) = \frac{1}{C_{\mathrm{Mah.}}}g(\|\boldsymbol{s}\|^2)$. We refer to the next theorem.

Theorem 6 ([44, pp. 37–38]): Suppose $\boldsymbol{S} = (S_1,\cdots,S_d) \sim \frac{1}{C_{\mathrm{Mah.}}}g(\|\boldsymbol{s}\|^2)$, $d \geq 2$. Consider the transformation to spherical coordinates for $\boldsymbol{S}$,

$$\begin{aligned}&\boldsymbol{S}(R,\boldsymbol{H}) = R\bar{\boldsymbol{s}}(\boldsymbol{H}),\\ &\bar{\boldsymbol{s}}(\boldsymbol{\eta}) = \begin{cases}\bar{s}_j = \left(\prod_{k=1}^{j-1}\sin\eta_k\right)\cos\eta_j,\ 1\leq j\leq d-1,\\ \bar{s}_d = \left(\prod_{k=1}^{d-2}\sin\eta_k\right)\sin\eta_{d-1},\end{cases}\end{aligned}\tag{37}$$

where $R \geq 0$, $H_k \in [0,\pi)$, $k = 1,\cdots,d-2$, $H_{d-1} \in [0,2\pi)$. Then R, $H_1,\cdots,H_{d-1}$ are independent and, respectively, have the following probability density functions

$$\begin{cases}h_r(r) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}\frac{1}{C_{\mathrm{Mah.}}}r^{d-1}g(r^2),\ r\geq 0,\\ h_{\eta_k}(\eta_k) = \frac{1}{B(\frac{1}{2},\frac{d-k}{2})}\sin^{d-k-1}\eta_k,\ 0\leq\eta_k<\pi, k=1,\cdots,d-2,\\ h_{\eta_{d-1}}(\eta_{d-1}) = \frac{1}{2\pi},\ 0\leq\eta_{d-1}<2\pi.\end{cases}\tag{38}$$

Conversely if R, $H_1,\cdots,H_{d-1}$ are independent and have probability density functions given by (38), and $\boldsymbol{S}$ is defined by (37), then $\boldsymbol{S} \sim \frac{1}{C_{\mathrm{Mah.}}}g(\|\boldsymbol{s}\|^2)$. Here, $B(\cdot,\cdot)$ is the beta function.

Note that R is the random variable with respect to radius, i.e., $R = \|\boldsymbol{S}\|$ and $\boldsymbol{H}$ is the random vector that follows

the uniform distribution on the unit hypersphere. The joint distribution of (38) [44, p. 38] is given by

$$h(r,\boldsymbol{\eta}) = \frac{1}{C_{\text{Mah.}}} g(r^2) r^{d-1} \prod_{k=1}^{d-2} \sin^{d-k-1}\eta_k. \tag{39}$$

From (36), (37), and (39), we have

$$\begin{aligned}
&\boldsymbol{V}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1} \mathbb{E}_{p(\boldsymbol{s})}\left[f'\left(\|\boldsymbol{S}\|^2\right)\boldsymbol{S}\right] \\
&= \boldsymbol{V}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1} \mathbb{E}_{h(r,\boldsymbol{\eta})}\left[f'(R^2)s(R,\boldsymbol{H})\right] \\
&= \boldsymbol{V}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1} \frac{1}{C_{\text{Mah.}}} \underbrace{\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi} \bar{\boldsymbol{s}}(\boldsymbol{\eta}) \prod_{k=1}^{d-2}\sin^{d-k-1}\eta_k d\boldsymbol{\eta}}_{=\boldsymbol{0}} \\
&\quad \cdot \int_0^{\infty} f'(r^2) g(r^2) r^d dr.
\end{aligned}$$

This means that the expected value of the uniform distribution on the unit hypersphere is zero. Therefore, if the following integral exists, then the unbiased estimating equation holds without any bias correction term

$$\begin{aligned}
&\int_0^{\infty} g(r^2) f'(r^2) r^d dr \\
&= \frac{1}{2}\int_0^{\infty} g(t) f'(t) t^{\frac{d-1}{2}} dt,
\end{aligned}$$

where we used integration by substitution as $t = r^2$. Conversely, the existence (finiteness) of the bias correction term requires that the absolute value of its each element also has a finite expectation. This requires that $f'(R^2)R$ has a finite expectation in the above discussion since $\boldsymbol{S}/R = \bar{s}(\boldsymbol{H})$ is always bounded. This means that the condition (18) is also necessary.

Appendix B
Property of Scale Family

The scale family is defined by

$$p(x|\theta) = \frac{1}{\theta} h\left(\frac{x}{\theta}\right),$$

where h is the probability density function and $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ is the scale parameter. Let us consider the following case,

$$h(x) = \frac{1}{C_{\text{SF}}} \bar{h}(x),$$

where C_{SF} is the normalization constant as follows,

$$C_{\text{SF}} = \int_0^{\infty} \bar{h}(x) dx.$$

That is, the scale family is given by

$$p(x|\theta) = \frac{1}{\theta}\frac{1}{C_{\text{SF}}} \bar{h}\left(\frac{x}{\theta}\right). \tag{40}$$

Theorem 7: The following relation holds with respect to the expected value $\mathbb{E}[X]$ and the normalization constant C_{SF},

$$\begin{aligned}
&\mathbb{E}_{p(x|\theta)}[X] = \theta < \infty \iff \\
&C_{\text{SF}} = \int_0^{\infty} \bar{h}(x) dx = \int_0^{\infty} x\bar{h}(x) dx < \infty.
\end{aligned}$$

Proof: From the definition and finiteness of the expected value $\mathbb{E}[X]$, we have

$$\begin{aligned}
&\infty > \theta = \mathbb{E}_{p(x|\theta)}[X] \\
&= \int_0^{\infty} \frac{x}{\theta}\frac{1}{C_{\text{SF}}}\bar{h}\left(\frac{x}{\theta}\right) dx = \frac{\theta}{C_{\text{SF}}}\int_0^{\infty} x\bar{h}(x) dx.
\end{aligned}$$

Here, the normalization constant C_{SF} must satisfy

$$C_{\text{SF}} = \int_0^{\infty} \bar{h}(x) dx = \int_0^{\infty} x\bar{h}(x) dx < \infty. \tag{41}$$

Therefore, we have

$$\mathbb{E}_{p(x|\theta)}[X] = \theta < \infty \Rightarrow (41).$$

Conversely, when we assume (41), we have

$$(41) \Rightarrow \mathbb{E}_{p(x|\theta)}[X] = \theta < \infty.$$

Therefore, Theorem 7 holds. □

If we set $\bar{h}(x) = \frac{1}{x} g\left(d_{\text{IS}}(x,1)\right)$, then, the scale family (40) reduces to the IS distribution (20). We immediately obtain Lemma 2 as a corollary to Theorem 7.

Appendix C
Robust Divergences

A. β-Divergence

The β-divergence between two probability density functions q and p is defined as the difference of β-cross-entropy,

$$D_\beta(q,p) = d_\beta(q,p) - d_\beta(q,q),$$

where it is defined by

$$d_\beta(q,p) = -\frac{1}{\beta}\int q(x) p(x|\theta)^\beta dx + \frac{1}{1+\beta}\int p(x|\theta)^{1+\beta} dx.$$

Ordinally, the probability distribution q is the data-generating distribution, and p is the statistical model. Thus, the β-cross-entropy is minimized to estimate the probability distribution p by minimizing. However, the empirical distribution is substituted for the empirical estimation since the true distribution q is unknown. The objective function to be minimized is given by the following equation, where the empirical distribution is substituted for q as

$$L_\beta(\theta) = -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} p(x_i|\theta)^\beta + \frac{1}{1+\beta}\int p(x|\theta)^{1+\beta} dx.$$

1) Exponential Distribution: The objective function and update rule for the β-divergence, assuming the exponential distribution for the statistical model, are given by

$$\begin{aligned}
L_\beta(\theta) &= -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\theta}\exp\left(-\frac{x_i}{\theta}\right)\right]^\beta + \frac{1}{(1+\beta)^2\theta^\beta}, \\
\theta &= \frac{\sum_{i=1}^{n}\exp(-\beta\frac{x_i}{\theta})x_i}{\sum_{j=1}^{n}\exp(-\beta\frac{x_j}{\theta}) - \frac{n\beta}{(1+\beta)^2}}.
\end{aligned}$$

2) Gamma Distribution: The objective function and update rule for the β-divergence, assuming the gamma distribution (28) with a known shape parameter $k > 0$ for the statistical model, are given by

$$L_\beta(\theta) = -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)} x_i^{k-1}\exp\left(-\frac{k}{\theta}x_i\right)\right]^\beta + \frac{1}{(1+\beta)^{k+\beta(k-1)+1}}\left(\frac{k}{\theta}\right)^\beta \frac{\Gamma(k+\beta(k-1))}{[\Gamma(k)]^{1+\beta}},$$
$$\theta = \frac{\sum_{i=1}^n x_i^{\beta(k-1)+1}\exp\left(-\beta\frac{k}{\theta}x_i\right)}{\sum_{j=1}^n x_j^{\beta(k-1)}\exp\left(-\beta\frac{k}{\theta}x_j\right) - M},$$

where

$$M = \frac{n\beta}{(1+\beta)^{k+\beta(k-1)+1}}\left(\frac{\theta}{k}\right)^{\beta(k-1)}\frac{\Gamma(k+\beta(k-1))}{\Gamma(k)}.$$

Note that when $0 < k < 1$, the tuning parameter β for the β-divergence is limited to the following ranges:

$$0 \le \beta < \frac{k}{1-k},\ (0 < k < 1).$$

3) Gaussian Distribution: Similarly, when the Gaussian distribution is assumed for the statistical model, the objective function and the update rules for the mean and variance parameters are given by

$$L_\beta(\mu,\sigma^2) = -\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^\beta + (1+\beta)^{-\frac{3}{2}}(2\pi\sigma^2)^{-\frac{\beta}{2}},$$
$$\mu = \frac{\sum_{i=1}^n \exp\left(-\beta\frac{(x_i-\mu)^2}{2\sigma^2}\right)x_i}{\sum_{j=1}^n \exp\left(-\beta\frac{(x_j-\mu)^2}{2\sigma^2}\right)},$$

and

$$\sigma^2 = \frac{\sum_{i=1}^n \exp\left(-\beta\frac{(x_i-\mu)^2}{2\sigma^2}\right)(x_i-\mu)^2}{\sum_{j=1}^n \exp\left(-\beta\frac{(x_j-\mu)^2}{2\sigma^2}\right) - \frac{n\beta}{(1+\beta)^{\frac{3}{2}}}}$$

respectively.

B. γ-Divergence

As in the case of β-divergence, the γ-divergence between two probability density functions is defined as the difference of the corresponding cross entropies as follows

$$D_\gamma(q,p) = d_\gamma(q,p) - d_\gamma(q,q),$$

where $d_\gamma(q,p)$ represents the γ-cross-entropy which is defined by

$$d_\gamma(q,p) = -\frac{1}{\gamma}\log\int q(x)p(x|\theta)^\gamma dx + \frac{1}{1+\gamma}\log\int p(x|\theta)^{1+\gamma}dx.$$

The objective function to be minimized is given by the following equation by replacing the true distribution with the empirical distribution,

$$L_\gamma(\theta) = -\frac{1}{\gamma}\log\left[\frac{1}{n}\sum_{i=1}^n p(x_i|\theta)^\gamma\right] + \frac{1}{1+\gamma}\log\int p(x|\theta)^{1+\gamma}dx.$$

1) Exponential Distribution: The objective function and update rule for the γ-divergence, assuming the exponential distribution for the statistical model, are given by

$$L_\gamma(\theta) = -\frac{1}{\gamma}\log\left[\frac{1}{n}\sum_{i=1}^n\left[\frac{1}{\theta}\exp\left(-\frac{x_i}{\theta}\right)\right]^\gamma\right] - \frac{1}{1+\gamma}(\gamma\log\theta + \log(1+\gamma)),$$
$$\theta = (1+\gamma)\frac{\sum_{i=1}^n \exp(-\gamma\frac{x_i}{\theta})x_i}{\sum_{j=1}^n \exp(-\gamma\frac{x_j}{\theta})}.$$

2) Gamma Distribution: The objective function and update rule for the γ-divergence, assuming the gamma distribution (28) with a known shape parameter $k > 0$ for the statistical model, are given by

$$L_\gamma(\theta) = -\frac{1}{\gamma}\log\left[\frac{1}{n}\sum_{i=1}^n\left[\left(\frac{k}{\theta}\right)^k\frac{1}{\Gamma(k)}x^{k-1}\exp\left(-\frac{k}{\theta}x\right)\right]^\gamma\right] - \frac{1}{1+\gamma}\Big[\gamma\log\theta - \gamma\log k + (k+\gamma(k-1))\log(1+\gamma) - \log\Gamma(k+\gamma(k-1))\Big] - \log\Gamma(k),$$
$$\theta = \frac{(1+\gamma)k}{k+\gamma(k-1)}\frac{\sum_{i=1}^n x_i^{\gamma(k-1)+1}\exp\left(-\gamma\frac{k}{\theta}x_i\right)}{\sum_{j=1}^n x_j^{\gamma(k-1)}\exp\left(-\gamma\frac{k}{\theta}x_j\right)}.$$

Note that when $0 < k < 1$, the tuning parameter γ for γ-divergence is limited to the following range:

$$0 \le \gamma < \frac{k}{1-k},\ (0 < k < 1).$$

3) Gaussian Distribution: Similarly, when the Gaussian distribution is assumed for the statistical model, the objective function and the update rules for the mean and variance parameters are given by

$$L_\gamma(\mu,\sigma^2) = -\frac{1}{\gamma}\log\left[\frac{1}{n}\sum_{i=1}^n\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^\gamma\right] - \frac{1}{2(1+\gamma)}\left[\gamma\log(2\pi\sigma^2) + \log(1+\gamma)\right],$$
$$\mu = \frac{\sum_{i=1}^n \exp\left(-\gamma\frac{(x_i-\mu)^2}{2\sigma^2}\right)x_i}{\sum_{j=1}^n \exp\left(-\gamma\frac{(x_j-\mu)^2}{2\sigma^2}\right)},$$

and

$$\sigma^2 = (1+\gamma)\frac{\sum_{i=1}^n \exp\left(-\gamma\frac{(x_i-\mu)^2}{2\sigma^2}\right)(x_i-\mu)^2}{\sum_{j=1}^n \exp\left(-\gamma\frac{(x_j-\mu)^2}{2\sigma^2}\right)}$$

respectively.

ACKNOWLEDGMENT

The authors are grateful of two anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. Hoboken, NJ, USA: Wiley, 2005.

[2] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Hoboken, NJ, USA: Wiley, 2009.

[3] A. Basu, H. Shioya, and C. Park, *Statistical Inference: The Minimum Distance Approach*. Boca Raton, FL, USA: CRC Press, 2011.

[4] L. Pardo, *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: CRC Press, 2006.

[5] A. Cichocki and S. Amari, "Families of Alpha- Beta- and Gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, Jun. 2010.

[6] S. Amari, "α-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4925–4931, Nov. 2009.

[7] S. Amari, *Information Geometry and Its Applications*. Tokyo, Japan: Springer, 2016.

[8] R. Beran, "Minimum Hellinger distance estimates for parametric models," *Ann. Statist.*, vol. 5, no. 3, pp. 445–463, May 1977.

[9] A. Basu and B. G. Lindsay, "Minimum disparity estimation for continuous models: Efficiency, distributions and robustness," *Ann. Inst. Stat. Math.*, vol. 46, no. 4, pp. 683–705, Dec. 1994.

[10] M. Broniatowski and A. Keziou, "Parametric estimation and tests through divergences and the duality technique," *J. Multivariate Anal.*, vol. 100, no. 1, pp. 16–36, Jan. 2009.

[11] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, Nov. 2010.

[12] D. Al Mohamad, "Towards a better understanding of the dual representation of phi divergences," *Stat. Papers*, vol. 59, no. 3, pp. 1205–1253, Sep. 2018.

[13] M. Broniatowski, A. Toma, and I. Vajda, "Decomposable pseudodistances and applications in statistical estimation," *J. Stat. Planning Inference*, vol. 142, no. 9, pp. 2574–2585, Sep. 2012.

[14] S. Jana and A. Basu, "A characterization of all single-integral, non-kernel divergence estimators," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 7976–7984, Dec. 2019.

[15] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, Sep. 1998.

[16] H. Fujisawa, "Normalized estimating equation for robust parameter estimation," *Electron. J. Statist.*, vol. 7, pp. 1587–1606, 2013.

[17] T. Mukherjee, A. Mandal, and A. Basu, "The B-exponential divergence and its generalizations with applications to parametric estimation," *Stat. Methods Appl.*, vol. 28, no. 2, pp. 241–257, Jun. 2019.

[18] P. Singh, A. Mandal, and A. Basu, "Robust inference using the exponential-polynomial divergence," *J. Stat. Theory Pract.*, vol. 15, no. 2, pp. 1–22, Jun. 2021.

[19] S. Roy, K. Chakraborty, S. Bhadra, and A. Basu, "Density power downweighting and robust inference: Some new strategies," *J. Math. Statist.*, vol. 15, no. 1, pp. 333–353, 2019.

[20] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-Boost and Bregman divergence," *Neural Comput.*, vol. 16, no. 7, pp. 1437–1481, Jul. 2004.

[21] S. Eguchi and Y. Kano, "Robustifing maximum likelihood estimation by psi-divergence," Inst. Stat. Math. (ISM), ISM Res. Memo, Tachikawa, Tokyo, Tech. Rep., 802, 2001.

[22] M. P. Windham, "Robustifying model fitting," *J. Roy. Stat. Soc., B, Methodol.*, vol. 57, no. 3, pp. 599–609, Sep. 1995.

[23] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu, "A comparison of related density-based minimum divergence estimators," *Biometrika*, vol. 88, no. 3, pp. 865–873, Oct. 2001.

[24] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *J. Multivariate Anal.*, vol. 99, no. 9, pp. 2053–2081, Oct. 2008.

[25] T. Kanamori and H. Fujisawa, "Affine invariant divergences associated with proper composite scoring rules and their applications," *Bernoulli*, vol. 20, no. 4, pp. 2278–2304, Nov. 2014.

[26] T. Kanamori and H. Fujisawa, "Robust estimation under heavy contamination using unnormalized models," *Biometrika*, vol. 102, no. 3, pp. 559–572, Sep. 2015.

[27] A. K. Kuchibhotla, S. Mukherjee, and A. Basu, "Statistical inference based on bridge divergences," *Ann. Inst. Stat. Math.*, vol. 71, no. 3, pp. 627–656, Jun. 2019.

[28] S. Ray, S. Pal, S. K. Kar, and A. Basu, "Characterizing the functional density power divergence class," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 1141–1146, Feb. 2023.

[29] A. Ghosh, I. R. Harris, A. Maji, A. Basu, and L. Pardo, "A generalized divergence for statistical inference," *Bernoulli*, vol. 23, no. 4A, pp. 2746–2783, Nov. 2017.

[30] A. Ghosh, "Asymptotic properties of minimum S-divergence estimator for discrete models," *Sankhya A*, vol. 77, no. 2, pp. 380–407, Aug. 2015.

[31] A. Ghosh and A. Basu, "The minimum S-divergence estimator under continuous models: The Basu–Lindsay approach," *Stat. Papers*, vol. 58, no. 2, pp. 341–372, Jun. 2017.

[32] A. Ghosh and A. Basu, "A new family of divergences originating from model adequacy tests and application to robust statistical inference," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5581–5591, Aug. 2018.

[33] A. Maji, A. Ghosh, and A. Basu, "The logarithmic super divergence and asymptotic inference properties," *Adv. Stat. Anal.*, vol. 100, no. 1, pp. 99–131, Jan. 2016.

[34] T. van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[35] A. Maji, A. Ghosh, A. Basu, and L. Pardo, "Robust statistical inference based on the C-divergence family," *Ann. Inst. Stat. Math.*, vol. 71, no. 5, pp. 1289–1322, Oct. 2019.

[36] F. Vonta, K. Mattheou, and A. Karagrigoriou, "On properties of the (Φ, a)-power divergence family with applications in goodness of fit tests," *Methodol. Comput. Appl. Probab.*, vol. 14, no. 2, pp. 335–356, Jun. 2012.

[37] S. Basak and A. Basu, "The extended Bregman divergence and parametric estimation," *Statistics*, vol. 56, no. 3, pp. 699–718, 2022.

[38] Y. Shkel and S. Verdú, "A coding theorem for f-separable distortion measures," *Entropy*, vol. 20, no. 2, pp. 1–16, Feb. 2018.

[39] M. Kobayashi and K. Watanabe, "Generalized Dirichlet-process-means for f-separable distortion measures," *Neurocomputing*, vol. 458, pp. 667–689, Oct. 2021.

[40] R. V. Lenth and P. J. Green, "Consistency of deviance-based M estimators," *J. Roy. Stat. Soc., Ser. B*, vol. 49, no. 3, pp. 326–330, Jul. 1987.

[41] A. M. Bianco, M. Garcia Ben, and V. J. Yohai, "Robust estimation for linear regression with asymmetric errors," *Can. J. Statist.*, vol. 33, no. 4, pp. 511–528, Dec. 2005.

[42] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[43] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *J. Multivariate Anal.*, vol. 11, no. 3, pp. 368–385, Sep. 1981.

[44] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions*. Boca Raton, FL, USA: CRC Press, 2018.

[45] D. Ferrari and Y. Yang, "Maximum Lq-likelihood estimation," *Ann. Statist.*, vol. 38, no. 2, pp. 753–783, Apr. 2010.

[46] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[47] S. Basak, A. Basu, and M. C. Jones, "On the 'optimal' density power divergence tuning parameter," *J. Appl. Statist.*, vol. 48, no. 3, pp. 536–556, 2021.

[48] J. Warwick and M. C. Jones, "Choosing a robustness tuning parameter," *J. Stat. Comput. Simul.*, vol. 75, no. 7, pp. 581–588, 2005.

[49] S. Sugasawa and S. Yonekura, "On selection criteria for the tuning parameter in robust divergence," *Entropy*, vol. 23, no. 9, pp. 1–10, Sep. 2021.

Masahiro Kobayashi (Member, IEEE) received the Ph.D. degree in engineering from Toyohashi University of Technology, Japan, in 2021. From 2019 to 2021, he was a Research Fellow at the Japan Society for the Promotion of Science (DC2). From July 2021 to August 2021, he was a Post-Doctoral Researcher at the Japan Society for the Promotion of Science (DC2). Since 2021, he has been a Faculty Member with the Information and Media Center, Toyohashi University of Technology, where he is currently an Assistant Professor. His research interests include statistical learning theory and robust statistics.

Kazuho Watanabe (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology in 2002, 2004, and 2006, respectively. From 2007 to 2008, he was a Post-Doctoral Fellow and a Research Associate with The University of Tokyo. From 2009 to 2013, he was an Assistant Professor with Nara Institute of Science and Technology. Since 2014, he has been a Faculty Member with the Department of Computer Science and Engineering, Toyohashi University of Technology, where he is currently an Associate Professor. His research interests include statistical machine learning theory and algorithms.