The extended Bregman divergence and parametric estimation

Sancharee Basak and Ayanendranath Basu
Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India

ABSTRACT
Minimization of suitable statistical distances (between the data and model densities) is a useful technique in the field of robust inference. Apart from the class of $\phi$-divergences, the Bregman divergence is extensively used for this purpose. However, since the data density must have a linear presence in the term involving both the data and model densities in this structure, several useful divergences cannot be captured by the usual Bregman form. We provide an extension of the Bregman divergence by considering an exponent of the density function as the argument rather than the density itself. Many useful divergences, that are not ordinarily Bregman divergences, can be accommodated within this extended description. Using this formulation, one can develop many new families of divergences which may be useful in robust inference. In particular, through an application of this extension, we propose the new class of the GSB divergence family. We explore the applicability of the minimum GSB divergence estimator in discrete parametric models. Simulation studies and real data examples are provided to demonstrate the performance of the estimator and to substantiate the theory developed.

1. Introduction
In the domain of statistical inference, there is an inherent trade-off between model efficiency and robustness. Often we slightly compromise the efficiency of the procedure to obtain better robustness. In the present age of big data, the robustness angle of statistical inference has to be dealt with greater care than ever before. Such methods do exist in the literature which allow full asymptotic efficiency simultaneously with strong robustness properties; see, e.g., [1–3]. In practice, however, one would be hard-pressed to find a procedure which matches the likelihood-based methods in terms of efficiency in small to moderate samples without inheriting any of its robustness limitations. Many of these trade-off issues are discussed in the canonical texts on robustness such as [4–6]; for a minimum divergence view of this issue, see [3].

Several types of divergences are used in minimum distance inference. Most are not mathematical metrics. They may not satisfy the triangle inequality or may not even be symmetric in their arguments. The only properties we demand of these measures are that
they are non-negative and are equal to zero if and only if the two arguments are identically equal. Sometimes we refer to these divergence measures as ‘statistical distances’ or, loosely, as ‘distances’ without any claim to metric properties.

Most density-based divergences belong to either the class of chi-square type distances (formally called $\phi$-divergences, $f$-divergences or disparities) or Bregman divergences. See [2,7,8] for a description of the divergences of the chi-square type and [9] for Bregman divergences. The Bregman divergence has been successfully used in many branches of natural science as well as in areas like information theory and computational geometry. The class of chi-square type distances between two densities $g$ and $f$ includes, for example, the likelihood disparity (LD), the Kullback-Leibler divergence (KLD) and the (twice, squared) Hellinger distance (HD), given by

$$\text{LD}(g,f) = \int g \log \left( \frac{g}{f} \right), \quad \text{KLD}(g,f) = \int f \log \left( \frac{f}{g} \right),$$

$$\text{HD}(g,f) = \frac{1}{2} \int (f^{1/2} - g^{1/2})^2,$$

respectively. Representative members of the class of Bregman divergences include the LD and the squared $L_2$ distance, where

$$L_2(g,f) = \int (g - f)^2. \quad (2)$$

The LD is the only common member between the chi-square and the Bregman classes.

In our parametric estimation scheme, the estimator corresponds to the parameter of the model density which is closest to the observed data density in terms of the given divergence, the observed data density being a non-parametric representative of the true unknown density based on the given sample. In case of chi-square type distances, the construction of the data density inevitably requires the use of an appropriate non-parametric smoothing technique, like kernel density estimation, in continuous models (the LD is the only exception). This makes the derivation of the asymptotic properties far more involved, and complicates the computational aspect of this estimation. On the other hand, all the minimum Bregman divergence estimators are M-estimators and, hence, they avoid this density estimation component.

In this paper, our primary aim is to extend the scope of the Bregman divergence by utilizing powers of densities as arguments, rather than the densities themselves; this leads to the class of the extended Bregman divergences. The use of the Bregman divergence in statistics is relatively recent; the class of density power divergences by [10], defined in Section 2, is a prominent example of Bregman divergences having significant applications in statistical inference. Many minimum divergence procedures have natural robustness properties against data contamination and outliers. The extended Bregman divergence allows us to express several existing divergence families as special cases of it, which is not possible through the ordinary Bregman divergence. Thus the extended Bregman idea can be used to generate large super-families of divergences containing, together with the existing divergences, many new and useful divergence families as special cases. This is the key idea of this work.

In Section 2, we propose the extended Bregman divergence. We demonstrate how it captures well known divergences beyond the ordinary Bregman class and give a potential route
for constructing new divergences. In Section 3, by considering a specific convex function and a particular exponent of the density, we construct a large super-family of divergences within the extended Bregman class. Several known divergence families are special cases of this super-family. Section 4 introduces the corresponding minimum distance estimator, while Section 5 studies its asymptotic properties. Section 6 explores the robustness properties of the estimator based on its influence function. A large simulation study is taken up in Section 7, and tuning parameter selection is considered in Section 8. The final section has some concluding remarks.

We now summarize what we believe to be the main achievements in this paper.

1. We provide a simple extension of the Bregman divergence by considering powers of densities (instead of the densities themselves) as arguments. Many divergences, not ordinarily members of the Bregman class, belong to this extended family, and the properties of the corresponding minimum distance estimators may be obtained from the general properties common to this extended family.

2. Ordinarily, all minimum Bregman divergence estimators are M-estimators. But, through this extension, several minimum distance estimators which are not M-estimators also become a part of this extended family.

3. The choice of a particular convex function and specific exponent produces a generalized super-family which we refer to as the GSB divergence family. The power divergence family of [11], the density power divergence (DPD) family of [10], the Bregman exponential divergence (BED) of [12] and the S-divergence family of [13] can all be brought under the umbrella of this super-family.

4. This GSB family consists of three tuning parameters. By simultaneously varying all of them, we can generate new divergences (and hence new minimum divergence estimators) which are outside the union of the BED and S-divergence families, but can potentially provide improved performance over both of them.

2. The extended Bregman divergence and special cases

Bregman [3] introduced his divergence as a measure of dissimilarity between any two vectors in the Euclidean space. In \( \mathbb{R}^p \), it has the form

\[
D_\psi (x, y) = \left\{ \psi (x) - \psi (y) - \langle \nabla \psi (y), x - y \rangle \right\},
\]

for any strictly convex function \( \psi : S \to \mathbb{R} \) and for any two \( p \)-dimensional vectors \( x, y \in S \), where \( S \) is a convex subset of \( \mathbb{R}^p \). Here, \( \nabla \psi (y) \) denotes the gradient of \( \psi \) with respect to its argument at \( y = (y_1, y_2, \ldots, y_p)^T \). Evidently, only the convexity property of the function \( \psi (\cdot) \) is necessary for the non-negativity of the divergence \( D_\psi (x, y) \). Hence, as long as \( \psi \) remains convex, any set of arguments whose equivalence translates to the equivalence of \( x \) and \( y \) can be used in the distance expression. This observation may be used to extend the Bregman divergence to have the form

\[
D_\psi (x, y) = \left\{ \psi (x^k) - \psi (y^k) - \langle \nabla \psi (y^k), x^k - y^k \rangle \right\}.
\]

Here, \( \nabla \psi (y^k) \) denotes the gradient of \( \psi \) with respect to its argument evaluated at \( y^k = (y_{1k}, y_{2k}, \ldots, y_{dk})^T \), \( k \) is any real non-negative exponent and \( \psi \) is a strictly convex function,
mapping $S$ to $\mathbb{R}$, $S$ being a convex subset of $\mathbb{R}^+$. As our purpose is to utilize this extension in the field of statistics where the arguments, being probability density functions, are inherently non-negative, restricting the domain of $\psi$ to $\mathbb{R}^+$ is not a hindrance. Many of the properties of the Bregman divergence in Equation (3), as described by [14], are retained by the extended version in Equation (4). However, we will not make use of these properties in this paper, so we do not discuss them further.

The Bregman divergence has significant applications in statistical inference for both discrete and continuous models. Given densities $g$ and $f$, the Bregman divergence between them (associated with a convex $\psi$) is given by

$$D_\psi(g,f) = \int \{ \psi(g(x)) - \psi(f(x)) - (g(x) - f(x)) \nabla \psi(f(x)) \} \, dx.$$  

(5)

By the strict convexity of $\psi$, the above integrand is non-negative and, therefore, so is the integral. Clearly the divergence equals zero if and only if the arguments $g$ and $f$ are identically equal. Well-known examples include the LD and the (squared) $L_2$ distance which correspond to $\psi(x) = x \log x$ and $\psi(x) = x^2$ respectively. In a real scenario, one uses $g$ as the data generating density and $f = f_\theta$ as the parametric model density.

The DPD class, a subfamily of Bregman divergences, has been proposed in [10]. It is generated by the function $\psi(x) = x^{\alpha+1} - x^\alpha$, indexed by a non-negative tuning parameter $\alpha$. As a function of $\alpha$, the density power divergence may be expressed as

$$\text{DPD}_\alpha(g,f) = \int \left\{ f^{\alpha+1}(x) - \left(1 + \frac{1}{\alpha}\right) g(x) f^\alpha(x) + \frac{1}{\alpha} g^{\alpha+1}(x) \right\} \, dx.$$  

(6)

For $\alpha = 1$, the above reduces to the (squared) $L_2$ distance between $g$ and $f$, whereas for $\alpha \downarrow 0$, one recovers the likelihood disparity (LD).

The DPD class is distinct from the power divergence class (PD), which has the form

$$\text{PD}_\lambda(g,f) = \frac{1}{\lambda(\lambda + 1)} \int \left\{ g(x) \left[ \left( \frac{g(x)}{f(x)} \right)^\lambda - 1 \right] \right\} \, dx, \quad \lambda \in \mathbb{R}.$$  

(7)

The PD class – see [11] – is a subfamily of chi-square type distances. The latter class of divergences has the general form

$$\rho(g,f) = \int C(\delta(x)) f(x) \, dx,$$  

(8)

where $C$ is a strictly convex function and $\delta(x) = \frac{g(x)}{f(x)} - 1$. The PD family corresponds to

$$C(\delta) = \frac{(\delta + 1)^{\lambda+1} - (\delta + 1)}{\lambda(\lambda + 1)} - \frac{\delta}{\lambda + 1}, \quad \lambda \in \mathbb{R}.$$  

(9)

Important special cases include the LD (obtained as $\lambda \to 0$) and the HD (obtained for $\lambda = -\frac{1}{2}$). The LD is the only common member between the PD and the DPD classes.
### Table 1. Different special cases of the Bregman divergence.

| Choice of convex function $\psi(x)$ | Divergence                        |
|-------------------------------------|-----------------------------------|
| $x^2$                               | (squared) $L_2$ Distance          |
| $x \log(x)$                         | Likelihood Disparity              |
| $x^{1+\alpha} - x$                  | Density Power Divergence (DPD)    |
| $-\log(x)/2\pi$                     | Itakura-Saito Distance            |
| $\frac{2(e^{\beta x} - \beta x - 1)}{\beta^2}$ | Bregman Exponential Divergence |

The Bregman exponential divergence (BED) [12], on the other hand, has the form

$$\text{BED}_\beta (g, f) = \frac{2}{\beta} \int \left\{ e^{\beta f(x)} \left( f(x) - \frac{1}{\beta} \right) - e^{\beta f(x)} g(x) + \frac{e^{\beta g(x)}}{\beta} \right\} dx.$$  \hspace{1cm} (10)

The defining function is $\psi(x) = \frac{2(e^{\beta x} - \beta x - 1)}{\beta^2}$ which is indexed by $\beta \in \mathbb{R}$. This family generates the (squared) $L_2$ distance in the limit $\beta \to 0$.

A list of some Bregman divergences useful in the context of statistical inference is presented in Table 1.

Let $G$ denote the true distribution which is modelled by the parametric family $F = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. Let $g$ and $f_\theta$ be the corresponding densities. We assume that both $G$ and $F_\theta$ belong to $\mathcal{G}$, the class of all cumulative distribution functions having densities with respect to a suitable dominating measure. Our aim is to estimate the unknown parameter $\theta$ by choosing the model density closest to the true density in the Bregman sense. The definition of Bregman divergences as given in Equation (5), useful as it is, does not include many well-known divergences which are extensively used in the literature for different purposes including parameter estimation. The PD family is a prominent example. An inspection of the Bregman form in Equation (5) indicates that the term which involves both densities $g$ and $f$ is of the form

$$\int g(x) \nabla \psi \left( f(x) \right) dx.$$  \hspace{1cm} (11)

Here, the density $g$ has exponent one. Given a random sample $X_1, X_2, \ldots, X_n$ from $G$, the term in Equation (11) can be empirically estimated by $\frac{1}{n} \sum \nabla \psi \left( f_\theta (X_i) \right)$ (with $f = f_\theta$ under the parametric model) so that one can construct an empirical version of the divergence without any non-parametric smoothing. But this restricts the class of divergences that are expressible in the Bregman form. An extension in the spirit of Equation (4) may allow the construction of richer classes of divergences. We define the extended Bregman divergence between two densities $g$ and $f$ as

$$D_{\psi}^{(k)} (g, f) = \int \left\{ \psi \left( g^k (x) \right) - \psi \left( f^k (x) \right) - \left( g^k (x) - f^k (x) \right) \nabla \psi \left( f^k (x) \right) \right\} dx.$$  \hspace{1cm} (12)

Apart from a strictly convex $\psi$, this formulation also depends on a positive index $k$ with which the density is exponentiated. For the rest of the paper, the notation $D_{\psi}^{(k)} (\cdot, \cdot)$ will refer to this general form in Equation (12); the divergence in Equation (5) is a special case for $k = 1$. Evidently, $D_{\psi}^{(k)} (g, f) \geq 0$ for any two densities $g$ and $f$ with respect to the same
measure. The strict convexity of $\psi$ and the non-negativity of the densities indicate that $D_{\psi}^{(k)}(g,f) = 0$ if and only if $g = f$.

In the following, we present some special cases of extended Bregman divergence.

(1) **S-Hellinger family** [13]. If we take $\psi(x) = \frac{2e^{\beta x}}{\beta^2}$ with $k = \frac{1+\alpha}{2}$, $\alpha \in (0,1)$ in Equation (12), it will generate an extension of the BED family having the form

$$\text{BED}_{\beta}^{(k)}(g,f) = \frac{2}{\beta} \int \left\{ e^{\beta f^k(x)} \left( f^k(x) - \frac{1}{\beta} \right) - e^{\beta g^k(x)} g^k(x) + \frac{e^{\beta g^k(x)}}{\beta} \right\} dx. \quad (13)$$

It can be easily shown that, as $\beta \to 0$ and $k = \frac{1+\alpha}{2}$, $\alpha \in (0,1)$, the application of L'Hospital's rule leads to the S-Hellinger Distance (SHD) family with the form

$$\text{SHD}_\alpha(g,f) = \frac{2}{1+\alpha} \int \left( g^{1+\alpha}(x) - f^{1+\alpha}(x) \right)^2 dx. \quad (14)$$

This was introduced by [13] as a special case of the S-divergence family. This family is ordinarily outside the Bregman divergence class, but through this extension, we can express this member of the S-divergence family as a (limiting) member of the extended BED class.

(2) **PD family** [11]. If we take $\psi(x) = \frac{x^{1+B}}{B}$, $A = 1 + \lambda$, $B = -\lambda$ and $\lambda \in \mathbb{R}$, with $k = A$ in Equation (12), we get the PD family introduced in Equation (7).

(3) **S-divergence family** [13]. If we take $\psi(x) = \frac{x^{1+B}}{B}$, $A = 1 + \lambda(1-\alpha)$, $B = \alpha - \lambda(1-\alpha)$, $A + B = 1 + \alpha$, $\alpha \geq 0$, $\lambda \in \mathbb{R}$ and $k = A$ in Equation (12), we get the S-divergence (SD) having the form

$$\text{SD}_{(\alpha,\lambda)}(g,f) = \int \left\{ \frac{1}{B} \left( g^{A+B}(x) - f^{A+B}(x) \right) - \left( g^A(x) - f^A(x) \right) \frac{A+B}{AB} f^B(x) \right\} dx. \quad (15)$$

This is a very useful divergence family in the domain of robust inference due to its capacity to generate many more robust estimator(s) compared to the DPD and PD families. The S-divergences are generally not Bregman divergences, but belong to the extended class.

For $k \neq 1$, $f^k$ and $g^k$ will generally no longer represent probability densities, and by extending the divergence idea to general positive measures, Amari [15] has suggested certain constructions where the power divergence has been exhibited in the Bregman divergence form for general measures. Also, see the discussion in [16]. We differ from the interpretation in these papers in that we still view the divergences presented here in Equation (12) as divergences between valid probability densities. Given any two probability densities, the expression in Equation (12) is non-negative, and equals zero if and only if the densities $g$ and $f$ are identical, irrespective of the value of $k$. In this sense, our approach is fundamentally different from that of [15].
For the ordinary Bregman divergence, the term in Equation (11), with $f = f_\theta$, may be approximated by

$$
\frac{1}{n} \sum_{i=1}^{n} \nabla \psi \left( f_\theta (X_i) \right),
$$

by replacing the theoretical expectation with the observed sample mean based on the random sample $X_1, X_2, \ldots, X_n$. Evidently the minimizer of the empirical version of the Bregman divergence is an M-estimator. But there are several useful divergences where this empirical representation is not possible, and such divergences generate estimators beyond the M-estimator class. See [17] for more discussion on this issue. We provide several examples where the extended Bregman class contains such divergences which are not covered by the ordinary Bregman form. Thus the structure of the extended class extends the scope much beyond that of the ordinary Bregman divergence.

### 3. Introducing a new divergence family

Our aim is to exploit the extended Bregman idea and generate rich new super families of divergences by choosing a suitable convex generating function and a suitable exponent. In particular, we use the convex function

$$
\psi(x) = e^{\beta x} + \frac{x^{1+\frac{\alpha}{\lambda}}}{B},
$$

$A = 1 + \lambda(1 - \alpha), B = \alpha - \lambda(1 - \alpha), A + B = 1 + \alpha, \alpha \geq -1, \beta, \lambda \in \mathbb{R}$, which, together with the exponent $k = A$, generates the divergence

$$
D^\ast (g,f) = \int \left\{ e^{\beta f^A} \left( \beta f^A - \beta g^A - 1 \right) + e^{\beta g^A} + \frac{1}{B} \left( g^{A+B} - f^{A+B} \right) 
- \left( g^A - f^A \right) \frac{A + B}{AB} f^B \right\} dx,
$$

which we refer to as the generalized S-Bregman (GSB) divergence. The divergence measure $D^\ast$ is also a function of $\alpha, \lambda$ and $\beta$, which we suppress for brevity. Furthermore, the above form of divergence will be defined only corresponding to those triplets $(\alpha, \lambda, \beta)$ for which $A \neq 0$ and $B \neq 0$.

Denote $A^* = A + B$. If we put $A^* = 0$ in the above expression with $A \neq 0$ and $B \neq 0$, we will get the extended BED family with parameter $\beta$ and exponent $k = A$. Moreover, if $A = 1$, i.e., $\lambda = 0$ then it will lead to the ordinary BED family with parameter $\beta$. On the contrary, if we put $\beta = 0$, it will lead to the S-divergence family with parameters $\alpha$ and $\lambda$ (in terms of $A$ and $B$). More specifically, when $\alpha = 0$ and $\beta = 0$, it leads to the PD family. On the other hand, $\beta = 0$ and $\lambda = 0$ leads to the DPD family. Thus, it acts as a connector between the BED and the S-divergence family.

We get several well-known divergences from the general form of GSB for particular choices of the tuning parameters $\alpha, \lambda$ and $\beta$. Some such choices are given in Table 2.
Table 2. Different divergences as special cases of GSB divergence.

| $\alpha$ | $\lambda$ | $\beta$ | Divergences |
|----------|-----------|---------|-------------|
| $\alpha = -1$ | $\lambda = 0$ | $\beta \in \mathbb{R}$ | Bregman Exponential Divergence$^a$ |
| $\alpha = 0$ | $\lambda \in \mathbb{R}$ | $\beta = 0$ | Power Divergence |
| $\alpha \in \mathbb{R}$ | $\lambda = 0$ | $\beta \in \mathbb{R}$ | Density Power Divergence |
| $\alpha = 0$ | $\lambda = -1$ | $\beta = 0$ | $S$-Divergence |
| $\alpha = 0$ | $\lambda = 0$ | $\beta = 0$ | Kullback-Liebler Divergence |
| $\alpha \in \mathbb{R}$ | $\lambda = 0$ | $\beta = 0$ | Likelihood Disparity |
| $\alpha = 0$ | $\lambda = 1$ | $\beta = 0$ | Pearson’s Chi-square Divergence |
| $\alpha = 0$ | $\lambda = -2$ | $\beta = 0$ | Neyman’s Chi-square Divergence |
| $\alpha = 1$ | $\lambda \in \mathbb{R}$ | $\beta = 0$ | (squared) $L_2$ Distance |

$^a$This is a constant time $B$-exponential divergence. It basically generates all the members of BED family corresponding to the same $\beta$ except (squared) $L_2$ distance, which occurs when $\beta \to 0$. However, as seen above, the (squared) $L_2$ distance remains a member of the GSB class for other choices of the tuning parameters.

4. The minimum GSB divergence estimator

Under the parametric set-up described in Section 2, we would like to identify the best fitting parameter $\theta^g$ by choosing the parameter of the model element which provides the closest match to the true density $g$ in terms of the given divergence. The minimum GSB divergence functional $T_{\alpha,\lambda,\beta} : \mathcal{G} \to \Theta$ is defined by the relation

$$D^* (g, f_{T_{\alpha,\lambda,\beta}}) = \min \{D^* (g, f_{\theta}) : \theta \in \Theta \},$$

provided the minimum exists. If the model family is identifiable, it follows that $D^* (g, f_{\theta}) = 0$, if and only if $g = f_{\theta}$. Thus, $T_{\alpha,\lambda,\beta} (F_{\theta}) = \theta$, uniquely. Hence, the functional $T_{\alpha,\lambda,\beta}$ is Fisher consistent. Given the density $g$, a straightforward differentiation of the GSB divergence of Equation (17) leads to the estimating equation

$$\int \left\{ A^2 \beta^2 e^{\beta f_{\theta}^A (x)} f_{\theta}^A (x) + A^* f_{\theta}^B (x) \right\} \left( f_{\theta}^A (x) - g^A (x) \right) u_{\theta} (x) \, dx = 0. \quad (18)$$

In practice, the true density $g$ is unknown, so one has to use a suitable non-parametric density estimator $\hat{g}$ for $g$. Under a discrete parametric set-up, the natural choice for $\hat{g}$ is the vector of relative frequencies as obtained from the sample data. Thus, assuming a discrete parametric model, and assuming, without loss of generality, that the support of the random variable is $\{0, 1, 2, \ldots, \}$, the estimating equation becomes

$$\sum_{x=0}^{\infty} A^2 \beta^2 e^{\beta f_{\theta}^A (x)} f_{\theta}^{2A} (x) u_{\theta} (x) + \sum_{x=0}^{\infty} A^* f_{\theta}^{A*} (x) u_{\theta} (x)$$

$$= \sum_{x=0}^{\infty} A^2 \beta^2 e^{\beta f_{\theta}^A (x)} \hat{g}^A (x) u_{\theta} (x) + \sum_{x=0}^{\infty} A^* f_{\theta}^B (x) \hat{g}^A (x) u_{\theta} (x). \quad (19)$$

For continuous models, on the other hand, some suitable non-parametric smoothing technique such as kernel density estimation is inevitable unless the exponent $A$ equals 1. In the latter case, one can replace the theoretical mean with the empirical sample mean and hence
for $A = 1$, Equation (19) can be reduced to

$$
\sum_{x=0}^{\infty} \beta^2 e^{\beta f_0(x)} f_0^2(x) u_\theta(x) + \sum_{x=0}^{\infty} (1 + B) f_0^{1+B}(x) u_\theta(x) = \frac{1}{n} \sum_{i=1}^{n} \beta^2 e^{\beta f_0(X_i)} f_0(X_i) u_\theta(X_i) + \frac{1}{n} \sum_{i=1}^{n} (1 + B) f_0^B(X_i) u_\theta(X_i). \tag{20}
$$

Clearly, the corresponding estimator is an M-estimator.

In accordance with the information on the first three rows of Table 2, we will refer to the parameters $\alpha, \lambda$ and $\beta$ as the DPD, the PD and the BED parameters, respectively.

### 5. Asymptotic properties of the minimum GSB divergence estimator

Here we concentrate on the asymptotic properties of our proposed estimator. As mentioned, we will focus on the discrete set-up throughout the rest of the paper. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from an unknown distribution $G$ with support $\chi = \{0, 1, 2, 3, \ldots\}$. We consider a parametric family of distributions $F = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$, also supported on $\chi$, to model $G$. We assume both $G$ and $F$ to have densities $g$ and $f_\theta$ with respect to the counting measure. Let $\theta^g = T_{\alpha, \beta, \lambda}(G)$ be the best fitting parameter. Let $r_n(x)$ be the relative frequency of the value $x$ in the sample; we will use the vector of relative frequencies as an estimate of $g$ throughout this paper. The proof of the asymptotic normality of our estimators, like any proof of this nature, depends on a bunch of regularity conditions. These conditions, which are primarily structured upon those of [5], [18] and [19], are listed as (A1) – (A7) in the Supplementary Material. The minimum GSB divergence estimator is obtained as a root of the estimating equation

$$
\sum_{x=0}^{\infty} \left\{ A \beta^2 e^{\beta f_0(x)} f_0^A(x) + \frac{A^*}{A} f_0^B(x) \right\} \left( f_0^A(x) - \hat{g}^A(x) \right) u_\theta(x) = 0
$$

$$
\Rightarrow \sum_{x=0}^{\infty} K(\delta(x)) \left( A^2 \beta^2 e^{\beta f_0(x)} f_0^{2A}(x) + A^* f_0^{A^*}(x) \right) u_\theta(x) = 0, \tag{21}
$$

where, $\delta(x) = \delta_{\hat{g}}(x) = \frac{\hat{g}(x)}{f_0(x)} - 1 = \frac{r_n(x)}{f_0(x)} - 1$, $K(\delta) = \frac{(\delta + 1)^{A-1}}{A}$ and $u_\theta(x)$ is the likelihood score function at $x$. We denote the minimum GSB divergence estimator obtained as a solution of the above equation as $\hat{\theta}$. Let

$$
I_g = I_{\alpha, \beta, A}(g)
$$

$$
= E_g \left( u_{\theta \delta}(X) u_{\theta \delta}^T(X) K' \left( \delta_{g}(X) \right) \left( A^* f_0^{A^*}(X) + A^2 \beta^2 e^{\beta f_0(x)} f_0^{2A-1}(X) \right) \right)
$$

$$
+ \sum_{x=0}^{\infty} K(\delta_{g}(x)) \left( A^2 \beta^2 e^{\beta f_0(x)} f_0^{2A}(x) + A^* f_0^{A^*}(x) \right) i_{\theta \delta}(x)
$$
where, \( X \) is a random variable having density \( g \), \( \text{Var}_g \) represents variance under the density \( g \), \( \delta_g(x) = \frac{g(x)}{f_0(x)} - 1 \), \( K'() \) is the derivative of \( K() \) with respect to its argument, \( \delta_g(x) = \frac{g(x)}{f_0(x)} - 1 \) and \( J_g(x) = -u_0'(x) \), the negative of the derivative of the score function with respect to the parameter.

**Theorem 5.1:** Under the above-mentioned set-up and under the regularity conditions (A1)–(A7), there exists a consistent sequence of roots \( \hat{\theta}_n \) of the estimating Equation (18). Moreover, the asymptotic distribution of \( \sqrt{n}(\hat{\theta}_n - \theta^*) \) is \( p \)-dimensional normal with mean 0 and \( J_g^{-1} \text{Var}_g J_g^{-1} \).

**Corollary 5.2:** When \( g = f_0 \) for some \( \theta \in \Theta \), then \( \sqrt{n}(\hat{\theta}_n - \theta) \sim N(0, J^{-1}VJ^{-1}) \) asymptotically, where,

\[
J = E_{f_0} \left\{ u_0(X) u_0^T(X) \left( A^* f_0^\alpha(x) + A^2 \beta^2 e^{\beta f_0^A(x)} f_0^{2A-1}(x) \right) \right\}
\]

\[
= \sum_{x=0}^\infty \left\{ u_0(x) u_0^T(x) \left( A^* f_0^\alpha(x) + A^2 \beta^2 e^{\beta f_0^A(x)} f_0^{2A-1}(x) \right) \right\} f_0(x)
\]

\[
V = \text{Var}_{f_0} \left\{ u_0(X) \left( A^* f_0^\alpha(X) + A^2 \beta^2 e^{\beta f_0^A(X)} f_0^{2A-1}(X) \right) \right\}
\]

\[
= A^* \sum_{x=0}^\infty f_0^{1+2\alpha}(x) u_0(x) u_0^T(x) + A^4 \beta^4 \sum_{x=0}^\infty e^{2\beta f_0^A(x)} f_0^{4A-1}(x) u_0(x) u_0^T(x) + 2A^* A^2 \beta^2 \sum_{x=0}^\infty e^{\beta f_0^A(x)} f_0^{2A+\alpha}(x) u_0(x) u_0^T(x) - \zeta \zeta',
\]

where, \( \zeta = \sum_{x=0}^\infty u_0(x)(A^* f_0^\alpha(x) + A^2 \beta^2 e^{\beta f_0^A(x)} f_0^{2A}(x)) \).

### 6. Influence analysis of the minimum GSB divergence estimator

Here we study the stability of our estimators through their influence function (IF), one of the most important heuristic tools of robustness. A simple differentiation of a contaminated version of Equation (18) leads to the expression

\[
\text{IF}(y, G, T_{\alpha, \beta}) = J_G^{-1} N_G(y),
\]
where,

\[
N_G (y) = \left( A^2 \beta^2 e^{\beta f^A_{\theta \theta}} (y) f^A_{\theta \theta} (y) + A^* f^B_{\theta \theta} (y) \right) g^{A-1} (y) u_{\theta \theta} (y) \\
- \sum_{x=0}^{\infty} \left( A^2 \beta^2 e^{\beta f^A_{\theta \theta}} (x) f^A_{\theta \theta} (x) + A^* f^B_{\theta \theta} (x) \right) g^A (x) u_{\theta \theta} (x),
\]

\[
J_G = A^2 \beta^2 \sum_{x=0}^{\infty} e^{\beta f^A_{\theta \theta}} (x) f^A_{\theta \theta} (x) \left( 2 f^A_{\theta \theta} (x) - g^A (x) \right) u_{\theta \theta} (x) u_{\theta \theta}^T (x)
+ A^* \sum_{x=0}^{\infty} f^B_{\theta \theta} (x) \left( A^* f^A_{\theta \theta} (x) - B g^A (x) \right) u_{\theta \theta} (x) u_{\theta \theta}^T (x)
+ A^2 \beta^2 \sum_{x=0}^{\infty} e^{\beta f^A_{\theta \theta}} (x) f^A_{\theta \theta} (x) \left( g^A (x) - f^A_{\theta \theta} (x) \right) i_{\theta \theta} (x)
- A^* \sum_{x=0}^{\infty} f^B_{\theta \theta} (x) \left( f^A_{\theta \theta} (x) - g^A (x) \right) i_{\theta \theta} (x).
\]

If the distribution \( G \) belongs to the model family \( \mathcal{F} \) with \( g = f_\theta \), then the influence function reduces to,

\[
IF (y, F_\theta, T_{\alpha, \lambda, \beta}) = J_{F_\theta}^{-1} N_{F_\theta} (y), \quad \text{where,}
\]

\[
J_{F_\theta} = \sum_{x=0}^{\infty} \left( A^2 \beta^2 e^{\beta f^A_{\theta \theta}} (x) f^2_{\theta \theta} (x) + A^* f^A_{\theta \theta} (x) \right) u_\theta (x) u_\theta^T (x),
\]

\[
N_{F_\theta} (y) = A^2 \beta^2 e^{\beta f^A_{\theta \theta}} (y) f^2_{\theta \theta} \left( y u_\theta (y) + A^* f^A_{\theta \theta} \left( y u_\theta (y) \right) \right)
- \sum_{x=0}^{\infty} A^2 \beta^2 e^{\beta f^A_{\theta \theta}} (x) f^2_{\theta \theta} (x) u_\theta (x) - \sum_{x=0}^{\infty} A^* f^A_{\theta \theta} (x) u_\theta (x). \quad (26)
\]

The influence function depends on all three tuning parameters. When the matrix \( J_{F_\theta} \) is non singular, the boundedness of the influence function depends on controlling the score \( u_\theta (y) \) in the first two terms of the numerator. In most parametric models including all exponential family models, \( f^A_{\theta} (y) u_\theta (y) \) remains bounded for any \( \tau > 0 \); for \( \tau = 0 \) this equals \( u_\theta (y) \) which is no longer bounded. For the second term of the numerator in Equation (26), boundedness is achieved when \( A^* > 1 \), i.e., when \( \alpha > 0 \). As \( f^A_{\theta} (y) \leq 1 \) for any \( y \) in the support of a discrete random variable, the first term of the numerator is bounded for any real \( \beta \) when \( 2A - 1 > 0 \), i.e., \( A > 1/2 \). Here we list the different possible cases for boundedness of the influence function.
(1) $\beta = 0$; here the first and third terms of the numerator vanish, and the only condition necessary is $A^* > 1$, i.e., $\alpha > 0$. This is the $S$-divergence case, and all minimum $S$-divergence functionals with $\alpha > 0$ have bounded influence. Here the allowable region for the triplet $(\alpha, \lambda, \beta)$ for bounded influence is $S_1 = (\alpha > 0, \lambda \in \mathbb{R}, \beta = 0)$.

(2) $\beta \neq 0, A = 0$. In this case also the first and third terms of the numerator drop out and the additional required condition is $\alpha > 0$. However, since $A = 1 + \lambda(1 - \alpha) = 0$, this implies $\lambda = -\frac{1}{1-\alpha}$. In this case the influence function is independent of $\beta$. Now the relevant region for the triplet is $S_2 = (\alpha > 0, \lambda = -\frac{1}{1-\alpha}, \beta \neq 0)$.

(3) Now suppose $A^* = 0$, without the components being individually zero. In this case the second and fourth terms get eliminated and we have $\alpha = -1$. In this case the condition $2A - 1 > 0$ translates to $\lambda > -\frac{1}{4}$. Here the corresponding region for the triplet is $S_3 = (\alpha = -1, \lambda \geq -\frac{1}{4}, \beta \neq 0)$.

(4) Now we allow all the terms $\beta$, $A$ and $A^*$ to be non-zero. In this case all the four terms of the numerator are non-vanishing. Then, beyond the condition on $\beta$, the required conditions are $\alpha > 0$ and $\lambda(1 - \alpha) > -\frac{1}{2}$. The region here is $S_4 = (\alpha > 0, \lambda(1 - \alpha) > -\frac{1}{2}, \beta \neq 0)$.

Combining all the cases, we see that the IF will be bounded if the triplet $(\alpha, \lambda, \beta) \in S = S_1 \cup S_2 \cup S_3 \cup S_4$. It is easily seen that the four constituent subregions are disjoint. We present some plots for bounded and unbounded influence functions for the minimum GSB functional under the Poisson($\theta$) model in Figure 1; the true data distribution is Poisson(3).

7. Simulation results

In this section we aim to demonstrate that by choosing non-zero values of the parameter $\beta$, we may be able to generate estimators that improve upon those provided by the existing standard, the class of $S$-divergences. In case of the latter family, Ghosh et al. [13] have empirically identified an elliptical subset of the tuning parameter space, with $\alpha \in [0.1, 0.6]$ and $\lambda \in [-1, -0.3]$, which represent good choices. We hope to show that for most $(\alpha, \lambda)$ choices (including the best ones) there is a better or competitive $(\alpha, \lambda, \beta)$ combination with a non-zero $\beta$, at least to the extent of the findings in these simulations. In the following we explain the simulation exercise systematically.

(1) The model, the sample size, the contamination and the replication: We consider the Poisson($\theta$) model, and choose samples of size 50 from the $(1 - \epsilon)\text{Poisson}(3) + \epsilon\text{Poisson}(10)$ mixture. The second component is the contaminant and $\epsilon \in [0, 1)$ is the contaminating proportion. The values 0, 0.05, 0.1 and 0.2 are considered for $\epsilon$, and at each contamination level, the samples are replicated 1000 times.
Figure 1. Examples of unbounded influence functions (left panel) and bounded influence functions (right panel) corresponding to $(\alpha, \lambda, \beta) \in$ each disjoint subsets contained in $S$.

(2) Minimum $S$-divergence estimator: As these are the basis of comparison, the minimum $S$-divergence estimators of the Poisson parameter are calculated over a cross-classified grid with $\alpha$ values in 0.1, 0.25, 0.4, 0.5, 0.6, 0.8, 1 and $\lambda$ values in $\{-1, -0.7, -0.5, -0.3, 0, 0.2, 0.5, 0.8, 1\}$, a total of 63 combinations, at each contamination level. The mean square errors (MSEs) are calculated against the target value of
3 over the 1000 replications in each case. The MSEs are presented in Table 3, where, in each cell, the empirical MSEs for $\epsilon = 0, 0.05, 0.1$ and 0.2 are presented in a column of four elements, in that order, followed by the corresponding combination of tuning parameters $(\alpha, \lambda, \beta = 0)$. We have included the parameter $\beta = 0$ parameter in each triplet, to indicate that the $S$-divergence is indeed a special case of the GSB divergence.

Note that there is no unique $(\alpha, \lambda)$ combination which produces an overall best result in terms of smallest MSE over all four levels of contamination.

(3) The minimum GSB divergence estimator: We now expand the exploration by considering, in addition, a grid of possible non-zero $\beta$ values at each $(\alpha, \lambda)$ combination to see if the results can be improved. To be conservative about our definition of improvement, we declare the existence of a ‘better’ triplet in the GSB sense if all the four MSEs corresponding to a $(\alpha, \lambda)$ combination within the $S$-divergence family in Table 3 are reduced by a suitable member of the GSB class strictly outside the $S$-divergence family (corresponds to a non-zero $\beta$).

Our exploration indicates that in a large majority of the 63 cells in Table 3 there is a minimum GSB divergence estimator with a non-zero $\beta$ value which improves (over all the four cells) the performance of the corresponding minimum $S$-divergence estimator. Interestingly it turns out that in practically all the cases where an improvement is observed it happens for a negative value of $\beta$ (it is observed to be zero in rare cases, but is never positive). A more detailed inspection indicates that in many of these cases, the improvement occurs at the value $\beta = -4$. In order to summarize the findings of this rather large exploration (presented in Table 4) in a systematic and meaningful manner, we first note the following different cases,

(a) (First Case) These are the cells where all the four mean square errors for the $S$-divergence case are reduced by the minimum GSB divergence estimator with the same values of $(\alpha, \lambda)$ and $\beta = -4$. These cells are highlighted with the blue colour in Table 4. (There are 18 such cells).

(b) (Second Case) These are the cells where all the four MSEs for the $S$-divergence case are reduced by a minimum GSB divergence estimator with $\beta = -4$ but with a different $(\alpha, \lambda)$ combination. These cells are highlighted in red in Table 4. (There are 39 such cells).

(c) (Third Case) These are the cells where all the four MSEs are reduced by a minimum GSB divergence estimator outside the $S$-divergence family, but with $\beta \neq -4$, and not necessarily the same $(\alpha, \lambda)$. These cells are highlighted in orange in Table 4. (There is one such cell).

(d) (Fourth Case) These are the cells where some triplet within the minimum GSB divergence class can improve upon the three MSEs under contamination $(\epsilon = 0.05, 0.1, 0.2)$ but not all the four MSEs simultaneously. While these are not ‘better’ triplets in the sense described earlier in the section, the pure data MSEs (not reported here) for these triplets are close to those of the $S$-divergence MSEs for these cells; in this sense these triplets are at least competitive. These cells are highlighted in green in Table 4. (There are three such cases).
Table 3. MSEs of the minimum divergence estimators within the $S$-divergence family for pure and contaminated data.

| $(\alpha, -1, 0)$ | MSE         | MSE         | MSE         | MSE         | MSE         | MSE         | MSE         |
|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| $(0.1, -0.7, 0)$  | 0.0751      | 0.0666      | 0.0673      | 0.0698      | 0.0729      | 0.0799      | 0.0876      |
| $(0.1, -0.3, 0)$  | 0.0592      | 0.0617      | 0.0657      | 0.0688      | 0.0721      | 0.0796      | 0.0876      |
| $(0.1, 0.0)$      | 0.0608      | 0.0621      | 0.0660      | 0.0691      | 0.0725      | 0.0798      | 0.0876      |
| $(0.1, 0.2, 0)$   | 0.0671      | 0.0638      | 0.0658      | 0.0685      | 0.0718      | 0.0792      | 0.0876      |
| $(0.1, 0.5, 0)$   | 0.0778      | 0.0676      | 0.0665      | 0.0685      | 0.0716      | 0.0790      | 0.0876      |
| $(0.1, 0.8, 0)$   | 0.0863      | 0.0717      | 0.0673      | 0.0686      | 0.0714      | 0.0789      | 0.0876      |
| $(0.1, 1.0)$      | 0.0944      | 0.0837      | 0.0803      | 0.0828      | 0.0859      | 0.0843      | 0.0876      |

On the whole, therefore, it turns out that we observe improvements in a vast majority of the 63 cells in all four rows of the column of MSEs in that cell by choosing $\beta = -4$. Even in the handful of cases (cells) where we do not have an improvement in all the rows of the
Table 4. MSEs of the minimum GSB divergence estimators under pure and contaminated data.

| α   | λ       | β   | MSE under pure data | MSE under contaminated data |
|-----|---------|-----|----------------------|-----------------------------|
| 0.0623 | 0.0696  | 0.0704 | 0.0708 | 0.0687 | 0.0720 | 0.0763 |
| 0.0816 | 0.0843  | 0.0855 | 0.0852 | 0.0833 | 0.0859 | 0.0892 |
| 0.1115 | 0.1056  | 0.1012 | 0.1028 | 0.1042 | 0.1060 | 0.1077 |
| 0.2831 | 0.2207  | 0.2119 | 0.2113 | 0.2110 | 0.2162 | 0.2115 |
| 0.0642 | 0.0681  | 0.0681 | 0.0696 | 0.0696 | 0.0720 | 0.0763 |
| 0.0816 | 0.0826  | 0.0826 | 0.0843 | 0.0843 | 0.0859 | 0.0892 |
| 0.1115 | 0.1043  | 0.1043 | 0.1055 | 0.1055 | 0.1060 | 0.1077 |
| 0.2514 | 0.2135  | 0.2135 | 0.2207 | 0.2207 | 0.2162 | 0.2115 |
| 0.0623 | 0.0623  | 0.0659 | 0.0678 | 0.0678 | 0.0696 | 0.0763 |
| 0.0816 | 0.0816  | 0.0825 | 0.0834 | 0.0834 | 0.0843 | 0.0892 |
| 0.1115 | 0.1115  | 0.1071 | 0.1061 | 0.1061 | 0.1055 | 0.1077 |
| 0.2831 | 0.2831  | 0.2417 | 0.2295 | 0.2295 | 0.2207 | 0.2115 |
| 0.0600 | 0.0600  | 0.0644 | 0.0644 | 0.0644 | 0.0755 | 0.0763 |
| 0.0845 | 0.0822  | 0.0819 | 0.0825 | 0.0834 | 0.0859 | 0.0892 |
| 0.1294 | 0.1154  | 0.1091 | 0.1071 | 0.1061 | 0.1060 | 0.1077 |
| 0.4049 | 0.3080  | 0.2602 | 0.2417 | 0.2295 | 0.2162 | 0.2115 |
| 0.0600 | 0.0600  | 0.0644 | 0.0644 | 0.0644 | 0.0779 | 0.0763 |
| 0.0845 | 0.0845  | 0.0817 | 0.0817 | 0.0817 | 0.0887 | 0.0892 |
| 0.1294 | 0.1294  | 0.1073 | 0.1073 | 0.1073 | 0.1077 | 0.1077 |
| 0.4049 | 0.4049  | 0.2492 | 0.2492 | 0.2492 | 0.2139 | 0.2115 |
| 0.0600 | 0.0600  | 0.0644 | 0.0644 | 0.0644 | 0.0779 | 0.0763 |
| 0.0845 | 0.0845  | 0.0817 | 0.0817 | 0.0817 | 0.0907 | 0.0892 |
| 0.1294 | 0.1294  | 0.1073 | 0.1073 | 0.1073 | 0.1092 | 0.1077 |
| 0.4049 | 0.4049  | 0.2492 | 0.2492 | 0.2492 | 0.2146 | 0.2115 |
| 0.0600 | 0.0600  | 0.0644 | 0.0644 | 0.0644 | 0.0696 | 0.0763 |
| 0.0845 | 0.0845  | 0.0817 | 0.0817 | 0.0817 | 0.0843 | 0.0892 |
| 0.1294 | 0.1294  | 0.1073 | 0.1073 | 0.1073 | 0.1055 | 0.1077 |
| 0.4049 | 0.4049  | 0.2492 | 0.2492 | 0.2492 | 0.2207 | 0.2115 |
| 0.0600 | 0.0600  | 0.0644 | 0.0644 | 0.0644 | 0.0696 | 0.0763 |
| 0.0845 | 0.0845  | 0.0817 | 0.0817 | 0.0817 | 0.0843 | 0.0892 |
| 0.1294 | 0.1294  | 0.1073 | 0.1073 | 0.1073 | 0.1055 | 0.1077 |
| 0.4049 | 0.4049  | 0.2492 | 0.2492 | 0.2492 | 0.2207 | 0.2115 |

In each cell of Table 4, we also present the particular (α, λ, β) combination which generates the mean square errors reported in that cell. In Figure 2, we provide a three-dimensional plot in the three-dimensional (α, λ, β) plane, where the region S has been expressed as a union of several colour-coded subregions representing the individual components. The triplets corresponding to the solutions reported in the cells of Table 4 all belong to the blue subregion of this figure, indicating that the improvements are provided by bounded influence estimators.
8. Selection of tuning parameters

While the GSB divergence provides several good options for parameter estimation, some guidance is needed on the method to be chosen in a given situation. One must have a data-based method for selecting the optimal procedure, where a higher proportion of data anomaly is dealt with a more robust member of the GSB class.

The works of [20–22] provide some algorithms for choosing data-driven ‘optimal’ tuning parameters for the DPD, which we denote as the HK (for Hong and Kim), the OWJ (for one-step Warwick-Jones) and the IWJ (for the iterated Warwick-Jones) algorithms, respectively. The idea is to construct an empirical approximation to the mean square error as a function of the tuning parameters (and a pilot estimator) and minimize it over the former. The IWJ algorithm [22] refines the OWJ algorithm [21] by extending the one-step process to a converged iterative process, thus eliminating the dependence on the pilot estimator (subject to a robust starting value). The HK algorithm ignores the bias part of the MSE and occasionally throws up highly non-robust solutions. We will implement these algorithms here for the GSB parameters. However, while we use the method proposed in [22] as a tool, there is no overlap between the present paper and [22] in terms of methodological development.

Here we take up two real data examples and consider the problem of selecting the ‘optimal’ tuning parameters in each case. The OWJ and IWJ algorithms considered here use the minimum $L_2$ distance estimator as the pilot. (The latter algorithm is pilot-independent, but for computational purposes it needs to commence with a suitable pilot). While the IWJ algorithm is our preferred method, we demonstrate the use of all the three algorithms in the following data sets.

Figure 2. The figure shows the region needed for bounded IF. Here, the grey, the orange, the green and the blue planes represent the boundaries of the sets $S_1, S_2, S_3$ and $S_4$, respectively.
Figure 3. Some significant fits for the Stolen Bases Data under the Poisson model. Here ‘clean data’ refer to the modified data after removing all 9 outliers.

Example 8.1 (Peritonitis Data): This example involves the incidence of peritonitis in 390 kidney patients. The data are presented in the Online Supplement. The observations at 10 and 12 may be regarded as mild outliers. A geometric model with success probability $\theta$ has been fitted to these frequency data. Here, the IWJ solution coincides with the HK solution where the estimate of success probability is 0.5110 corresponding to $(\alpha, \lambda, \beta) = (0.41, -0.84, -3.5)$. The OWJ solution gives a slightly different success probability of 0.5105 corresponding to $(\alpha, \lambda, \beta) = (0.17, -0.60, -3)$. In case of clean data these IWJ, OWJ and HK estimates will be 0.5044, 0.5061 and 0.5029 corresponding to $(\alpha, \lambda, \beta) = (0.47, -1, -2), (0.29, -1, -1)$ and $(0.55, -1, -3)$, respectively, being slightly different from each other. On the contrary, the MLEs for the full dataset and the (two) outlier deleted dataset are 0.4962 and 0.5092, respectively.

Now we consider a more recent dataset for the implementation of our new proposal.

Example 8.2 (Stolen Bases Data): In ‘Major League Baseball (MLB) Player Batting Stats’ for the 2019 MLB Regular Season, obtained from ESPN.com, one variable of interest is the number of Stolen Bases (SB) awarded to the top 40 Home Run (HR) scorers of the American League (AL). These data, containing three extreme and six moderate outliers, could be well-modelled by the Poisson distribution if not for the outliers. We wish to estimate $\theta$, the average number of Stolen Bases (SB) awarded to the MLB batters of the AL throughout the regular season. The ‘optimal’ estimates, derived from the implementation of the three algorithms under the Poisson model, are presented in Table 5. The fitted curves for some of these optimal estimates are given in Figure 3. It is clear that except for the full data MLE, all the other estimators primarily model the main bulk of the data and sacrifice the outliers.
Table 5. Optimal estimates in different cases for the Stolen Bases Data.

| data                      | method | optimal $\hat{\theta}$ | optimal $(\alpha, \lambda, \beta)$ |
|---------------------------|--------|-------------------------|-------------------------------------|
| Full data (with outliers) | IWJ    | 2.6270                  | $(0.65, -0.98, -8)$                  |
|                           | OWJ    | 2.5086                  | $(0.73, -1, -8)$                     |
|                           | HK     | 2.6409                  | $(0.65, -1, -8)$                     |
|                           | MLE    | 4.875                   | $(0, 0, 0)$                          |
| excluding 9 outliers      | IWJ    | 2.3949                  | $(0.01, 1, 0)$                       |
|                           | OWJ    | 2.3229                  | $(0.25, 1, 0)$                       |
|                           | HK     | 2.6918                  | $(0.45, -1, -8)$                     |
|                           | MLE    | 2.3871                  | $(0, 0, 0)$                          |

9. Concluding remarks

In this paper, we have provided an extension of the ordinary Bregman divergence which has direct applications to developing new classes of divergence measures, and, in turn, in providing more options for minimum distance inference with better mixes of model efficiency and robustness. In the second part of the paper, we have made use of the suggested approach in generating a particular super-family of divergences which seems to work very well in practice and provides new minimum divergence techniques which appear to improve the performance of the S-divergence based procedures in many cases. Since the results presented here are based on a single study, more research will certainly be needed to decide to what extent the observed advantages of the procedures considered here can be generalized, but clearly there appears to be enough evidence to suggest such explorations are warranted.

Apart from the search for other divergences, several possible follow ups of this research immediately present themselves. This paper is restricted to discrete parametric models. An obvious follow up step is to suitably handle the case of continuous models, where the construction of the divergence is much more complicated. Another extension will be to extend the procedures to more complicated data structures beyond the independently and identically distributed data scenarios. Yet another extension would be to apply this and other similarly developed super-divergences in the area of robust testing of hypothesis. We hope to take up all of these in our future work.

The GSB subfamily with $\beta = -4$ also needs some attention, and we hope to take it up in the future. For the time being we have presented the results for our real data examples for the $\beta = -4$ subfamily of GSB in the Online Supplement.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Csiszár I. Eine informations theoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von Markoffschen ketten. Publ Math Inst Hungar Acad Sci. 1963;3:85–107.
[2] Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. J R Stat Soc B. 1966;28:131–142.
[3] Bregman LM. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput Math Math Phys. 1967;7:200–217.
[4] Beran R. Minimum Hellinger distance estimates for parametric models. Ann Statist. 1977;5:445–463.

[5] Lindsay BG. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann Statist. 1994;22:1081–1114.

[6] Basu A, Shioya H, Park C. Statistical inference: the minimum distance approach. Boca Raton (FL): CRC Press; 2011.

[7] Hampel FR, Ronchetti EM, Rousseeuw PJ, et al. Robust statistics: the approach based on influence functions. New York (NY): John Wiley and Sons; 1986.

[8] Huber PJ, Ronchetti EM. Robust statistics. Hoboken (NJ): John Wiley and Sons; 2009.

[9] Maronna RA, Martin RD, Yohai VJ. Robust statistics. Chichester: John Wiley and Sons; 2006.

[10] Basu A, Harris IR, Hjort NL, et al. Robust and efficient estimation by minimising a density power divergence. Biometrika. 1998;85:549–559.

[11] Cressie N, Read TRC. Multinomial goodness-of-fit tests. J R Stat Soc B. 1984;46:440–464.

[12] Mukherjee T, Mandal A, Basu A. The B-exponential divergence and its generalizations with applications to parametric estimation. Statist Methods Appl. 2019;28:241–257.

[13] Ghosh A, Harris IR, Maji A, et al. A generalized divergence for statistical inference. Bernoulli. 2017;23:2746–2783.

[14] Banerjee A, Merugu S, Dhillon IS, et al. Clustering with Bregman divergences. J Mach Learn Res. 2005;6:1705–1749.

[15] Amari S. Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Trans Inf Theory. 2009;55:4925–4931.

[16] Gutmann M, Hirayama J. Bregman divergence as general framework to estimate unnormalized statistical models. arXiv preprint arXiv:1202.3727. 2012.

[17] Jana S, Basu A. A characterization of all single-integral, non-kernel divergence estimators. IEEE Trans Inf Theory. 2019;65:7976–7984.

[18] Lehmann EL. Theory of point estimation. New York (NY): John Wiley & Sons; 1983.

[19] Ghosh A. Asymptotic properties of minimum S-divergence estimator for discrete models. Sankhya A. 2015;77:380–407.

[20] Hong C, Kim Y. Automatic selection of the tuning parameter in the minimum density power divergence estimation. J Korean Stat Soc. 2001;30:453–465.

[21] Warwick J, Jones MC. Choosing a robustness tuning parameter. J Stat Comput Simul. 2005;75:581–588.

[22] Basak S, Basu A, Jones MC. On the ‘optimal’ density power divergence tuning parameter. J Appl Stat. 2021;48:536–556.