THE GENERALIZED KULLBACK-LEIBLER DIVERGENCE AND ROBUST INFERENCE

CHANSEOK PARK[a,*] and AYANENDRANATH BASU[b,†]

[a]*Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA;*
[b]*Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India*

(Received 9 April 2001; In final form 11 July 2002)

This paper examines robust techniques for estimation and tests of hypotheses using the family of generalized Kullback-Leibler (GKL) divergences. The GKL family is a new group of density based divergences which forms a subclass of disparities defined by Lindsay (1994). We show that the corresponding minimum divergence estimators have a breakdown point of 50% under the model. The performance of the proposed estimators and tests are investigated through an extensive numerical study involving real-data examples and simulation results. The results show that the proposed methods are attractive choices for highly efficient and robust methods.

Keywords: Disparity; Breakdown point; Empty cell; Pearson residual; Residual adjustment function

1 INTRODUCTION

We consider the general setting of inference under a parametric class $\mathcal{F}_\Theta = \{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$. Let G denote the true distribution belonging to $\mathcal{G}$, the class of all distributions having probability density functions (pdf's) with respect to a dominating measure. We will assume that the model $\mathcal{F}_\Theta$ is a subclass of $\mathcal{G}$. To deal with the general case we also assume, for discrete models, that the distributions have countable support $\{0, 1, 2, \ldots\}$. Throughout this paper we will let the corresponding lower case letters denote the pdf's of the cumulative distribution functions (CDFs), *e.g.*, the pdf's of G, F_θ and G_n will be g, f_θ and g_n respectively.

In parametric estimation one wishes to estimate θ efficiently when the model is correct and robustly in case the true distribution is close to but not necessarily in it. Similarly, in testing of hypotheses it is desirable to have a procedure which has high power under the model simultaneously with high stability in terms of level and power under small departures from the model. Beran (1977) first demonstrated that the simultaneous goals of asymptotic efficiency and robustness can be achieved by using the minimum Hellinger distance estimator (MHDE). Other authors, such as Tamura and Boos (1986) and Simpson (1987; 1989) have further pursued this line of research and established other desirable properties of the MHDE.

[*] Corresponding author. E-mail: cspark@ces.clemson.edu
[†] His research was done while he was visiting Department of Statistics, Pennsylvania State University.

Lindsay (1994) generalized the work based on Hellinger distance (HD) to a general class of disparities generating estimators that are both robust and first order efficient. A disparity is a measure of discrepancy between a nonparametric density estimator and the model density. In this paper we develop a subclass of disparities called the generalized Kullback-Leibler (GKL) divergence. The inference properties of the corresponding minimum divergence procedures is the subject of the current study. The emphasis is on quantitative investigations – the efficiency and robustness properties of the procedures are studied through an extensive numerical study. However, the asymptotic breakdown point of the estimators is also theoretically established. A sequel paper will deal with the remaining theoretical properties.

The rest of the paper is organized as follows: Section 2 provides a brief introduction to minimum disparity estimation. In Section 3 we introduce the GKL family. In Section 4 we study the influence function, the breakdown point, and the GKL tests statistics. Examples and simulation results illustrating the performance of the procedures are given in Sections 5 and 6 respectively. Section 7 presents some concluding remarks.

2 MINIMUM DISPARITY ESTIMATION

For a random sample $X_1, X_2, \ldots, X_n$ from distribution G, let

$$g_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n}\right) \tag{1}$$

define a nonparametric density estimator of g, where w is a smooth family of *kernel* functions with bandwidth h_n. For discrete models, we take g_n to be the empirical density function, defined as $g_n(x)$ = the proportion of x-values in the sample for $x = 0, 1, 2, \ldots$. Let G_n be the CDF of g_n. Define the *Pearson residual* at a point x as

$$\delta_P(x) = \frac{g_n(x) - f_\theta(x)}{f_\theta(x)}. \tag{2}$$

We denote the Pearson residual δ_P by δ for brevity. Let $C(\cdot)$ be a real-valued, thrice differentiable convex function on $[-1, \infty)$ with $C(0) = 0$. Following Lindsay (1994), construct the *disparity* ρ_C (between g_n and f_θ) defined as

$$\rho_C(g_n, f_\theta) = \int C(\delta) f_\theta(x), \tag{3}$$

where the integral is with respect to the dominating measure. Under the assumptions the disparity ρ_C is nonnegative and equals zero if and only if $g_n \equiv f_\theta$. Under appropriate choices of disparities the "minimum disparity estimators" have attractive efficiency and robustness features. The class of disparities include the likelihood disparity (LD) and the squared Hellinger distance, defined by $\mathrm{LD}(g_n, f_\theta) = \int g_n(x) \log(g_n/f_\theta)$ and $\mathrm{HD}(g_n, f_\theta) = \int (g_n(x)^{1/2} - f_\theta(x)^{1/2})^2$ respectively. The LD is a version of the Kullback-Leibler divergence, and in the discrete case it is minimized by the maximum likelihood estimator (MLE) of θ; however, by the Kullback-Leibler divergence we will refer to $\mathrm{KL}(g_n, f_\theta) = \int f_\theta(x) \log(f_\theta(x)/g_n(x))$.

Let ∇ represent the gradient with respect to θ. For any real valued function $a(x)$ we will let $a'(x)$ and $a''(x)$ denote its first and second derivatives with respect to x. Minimization of the disparity $\rho_C(g_n, f_\theta)$ over $\theta \in \Theta$ gives the minimum disparity estimator corresponding to the $C(\cdot)$ function; the Hellinger distance produces the MHDE. Under differentiability of the model, the minimum disparity estimating equation becomes

$$-\nabla\rho_C = \int A(\delta)\nabla f_\theta(x) = 0, \tag{4}$$

where $A(\delta) \equiv (\delta + 1)C'(\delta) - C(\delta)$. The function $A(\delta)$ is an increasing function on $[-1, \infty)$, and without affecting the estimating properties of the disparity ρ_C, it can be redefined to satisfy $A(0) = 0$ and $A'(0) = 1$. This standardized function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity. The estimating equation for the likelihood disparity (the likelihood equation in the discrete case) is given by

$$-\nabla\rho_C = \int \delta\nabla f_\theta(x) = 0, \tag{5}$$

i.e. $A(\delta) = \delta$ for the likelihood disparity. Under the standardizations described above, the leading term (in a Taylor series expansion) of the minimum disparity estimating function in (4) matches that for the likelihood disparity, indicating that establishing the asymptotic efficiency of the minimum disparity estimators corresponds to demonstrating that the remainder term is "small" in the limit. Since the estimating equations of the minimum disparity estimators – as given in Eq. (4) – are otherwise equivalent, their distinctive features are governed by the form of their RAF. Thus, for example, RAFs for which $A(\delta) \ll \delta$ are able to strongly downweight the effect of large outlying observations (which manifest themselves as large positive values of δ) relative to maximum likelihood. The value $A_2 = A''(0)$ is called the curvature parameter (Lindsay, 1994) of the RAF, and is a measure of how fast the function curves away from the line $A(\delta) = \delta$ at $\delta = 0$. Large negative values of A_2 provide greater downweighting effect relative to maximum likelihood estimation, while $A_2 = 0$ indicates a form of second order efficiency of the estimator in the sense of Rao (1961;1962). For the likelihood disparity and the Hellinger distance A_2 equals 0 and $-1/2$ respectively.

3 THE GENERALIZED KULLBACK-LEIBLER DIVERGENCE

We introduce the new family of generalized Kullback-Leibler divergences (GKLs) between two densities $g(\cdot)$ and $f(\cdot)$ indexed by a single parameter $\tau \in [0, 1]$ as:

$$\mathrm{GKL}_\tau(g, f) = \int\left[\frac{g(x)}{\bar\tau}\log\left(\frac{g(x)}{f(x)}\right) - \left(\frac{g(x)}{\bar\tau} + \frac{f(x)}{\tau}\right)\log\left(\tau\frac{g(x)}{f(x)} + \bar\tau\right)\right], \quad \bar\tau = 1 - \tau,$$

which can also be written as

$$\begin{aligned}\mathrm{GKL}_\tau(g, f) &= \frac{1}{\tau\bar\tau}\int[\tau\varphi(g(x)) + \bar\tau\varphi(f(x)) - \varphi(\tau g(x) + \bar\tau f(x))]\\ &= \int D(g(x), f(x)),\end{aligned}$$

where $\varphi(\cdot)$ is defined to be

$$\varphi(t) = \begin{cases} 0; & t = 0 \\ t \log t; & t \in (0, 1] \end{cases}.$$

The divergences for $\tau = 0$ and $\tau = 1$ are defined by the limiting cases as $\tau \to 0$, and $\tau \to 1$ respectively. Notice that $D(g(x), f(x))$ is non-negative by the convexity of $\varphi(\cdot)$, and the divergence equals zero only when $g(x) \equiv f(x)$, identically. The divergences $\mathrm{GKL}_{\tau=0}$ and $\mathrm{GKL}_{\tau=1}$ are the likelihood disparity (LD) and the Kullback-Leibler divergence (KL) respectively. In Figure 1(a), we present the RAFs of several members of the GKL family. It is quite obvious how the RAF functions for large values of τ strongly downweight the effect of the large outliers. Under the notation of Section 2, we will denote by MGKLE_τ the minimum disparity estimator which minimizes $\mathrm{GKL}_\tau(g_n, f_\theta)$ over θ. The RAF $A_\tau(\cdot)$ and the $C_\tau(\cdot)$ function for the GKL_τ divergence are given by

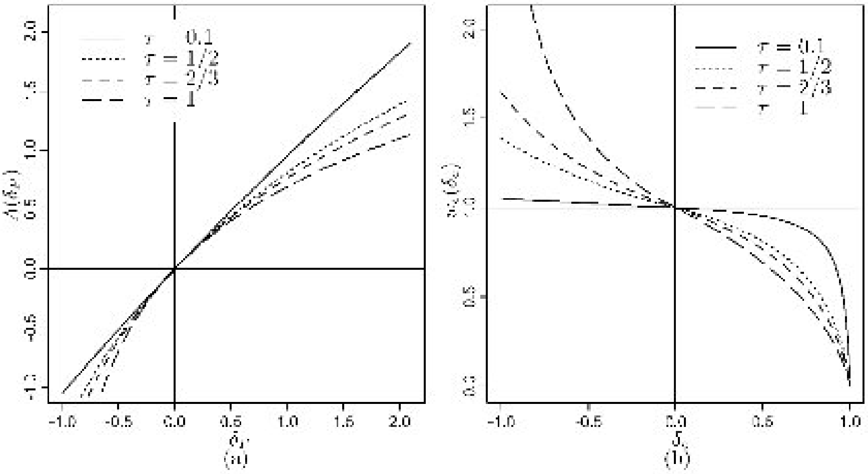
$$A_\tau(\delta) = \frac{1}{\tau}\log(\tau\delta + 1), \tag{6}$$

$$C_\tau(\delta) = \frac{\delta + 1}{1 - \tau}\log(\delta + 1) - \frac{\tau\delta + 1}{\tau(1 - \tau)}\log(\tau\delta + 1). \tag{7}$$

A motivation for the construction of the GKL family, which produces a smooth bridge between LD and KL may be provided as follows. It is easily seen that

$$\mathrm{GKL}_\tau(g, f) = \min_p\{\tau\mathrm{LD}(g, p) + (1 - \tau)\mathrm{KL}(p, f)\}$$

where the minimization is over the density p. The right hand side of the above equation is actually the solution of the likelihood ratio testing problem which minimizes the likelihood disparity $\mathrm{LD}(g, p)$ subject to $p \in B_f = \{p\colon \mathrm{KL}(p, f) \le c\}$, and τ is an appropriate function of c. A reversal of the roles of LD and KL generates the celebrated power divergence family of Cressie and Read (1984).



For better understanding the robustness of the methods based on the GKL divergence, we also present the combined weight function $w_c(\delta_c)$ (Park *et al.*, 2002) for the GKL_τ family for different values of τ in Figure 1(b). The combined weight function $w_c(\delta_c)$ represent the relative impact of the observation in the estimating equation compared to maximum likelihood. To keep the notation clear, we denote the Pearson residual δ by δ_P, and define the Neyman residual $\delta_N(x) = [g_n(x) - f_\theta(x)]/g_n(x)$. The combined residual δ_c is defined as

$$\delta_c(x) = \begin{cases} \delta_P(x); & d \le f_\theta \\ \delta_N(x); & d > f_\theta \end{cases}$$

and the combined weight function $w_c(\delta_c)$ is

$$w_c(\delta_c) = \begin{cases} \dfrac{A(\delta_c)}{\delta_c}; & -1 \le \delta_c < 0 \\ A'(0); & \delta_c = 0 \\ \dfrac{1-\delta_c}{\delta_c} A\left(\dfrac{\delta_c}{1-\delta_c}\right); & 0 < \delta_c < 1 \\ A'(\infty); & \delta_c = 1 \end{cases} . \tag{8}$$

On the positive side of the δ_c axis, this amounts to looking at the weights as a function of the Pearson residuals but in the Neyman scale. For better robustness, it is desirable that the weight functions converge to 0 as $\delta_c \to 1$ which happens for all the members of the GKL family considered in this figure. Notice that the graphs of the combined weight function are defined over a bounded interval of values of δ_c (unlike the RAF), and hence allows the graphical investigation of the method at either end of the range.

We conclude the section with the following boundedness properties of the divergences which are proved in the Appendix. The boundedness results are also useful in establishing the breakdown properties of the corresponding estimators later in Section 4.

LEMMA 1 *Denote* $D(g,f) = \dfrac{1}{\tau\bar{\tau}}[\tau\varphi(g) + \bar{\tau}\varphi(f) - \varphi(\tau g + \bar{\tau} f)], \bar{\tau} = 1 - \tau$. *Then*

$$0 \le D(g,f) \le D(0,f)\mathbb{I}(g \le f) + D(g,0)\mathbb{I}(f < g),$$

where $\mathbb{I}(\cdot)$ *is the indicator function.*

THEOREM 2 *The* GKL_τ *divergence is bounded for* $0 < \tau < 1$. *In particular*

$$0 \le \mathrm{GKL}_\tau(g,f) \le \frac{1}{\tau}\log\left(\frac{1}{\bar{\tau}}\right) + \frac{1}{\bar{\tau}}\log\left(\frac{1}{\tau}\right).$$

The left equality holds when $g(\cdot) \equiv f(\cdot)$, *and the right equality holds when the distributions are singular, i.e.* $\{x : f(x) > 0\} \cap \{x : g(x) > 0\}$ *is a set of measure 0 with respect to the dominating measure.*

4 INFLUENCE FUNCTION, BREAKDOWN RESULTS, AND DISPARITY TESTS

4.1 Influence Function and Standard Error

Consider the contaminated version of the true density g defined by $g_\varepsilon(x) = (1-\varepsilon)g(x) + \varepsilon\mathbb{I}_\xi(x)$, where ε is the contamination proportion and $\mathbb{I}_\xi(x)$ represents the indicator function for the set containing only ξ. Let G and G_ε be the corresponding distributions. Suppose $T_\tau(\cdot)$ is the MGKLE$_\tau$ functional. Its influence function $\mathrm{IF}_{\tau,G}$ at G is defined by

$$\mathrm{IF}_{\tau,G}(\xi) = \left.\frac{\partial T_\tau(G_\varepsilon)}{\partial_\varepsilon}\right|_{\varepsilon=0}.$$

A straightforward differentiation of the estimating Eq. (4) shows the influence function of T_τ to be

$$\mathrm{IF}_{\tau,G}(\xi) = \left[\int A'_\tau(\delta)u_\theta u_\theta^T g - \int A_\tau(\delta)\nabla^2 f_\theta\right]^{-1}\left[A'_\tau(\delta(\xi))u_\theta(\xi) - \int A'_\tau(\delta)u_\theta g\right],$$

where $\theta = T_\tau(G)$, $u_\theta(x) = \nabla \log f_\theta(x)$, $\delta(x) = (g(x) - f_\theta(x))/f_\theta(x)$, and $A_\tau(\cdot)$ is as in (6). As an immediate consequence of the above result we note that if G is a model point F_θ, then the influence function of the T_τ functional reduces to $I^{-1}(\theta)u_\theta(\xi)$ suggesting that the MGKLE$_\tau$ is asymptotically fully efficient at the model.

On the other hand, being equal to the influence function of the MLE, the influence function of the T_τ estimator is potentially unbounded. Thus the robustness of the MGKLE$_\tau$ cannot be described through the traditional bounded influence approach. Beran (1977) claims that for evaluating the robustness of a functional with respect to a gross-error model, one should consider the *α-influence function* instead of the influence function, unless the former converges to the latter uniformly. Let $f_{\theta,\alpha,z} \equiv (1-\alpha)f_\theta + \alpha\eta_z$, where η_z denotes the uniform density on the interval $(z-\varepsilon, z+\varepsilon)$, $\varepsilon > 0$ arbitrarily small, $\theta \in \Theta$, $\alpha \in (0,1)$, $z \in \mathbb{R}$. Let $F_{\theta,\alpha,z}$ denote the corresponding CDF. The α-influence function is the difference quotient $\alpha^{-1}[T_\tau(F_{\theta,\alpha,z}) - \theta]$. The influence function of T_τ is the limit of the above difference quotient as $\alpha \to 0$ (with a slight modification of Hampel's definition to accommodate functionals defined on the space of distributions having densities with respect to the Lebesgue measure). Beran (1977) proved that the α-influence function for the minimum Hellinger distance is a bounded continuous function of z for each fixed α; Eslinger and Woodward (1991) empirically demonstrated the same for the normal model. A similar graphical investigation (not presented here) of the α-influence function of the MGKLE$_\alpha$ show that they have a similar boundedness property for all $\alpha \in (0,1)$ under the normal model. Since a functional with well behaved α-influence functions can have an unbounded influence function "there is no intrinsic conflict between robustness of an estimator and asymptotic efficiency" (Beran, 1977).

4.2 Breakdown Point Analysis

The breakdown point of a statistical functional is roughly the smallest fraction of contamination in the data that may cause an arbitrarily extreme value in the estimate. Here we establish the breakdown point of the MGKLE$_\tau$ functional under the following set up. Let $T_\tau(G)$ be the

MGKLE$_\tau$ functional at the true distribution G. For $\varepsilon \in (0, 1)$, consider the contamination model,

$$H_{\varepsilon,m} = (1 - \varepsilon)G + \varepsilon K_m,$$

where $\{K_m\}$ is a sequence of contaminating distributions, and $h_{\varepsilon,m}, g$ and k_m are the corresponding densities with respect to the dominating measure. Given a contamination sequence $\{K_m\}$ we will say that there is breakdown in T_τ for ε level contamination if $\lim_{m\to\infty} |T_\tau(H_{\varepsilon,m})| = \infty$, in which case we are interested in $\inf\{\varepsilon : \lim_{m\to\infty} |T_\tau(H_{\varepsilon,m})| = \infty\}$. We write below $\theta_m = T_\tau(H_{\varepsilon,m})$, suppressing the τ and ε subscripts for brevity.

We develop the following conditions for the breakdown point analysis. The conditions put appropriate structure on the model and on the contamination sequence which allows us to determine the behavior of the divergences under extreme forms of contamination.

DEFINITION 1 *A contaminating sequence of densities $\{k_m\}$ will be called an outlier sequence relative to truth* g(x) *and model $f_\theta(x)$ if*:

Al $\int \min\{g(x), k_m(x)\} \to 0$ *as $m \to \infty$. That is, the contamination distribution becomes asymptotically singular to the true distribution.*

A2 $\int \min\{f_\theta(x), k_m(x)\} \to 0$ *as $m \to \infty$ uniformly for $|\theta| \le c$, for any fixed c. That is the contamination distribution is asymptotically singular to the specified models.*

Finally we assume that

A3 $\int \min\{g(x), f_{\theta_m}(x)\} \to 0$ *as $m \to \infty$ if $|\theta_m| \to \infty$ as $m \to \infty$. That is, large values of the parameter θ give distributions which become singular to the true distribution.*

Intuitively, outlier sequences represent the worst possible type of contamination sequences. The proof of the foliowing theorem, establishing the breakdown point of the MGKLE$_\tau$ functional under an outlier sequence, is given in Appendix.

THEOREM 3 *Let $\{k_m\}$ be any outlier sequence of densities with respect to the true distribution and the model* (*i.e.* $\{k_m\}$ *satisfies conditions* **Al** *and* **A2**)*. In addition suppose that the model satisfies condition* **A3** *in relation to the true distribution. If the true distribution belongs to the model, then, for any $\varepsilon < 1/2$,* $\limsup_{m\to\infty} |T_\tau(H_{\varepsilon,m})| < \infty$ *where T_τ is the* MGKLE$_\tau$ *functional. When the true distribution does not belong to the model,* $\limsup_{m\to\infty} |T_\tau(H_{\varepsilon,m})| < \infty$ *whenever* $\varepsilon^* = \inf\{\varepsilon\colon a_1(\varepsilon) \le a_2(\varepsilon)\}$, *where,* $a_1(\varepsilon) = (1 - \varepsilon)(1/\bar{\tau})\log(1/\tau) + C_\tau(\varepsilon - 1)$, $a_2(\varepsilon) = \varepsilon(1/\bar{\tau})\log(1/\tau) + \mathrm{GKL}_\tau((1 - \varepsilon)g, f_{\theta^*})$, *$C_\tau(\cdot)$ is as in* (7), *and θ^* is the minimizer of* $\mathrm{GKL}_\tau((1 - \varepsilon)g, f_\theta)$.

4.3 The Generalized Kullback-Leibler Divergence Tests

Because of the lack of robustness of the likelihood ratio tests (LRTs), alternative robust tests have received a lot of attention in the literature. Simpson's (1989) Hellinger deviance test (HDT) is robust under data contamination and efficient under the model. Here, we study analogs of the LRT based on the GKL$_\tau$. Under the parametric setup given in Section 1, let the null hypotheses of interest be H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta\backslash\Theta_0$, where Θ_0 is a proper subset of Θ. Let the functional $T_{\tau,0}$ and T_τ be defined as $T_{\tau,0}(G) \equiv \theta^*_{\tau,G} \in \Theta_0$, and $T_\tau(G) \equiv \theta_{\tau,G}$ which satisfy

$$\mathrm{GKL}_\tau(g, \theta^*_{\tau,G}) = \min_{\theta\in\Theta_0} \mathrm{GKL}_\tau(g, f_\theta) \quad \text{and} \quad \mathrm{GKL}_\tau(g, \theta_{\tau,G}) = \min_{\theta\in\Theta} \mathrm{GKL}_\tau(g, f_\theta)$$

respectively. For a random sample of size n and a kernel density estimate g_n (CDF G_n), denote the estimators $T_{\tau,0}(G_n)$ and $T_\tau(G_n)$ under the null and under no restriction by $\hat\theta^*$ and $\hat\theta$, respectively. Define the generalized Kullback-Leibler divergence test (GKLDT$_\tau$) statistic as

$$2n\{\mathrm{GKL}_\tau(g, f_{\hat\theta^*}) - \mathrm{GKL}_\tau(g, f_{\hat\theta})\}.$$

In discrete models, the GKLDT$_\tau$ for all τ have asymptotic $\chi^2(r)$ distributions under H_0 where r is the number of independent restrictions imposed on Θ by the null hypothesis (Lindsay, 1994). Our sequel paper will theoretically establish the asymptotic null distribution of the GKLDT$_\tau$ statistics for continuous distributions, and here we numerically study the properties of the disparity tests under a variety of settings.

4.4 The Penalized Disparities

Although the estimators and the tests within the GKL family provide a high degree of downweighting for outliers, particularly for values of τ close to 1, the structure of the divergences inevitably leads to an inflation of the effect of the "inliers", points with less observations than predicted under the model. This is clearly apparent from the form of the RAFs and weight functions presented in Figure 1. While the more robust RAFs within this family move away from the linear curve for likelihood as $\delta(=\delta_P)$ increases in positive magnitude exhibiting a much flatter and dampened growth leading to robustness, the same curves also move away from the linear curve as δ increases in negative magnitude, but in the wrong direction, and hence magnifies the effect of inliers. Figure 1(b) gives a different representation of the same phenomenon, where the weight functions all converge to 0 as $\delta_c \to 1$, but those for larger values of τ lead to an unacceptably high value of the weight in the left tail as $\delta_c \to -1$.

In particular one can write the GKL$_\tau$ divergence in the form

$$\mathrm{GKL}_\tau(g,f) = \int_{g>0}\left[\frac{g(x)}{\bar\tau}\log\left(\frac{g(x)}{f(x)}\right) - \left(\frac{g(x)}{\bar\tau} + \frac{f(x)}{\tau}\right)\log\left(\tau\frac{g(x)}{f(x)} + \bar\tau\right)\right] - \frac{1}{\tau}\log\,\bar\tau\int_{g=0} f(x)$$

where the second term on the right is the contribution of the observations with $g = 0$ to the divergence (contribution of the empty cells for the discrete case when $g = g_n$ represents the sample relative frequencies). This coefficient $-1/\tau \log\,\bar\tau = -\log(1-\tau)^{1/\tau}$ converges to 1 as $\tau \to 0$, but becomes arbitrarily large as $\tau \to 1$, showing that the divergence puts a very high weight on the set $\{x: g(x) = 0\}$. In this paper we will also consider the following "penalized" version of the GKL$_\tau$ divergence, to be called the pGKL$_\tau$ divergence, defined as

$$\mathrm{pGKL}_\tau(g,f) = \int_{g>0}\left[\frac{g(x)}{\bar\tau}\log\left(\frac{g(x)}{f(x)}\right) - \left(\frac{g(x)}{\bar\tau} + \frac{f(x)}{\tau}\right)\log\left(\tau\frac{g(x)}{f(x)} + \bar\tau\right)\right] + \int_{g=0} f(x).$$

Notice that the weight of the set $\{x: g(x) = 0\}$ has been redefined to be 1 (which is the ordinary weight for the likelihood disparity where $\tau = 0$) for all values of τ. Our simulations will show that for large values of τ the use of the penalized divergences instead of the natural ones can lead to a fair improvement in small sample efficiency without significantly altering the robustness properties. Notice that the penalized divergence is also nonnegative by Lemma 1, and the estimation functional minimizing $\text{pGKL}_\tau(g_n, f_\theta)$ is Fisher consistent. See Harris and Basu (1994) and Basu and Basu (1998) for other applications of penalized divergences.

5 NUMERICAL RESULTS

5.1 Examples

5.1.1 Drosophila Assay

First, we consider a part of an experiment originally reported by Woodruff *et al.* (1984), and analyzed by Simpson (1987). The frequencies of frequencies of daughter flies carrying a recessive lethal mutation on the X-chromosome are considered where the male parents have been exposed to a certain degree of a chemical. Roughly hundred daughter flies were sampled for each male. This particular experiment resulted in $(x_i, f_i) = (0, 23)$, (1, 7), (2, 3), (91, 1) for one experimental run, where x_i is the number of daughters carrying the recessive lethal mutation and f_i is the number of male parents having x_i such daughters. We will refer to this as the *Drosophila Data I*. The estimators of θ under a parametric Poisson (θ) model corresponding to $\tau = 0.1$, 0.3, 0.5, 0.7, 0.9, 0.99 for the Drosophila Data I are presented in Table I together with the MLE $(\tau = 0)$. Notice that all the estimators with $\tau > 0$, including those with very small values of τ are successful in completely discarding the outlier (91). When this outlier is removed the MLE of the remaining data is 0.394, much in the vicinity of the other MLGKE_τs and pMGKLE_τs obtained with the full data.

The second example also involves data from Woodruff *et al.* (1984). The responses now are the frequencies of daughter flies having a recessive lethal mutation on the X-chromosome where the male parent was either exposed to a dose of chemical or to control conditions. This data set, also analyzed by Simpson (1989, Table 5) will be referred to as the *Drosophila Data II*. The responses are modeled as Poissons with mean θ_1 (control), and θ_2 (exposed) respectively. For testing $H_0\colon \theta_1 \geq \theta_2$ against $H_1\colon \theta_1 < \theta_2$, a two sample signed version of the GKLDT_τ (or its penalized version) is appropriate. Suppose that random samples of size n_i are available from the population with density $f_{\theta_i}(\cdot)$ and let $d_i(\cdot)$ be the empirical density of ith sample, $i = 1, 2$. For a disparity $\rho(\cdot)$ between two densities, define the overall disparity for the two sample case as

$$D = D(\theta_1, \theta_2) = \frac{1}{n_1 + n_2}\{n_1\rho(d_1, f_{\theta_1}) + n_2\rho(d_2, f_{\theta_2})\}.$$

Given the disparity test statistic $t_n = 2n(\hat{D}_0 - \hat{D})$, where $\hat{D}_0$ and $\hat{D}$ are evaluated at the minimizers of $D(\cdot, \cdot)$ under the null and without any restrictions respectively, the signed divergence statistic is given by $s_n = t_n^{1/2}\text{sign}(\hat{\theta}_2 - \hat{\theta}_1)$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the minimum disparity estimators of the parameters θ_1 and θ_2. It follows from Sarkar and Basu

TABLE I The Estimated Parameters Under the Poisson Model for the Drosophila Data I.

τ	*0*	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*	*0.99*
GKL_τ	3.059	0.390	0.383	0.374	0.363	0.345	0.316
pGKL_τ	3.059	0.391	0.387	0.382	0.378	0.373	0.372

TABLE II The Signed Divergence Statistics (s_n) and Their p-values for the Drosophila Data II.

	All observations				*Outliers deleted*			
	GKLDT$_\tau$		*pGKLDT*$_\tau$		*GKLDT*$_\tau$		*pGKLDT*$_\tau$	
τ	s_n	*p-value*	s_n	*p-value*	s_n	*p-value*	s_n	*p-value*
0	2.595	0.002	2.595	0.002	1.099	0.136	1.099	0.136
0.1	0.963	0.168	0.963	0.168	1.000	0.159	1.001	0.158
0.3	0.753	0.226	0.755	0.225	0.795	0.213	0.796	0.213
0.5	0.608	0.271	0.612	0.270	0.657	0.256	0.660	0.255
0.7	0.509	0.306	0.514	0.304	0.563	0.287	0.568	0.285
0.9	0.436	0.331	0.446	0.328	0.497	0.310	0.504	0.307
0.99	0.402	0.344	0.421	0.337	0.468	0.320	0.482	0.315

(1995) that the signed two sample GKLDT$_\tau$ is asymptotically equivalent to the signed likelihood ratio test. For the full data and the reduced data (after removing the two large observations from the treated group) the signed divergences and the associated p-values using the standard normal approximation are given in Table II. The apparently significant result for the likelihood ratio test ($\tau = 0$) is due to the presence of the two large outliers, as the removal of these observations changes the conclusion of the likelihood ratio test. The conclusion of the other tests are not affected by the presence or absence of the outliers.

5.1.2 *The Number of Cases of Peritonitis*

The next example involves the incidence of peritonitis on $n = 390$ kidney patients (Tab. III). A glimpse of the data suggests that a *geometric* model with θ around $1/2$ may fit the data well. The data set, provided by Prof. P. W. M. John, was previously analyzed by Basu and Basu (1998). The observed frequency (O_k) of the number of cases of peritonitis (k) is modeled by the geometric distribution with success probability θ. For an estimate $\hat{\theta}$, the expected frequencies are then obtained as $E_k = n\hat{\theta}(1 - \hat{\theta})^k$. The largest number of cases of peritonitis is $k = 12$, so we merged all the expected frequencies for $k \geq 12$. To assess the goodness-of-fit of the model, we use the log likelihood ratio statistic which is given for this data as

$$G^2 = 2\sum_{k=0}^{12} O_k \log\left(\frac{O_k}{E_k}\right).$$

In this example the fit provided by the MLE is excellent (Tab. III); those for the minimum disparity estimators are also remarkably good, particularly those based on the penalized disparities. The two marginally large observations at 10 and 12 have little impact since the sample size is so large. This example shows that when the data roughly follows the model the proposed methods are close to likelihood based ones in performance.

5.1.3 *Determinations of the Parallax of the Sun*

We consider Short's data (Stigler, 1977, Data Set 2) for the determination of the parallax of the sun, the angle subtended by the earth's radius, as if viewed and measured from the surface of the sun. From this angle and available knowledge of the physical dimensions of the earth, the mean distance from earth to sun can be easily determined. To carry out the GKL$_\tau$ and pGKL$_\tau$ estimation, we have used the kernel density function, defined in (1),

TABLE III The Observed Frequencies (O_k) of the Number of Cases (k) of Peritonitis for each of 390 Kidney Patients and the Expected Frequencies Under Different Methods with the Goodness-of-fit Likelihood Ratio Statistics (G^2).

k	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12+*	G^2
O_k	*199*	*94*	*46*	*23*	*17*	*4*	*4*	*1*	*0*	*0*	*1*	*0*	*1*	–
τ							GKL$_\tau$							
0	193.5	97.5	49.1	24.7	12.5	6.3	3.2	1.6	0.8	0.4	0.2	0.1	0.1	10.4
0.1	195.3	97.5	48.7	24.3	12.1	6.1	3.0	1.5	0.8	0.4	0.2	0.1	0.1	10.5
0.3	196.9	97.5	48.3	23.9	11.8	5.9	2.9	1.4	0.7	0.4	0.2	0.1	0.1	10.7
0.5	198.2	97.5	47.9	23.6	11.6	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.1	10.9
0.7	199.5	97.4	47.6	23.2	11.4	5.5	2.7	1.3	0.6	0.3	0.2	0.1	0.1	11.2
0.9	201.8	97.4	47.0	22.7	11.0	5.3	2.6	1.2	0.6	0.3	0.1	0.1	0.1	11.9
0.99	205.7	97.2	45.9	21.7	10.3	4.9	2.3	1.1	0.5	0.2	0.1	0.1	0.0	13.6
τ							pGKL$_\tau$							
0	193.5	97.5	49.1	24.7	12.5	6.3	3.2	1.6	0.8	0.4	0.2	0.1	0.1	10.4
0.1	195.1	97.5	48.7	24.3	12.2	6.1	3.0	1.5	0.8	0.4	0.2	0.1	0.1	10.5
0.3	196.5	97.5	48.4	24.0	11.9	5.9	2.9	1.5	0.7	0.4	0.2	0.1	0.1	10.6
0.5	197.2	97.5	48.2	23.8	11.8	5.8	2.9	1.4	0.7	0.3	0.2	0.1	0.1	10.8
0.7	197.8	97.5	48.0	23.7	11.7	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.1	10.9
0.9	198.3	97.5	47.9	23.5	11.6	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.1	11.0
0.99	198.5	97.5	47.9	23.5	11.5	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.1	11.0

TABLE IV Fits of a Normal $N(\mu, \sigma^2)$ Model to Short's Data Using the Maximum Likelihood (ML), Maximum Likelihood Without Outlier (ML-O), and Minimum GKL$_\tau$ and pGKL$_\tau$ Estimation.

	τ	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*	*0.99*	*ML-O*	*ML*
GKL$_\tau$	$\hat{\mu}$	8.518	8.382	8.379	8.383	8.388	8.391	8.541	8.378
	$\hat{\sigma}$	0.550	0.355	0.336	0.323	0.308	0.292	0.552	0.846
pGKL$_\tau$	$\hat{\mu}$	8.520	8.384	8.382	8.385	8.389	8.392	–	–
	$\hat{\sigma}$	0.554	0.362	0.345	0.335	0.326	0.322	–	–

with the Epanechnikov kernel ($w(x) = (3/4)(1 - x^2)$, if $|x| < 1$, and $w(x) = 0$, otherwise). Following Devroye and Györfi (1985, pp. 107–108), the mean L_1 criterion with the Epanechnikov kernel and Gaussian $f(\cdot)$ leads to an optimal bandwidth of the form

$$h_n = (15e)^{1/5}\left(\frac{\pi}{32}\right)^{1/10}\sigma n^{-1/5} = 1.66\sigma n^{-1/5},$$

where σ is the standard deviation. If the standard deviation is not specified, one can use $\hat{h}_n = 1.66\hat{\sigma}n^{-1/5}$, where $\hat{\sigma} = \mathrm{MAD} = \mathrm{median}(|X_i - \mathrm{median}(X_i)|)/0.674$.

For Short's data, Table IV gives the values of the maximum GKL$_\tau$ and pGKL$_\tau$ estimates of μ and σ for various values of τ under the normal model, as well as MLEs for the all observations and those after deleting the biggest outlier 5.76. Removal of the large outlier 5.76 reduces the MLE of σ from 0.846 to 0.552. All the MGKLE$_\tau$s successfully downweight the largest outlier 5.76. In addition, the strong downweighting properties of the MGKLE$_\tau$s for large values of τ tend to discard the more moderate outliers 9.71 and 9.87. Fitted normal densities are shown in Figure 2 along with the Epanechnikov kernel density estimate and the normal density fit by maximum likelihood and GKL$_{\tau=0.5}$. The superiority of the robust fit is evident.

5.1.4 *Measurements of the Passages of the Light*

We next consider Newcomb's light speed data (Stigler, 1977). The data were analyzed by Brown and Hwang (1993), who attempted to fit the "best approximating normal distribution"

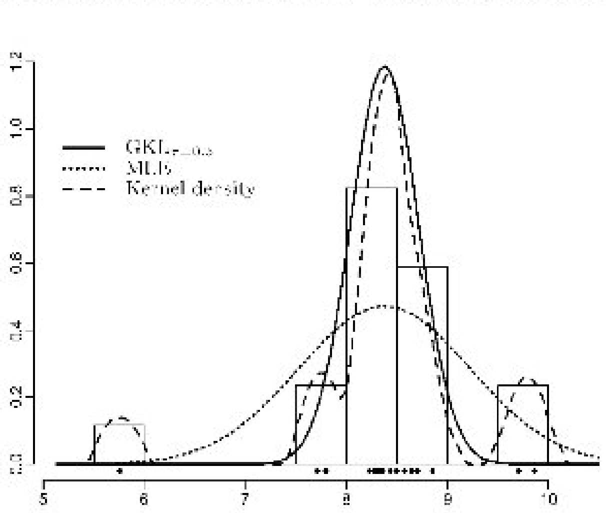


FIGURE 2 Normal density fits to Short's 1763 determinations of the parallax of the sun (Data Set 2).

TABLE V Fits of a Normal $N(\mu, \sigma^2)$ Model to Newcomb's Data Using the Maximum Likelihood (ML), Maximum Likelihood Without Outlier (ML-O), and Minimum GKL$_\tau$ and pGKL$_\tau$ Estimation.

	τ	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*	*0.99*	*ML-O*	*ML*
GKL$_\tau$	$\hat{\mu}$	27.75	27.74	27.74	27.73	27.72	27.71	27.75	26.21
	$\hat{\sigma}$	5.24	5.21	5.18	5.14	5.06	4.94	5.04	10.66
pGKL$_\tau$	$\hat{\mu}$	27.75	27.74	27.74	27.73	27.72	27.72	–	–
	$\hat{\sigma}$	5.25	5.23	5.22	5.20	5.19	5.18	–	–

to the corresponding histogram. For the Newcomb data, Table V gives the values of the GKL$_\tau$ and pGKL$_\tau$ estimate of μ and σ for various values of τ under the normal model, as well as MLEs for the full data (ML), and those after deleting -44, and -2, the two obvious outliers (ML-O).

We have used the Epanechnikov kernel with bandwidth $\hat{h}_n = 1.66\,\mathrm{MAD}n^{-1/5}$. Notice that these estimators exhibit strong outlier resistance properties even for quite small values of τ. A graphic representation is provided in Figure 3, where the normal densities $N(\hat{\mu}, \hat{\sigma}^2)$, for $\tau = 0$ and 0.5 are superimposed on a histogram of the Newcomb data, together with the kernel density estimator. With the robust estimator, the estimated normal density fits the main body of histogram very well, unlike the result obtained with the maximum likelihood estimator.

5.1.5 Telephone-line Faults

Welch (1987) considered data from an experiment to test a method of reducing faults on telephone lines. This data set is presented and analyzed in Simpson (1989) using the Hellinger distance. For the telephone-line fault data, the next table gives the values of the GKL$_\tau$ and pGKL$_\tau$ estimates of μ and σ for various values of τ under the normal model, as well as MLEs for the full data (ML), and those after deleting the large outlier -988.

Once again we have used the Epanechnikov kernel with bandwidth $\hat{h}_n = 1.66\,\mathrm{MAD}n^{-1/5}$. A graphic representation is provided in Figure 4, where the normal densities $N(\hat{\mu}, \hat{\sigma}^2)$, for $\tau = 0$, 0.5 are superimposed on a histogram of the telephone-line faults data, together

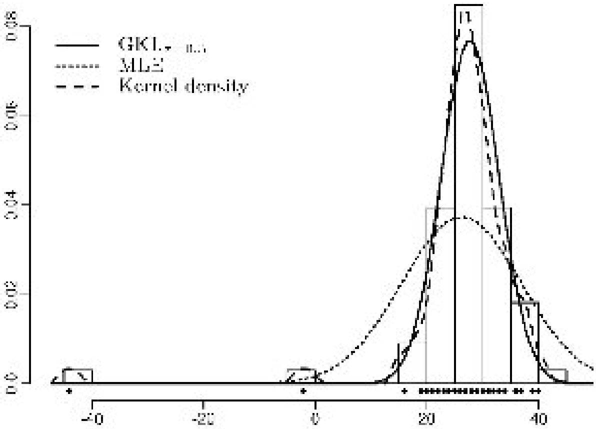


FIGURE 3 Normal density fits for Newcomb data.

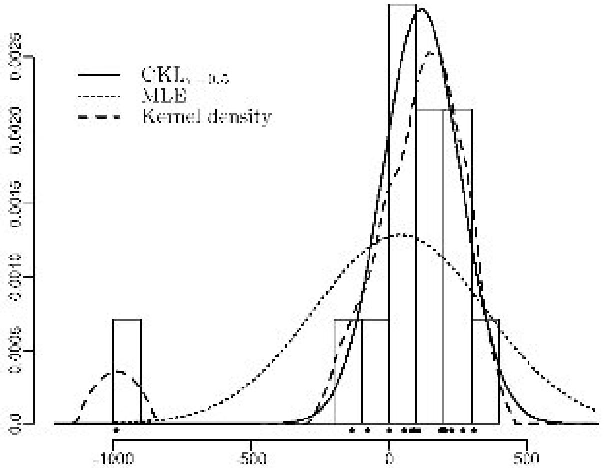


FIGURE 4 Normal density fits for the telephone-line fault data.

with the kernel density estimate. Once again the robustness of the proposed estimators is clear from Table VI and Figure 4.

For the telephone fault data we also computed the disparity test statistics for testing the hypothesis $H_0\colon \mu = 0$ against $H_1\colon \mu > 0$ under the normal model with σ unspecified. The likelihood ratio test (the one sided t test in this case) produces p-values of 0.3292 and 0.0038 for the above data with and without the outlier respectively – the large outlier reverses the conclusion and forces the acceptance of H_0. The p-values for the corresponding signed divergences of the GKLDT$_\tau$s (with z critical values) are presented in Table VII. Clearly these p-values are only minimally affected by the outlier, and the presence of the latter does not change the conclusion. Use of the critical values of the $t(13)$ distribution increases the p-values slightly, but does not change the conclusions.

5.2 Simulation Results

To keep a clear focus, we restrict the simulations to a Poisson model. In the first study, the data are generated from the Poisson distribution with mean 5, and modeled as the Poisson(θ) distribution. Here, as well as in the rest of the paper, three sample sizes $n = 20, 50, 100$ are considered. In Table VIII, we have presented the bias and the mean square errors of the estimators of θ obtained by minimizing the GKL$_\tau$ and pGKL$_\tau$ for several values of τ for pure Poisson data as well as contaminated Poisson data (discussed later) with 5000 replications. It is clear that the small sample efficiency at the model is a decreasing function of τ. The penalized versions provide improved efficiency for larger values of τ under the model,

TABLE VI Fits of a Normal $N(\mu, \sigma^2)$ to the Telephone-line Fault Data Using the Maximum Likelihood (ML), Maximum Likelihood Without Outlier (ML-O), and Minimum GKL$_\tau$ and pGKL$_\tau$ Estimation.

	τ	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*	*0.99*	*ML-O*	*ML*
GKL$_\tau$	$\hat{\mu}$	117.9	117.8	117.7	117.5	117.1	116.3	117.92	38.93
	$\hat{\sigma}$	143.6	142.5	141.1	139.2	135.8	130.5	127.61	310.23
pGKL$_\tau$	$\hat{\mu}$	118.0	118.2	118.3	118.4	118.4	118.3	–	–
	$\hat{\sigma}$	143.9	143.6	143.2	142.8	142.2	141.8	–	–

TABLE VII The p-values for the Test of H_0: $\mu=0$ Against H_1: $\mu>0$ with σ Unspecified.

	τ	*0.1*	*0.3*	*0.5*	*0.7*	*0.9*	*LRT*
All observations	GKLDT$_\tau$	0.0050	0.0050	0.0048	0.0044	0.0038	0.3292
	pGKLDT$_\tau$	0.0047	0.0042	0.0036	0.0028	0.0017	
Outlier deleted	GKLDT$_\tau$	0.0048	0.0050	0.0051	0.0049	0.0045	0.0038
	pGKLDT$_\tau$	0.0046	0.0043	0.0038	0.0031	0.0021	

particularly for smaller sample sizes. At sample size $n=100$, the efficiency of the pGKL$_{0.5}$ estimator is over 95% compared to the MLE.

Next, data are generated from the 0.9Poisson(5) + 0.1Poisson(15) mixture, and the assumed model is Poisson(θ). Once again the estimates of θ, and their mean square errors around the target value of $\theta=5$ are computed. The results, presented in Table VIII, show that the more robust estimators (corresponding to larger values of τ) now start doing better. The penalized estimators are close to or better in performance than the ordinary versions in most cases, showing that their robustness has not been compromised by the effect of the penalty.

Next, we generated data from Poisson distributions with θ in the range (3.5, 6.5), and determined the power of each of the tests for the null H_0: $\theta=5$ against the two sided alternative based on both the chi-square critical values and empirically determined critical values. The results for the nominal level $\gamma=0.05$ are presented in Figure 5 for a few values of τ and are based on sample size 50 with 1000 replications. The thick dashed line represents the

TABLE VIII Estimated Biases and Mean Square Errors of the Estimators Under Consideration. 5000 Random Samples were Drawn from Poisson(5) and 0.1Poisson(5)+0.9Poisson(15) with Sample Size $n=20$, 50, 100.

	Poisson(5)				*0.1Poisson(5)+0.9Poisson(15)*			
	GKL$_\tau$		*pGKL$_\tau$*		*GKL$_\tau$*		*pGKL$_\tau$*	
τ	*Bias*	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*
				Sample size $n=20$				
0	0.0079	0.2493	0.0079	0.2493	1.0038	1.7522	1.0038	1.7522
0.1	−0.0268	0.2550	−0.0193	0.2542	0.2394	0.4950	0.2517	0.5035
0.3	−0.0693	0.2688	−0.0424	0.2631	0.1185	0.4195	0.1563	0.4319
0.5	−0.1089	0.2876	−0.0552	0.2710	0.0478	0.4042	0.1187	0.4148
0.7	−0.1562	0.3176	−0.0627	0.2787	−0.0202	0.4126	0.0984	0.4107
0.9	−0.2360	0.3904	−0.0658	0.2872	−0.1178	0.4672	0.0881	0.4144
0.99	−0.3498	0.5501	−0.0657	0.2916	−0.2450	0.6150	0.0861	0.4187
				Sample size $n=50$				
0	0.0020	0.1005	0.0020	0.1005	1.0099	1.3133	1.0099	1.3133
0.1	−0.0174	0.1023	−0.0130	0.1022	0.2839	0.2547	0.2902	0.2584
0.3	−0.0423	0.1061	−0.0266	0.1048	0.1749	0.1882	0.1933	0.1942
0.5	−0.0662	0.1115	−0.0350	0.1075	0.1193	0.1691	0.1532	0.1761
0.7	−0.0959	0.1205	−0.0409	0.1106	0.0716	0.1628	0.1279	0.1686
0.9	−0.1490	0.1438	−0.0452	0.1146	0.0083	0.1712	0.1090	0.1669
0.99	−0.2361	0.2046	−0.0467	0.1171	−0.0758	0.2180	0.1015	0.1681
				Sample size $n=100$				
0	0.0003	0.0503	0.0003	0.0503	1.0061	1.1591	1.0061	1.1591
0.1	−0.0120	0.0507	−0.0093	0.0507	0.3286	0.1986	0.3309	0.1999
0.3	−0.0285	0.0520	−0.0191	0.0516	0.2241	0.1326	0.2307	0.1348
0.5	−0.0445	0.0541	−0.0258	0.0526	0.1759	0.1119	0.1878	0.1145
0.7	−0.0645	0.0577	−0.0311	0.0539	0.1402	0.1018	0.1599	0.1043
0.9	−0.1010	0.0675	−0.0356	0.0557	0.1021	0.0988	0.1382	0.0985
0.99	−0.1655	0.0963	−0.0374	0.0567	0.0616	0.1134	0.1294	0.0970

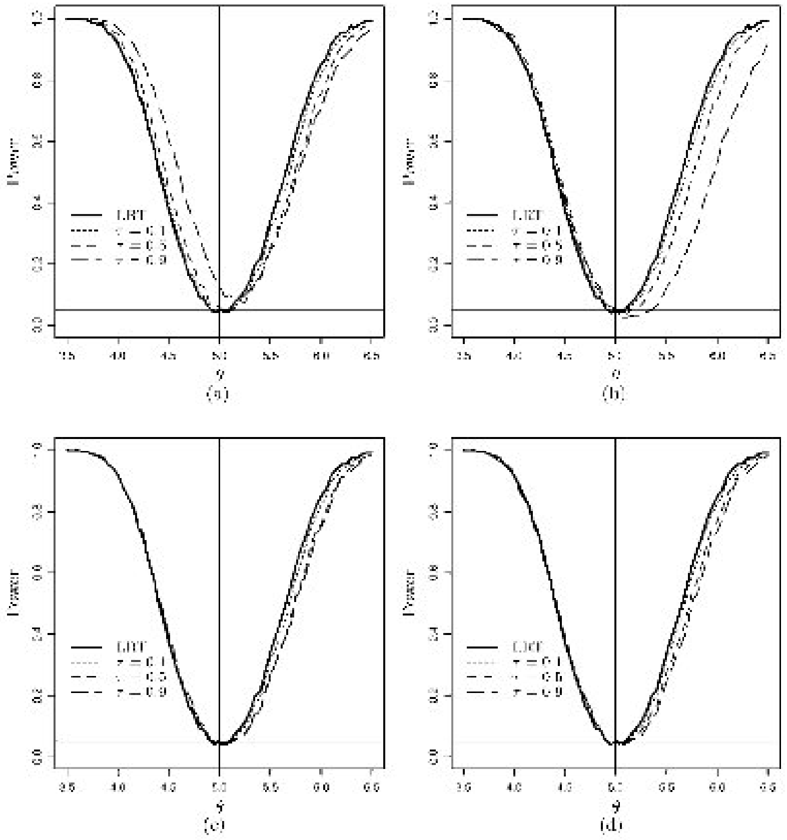


FIGURE 5 Estimated powers for the tests under consideration testing $H_0: \theta = 5$ versus $H_1: \theta \neq 5$ with level $\gamma = 0.05$. 1000 random samples were drawn from Poisson(θ) with sample size $n = 50$. (a) GKL$_\tau$ based on chi-square critical value; (b) GKL$_\tau$ based on empirical critical values; (c) pGKL$_\tau$ based on chi-square critical value; (d) pGKL$_\tau$ based on empirical critical values.

likelihood ratio test for each case. Notice that, particularly for the penalized divergences, the tests are very close to the likelihood ratio tests, and the nominal levels are very close to the true levels (when using chi-square critical values). This is particularly encouraging since in actual practice when one wants to use these tests determining empirical critical values for each individual case is obviously not practical.

We next looked at the powers of the methods under contamination. Data are now generated from $0.9\text{Poisson}(\theta) + 0.1\text{Poisson}(15)$ mixture. The results for the power calculation (both with chi-square critical values and empirical critical values determined from the pure data) for the same null hypothesis and for the nominal level $\gamma = 0.05$ are presented in Figure 6 and are based on sample size 50 with 1000 replications. For comparison purposes the power curve of the likelihood ratio test for the no contamination case is presented with the other graphs as the thick solid line. While the power curve of the likelihood ratio test under contamination shows a dramatic shift with substantial loss of power at several cases, the other curves are largely unchanged in comparison, demonstrating the relative stability of these test statistics under contamination.

This suggests that the use of robust tests based on the GKL$_\tau$ family can provide attractive alternatives to the likelihood ratio test, and have very good power under pure data and good stability in level and power when the model is diffused with noise.

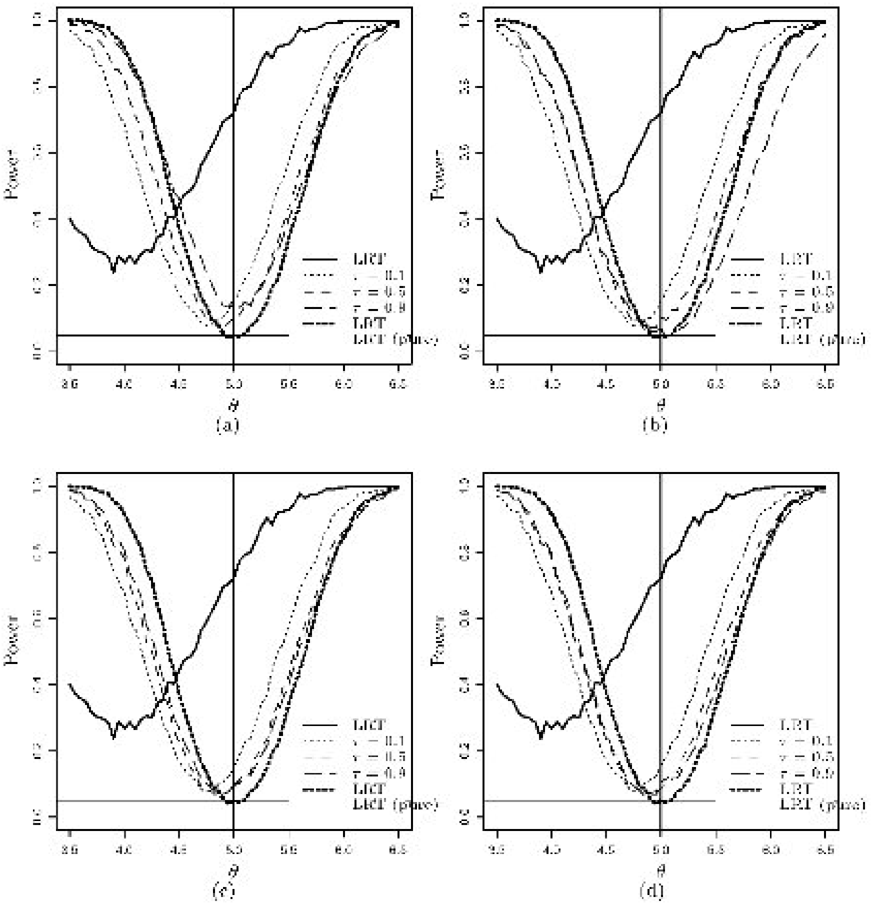


FIGURE 6 Estimated powers for the tests under consideration testing H_0: $\theta = 5$ versus $H_1 : \theta \neq 5$ with level $\gamma = 0.05$. 1000 random samples were drawn from 0.9Poisson(θ)+0.1Poisson(15) with sample size $n = 50$. (a) GKL$_\tau$ based on chi-square critical value; (b) GKL$_\tau$ based on empirical critical values; (c) pGKL$_\tau$ based on chi-square critical value; (d) pGKL$_\tau$ based on empirical critical values.

5.3 Bias Plots for Contaminated Distributions

In order to investigate more robustness properties of the proposed methods, we have also looked at the global minimum of the GKL divergence when the assumed model is Poisson but the true distribution is a contaminated version of a Poisson density. The bias plots of the true functionals are studied as functions of z, the mean of the contaminating distribution (which is also taken to be a Poisson), as well as functions of ε, the contaminating proportion.

Let Poisson(M) represent a Poisson distribation with mean M. For the Poisson model, let $T(F)$ be the GKL functional (dropping the τ subscript for brevity) representing the population mean. In our first study we have computed "ε-influence function"

$$\begin{aligned}\mathrm{IF}_\varepsilon &= \frac{1}{\varepsilon}[T\{(1-\varepsilon)\mathrm{Poisson}(5) + \varepsilon\mathrm{Poisson}(z)\} - T\{\mathrm{Poisson}(5)\}] \\ &= \frac{1}{\varepsilon}[T\{(1-\varepsilon)\mathrm{Poisson}(5) + \varepsilon\mathrm{Poisson}(z)\} - 5]. \qquad (9)\end{aligned}$$

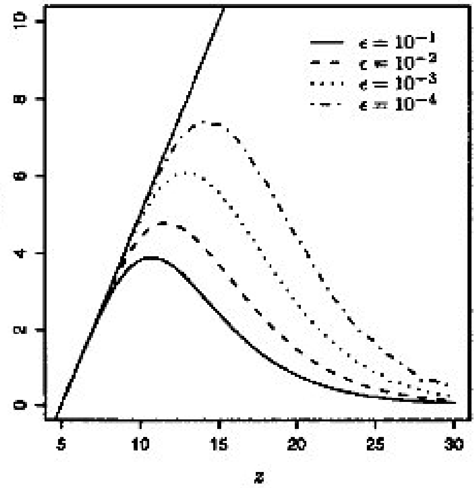


FIGURE 7 Influence function and ε-influence functions.

Notice that one gets the ordinary influence function when one takes the limit of the difference quotient (a standardized version of the bias) on the right hand side of (9) as $\varepsilon \to 0$. In our specific case we have chosen $\tau = 0.5$, allowed z to vary from 5 to 30, and used four different values of ε. The results are given in Figure 7. Notice that in each case the ε-influence function has a redescending nature, unlike the actual influence function, which is unbounded (see Sec. 4.1) and shown on the same figure as the solid black line. The above behavior of the GKL is similar to that of the Hellinger distance, where the ε-influence function turns out to be a bounded continuous function even when the influence function is unbounded (Beran, 1977).

As a second example we looked at the effect of increasing the contaminating proportion on the minimum GKL functional $T(\cdot)$. Once again we have computed the functional corresponding to $\tau = 0.5$ at the contaminated distribution $(1-\varepsilon)$Poisson(5) + εPoisson(z), but now z is kept fixed and ε is allowed to vary. The value of the functional corresponding to two different scenarios ($z = 15$ and 20) are plotted in Figures 8(a) and (c). In either case, the relative stability of the functional for values of $\varepsilon < 0.5$ is clear. Around $\varepsilon = 0.5$ there is a jump in the functional to a value close to the true value of the functional at the contaminating component. Notice that for $\varepsilon < 0.5$, the functional becomes more stable under contamination as the contaminating component is further removed from the true one. The solid black line representing the change in the maximum likelihood functional provides the basis for comparison. Figures 8(b) and (d) provide a blown-up corner of the plot of the functionals close to the value $\varepsilon = 0$.

6 CONCLUDING REMARKS

We have numerically demonstrated that inference procedures based on the minimized GKL_τ divergence can provide attractive alternatives to classical inference procedures in many situations. The procedure leads to 50% estimation breakdown under outlier sequences, and is fairly competitive in performance in relation with the optimal methods.

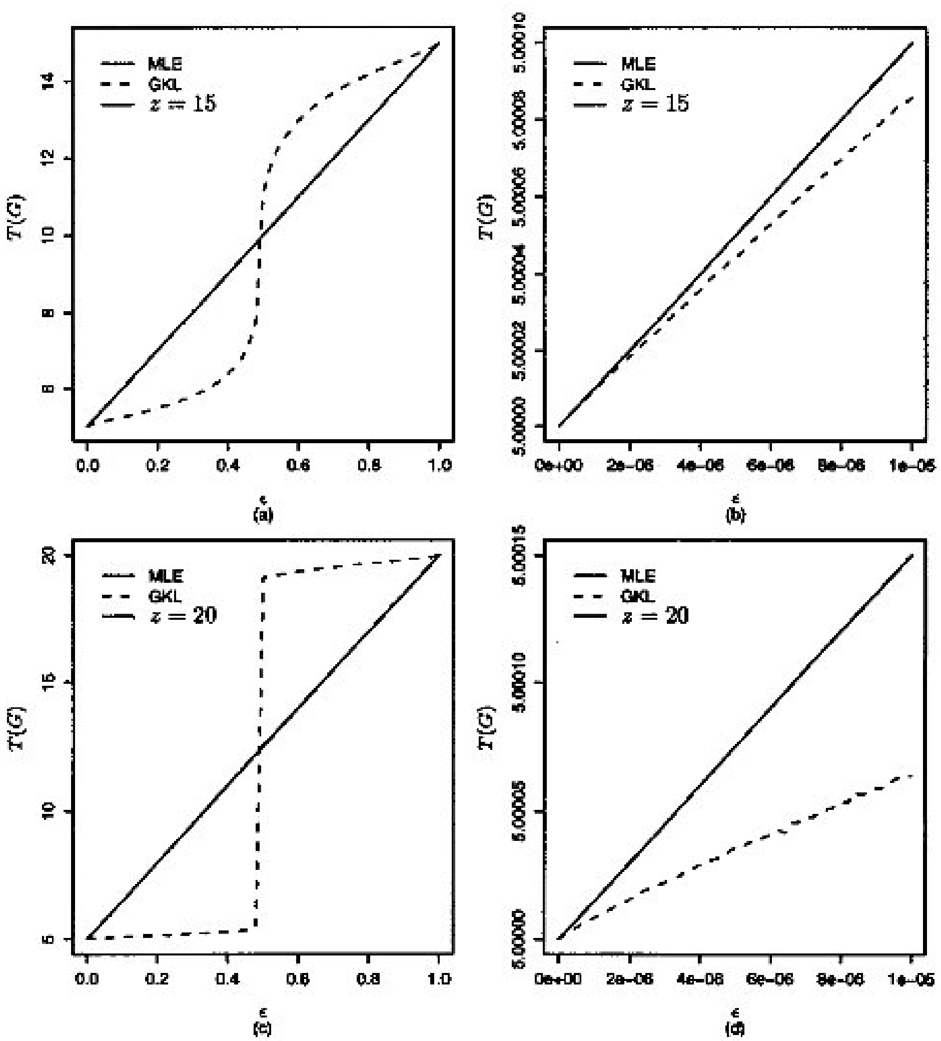


FIGURE 8 Bias plot for the MGKLE$_\tau$ with $G = (1 - \varepsilon)\text{Poisson}(5) + \varepsilon\text{Poisson}(z)$.

A penalized version generally produces better efficiency results for large values of τ, apparently without compromising their robustness properties. The asymptotic optimality of the methods are direct consequences of Lindsay (1994) in discrete models. The corresponding theoretical properties under continuous models will be dealt with by the authors in a sequel paper.

We conclude with a brief discussion on the possible selection of τ in a practical situation on the basis of our numerical studies. Although not entirely comprehensive, our studies appear to suggest the following features for the disparity:

(a) For pure data, the estimator corresponding to $\tau = 0$ (which actually represents the maximum likelihood estimator) performs the best, and the performance of the estimators become relatively poor as the value of τ increases;

(b) there is little difference between the ordinary and penalized estimators for relatively smaller values of τ, but there appears to be a substantial improvement in the attained mean square error due to the penalty for values of τ close to 1;
(c) under contamination, the performance of the maximum likelihood estimator ($\tau = 0$) is seriously distorted, but the estimators are fairly robust for all other values of τ;
(d) for the contaminated examples, the best performer varies over the sample size.

We hope that the above observations can provide some guideline to practitioners for choosing the value of τ in particular problems depending on specific needs. A universal recommendation suitable for all situations appears impossible at the moment. However, based on the above observations we make the following compromise recommendations: (i) wherever the experimenter is not sure about the quality of the data, the minimum GKL estimator corresponding to $\tau = 0.5$ may be used; the penalty, which appears to make this particular choice function slightly better at the model and slightly worse under contamination can be used at the experimenters discretion; (ii) however, whenever the experimenter is fairly certain about the purity of the data, we recommend the choice of a very small value of τ (but not $\tau = 0$, since it can lead to disaster if the experimenter has erred in his/her judgment); (iii) when the experimenter has a strong reason to believe that the data may contain a fair amount of contamination, we recommend the use of a very large value of τ together with the penalty.

References

Basu, A. and Basu, S. (1998). Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, **8**, 841–860.

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, **5**, 445–463.

Brown, L. D. and Hwang, J. T. G. (1993). How to approximate a histogram by a normal density. *The American Statistician*, **47**, 251–255.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, **46**, 440–464.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, New York.

Eslinger, P. W. and Woodward, W. A. (1991). Minimum Hellinger distance estimation for normal models. *Journal of Statistical Computation and Simulation*, **39**, 95–114.

Harris, I. R. and Basu, A. (1994). Hellinger distance as a penalized log likelihood. *Communications in Statistics: Simulation and Computation*, **23**, 1097–1113.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, **22**, 1081–1114.

Park, C., Basu, A. and Lindsay, B. G. (2002). The residual adjustment function and weighted likelihood: A graphical interpretation of robustness of minimum disparity estimators. *Computational Statistics and Data Analysis*, **39**, 21–33.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp.*, Vol. I. University of California Press, Berkeley, pp. 531–546.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society B*, **24**, 46–72.

Sarkar, S. and Basu, A. (1995). On disparity based robust tests for two discrete populations. *Sankhya B*, **57**, 353–364.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, **82**, 802–807.

Simpson, D. G. (1989). Hellinger deviance test: Efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, **84**, 107–113.

Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, **5**, 1055–1098.

Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, **81**, 223–229.

Welch, W. J. (1987). Rerandomizing the median in matched-pairs designs. *Biometrika*, **74**, 609–614.

Woodruff, R. C., Mason, J. M., Valencia, R. and Zimmering, A. (1984). Chemical mutagenesis testing in drosophila – I: Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental Mutagenesis*, **6**, 189–202.

APPENDIX

Proof of Lemma 1 Since $\varphi(\cdot)$ is strictly convex on (0, 1), we have $D(g,f) \geq 0$ with the equality only when $g = f$. For fixed f and $g \in (0,f)$, look at $D(g,f)$ as a function of g.

$$\frac{\partial}{\partial g} D(g,f) = \frac{1}{\bar{\tau}} \log\left(\frac{g}{\tau g + \bar{\tau} f}\right) < 0, \quad \forall g \in (0,f).$$

Since $D(\cdot,f)$ is strictly decreasing for $g \in (0,f)$ and right-continuous at $g = 0$, we have $D(g,f) \leq D(0,f)$ for $g \in (0,f)$ with the equality only when $g = 0$. Similarly, $D(g,f) \leq D(g,0)$ for $f \in (0,g)$. ■

Proof of Theorem 2 It follows from Lemma 1 that

$$\begin{aligned}\int D(g(x),f(x)) &\leq \int_{g<f} D(0,f(x)) + \int_{f<g} D(g(x),0)\\ &\leq \int D(0,f(x)) + \int D(g(x),0) = \frac{1}{\tau}\log\left(\frac{1}{\bar{\tau}}\right) + \frac{1}{\bar{\tau}}\log\left(\frac{1}{\tau}\right).\end{aligned}$$

It is easily shown that the equality holds only when the two densities have disjoint support almost everywhere. ■

Proof of Theorem 3 Let θ_m be the minimizer of $\mathrm{GKL}_\tau(h_{\varepsilon,m},f_\theta)$. Given a level of contamination ε suppose, if possible, breakdown occurs, that is there exists a sequence $\{K_m\}$ such that $|\theta_m| \to \infty$ where $\theta_m = T_\tau(H_{\varepsilon,m})$. Define $A_m = \{x: g(x) > \max(k_m(x), f_{\theta_m}(x))\}$.

$$\begin{aligned}\mathrm{GKL}_\tau(h_{\varepsilon,m},f_{\theta_m}) &= \int D(h_{\varepsilon,m}(x), f_{\theta_m}(x))\\ &= \int_{A_m} D(h_{\varepsilon,m}(x), f_{\theta_m}(x)) + \int_{A_m^c} D(h_{\varepsilon,m}(x), f_{\theta_m}(x)),\end{aligned}$$

From **A1**, $\int_{A_m} k_m(x) \to 0$, and from **A3**, $\int_{A_m} f_{\theta_m}(x) \to 0$ as $m \to \infty$. Similarly from **A1** and **A3**, $\int_{A_m^c} g(x) \to 0$ as $m \to \infty$. Thus under $g(\cdot)$, the set A_m^c converges to a set of zero probability, while under $k_m(\cdot)$ and $f_{\theta_m}(\cdot)$, the set A_m converges to a set of zero probability. Thus on A_m, $D(h_{\varepsilon,m}(x), f_{\theta_m}(x)) \to D((1-\varepsilon)g(x),0)$ as $m \to \infty$ and

$$\left|\int_{A_m} D(h_{\varepsilon,m}(x), f_{\theta_m}(x)) - \int_{g>0} D((1-\varepsilon)g(x),0)\right| \to 0$$

by dominated convergence theorem and Lemma 1. Notice that $\int_{g>0} D((1-\varepsilon)g(x),0) = \int D((1-\varepsilon)g(x),0) = (1-\varepsilon)(1/\bar{\tau})\log(1/\tau)$. Similarly we have

$$\left|\int_{A_m^c} D(h_{\varepsilon,m}(x), f_{\theta_m}(x)) - \int D(\varepsilon k_m(x), f_{\theta_m}(x))\right| \to 0.$$

Notice that $\int D(\varepsilon k_m(x),\ f_{\theta_m}(x)) \geq C_\tau(\varepsilon - 1)$ by Jensen's inequality. It follows that

$$\liminf_{n\to\infty} \mathrm{GKL}_\tau(h_{\varepsilon,n}, f_{\theta_m}) \geq C_\tau(\varepsilon - 1) + (1-\varepsilon)\frac{1}{\bar{\tau}}\log\left(\frac{1}{\tau}\right).$$

We will denote the right hand side of the above inequality by $a_1(\varepsilon)$.

We will have a contradiction to our assumption of the existence of a sequence $\{k_m\}$ for which breakdown occurs if we can show that there exists a constant value θ^* in the parameter space such that

$$\limsup_{m\to\infty} \mathrm{GKL}_\tau(h_{\varepsilon,m}, f_{\theta^*}) < a_1(\varepsilon) \tag{10}$$

as then the $\{\theta_m\}$ sequence above could not minimize GKL_τ for every m. We will show that this is true for all $\varepsilon < 1/2$ under the model where θ^* is the minimizer of $\int D((1-\varepsilon)g(x), f_\theta(x))$. Using analogous techniques, assumptions **A1**, **A2**, and Lemma 1 we obtain, for any fixed θ,

$$\begin{aligned}\lim_{m\to\infty} \mathrm{GKL}_\tau(h_{\varepsilon,m}, f_\theta) &= \frac{\varepsilon}{\bar{\tau}}\log\left(\frac{1}{\tau}\right) + \int D((1-\varepsilon)g(x), f_\theta(x)) \\ &\geq \frac{\varepsilon}{\bar{\tau}}\log\left(\frac{1}{\tau}\right) + \inf_\theta \int D((1-\varepsilon)g(x), f_\theta(x)).\end{aligned} \tag{11}$$

with equality for $\theta = \theta^*$. Let $a_2(\varepsilon) = (1/\bar{\tau})\log(1/\tau) + \int D((1-\varepsilon)g(x), f_{\theta^*}(x))$. Notice from (11) that among all fixed θ the divergence $\mathrm{GKL}_\tau(h_{\varepsilon,m}, f_\theta)$ is minimized in the limit by θ^*.

If $g(\cdot) = f_{\theta_t}(\cdot)$, that is the true distribution belongs to the model, $\int D((1-\varepsilon)f_{\theta_t}(x), f_{\theta_t}(x)) = C_\tau(-\varepsilon)$ which is also the lower bound (over $\theta \in \Theta$) for $\int D((1-\varepsilon)f_{\theta_t}(x), f_\theta(x))$. Thus in this case $\theta^* = \theta_t$, and from (11),

$$\lim_{m\to\infty} \mathrm{GKL}_\tau(h_{\varepsilon,m}, f_{\theta^*}) = \lim_{m\to\infty} \mathrm{GKL}_\tau(h_{\varepsilon,m}, f_{\theta_t}) = \frac{\varepsilon}{\bar{\tau}}\log\left(\frac{1}{\tau}\right) + C_\tau(-\varepsilon). \tag{12}$$

As a result asymptotically there is no breakdown for ε level contamination when $a_3(\varepsilon) < a_1(\varepsilon)$, where $a_3(\varepsilon)$ is the right hand sides of Eq. (12). Note that $a_1(\varepsilon)$ and $a_3(\varepsilon)$ are strictly decreasing and increasing respectively in ε, and $a_1(1/2) = a_3(1/2)$, so that asymptotically there is no breakdown and $\limsup_{m\to\infty} |T_\tau(H_{\varepsilon,m})| < \infty$ for $\varepsilon < 1/2$.

More generally when $g(\cdot)$ is not in the model, asymptotically there is no breakdown for $\varepsilon < \varepsilon^*$, where ε^* is given by $\varepsilon^* = \inf\{\varepsilon\colon a_1(\varepsilon) \leq a_2(\varepsilon)\}$. ■