A Minimum Distance Weighted Likelihood Method of Estimation

Arun Kumar Kuchibhotla, Ayanendranath Basu

Interdisciplinary Statistical Research Unit,
203 B.T.Road,
Kolkata - 108.

Abstract: Over the last several decades, minimum distance (or minimum divergence, minimum disparity, minimum discrepancy) estimation methods have been studied in different statistical settings as an alternative to the method of maximum likelihood. The initial motivation was probably to exhibit that there exists other estimators apart from the maximum likelihood estimator (MLE) which has full asymptotic efficiency at the model. As the scope of and interest in the area of robust inference grew, many of these estimators were found to be particularly useful in that respect and performed better than the MLE under contamination. Later, a weighted likelihood variant of the method was developed in the same spirit, which was substantially simpler to implement. In the statistics literature the method of minimum disparity estimation and the corresponding weighted likelihood estimation methods have distinct identities. Despite their similarities, they have some basic differences. In this paper we propose a method of estimation which is simultaneously a minimum disparity method and a weighted likelihood method, and may be viewed as a method that combines the positive aspects of both. We refer to the estimator as the minimum distance weighted likelihood (MDWL) estimator, investigate its properties, and illustrate the same through real data examples and simulations. We briefly explore the applicability of the method in robust tests of hypothesis.

MSC 2010 subject classifications: Primary 62F35, 62F12, 62F10; secondary 62F03.
Keywords and phrases: Disparity, Discrepancy, Minimum disparity estimation, Weighted Likelihood, Kernel Density Estimation, Integral Density Functional.

1. Introduction

Statistical inference based on density based distances has a long history and dates back at least to Pearson (1900). The maximum likelihood method itself may be viewed as a minimum distance method and therefore represents a particular case of minimum distance inference procedures. However, barring the maximum likelihood estimator, research activity in density based minimum distance estimation has been somewhat sporadic till the 1960s. Rao provided a rigorous treatment of first order efficiency (and the relatively more complicated second order efficiency) in the 1960s; see Rao (1961, 1962, 1963). The description of general chi-square distances in the form of phi-divergences was considered independently by Csiszár (1963) and Ali and Silvey (1966). There was a proliferation in the area of minimum distance methods of estimation in the 1970s, as evidenced by the bibliographic collection of Parr (1981); these include the work of Robertson (1972), Fryer and Robertson (1972) and Berkson (1980), among others.

It is important to point out that the measures of “distance” which we consider here are not necessarily mathematical metrics. Some of these measures are not symmetric in their arguments and some do not satisfy the triangle inequality. The only properties that we demand are that the measure should be non-negative and should equal zero if and only if the arguments are identically equal. For the sake of an unified notation, we refer to all such measures loosely as statistical distances or simply distances. Most of the measures that we consider will be statistical distances in the above sense. In particular, we will consider the class of disparities which is essentially the same as the class of ϕ-divergences.

The primary consideration in the literature till the late 1970s appears to have been the construction of a parallel method of estimation which is as efficient (or close in efficiency to) the method of maximum likelihood. The robustness angle originated with Beran (1977); he studied minimum Hellinger distance estimation in case of general continuous parametric models and proved asymptotic first order efficiency of the parameter estimator that minimizes the Hellinger distance between a kernel density estimator and a density from the model family. This work assumed somewhat restrictive conditions, but did exhibit many nice robustness properties of the estimator and subsequent research in minimum distance estimation was very significantly influenced by it. Since then, minimum disparity estimation has been studied from both the robustness and efficiency perspectives.

In case of discrete models, minimum disparity estimation was rigorously studied by Lindsay (1994) who established first order efficiency under fairly general conditions on the class of disparities. In the case of continuous models, Park and Basu (2004) provided a general framework albeit under somewhat restrictive conditions on the distance. Their approach excluded some common disparities such as the Pearson’s chi-square, the Hellinger distance and the likelihood disparity. But they did show that there are many families of disparities which satisfied their conditions and led to robust inference with full efficiency. Recently, this framework was extended to include all the popular disparities by Kuchibhotla and Basu (2015). Minimum disparity estimation has been used in various statistical scenarios. See Basu et al. (2011) for more details on applications of minimum disparity estimation.

The minimum disparity estimators derive their robustness from the fact that they downweight the outliers in the data. Using this idea, Markatou et al. (1998) proposed weighted likelihood estimating equations with smaller weights for outliers in the data. They also showed that one can choose weights corresponding to a minimum disparity estimation procedure. But their estimators are not minimizers of a proper objective function, although this estimation procedure has been extended to different scenarios. See Basu et al. (2011) for more details.

These two methods, minimum disparity estimation which describes a minimization problem and weighted likelihood estimation which describes a root solving problem, have been dealt with separately in the literature. Each of these methods have certain advantages which are specific to it. In case of minimum disparity estimation, the advantages are as follows: (i) the method has a valid objective function; (ii) selection of the correct root, when there are multiple zeros of the estimating function, is automatic; (iii) one can easily generate robust analogues of the likelihood ratio type tests; see, for example, Simpson (1989), and (iv) the presence of the objective function allows one to study the breakdown properties of the estimator using routine techniques without requiring a parallel disparity measure, e.g., Markatou et al. (1998, Sec. 8.2, Appx. A.3). On the other hand, the advantages of the weighted likelihood estimating equation method are as follows: (i) the estimating equation is now a sum of the observed data points, rather than an integral over the whole support, so that all the numerical evaluations related are substantially simpler, particularly in multivariate/multiparameter situation; (ii) the form of the estimating equation readily leads to an iterative reweighting algorithm similar to the iteratively reweighted least squares, and the computation of the estimator can avoid the evaluation of the second derivative Hessian matrix; a weighted likelihood equation with given fixed weights can be solved at one step for most common parametric models; Basu and Lindsay (2004) have described some of the simplifications that similar algorithms lead to in case of exponential families, and (iii) the final fitted weights give a measure of "fitness" of each individual observation in terms of their compatibility with the rest of the data given the parametric model, e.g., Markatou et al. (1998). Unfortunately, in either case, the advantages of the method are not shared by the other.

In this paper we present a formulation where the minimum disparity estimation procedure can be equivalently described as a weighted likelihood estimation procedure, so that the advantages of the two methods are combined in this particular formulation. We believe that this formulation increases the scope and the applicability of both these methods. We also provide a general proof for our *minimum distance weighted likelihood* estimator under fairly accessible conditions following the approach of Kuchibhotla and Basu (2015). Taken together with the latter paper, the current manuscript provides the general continuous analogue to Lindsay (1994) and integrates it within the weighted likelihood set up.

The outline of the rest of the paper is as follows. In Section 2, we introduce the procedure of minimum disparity estimation and also the modification to view this as a weighted likelihood procedure. In Section 3, we derive the asymptotic results under fairly general conditions in the spirit of Kuchibhotla and Basu (2015). In Section 4, we study the robustness properties of our estimators. In Section 5, we introduce a computational algorithm for fitting finite mixture models which is similar to the EM algorithm in this case. In Section 6, we apply our methodology on some real datasets. In Section 7, we discuss applications of our procedure in robust testing of hypothesis. In Section 8, we conclude with some remarks and ideas about future directions. The lengthier proofs and several additional real data examples are provided in Supplementary material.

2. Minimum Disparity and Weighted Likelihood

Let $\mathcal{G}$ represent the class of all distribution functions having densities with respect to the Lebesgue measure. We assume that the true distribution G and the model $\mathcal{F}_\theta = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ belong to $\mathcal{G}$. Let g and f_θ be the corresponding densities. Here we do not assume that $g = f_{\theta_0}$ for some $\theta_0 \in \Theta$, but assume that g is "close" to f_{θ_g} for some $\theta_g \in \Theta$ in some appropriate sense. Let $X_1, X_2, \ldots, X_n$ be a random sample from G which is modelled by $\mathcal{F}_\theta$. Our aim is to estimate the parameter θ_g by choosing the model density which gives the "closest" fit to the data.

Let C be a thrice differentiable convex function defined on $[-1, \infty)$, satisfying $C(0) = 0$. Define

$$\rho_C(g, f_\theta) = \int C\left(\frac{g(x)}{f_\theta(x)} - 1\right) f_\theta(x)dx. \tag{2.1}$$

This form describes the class of disparities between the densities g and f_θ. A simple application of Jensen's inequality shows that $\rho_C(g, f_\theta) \geq 0$ with equality if and only if $g = f_\theta$ identically. Without changing the disparity, one can standardize the disparity generating function $C(\cdot)$ by requiring $C'(0) = 0$ and $C''(0) = 1$. We denote by θ_g or $T(G)$, the "best fitting parameter" which minimizes $\rho_C(g, f_\theta)$ over all $\theta \in \Theta$. We consider the minimum disparity estimator $\hat{\theta}_n$ of θ_g defined by

$$\hat{\theta}_n := \arg\min_\theta \rho_C(g_n, f_\theta), \tag{2.2}$$

where g_n is a kernel density estimator obtained from the sample. Under differentiability of the model, $\hat{\theta}_n$ can be obtained as a root of the equation

$$\int A(\delta_n(x))\nabla f_\theta(x)dx = 0, \tag{2.3}$$

where ∇ represents the gradient with respect to θ,

$$A(\delta) = C'(\delta)(\delta + 1) - C(\delta) \quad \text{and} \quad \delta_n(x) + 1 = g_n(x)/f_\theta(x).$$

Here C' represents the derivative of C with respect to its argument. Primes and double primes will be employed to denote the first and the second derivatives of relevant functions throughout the manuscript. The function $A(\cdot)$ is called the residual adjustment function (RAF) of the disparity and δ_n is referred to as the Pearson residual. Convexity of C implies that the function $A(\cdot)$ is an increasing function. The function $A(\cdot)$ plays a very crucial role in determining the robustness properties of the estimator. Under the standardization of $C(\cdot)$, we get $A(0) = 0$ and $A'(0) = 1$. See Basu et al. (2011) for more details.

Observe that the objective function $\rho_C(g_n, f_\theta)$ is the same as the one in (2.1) except that g is replaced by its nonparametric density estimator g_n; $\rho_C(g_n, f_\theta)$ is a natural estimator of $\rho_C(g, f_\theta)$. Having both the objective function and the estimating function defined in terms of an integral can make the estimation procedure difficult for a practitioner because of integral calculations at every iterative step; one also needs to look for convergence of the numerical integral to the actual one. This can be particularly difficult if the observed data is in a higher dimension and the objective function involves multiple integrals.

Define, $\delta(x) + 1 = g(x)/f_\theta(x)$. Notice that

$$\begin{aligned}\rho_C(g, f_\theta) = \int C(\delta) f_\theta dx &= \int \{C(\delta) + k\delta\} f_\theta dx \\ &= \int \frac{\{C(\delta) + k\delta\} f_\theta}{g} g dx \\ &= \int \frac{C(\delta) + k\delta}{\delta + 1} g dx \\ &= \mathbb{E}_g\left[\frac{C(\delta) + k\delta}{\delta + 1}\right],\end{aligned}$$

for every $k \in \mathbb{R}$, since $\int \delta(x) f_\theta(x) dx = 0$. So, consider the modification

$$\mathbb{E}_g\left[\frac{C(\delta(X)) + k\delta(X)}{\delta(X) + 1}\right] \approx \mathbb{E}_g\left[\frac{C(\delta(X)) + k\delta(X)}{\delta(X) + 1}\mathbb{1}_{\{X \in A_n\}}\right], \tag{2.4}$$

for some sequence of sets $A_n \uparrow \mathbb{R}$ as $n \uparrow \infty$. In practice, we take

$$\frac{1}{n}\sum_{i=1}^{n} \frac{C(\delta_n(X_i)) + k\delta_n(X_i)}{\delta_n(X_i) + 1}\mathbb{1}_{\{X_i \in A_n\}},$$

as an empirical estimate of the right hand side of Equation (2.4). In this paper, we use

$$A_n = \{x : g_n(x) > \gamma_n/2\}$$

for some $\gamma_n \downarrow 0$ as $n \uparrow \infty$ at some rate to be mentioned later. Here we are trimming the tails in order to avoid dividing by small values since having $g_n(x)$ in the denominator might cause numerical instability for x in the tails. We anticipate that it is also possible to proceed without trimming since the denominator actually contains $g_n(x)$ and $f_\theta(x)$ both of which converge to zero as $x \to \infty$, but we do not deal with this in this paper. However in models like the normal or the exponential where the tails decay exponentially such trimmings are generally not necessary.

Instead of considering the criterion function on the right hand side of (2.2) as an estimate of $\rho_C(g, f_\theta)$, consider the right hand side of (2.4) as the estimate. Since, the objective function now is an average over the sample at hand, the estimating function will also be an average. It is easy to see that the estimating equation is given by

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_{n,i}\frac{A(\delta_n(X_i)) + k}{\delta_n(X_i) + 1} u_\theta(X_i) = 0, \tag{2.5}$$

where $\kappa_{n,i} = \mathbb{1}_{\{X_i \in A_n\}}$, and $u_\theta(x) = \nabla \ln f_\theta(x)$. Denote by $\Psi_n(\theta)$ the expression on the left hand side of Equation (2.5); $\Psi_n(\theta)$ is our estimating function.

Comparing with the ordinary likelihood score equation

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) = 0$$

it may be seen that Equation (2.5) is basically a weighted likelihood estimating equation. This form of the estimating equation makes it clear why one would require $|A(\delta)| \leq |\delta|$ in order to get an estimator which is robust to both outliers and inliers. Also, note that if this inequality holds the weights will all be bounded by 1. In this paper, we take $k = 1$, but the proofs will go through with any real constant. One might want to take $k = -A(-1) = C(-1)$ to make sure that the weights are all non-negative. If we take $k = 1$ and $A(t) = t$ for $t \in [-1, \infty)$, then the estimating equation will exactly coincide with the likelihood equation whenever $\kappa_{n,i}$'s are all equal to one. We refer to the estimator obtained as the solution of Equation (2.5) as the minimum distance weighted likelihood (MDWL) estimator.

This type of estimation of integral functionals of density, in which we replace the expectation by an average and the unknown density by a nonparametric density estimator, is not entirely new in the density functional estimation literature. Joe (1989) used this idea in estimating functionals of the type $\int J(f)f dx$ for some thrice differential function J and gave expressions for the bias and the variance of the estimator. Giné and Mason (2008) also used the same idea for functionals of the type

$$\int \phi(x, F(x), F^{(1)}(x), \ldots, F^{(k)}(x))dF(x),$$

for some twice differentiable function ϕ with certain boundedness assumptions where F is the unknown distribution function. They proved uniform in bandwidth asymptotic normality of their estimator. See also the references therein.

We will not get into a geometric description of the robustness of the proposed estimator obtained as a solution of Equation (2.5), as all the interpretations and insights provided by Lindsay (1994) and Markatou et al. (1998) also remain valid in our context. Clearly, a residual adjustment function $A(\delta)$, which exhibits a severely dampened response to increasing δ exhibits greater local robustness. On the other hand the coefficient of $u_\theta(X_i)$ may be looked upon as weights and therefore as a measure of the fitness of the observation X_i in the parametric estimation scheme. In this respect the method of estimation described in this section can be considered to be a minimum distance estimation method as well as a weighted likelihood estimation method. Thus, although the estimator is generated by a legitimate optimization process, it automatically generates a measure of fitness corresponding to each X_i as described above. More generally the MDWL combines the positive aspects of minimum disparity and weighted likelihood estimation.

See Lindsay (1994), Markatou et al. (1998) and Basu et al. (2011) for an expanded discussion on the role of the residual adjustment function in robust estimation.

3. Asymptotic Results

Before proceeding to the assumptions and the proof of asymptotic normality, we provide a short discussion on the existing literature which deals with estimating functions like $\Psi_n(\theta)$ which are averages over some function with a nonparametric function estimate involved. In econometrics and empirical processes theory, these type of estimators are called semi-parametric M-estimators. Newey and McFadden (1994) discuss different assumptions under which one can prove asymptotic normality of these estimators. Andrews (1994) also gives asymptotic results via stochastic equicontinuity. These procedures applied in our case will readily lead to asymptotic normality of the estimator but will be under a restrictive class of disparities. In the case of weighted likelihood estimating equation also, the proofs that are available in literature use restrictive boundedness assumptions on the residual adjustment function. We will provide a new proof, along the lines of Kuchibhotla and Basu (2015), which operates on assumptions similar to those in the latter paper.

We use the theorems of Yuan and Jennrich (1998) to prove the asymptotic normality of the estimator. So, we only prove asymptotic normality of $\Psi_n(\theta)$ and uniform convergence of the derivative of $\Psi_n(\theta)$ to a non-random function of θ. We refer the reader to Yuan and Jennrich (1998) for more details. Our necessary assumptions are detailed in the next subsection.

3.1. Assumptions

The nonparametric density estimator g_n based on independent and identically distributed observations $X_1, X_2, \ldots, X_n$ is given by

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

where K is the kernel function and h_n is the bandwidth. In the following $A(\delta)$ will represent the residual adjustment function of the disparity ρ_C.

(A1) $A''(\delta)(\delta+1)^\alpha$ is bounded for some fixed α, i.e, $|A''(\delta)(\delta+1)^\alpha| \le M < \infty$ for some α and for all $\delta \ge -1$, where $1+\delta(x) = g(x)/f_\theta(x)$. All the other instances of α in the assumptions relate to this specific value.

(A2) The support of f_θ is independent of θ and is same as the support of g.

(A3) The density g is twice differentiable. Also, the first and the second derivatives g', g'' are bounded

(A4) The kernel K is symmetric and has a compact support denoted by Ω; $h_n \to 0, nh_n \to \infty$, as $n \to \infty$.

(A5) The trimming sequence $\{\gamma_n\}$ is assumed to satisfy, in conjunction with the bandwidth sequence, the following: $h_n/\gamma_n \to 0$, $n^{1/2}h_n^2/\gamma_n \to 0$ and $nh_n\gamma_n^2/\ln(1/h_n) \to \infty$ as $n \to \infty$.

(A6) Let $D_n := \{x : g(x) \geq \gamma_n\}$. Then trimming sequence $\{\gamma_n\}$ satisfies, in association with A, the following conditions:

$$\int_{D_n^c} A(\delta(x))\nabla f_\theta(x)dx = o_p(n^{-1/2}),$$
$$\int_{D_n^c} |A'(\delta(x))\nabla f_\theta(x)|dx = o_p(n^{-1/2}),$$
$$\int_{D_n^c} \frac{f_\theta^{\alpha-1}(x)}{g^{\alpha-2}(x)}|u_\theta(x)|dx = O(1).$$

Here D_n^c represents the complement of D_n.

(A7) The random vectors $[A(\delta(X))+1]\nabla f_\theta(X)/g(X)$, $A'(\delta(X))u_\theta(X)$ and $f_\theta^{\alpha-1}(X)u_\theta(X)/g^{\alpha-1}(X)$ have component-wise finite moments of some order strictly greater than 2.

(A8) There exists a compact subset Θ_0 of Θ, which is a neighbourhood of θ_g such that

$$T_\theta|[A(\delta(X))+1]\nabla_2 f_\theta(X)|/g(X), \qquad T_\theta|A'(\delta(X))\nabla_2 f_\theta(X)|/f_\theta(X),$$
$$T_\theta f_\theta^{\alpha-1}(X)|\nabla_2 f_\theta(X)|/g^{\alpha-1}(X), \quad T_\theta f_\theta^{\alpha-1}(X)u_\theta(X)u_\theta^\top(X)/g^{\alpha-1}(X)$$

are all finite. Here T_θ is used to denote the operator $E\sup_{\theta\in\Theta_0}$. Also, $\int|A(\delta(x))\nabla_2 f_\theta(x)|dx$ and $\int|A'(\delta(x))|g(x)u_\theta(x)u_\theta^T(x)dx$ are both finite for $\theta \in \Theta_0$.

(A9) $V(\theta)$ is finite and positive definite and $B(\theta)$ is non-zero for $\theta = T(G)$, where

$$\begin{aligned} V(\theta) &= \lim_{n\to\infty} \mathrm{Var}\left[\int K_{h_n}(x-X_1)A'(\delta(x))u_\theta(x)dx\right] \\ &= \mathrm{Var}\left[A'(\delta(X_1))u_\theta(X_1)\right], \\ B(\theta) &= \int A(\delta(x))\nabla_2 f_\theta dx - \int A'(\delta(x))(\delta(x)+1)u_\theta u_\theta^T f_\theta dx. \end{aligned}$$

Here ∇_2 represents the second derivative with respect to θ.

3.2. Estimating Function

For the trimming sequence $\{\gamma_n\}$ satisfying the assumption **(A5)**, define $\kappa_i = \mathbb{1}_{\{g(X_i)\geq\gamma_n\}}$. Notice that the definition of $\kappa_{n,i}$ involves the kernel density estimate g_n and κ_i is based on the actual density g. For simplicity of notation, we will drop the subscript n from h_n, unless specifically demanded by the situation.

Define the function

$$T_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \kappa_i \frac{C(\delta_n(X_i)) + \delta_n(X_i)}{\delta_n(X_i)+1}.$$

Here δ and δ_n both depend on θ. The corresponding estimating function (derivative) is given by

$$\nabla T_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \kappa_i \frac{A(\delta_n(X_i)) + 1}{\delta_n(X_i) + 1} u_\theta(X_i).$$

In proving asymptotic normality of the estimating function $\Psi_n(\theta)$, we will closely follow the method of Lewbel (1997) and will prove the following:

- *Step 1:*

$$n^{1/2}\left(\nabla T_n(\theta) - \int A(\delta(x))\nabla f_\theta(x)dx\right) \xrightarrow{\mathcal{L}} N(0, V(\theta)).$$

- *Step 2:*

$$\nabla T_n(\theta) - \Psi_n(\theta) = o_p(n^{-1/2}).$$

We now state three propositions which will be used in the proofs of the lemmas and theorems related to step 1. Proofs of all these results are deferred to the Supplementary Material.

Proposition 3.1. *If w is a measurable function such that $w(X_1)$ has finite mean and $w(X_1)g(X_1)$ has finite second moment, then*

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)\left[g_n(X_i) - g(X_i)\right] - \int w(y)\left[g_n(y) - g(y)\right]g(y)dy = o_p(n^{-1/2}).$$

Remark 3.1 Here the assumption that $w(X_1)$ has finite mean is used only to control the asymptotic bias. This assumption can be relaxed if instead of $w(x)$, we have $w(x)\mathbb{1}_{\{x\in D_n\}}$. This relaxation results in,

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_i w(X_i)\left[g_n(X_i) - g(X_i)\right] - \int_{D_n} w(y)\left[g_n(y) - g(y)\right]g(y)dy = o_p(n^{-1/2}).$$

Proposition 3.2. *If t is a measurable function such that $t(X_1)$ has finite second moment, then*

$$\mathcal{S}_n := \frac{1}{n}\sum_{i=1}^{n} \kappa_i t(X_i) - \int_{D_n} t(x)g_n(x) = o_p(n^{-1/2}). \tag{3.1}$$

Proposition 3.3. *If w is a measurable function such that $w(X_1)/g(X_1)$ has finite expectation, then*

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_i w(X_i)\left[\frac{1}{g_n(X_i)} - \frac{1}{g(X_i)}\right] = -\frac{1}{n}\sum_{i=1}^{n} \kappa_i \frac{w(X_i)}{g(X_i)}\frac{g_n(X_i) - g(X_i)}{g(X_i)} + o_p(n^{-1/2}).$$

Lemma 3.1. *Under the assumptions (A1)-(A7),*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in D_n\}} W_n(X_i)\frac{\nabla f_\theta(X_i)}{g_n(X_i)} = o_p(n^{-1/2}),$$

where $W_n(x) = \left[A(\delta_n(x)) - A(\delta(x)) - A'(\delta(x))(\delta_n(x) - \delta(x))\right].$

Theorem 3.1. *Under the assumptions (A1) - (A7),*

$$\nabla T_n(\theta) - \int A(\delta(x))\nabla f_\theta(x)dx = \int A'(\delta(x))\left[g_n(x) - g(x)\right] u_\theta(x)dx + o_p(n^{-1/2}).$$

The following theorem presents the statement of step 2.

Theorem 3.2. *Under the assumptions (A1)-(A7),* $\nabla T_n(\theta) - \Psi_n(\theta) = o_p(n^{-1/2})$.

Corollary 3.1. *Under the assumptions (A1)-(A7),*

$$n^{1/2}\left(\Psi_n(\theta) - \int A(\delta(x))\nabla f_\theta(x)dx\right) \overset{\mathcal{L}}{\to} N(0, V(\theta)).$$

Proof. By Theorem 3.1 and Theorem 3.2, we have that the asymptotic distribution of

$$n^{1/2}\left(\Psi_n(\theta) - \int A(\delta(x))\nabla f_\theta(x)dx\right)$$

is same as that of

$$n^{1/2}\int A'(\delta(x))u_\theta(x)\left[g_n(x) - g(x)\right]dx.$$

Also, note that by Proposition 3.2, we have that

$$\int A'(\delta(x))u_\theta(x)g_n(x)dx - \frac{1}{n}\sum_{i=1}^{n} A'(\delta(X_i))u_\theta(X_i) = o_p(n^{-1/2}).$$

By central limit theorem for iid random variables and the assumption that $V(\theta)$ is non-singular, we get that

$$n^{1/2}\int A'(\delta(x))\nabla f_\theta(x)\left[g_n(x) - g(x)\right]dx \overset{\mathcal{L}}{\to} N(0, V(\theta)).$$

See Kuchibhotla and Basu (2015) for more details. □

Remark 3.2 Theorems 3.1 and 3.2 combined proves that

$$\Psi_n(\theta) - \int A(\delta(x))\nabla f_\theta(x)dx = \int A'(\delta(x))\left[\delta_n(x) - \delta(x)\right]\nabla f_\theta(x)dx + o_p(n^{-1/2}).$$

Proposition 1 of Kuchibhotla and Basu (2015) proves that

$$\int [A(\delta_n(x)) - A(\delta(x))]\nabla f_\theta(x)dx$$
$$= \int A'(\delta(x))\left[\delta_n(x) - \delta(x)\right]\nabla f_\theta(x)dx + o_p(n^{-1/2}).$$

These two statements compared yields,

$$\Psi_n(\theta) - \int A(\delta_n(x))\nabla f_\theta(x)dx = o_p(n^{-1/2}).$$

Under $g = f_{\theta_0}$ we have, by Theorems 3.1 and 3.2 and using $A'(0) = 1$,

$$\Psi_n(\theta_0) = \int \delta_n(x)\nabla f_{\theta_0}(x)dx + o_p(n^{-1/2}) = \int g_n(x)u_{\theta_0}(x)dx + o_p(n^{-1/2}).$$

Thus, by Proposition 3.2, we get

$$\Psi_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} u_{\theta_0}(X_i) + o_p(n^{-1/2}). \tag{3.2}$$

3.3. The Derivative of the Estimating Function

The derivative of the estimating function $\Psi_n(\theta)$ is given by,

$$\nabla\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in A_n\}}\left[\frac{A(\delta_n(X_i))+1}{g_n(X_i)}\nabla_2 f_\theta(X_i) - A'(\delta_n(X_i))u_\theta(X_i)u_\theta^\top(X_i)\right].$$

We will prove uniform (in θ) convergence of $\nabla\Psi_n(\theta)$ to a non-stochastic function in the following sequence of Lemmas.

Lemma 3.2. *Under assumptions **(A5)** and **(A8)**,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in A_n\}}\left[\frac{A(\delta_n(X_i))+1}{g_n(X_i)}\nabla_2 f_\theta(X_i) - \frac{A(\delta(X_i))+1}{g(X_i)}\nabla_2 f_\theta(X_i)\right] = o_p(1),$$

uniformly in $\theta \in \Theta_0$.

Lemma 3.3. *Under assumptions **(A5)** and **(A8)**,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in A_n\}}\left[A'(\delta_n(X_i)) - A'(\delta(X_i))\right]u_\theta(X_i)u_\theta^\top(X_i) = o_p(1),$$

uniformly in $\theta \in \Theta_0$.

For notational ease, define

$$K_\theta(x) = \frac{A(\delta(x))+1}{g(x)}\nabla_2 f_\theta(x) - A'(\delta(x))u_\theta(x)u_\theta^\top(x)$$

for the next Lemma.

Lemma 3.4. *Under assumptions **(A3)** - **(A5)** and **(A8)**,*

$$\frac{1}{n}\sum_{i=1}^{n} K_\theta(X_i)\left[\mathbb{1}_{\{X_i\in A_n\}} - \mathbb{1}_{\{X_i\in D_n\}}\right] = o_p(1),$$

uniformly in $\theta \in \Theta_0$.

Theorem 3.3. *Under assumptions **(A1)** - **(A9)**, $\nabla\Psi_n(\theta) \xrightarrow{P} B(\theta)$, uniformly in $\theta \in \Theta_0$.*

Proof of Theorem 3.3. This theorem follows from Lemmas 3.2, 3.3 and 3.4. Here we also need assumption **(A9)** in order to ensure that the tail integral of $B(\theta)$ converges to zero. □

Theorem 3.4. *Under assumptions* ***(A1)*** *-* ***(A9)****, there exists a zero of* $\Psi_n(\theta)$*,* $\hat{\theta}_n$*, which converges almost surely to* θ_g *and*

$$n^{1/2}(\hat{\theta}_n - \theta_g) \overset{\mathcal{L}}{\to} N(0, B^{-1}(\theta_g)V(\theta_g)B^{-1}(\theta_g)).$$

Proof of Theorem 3.4. The proof follows from Corollary 3.1 and Theorem 3.3 using Theorems 1, 2 and 4 of Yuan and Jennrich (1998). □

Remark 3.3 Theorem 3.4 parallels the general asymptotic normality results of Lindsay (1994), Park and Basu (2004), Markatou et al. (1998) and Kuchibhotla and Basu (2015).

Remark 3.4 Under the model, $g = f_{\theta_0}$ with $\theta_0 \in \Theta$, we get $\theta_g = \theta_0$ and $B(\theta_0) = -I(\theta_0)$ and $V(\theta_0) = I(\theta_0)$, where $I(\theta_0)$ represents the Fisher information matrix. Therefore, in this case, we get

$$n^{1/2}(\hat{\theta}_n - \theta_0) \overset{\mathcal{L}}{\to} N(0, I^{-1}(\theta_0)).$$

Remark 3.5 Since, $\Psi_n(\hat{\theta}_n) = 0$, a Taylor series expansion of $\Psi_n(\theta)$ with respect to θ around θ_0, exhibits,

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \nabla\Psi_n(\theta^*)(\hat{\theta} - \theta_0),$$

for some θ^* belonging to the line joining the points $\hat{\theta}_n$ and θ_0. Hence, we get

$$\hat{\theta}_n - \theta_0 = -\left[\nabla\Psi_n(\theta^*)\right]^{-1}\Psi_n(\theta_0).$$

Now by Theorem 3.3,

$$-\left[\nabla\Psi_n(\theta^*)\right]^{-1} = -\left[B(\theta_0)\right]^{-1} + o_p(1),$$

Hence, under the model $g = f_{\theta_0}$ we get, by Equation (3.2), the representation

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{1/2}\left[I^{-1}(\theta_0)\right]\frac{1}{n}\sum_{i=1}^{n} u_{\theta_0}(X_i) + o_p(1), \tag{3.3}$$

proving that the estimator is first order efficient and

$$n^{1/2}(\hat{\theta}_n - \hat{\theta}_{ML}) = o_p(1), \tag{3.4}$$

where $\hat{\theta}_{ML}$ represents the unrestricted maximum likelihood estimator.

Remark 3.6 Remark 3.2 paired with the arguments in Remark 3.5 also proves that $n^{1/2}(\hat{\theta}_{MDWL} - \hat{\theta}_{MD}) = o_p(1)$ under any distribution G, where $\hat{\theta}_{MDWL}$ and $\hat{\theta}_{MD}$ represents the MDWL estimator and the minimum disparity estimator corresponding to the common disparity generating function $C(\cdot)$.

4. Robustness Properties

We will now provide asymptotic robustness results for the MDWL estimator which is a zero of the equation $\Psi_n(\theta) = 0$. We follow the approaches of Lindsay (1994) and Park and Basu (2004).

4.1. First and Higher Order Influence Analysis

The influence function, loosely speaking, calculates the derivative of the estimator functional, at zero, with respect to the proportion of contamination at a given point y. Lindsay (1994) demonstrated that the first order influence function of the minimum disparity estimator is the same as that of the maximum likelihood estimator under the model and therefore the first order influence function is not a very good indicator of the robustness of these estimators. Lindsay (1994) suggested taking one more derivative of the functional and demonstrated that the second order influence function can better approximate, often substantially, the bias of the estimator compared to the first. The theorem below gives the influence functions of the first and second order and expresses the second as a function of the first.

Theorem 4.1. *The influence function of the minimum disparity estimator functional T at G has the form $T'(y) = D^{-1}N$, where*

$$\begin{aligned} N &= A'(\delta(y))u_{\theta_g}(y) - E\left[A'(\delta(X))u_{\theta_g}(X)\right] \\ D &= E\left[A'(\delta(X))u_{\theta_g}(X)u_{\theta_g}^\top(X)\right] - \int A(\delta(x))\nabla_2 f_{\theta_g}(x)dx. \end{aligned}$$

Let $T(y) = \theta_\epsilon$ represent the functional corresponding to the contaminated density $g_\epsilon = (1-\epsilon)g + \epsilon\Delta_y$, Δ_y represents the density of a random variable degenerate at y and $\theta_g = T(G)$. Moreover, if $g = f_\theta$ for some $\theta \in \Theta \subset \mathbb{R}$, then

$$T''(y) = T'(y)[m_1(y) + A''(0)m_2(y)]/I(\theta),$$

where $I(\theta)$ represents the Fisher information and

$$\begin{aligned} m_1(y) &= 2\nabla u_\theta(y) - 2E[\nabla u_\theta(X)] + T'(y)E[\nabla_2 u_\theta(X)], \\ m_2(y) &= \frac{I(\theta)}{f_\theta(y)} + E[u_\theta^3(X)]\frac{u_\theta(y)}{I(\theta)} - 2u_\theta^2(y). \end{aligned}$$

Here $T'(y)$ and $T''(y)$ are the first and the second derivative of the functional $T(y)$ evaluated at $\epsilon = 0$.

Proof of Theorem 4.1. Direct differentiation of the estimating equation corresponding to the contaminated density g_ϵ gives

$$D_\epsilon \frac{\partial}{\partial\epsilon}\theta_\epsilon = N_\epsilon, \tag{4.1}$$

where

$$D_\epsilon = \int A'(\delta_\epsilon(x))g_\epsilon(x)u_{\theta_\epsilon}(x)u_{\theta_\epsilon}^\top(x)dx - \int A(\delta_\epsilon(x))\nabla_2 f_{\theta_\epsilon(x)}$$

$$N_\epsilon = A'(\delta_\epsilon(y))u_{\theta_\epsilon}(y) - \int A'(\delta_\epsilon(x))u_{\theta_\epsilon}(x)g_\epsilon(x)dx,$$

which immediately leads to the formula for influence function at the model by evaluating at $\epsilon = 0$.

Differentiating Equation (4.1) a second time with respect to ϵ, gives

$$D_\epsilon T''(y) + \frac{\partial}{\partial\epsilon}D_\epsilon T'(y) = \frac{\partial}{\partial\epsilon}N_\epsilon.$$

Hence,

$$T''(y) = D_\epsilon^{-1}\left[\frac{\partial}{\partial\epsilon}N_\epsilon - \frac{\partial}{\partial\epsilon}D_\epsilon T'(y)\right].$$

Calculating the required derivatives and evaluating them at $\epsilon = 0$ using the assumption $g = f_\theta$ implies the stated result. See Basu et al. (2011, p. 134) for more detailed calculations. □

Remark 4.1 We can do a second order influence function study using the MDWL estimators along the lines of the analysis done by Lindsay (1994) and can produce results, examples and graphs similar to those presented by Basu et al. (2011); exactly the same kind of interpretations hold, and the second order predicted bias of our estimators demonstrate similar improvements as presented by these authors. The interested reader can look up the description in Section 4.4 of Basu et al. (2011). Our estimators exhibit exactly the same kind of improvements. For brevity, we refrain from presenting such results in this paper.

Remark 4.2 Similar calculations can also be done in higher dimensions of θ but the derivative expressions will get more complicated and also interpreting the result would be harder.

4.2. Breakdown Point

The breakdown point of a statistical functional can be thought of as the smallest fraction of contamination in the data that may cause an extreme change in the functional. We can derive asymptotic breakdown points for our estimators using the results of Park and Basu (2004) which were given under fairly general conditions. The key conditions on the disparity in the Park and Basu (2004) approach are the finiteness of $C(-1)$ and $C'(\infty)$.

Consider the contamination model,

$$H_{\varepsilon,n} = (1-\varepsilon)G + \varepsilon K_n,$$

where $\{K_n\}$ is a sequence of contaminating distributions. Let $h_{\varepsilon,n}, g, k_n$ represent the corresponding densities. Following Simpson (1987), we state that breakdown occurs for the functional T at ε level contamination if there exists a contaminating sequence $\{K_n\}$ such that $|T(H_{\varepsilon,n}) - T(G)| \to \infty$ as $n \to \infty$. Under the conditions stated below, Theorem 4.1 of Park and Basu (2004) is directly applicable in case of the MDWL estimators. For the sake of completeness, we state the result below without repeating the proof. Define $\theta_n = T(H_{\varepsilon,n})$.

The list of assumptions needed for this theorem in respect of the contaminating sequence $\{k_n\}$, the truth g and the model f_θ are as follows.

(B1) $\int \min\{g(x), k_n(x)\}dx \to 0$ as $n \to \infty$.
(B2) $\int \min\{f_\theta(x), k_n(x)\}dx \to 0$ as $n \to \infty$ uniformly for $|\theta| \le c$, for any fixed c.
(B3) $\int \min\{g(x), f_{\theta_n}(x)\}dx \to 0$ as $n \to \infty$ if $|\theta_n| \to \infty$ as $n \to \infty$.
(B4) $C(-1)$ and $C'(\infty)$ are finite.

Theorem 4.2 (Theorem 4.1, Park and Basu (2004))**.** *Under the assumptions **(B1)** - **(B4)** above, the asymptotic breakdown point of the MDWL estimator is atleast* $1/2$ *at the model.*

5. Computational Algorithms

We have already pointed out that the weighted likelihood representation allows us to use a simple fixed point iterative reweighting algorithm for the evaluation of the estimators. While in our actual illustrations we will restrict ourselves to standard parametric models in this paper, we also describe here a appropriate computational algorithm to fit the MDWL method for finite mixture models. Our proposal may be considered to belong to the class of MM (Majorization–Minimization) algorithms; we primarily deal with the minimization part only. For a fuller description of the MM method, see Hunter and Lange (2004). One of the many advantages offered by MM is that it ensures a descent property and thus offers a numerically stable algorithm.

Let $\theta^{(m)}$ represent a fixed value of the parameter θ, and let $h(\theta|\theta^{(m)})$ denote a real-valued function of θ whose form depends on $\theta^{(m)}$. We say $h(\theta|\theta^{(m)})$ majorizes a function $k(\theta)$, if

$$k(\theta) \le h(\theta|\theta^{(m)}) \text{ for all } \theta \ne \theta^{(m)} \quad \text{and} \quad k(\theta^{(m)}) = h(\theta^{(m)}|\theta^{(m)}).$$

We now minimize the majorizing function instead of the function itself. Define,

$$\theta^{(m+1)} = \text{argmin}_\theta h(\theta|\theta^{(m)}).$$

This implies that

$$k(\theta^{m+1}) \le h(\theta^{(m+1)}|\theta^{(m)}) \le h(\theta^{(m)}|\theta^{(m)}) \le k(\theta^m).$$

Hence, the descent property of the algorithm follows. Possibly, the most difficult part in applying this technique is to get a “simple” majorizing function. The EM algorithm which was brought into limelight by Dempster et al. (1977) can be shown to be a special case of MM. See (Lange, 2010, pg. 226) for more details.

5.1. Finite Mixture Models

Cutler and Cordero-Braña (1996) proposed an EM-type algorithm called HMIX for fitting finite mixture models in the continuous case using Hellinger distance. Karlis and Xekalaki (1998) also proposed an EM-type algorithm called HELMIX for Poisson mixture models using Hellinger distance. These two algorithms are very similar and convergence properties of these two iterative algorithms were exhibited by Monte Carlo studies, but no theoretical properties were derived. Fujisawa and Eguchi (2006) also proposed an EM-type algorithm in the case of density power divergences introduced by Basu et al. (1998). In this section, we present an algorithm for fitting finite mixture model with distributions in the mixture having densities with respect to some common dominating measure using our minimum distance weighted likelihood method. This algorithm is not specific to the Hellinger distance and it is straightforward to extend this algorithm to all minimum disparity estimation procedures. Also, we show that this algorithm has the descent property similar to the ascent property of the EM algorithm. In particular, we show that all these algorithms belong to a class of algorithms governed by MM methodology.

Getting back to the minimization problem at hand, if C is a convex function, then the map Y defined by $t \mapsto tC(-1+[a/t])$ is also convex for all $a \geq 0$. Thus, for any two vectors $u, v \in \mathbb{R}^k$ with all components non-negative, we get by convexity,

$$Y(u^\top v) \leq \sum_{j=1}^{k} \frac{u_j^{(m)} v_j^{(m)}}{u^{(m)\top} v^{(m)}} Y\left(\frac{u^{(m)\top} v^{(m)}}{u_j^{(m)} v_j^{(m)}} u_j v_j\right), \tag{5.1}$$

for any two vectors $u^{(m)}, v^{(m)} \in \mathbb{R}^k$ with all components non-negative. Now, take $u^\top v = \sum_{j=1}^{k} w_j f_j(x;\theta_j)$ where $w_j \geq 0$ for $1 \leq j \leq k$, $\sum_{j=1}^{k} w_j = 1$, $f_j(x;\theta_j)$ represents a probability density evaluated a fixed point x with parameter θ_j. Note that here we are not assuming that the densities in the mixture model are from the same parametric family. In this case, $u^{(m)}$ and $v^{(m)}$ can be taken as vector of weights in the past iterate and as vector of probability densities with parameters obtained in the past iterate respectively. That is $u^{(m)} = (w_1^m, w_2^{(m)}, \ldots, w_k^{(m)})$ and $v^{(m)} = v^{(m)}(x) = (f_1(x;\theta_1^{(m)}), f_2(x;\theta_2^{(m)}), \ldots, f_k(x;\theta_k^{(m)}))$. Define, following Cutler and Cordero-Braña (1996),

$$a_j(x;\theta^{(m)}) = \frac{w_j^{(m)} f_j(x;\theta_j^{(m)})}{u^{(m)\top} v^{(m)}(x)}, \quad \text{and} \quad \tilde{f}_j(x;\theta_j, w_j) = \frac{w_j f_j(X_i;\theta_j)}{a_j(X_i;\theta^{(m)})}.$$

Hence using inequality (5.1) and the definitions, we get

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_{n,i} f(X_i) \frac{C(\delta_n(X_i)) + \delta_n(X_i)}{g_n(X_i)} \leq \sum_{j=1}^{k} Z(\theta_j, w_j), \tag{5.2}$$

where

$$Z(\theta_j, w_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{f}_j(X_i;\theta_j, w_j)}{g_n(X_i)} C\left(\frac{g_n(X_i)}{\tilde{f}_j(X_i;\theta_j, w_j)} - 1\right) + 1 - \frac{\tilde{f}_j(X_i;\theta_j, w_j)}{g_n(X_i)}.$$

Taking the right hand side of inequality (5.2) as a majorizer $h(\theta|\theta^{(m)})$, we get an iterative algorithm which is as explained above. By the descent property of MM, this algorithm also has a descent property. Also, observe that the majorizer has θ_j in only the j-th term so that minimizing the majorizer can be shifted to minimizing the j-th term with respect to θ_j once new weights are found. Note that the majorizer has to be minimized under the constraint of sum of weights equal to one. This can be achieved by the Lagrange multiplier method. The algorithm discussed here which we will refer to as DMIX becomes HMIX and HELMIX algorithm when applied to those specific cases.

Convergence results for MM algorithms were derived by Vaida (2005). Vaida (2005) proved that under certain regularity conditions on the majorizer the sequence of MM iterations converge to an element of the set of stationary points of the actual function which is being minimized. In particular, if the majorizer has a unique minimum at the stationary points of the actual function, then the MM algorithm is convergent. See Theorems 2 and 4 and comments before Section 6 of Vaida (2005). These results prove the convergence of our algorithm.

6. Real Data and Simulation Studies

In this section, we apply our estimation procedure on some real datasets. All the robust estimator presented in this section are obtained as solutions to the corresponding minimum distance weighted likelihood estimating equations.

6.1. Newcomb: Speed of Light Data

In 1882, Simon Newcomb set up an experiment which measured the amount of time required for light to travel a distance of 7442 metres. The data are recorded as deviations from 24,800 nanoseconds. There are two unusually low measurements (-44 and -2) and then a cluster of measurements that seems to be approximately symmetrically distributed. For a full description of Newcomb's data, see Stigler (1977).

The histogram of Newcomb's data and normal density fits given by the maximum likelihood estimator and the minimum symmetric chi-square estimator (see Lindsay (1994), Markatou et al. (1998)) are presented in Figure 1. For comparison, we also present a kernel density fit to the given data in 1. The estimators corresponding to the symmetric chi-square (SCS), the Hellinger distance (HD) and the negative exponential disparity (NED) obtained using our methodology are given in Table 1. Taken together, Figure 1 and Table 1 demonstrate that our proposed estimators successfully discount the effect of the large outliers, unlike the MLE, and lead to much more stable inference. Here and in all other datasets presented, we used Epanechnikov kernel with optimal bandwidth for nonparametric density estimation.

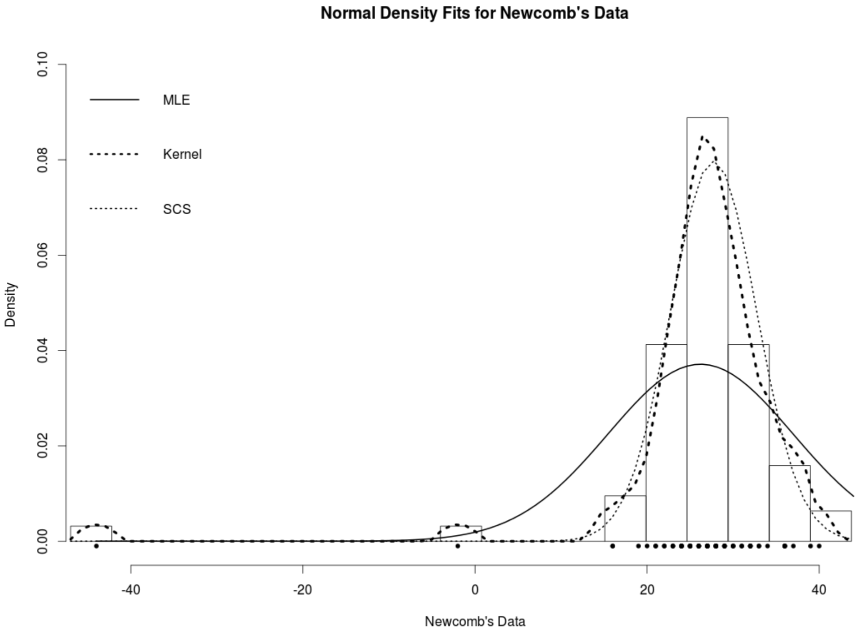

Fig 1. *Normal Density Fits for Newcomb's Data.*

6.2. Melbourne: Daily Rainfall Data

This dataset is taken from Staudte and Sheather (1990). Rainfall varies with the seasons in Melbourne, Australia. For the sake of time homogeneity, we restrict attention to the winter months of June, July, and August. During this rainy season roughly half the days have no measurable rainfall, and we will hereafter restrict attention to "rain days," those in which there is at least one millimeter of measured rainfall. The distribution of the daily rainfall for the winter months of 1981-1983 can be approximated by an exponential distribution as suggested by the histogram in Figure 2. Since there is some day-to-day dependence, a Markov model is more appropriate if one wants to use all the information. However, we will select every fourth rain day observation from the data in Table C.2 of the Appendix of Staudte and Sheather (1990) and assume independence as was also done by Staudte and Sheather (1990). The measurements in millimeter are:

1. 1981: 6.4, 4.0, 3.2, 3.2, 8.2, 11.8, 6.0, 0.2, 4.2, 2.8, 0.6, 2.0, 16.4.
2. 1982: 0.4, 8.4, 1.0, 7.4, 0.2, 4.6, 0.2.
3. 1983: 0.2, 0.2, 0.8, 0.2, 9.8, 1.2, 1.0, 0.2, 30.2, 1.4, 3.0 .

The value 30.2 is a clear outlier and stands out in the histogram. The exponential density fits given by maximum likelihood and symmetric chi-square with and without the outliers are shown in Figures 2 and 3 respectively. The estimators of the mean parameter given by the symmetric chi-square (SCS), the Hellinger distance (HD) and the negative exponential disparity (NED) using our methodology are given in Table 1. It is clear the that effect of outlier has been largely arrested by our robust estimators unlike the MLE. On the other hand, when the outlier is removed, all the estimators including the MLE are closely clustered together.

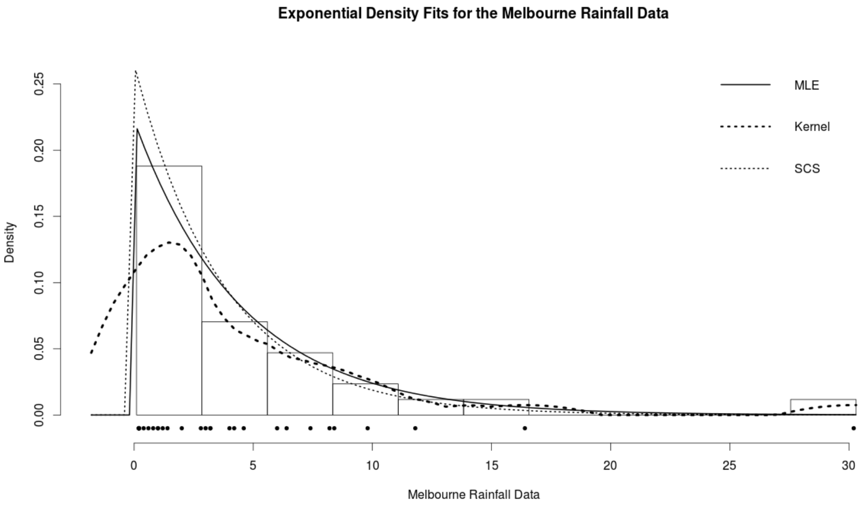

Fig 2. *Exponential Density Fits for the Melbourne Rainfall Data (with the outlier).*

The estimators obtained by using the Hellinger distance, the symmetric chi-square disparity and the negative exponential disparity for the datasets presented in Sections 6.1-6.2 are given in Table 1. The two rows for the Newcomb

Table 1
Estimates for the Newcomb and the Melbourne Datasets

Data	MLE	HD	SCS	NED
Newcomb	26.212121	27.728633	27.725862	27.745055
	10.745325	5.011576	5.000903	5.032101
Melbourne	4.496774	4.063245	3.777422	3.475987
Melbourne($-O$)	3.640000	3.730983	3.631070	3.448983

data represent the estimates of mean (μ) and the standard deviation (σ) in the normal density. Melbourne($-O$) represents the Melbourne data obtained after deleting the outlier 30.2.

6.3. Simulation Studies

The following tables give the MSE of the estimated parameters under the normal model, based on 125 samples, each containing 100 observations from

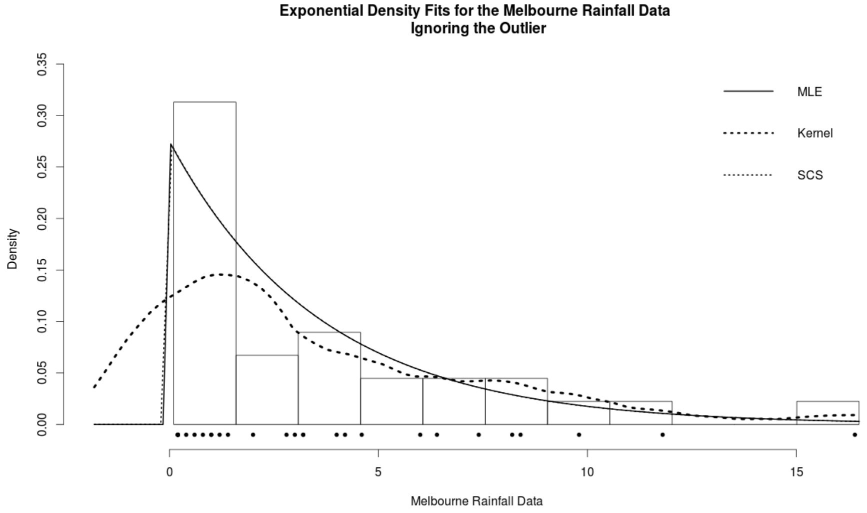

Fig 3. *Exponential Density Fits for the Melbourne Rainfall Data Without the Outlier.*

$(1-\varepsilon)N(0,1)+\varepsilon N(10,1)$ for $\varepsilon = 0$, 0.0, 0.10, 0.15, 0.20, 0.25. Our targets are the parameters of the larger, $N(0,1)$ component. We used Epanechnikov kernel with optimal bandwidth for nonparametric density estimation. The observed mean square error for the mean parameter is computed against the target 0, while the observed MSE of the parameter of standard deviation is computed against the target value of 1. The tabled values show that all the MDWL estimators are highly successful in ignoring the smaller, contaminating component, unlike the MLE.

Table 2
MSEs of the MLE and the MDWL estimates of the Mean parameter

Error(ε)	HD	SCS	NED	MLE
0%	0.009103	0.009236	0.009028	0.009034
5%	0.012910	0.013133	0.013112	0.311434
10%	0.010239	0.010376	0.010091	1.056103
15%	0.011506	0.011516	0.011650	2.369552
20%	0.011384	0.011493	0.011184	4.074779
25%	0.014786	0.015430	0.015732	6.368500

7. Hypothesis Testing

A popular and useful statistical tool for the hypothesis testing problem is the likelihood ratio test. The likelihood ratio test statistic is constructed as twice the difference of the unconstrained maximum log likelihood and the maximum log likelihood under the null hypothesis. In the language of disparities, the test

Table 3

MSEs of the MLE and the MDWL estimates of the scale parameter

Error(ε)	HD	SCS	NED	MLE
0%	0.005566	0.006390	0.006209	0.004927
5%	0.005589	0.006191	0.006052	2.130894
10%	0.005653	0.006171	0.006163	4.745307
15%	0.006299	0.006524	0.006582	7.348949
20%	0.006411	0.006915	0.007069	9.768701
25%	0.007965	0.008243	0.008178	11.910837

statistic is constructed by taking the difference between the minimum of the likelihood disparity under the null and that without any constraint. Under certain regularity conditions, the likelihood ratio test enjoys some asymptotic optimality properties.

However, as in the case of the maximum likelihood estimator, the likelihood ratio test exhibits poor robustness properties in many cases. As an alternative to the likelihood ratio test, Simpson (1989) introduced the Hellinger deviance test which was later generalized to disparity difference tests, in a unified way; see eg. Lindsay (1994) and Basu et al. (2011).

The set up under which we deal with the problem of hypothesis testing is as follows. We assume the parametric set up of Section 2 and let independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ be available from the true distribution G. The hypothesis testing problem under consideration is

$$H_0 \ : \ \theta \in \Theta_0 \quad \text{and} \quad H_1 \ : \ \theta \in \Theta \setminus \Theta_0,$$

for a proper subset Θ_0 of Θ. We define the empirical divergence to be

$$\rho_C(g_n, f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{C(\delta_n(X_i)) + \delta_n(X_i)}{\delta_n(X_i) + 1} \mathbb{1}_{\{X_i \in A_n\}}.$$

As an analog of the likelihood ratio test, define the test statistic,

$$W_C(g_n) := 2n\left[\rho_C(g_n, f_{\hat{\theta}_0}) - \rho_C(g_n, f_{\hat{\theta}})\right], \tag{7.1}$$

where $\hat{\theta}$ and $\hat{\theta}_0$ denote the unrestricted minimizer of $\rho_C(g_n, f_\theta)$ and the minimizer under the constraint of $\theta \in \Theta_0$ and g_n is the kernel density estimate.

We will now present the main theorem of this section which establishes the asymptotic distribution of W_C.

Theorem 7.1. *Under the model f_{θ_0}, $\theta_0 \in \Theta_0$ and assumptions **(A1)** - **(A9)**, the limiting null distribution of the test statistic $W_C(g_n)$ is χ^2_r, where r is the number of restrictions imposed by the null hypothesis H_0.*

proof of Theorem 7.1. A Taylor series expansion of $\rho_C(g_n, f_{\hat{\theta}_0})$ with respect to

θ around $\hat{\theta}$, gives

$$\begin{aligned}W_C(g_n) &= 2n\left[\rho_C(g_n, f_{\hat{\theta}_0}) - \rho_C(g_n, f_{\hat{\theta}})\right]\\ &= 2n\left[(\hat{\theta}_0 - \hat{\theta})^\top \nabla\rho_C(g_n, f_{\hat{\theta}})\right]\\ &\quad + 2n\left[\frac{1}{2}(\hat{\theta}_0 - \hat{\theta})^\top \nabla_2\rho_C(g_n, f_{\theta^*})(\hat{\theta}_0 - \hat{\theta})\right],\end{aligned}$$

where θ^* belongs to the line joining $\hat{\theta}_0$ and $\hat{\theta}$. Note that the first term in the last expression is zero as $\hat{\theta}$ is the minimizer of ρ_C over Θ. So, we only need to deal with the second term in the expansion. Now

$$\begin{aligned}W_C(g_n) &= n\left[(\hat{\theta}_0 - \hat{\theta})^\top I(\theta_0)(\hat{\theta}_0 - \hat{\theta})\right]\\ &\quad + n\left[(\hat{\theta}_0 - \hat{\theta})^\top \{\nabla_2\rho_C(g_n, f_{\theta^*}) - I(\theta_0)\}(\hat{\theta}_0 - \hat{\theta})\right]. \end{aligned} \tag{7.2}$$

Under the model f_{θ_0}, $n^{1/2}(\hat{\theta}_0 - \theta_0)$ and $n^{1/2}(\hat{\theta} - \theta_0)$ are both $O_p(1)$. Thus, $n^{1/2}(\hat{\theta}_0 - \hat{\theta}) = O_p(1)$. By Theorem 3.3, $\nabla_2\rho_C(g_n, f_\theta) = \nabla\Psi_n(\theta)$ converges to $-B(\theta)$ uniformly in $\theta \in \Theta_0$. Note that $B(\theta_0) = -I(\theta_0)$ under $g = f_{\theta_0}$. Since $\hat{\theta}_0 - \theta_0 = o_p(1)$ and $\hat{\theta} - \theta_0 = o_p(1)$, $\theta^* \in \Theta_0$ for large enough n and so

$$\begin{aligned}|\nabla_2\rho_C(g_n, f_{\theta^*}) - I(\theta_0)| &\le |\nabla_2\rho_C(g_n, f_{\theta^*}) + B(\theta^*)| + |-B(\theta^*) - I(\theta_0)|\\ &\le \sup_{\theta\in\Theta_0}|\nabla_2\rho_C(g_n, f_\theta) + B(\theta)| + |B(\theta^*) + I(\theta_0)| \xrightarrow{P} 0.\end{aligned}$$

Hence, by the arguments above, the second term on the right hand side of Equation (7.2) converges in probability to zero. By Equations (3.3) and (3.4), we have

$$n^{1/2}(\hat{\theta} - \hat{\theta}_0) = n^{1/2}(\hat{\theta}_{ML} - \hat{\theta}_{0,ML}) + o_p(1),$$

where $\hat{\theta}_{ML}$ and $\hat{\theta}_{0,ML}$ are the unrestricted and constrained maximum likelihood estimators. Hence, $W_C(g_n)$ is equivalent to the likelihood ratio test statistic under the model f_{θ_0} in the sense that

$$W_C(g_n) - n\left[(\hat{\theta}_{0,ML} - \hat{\theta}_{ML})^\top I(\theta_0)(\hat{\theta}_{0,ML} - \hat{\theta}_{ML})\right] = o_p(1). \tag{7.3}$$

From the theory of likelihood ratio test, we conclude that W_C converges in distribution to a χ^2_r as $n \to \infty$ as stated. See Serfling (1980, Section 4.4.4) for a complete discussion on likelihood ratio test. □

Theorem 7.2. *The conditions of Theorem 7.1 and the additional assumption that the parametric family $\mathcal{F}_\theta$ satisfies the local asymptotic normality (LAN) condition indicate that under f_{θ_n} and as $n \to \infty$*

$$W_C(g_n) - 2\sum_{i=1}^n\left[\log f_{\hat{\theta}_{ML}}(X_i) - \log f_{\hat{\theta}_{0,ML}}(X_i)\right] \xrightarrow{P} 0,$$

where $\theta_n = \theta_0 + \tau n^{-1/2}$.

Proof of Theorem 7.2. Under the assumptions of Theorem 7.1, Equation (7.3) implies the stated claim under f_{θ_0}, since the Wald test statistic is equivalent to the likelihood ratio test statistic under the null. See Serfling (1980, pg. 158 – 160) for more details. By LAN condition, we have that f_{θ_n} is contiguous to f_{θ_0} and so convergence in probability under f_{θ_0} implies convergence in probability under f_{θ_n}. Hence the proof is complete. □

The following theorem explores the stability of the limiting distribution of the test statistic $W_C(g_n)$ under contamination. For this theorem the null hypothesis under consideration is $H_0 : \theta_g = \theta_0$, where the unknown true distribution G may or may not be in the model.

Theorem 7.3. *Under assumptions **(A1)** - **(A9)**, under the null hypothesis, we have the following*

$$W_C(g_n) - Y_{2n} = Y_1 + o_p(1),$$

where $Y_1 \sim \chi^2_p$ and $\lim_{g\to f_{\theta_0}} Y_{2n} = 0$ for any C. Here by $g \to f_{\theta_0}$, we mean the convergence in L_1 sense. The rate at which the convergence to 0 of Y_{2n} holds depend on the form of C. See Remark 7.1 for more details.

Proof of Theorem 7.3. The proof of this theorem closely follows the proof of Theorem 7.1. As in Theorem 7.1, we get by Taylor series expansion of the test statistic around $\hat{\theta}_n$,

$$\begin{aligned} W_C(g_n) &= 2n\left[\rho_C(g_n, f_{\theta_0}) - \rho_C(g_n, f_{\hat{\theta}_n})\right] \\ &= n(\theta_0 - \hat{\theta}_n)^\top \nabla_2 \rho_C(g_n, f_{\theta^*})(\theta_0 - \hat{\theta}_n), \end{aligned}$$

where θ^* belongs to the line joining θ and θ_0. By Theorem 3.3, $\nabla_2\rho_C(g_n, f_{\theta^*})$ converges in probability to $B(\theta_0)$ under the null hypothesis. Hence, we have

$$W_C(g_n) = -n(\theta_0 - \hat{\theta}_n)^\top B(\theta_0)(\theta_0 - \hat{\theta}_n) + o_p(1).$$

Note that

$$-B(\theta_0) = B(\theta_0)V^{-1}(\theta_0)B(\theta_0) - B(\theta_0)\left[V^{-1}(\theta_0) + B^{-1}(\theta_0)\right]B(\theta_0).$$

By, Theorem 3.4, we get

$$n(\hat{\theta}_n - \theta_0)B(\theta_0)V^{-1}(\theta_0)B(\theta_0)(\hat{\theta}_n - \theta_0) = Y_1 + o_p(1),$$

where $Y_1 \sim \chi^2_p$. The remaining term given by

$$Y_{2n} = -n(\hat{\theta}_n - \theta_0)B(\theta_0)\left[V^{-1}(\theta_0) + B^{-1}(\theta_0)\right]B(\theta_0)(\hat{\theta}_n - \theta_0),$$

becomes zero if $g = f_{\theta_0}$ and stays close to zero as $g \approx f_{\theta_0}$. □

Remark 7.1 This result extends Theorem 6 of Lindsay (1994), which was in the case of a scalar parameter. In our case, if $p = 1$, both $B = -B(\theta_0)$ and $V = V(\theta_0)$ are scalars, so that

$$W_C(g_n) = \frac{V}{B} X_n + o_p(1),$$

where $X_n \overset{\mathcal{L}}{\to} \chi^2_1$ under H_0. Thus V/B, as a function of the true density g, and the disparity generating function $C(\cdot)$ represents the inflation in the χ^2 distribution, and can be legitimately called the χ^2 inflation factor. This is exactly the same as the inflation factor described in Theorem 6, part (ii) of Lindsay (1994). When $g = f_{\theta_0}$ is the true distribution, $V = B$ so that there is no inflation. However, when the true distribution is a point mass mixture contamination Lindsay (1994) demonstrated, using the binomial model for illustrations, that the inflation factor for the likelihood ratio test rises sharply with the contamination proportion, whereas for the Hellinger deviance test this rise is significantly dampened in comparison. Our inflation factor calculations in the normal mean model exhibit improvements of similar order between the likelihood ratio test and other robust tests, although we do not present the actual numbers here.

In the multidimensional case, however, the relation is not so simple as now it requires comparison between the matrices $B(\theta_0)V^{-1}(\theta_0)B(\theta_0)$ and $-B(\theta_0)$, rather than the between scalars. While we have presented the essential result, it could be interest to develop a single quantitative measure of inflation for the multidimensional case in the future.

8. Conclusions and Future Work

This paper demonstrates that the minimum disparity estimation procedure can be simultaneously viewed as a weighted likelihood estimation procedure and also gives a proof of asymptotic normality of the MDWL estimator under fairly general conditions on the family of disparities. For example, all the disparities presented in Table 2.1 of Basu et al. (2011) satisfy our assumptions but not all of them satisfy the assumptions of Markatou et al. (1998). We also generalize the proof of asymptotic normality due to Kuchibhotla and Basu (2015) by appropriately modifying their assumption (A8) which may not be satisfied over the whole real line. In the proof presented here, we trim the kernel density estimator so that such an assumption is valid on this trimmed set. Hence the proof of Kuchibhotla and Basu (2015) may be carried out with assumption (A8) and with an inclusion of trimming parameter as is done here.

As the proof presented here involves a trimming parameter, the application of this method involves choosing such a parameter. This choice will certainly require further research. We anticipate that the proof can be done without trimming since the numerator $C(\delta_n(x))f_\theta(x)$ and the denominator $g(x)$ converges to zero as $|x| \to \infty$ at an approximately same rate if $\theta = \theta_g$.

Also, the proof explicitly uses the form of the nonparametric density estimator used, namely, the kernel density estimator. But using the techniques from semiparametric M-estimation or those of empirical processes, we feel that a proof can be done without explicitly using the form of density estimator. Density estimators based on spacings are easier to calculate numerically than the kernel density estimator. We think that this method might give a competitive alternative to the one with kernel density estimator.

Finally, we note that in the context of minimum disparity estimation, we now that have two estimated estimating functions (sample versions), the usual one involving integrals and the other as introduced here, with both having the same asymptotic robustness properties because of the same population objective function. It would therefore be appropriate to have a detailed simulation study comparing the small sample properties of the two corresponding estimators. Small sample theoretical properties like finite sample breakdown point or expected finite sample breakdown point of these two estimators would give a better comparison of their capabilities.

Acknowledgements

The authors thank Dr. Arijit Chakrabarti and Mr. Promit Kumar Ghosal, both of the Indian Statistical Institute, for helpful discussions.

Supplementary Material

Supplement to “A Minimum Distance Weighted Likelihood Method of Estimation”

(). Propositions and Theorems which have not been proved in this manuscript are provided in the supplementary material. Several additional real data examples have also been included in the supplementary material.

References

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131142.

Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72.

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

Basu, A. and Lindsay, B. G. (2004). The iteratively reweighted estimating equation in minimum distance problems. *Comput. Statist. Data Anal.*, 45(2):105–124.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference:The minimum distance approach*, volume 120 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, 5(3):445–463.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, 8(3):457–487.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl*, 8:85108.

Cutler, A. and Cordero-Braña, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.*, 91(436):1716–1723.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Fryer, J. G. and Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika*, 59(3):pp. 639–648.

Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *J. Statist. Plann. Inference*, 136(11):3989–4011.

Giné, E. and Mason, D. M. (2008). Uniform in bandwidth estimation of integral functionals of the density function. *Scand. J. Statist.*, 35(4):739–761.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37.

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41(4):683–697.

Karlis, D. and Xekalaki, E. (1998). Minimum hellinger distance estimation for poisson mixtures. *Computational Statistics and Data Analysis*, 29(1):81 – 103.

Kuchibhotla, A. K. and Basu, A. (2015). A general set up for minimum disparity estimation. *Statistics & Probability Letters*, 96:68 – 74.

Lange, K. (2010). *Numerical analysis for statisticians*. Statistics and Computing. Springer, New York, second edition.

Lewbel, A. (1997). Semiparametric estimation of location and other discrete choice moments. *Econometric Theory*, 13(1):32–51.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.*, 22(2):1081–1114.

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *J. Amer. Statist. Assoc.*, 93(442):740–750.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam.

Park, C. and Basu, A. (2004). Minimum disparity estimation: asymptotic normality and breakdown point results. *Bull. Inform. Cybernet.*, 36:19–33.

Parr, W. C. (1981). Minimum distance estimation:a bibliography. *Communications in Statistics - Theory and Methods*, 10(12):1205–1224.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, L:157175.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 531–545. Univ. California Press, Berkeley, Calif.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B*, 24:46–72.

Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā Ser. A*, 25:189–206.

Robertson, C. A. (1972). On minimum discrepancy estimators. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 34(2):pp. 133–144.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.*, 82(399):802–807.

Simpson, D. G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.*, 84(405):107–113.

Staudte, R. G. and Sheather, S. J. (1990). *Robust estimation and testing*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Stigler, S. M. (1977). Do robust estimators work with real data? *Ann. Statist.*, 5(6):1055–1098.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statist. Sinica*, 15(3):831–840.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.*, 65(2):245–260.