Sankhyā : The Indian Journal of Statistics
2008, Volume 70-B, Part 2, pp. 0-12

Minimum Hellinger Distance Estimation with Inlier Modification

Rohit Kumar Patra
Indian Statistical Institute, Kolkata, India
Abhijit Mandal
Indian Statistical Institute, Kolkata, India
Ayanendranath Basu
Indian Statistical Institute, Kolkata, India

Abstract

Inference procedures based on the Hellinger distance provide attractive alternatives to likelihood based methods for the statistician. The minimum Hellinger distance estimator has full asymptotic efficiency under the model together with strong robustness properties under model misspecification. However, the Hellinger distance puts too large a weight on the inliers which appears to be the main reason for the poor efficiency of the method in small samples. Here some modifications to the inlier part of the Hellinger distance are provided which lead to substantial improvements in the small sample properties of the estimators. The modified divergences are members of the general class of disparities and satisfy the necessary regularity conditions so that the asymptotic properties of the resulting estimators follow from standard theory. In limited simulations the proposed estimators exhibit better small sample performance at the model and competitive robustness properties in relation to the ordinary minimum Hellinger distance estimator. As the asymptotic efficiencies of the modified estimators are the same as that of the ordinary estimator, the new procedures are expected to be useful tools for applied statisticians and data analysts.

AMS (2000) *subject classification.* Primary To be filled.
Keywords and phrases. Hellinger distance, inliers, inlier modified Hellinger distance, asymptotic distribution.

1 Introduction

In recent times, density based divergences have been studied in the context of discrete models by Cressie and Read (1984) and Lindsay (1994).

Pardo (2006) provides a good general reference for results relating to density based divergences in discrete models. The maximum likelihood estimator is known to be fully asymptotically efficient under standard regularity conditions, but has very poor robustness properties. On the other hand classical robust estimators – based on M-estimation and its extensions – usually sacrifice first order efficiency at the model to achieve their robustness (see, eg. Hampel et al. 1986). Estimators which combine full asymptotic efficiency at the model with strong stability properties can have great practical value.

Some density-based minimum distance estimators have been shown to attain first-order efficiency at the model together with strong robustness properties. Within the class of minimum divergence procedures, the methods based on the minimum Hellinger distance stand out in terms of their popularity and often represent the standard against which other minimum divergence procedures are judged. Beran (1977) and Simpson (1987, 1989) have provided much of the basic background and properties of minimum Hellinger distance inference. Lindsay (1994) has considered a larger class of divergences, called disparities, which includes the Hellinger distance; Lindsay has also provided general conditions under which the minimum disparity estimators have full asymptotic efficiency at the model.

The popularity of the minimum Hellinger distance procedures are partially tempered by the relatively poor efficiency of these methods compared to the likelihood based methods in small samples. It appears that the diminished small sample efficiency is due to the large weight attached to the "inliers" (discussed in the next section) by the Hellinger distance. Our aim in this paper is to modify the Hellinger distance in such a manner that the small sample efficiency of the resulting estimator is improved while the robustness properties remain intact. Among others, the minimum penalized Hellinger distance estimator (Harris and Basu, 1994; Basu and Basu, 1998) and the minimum combined distance estimators (Park et al., 1995) have also been shown to provide reasonable solutions to the inlier problem. However, the modifications made in these cases result in divergences which do not satisfy the defining conditions of a disparity, and the general approach of Lindsay (1994) is no longer directly applicable in determining the asymptotic properties of these estimators. The asymptotic distributions of the minimum penalized Hellinger distance estimator and minimum combined disparity estimator are yet to be theoretically established in the literature.

In this paper, we will discuss the suggested modification of the Hellinger distance in connection with discrete models only. The method is applicable

to the continuous case in principle, but requires additional accessories such as kernel density estimation or other nonparametric smoothing techniques. We will consider such models in a sequel paper.

2 The Class of Disparities and the Hellinger Distance

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables from the distribution with probability mass function (p.m.f.) $f(x)$ with $x \in \mathcal{X}$. Consider a discrete parametric model with p.m.f. $m_\theta(x)$ where
$\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T \in \Theta \subset \mathbb{R}^p$, and which will be expected to model the data. Here the true data generating p.m.f. f may or may not be a member of the family $\{m_\theta : \theta \in \Theta\}$.

For any $x \in \mathcal{X}$, suppose $d_n(x)$ be the proportion of sample observations at x. The Pearson residual function $\delta_n(x)$ at x is defined as

$$\delta_n(x) = \frac{[d_n(x) - m_\theta(x)]}{m_\theta(x)} \ . \tag{2.1}$$

We will drop the subscript n from $\delta_n(x)$ and $d_n(x)$ whenever there is no scope of confusion. Suppose that $G(\cdot)$ is a real-valued, thrice-differentiable, strictly convex function on $[-1, \infty)$, with $G(0) = 0$. We will consider density based divergences, called disparities, generated by the function G which are denoted by $\rho_G(d, m_\theta)$ and defined as

$$\rho_G(d, m_\theta) = \sum_{x \in \mathcal{X}} G(\delta(x)) m_\theta(x) \ . \tag{2.2}$$

There are several important subfamilies of the class of disparities which include the Cressie-Read (1984) family of power divergences, indexed by a parameter $\lambda \in \mathbb{R}$, having the form

$$\begin{aligned} I^\lambda(d, m_\theta) &= \frac{1}{\lambda(\lambda+1)} \sum_x d(x) \left\{ \left(\frac{d(x)}{m_\theta(x)} \right)^\lambda - 1 \right\} \\ &= \sum_x \left[\frac{(\delta(x)+1)^{\lambda+1} - (\delta(x)+1)}{\lambda(\lambda+1)} - \frac{\delta(x)}{\lambda+1} \right] m_\theta(x). \end{aligned} \tag{2.3}$$

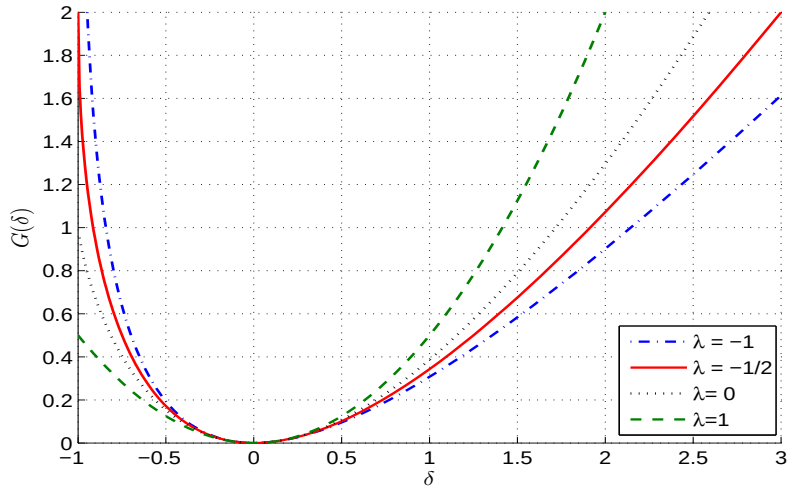
The disparities for the cases $\lambda = 0$ and $\lambda = -1$ are defined by the continuous limits of the above expressions as $\lambda \to 0$ and $\lambda \to -1$ respectively. The

likelihood disparity is generated by $\lambda = 0$ which is minimized by maximum likelihood estimator. The likelihood disparity has the form

$$\mathrm{LD}(d, m_\theta) = \sum_x [d(x) \log(d(x)/m_\theta(x)) + (m_\theta(x) - d(x))].$$

The graphs of the $G(\cdot)$ functions for several members of the power divergence family are presented in Figure 1. Notice that the flatter the graphs are to the right of the point $\delta = 0$, the more sharply they rise on the left.

Figure 1: Plots of the $G(\delta)$ function for several members of the power divergence family corresponding to $\lambda = -1, -\frac{1}{2}, 0, 1$.



The minimum disparity estimator $\hat{\theta}_n$ corresponding ρ_G is defined to be the value of θ which minimizes ρ_G over $\theta \in \Theta$. Thus

$$\rho_G(d, m_{\hat{\theta}_n}) = \min_{\theta \in \Theta} \rho_G(d, m_\theta).$$

Under differentiability of the model this is equivalent to solving the estimating equation

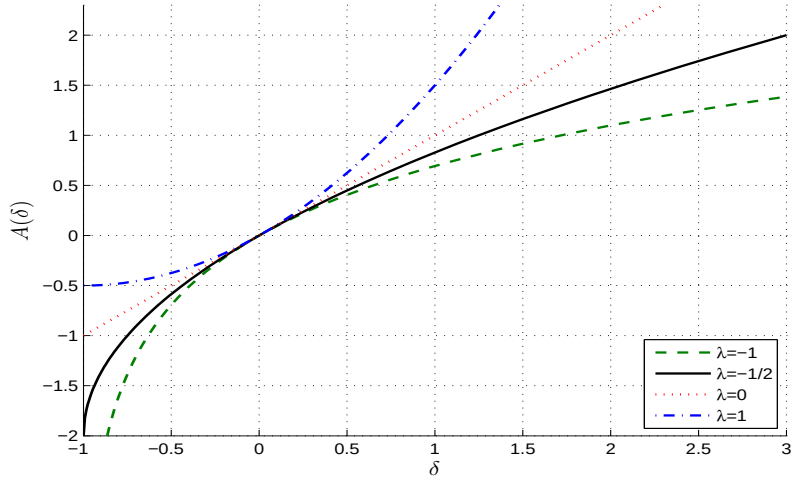
$$-\nabla \rho_G(d, m_\theta) = \sum_x \left[G'(\delta(x)) \frac{d(x)}{m_\theta(x)} - G(\delta(x)) \right] \nabla m_\theta(x) = 0, \qquad (2.4)$$

where ∇ denotes the gradient with respect to θ and G' is the first derivative of G. The estimating equation can be written as

$$\sum_x A(\delta(x)) \nabla m_\theta(x) = 0 \ , \quad \text{where } A(\delta) = (\delta + 1) G'(\delta) - G(\delta). \qquad (2.5)$$

The function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity. It is strictly increasing on $[-1, \infty)$, and can be redefined to satisfy $A(0) = 0$ and $A'(0) = 1$ without changing the estimating properties of the disparity. This is achieved if the corresponding G functions is standardized so that $G'(0) = 0$ and $G''(0) = 1$. The residual adjustment functions of several members of the power divergence family are presented in Figure 2.

Figure 2: Plots of the RAF $A(\delta)$ for several members of the power divergence family corresponding to $\lambda = -1, \ -\frac{1}{2}, \ 0, \ 1$.



To analyze the robustness properties of the minimum disparity estimators, one has to characterize the outliers probabilistically. If an observation x in the sample space is associated with a large positive value of $\delta(x)$, it will be called an outlier in the sense that the actual observed proportion at x is much larger than what is predicted by the model. For robust estimation, one should choose such disparities which give very small weights to the observations having large positive values of δ; for such cases, the RAF $A(\delta)$ would exhibit a severely dampened response to increasing δ. For a qualitative description, one can take the RAF of the likelihood disparity $A_{\mathrm{LD}}(\delta)$ as the basis for comparison. For this disparity $G_{\mathrm{LD}}(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ and $A_{\mathrm{LD}}(\delta) = \delta$, and thus to compare the other minimum disparity estimators with the maximum likelihood estimator, one must focus on how their RAFs depart from linearity for large positive δ. The RAFs for the disparities with large negative λ are expected to perform better in terms of robustness; however, in these cases the functions curve sharply down on the right hand side

of the δ axis. We will refer to the curves (of the G function as well as the A function) on the right hand side of the point $\delta = 0$ as the outlier part of the function, while the corresponding left hand side will be referred to as the "inlier" part.

Inliers, corresponding to negative values of δ, are those points which have less data than expected under the model. Unfortunately, the minimum disparity estimators which have better stability properties under the presence of outliers are often highly sensitive to inliers. Figure 2 demonstrates that the residual adjustment functions corresponding to large negative values of λ are quite flat on the outlier part, but rise steeply in the inlier part. Empirical studies have shown that this often leads to a substantial deficiency in the small sample performance of the estimators at the model; this is unfortunate, since these estimators are otherwise desirable in terms of their robustness properties. See Lindsay (1994) and Basu and Basu (1998) for a more detailed discussion of the inlier problem.

In this paper, we will attempt to provide one solution to the inlier problem without jeopardizing the asymptotic efficiency or the robustness properties of the corresponding estimators. To keep the presentation short and focused, we will illustrate our proposal thorough the Hellinger distance (a member of the power divergence family corresponding to $\lambda = -1/2$). In general such modifications will be useful for all disparities which are sensitive to "inliers".

The (twice, squared) Hellinger distance between d and m_θ is defined as

$$\mathrm{HD}(d, m_\theta) = 2\sum_x (d^{1/2}(x) - m_\theta^{1/2}(x))^2, \tag{2.6}$$

and the corresponding G and A functions are given as

$$G_{\mathrm{HD}}(\delta) = 2(\sqrt{\delta+1}-1)^2, \quad A_{\mathrm{HD}}(\delta) = 2(\sqrt{\delta+1}-1).$$

The plots of the $G_{\mathrm{HD}}(\cdot)$ and $A_{\mathrm{HD}}(\cdot)$ functions are provided in Figures 1 and 2 corresponding to $\lambda = -1/2$.

3 The Inlier Modified Hellinger Distance

We have observed that the $G_{\mathrm{HD}}(\cdot)$ and the $A_{\mathrm{HD}}(\cdot)$ functions are quite sensitive to inliers. In order to reduce this sensitivity, we propose to shrink $G_{\mathrm{HD}}(\delta)$ towards zero for negative values of δ; however, while doing this modification we keep the outlier part of the function intact, and also ensure that

the resulting $A(\delta)$ function is continuously differentiable up to the second order at $\delta = 0$.

Letting IMHD stand for "Inlier Modified Hellinger Distance", we define the function

$$G^{\gamma}_{\text{IMHD}}(\delta) = \begin{cases} \dfrac{G_{\text{HD}}(\delta)}{(1+\delta^2)^{\gamma}}, & \delta \leq 0, \gamma \geq 0, \\ G_{\text{HD}}(\delta), & \delta > 0, \end{cases} \tag{3.1}$$

which will be used to generate the family of inlier modified Hellinger distances; IMHD$^{\gamma}$ will represent the disparity within this family indexed by the tuning parameter γ. Note that $\gamma = 0$ leads to no shrinkage and recovers the original Hellinger distance, while the amount of shrinkage increases with increasing γ. It can be easily verified that $G^{\gamma'''}_{\text{IMHD}}(\delta)$, the third derivative of the function $G^{\gamma}_{\text{IMHD}}(\delta)$, is continuous at $\delta = 0$ (and hence the corresponding second derivative $A^{\gamma''}_{\text{IMHD}}(\delta)$ of the residual adjustment function is continuous at $\delta = 0$). This is true more generally if the denominator of the inlier part of the function defined in (3.1) is replaced by $(1+\delta^2 P(\delta))^{\gamma}$ where $P(\delta)$ is a polynomial in δ. This would not be the case, for example, if the denominator of the inlier part is replaced by $(1+\delta P(\delta))^{\gamma}$ instead. Since the second derivative of the residual adjustment function – also called the estimation curvature of the disparity – is a core component in the theoretical development of the asymptotic properties of the minimum disparity estimators, it is important to control it appropriately. The estimation curvature is an useful indicator of the second order efficiency and the robustness of minimum disparity estimators, and Lindsay's (1994) proof of the asymptotic distribution of the minimum disparity estimator depends critically on, among other things, the boundedness of $A''(\delta)(1+\delta)$. When this condition is violated, as in the case of minimum combined disparity estimators (Park et al. 1995), the asymptotic distribution no longer follows automatically from the results of Lindsay.

Among specific members of the IMHD family, we will focus, in particular, on the IMHD^{1} and IMHD$^{0.5}$ measures. The IMHD^{1} measure shrinks the original $G(\cdot)$ function in such a way that it becomes identical to that of G_{LD} at $\delta = -1$. The G function for the measure IMHD$^{0.5}$ has a somewhat better correspondence with G_{LD} over a larger range of the inlier part, although it exhibits a relatively sharper dip near the left boundary. The IMHD$^{0.5}$ measure is also intuitive in that its denominator $(1+\delta^2)^{1/2}$ is of the order $O(\delta)$ which is of the same order as the numerator, and hence has the same scale factor.

Figure 3: Comparison plots of $G(\delta)$ vs δ for for MIMHDE^1,MIMHDE$^{0.5}$ and MHDE

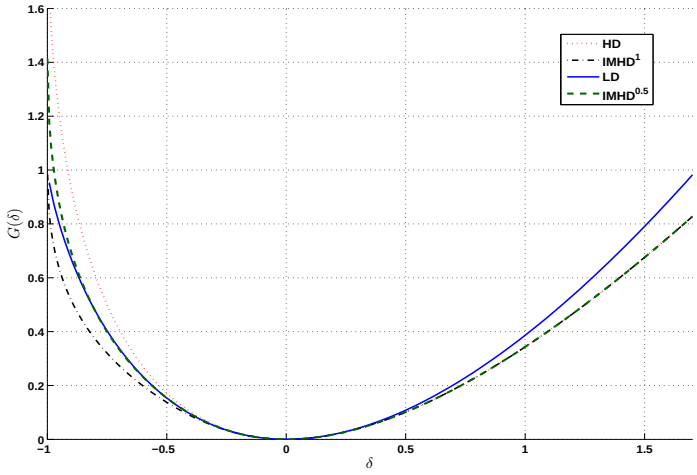
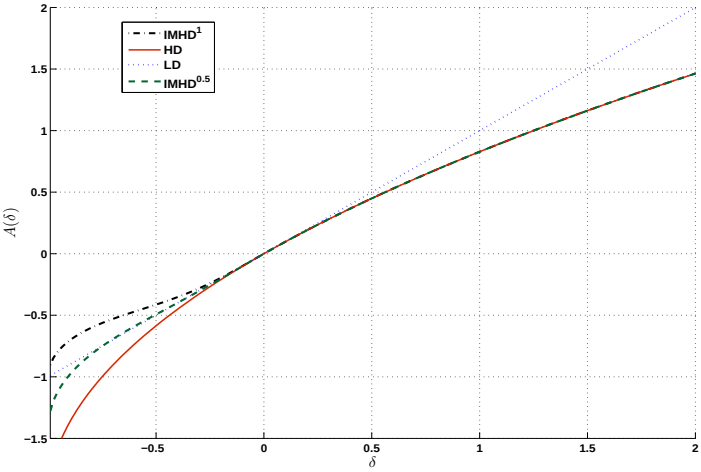


Figure 4: Comparison plots of $A(\delta)$ vs δ for for MIMHDE^1,MIMHDE$^{0.5}$ and MHDE



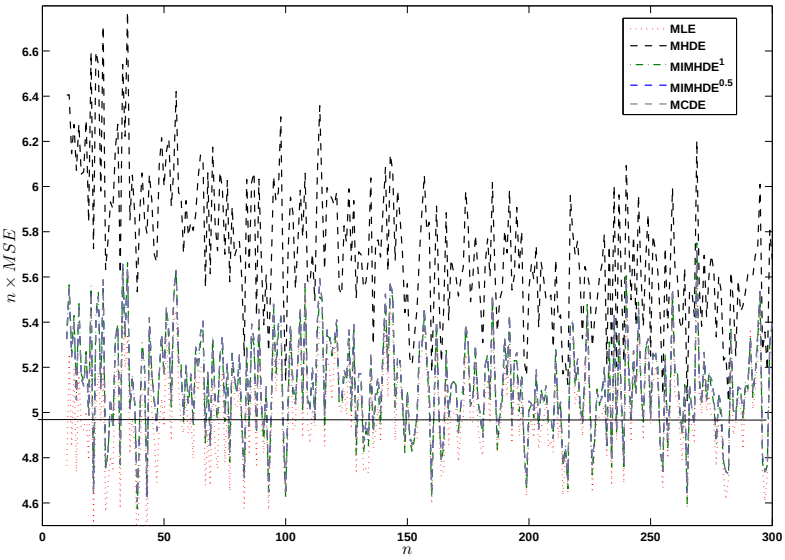
It is not easy to analytically show that the function $G^{\gamma}_{\text{IMHD}}(\delta)$ is convex on $[-1, \infty)$; however, direct evaluation shows that this is indeed the case, at least for $\gamma \in [0, 1]$. In Figures 3 and 4 we have presented the plots of $G^{0.5}_{\text{IMHD}}$, G^{1}_{IMHD} and $A^{0.5}_{\text{IMHD}}$, A^{1}_{IMHD} functions, together with the corresponding curves

for the likelihood disparity and the ordinary Hellinger distance. The reduced sensitivity of the IMHD$^\gamma$ curves in the inlier side for $\gamma > 0$ is clearly apparent.

We denote the minimizer of IMHD$^\gamma$ as the minimum inlier modified Hellinger distance estimator at γ (MIMHDE$^\gamma$). The $G^\gamma_{\text{IMHD}}(\delta)$ function satisfies all the conditions of Lindsay (1994) for $\gamma \in [0, 1]$ (and possibly beyond); from Theorem 33 of the above paper it follows that for each $\gamma \in [0, 1]$, the MIMHDE$^\gamma$ is asymptotically efficient at the model with the same asymptotic distribution as the MLE or the MHDE.

Since $G^\gamma_{\text{IMHD}}(-1)$ and $G^{\gamma'}_{\text{IMHD}}(\infty)$ are finite for all $\gamma > 0$, the MIMHDE$^\gamma$ inherits the asymptotic breakdown properties described in Theorem 4.1 of Park and Basu (2004). In particular, it enjoys 50% breakdown at the model under the conditions of the above theorem.

Figure 5: Plot of $n\times$MSE in 1000 replications for data generated from the Poisson(5) distribution.



4 Simulations

In this section we present the results of a small simulation experiment to illustrate the performance of MIMHDE^1 and MIMHDE$^{0.5}$ under the Poisson

model. For the purpose of comparison we also present the corresponding results for three other estimators; (i) the maximum likelihood estimator (MLE); (ii) the ordinary minimum Hellinger distance estimator (MHDE), and (iii) the minimum combined disparity estimator (MCDE) which uses the LD for $\delta \leq 0$ and HD for $\delta > 0$. The last estimator would be the ideal one for our purpose, if not for the fact that its asymptotic distribution does not follow from the established results and the second derivative of its residual adjustment function at $\delta = 0$ is undefined.

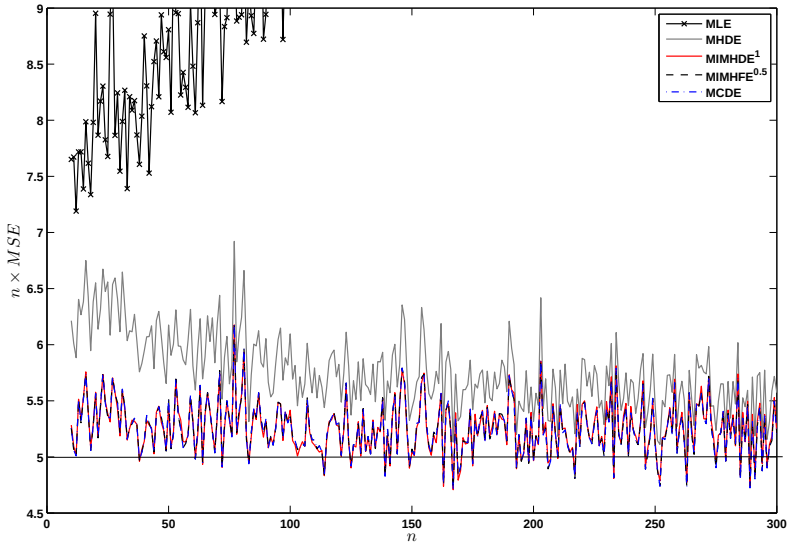
Samples of size n were generated randomly from a Poisson distribution with mean parameter $\theta = 5$ for several different values of n. All the five different estimators of θ were computed for each sample, and the mean square error of the estimates were evaluated at each sample size around the true value of 5 using 1000 replications. The plot of the mean square errors (times n) are presented in Figure 5 for values of n between 10 and 300. It is clearly observed that the MCDE and the inlier modified estimators have a much improved performance, and are quite competitive with the MLE even in very small samples. However the mean square error plot of the MHDE appears to be well above the other four plots. Clearly, the MHDE has a much inferior performance in small samples, and is substantially poorer than the other estimators even at a sample size of 300. The performance of the three estimators, MCDE, MIMHDE$^{0.5}$ and MIMHDE^{1} are so close that for any practical purpose there is nothing to choose between them in this case.

To compare the robustness of the estimators, we next generated data from the 0.99Poisson(5)+0.01Poisson(20) mixture and calculated our five estimators as in the previous case assuming a Poisson model. The mean square error of each estimator is calculated against the target value of 5 in 1000 replications and (n times) the MSE curves are presented as a function of the sample size n in Figure 6. The MLE is strongly affected by the presence of the contaminating component, but it is clear that the inlier modification has not compromised the robustness of our proposed estimators in this case, which are extremely competitive or better than the MCDE and the ordinary MHDE.

5 Concluding Remarks

In this paper we have proposed a modification to the Hellinger distance to increase its small sample efficiency. Several members of the corresponding

Figure 6: Plot of the $n\times$MSE for contaminated Poisson data



family of estimators appear to remove some of the deficiencies of the minimum Hellinger distance estimator without compromising its robustness.

For the data analyst and the statistician, it will be useful if we can make an empirical recommendation on which of the methods to choose in a practical situation based on our simulations. It is well known – and clear in our numerical investigation – that the performance of the MLE takes a major hit under data contamination. On the other hand, the performance of the ordinary minimum Hellinger distance estimator is substantially worse under pure data compared to the other four estimators. The three estimators that perform reasonably in all the situations considered here are the MCDE, MIMHDE$^{0.5}$ and MIMHDE^{1}. These three estimators are so close in performance in all the situations we looked at that it is not possible to separate them in terms of their performance alone. Given that the asymptotics for the MCDE is not yet well developed, and that the MIMHDE^{1} uses a different scale factor in the denominator of (3.1), our preference will be for MIMHDE$^{0.5}$. We hope that more extensive future studies will shed more light on the hierarchy between these estimators in different situations.

The idea of dividing $G(\delta)$ by $(1+\delta^2 P(\delta))^{\gamma}$ where $P(\delta)$ is a polynomial in δ will also require further follow up to determine the effects of the additional

terms. Further modifications of this denominator term to increase the degree of smoothness of the $G(\delta)$ curve at $\delta = 0$, which has both theoretical and numerical implications, may also be looked into in the future.

Acknowledgements. The authors gratefully acknowledge the comments of a referee, which led to a improved version of the paper.

References

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

Basu, A. , Harris, I. R., Basu, S. (1996). Test of hypothesis in discrete models based on the penalized Hellinger distance. *Statist. Prob. Lett.* **27**, 367–373.

Beran, R. (1977) Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Statist.* **5**, 445–463.

Cressie, N., Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. B* **46**, 440–464.

Hampel, F. R., Ronchetti, E. M. , Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley, New York .

Harris, I. R, Basu, A. (1994). Hellinger distance as a penalized log likelihood. *Commun. -Statist. Simula. Computa.* **23**, 1097–1113.

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

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures.* Chapman & Hall/CRC, Taylor & Francis, Boca Raton, Florida .

Park, C., Basu, A. (2004). Minimum disparity estimation: asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics. Special issue in honour of Professor Takashi Yanagawa* **36**, 19-34.

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

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

Rohit Kumar Patra
Indian Statistical Institute
203 B.T. Road, Kolkata 700 018
India.
E-mail: bst0514@isical.ac.in

Abhijit Mandal
Applied Statistics Unit,
Indian Statistical Institute
203 B.T. Road, Kolkata 700 018
India.
E-mail: abhi_r@isical.ac.in

Ayanendranath Basu
Bayesian and Interdisciplinary
Research Unit
Indian Statistical Institute
203 B.T. Road, Kolkata 700 018
India.
E-mail: ayanbasu@isical.ac.in

Paper received September 2008; revised March 2009.