Sankhyā : The Indian Journal of Statistics
2010, Volume 72-A, Part 2, pp. 376-406

Minimum Disparity Inference and the Empty Cell Penalty: Asymptotic Results

Abhijit Mandal
Indian Statistical Institute, Kolkata, India
Ayanendranath Basu
Indian Statistical Institute, Kolkata, India
Leandro Pardo
Complutense University of Madrid, Madrid, Spain

Abstract

Inference procedures based on the Hellinger distance and other disparities provide attractive alternatives to likelihood based methods for the statistician. The minimum disparity estimators are asymptotically efficient under the model. Several members of this family also have strong robustness properties under model misspecification. Similarly, the disparity difference tests have the same null distribution as the likelihood ratio test but are often superior than the latter in terms of robustness properties. However, many disparities including the Hellinger distance put large weights on the empty cells which appears to be responsible for a somewhat poor efficiency of the corresponding methods in small samples. An artificial empty cell penalty has been shown to greatly improve the small sample properties of these procedures. However all studies involving the empty cell penalty have so far been empirical, and there are no results on the asymptotic properties of the minimum penalized disparity estimators and the corresponding tests. In view of the usefulness of these procedures this is a major gap in theory, which we try to fill through the present work.

AMS (2000) *subject classification.* Primary 62F12, 62F05; Secondary 62F35.
Keywords and phrases. Hellinger distance, empty cell penalty, minimum penalized disparity estimator, disparity difference test, asymptotic distribution.

1. **Introduction**

In recent times, density based divergences have been studied in the context of discrete models by, among others, Cressie and Read (1984) and Lindsay (1994). Pardo (2006) provides a good general reference for results relating to density based divergences in discrete models. 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), Tamura and Boos (1986), and Simpson (1987, 1989) provide much of the basic background and properties of minimum Hellinger distance inference.

Lindsay (1994) considered a large class of density based divergences called *disparities*. This includes the Hellinger distance and many other common density based divergences. The popularity of the minimum Hellinger distance procedures and those based on other robust disparities are partially tempered by a relatively poor efficiency of these methods compared to the likelihood based methods in small samples. However this trade off between robustness and small sample efficiency appears to be primarily caused, at least in discrete models, by the large weight that the Hellinger distance and the other robust disparities put on the empty cells. It has been empirically observed that an artificial empty cell penalty can greatly improve the small sample performance of the minimum disparity methods (see, e.g., Harris and Basu, 1994, Basu, Harris and Basu, 1996, Basu and Basu, 1998). In view of the improvements that result from the imposition of this penalty, the asymptotic properties of the minimum penalized disparity estimators and the corresponding tests are of great practical importance. However, such asymptotic properties of these methods have not been theoretically determined so far. In this paper we provide these asymptotic results.

The minimum disparity estimator is a member of the class of best asymptotically normal (BAN) estimators investigated and discussed by Lindsay (1994). The asymptotic distribution of the penalized estimators is an important piece in the theoretical results concerning BAN minimum distance estimators. Because the empty cell penalty is applied on a random subset of the sample space with vanishing probability, a theoretical comparison of the ordinary and penalized estimators based on higher order asymptotic properties appears to be very difficult. Yet there is overwhelming empirical evidence of improved performance of the penalized estimators in small samples over the ordinary estimators. Limited glimpses of such improved performance are provided in Section 7 with some heuristic justifications. However, the authors acknowledge that the scope remains for a more theoretical investigation about the precise source of the improvement in efficiency due to the application of the penalty.

2. Minimum Disparity Inference and the Empty Cell Penalty

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed observations from a discrete distribution F having probability mass function f with respect to the appropriate dominating measure. As we are dealing with density based divergences, we will represent our chosen model in terms of the corresponding probability mass functions $m_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $x \in \mathcal{X}$, the sample space.

Suppose $d_n(x)$ be the proportion of sample observations at x. The Pearson residual function $\delta_{n\theta}(x)$ at x is defined as

$$\delta_{n\theta}(x) = \frac{d_n(x) - m_\theta(x)}{m_\theta(x)}.$$

When there is no scope of ambiguity we will simply write $\delta_{n\theta},\ d_n$ and m_θ to denote the functions $\delta_{n\theta}(x),\ d_n(x)$ and $m_\theta(x)$ respectively.

Suppose that $G(\cdot)$ is a real-valued, thrice differentiable, strictly convex function on $[-1, \infty)$, with $G(0) = 0$. The disparity between $d_n(x)$ and $m_\theta(x)$ based on G is denoted by $\rho_G(d_n, m_\theta)$ and defined as

$$\rho_G(d_n, m_\theta) = \sum_{x \in \mathcal{X}} G(\delta_{n\theta}(x)) m_\theta(x). \tag{2.1}$$

The convexity of G and Jensen's inequality immediately imply, together with the condition $G(0) = 0$, that the disparity in (2.1) is non-negative. We can get many well known disparities by choosing specific forms of the function G. For example, $G(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ generates the well known likelihood disparity (LD) given by

$$\begin{aligned} \mathrm{LD}(d_n, m_\theta) &= \sum_{x \in \mathcal{X}} \left[d_n \log\left(\frac{d_n}{m_\theta}\right) + (m_\theta - d_n) \right] \\ &= \sum_{x \in \mathcal{X}} d_n \log\left(\frac{d_n}{m_\theta}\right), \end{aligned}$$

which is a form of the Kullback–Leibler divergence (Kullback and Leibler, 1951). Here log denotes the natural logarithm. The (twice, squared) Hellinger distance (HD) has the form

$$\mathrm{HD}(d_n, m_\theta) = 2 \sum_{x \in \mathcal{X}} \left(d_n^{1/2} - m_\theta^{1/2} \right)^2,$$

which corresponds to $G(\delta) = 2[(\delta+1)^{1/2} - 1]^2$.

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

$$\begin{aligned} I^{\lambda}(d_n, m_\theta) =& \frac{1}{\lambda(\lambda+1)} \sum_{x\in\mathcal{X}} d_n \left\{ \left(\frac{d_n}{m_\theta}\right)^{\lambda} - 1 \right\} \\ =& \sum_{x\in\mathcal{X}} \left[\frac{(\delta_{n\theta}+1)^{\lambda+1} - (\delta_{n\theta}+1)}{\lambda(\lambda+1)} - \frac{\delta_{n\theta}}{\lambda+1} \right] m_\theta. \end{aligned} \tag{2.2}$$

The power divergence family contains many well known divergences as special cases. The likelihood disparity corresponds to $\lambda = 0$ defined via the continuous limit of the quantity in the right hand side of (2.2) as $\lambda \to 0$. The (twice, squared) Hellinger distance is also a member of the power divergence family for $\lambda = -0.5$.

Let $\hat{\theta}_n$ be the estimator of θ that minimizes ρ_G over $\theta \in \Theta$; then $\hat{\theta}_n$ is the *minimum disparity estimator* (MDE) of θ corresponding to ρ_G. Provided it exists, $\hat{\theta}_h$ satisfies

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

Notice that the minimizer of the likelihood disparity is the maximum likelihood estimator, so that the latter estimator is a member of the class of minimum disparity estimators. Under differentiability of the model, the estimating equation for θ is of the form

$$-\nabla\rho_G(d_n, m_\theta) = \sum_x \left[G'(\delta_{n\theta}) \frac{d_n}{m_\theta} - G(\delta_{n\theta}) \right] \nabla m_\theta = 0, \tag{2.3}$$

where $\nabla = (\nabla_1, \nabla_2, \ldots, \nabla_p)^T$ denotes the gradient operator with respect to θ, and G' is the first derivative of G. The estimating equation above can be written as

$$\sum_x A_G(\delta_{n\theta}(x)) \nabla m_\theta(x) = 0, \tag{2.4}$$

where

$$A_G(\delta) = (1+\delta)G'(\delta) - G(\delta).$$

The function A_G is called the *residual adjustment function* (RAF) of the disparity. Since $A'_G(\delta) = (1+\delta)G''(\delta)$ and G is strictly convex, the function $A_G(\delta)$ is a strictly increasing function on $\delta \in [-1, \infty)$. As $\sum_x \nabla m_\theta(x) = 0$,

we can redefine the function $A_G(\delta)$ by $A_G(\delta) = A_G(\delta) - A_G(0)$, so that $A_G(0) = 0$. Similarly without changing the solution of the estimating equation we can rescale the function A_G to make $A'_G(0) = 1$ (since $A'_G(0) = G''(0) > 0$). These two conditions are automatic if, in addition to its usual properties, the associated G function satisfies

$$G'(0) = 0, \text{ and } G''(0) = 1. \tag{2.5}$$

Notice also that for any disparity satisfying $G(0) = 0$ and $G'(0) = 0$, the convexity of G guarantees that each term in the summand of (2.1) is itself non-negative. Equation (2.4) shows that within the class of minimum disparity estimators the estimating equation of the MDE can be distinguished by the function $A_G(\cdot)$. Thus the theoretical properties of the MDE are controlled by the nature of the RAF. For more details see Lindsay (1994) and Basu, Harris and Basu (1997).

To analyze the robustness properties of these minimum disparity estimators, one has to characterize the outliers probabilistically. If an observation x in the sample space has a large positive value of $\delta_{n\theta}(x)$, it will be called an outlier in the sense that the actual observed proportion is much larger at that point 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 $\delta_{n\theta}$. For such disparities, the RAF $A_G(\delta)$ would exhibit a severely dampened response to increasing δ. For a qualitative description, one can take the RAF of the likelihood disparity $A_{LD}(\delta)$ as the basis for comparison. For this disparity, $A_{LD}(\delta) = \delta$, 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. A graph of the RAFs of some of the common disparities within the power divergence family is given in Figure 1. Disparities with large negative values of λ, for which the RAFs curve sharply down on the right hand side of the δ axis, are expected to perform better in terms of robustness.

It may be noted from Figure 1 that the robust disparities also curve down sharply on the left tail, and put large negative weights on the empty cells (i.e. at $\delta = -1$). In small sample sizes, where a large number of empty cells is likely, this large negative weight on the the empty cells appears to have a strong adverse impact on the performance of the estimator. Therefore an artificial empty cell penalty may be expected to give better small sample performance. In the subsequent sections we will show that the penalty does not affect the asymptotic distribution of the minimum disparity estimator or the corresponding null distribution of the disparity difference test statistic.

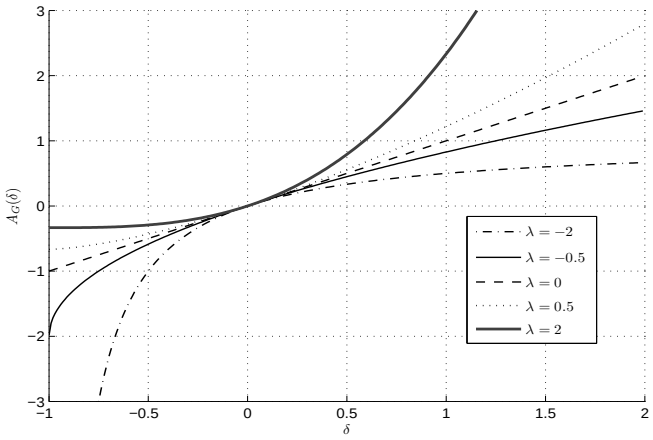

Figure 1: Plot of the RAFs $A_G(\delta)$ for different values of λ for the power divergence family.

3. **The Penalized Disparity**

Consider the setup of the previous section where $X_1, X_2, \ldots, X_n$ are n independent and identically distributed observations from a discrete distribution having probability mass function modeled by $\{m_\theta(x)\}$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $x \in \mathcal{X}$. Assume that the disparity generating function $G(\cdot)$ satisfies the conditions in (2.5) in addition to its usual properties. The disparity in (2.1) can be rewritten as

$$\rho_G(d_n, m_\theta) = \sum_{x:d_n(x)>0} G(\delta_{n\theta}(x))m_\theta(x) + G(-1) \sum_{x:d_n(x)=0} m_\theta(x). \qquad (3.1)$$

Now the penalized disparity for the tuning parameter h between the densities d_n and m_θ is defined as

$$\rho_{G_h}(d_n, m_\theta) = \sum_{x:d_n(x)>0} G(\delta_{n\theta}(x))m_\theta(x) + h \sum_{x:d_n(x)=0} m_\theta(x),\ h > 0. \quad (3.2)$$

It is clear that the penalized disparity in (3.2) is non-negative; also evident is the fact that if the probability mass functions d_n and m_θ are identically equal the penalized disparity must equal zero. Again, for $h > 0$, two probability mass functions which are not identically equal must necessarily produce a

positive penalized disparity. If the support of m_θ is independent of θ, the range of h can be increased to include $h = 0$.

Comparing with the likelihood disparity

$$\mathrm{LD}(d_n, m_\theta) = \sum_{x:d_n(x)>0} \left[d_n(x) \log \left(\frac{d_n(x)}{m_\theta(x)} \right) + (m_\theta(x) - d_n(x)) \right] + \sum_{x:d_n(x)=0} m_\theta(x),$$

we observe that the penalized disparity in (3.2) will put the same weight on the empty cells as given by the likelihood disparity when the tuning parameter h equals 1. The natural weight applied on the empty cells by the ordinary disparity corresponds to $h = G(-1)$.

The minimum penalized disparity estimator (MPDE) $\hat{\theta}_n^h$ is obtained by minimizing $\rho_{G_h}(d_n, m_\theta)$ over $\theta \in \Theta$, where ρ_{G_h} is as given in (3.2). So

$$\rho_{G_h}(d_n, m_{\hat{\theta}_n^h}) = \min_{\theta \in \Theta} \rho_{G_h}(d_n, m_\theta),$$

provided such a minimum exists. We will suppress the superscript h in $\hat{\theta}_n^h$, whenever there is no scope for confusion.

For a parametric model m_θ with infinite support, the set $\{x : d_n(x) = 0\}$ is also infinite where $d_n(x)$ is based on a sample of size n as described above. Thus the proper control of the term $\sum_{x:d_n(x)=0} m_\theta(x)$ can lead to large benefits when the natural weight is too large.

In general the penalized disparity is a function of h. All the results and proofs in the rest of this paper correspond to general values of h. In Section 7 we provide a thorough numerical investigation involving the role of h.

4. Consistency of the Minimum Penalized Disparity Estimator

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed observations from a discrete distribution within the model family having probability mass function $m_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $x \in \mathcal{X}$. Let $\theta^0 \in \Theta$ be the true value of the parameter. We define the score function and its successive derivatives as $u_{i\theta}(x) = \nabla_i \log m_\theta(x)$, $u_{ij\theta}(x) = \nabla_{ij} \log m_\theta(x)$ and $u_{ijk\theta}(x) = \nabla_{ijk} \log m_\theta(x)$. Here and elsewhere we will assume that the family $\{m_\theta\}$ is identifiable in the sense that $\theta_1 \neq \theta_2$ implies that the set $A = \{x : m_{\theta_1}(x) \neq m_{\theta_2}(x)\}$ has positive measure.

Consider an ordinary disparity $\rho_G(d_n, m_\theta)$ and the corresponding penalized disparity $\rho_{G_h}(d_n, m_\theta)$. Let A_G represents the residual adjustment function of the ordinary disparity. We define the function A_h as

$$A_h(\delta) = \begin{cases} A_G(\delta), & \text{if } \delta > -1, \\ -h, & \text{if } \delta = -1, \end{cases} \tag{4.1}$$

and denote it as the residual adjustment function of the penalized disparity. Note that the estimating equation of the minimum penalized estimator is given by

$$\sum_x A_h(\delta_{n\theta}(x))\nabla m_\theta = 0.$$

We prove the consistency of the minimum penalized disparity estimator in the following theorem.

THEOREM 4.1. *Assume that the following conditions are satisfied.*

(A0) *The probability mass function m_θ of the observations X have common support so that the set $\mathcal{X} = \{x : m_\theta(x) > 0\}$ is independent of θ.*

(A1) *There exists an open subset ω of Θ for which the true parameter θ^0 is an interior point, and for almost all x the density $m_\theta(x)$ admits all third derivatives of the type $\nabla_{ijk} m_\theta(x)$ for all $\theta \in \omega$.*

(A2) *The first two derivatives of $m_\theta(x)$ with respect to θ satisfy the following equations*

$$E_\theta[u_{i\theta}(X)] = 0 \text{ for all } i = 1, 2, \ldots, p,$$

and

$$I_{jk}(\theta) = E_\theta[u_{j\theta}(X)u_{k\theta}(X)] = -E_\theta\left[u_{jk\theta}(X)\right], \text{ for all } j, k,$$

where the matrix $I(\theta) = ((I_{jk}(\theta)))_{p\times p}$ is the Fisher information matrix at m_θ. We assume that $I(\theta)$ is positive definite for all $\theta \in \omega$.

(A3) *The quantities*

$$\sum_x m_\theta^{\frac{1}{2}}(x)|u_{i\theta}(x)|, \sum_x m_\theta^{\frac{1}{2}}(x)|u_{i\theta}(x)u_{j\theta}(x)| \text{ and } \sum_x m_\theta^{\frac{1}{2}}(x)|u_{ij\theta}(x)|$$

are bounded for all i and j and all $\theta \in \omega$.

(A4) *For almost all x there exist functions $M_{ijk}(x), M_{ij,k}(x), M_{i,j,k}(x)$ that dominate $|u_{ijk\theta}(x)|$, $|u_{ij\theta}(x)u_{k\theta}(x)|$ and $|u_{i\theta}(x)u_{j\theta}(x)u_{k\theta}(x)|$ for all i, j, and k, and that are uniformly bounded in expectation E_θ for all $\theta \in \omega$.*

(A5) The RAF $A_G(\delta)$ is such that $A'_G(\delta)$ and $(1+\delta)A''_G(\delta)$ are bounded in absolute value by positive real constants M and N respectively on $[-1, \infty)$.

Under the above conditions, with probability tending to 1 as $n \to \infty$, the minimum penalized disparity estimating equation

$$\frac{\partial}{\partial\theta}\rho_{G_h}(d_n, m_\theta) = 0 \tag{4.2}$$

has a root $\hat{\theta}_n$ such that $\hat{\theta}_n$ is consistent for estimating θ^0.

PROOF. Suppose Q_a is the sphere with center at the true parameter θ^0 and radius $a > 0$. We will first show that for any sufficiently small a, $\rho_{G_h}(d_n, m_{\theta^0}) < \rho_{G_h}(d_n, m_\theta)$, for all θ on the surface of Q_a, with probability tending to 1. Thus the disparity measure $\rho_{G_h}(d_n, m_\theta)$ has a local minimum in the interior of Q_a with probability tending to 1. Hence the minimum disparity estimating equation (4.2) has a solution $\hat{\theta}_n(a)$ within Q_a with probability tending to 1.

Taking a Taylor series expansion of $\rho_{G_h}(d_n, m_\theta)$ about θ^0 we get

$$\begin{aligned}
&\rho_{G_h}(d_n, m_{\theta^0}) - \rho_{G_h}(d_n, m_\theta) \\
=& -\Big\{ \sum_j (\theta_j - \theta_j^0)\nabla_j\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} \\
&\quad + \frac{1}{2}\sum_{j,k}(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} \\
&\quad + \frac{1}{6}\sum_{j,k,l}(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)(\theta_l - \theta_l^0)\nabla_{jkl}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*}\Big\} \\
=& S_1 + S_2 + S_3 \text{ (say)},
\end{aligned} \tag{4.3}$$

where θ^* lies on the line segment joining θ and θ^0; θ_j and θ_j^0 are the j-th components of the indicated vectors. From Lemma 9.1 we get

$$\begin{aligned}
\nabla_j\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} &= -\sum_x d_n(x)u_{j\theta^0}(x) + o_p(n^{-1/2}) \\
&\to 0, \text{ as } n \to \infty,
\end{aligned} \tag{4.4}$$

so that the left hand side of (4.4) is bounded in absolute value by a^2 with probability tending to 1. Thus on Q_a we have

$$|S_1| < pa^3 \tag{4.5}$$

with probability tending to 1.

Now from Lemma 9.2, we get

$$\begin{aligned}\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} =& \sum m_{\theta^0}(x)u_{j\theta^0}(x)u_{k\theta^0}(x) + o_p(1)\\ =& I_{jk}(\theta^0) + o_p(1).\end{aligned}$$

Hence

$$\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} \longrightarrow I_{jk}(\theta^0) \text{ as } n \to \infty.$$

So $\left|\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} - I_{jk}(\theta^0)\right|$ is bounded by a with probability tending to 1. Now

$$\begin{aligned}2S_2 =& \sum_{j,k}(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)\left\{-\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} - (-I_{jk}(\theta^0))\right\}\\ &+ \sum_{j,k}(-I_{jk}(\theta^0))(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)\\ <& p^2a^3 + (\theta - \theta^0)^T(-I(\theta^0))(\theta - \theta^0) \qquad (4.6)\end{aligned}$$

with probability tending to 1. The second term in the right hand side of (4.6) is a negative definite quadratic form in the variables $(\theta_j - \theta_j^0)$. Letting λ_1 be the largest eigenvalue of $-I(\theta^0)$, this quadratic form is less than $\lambda_1 a^2$. Combining the two terms we see that there exists $c > 0$ and $a_0 > 0$ such that for $a < a_0$,

$$S_2 < -ca^2 \qquad (4.7)$$

with probability tending to 1. For the cubic term S_3 notice that

$$S_3 \leq \frac{a^3}{6}\sum_{j,k,l}|\nabla_{jkl}\ \rho_{G_h}(d_n, m_\theta)|\Big|_{\theta=\theta^*}. \qquad (4.8)$$

From Lemma 9.3 we can find a positive number γ, such that with probability tending to 1

$$|\nabla_{jkl}\rho_{G_h}(d_n, m_\theta)|\Big|_{\theta=\theta^*} < \gamma < \infty,$$

for all j, k and l. So from (4.8) it follows that

$$S_3 < ba^3, \qquad (4.9)$$

for a positive constant b. Hence combining (4.5), (4.7) and (4.9) we get from (4.3)

$$S_1 + S_2 + S_3 < -ca^2 + (b + p)a^3 \qquad (4.10)$$

with probability tending to 1, which is less that zero for $a < c/(b+p)$.

Thus for any sufficiently small a there exists a sequence of roots $\hat{\theta}_n = \hat{\theta}_n(a)$ to equation (4.2) such that

$$P_{\theta^0}(||\hat{\theta}_n - \theta^0|| < a) \to 1 \text{ as } n \to \infty, \tag{4.11}$$

where $||\cdot||$ represents the L_2 norm. Let θ_n^* be the root closest to θ^0 (it exists as limit of a sequence of roots is again a root by the continuity of $\rho_{G_h}(d_n, m_\theta)$). Then clearly $P_{\theta^0}(||\theta_n^* - \theta^0|| < a) \to 1$ as $n \to \infty$, and the sequence θ_n^* does not depend on a. Hence there exists a root of (4.2) which tends to the true value θ^0 in probability. □

5. Asymptotic Distribution of the Minimum Penalized Disparity Estimator

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed observations from a discrete distribution within the model family having probability mass function $m_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $x \in \mathcal{X}$. Let $\theta^0 \in \Theta$ be the true value of the parameter.

THEOREM 5.1. *Under Assumptions (A0)–(A5) of Theorem 4.1, the minimum penalized disparity estimator, $\hat{\theta}_n^h$, satisfies*

$$n^{1/2}(\hat{\theta}_n^h - \theta^0) \overset{a}{\sim} N_p\left(0, I^{-1}(\theta^0)\right),$$

where the $\overset{a}{\sim}$ notation represents asymptotic distribution, and $I(\theta)$ is the Fisher information matrix as defined in assumption (A2) of Theorem 4.1.

PROOF. Let us assume the set up and notation of Section 4. Let

$$\sum_x A_h(\delta_{n\theta}(x))\nabla m_\theta(x) = 0 \tag{5.1}$$

be the penalized estimating equation, where A_h is as defined in (4.1). Equation (5.1) is solved by $\hat{\theta}_n^h$, the minimum penalized disparity estimator of θ. Let us expand the j-th component of the left hand side of the above equation around $\theta = \theta^0$. This gives

$$\begin{aligned}
&\sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x) \\
&= \sum_x A_h(\delta_{n\theta^0}(x))\nabla_j m_{\theta^0}(x) + \sum_k (\theta_k - \theta_k^0)\nabla_k \sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^0}
\end{aligned}$$

$$+\frac{1}{2}\sum_{kl}(\theta_k-\theta_k^0)(\theta_l-\theta_l^0)\nabla_{kl}\sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^*}, \quad (5.2)$$

where θ^* lies on the line segment joining θ and θ^0. We replace θ with $\hat{\theta}_n^h$, the minimum penalized disparity estimator, so that the the left hand side of the above equation becomes zero. Multiplying by $n^{1/2}$ and rearranging terms, the equation (5.2) can then be rewritten as

$$\begin{aligned}&-n^{1/2}\sum_x A_h(\delta_{n\theta^0}(x))\nabla_j m_{\theta^0}(x)\\ =&n^{1/2}\sum_k(\hat{\theta}_{nk}^h-\theta_k^0)\Big[\nabla_k\sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^0}\\ &+\frac{1}{2}\sum_l(\hat{\theta}_{nl}^h-\theta_l^0)\nabla_{kl}\sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^*}\Big]. \quad (5.3)\end{aligned}$$

Now from Lemma 9.1,

$$\sum_x A_h(\delta_{n\theta^0}(x))\nabla_j m_{\theta^0}(x)=\sum_x d_n(x)u_{j\theta^0}(x)+o_p\left(\frac{1}{\sqrt{n}}\right),$$

and hence

$$\begin{aligned}-n^{1/2}\sum_x A_h(\delta_{n\theta^0}(x))\nabla_j m_{\theta^0}(x)&=-n^{1/2}\sum_x d_n(x)u_{j\theta^0}(x)+o_p(1)\\ &=-n^{1/2}\frac{1}{n}\sum_{i=1}^n u_{j\theta^0}(X_i)+o_p(1), \quad (5.4)\end{aligned}$$

which shows that $-n^{1/2}\sum_x A_h(\delta_{n\theta^0}(x))\nabla_j m_{\theta^0}(x)\overset{a}{\sim}N(0,I_{jj}(\theta^0))$. Thus the p dimensional vector

$$Y=-n^{1/2}\Big(\sum_x A_h(\delta_{n\theta^0}(x))\nabla_1 m_{\theta^0}(x),\ldots,\sum_x A_h(\delta_{n\theta^0}(x))\nabla_p m_{\theta^0}(x)\Big)^T$$

is asymptotically equivalent to

$$Z_{\theta^0}=-n^{1/2}\Big(\frac{1}{n}\sum_i u_{1\theta^0}(X_i),\ldots,\frac{1}{n}\sum_i u_{p\theta^0}(X_i)\Big)^T, \quad (5.5)$$

and has an asymptotic multivariate normal distribution with mean vector 0 and covariance matrix $I(\theta^0)$.

From Lemma 9.2 it follows that

$$\nabla_k \sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^0} - (-I_{jk}(\theta^0))$$

converges to zero in probability. Again from Lemma 9.3, the quantity

$$\nabla_{kl} \sum_x A_h(\delta_{n\theta}(x))\nabla_j m_\theta(x)\Big|_{\theta=\theta^*}$$

is bounded in probability. Since, by Theorem 4.1, $\hat{\theta}_n^h$ is consistent for θ^0, the bracketed quantity on the right hand side of (5.3) goes to $-I_{jk}(\theta^0)$ in probability. It then follows from Lehmann (1983, Lemma 4.1) that the asymptotic distribution of $n^{1/2}(\hat{\theta}_n^h - \theta_0)$ is multivariate normal with mean vector 0 and covariance matrix $I^{-1}(\theta^0)I(\theta^0)I^{-1}(\theta^0) = I^{-1}(\theta^0)$. This establishes the required result. □

Essentially the above shows that the $n^{1/2}(\hat{\theta}_n^h - \theta^0)$ is asymptotically equivalent to $I^{-1}(\theta^0)Z_{\theta^0}$, where Z_{θ^0} is defined in equation (5.5), and hence is also asymptotically equivalent to $n^{1/2}(\hat{\theta}_n^{ML} - \theta^0)$ in the sense

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

where $\hat{\theta}_n^{ML}$ is the maximum likelihood estimator of θ.

Other approaches, such as the one by Seo and Lindsay (2009) may also be adopted to prove the consistency and asymptotic normality of the minimum penalized disparity estimators.

6. Asymptotic Null Distribution of the Penalized Disparity Difference Test Statistic

Now consider the parametric hypothesis testing problem under the set up described in Section 4. Let Θ_0 be the subset of Θ with $r \leq p$ restrictions on the vector θ such that $C_i(\theta) = 0,\ i = 1, 2, \cdots, r$. Let us present the composite null hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta - \Theta_0$. Under the null H_0 may be described by the parameter $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma^{p-r})^T$ with $p-r$ independent components, where $\gamma \in \Gamma \subseteq \mathbb{R}^{p-r}$. In this case there exists a function $\eta : \mathbb{R}^{p-r} \to \mathbb{R}^p$ such that $\theta = \eta(\gamma)$, where $\theta \in \Theta_0$ and $\gamma \in \Gamma$. If H_0 is true, then there exists a $\gamma^0 \in \Gamma$ such that $\theta^0 = \eta(\gamma^0)$, where $\theta^0 \in \Theta$ is the true value of θ. We assume that η has continuous second derivatives in an open set containing γ^0. Suppose the first derivative $\dot{\eta}(\gamma)$ of order $p \times (p-r)$

has full column rank at $\gamma = \gamma^0$. Under the set up of the previous sections the ordinary disparity difference test statistic is given by

$$T_n = 2n \left\{ \rho_G(d_n, m_{\hat{\theta}_0}) - \rho_G(d_n, m_{\hat{\theta}}) \right\}, \tag{6.1}$$

and the penalized disparity difference test statistic is given by

$$T_n^p = 2n \left\{ \rho_{G_h}(d_n, m_{\hat{\theta}_0}) - \rho_{G_h}(d_n, m_{\hat{\theta}}) \right\}, \tag{6.2}$$

where, depending on the case, $\hat{\theta}$ represents the unrestricted minimizer of ρ_G or ρ_{G_h}, while $\hat{\theta}_0$ is the corresponding minimizers under the null.

THEOREM 6.1. *Under assumptions (A0)–(A5) in Theorem 4.1, the null distribution of the penalized disparity difference test statistic tends to a χ^2 distribution with r degrees of freedom.*

PROOF. Let $\theta^0 \in \Theta$ be the true value of θ. In the unrestricted situation we get from Theorem 5.1 that

$$n^{1/2}(\hat{\theta} - \theta^0) = I^{-1}(\theta^0) Z_{\theta^0} + o_p(1),$$

where Z_{θ^0} is defined in (5.5). If $\hat{\theta}^{ML}$ is the unrestricted maximum likelihood estimate of θ, then

$$n^{1/2}(\hat{\theta} - \theta^0) = n^{1/2}(\hat{\theta}^{ML} - \theta^0) + o_p(1). \tag{6.3}$$

If H_0 is true, then there exists a $\gamma^0 \in \Gamma$ such that $\theta^0 = \eta(\gamma^0)$. Let $\hat{\gamma}^0$ be the minimum penalized disparity estimate of γ under H_0, then $\hat{\theta}_0 = \eta(\hat{\gamma}^0)$. Suppose $\hat{\gamma}^{0ML}$ is the maximum likelihood estimate of γ under H_0. By Theorem 5.1 it is easy to show that, under H_0,

$$n^{1/2}(\hat{\gamma}^0 - \gamma^0) = n^{1/2}(\hat{\gamma}^{0ML} - \gamma^0) + o_p(1). \tag{6.4}$$

So using delta method we get

$$n^{1/2}(\hat{\theta}_0 - \theta^0) = n^{1/2}(\hat{\theta}_0^{ML} - \theta^0) + o_p(1), \tag{6.5}$$

where $\hat{\theta}_0^{ML}$ is the maximum likelihood estimate of θ under H_0. Combining (6.3) and (6.5) we get

$$n^{1/2}(\hat{\theta}_0 - \hat{\theta}) = n^{1/2}(\hat{\theta}_0^{ML} - \hat{\theta}^{ML}) + o_p(1). \tag{6.6}$$

From Therorem 4.4.4 of Serfling (1980), $n^{1/2}(\hat{\theta}_0^{ML} - \hat{\theta}^{ML})$ is $O_p(1)$, so

$$n^{1/2}(\hat{\theta}_0 - \hat{\theta}) = O_p(1). \tag{6.7}$$

Now taking a Taylor series expansion of (6.2) about $\hat{\theta}_0 = \hat{\theta}$ gives

$$\begin{aligned} T_n^p =& 2n\{\rho_{G_h}(d_n, m_{\hat{\theta}_0}) - \rho_{G_h}(d_n, m_{\hat{\theta}})\} \\ =& 2n \sum_j (\hat{\theta}_{0j} - \hat{\theta}_j) \nabla_j \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\hat{\theta}} \\ &+ n \sum_{j,k} (\hat{\theta}_{0j} - \hat{\theta}_j)(\hat{\theta}_{0k} - \hat{\theta}_k) \nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*}, \end{aligned} \tag{6.8}$$

where θ^* lies on the line segment joining $\hat{\theta}_0$ and $\hat{\theta}$; $\hat{\theta}_{0j}$ and $\hat{\theta}_j$ are the j-th components of the indicated vectors. Since $\nabla_j \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\hat{\theta}} = 0$, we get

$$\begin{aligned} T_n^p =& n \sum_{j,k} (\hat{\theta}_{0j} - \hat{\theta}_j)(\hat{\theta}_{0k} - \hat{\theta}_k) \nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*} \\ =& n(\hat{\theta}_0 - \hat{\theta})^T I(\theta^0)(\hat{\theta}_0 - \hat{\theta}) \\ &+ n(\hat{\theta}_0 - \hat{\theta})^T [\nabla_2 \rho_{G_h}(d_n, m_{\theta^*}) - I(\theta^0)](\hat{\theta}_0 - \hat{\theta}), \end{aligned} \tag{6.9}$$

where

$$\nabla_2 \rho_{G_h}(d_n, m_{\theta^*}) = ((\nabla_{jk} \rho_{G_h}(d_n, m_\theta)))_{p\times p}\Big|_{\theta=\theta^*}.$$

Now we will show that

$$\nabla_2 \rho_{G_h}(d_n, m_{\theta^*}) - I(\theta^0) = o_p(1).$$

For this we need to establish, for each j and k,

$$\nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*} = I_{jk}(\theta^0) + o_p(1). \tag{6.10}$$

Using Taylor series expansion we get

$$\begin{aligned} \nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*} =& \nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} \\ &+ \sum_l (\theta_l^* - \theta_l^0) \nabla_{jkl} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^{**}}, \end{aligned} \tag{6.11}$$

where θ^{**} lies on the line segment joining θ^* and θ^0. From Lemma 9.3 we find that $\nabla_{jkl} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^{**}}$ is bounded in probability. Again from Theorem 4.1 we get $(\hat{\theta} - \theta^0) = o_p(1)$ and equation (6.5) gives $(\hat{\theta}_0 - \theta^0) = o_p(1)$. Therefore $(\theta^* - \theta^0) = o_p(1)$. So

$$\sum_l (\theta_l^* - \theta_l^0) \nabla_{jkl} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^{**}} = o_p(1).$$

Hence equation (6.11) can be written as

$$\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*} = \nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} + o_p(1).$$

Thus using Lemma 9.2 we get

$$\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^*} = I_{jk}(\theta^0) + o_p(1). \tag{6.12}$$

Now using (6.7) equation (6.9) reduces to

$$T_n^p = n(\hat{\theta}_0 - \hat{\theta})^T I(\theta^0)(\hat{\theta}_0 - \hat{\theta}) + o_p(1). \tag{6.13}$$

Using (6.6) we get from the above equation

$$T_n^p = n(\hat{\theta}_0^{ML} - \hat{\theta}^{ML})^T I(\theta^0)(\hat{\theta}_0^{ML} - \hat{\theta}^{ML}) + o_p(1). \tag{6.14}$$

From the proof of Theorem 4.4.4 of Serfling (1980) it follows that the asymptotic distribution of $n(\hat{\theta}_0^{ML} - \hat{\theta}^{ML})^T I(\theta^0)(\hat{\theta}_0^{ML} - \hat{\theta}^{ML})$ is chi-square with r degrees of freedom. Thus the proof is complete. □

7. **The Role of** h

In this section we present the results of some numerical investigations providing confirmation of the improved performance due to application of the penalty over a moderately wide range of models, parameters and sample sizes. We use the Hellinger distance and its penalized versions for illustration.

7.1. *A Data Dependent Choice for the Penalty Weight.* Theorem 5.1 shows that MPDEs are BAN estimators irrespective of the choice of h. It verifies our intuition that when the sample size is large, the probability of the empty cells eventually becomes sufficiently small, and the amount of the empty cell penalty does not affect the asymptotic distribution of the estimator. But empirical studies show that in small sample sizes, where the number of empty cells may be large, the estimators are, in fact, quite sensitive to the choice of h. In order to choose the optimum penalty we need to express the mean square error (MSE) of the MPDE for a fixed sample size as a function of h and then choose that h for which the MSE is minimum. This is a very difficult thing to do in practice; however in this section we present an approximate method with the aim of minimizing the error in estimation based on Rao's (1963) expression for the estimated bias in multinomial model for

the power divergence family, which determines an estimate of the optimal choice of h.

Suppose $\hat{\theta}_n^h$ denotes the MPDE obtained as a solution of equation (4.2). Let us denote the estimating function $\frac{\partial}{\partial\theta}\rho_{G_h}(d_n, m_\theta)$ by $f(d_n, \theta, h)$. Our aim here is to linearly approximate the change in $\hat{\theta}_n^h$ as a function of h. From Lemma 9.4 we get

$$\eta(d_n, \hat{\theta}_n^h, h) = -\left(\frac{\frac{\partial}{\partial h} f(d_n, \theta, h)}{\frac{\partial}{\partial\theta} f(d_n, \theta, h)}\right)\Bigg|_{\theta=\hat{\theta}_n^h}, \tag{7.1}$$

where $\eta(d_n, \hat{\theta}_n^h, h) = \frac{\partial\hat{\theta}_n^h}{\partial h}$ is the rate of change of $\hat{\theta}_n^h$ as a function of h. For the power divergence family in equation (2.2) we get, after some simple algebra,

$$\frac{\partial}{\partial h} f(d_n, \theta, h)\Big|_{\theta=\hat{\theta}_n^h} = -\sum_x m'_{\hat{\theta}_n^h}(x) I(d_n(x)),$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta} f(d_n, \theta, h)\Big|_{\theta=\hat{\theta}_n^h} = &-\sum_x \left(\frac{d_n(x)}{m_{\hat{\theta}_n^h}(x)}\right)^{\lambda+1} \frac{m'^2_{\hat{\theta}_n^h}(x)}{m_{\hat{\theta}_n^h}(x)} \\ &+ \frac{1}{\lambda+1}\sum_x \left(\frac{d_n(x)}{m_{\hat{\theta}_n^h}(x)}\right)^{\lambda+1} m''_{\hat{\theta}_n^h}(x) \\ &+ \left(\frac{1}{\lambda+1} - h\right)\sum_x m''_{\hat{\theta}_n^h}(x) I(d_n(x)),\end{aligned}$$

where $I(y) = 1$ if $y = 0$ and 0 otherwise.

The expression $\eta(d_n, \hat{\theta}_n^h, h)$ may be viewed as the increment in the estimator for unit change in h. Let $\hat{\theta}_n$ be the ordinary minimum disparity estimator without using the penalty, and let θ^0 be the true value of θ. If we use the linear approximation based on (7.1) the 'optimal' h, which will eliminate the error $(\hat{\theta}_n - \theta^0)$ in estimation, is

$$h_{opt} = G(-1) - \frac{(\hat{\theta}_n - \theta^0)}{\eta(d_n, \hat{\theta}_n, G(-1))}. \tag{7.2}$$

As h_{opt} is a function of θ, we can not compute it directly. So we replace $(\hat{\theta}_n - \theta^0)$ in (7.2) by the estimated bias of the ordinary minimum disparity estimator. In calculating the bias for a real-valued parameter (i.e. when

dimension of θ is unity) we follow the approach of Rao (1963). Putting $\lambda = -(k+1)$ in Haldane's minimum discrepancy (see page 204 of Rao, 1963) we get the bias for the power divergence family as

$$E(\hat{\theta}_n - \theta^0) = \frac{1}{2nI(\theta^0)} \left\{ \lambda \sum_x \frac{m'_{\theta^0}(x)}{m_{\theta^0}(x)} - \frac{\lambda\mu_{30} + \mu_{11}}{I(\theta^0)} \right\} + o\left(\frac{1}{n}\right), \qquad (7.3)$$

where

$$\mu_{uv} = \sum_x m_{\theta^0} \left(\frac{m'_{\theta^0}}{m_{\theta^0}}\right)^u \left(\frac{m''_{\theta^0}}{m_{\theta^0}}\right)^v,$$

and $I(\theta^0)$ is the Fisher information defined in Theorem 4.1.

We estimate $E(\hat{\theta}_n - \theta^0)$ by the first term on the right hand side of (7.3) evaluated at $\theta^0 = \hat{\theta}_n$, and replace $(\hat{\theta}_n - \theta^0)$ in (7.2) by the estimated value. If the estimated h_{opt} comes out to be negative it is replaced by zero. We then follow it up by finding the minimum penalized disparity estimator $\hat{\theta}_n^{h_{opt}}$, and choose that as our final estimator. Thus this case involves the use of a data dependent penalty.

In Figure 2 we present the mean square errors of four different estimators in a particular numerical example. Data are randomly generated from a

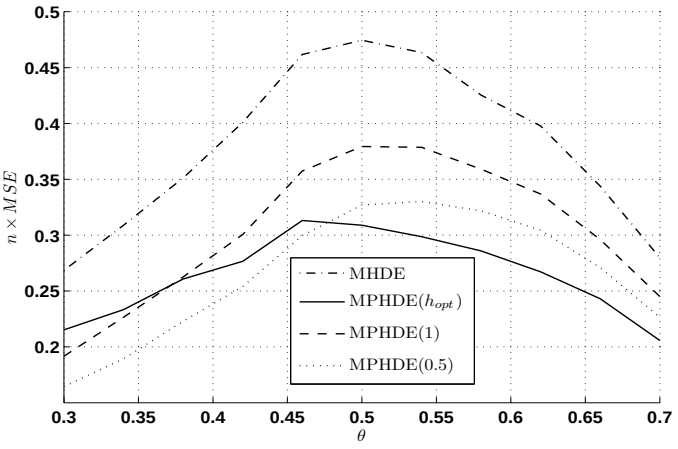

Figure 2: The MSE (multiplied by n) of the estimators for the truncated geometric distribution with parameter θ, where the number of cells is 5 and $n = 15$.

truncated geometric distribution with 5 cells having probability mass function

$$m_\theta(x) = \begin{cases} \theta(1-\theta)^x, & \text{if } x = 0, 1, 2, 3, \\ 1 - \sum_{x=0}^{3} m_\theta(x), & \text{if } x = 4. \end{cases}$$

The values of the parameter θ are chosen to be between 0.3 and 0.7. Samples of size 15 are drawn each time, and the exercise is repeated 1000 times. The four estimators chosen are (a) the ordinary minimum Hellinger distance estimator (MHDE) of θ, (b) the minimum penalized Hellinger distance estimator (MPHDE) of θ based on the estimated optimal choice of h_{opt} as in (7.2), (c) the minimum penalized Hellinger distance estimator of θ based on a fixed penalty of $h = 1$, and (d) the minimum penalized Hellinger distance estimator of θ based on a fixed penalty of $h = 0.5$.

The plot shows that all the minimum penalized Hellinger distance estimators of θ provide substantial improvements on the ordinary minimum Hellinger distance estimator. It appears that the MPHDE for $h = 1$ is slightly inferior than the other two penalized estimators. The estimator in (b) does not necessarily beat out the one in (d), and the performance of the estimators in (b) and (d) appear to be competitive. We realize that the estimation of the optimal choice of h in (7.2) involves some approximations and perhaps it should be viewed generally as a 'good, suitable, data dependent' choice.

The distribution of the optimal choice of h, at least in this example, also reveals that the values of h_{opt} are almost always smaller than 2, the natural empty cell weight of the Hellinger distance, although there is some variation based on the true value of θ. For $\theta = 0.5$, for example, the median value of h_{opt} equals 0.623 over the 1000 replications. In this illustration, reducing the weight of the empty cells clearly improves the performance of the minimum Hellinger distance estimator.

7.2. *3D Plots for Three Models.* To get better insight about the nature of variation in the MSE of the estimators, we present 3D plots of the MSE surface for three different models – binomial, Poisson and geometric. We use 2000 replications in each cases.

In the first example data are generated from a binomial $(15, \theta)$ distribution, so that

$$m_\theta(x) = \begin{cases} \binom{15}{\theta}\theta^x(1-\theta)^{15-x}, & \text{if } x = 0, 1, \ldots, 15, \\ 0, & \text{otherwise.} \end{cases}$$

We have taken different values of θ in $(0, 1)$. Figure 3 shows the 3D plot of the MSEs of the MPHDEs for different fixed values of h and θ. The sample size is 20.

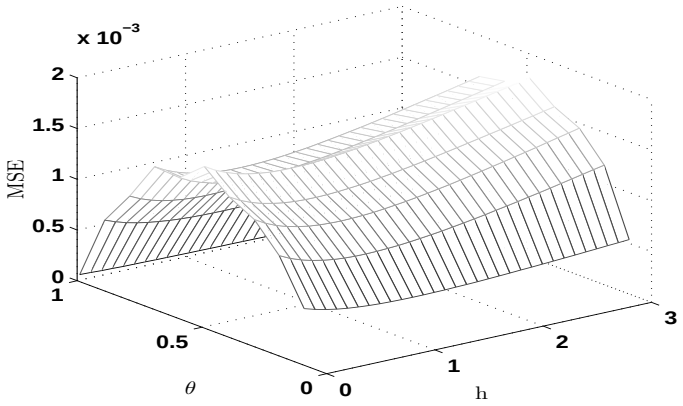

Figure 3: MSE of the MPHDEs using different values of h for the binomial model with parameters 15 and θ, where $n = 20$.

In Figure 4 we have plotted the MSEs in case of Poisson distribution where the values of mean parameter θ are $3, 4, \cdots, 10$. The sample size is $n = 20$ in each replication. Similarly in Figure 5 we have presented the MSE surface in case of geometric model with $\theta \in (0, 1)$ and sample size $n = 15$. In practically all the cases it appears that the optimum value of h is close to 0.5 or slightly higher when one takes a fixed h approach. The surfaces clearly slope downward at $h = 2$ for each of the cases considered. By the time the penalty weight slides down to 1, the gain is already substantial over the natural weight $h = 2$ in each case. The gain seems to be further enhanced at $h = 0.5$. In almost all of these cases the mean square error surface appears to reach a minimum at some point between 0.8 and 0.5 before curving upward again.

7.3. *The Score Functions.* We also look at the contributions of the different terms in the score equation to get an idea of how the estimating function changes with h. For this we look at the estimating equation of the penalized

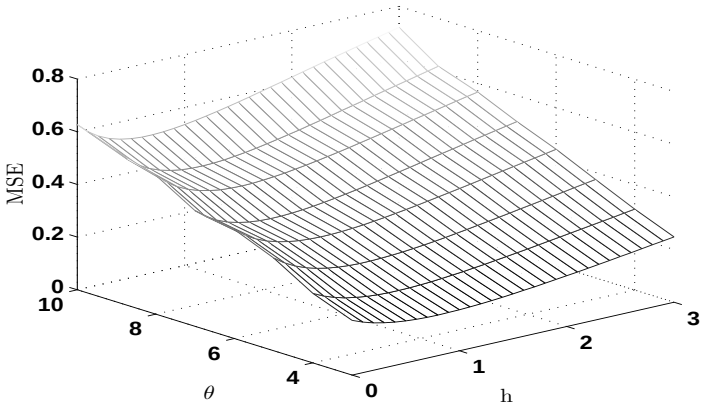

Figure 4: MSE of the MPHDEs using different values of h for the Poisson model with parameter θ, where $n = 20$.

estimator

$$\sum_{x:d_n(x)>0} A_G(\delta_{n\theta}(x))\nabla m_\theta(x) + h \sum_{x:d_n(x)=0} \nabla m_\theta(x) = 0 \qquad (7.4)$$

at the true parameter value $\theta = \theta^0$. Denoting

$$S_1 = \sum_{x:d_n(x)>0} A_G(\delta_{n\theta^0}(x))\nabla m_{\theta^0}(x), \text{ and } S_2 = \sum_{x:d_n(x)=0} \nabla m_{\theta^0}(x),$$

our aim is to identify the value of h for which the score function $S_1 - hS_2$ closest to zero.

We look at our three different models, and consider two sets of parameters. In each case we present the average values of the score functions and the individual terms S_1 and S_2 in a simulation of 1000 replications. The results are given in Tables 1 and 2. We present the average values (over the 1000 replications) of S_1, S_2 and corresponding average score for three disparities, including the ordinary Hellinger distance ($h = 2$) and the penalized Hellinger distances at $h = 1$ and 0.5. Since the average score $S_1 - hS_2$ vanishes at $h = S_1/S_2$, the 'ideal' h for each of the models, calculated according to this criterion is also presented in the last row of each table. Among the three values of h, $h = 0.5$ takes the average score closest to zero. In each of the above cases the ideal h lies between 0.6 and 0.75.

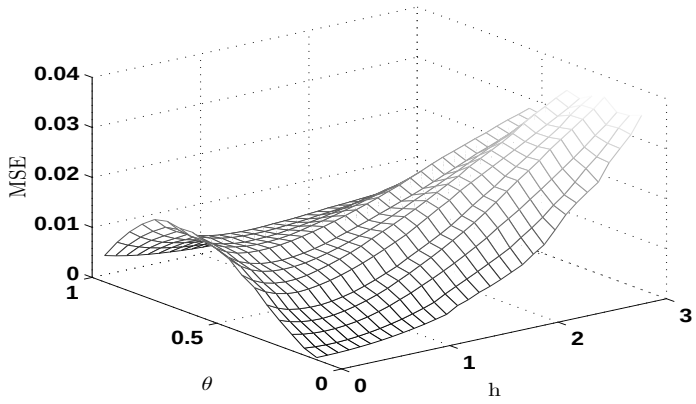

Figure 5: MSE of the MPHDEs using different values of h for the geometric model with parameter θ, where $n = 15$.

7.4. *The Role of the Sample Size n.* We conclude with a small simulation example demonstrating the behavior of the penalized estimators and the penalized disparity difference tests in the geometric model as a function of increasing sample size. Data are randomly generated from a geometric distribution with parameter $\theta = 0.5$. First we compare the empirical mean square errors of the ordinary MHDE, the MPHDEs with $h = 1$, and $h = 0.5$, and the MLE. Next we consider the hypothesis $H_0 : \theta = 0.5$ against $H_1 : \theta \neq 0.5$ and compare the observed levels of the disparity difference tests based on the Hellinger distance, the corresponding penalized tests with $h = 1$ and $h = 0.5$, and the likelihood ratio test (LRT) statistics. As the value of θ is taken to be 0.5 the Fisher information $I(\theta)$ in this case is $1/(0.5^2(1-0.5)) = 8$, and n times the asymptotic variance of the estimators are all the four equal to $1/8 = 0.125$.

The empirical mean square errors of each of these estimates are computed for 1000 replications at each sample size between 10 and 200. These observed MSEs (multiplied by n) are plotted as a function of the sample size, and presented in the same graph (Figure 6) for comparison. The straight line in the figure is the theoretical asymptotic variance of $\sqrt{n}$ times of the estimators. Clearly the MSEs of the MPHDEs are substantially closer to that of the maximum likelihood estimator compared to that of the MHDE. The penalty weight $h = 0.5$ again provides a slight improvement on $h = 1$.

Table 1: Score functions of the ordinary Hellinger distance (HD) and the penalized Hellinger distance (PHD) estimators, where sample size is 20.

	Geo(0.3)	Bin(20,0.3)	Poi(3)
S_1	−0.581	0.194	0.033
S_2	−0.931	0.290	0.048
HD Score: $S_1 - 2S_2$	1.281	−0.386	−0.063
PHD ($h = 1$) Score: $S_1 - S_2$	0.350	−0.096	−0.015
PHD ($h = 0.5$) Score: $S_1 - S_2/2$	−0.115	0.049	0.009
Ideal h	0.624	0.669	0.687

Table 2: Score functions of the ordinary Hellinger distance (HD) and the penalized Hellinger distance (PHD) estimators, where sample size is 15.

	Geo(0.5)	Bin(10,0.7)	Poi(5)
S_1	−0.319	−0.207	0.022
S_2	−0.524	−0.287	0.029
HD Score: $(S_1 - 2S_2)$	0.729	0.367	−0.036
PHD ($h = 1$) Score: $S_1 - S_2$	0.205	0.080	−0.007
PHD ($h = 0.5$) Score: $S_1 - S_2/2$	−0.057	−0.063	0.007
Ideal h	0.609	0.721	0.754

The most remarkable point in the graph is the extremely poor small sample performance of the ordinary MHDE. This estimator appears unsatisfactory, from the efficiency stand point, even at a sample size of 200. This deficiency appears to be almost completely eliminated by the proper choice of the penalty.

In Figure 7 the observed levels, computed as the proportion of test statistics exceeding the chi-square critical value, are plotted for the disparity difference test based on the ordinary Hellinger distance (the test in (6.1) with $\rho_G = \text{HD}$), the corresponding penalized tests in (6.2) with $h = 1$ and $h = 0.5$, and the likelihood ratio test. Once again it is observed that the levels of the penalized disparity difference tests are quite close to those of the LRT (and to the nominal value of 0.05) except at very small sample sizes, where $h = 0.5$ appears to produce a more conservative test. The observed level of the ordinary disparity difference test based on the Hellinger distance remains significantly higher than the nominal level even at a sample size of 200, while

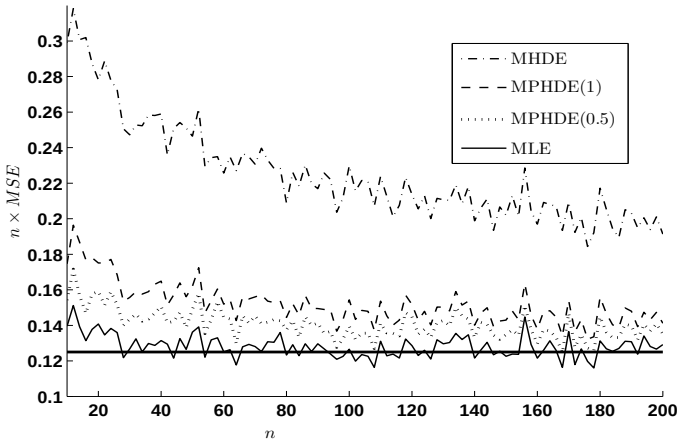

Figure 6: MSE (multiplied by n) of the MHDE, the MPHDE with $h = 1$, the MPHDE with $h = 0.5$ and the MLE.

such deficiencies are again almost entirely eliminated by appropriate choice of the penalty. The label DDT represents the ordinary disparity difference test based on the Hellinger distance in Figure 7, while the labels PDDT represent the indicated penalized disparity difference tests.

8. **Concluding Remarks**

In this paper we have established the asymptotic results about penalized minimum disparity methods, and provided a series of numerical investigations to get a good idea about the nature of improvement due to the application of the penalty. In most examples the optimal value of the penalty in the minimum Hellinger distance estimation turned out to be in the range 0.5 to 0.8. In general $h = 0.5$ appears to be a good choice for the minimum Hellinger distance estimator. In case of hypothesis testing problems, however, $h = 0.5$ appears to be a bit conservative for small samples, and $h = 1$ appears to do better in terms of the closeness of the achieved level with the nominal level.

We feel that this present work gives enough indications to suggest that further investigation about the use of penalized disparity methods in minimum disparity inference is well warranted. Also it is very likely that the

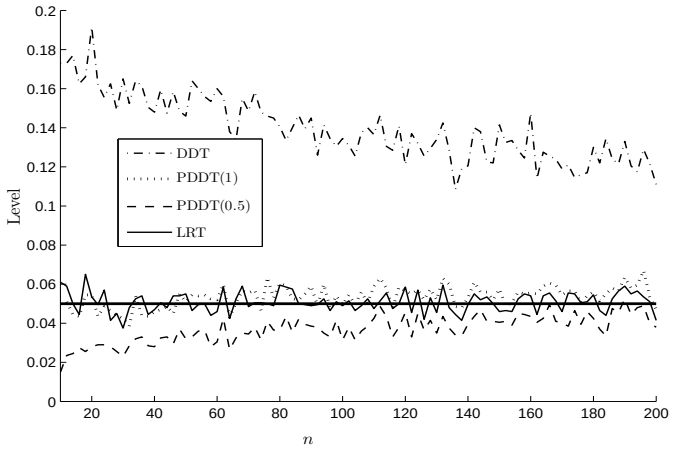

Figure 7: Observed levels of the DDT, the PDDT with $h = 1$, the PDDT with $h = 0.5$ and the likelihood ratio test (LRT) statistics. The nominal level is 0.05.

choice of the optimal penalty weight will be a function of the original disparity, which may be confirmed by future studies.

9. **Appendix**

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed observations from a discrete distribution modeled by the family having probability mass function $m_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $x \in \mathcal{X}$. Let $\theta^0 \in \Theta$, be the true value of the parameter. Denote

$$\begin{aligned} A_j^{(n)} =& \nabla_j \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0}, \\ B_{jk}^{(n)} =& \nabla_{jk} \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} \text{ and} \\ C_{jkl}^{(n)}(\theta) =& \nabla_{jkl} \ \rho_{G_h}(d_n, m_\theta). \end{aligned}$$

Then, under assumptions (A0)–(A5) as presented in the statement of Theorem 4.1, the following results hold.

LEMMA 9.1. $A_j^{(n)} = -\sum_x d_n(x) u_{j\theta^0}(x) + o_p\left(n^{-1/2}\right)$.

LEMMA 9.2. $B_{jk}^{(n)} = \sum_x m_{\theta^0}(x) u_{j\theta^0}(x) u_{k\theta^0}(x) + o_p(1)$.

LEMMA 9.3. *With probability tending to 1 there exists a finite number γ such that*

$$\left|C_{jkl}^{(n)}(\theta)\right| < \gamma,$$

for all j, k and l and for all $\theta \in \omega$, where ω is defined in (A1) of Theorem 4.1.

LEMMA 9.4. *Suppose $\hat{\theta}_n^h$ is the MPDE, which is a root of equation (4.2). Let us denote the function $\frac{\partial}{\partial\theta}\rho_{G_h}(d_n, m_\theta)$ by $f(d_n, \theta, h)$. Then, for the multinomial model*

$$\frac{\partial\hat{\theta}_n^h}{\partial h} = -\left(\frac{\frac{\partial}{\partial h}f(d_n,\theta,h)}{\frac{\partial}{\partial\theta}f(d_n,\theta,h)}\right)\Bigg|_{\theta=\hat{\theta}_n^h}.$$

PROOF OF LEMMA 9.1. Let

$$\begin{aligned} R_n(\theta) =& \rho_G(d_n, m_\theta) - \rho_{G_h}(d_n, m_\theta) \\ =& (G(-1) - h) \sum_{x:d_n(x)=0} m_\theta(x) \\ =& (G(-1) - h) \sum_x m_\theta(x) I(d_n(x)), \end{aligned} \tag{9.1}$$

where $I(y) = 1$ if $y = 0$ and 0 otherwise. We use the notation $R_{jn}(\theta) = \nabla_j R_n(\theta)$ and $m_{j\theta}(x) = \nabla_j m_\theta(x)$, where ∇_j represents the gradient with respect to θ_j. Then

$$\begin{aligned} R_{jn}(\theta) =& (G(-1) - h) \sum_x m_{j\theta}(x) I(d_n(x)) \\ =& (G(-1) - h) \sum_x m_\theta(x) u_{j\theta}(x) I(d_n(x)), \end{aligned}$$

where $u_{j\theta}(x) = \nabla_j \log m_\theta(x)$. So

$$\begin{aligned} & E\left[n^{1/2}|R_{jn}(\theta)|\right] \\ =& n^{1/2}|G(-1) - h| \sum_x |m_\theta(x) u_{j\theta}(x)| E\left[I(d_n(x))\right] \\ =& n^{1/2}|G(-1) - h| \sum_x |m_\theta(x) u_{j\theta}(x)| \{1 - m_\theta(x)\}^n \\ =& |G(-1) - h| \sum_x \left|m_\theta(x)^{1/2} u_{j\theta}(x)\right| \left[n^{1/2} m_\theta(x)^{1/2} \{1 - m_\theta(x)\}^n\right]. \end{aligned}$$

Suppose $g_n(x) = n^{1/2}x^{1/2}(1-x)^n$, where $0 < x < 1$. Note $g_n(x) \to 0$ for all $0 < x < 1$ as $n \to \infty$. Now

$$\max_{0<x<1} \quad g_n(x) = \frac{1}{\sqrt{2}} \left\{ \frac{2n}{2n+1} \right\}^{\frac{2n+1}{2}} \leq \frac{1}{\sqrt{2}}.$$

Therefore, using assumption (A3) it follows from dominated convergence (DCT) theorem that

$$E\left[n^{1/2}|R_{jn}(\theta)|\right] \to 0 \quad \text{as } n \to \infty.$$

Thus it follows from Markov's inequality that

$$R_{jn}(\theta) = o_p\left(n^{-1/2}\right). \tag{9.2}$$

Differentiating (9.1) with respect to θ_j we get

$$\nabla_j \ \rho_G(d_n, m_\theta) - \nabla_j \ \rho_{G_h}(d_n, m_\theta) = R_{jn}(\theta). \tag{9.3}$$

By substituting θ^0 for θ in equations (9.2) and (9.3) we get

$$\nabla_j \ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} = \nabla_j \ \rho_G(d_n, m_\theta)\Big|_{\theta=\theta^0} + o_p\left(n^{-1/2}\right). \tag{9.4}$$

Lindsay (1994) has shown that, under assumptions (A0)–(A5),

$$\nabla_j \ \rho_G(d_n, m_\theta)\Big|_{\theta=\theta^0} = -\sum_x d_n(x) u_{j\theta^0}(x) + o_p\left(n^{-1/2}\right). \tag{9.5}$$

Combining (9.4) and (9.5) the lemma is proved. □

Proof of Lemma 9.2. Here we consider the second order partial derivative of $R_n(\theta)$. From (9.1) we have

$$R_{jkn}(\theta) = \nabla_{jk} R_n(\theta) = (G(-1) - h) \sum_x m_{jk\theta}(x) I(d_n(x)),$$

where $m_{jk\theta}(x) = \nabla_{jk} m_\theta(x)$. Then

$$E\left[|R_{jkn}(\theta)|\right] = |G(-1) - h| \sum_x |m_{jk\theta}(x)| E\left[I(d_n(x))\right]$$

$$= |G(-1) - h| \sum_x |m_{jk\theta}(x)| \{1 - m_\theta(x)\}^n$$

$$=|G(-1)-h|\sum_x \Big|m_\theta(x)u_{j\theta}(x)u_{k\theta}(x)+m_\theta(x)u_{jk\theta}(x)\Big|\{1-m_\theta(x)\}^n.$$

Now $\{1-m_\theta(x)\}^n \to 0$ as $n\to\infty$. Again $\{1-m_\theta(x)\}^n \le 1$. Therefore, using assumption (A3) it follows from DCT that

$$E\left[|R_{jkn}(\theta)|\right] \to 0 \quad \text{as } n\to\infty.$$

Hence using Markov's inequality we can prove that

$$R_{jkn}(\theta) = o_p(1). \tag{9.6}$$

Differentiating (9.1) with respect to θ_j and θ_k we get

$$\nabla_{jk}\ \rho_G(d_n, m_\theta) - \nabla_{jk}\ \rho_{G_h}(d_n, m_\theta) = R_{jkn}(\theta). \tag{9.7}$$

By substituting θ^0 for θ in equations (9.6) and (9.7) we get

$$\nabla_{jk}\ \rho_{G_h}(d_n, m_\theta)\Big|_{\theta=\theta^0} = \nabla_{jk}\ \rho_G(d_n, m_\theta)\Big|_{\theta=\theta^0} + o_p(1). \tag{9.8}$$

Lindsay (1994) has shown that, under assumptions (A0)–(A5),

$$\nabla_{jk}\ \rho_G(d_n, m_\theta)\Big|_{\theta=\theta^0} = \sum_x m_{\theta^0}(x)u_{j\theta^0}(x)u_{k\theta^0}(x) + o_p(1). \tag{9.9}$$

Combining (9.8) and (9.9), the lemma is proved. □

PROOF OF LEMMA 9.3.

$$\begin{aligned}
&\nabla_{jkl}\ \rho_{G_h}(d_n, m_\theta)\\
=&-\sum_x \xi(d_n(x))\Big[A''_G(\delta_{n\theta})(1+\delta_{n\theta})^2 u_j u_k u_l m_\theta\\
&\quad - A'_G(\delta_{n\theta})(1+\delta_{n\theta})\{u_j u_{kl}+u_k u_{jl}+u_l u_{jk}+u_j u_k u_l\}m_\theta\\
&\quad + A_G(\delta_{n\theta})\{u_j u_{kl}+u_k u_{jl}+u_l u_{jk}+u_j u_k u_l+u_{jkl}\}m_\theta\Big],
\end{aligned}$$

where $\xi(d_n(x)) = 1-\left(1-\frac{h}{G(-1)}\right)I(d_n(x))$, and $I(y)=1$ if $y=0$ and zero otherwise. So

$$\begin{aligned}
&\nabla_{jkl}\ \rho_{G_h}(d_n, m_\theta)\\
=&-\sum_x \xi(d_n(x))\Big[A''_G(\delta_{n\theta})(1+\delta_{n\theta})^2 u_j u_k u_l m_\theta
\end{aligned}$$

$$- A'_G(\delta_{n\theta})(1+\delta_{n\theta})\{u_j u_{kl} + u_k u_{jl} + u_l u_{jk} + u_j u_k u_l\} m_\theta$$
$$+ \delta_{n\theta} A'_G(\delta^*)\{u_j u_{kl} + u_k u_{jl} + u_l u_{jk} + u_j u_k u_l + u_{jkl}\} m_\theta \Big],$$

where δ^* lies on the line segment joining 0 and $\delta_{n\theta}$. Let $c = \max\left\{1, \frac{h}{G(-1)}\right\}$. So $\xi(d_n) \le c$ for all d_n. From assumption (A5) we get

$$\begin{aligned}
&|\nabla_{jkl}\ \rho_{G_h}(d_n, m_\theta)| \\
\le & cN \sum_x \left|(1+\delta_{n\theta}) u_j u_k u_l m_\theta\right| \\
&+ cM \sum_x \left|(1+\delta_{n\theta})\{u_j u_{kl} + u_k u_{jl} + u_l u_{jk} + u_j u_k u_l\} m_\theta\right| \\
&+ cM \sum_x \left|\delta_{n\theta}\{u_j u_{kl} + u_k u_{jl} + u_l u_{jk} + u_j u_k u_l + u_{jkl}\} m_\theta\right|. \qquad (9.10)
\end{aligned}$$

Now applying the Central Limit Theorem and using (A4) it can be shown that each term in (9.10) is bounded in probability for all j, k and l and for all $\theta \in \omega$, where ω is defined in (A1). □

Proof of Lemma 9.4. The estimating equation is given by

$$f(d_n, \theta, h) = 0, \qquad (9.11)$$

which results in a functional $\theta = T(d_n, h)$. Hence for a r cell multinomial the differential of θ can be expressed as

$$d\theta = \sum_{u=1}^{r} \frac{\partial T}{\partial d_n(u)} dd_n(u) + \frac{\partial T}{\partial h} dh. \qquad (9.12)$$

Similarly from equation (9.11) we get

$$df = \sum_{u=1}^{r} \frac{\partial f}{\partial d_n(u)} dd_n(u) + \frac{\partial f}{\partial h} dh + \frac{\partial f}{\partial \theta} d\theta. \qquad (9.13)$$

Substituting $d\theta$ from (9.12) in equation (9.13) we get

$$df = \sum_{u=1}^{r} \left(\frac{\partial f}{\partial d_n(u)} + \frac{\partial f}{\partial \theta}\frac{\partial T}{\partial d_n(u)}\right) dd_n(u) + \left(\frac{\partial f}{\partial h} + \frac{\partial f}{\partial \theta}\frac{\partial T}{\partial h}\right) dh.$$

As $df = 0$, it must follow that the coefficient of dh is identically zero, i.e.,

$$\frac{\partial f}{\partial h} + \frac{\partial f}{\partial \theta}\frac{\partial T}{\partial h} = 0.$$

Hence

$$\frac{\partial\theta}{\partial h} = \frac{\partial T}{\partial h} = -\frac{\frac{\partial f}{\partial h}}{\frac{\partial f}{\partial \theta}}.$$

For the specific case where d_n and h are specified, we have

$$\frac{\partial\hat{\theta}_n^h}{\partial h} = \frac{\partial T(d_n,h)}{\partial h} = -\left(\frac{\frac{\partial f(d_n,\theta,h)}{\partial h}}{\frac{\partial f(d_n,\theta,h)}{\partial \theta}}\right)\Bigg|_{\theta=\hat{\theta}_n^h},$$

where $\hat{\theta}_n^h$ solves $f(d_n,\hat{\theta}_n^h,h)=0$. □

Acknowledgment. The authors gratefully acknowledge the comments of a referee which helped to significantly improve the presentation of the paper. This work of the third author was partially supported by Grant MTM2009–10072.

Reference

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

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

BASU, A. and HARRIS, I.R. and BASU, S. (1997). Minimum distance estimation: The approach using density-based distances. In *Handbook of Statistics*, **15**, 21–48. North-Holland, Amsterdam.

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

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

HARRIS, I.R. and BASU, A. (1994). Hellinger distance as a penalized log likelihood. *Comm. Statist. Simulation and Comput.*, **23**, 1097–1113.

KULLBACK, S. and LEIBLER, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79-86.

LEHMANN, E.L. (1983). *Theory of Point Estimation*. John Wiley & Sons Inc., New York.

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.

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

SEO, D. and LINDSAY B.G. (2009). A universally consistent modification of maximum likelihood. Preprint.

Serfling R.J. (1980). *Approximation Theorems of Mathematical statistics.* Wiley, New York.

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.

Tamura, R.N. and Boos, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.*, **81**, 223–229.

Abhijit Mandal
Applied Statistics Unit,
Indian Statistical Institute,
203 B.T. Road, Kolkata 700108,
India.
E-mail: abhi_r@isical.ac.in

Ayanendranath Basu
Bayesian and Interdisciplinary
Research Unit,
Indian Statistical Institute,
203 B.T. Road, Kolkata 700108,
India.
E-mail: ayanbasu@isical.ac.in

Leandro Pardo
Department of Statistics and O.R. I,
Complutense University of Madrid,
Plaza De Ciencias 3–28040, Madrid,
Spain.
E-mail: lpardo@mat.ucm.es

Paper received June 2008; revised November 2009.