Statistica Sinica **8**(1998), 841-860

PENALIZED MINIMUM DISPARITY METHODS FOR MULTINOMIAL MODELS

Ayanendranath Basu and Srabashi Basu

Indian Statistical Institute

Abstract: Robust estimation of the probability vector is an important problem for the finite k cell multinomial model. When the probability vector is unrestricted, its estimate is equal to the vector of the observed proportions for all minimum disparity estimators (Lindsay (1994)). But, when the probabilities are functions of a parameter $\boldsymbol{\theta}$ of dimension smaller than k, the estimates may differ significantly for different disparities. In particular, some procedures like the minimum Hellinger distance method may be substantially superior to the maximum likelihood estimator (MLE) or the minimum (Pearson's) chi-square estimator under systematic deviations from the model. All the minimum disparity estimators have optimal asymptotic efficiency properties. However, in many subclasses of disparities such as the Cressie-Read family more robust members of the class generally suffer a significant loss in small sample efficiency. In this paper we consider a correction which can lead to appreciable improvements in the small sample properties of the procedures, generally keeping their robustness properties intact. Exact results are presented for several multinomial models and a number of data examples are also considered.

Key words and phrases: Asymptotic efficiency, blended weight Hellinger disparity, Cressie-Read disparity, empty cells, residual adjustment function.

1. Introduction

Consider a random variable X having a multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$. Given a random observation from the distribution of X, resulting in a vector of sample proportions $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_k)$, we want to estimate $\boldsymbol{p}$ robustly.

In discrete models density-based divergences have been studied by, among others, Cressie and Read (1984), and Lindsay (1994). The former introduced a family of divergences indexed by a single parameter λ which includes many important density-based divergences such as the Pearson's and Neyman's chi-squares, the Hellinger distance and the Kullback-Leibler divergence. Cressie and Read used their family primarily for testing goodness-of-fit in discrete multivariate models. Lindsay considered a larger class of density-based divergences, called disparities, and studied the associated estimators and tests of hypotheses.

In discrete models all the minimum disparity estimators are first order efficient, but many have better robustness properties than the maximum likelihood estimator (Lindsay (1994)). Although the class of disparities is large and includes the Cressie-Read family, several members of the latter class remain as the more popular density-based robust alternatives to the maximum likelihood estimator. In particular, an appealing justification of robustness of the minimum Hellinger distance estimator has been provided by Simpson (1987).

A disparity is a nonnegative measure of discrepancy between two densities (see Section 2 for a formal definition) which assumes its minimum value zero when the densities are identical. In minimum disparity estimation one minimizes a disparity between the observed proportions and the model probability vector. Thus any minimum disparity estimator of $\boldsymbol{p}$ equals $\hat{\boldsymbol{p}}$ if the parameter space is the entire k-dimensional probability simplex S_k. Differences in the estimates will only be observed when the parameter space is a restricted subset of S_k. In this paper we consider multinomial models where the probability vector $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$ is a function of a q-dimensional parameter $\boldsymbol{\theta}$, $q < k$.

As mentioned before, all minimum disparity estimators, including those within the Cressie-Read family, are asymptotically first order efficient under the model. In addition several estimators in this family have desirable robustness properties which have been studied in various settings by several authors including Simpson (1987) and Lindsay (1994). Simpson (1989) and Lindsay (1994) also studied the use of minimized disparities in parametric hypothesis testing and discovered nice robustness properties of the corresponding tests. At the same time, however, many of the more robust estimators can be substantially poor (in terms of efficiency) compared to the maximum likelihood estimator when the sample size is small; this phenomenon is particularly noticeable for some of the more robust minimum disparity estimators within the Cressie-Read family. For the test based on the Hellinger distance, Simpson (1989) observed that in the Poisson model "it requires rather large sample sizes for the chi-squared approximation to be at all reasonable" (see Simpson (1989), Table 3).

This unfortunate trade-off between robustness and small sample efficiency appears to be partly due to the disproportionately large weight that these disparities put on the empty cells. In this paper we consider an empty cell penalty for the minimum disparity estimators in multinomial models which does not alter the asymptotic properties of the estimators but has been empirically observed to improve the small sample performances of some of these procedures (Harris and Basu (1994), Basu, Harris and Basu (1996)). We emphasize, however, that the purpose of this paper is not just to develop another class of robust procedures. The robustness of the minimized disparity procedures, as well as

their asymptotic efficiencies have already been studied in some detail (Simpson (1987, 1989), Lindsay (1994), see also Cressie and Read (1984), Read and Cressie (1988)). Since the penalized estimators considered are asymptotically equivalent to the ordinary estimators, here we focus entirely on the *small sample properties* of these estimators in multinomial models. In particular we demonstrate that the penalty can often significantly improve the small sample performance of the estimators without compromising their robustness properties. All the numbers and figures presented here correspond to *exact computations*, rather than Monte-Carlo results. The relevant quantities are calculated by enumerating all possible samples and determining their probabilities under the true distribution. Exact computations such as these can be extremely valuable in distinguishing between estimators whose large sample properties are identical. Such exact calculations have also been considered by Cressie and Read (1984), Basu and Sarkar (1994a), Basu and Basu (1995) and Shin, Basu and Sarkar (1995) in different contexts.

2. Disparity Based Inference and the Empty Cell Penalty

We begin by considering minimum disparity inference in its general setting. Let $f_\theta(x)$ be a parametric density defined on a countable set taken to be {1, 2, 3, ...} without loss of generality. We are interested in estimating the parameter $\boldsymbol{\theta}$ robustly. The parameter space Ω is a subset of R^q. Let $X_1, \ldots, X_n$ be a random sample from the distribution of $f_\theta(x)$ and $d(x), x = 1, 2, \ldots$ be the observed proportion of the value x among the n sample observations. For each x define $\delta(x) = d(x)/f_\theta(x) - 1$ to be the 'Pearson Residual' at x. For any strictly convex, thrice differentiable function $G(\cdot)$ defined on $[-1, \infty)$ with $G(0) = 0$, we will call $\rho_G(\boldsymbol{d},\ \boldsymbol{f}_\theta) = \sum_{x=1}^{\infty} G(\delta(x)) f_\theta(x)$ to be a disparity between $\boldsymbol{d} = (d(1), d(2), \ldots)$ and $\boldsymbol{f}_\theta = (f_\theta(1), f_\theta(2), \ldots)$. The properties of the function $G(\cdot)$ immediately suggest that $\rho_G(\boldsymbol{d}, \boldsymbol{f}_\theta)$ is non-negative and is equal to 0 if and only if $\boldsymbol{d} \equiv \boldsymbol{f}_\theta$. The estimates of $\boldsymbol{\theta}$ obtained by minimizing the members of the class of disparities are minimum disparity estimators. Equating the negative of $\partial \rho_G / \partial \boldsymbol{\theta}$ to zero, one gets the minimum disparity estimating equation as $\sum_{x=1}^{\infty} A(\delta(x)) \nabla f_\theta(x) = 0$, where ∇ represents the gradient with respect to $\boldsymbol{\theta}$, and $A(\delta) = G'(\delta)(\delta + 1) - G(\delta)$. The disparities can be redefined without changing their estimating properties so that $A(0) = 0$ and $A'(0) = 1$. In this form the function $A(\delta)$ is called the residual adjustment function of the disparity (or the corresponding estimator). Since the estimating equations are otherwise equivalent, the theoretical properties of the estimators are controlled by the form of the residual adjustment function. See Lindsay (1994) for more details.

The Cressie-Read family of disparities $I^{\lambda}(\boldsymbol{d}, \boldsymbol{f}_\theta)$ defined as

$$I^{\lambda}(\boldsymbol{d}, \boldsymbol{f}_\theta) = \frac{1}{\lambda(\lambda+1)} \sum_{x=1}^{\infty} d(x)\Big[(\frac{d(x)}{f_\theta(x)})^{\lambda} - 1\Big] \tag{2.1}$$

belongs to the class of disparities with $G(\delta) = [\lambda(\lambda+1)]^{-1}\{(\delta+1)^{\lambda+1}-1\}$. Harris and Basu (1997) have considered the Cressie-Read disparity in the form

$$I^{\lambda}_*(\boldsymbol{d}, \boldsymbol{f}_\theta) = \sum_{x=1}^{\infty} \Big[\frac{d(x)[(\frac{d(x)}{f_\theta(x)})^{\lambda} - 1]}{\lambda(\lambda+1)} + \frac{(f_\theta(x) - d(x))}{\lambda+1}\Big]. \tag{2.2}$$

As $\boldsymbol{d}$ and $\boldsymbol{f}_\theta$ are both densities, the second term in the right hand side of (2.2) does not contribute anything to the disparity, so that this definition is equivalent to the original definition of Cressie and Read given in (2.1). However, each term in the summand of (2.2) is non-negative, and the corresponding residual adjustment function automatically satisfies $A(0) = 0$ and $A'(0) = 1$. Note that for the disparity in (2.2), the $G(\delta)$ and $A(\delta)$ functions have the form

$$G(\delta) = \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}, A(\delta) = \frac{(\delta+1)^{\lambda+1}}{\lambda+1} - \frac{1}{\lambda+1}. \tag{2.3}$$

For the cases $\lambda = 0$ and $\lambda = -1$, the divergences have to be defined as the limiting cases as $\lambda \to 0$ and $\lambda \to -1$. When these limits are evaluated, one gets

$$I^{0}_*(\boldsymbol{d}, \boldsymbol{f}_\theta) = \sum_{x=1}^{\infty} \Big[d(x) \log \frac{d(x)}{f_\theta(x)} + (f_\theta(x) - d(x))\Big],$$

and I^{-1}_* can be obtained from I^{0}_* by interchanging $\boldsymbol{d}$ and $\boldsymbol{f}_\theta$. The minimizer of I^{0}_* is the maximum likelihood estimator of $\boldsymbol{\theta}$. We will call I^{0}_* the likelihood disparity. Similarly $I^{1}_*, I^{-1/2}_*$ and I^{-2}_*, given by

$$I^{1}_* = \sum_{x=1}^{\infty} \frac{[d(x) - f_\theta(x)]^2}{2f_\theta(x)},\ I^{-1/2}_* = 2\sum_{x=1}^{\infty} [d^{1/2}(x) - f^{1/2}_\theta(x)]^2,\ I^{-2}_* = \sum_{x=1}^{\infty} \frac{[d(x) - f_\theta(x)]^2}{2d(x)}$$

are the Pearson's chi-square, the squared Hellinger distance and the Neyman's chi-square respectively. Note that the versions of Pearson's and Neyman's chi-squares considered here have an extra factor of 1/2 so that the residual adjustment function has the right properties; likewise for the squared Hellinger distance.

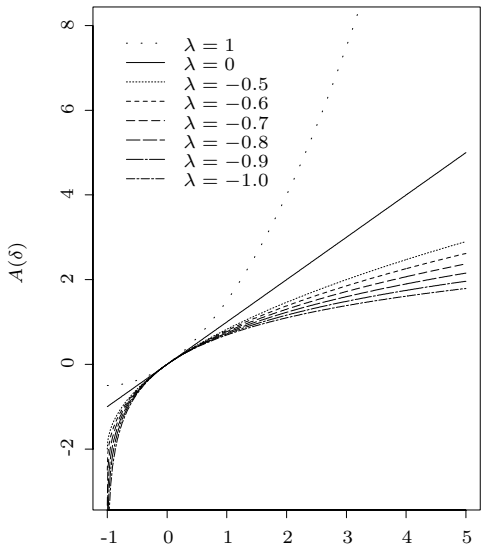

Figure 1. Residual adjustment functions of different disparities.

In Figure 1 we have plotted the residual adjustment functions of the Pearson's chi-square ($\lambda = 1$), the likelihood disparity ($\lambda = 0$), as well as of the disparities corresponding to $\lambda = -0.5, -0.6, -0.7, -0.8, -0.9, -1.0$. It may be observed that the residual adjustment functions of the disparities corresponding to the larger negative values of λ provide greater downweighting for observations with large positive values of δ. Such values of δ correspond to much larger observed frequencies than expected under the model. In this sense the residual adjustment function acts much like the ψ function in robust M-estimation. Geometrically, it can be seen that larger negative values of the curvature (measured by the second derivative) of the residual adjustment function at 0, lead to greater downweighting. Lindsay (1994) provides more theoretical justification to establish that larger values of the curvature in the negative magnitude lead to greater robustness. For the Cressie-Read family this curvature is equal to λ, so that larger negative values of λ lead to higher negative curvature. The Pearson's chi-square, on the other hand, *magnifies* the effect of large outlying values (i.e. has large *positive* curvature) and therefore may be expected to perform poorly in terms of robustness.

In this paper our interest is in comparing the robust minimum disparity estimators of the Cressie-Read family in the range $\lambda \in [-0.5, -1)$ with the traditional estimators such as minimum (Pearson's) chi-square estimator and the maximum likelihood estimator in multinomial models. Note that the Hellinger distance corresponds to $\lambda = -0.5$, so that the robust disparities considered here

provide a downweighting for the observations inconsistent with the model at a rate equal to or higher than the Hellinger distance. (For values of $\lambda \leq -1$ there are some practical problems as we discuss later.)

It should be emphasized that one cannot use the influence function approach for assessing the robustness of the above estimators; their influence functions are the same as that of the maximum likelihood estimator - otherwise the estimators cannot have full asymptotic efficiency. However there are several appealing robustness features in these estimators which we now briefly describe (a more detailed description of these features is provided in Lindsay (1994)). Firstly, there is an inherent dampened response to outliers, as measured by a second derivative generalization of the influence function idea. For estimators with large negative values of A_2, the quadratic approximation to the bias function of the estimator under contamination can be significantly smaller than the linear approximation (Lindsay, Section 4). Secondly, there is a high degree of stability of the disparity measures and the solutions to the estimating equations when outliers are added to the data. For example, when the data contain extreme outliers with small contaminating fractions, the Cressie-Read divergence from the contaminated data to the model is close to that obtained by simply deleting the outlier from the sample provided $\lambda < 0$ (Lindsay, Sections 6.2 and 6.3). And finally, under certain general conditions the minimum disparity estimators have asymptotic breakdown points of 50% (Lindsay, Section 6.4).

It has been noted before (Harris and Basu (1994), Basu and Sarkar (1994b), Basu, Harris and Basu (1996)) that for the model more robust minimum disparity estimators often fare more poorly than the maximum likelihood estimator if the sample size is small. Part of this behavior can be attributed to the manner in which the disparities treat the empty cells. Lindsay (1994) recognized that the treatment of the *Pearson inliers*, cells with *lower* observed frequency than expected under the model, can be the source of a problem for the minimum Hellinger distance estimator. Empty cells represent the extreme cases of inliers. To illustrate the empty cell problem in the case of the I_*^λ disparity let $\lambda > -1$, and write the disparity as the sum of two components with

$$I_*^\lambda(\boldsymbol{d}, \boldsymbol{f}_\theta) = \sum_{x:d(x)\neq 0} \Big\{ \frac{d(x)}{\lambda(\lambda+1)} \Big[\big(\frac{d(x)}{f_\theta(x)}\big)^\lambda - 1 \Big] + \frac{(f_\theta(x) - d(x))}{\lambda+1} \Big\} + \frac{1}{\lambda+1} \sum_{x:d(x)=0} f_\theta(x). \tag{2.4}$$

The second component in the above disparity, $(\lambda+1)^{-1} \sum_{x:d(x)=0} f_\theta(x)$, is the contribution of the empty cells, and can become very large for values of λ close

to -1. This can also be seen from Figure 1, and equation (2.3). There is a sharp decrease in the left tail of the graphs (note that the empty cells correspond to $\delta = -1$) if λ has a large negative value. To counter this problem we can alternatively consider the *penalized* family of disparities

$$P^{\lambda}(\boldsymbol{d}, \boldsymbol{f}_\theta) = \sum_{x:d(x)\neq 0} \Big\{ \frac{d(x)}{\lambda(\lambda+1)} \Big[(\frac{d(x)}{f_\theta(x)})^{\lambda} - 1 \Big] + \frac{(f_\theta(x) - d(x))}{\lambda+1} \Big\} + \sum_{x:d(x)=0} f_\theta(x). \tag{2.5}$$

The above is obtained from (2.4) by applying a penalty for the empty cells; this penalty changes the weight of the empty cells from $(\lambda+1)^{-1}$ to 1. For all λ, therefore, the penalized disparities put the same weight on the empty cells as $I^0_*(d, f_\theta)$ does. As the number of empty cells asymptotically goes to zero, this penalty does not affect the asymptotic distribution of the estimators; the downweighting properties of the disparities also remain intact. In Section 3 we compare the minimum disparity estimators obtained through the penalized version P^{λ} with those estimators that minimize I^{λ}_*.

For $\lambda = -1$, $A(-1) = -\infty$, and the disparity is not defined if there is a single empty cell. This is true for all values of $\lambda \leq -1$ and this makes it impossible to do the exact computations that we have considered in this paper for such disparities. The representation (2.4) of the Cressie-Read disparities is valid only for $\lambda > -1$. However, the penalized Cressie-Read disparity given in (2.5) is well defined for all values of λ irrespective of the number of empty cells.

Next we turn our attention to hypothesis testing problems using penalized disparities. Consider the simple null hypothesis $H_0 : \boldsymbol{\theta} = \theta_0$, and define the statistic $T^{\lambda} = 2n[I^{\lambda}_*(\boldsymbol{d}, \boldsymbol{f}_{\theta_0}) - I^{\lambda}_*(\boldsymbol{d}, \boldsymbol{f}_{\hat{\theta}})]$, where $\hat{\boldsymbol{\theta}}$ represents the minimizer of I^{λ}_* (with the disparities at $\lambda = 0, -1$ being defined in the usual limiting sense). It may be checked easily that T^0 equals the negative of twice the log likelihood ratio; it is well known that this has an asymptotic $\chi^2(q)$ distribution under the null hypothesis (e.g. Serfling (1980)). Simpson (1989) showed that the 'Hellinger deviance test statistic' $T^{-1/2}$ is asymptotically equivalent to T^0 under the null hypothesis. Lindsay (1994) went one step further, and proved this asymptotic equivalence for all disparities under some general conditions. Consequently, all the T^{λ} statistics are asymptotically $\chi^2(q)$ under the null for $\lambda > -1$. As in the estimation problem we restrict the parameter λ to be in this range since otherwise the disparity is not defined if there are any empty cells, and the moments of the test statistics do not exist.

The T^{λ} statistics corresponding to larger negative values of λ can generally perform much better than the likelihood ratio statistic T^0 or the Pearson's chi-square based statistic T^1 in keeping the level and the power of the tests stable

under contamination. In particular the Hellinger deviance test $T^{-1/2}$ has been studied by several authors (Simpson (1989), Lindsay (1994), Basu, Harris and Basu (1996)) which demonstrate the desirable robustness properties of this test. For small samples, the chi-square approximation for this test statistic under the null hypothesis, however, can be quite poor, with the observed levels being considerably inflated compared to the nominal levels; consequently, the confidence intervals obtained by inverting the test statistic also have lower confidence coefficients than the nominal one.

Here we discuss an alternative test statistic based on the penalized disparities. Define the penalized family of test statistics

$$T_p^\lambda = 2n[P^\lambda(\boldsymbol{d}, \boldsymbol{f}_{\theta_0}) - P^\lambda(\boldsymbol{d}, \boldsymbol{f}_{\hat{\theta}})],$$

where $\hat{\boldsymbol{\theta}}$ represents the minimizer of P^λ. As they differ only in the empty cells, the families T^λ and T_p^λ have the same asymptotic distribution under the null hypothesis, and the same asymptotic breakdown properties.

The testing procedures described in this section extend to the case of the composite null hypothesis using the techniques of Serfling (1980). The tests based on T^λ, T_p^λ again have the same asymptotic distribution as the likelihood ratio test under the null hypothesis.

In the following section we present several exact computations for disparity based methods in the multinomial model. A random sample of n categorical observations on k categories with probabilities $p_1, \ldots, p_k$ generates a multinomial observation X with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$. For the rest of the paper $\hat{\boldsymbol{p}}$ will replace $\boldsymbol{d}$, the vector of observed proportions, and $\boldsymbol{p}(\boldsymbol{\theta})$ will replace the probability function $\boldsymbol{f}_\theta$.

3. Numerical Studies and Data Examples

3.1. Exact computations

Consider a multinomial random variable $\boldsymbol{X}$ with $n = 20$ and $k = 4$; suppose that the probability vector $\boldsymbol{p} = (p_1, p_2, p_3, p_4)$ is the function of a single parameter θ. In order to avoid overly intensive computation one has to choose a suitable small value of k; in this paper we have chosen k to be equal to 4 for our exact computations. To obtain the exact probability distribution of an estimator $\hat{\theta}$ one can enumerate all possible sample combinations in the sample space D given by $D = \{(x_1, x_2, x_3, x_4) : \sum x_i = 20$, each x_i is an integer between 0 and 20 $\}$; the distinct values of the parameter estimate and their exact probabilities can then be calculated using the multinomial probability function under any given true value of θ.

In this paper, we have considered two particular structures on the multinomial cell probabilities. For the first case we assumed that the cell probabilities are generated by a $Poisson(\theta)$ distribution - i.e. $p_1 = \exp(-\theta), p_2 = \theta\exp(-\theta), p_3 = [\theta^2\exp(-\theta)]/2$ and $p_4 = (1 - p_1 - p_2 - p_3)$. (For the general k cell multinomial this structure can be generalized by generating $p_1, \ldots, p_{k-1}$ using the *Poisson* probability function, and letting $p_k = 1 - \sum_{i=1}^{k-1} p_i$.) The other structure is generated by a $geometric(\theta)$ distribution - i.e. $p_1 = \theta, p_2 = \theta(1-\theta), p_3 = \theta(1-\theta)^2$ and $p_4 = (1-\theta)^3$. As in the *Poisson* case, this can again be generalized to the k cell multinomial. We will refer to these structures as the *Poisson* model and the *geometric* model respectively. The reason for using these models is that they are the two most common count data models.

For each possible sample point $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$, and for each value of λ considered, we calculate two estimates of θ by minimizing the disparity I_*^λ and the penalized disparity P^λ; let these be denoted by $\hat{\theta}_I$ and $\hat{\theta}_P$ respectively. For each estimator $\hat{\theta}$ (which takes the value $\hat{\theta}(\boldsymbol{x})$ under the sample $\boldsymbol{x}$) and for $\lambda = 1, 0, -0.5, -0.6, -0.7, -0.8$ and -0.9, we compute the *exact* mean square error (MSE) of $\hat{\theta}$ under the true value θ as $\sum(\hat{\theta}(\boldsymbol{x}) - \theta)^2 P_\theta(\boldsymbol{x})$, where the sum is over the sample space D, and $P_\theta(\boldsymbol{x})$ is the probability of the sample $\boldsymbol{x}$ under the cell probability vector $\boldsymbol{p}$.

Contamination is introduced in the *Poisson* model by defining $\boldsymbol{p}^* = (p_1^*, p_2^*, p_3^*, p_4^*)$, $p_i^* = (1-\epsilon)p_i$, $i = 1, 2, 3$, $p_4^* = (1-\epsilon)p_4 + \epsilon$, where $\boldsymbol{p} = (p_1, \ldots, p_4)$ represent the probabilities under a target value θ of the parameter. The exact mean square is then computed around the target value as before, but the probabilities $P_\theta(\boldsymbol{x})$ are now calculated under the cell probability vector $\boldsymbol{p}^*$. In the *geometric* model, the cell probabilities are redefined as $\boldsymbol{p}^* = (p_1^*, p_2^*, p_3^*, p_4^*)$, $p_1^* = (1-\epsilon)p_1 + \epsilon$, $p_i^* = (1-\epsilon)p_i$, $i = 2, 3, 4$. A similar procedure is then adopted to calculate the exact MSE under this model. For the *Poisson* model the parameter space of θ was restricted to [0, 5] and for the *geometric* model the parameter space was [0, 1] to make them compact.

The results comparing the performances of $\hat{\theta}_I$ and $\hat{\theta}_P$ for different values of λ under the *Poisson* model are presented in Table 1, where the true multinomial cell frequencies are generated by the $Poisson(0.5)$ distribution. When the model holds, (i.e. when $\epsilon = 0$), it can be observed that the MSE corresponding to the ordinary minimum disparity estimators with large negative values of λ are very high compared to likelihood disparity and the Pearson's chi-square. However, when we consider the penalized disparities, the obtained MSEs for the robust minimum disparity estimators are extremely competitive with the cases $\lambda = 1$ and $\lambda = 0$. Since the penalized procedures replace the weights of the empty cells by that of I_*^0 the estimates minimizing I_*^0 and P^0 are identical.

Under moderate contaminations ($\epsilon = 10\%$) the performance of MLE is worse than the estimates corresponding to large negative values of λ (see the fourth and fifth columns of Table 1). In this case the MSEs for the penalized estimators are competitive with or are better than the ordinary estimators - suggesting that the robustness property is not compromised by the empty cell penalty. It appears, therefore, that the penalized disparity procedure is a judicious choice from both efficiency and robustness standpoints in these cases. The findings under a *geometric* model where the true probabilities are generated by the *geometric*(0.1) distribution are very similar to those of Table 1, and are not presented here for brevity.

Table 1. Exact MSE for estimation of θ by minimum Cressie-Read disparities and their penalized versions. Multinomial proportions are generated by the *Poisson*(0.5) distribution with and without contamination (contaminating proportion is ϵ). MSE($\hat{\theta}$) represents the mean square error of the estimator $\hat{\theta}$.

	$\epsilon = 0.0$		$\epsilon = 0.1$	
λ	MSE($\hat{\theta}_I$)	MSE($\hat{\theta}_P$)	MSE($\hat{\theta}_I$)	MSE($\hat{\theta}_P$)
1	2.9999×10^{-2}	2.7353×10^{-2}	0.190132	0.182801
0	2.5267×10^{-2}	2.5267×10^{-2}	0.138909	0.138909
-0.5	2.8077×10^{-2}	2.5872×10^{-2}	0.100597	0.107853
-0.6	3.0385×10^{-2}	2.6155×10^{-2}	0.097386	0.102089
-0.7	3.4238×10^{-2}	2.6477×10^{-2}	0.097371	0.096688
-0.8	4.1254×10^{-2}	2.6780×10^{-2}	0.100372	0.091990
-0.9	5.7015×10^{-2}	2.7092×10^{-2}	0.108290	0.087574

Next we look at the performances of the statistics T^λ and T_P^λ in testing the null hypothesis $H_0 : \theta = \theta_0$ under the model i.e. when the probability vector is actually generated by the parameter θ_0, and under contamination, i.e. when the probability vector $\boldsymbol{p}^*$ is obtained by contaminating the vector generated by θ_0 in the manner described above. We specify θ_0 to be 0.5 for the *Poisson* model and 0.1 for the *geometric* model. The results for the *Poisson* model are presented in Table 2. We calculate the exact probability of the test statistic to exceed the 10%, 5% and the 1% critical points of the $\chi^2(1)$ distribution under true model and under contamination. The following results deserve mention. Under the model true levels are considerably inflated for the ordinary robust disparities compared to the nominal levels. However, under the same conditions, the chi-square approximation seems to work much better for the penalized disparity statistic T_P^λ. Under contamination, the penalized statistics corresponding to very large negative values of λ present themselves as the better choices. Similar results, not presented here, were obtained for the *geometric* model.

Table 2. Observed levels for testing $H_0 : \theta = 0.5$ using the statistics T^λ and T_P^λ. Multinomial proportions are generated by the $Poisson(0.5)$ distribution with and without contamination (contaminating proportion is ϵ).

		$\epsilon = 0.0$			$\epsilon = 0.1$		
λ	Statistic	10%	5%	1%	10%	5%	1%
1	T^λ	0.140744	0.089645	0.034734	0.727659	0.702272	0.583755
	T_P^λ	0.159680	0.093932	0.033016	0.726892	0.697912	0.572514
0	T^λ	0.112588	0.053842	0.015813	0.448896	0.356221	0.212406
	T_P^λ	0.112588	0.053842	0.015813	0.448896	0.356221	0.212406
−0.5	T^λ	0.165402	0.103695	0.028159	0.283702	0.203426	0.079067
	T_P^λ	0.112687	0.048056	0.013241	0.285249	0.193096	0.076950
−0.6	T^λ	0.196754	0.145951	0.053768	0.282037	0.198189	0.072357
	T_P^λ	0.111387	0.050883	0.013852	0.253021	0.167504	0.061999
−0.7	T^λ	0.268835	0.207214	0.092556	0.341149	0.236601	0.084997
	T_P^λ	0.110517	0.051159	0.015752	0.232652	0.149522	0.052900
−0.8	T^λ	0.303311	0.246025	0.195763	0.421079	0.326488	0.191351
	T_P^λ	0.110914	0.050664	0.016068	0.221235	0.132459	0.047979
−0.9	T^λ	0.529700	0.370792	0.240500	0.442919	0.366155	0.284735
	T_P^λ	0.110354	0.057566	0.016706	0.197385	0.125525	0.043968

The performance of the Pearson's chi-square statistics is much worse than even the likelihood based statistics under contamination (see Tables 1-2). This is not surprising given the manner in which the statistic treats the large outliers (see Figure 1). Under the model the performance of these statistics is slightly worse than the likelihood based procedures. Note that for the Pearson's chi-square the weight of the empty cell is actually smaller than that of the likelihood disparity. The empty cell correction given in (2.5), therefore, actually increases the weight of the empty cells and is not necessarily expected to improve the performance of the method.

To better understand the improvement in the performance of the test statistics due to the penalty we looked at the histograms of the exact null distribution of the test statistics T^λ and T_P^λ with the $\chi^2(1)$ density superimposed. The null hypothesis considered is $H_0 : \theta = 0.1$ under the geometric model. Cell frequencies are generated by the $geometric(0.1)$ distribution with $n = 20$ and $k = 4$. The height of each bar represents the exact probability for the test statistic to lie between the respective end points. In particular we looked at the histograms of T^0, $T^{-3/4}$ and $T_P^{-3/4}$. Our interest is in the right hand tail area of the histograms, and how well the $\chi^2(1)$ density approximates it. Figure 2 shows the

histogram of $T^{-3/4}$, where the poor approximation to the relatively heavy tail of the statistic provided by the $\chi^2(1)$ density is evident. The vertical dashed line on the histogram corresponds to the 5% critical point of $\chi^2(1)$. On the other hand the right tails of the histograms of T^0 and $T_p^{-3/4}$ (not presented here for brevity) around and beyond the 5% critical point are very well approximated by the overlaid density, leading to very high agreement in the observed and nominal levels. The observations were similar for the 10% and 1% critical values. We also investigated the histograms for other values of λ in the $[-0.5, -1)$ range for this case. At each instance the T^λ statistic was poorly approximated by the $\chi^2(1)$ limit, with the approximation getting worse as λ approached -1. The distributions of all the penalized statistics were much improved and very close to that of the statistic T^0.

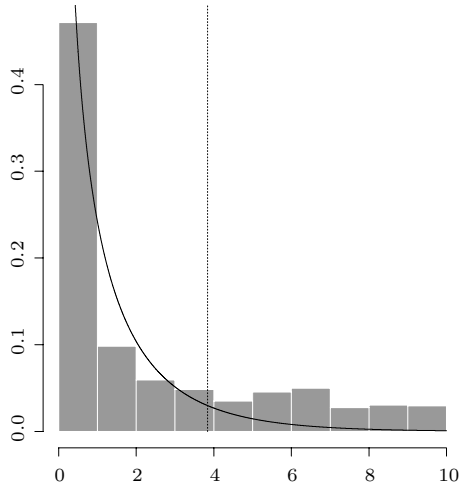

Figure 2. Density function of $\chi^2(1)$ distribution superimposed on the histogram of the null distribution of $T^{-0.75}$ for testing $H_0 : \theta = 0.1$ under the *geometric* model. Vertical dotted line denotes 5% critical point of $\chi^2(1)$ distribution.

3.2. Examples

We applied these procedures to several sets of real data. In the first example, the penalized estimation method is applied on chemical mutagenecity data previously analyzed by Simpson (1987) in the context of minimum Hellinger distance estimation. In the sex linked recessive lethal test in drosophila (fruit flies), male flies are exposed to different doses of a chemical to be screened. They are

then mated with unexposed females and the number of daughter flies carrying a recessive lethal mutation on the X choromosome is noted; the results of one such experiment is presented in Table 3, where the observed numbers of frequencies are recorded. Note that there is a very large outlier (having value 91) in the data. Simpson considered a *Poisson* fit for this data. We have considered a *Poisson* model under the k cell multinomial (with $k = 20$ and 30) for this data. In each case the large outlier was classified in the kth cell. These values of k were chosen so that we could study the effect of the large outlier in a cell widely separated from the rest of the data; two values of k were chosen to exhibit that moving the large outlier further away fails to corrupt the robust estimates.

Table 3. Observed distribution of the number of daughters in drosophila fruit flies carrying a recessive lethal mutation on the X chromosome. Expected frequencies (I), (II) and (III) represent the minimum I_*^0, $I_*^{-0.9}$ and $P^{-0.9}$ fits respectively. ($k = 20$).

No. of daughters	0	1	2	3	4	≥ 5
Observed Frequency	23	7	3	0	0	1(91)
Expected Frequency (I)	13.2	12.5	5.9	1.8	0.4	0.2
Expected Frequency (II)	25.8	7.1	1.0	0.1	0.0	0.0
Expected Frequency (III)	23.7	8.6	1.6	0.2	0.0	0.0

For $k = 20$, the outlier in the last cell of the multinomial exerts a heavy influence on the MLE as well as the minimum Pearson's chi-square estimate of θ (see Table 4). But the other estimates of θ obtained by minimizing $I_*^\lambda, \lambda = -0.5, -0.6, -0.7, -0.8, -0.9$ have been able to ignore this large value successfully. This phenomenon is even more apparent in the 30 cell multinomial. However, an interesting thing to note is that the penalized estimators behave much more uniformly than the ordinary minimum disparity estimators. Over the entire range $\lambda \in [-0.5, -1)$, the difference between the penalized estimators appears to be marginal, unlike the ordinary minimum disparity estimators. For both $k = 20$ and $k = 30$, the MLE of θ after the removal of the large outlier was 0.3939. In this case, therefore, both sets of robust estimators successfully ignore the outlier, but the penalized estimators succeed in keeping the estimates much closer to the outlier deleted MLE. In Table 3 we have also provided the expected frequencies for the 20 cell multinomial (the sixth cell representing the indicated '$\geq$' frequency) under three different estimates: the MLE (I), the ordinary minimum disparity estimator with $\lambda = -0.9$ (II), and the penalized minimum disparity estimator with $\lambda = -0.9$ (III). Both the robust methods give improved fits over the MLE, but the penalized estimator appears to fit the data better.

Table 4. Estimates of the parameter θ for the k cell multinomial under the $Poisson(\theta)$ model for the drosophila data.

	$k = 20$		$k = 30$	
λ	$\hat{\theta}_I$	$\hat{\theta}_P$	$\hat{\theta}_I$	$\hat{\theta}_P$
1	5.8482	9.3129	9.5175	13.2565
0	0.9424	0.9424	1.2360	1.2360
−0.5	0.3637	0.3774	0.3637	0.3774
−0.6	0.3532	0.3738	0.3532	0.3738
−0.7	0.3390	0.3700	0.3390	0.3700
−0.8	0.3173	0.3661	0.3173	0.3661
−0.9	0.2763	0.3621	0.2763	0.3621

The next data set on the incidence of peritonitis on 390 kidney patients (Table 5) was provided by Professor P. W. M. John (personal communication). A visual inspection suggests that a *geometric* distribution with θ around 0.5 may model the data well. While the largest observed frequency is 12, we chose a *geometric* model under $k = 20$ so that the observed sample has several empty cells. There are a few moderately large values in the data, but there are no extreme outliers (note that the sample size 390 is fairly large). In this case the estimates (not presented here for brevity) do not show any dramatic outlier effect. However the penalized estimates are much closer to the MLE than the ordinary ones. This is because the data presents several empty cells under the 20 cell multinomial, despite the large sample size. This improvement due to the penalty can be noted in Table 5 also, where we have provided the expected frequencies for the same three different methods as in Table 3. In this case the MLE fits the data well, and the fit provided by the penalized minimum disparity estimator appears to be better than the ordinary minimum disparity estimator.

The third data set has been analyzed by Rao (1973), pp 371-374. Every human being may be classified into one of four blood groups O, A, B and AB. The inheritance of these is controlled by one of three genes O, A and B, of which O is recessive to A and B. If p and q are the relative frequencies of the blood groups A and B, and the relative frequency of O is given by $r = 1 - p - q$, then the expected probabilities of the four groups in random mating are given by $\Pr(O) = r^2, \Pr(A) = p^2 + 2pr, \Pr(B) = q^2 + 2qr$ and $\Pr(AB) = 2pq$. The sample frequencies of the blood groups O, A, B, and AB, based on 435 observations are 17, 176, 182 and 60 respectively. We let $\boldsymbol{\theta} = (p, q)$ and find the minimum disparity estimate of $\boldsymbol{\theta}$. The results (not presented here) are extremely close, showing that for large sample sizes there is a very good agreement between different methods. Since there are no empty cells, $\hat{\boldsymbol{\theta}}_I$ and $\hat{\boldsymbol{\theta}}_P$ are the same in each case. The expected frequencies for the MLE and the minimum disparity estimator with $\lambda = -0.9$ provide almost identical fits (not presented here).

Table 5. Observed distribution of the number of cases of peritonitis for each of 390 kidney patients. Expected frequencies (I), (II) and (III) represent the minimum I_*^0, $I_*^{-0.9}$ and $P^{-0.9}$ fits respectively.

No. of cases	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed Frequency	199	94	46	23	17	4	4	1	0	0	1	0	1
Expected Frequency (I)	193.5	97.5	49.1	24.7	12.5	6.3	3.2	1.6	0.8	0.4	0.2	0.1	0.1
Expected Frequency (II)	212.3	96.7	44.1	20.1	9.1	4.2	1.9	0.9	0.4	0.2	0.1	0.0	0.0
Expected Frequency (III)	198.2	97.5	47.9	23.6	11.6	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.0

4. An Alternative Family

4.1. The blended weight Hellinger disparity

As we have discussed, the Cressie-Read family represents a very rich subclass of disparities. In the present paper we have concentrated on this subfamily keeping in mind their wide familiarity relative to some other subclasses of disparities, as well as the fact that they contain most of the well known density-based divergences. For $\lambda \leq -1$, however, the popularity of the Cressie-Read family is tempered by the fact that the disparities are not defined if there is even one empty cell. In this section we briefly discuss another subclass of disparities which is lesser known than the Cressie-Read family, although some properties have been studied in limited set ups (Basu and Sarkar (1994b); Shin, Basu and Sarkar (1995)). This is the family of blended weight Hellinger disparities, and is defined as (in the notation of Section 2)

$$BWHD^{\alpha}(\boldsymbol{d}, \boldsymbol{f}_\theta) = \frac{1}{2}\sum_{x=1}^{\infty}\Big(\frac{(d(x) - f_\theta(x))}{\alpha d^{1/2}(x) + (1-\alpha) f_\theta^{1/2}(x)}\Big)^2, -\infty < \alpha < \infty. \quad (4.1)$$

This family does not have the above empty cell limitation of the Cressie-Read family (except for the case $\alpha = 1$) and generates the Pearson's chi-square, the Hellinger distance, and the Neyman's chi-square for $\alpha = 0, 1/2$ and 1 respectively. Here we present a collection of results which show that for each disparity in the Cressie-Read family, there is a corresponding member within the $BWHD^{\alpha}$ class which is extremely close to it - so that even when the disparities in the Cressie- Read families are incomputable there are other alternatives with similar properties. (Of course one can also use the $BWHD$ when the corresponding Cressie-Read disparities are defined as well.)

Consider the Cressie-Read divergence I^λ under the multinomial set up of Section 3, and let $W_i = n^{1/2}(\hat{p}_i - p_i(\boldsymbol{\theta}))$. Expanding $2nI^\lambda(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta}))$ in a Taylor series one gets

$$2nI^\lambda(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})) = \sum_{i=1}^{k} \frac{W_i^2}{p_i(\boldsymbol{\theta})} + \frac{\lambda - 1}{3n^{1/2}} \sum_{i=1}^{k} \frac{W_i^3}{p_i^2(\boldsymbol{\theta})} + \frac{(\lambda-2)(\lambda-1)}{12n} \sum_{i=1}^{k} \frac{W_i^4}{p_i^3(\boldsymbol{\theta})} + O_p(n^{-3/2}). \tag{4.2}$$

(See Read and Cressie (1988), p. 176 for the derivation.) A similar expansion for the $BWHD^\alpha$, derived by Basu and Sarkar (1994a) has the form

$$2nBWHD^\alpha(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})) = \sum_{i=1}^{k} \frac{W_i^2}{p_i(\boldsymbol{\theta})} - \frac{\alpha}{n^{1/2}} \sum_{i=1}^{k} \frac{W_i^3}{p_i^2(\theta)} + \frac{3\alpha^2 + \alpha}{4n} \sum_{i=1}^{k} \frac{W_i^4}{p_i^3(\boldsymbol{\theta})} + O_p(n^{-3/2}). \tag{4.3}$$

Comparing (4.2) and (4.3) it can be seen that for any given θ and $\hat{\boldsymbol{p}}$, the Cressie-Read disparity with λ is exactly equivalent (upto $O_p(n^{-3/2})$ terms) to the blended weight Hellinger disparity with $\alpha = (1 - \lambda)/3$. (Also see Shin, Basu and Sarkar (1995)).

In the context of multivariate goodness-of-fit tests, Cressie and Read have derived the first three moments of the test statistic $2nI^\lambda(\hat{\boldsymbol{p}}, \boldsymbol{p}_0)$ under the null hypothesis $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$, where $\boldsymbol{p}_0$ is completely specified. (Note that they derive the moments only for the range $\lambda > -1$, since the moments do not exist otherwise). Since their moment calculations are done excluding the $O_p(n^{-3/2})$ term in equation (4.2), a corresponding calculation using the expression in (4.3) gives the moments for the $BWHD$. For each $\lambda > -1$, the corresponding member in the $BWHD$ family (with $\alpha = (1 - \lambda)/3$), has the same first three moments (excluding the $O_p(n^{-3/2})$ terms). Thus, the Cressie-Read family I^λ may often be well approximated by $BWHD^{\alpha=(1-\lambda)/3}$. Unlike I^λ however, empty cells alone cannot make the $BWHD^{\alpha=(1-\lambda)/3}$ undefined except for the case $\alpha = 1$.

To visually illustrate this equivalence we have plotted, on the same graph, the residual adjustment functions of the Cressie-Read disparities corresponding to λ, and those for the blended weight Hellinger disparities with $\alpha = (1 - \lambda)/3$. In Figures 3(a) - (d), we represent this for four different combinations of (λ, α) values which we chose to be (0, 1/3), (−0.2, 0.4), (−0.6, 8/15) and (−0.8, 0.6). In all the four cases the correspondence between the functions can be observed to be extremely close (except in the extreme left tail of the plot 3(d), where it can be seen that the Cressie-Read family puts a larger weight on the empty cell relative to the corresponding $BWHD^\alpha$).

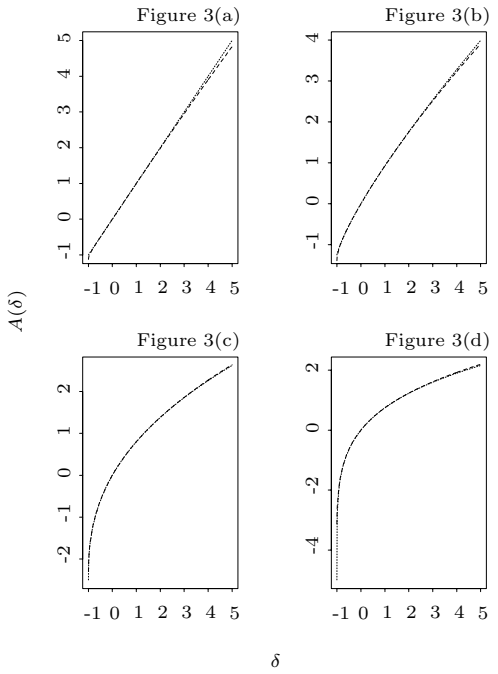

Figure 3. Residual adjustment functions for Cressie-Read family and $BWHD$ family for different combinations of (λ, α). (a) (0, 1/3), (b) $(-0.2, 0.4)$, (c) $(-0.6, 8/15)$, (d) $(-0.8, 0.6)$.

Parallel to the penalized Cressie-Read family we may define the penalized $BWHD^\alpha$ family ($PBWHD^\alpha$) by applying the same penalty on the empty cells. Write (4.1) as

$$\frac{1}{2}\sum_{x:d(x)\neq 0}\Big(\frac{(d(x)-f_\theta(x))}{\alpha d^{1/2}(x)+(1-\alpha)f_\theta^{1/2}(x)}\Big)^2+\frac{1}{2}\sum_{x:d(x)=0}\frac{f_\theta(x)}{(1-\alpha)^2} \tag{4.4}$$

and replacing $[2(1-\alpha)^2]^{-1}$ by unity we define the $PBWHD$ as

$$PBWHD^\alpha(\boldsymbol{d},\boldsymbol{f_\theta})=\frac{1}{2}\sum_{x:d(x)\neq 0}\Big(\frac{(d(x)-f_\theta(x))}{\alpha d^{1/2}(x)+(1-\alpha)f_\theta^{1/2}(x)}\Big)^2+\sum_{x:d(x)=0}f_\theta(x). \tag{4.5}$$

This family has all the asymptotic properties of the $BWHD^\alpha$ family and is also very close to the penalized Cressie-Read family for corresponding values $\alpha=(1-\lambda)/3$. For $\lambda=-0.5$, i.e. $\alpha=0.5$ the $BWHD^\alpha$ is identical to the Cressie-Read disparity, so that the corresponding penalized disparities are identical also.

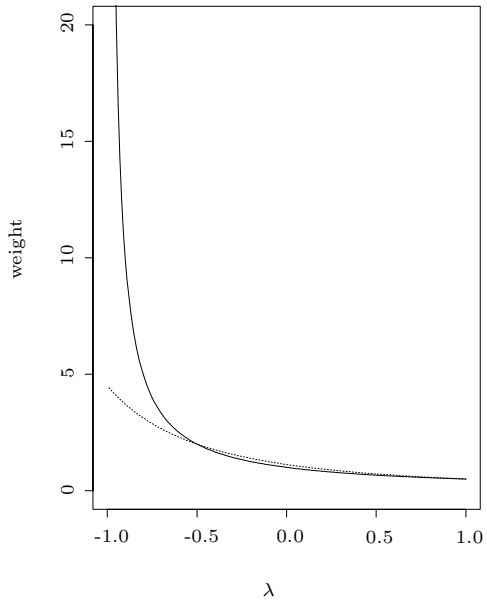

Figure 4. Weights of the empty cells for the Cressie-Read and $BWHD$ families plotted against λ (using the relation $\alpha = (1-\lambda)/3$ for the $BWHD$).

Despite its close correspondence with the Cressie-Read family, there are some differences in the treatment of empty cells within the $BWHD$ and the former. Some indication of this is provided by the left tails of the residual adjustment function of the disparities in Figure 3(d). To investigate this further, in Figure 4 we have plotted the weight of the empty cells, $(\lambda+1)^{-1}$ and $[2(1-\alpha)^2]^{-1}$ respectively, as a function of λ in the range $(-1, 1]$ for the Cressie-Read and the $BWHD$. (For the latter family we have plotted at λ the weight of the empty cell for the corresponding disparity $BWHD^{(1-\lambda)/3}$). From the graph it is clear that in the range $\lambda \in [-0.5, 1]$ the two disparities of the two families give approximately the same weight to the empty cells, but when λ is smaller than -0.5, the weight in the Cressie-Read family increases much faster. Therefore one might expect the small sample behavior of the minimum Cressie-Read disparities to deteriorate much faster for $\lambda \leq -0.5$. However the two penalized disparities appear to behave similarly. We demonstrate this with another exact computation.

4.2. Exact computation

Consider a multinomial distribution with $n = 10$ and $k = 4$, where the cell probabilities are generated by the *geometric*(θ) model (Section 3.1). At the true parameter $\theta = 0.1$, we compare exact MSEs for the estimators minimizing the ordinary and penalized Cressie-Read disparity, the $BWHD$ and the $PBWHD$

for several values of λ (and equivalent α) (Table 6). But for $-1 < \lambda < -0.5$ the $BWHD$ gives more precise estimates of the model parameter as is evident from the smaller values of the MSEs in the fourth column of Table 6 compared to the second. For $\lambda < -1$ exact calculations with the Cressie-Read disparitises are not possible, but the penalized family poses no such problems. For all values of $\lambda \leq -0.5$ the penalty leads to considerable reduction in the MSE in both families. The penalized estimators for the two classes behave similarly.

Table 6. Comparison of the exact MSEs of the minimum I_*^λ estimators, the corresponding minimum $BWHD^\alpha$ estimators, and the penalized versions for the 4 cell multinomial under the *geometric* model. True probabilities are generated by the *geometric*(0.1) distribution. The ordinary and penalized estimates of the $BWHD^\alpha$ family are denoted by $\hat{\theta}_B$ and $\hat{\theta}_{PB}$ respectively.

$\lambda(\alpha)$	$MSE(\hat{\theta}_I)$	$MSE(\hat{\theta}_P)$	$MSE(\hat{\theta}_B)$	$MSE(\hat{\theta}_{PB})$
0 (1/3)	0.003631	0.003631	0.003636	0.003631
−0.5(0.5)	0.004306	0.003630	0.004306	0.003630
−0.6(8/15)	0.004883	0.003652	0.004720	0.003653
−0.7(17/30)	0.005762	0.003695	0.005158	0.003695
−0.8(0.6)	0.007438	0.003745	0.005709	0.003740
−0.9(19/30)	0.010111	0.003807	0.006558	0.003792
−1.1(0.7)	—	0.003974	0.009025	0.003928
−1.3(23/30)	—	0.004256	0.010238	0.004117
−1.5(5/6)	—	0.004583	0.010759	0.004421
−1.7(0.9)	—	0.004786	0.010814	0.004666
−1.9(29/30)	—	0.004879	0.015892	0.004873

5. Concluding Remarks

In this paper we have provided an extensive study on the effects of an empty cell penalty on some density-based robust minimum disparity estimators for multinomial models. The aim was not just to find another robust estimator, but to find a robust estimator with good small sample efficiency. It appears that the penalized estimators studied do achieve good small sample efficiency in many cases, without compromising the robustness properties of the ordinary minimum disparity estimators.

References

Basu, A. and Sarkar, S. (1994a). On disparity based goodness-of-fit tests for multinomial models. *Statist. Probab. Lett.* **19**, 307-312.

Basu, A. and Sarkar, S. (1994b). The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *J. Statist. Comput. Simul.* **50**, 173-185.

Basu, A., Harris, I. R. and Basu, S. (1996). Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statist. Probab. Lett.* **27**, 367-373.

Basu, S. and Basu, A. (1995). Comparison of several goodness-of-fit tests for the kappa statistic based on exact power and coverage probability. *Statistics in Medicine* **14**, 347-356.

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

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

Harris, I. R. and Basu, A. (1997). A generalized divergence measure. Technical Report, Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India.

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

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* John Wiley, New York.

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer-Verlag, New York.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley, New York.

Shin, D. W., Basu, A. and Sarkar, S. (1995). Comparisons of the blended weight Hellinger distance based goodness-of-fit test statistics. *Sankhya Ser. B* **57**, 365-376.

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.

Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India.
E-mail: ayanbasu@isical.ac.in

Stat-Math Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India.
E-mail: srabashi@isical.ac.in

(Received August 1996; accepted May 1997)