ELSEVIER

Journal of Statistical Planning and Inference 93 (2001) 93–115

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Robust inference for finite poisson mixtures

Dimitris Karlis, Evdokia Xekalaki *

Department of Statistics, Athens University of Economics and Business, 76 Patision Str., 10434, Athens, Greece

Received 21 October 1999; received in revised form 21 June 2000; accepted 3 August 2000

Abstract

Inference for mixture models based on likelihood estimates suffers from lack of robustness. The presence of a few spurious observations may lead to incorrect decisions. In this paper we consider robust alternatives to the likelihood inference for finite Poisson mixtures based on the minimum Hellinger distance estimates. A new test, the Hellinger deviance test, is proposed for testing the Poisson hypothesis versus a Poisson mixture hypothesis. Moreover, diagnostics based on the Hellinger gradient function in order to examine for the presence of a mixture are described. Semiparametric estimation is also discussed. All these inferential procedures combine both efficiency when the model is correct and robustness when the model is incorrect, and make the minimum Hellinger distance methodology a competitive alternative to the maximum likelihood methodology.

MSC: 62F35; 62G35

Keywords: Gradient function; Influence; Hellinger deviance test; Robustness; Likelihood ratio test; Hellinger gradient function; Semiparametric estimation

1. Introduction

The Poisson distribution plays a prominent role in modeling discrete count data mainly because of its descriptive adequacy as a model when only randomness is present and the underlying population is homogeneous. Unfortunately, this is not a realistic assumption to make in modeling many real data sets. Poisson mixtures are well-known counterparts to the simple Poisson distribution for the description of inhomogeneous populations. Of special interest are populations consisting of a finite number of homogeneous subpopulations. In these cases the probability distribution of the population

* Corresponding author.
E-mail address: exek@aueb.gr (E. Xekalaki).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00207-X

can be regarded as a k-finite mixture of Poisson distributions defined by the probability function

$$f_{P_k}(x) = \sum_{i=1}^{k} p_i \frac{\exp(-\lambda_i)\lambda_i^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{1}$$

where P_k refers to the mixing distribution with k support points, which gives positive probability p_j to the point λ_j.

Rider (1962) described moment estimation for the parameters, while Hasselblad (1969) proposed an iterative scheme for maximum likelihood (ML) estimation for finite Poisson mixtures, which is an EM type algorithm, described later more formally by Dempster et al. (1977). The ML estimation suffers from high variability when the mixture components are close. However, the easily applicable EM algorithm has made the ML estimation the most popular method of estimation.

Recently, Karlis and Xekalaki (1998) proposed minimum Hellinger distance (MHD) estimation. MHD estimators are almost fully efficient when the model is correct and very robust when the model is not well specified or when some outliers have contaminated the data. This makes the MHD method a viable alternative to the commonly used ML method.

Apart from the estimation of the parameters, testing whether a mixed Poisson model is more adequate than a simple Poisson model, has been considered. A variety of such tests have been proposed in the literature. Among them the likelihood ratio test (LRT) for finite mixtures has been widely used. A detailed review of this test is given by McLachlan and Basford (1988) and Karlis and Xekalaki (1999). In the sequel, this test forms the basis for the development of inferential procedures using the Hellinger distance.

Robustness issues for finite mixtures have been overlooked in the Poisson case. There are only a few papers (see, e.g., Cutler and Cordero-Brana, 1996; Markatou, 2000 and the references therein) mainly focused on the case of normal mixtures. Our aim in this paper is to develop inferential procedures based on the Hellinger distance that have interesting robustness properties.

In Section 2 the MHD method for finite Poisson mixtures is described. Some interesting theorems for MHD estimation in finite mixtures are shown, extending the MHD method to the semiparametric case, where the number of support points is not known a priori. A test procedure referred to as the Hellinger deviance test (HDT) for testing the Poisson assumption against a 2-finite Poisson assumption is proposed in Section 3. Critical points of the null distribution of the test statistic derived via extensive simulation are reported in Section 4. The power of the test is compared to the power of the LRT in Section 5. The power of the HDT is almost the same as that of the LRT and in some cases larger, making the HDT a preferable procedure. The robustness of the HDT is examined in Section 6, while in Section 7 diagnostics for Poisson mixtures based on the Hellinger gradient function are proposed. These are more robust than the diagnostics proposed by Lindsay and Roeder (1992), and can thus be used in practice. Concluding remarks are made in Section 8.

2. Semiparametric estimation using the minimum Hellinger distance

Minimum distance estimation methods have attained a lot of attention as alternative methods of estimation to the commonly used ML method. In general, minimum distance estimation methods are appealing since they can be both robust and efficient. This trade off between robustness and efficiency is sometimes very useful for parametric estimation. Lindsay (1994) demonstrated how a general family of minimum distance estimation methods could be used for the estimation of discrete probability functions.

The MHD method is an interesting competitor to the ML method. Its study has revealed appealing efficiency and robustness properties, as can be seen from the results of Beran (1977) for parametric models, Eslinger and Woodward (1991), for normal models, Simpson (1987) and Lindsay (1994), for discrete distributions, Cutler and Cordero-Brana (1996) and Woodward et al. (1995), for normal mixtures and Karlis and Xekalaki (1998), for finite Poisson mixtures.

Suppose that $d(x)$ is the relative frequency of the value x from a sample of size n and $f_\theta(x)$ is the probability under the assumed model that the random variable X takes the value x, where θ denotes the vector of parameters of interest. The MHD estimators for discrete mixing distributions can be defined through the vector $\theta_{\min}$ which minimizes the Hellinger distance D given by

$$D(d, f_\theta) = \sum_{x=0}^{\infty} [\sqrt{d(x)} - \sqrt{f_\theta(x)}]^2 \tag{2a}$$

or equivalently by

$$D(d, f_\theta) = 2 - 2\sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)}. \tag{2b}$$

In the sequel, we refer to the estimation of the mixing distribution rather than to the estimation of the vector of parameters. Estimation of the mixing distribution requires minimizing (2a) or (2b) or equivalently maximizing the function

$$\varphi(P) = \sum_{x=0}^{\infty} \sqrt{d(x) f_P(x)}. \tag{3}$$

The representation in (3) is very useful since it allows for a comparison to the likelihood method where maximization of the function

$$L(P) = \sum_{x=0}^{\infty} d(x) \ln(f_P(x)) \tag{4}$$

is required.

We may extend the MHD method so as to be applicable to the case of semiparametric estimation of the mixing distribution (see Lindsay and Roeder, 1995). By the term semiparametric we refer to the case where the number of support points is unknown. The case of semiparametric ML estimation for mixture models has been treated by several authors (Simar, 1976; Laird, 1978; Lindsay, 1983; Lesperance and Kalbfleisch,

1992; Bohning, 1995 among others). The development of semiparametric estimation methods based on the MHD method is based on the similarity of (3) to (4).

Lindsay (1983a, b) provided the *General Mixture Maximum Likelihood Theorem* which gives sufficient and necessary conditions for the application of the ML estimation method in mixture models. If the mixing distribution is assumed to be continuous, we are restricted to estimate a finite-step approximation of the mixing distribution (see, e.g. Laird, 1978). Lindsay (1983a, b) provided the conditions for both, the case of a known number of support points and the case of an unknown number of support points. In fact, his theorems are generalizations of the general equivalence theorem for designs, given by Whittle (1973).

If k, the number of support points, is known, the aim is to maximize (3) with respect to all measures P with k support points. If k is unknown, maximization is considered over all measures P with a finite support. Whittle (1973) considered the case of concave distances ρ, showing that

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}e^2}\rho[(1-e)P+eG]\right]_{e=0} \leqslant 0$$

for all measures P, and G is a sufficient condition for their concavity.

It can be verified that the Hellinger distance ϕ satisfies this condition. Hence, the results of Whittle (1973) can be extended so as to lead to a general theorem on minimum Hellinger estimation for mixtures. It is very helpful to define the directional derivative of a general distance ρ at P to the direction of an alternative measure G by

$$D(P,G) = \lim_{e\to 0} \frac{\rho((1-e)P+eG)-\rho(P)}{e}. \tag{5}$$

For the Hellinger distance, the directional derivative $H(P,G)$ is given as

$$H(P,G) = \sum_{x=0}^{\infty} \sqrt{d(x)}\left[\frac{f_G(x)-f_P(x)}{\sqrt{f_P(x)}}\right].$$

Of special interest is the case where the measure G is a degenerate distribution at θ, i.e. it gives positive probability only at the point θ. In this case, the directional derivative is given by

$$\begin{aligned} H(P,\theta) &= \sum_{x=0}^{\infty} \sqrt{d(x)}\left[\frac{f(x\mid\theta)-f_P(x)}{\sqrt{f_P(x)}}\right] \\ &= \sum_{x=0}^{\infty} \sqrt{d(x)}\left[\frac{f(x\mid\theta)}{\sqrt{f_P(x)}} - \sqrt{f_P(x)}\right]. \end{aligned} \tag{6}$$

In the sequel, we refer to (6) as the Hellinger-gradient to distinguish it from the gradient function used in ML estimation and is defined by

$$D_{\mathrm{LIK}}(P,\theta) = \sum_{x=0}^{\infty} d(x)\left[\frac{f(x\mid\theta)}{f_P(x)} - 1\right]. \tag{7}$$

In both (6) and (7), $f(x\,|\,\theta)$ denotes the probability function of the simple Poisson distribution. The Hellinger gradient function can play an important role in minimum Hellinger estimation. The following theorem generalizes the results of Whittle (1973) and Lindsay (1983a):

Theorem 1. *The mixing distribution P is the MHD estimate of the mixing distribution if and only if*
(a) $H(P,\theta)\leqslant 0$, *for all* θ,
(b) $H(P,\theta)=0$, *for all* θ *in the support of* P,
(c) $H'(P,\theta)=0$ *for all* θ *in the support of* P *and*
(d) $H''(P,\theta)\leqslant 0$ *for all* θ *in the support of* P,

where the primes denote derivatives with respect to θ.

Proof. From conditions (a) and (b) it follows that all the support points are maxima of the gradient function. Hence, conditions (c) and (d) also hold. Since the only assumption of concavity is met, the theorem has been established. □

This theorem provides the conditions for the mixing distribution P to be the semiparametric MHD estimate. In this case the support is not restricted. In the case of restricted support (i.e. the case of a fixed number of support points), the following theorem can be shown:

Theorem 2. *The mixing distribution P is the MHD estimate of the mixing distribution with restricted support size if the following conditions are met*:
(a) $H(P,\theta)=0$, *for all* θ *in the support of* P.
(b) $H'(P,\theta)=0$, *for all* θ *in the support of* P.

Proof. For a k-finite Poisson mixture the estimating equations are derived by equating the first partial derivatives of (2a) with respect to the parameters with 0. Hence, the estimating equations are

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}(f(x\,|\,\lambda_j)-f(x\,|\,\lambda_k))=0,\quad j=1,2,\ldots,k, \tag{8}$$

and

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}p_j(f(x-1\,|\,\lambda_j)-f(x\,|\,\lambda_j))=0,\quad j=1,2,\ldots,k, \tag{9}$$

where $f(x\,|\,\theta)=\exp(-\theta)\theta^x/x!$, i.e. the probability function of a simple Poisson distribution.

From (8) we may derive easily that

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f(x\,|\,\lambda_j)=\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f(x\,|\,\lambda_k). \tag{10}$$

Table 1
The number of defaulted instalments in a Spanish bank (Dionne et al., 1996)

x	Frequency	x	Frequency	x	Frequency	x	Frequency
0	3002	9	53	18	8	27	0
1	502	10	41	19	6	28	1
2	187	11	28	20	3	29	1
3	138	12	34	21	0	30	1
4	233	13	10	22	1	31	1
5	160	14	13	23	0	32	0
6	107	15	11	24	1	33	0
7	80	16	4	25	0	34	1
8	59	17	5	26	0		

Also multiplying (8) with p_j and summing up we obtain that

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} f_\theta(x) = \sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} f(x \mid \lambda_k). \tag{11}$$

Combining (10) and (11) we obtain that

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} (f(x \mid \lambda_j) - f_\theta(x)) = 0 \quad \text{for } j = 1, \ldots, k$$

which is the Hellinger gradient function, i.e. we have proved the first condition.

Since for the Poisson distribution it holds that $f'(x \mid \lambda) = f(x-1 \mid \lambda) - f(x \mid \lambda)$ then (9) reduces to the second condition, i.e. to that the derivative of the Hellinger gradient function is 0 for all j. □

The key idea in Theorems 1 and 2 is that in the case of the unrestricted support the support points are the maxima of the Hellinger gradient function, while in the case of restricted support, the support points are not necessarily maxima, but they may also be minima or saddle points.

Note also that Bohning and Hoffman (1982) described distance-type estimation methods for probabilities and they gave theorems which simply require the concavity of these distances.

Example. Consider the data used in Dionne et al. (1996) concerning the number of defaulted installments in a financial institution in Spain, presented in Table 1. The data show great overdispersion. Thus, a Poisson mixture seems to be plausible for modeling the situation. In Fig. 1, one can see the Hellinger gradient function, plotted for several values of k. The semiparametric MHD estimate of the mixing distribution has $k = 6$ support points. For $k = 6$ all the support points are the local maxima of the Hellinger gradient function. Considering any additional support points would be redundant as either their mixing probabilities would be equal to 0 or the new support points would coincide with existing points. Fig. 1 shows that the MHD estimate of the mixing distribution cannot have more than six support points.

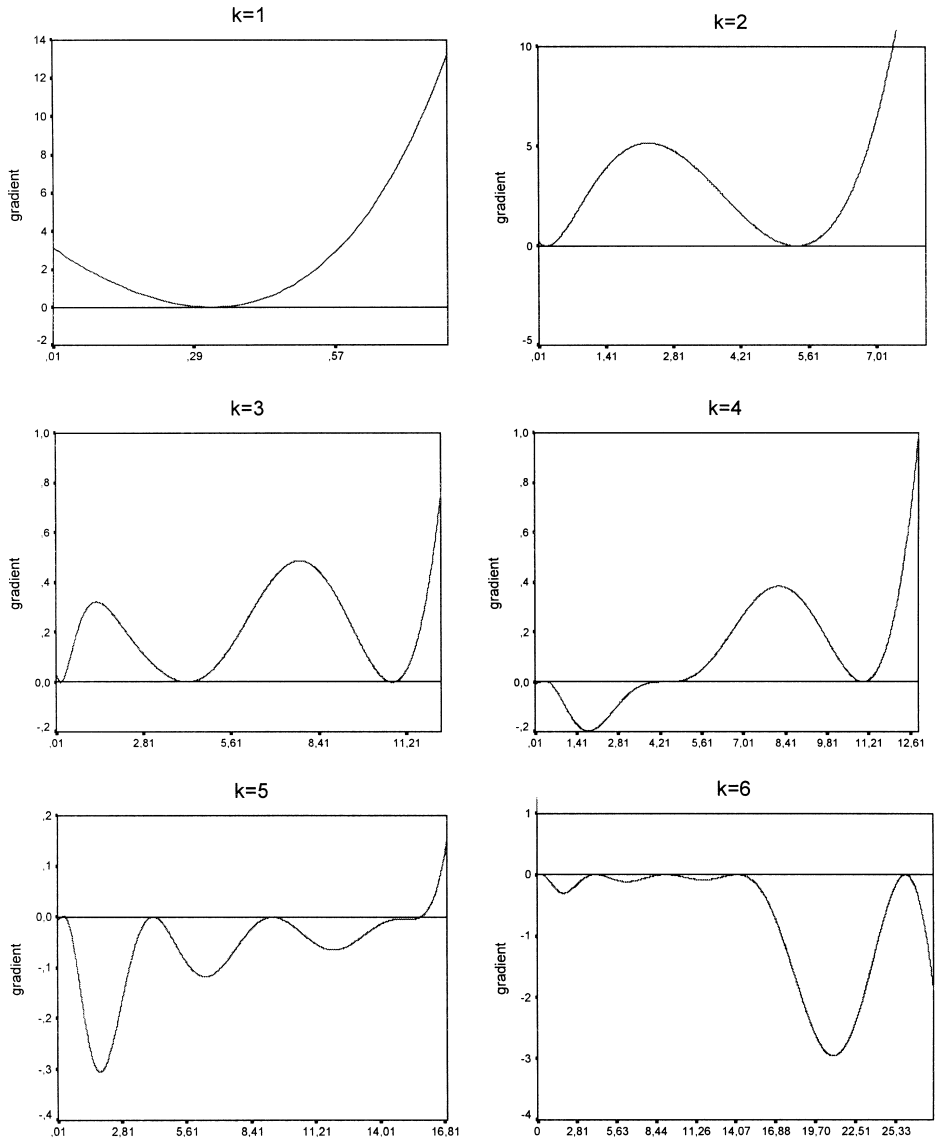

Fig. 1. The Hellinger gradient function plotted for the data of Table 1, for different values of k.

3. The Hellinger deviance test (HDT)

The LRT computes the statistic $L = 2[L_1 - L_0]$, where L_i, $i = 0, 1$, are the maximized loglikelihoods for the hypotheses in H_i. However, in the case of mixtures the asymptotic result for a χ^2 distribution with degrees of freedom equal to the difference in the numbers of parameters under the two hypotheses does not apply because the parameters

in the alternative hypothesis lie on the boundary of the parameter space whence the regularity conditions do not hold. Self and Liang (1987) showed that the asymptotic distribution is a mixture of a degenerate distribution at 0, and a χ^2 (1) distribution. Bohning et al. (1994) and Lindsay (1995) showed that this approximation is not valid in general. In order to overcome this problem of a null distribution of an unknown form, many authors tried to construct it via simulation (see, e.g., Symons et al., 1983; Bohning et al., 1994).

The robustness of the Hellinger distance makes it a potential tool for constructing a test statistic. It would be helpful to derive a test statistic, which would measure the improvement of the Hellinger distance if one new component were added. Since the influence of an outlier on this distance is much less than on the likelihood, a test based on the Hellinger distance is expected to be more robust against outliers.

Simpson (1989) proposed the use of the Hellinger distance analogues of the likelihood ratio tests for parametric inference. The Hellinger deviance test (HDT) statistic proposed is given by

$$\mathrm{HDT} = 4n[\mathrm{H}_0 - \mathrm{H}_1]$$

where H_i, $i=0,1$, are the minimized Hellinger distances for the distributions under the two hypotheses (see Simpson, 1989). Under some regularity conditions, the asymptotic distribution of the HDT is a χ^2 distribution with degrees of freedom equal to the difference in the numbers of parameters under the two hypotheses. This resembles the well-known LRT, discussed above. Again, however, the regularity conditions are not satisfied, making the asymptotic result irrelevant. Simpson (1989) showed that the HDT converges in probability to the LRT. This property indicates that the two tests have asymptotically the same properties.

The ambiguity for the distribution of the test statistic limits the usefulness of the test. In order to overcome this difficulty we propose the use of a bootstrap test. This means that we construct the null distribution via parametric bootstrap. The test proceeds as follows:

Step 1: Find the MHD estimates of the parameters of the simple Poisson distribution and the 2-finite Poisson mixture, say θ_H and θ_2 respectively and calculate the HDT statistic, say H_{obs}. The estimates can be easily obtained via the iterative algorithm given in Appendix A.

Step 2: Simulate B bootstrap samples of size n (n is the sample size of the data set) from the Poisson distribution with parameter θ_H, and for each bootstrap sample calculate the value of the HDT statistic, say T_j, $j=1,\ldots,B$.

Step 3: The p-value of the test will be the proportion of the values T_j of the HDT statistic that exceed the observed value H_{obs}.

The above scheme can be extended to test H_0: the data come from a k-finite Poisson mixture, against H_1: the data come from a $(k+1)$-finite Poisson mixture, by replacing θ_H by θ_k and θ_2 by θ_{k+1}. We focus our attention to testing the simple Poisson

hypothesis. Note that Beran (1988) has shown that bootstrap tests cannot be inferior to tests based on asymptotic results, recommending the use of bootstrap tests in cases where an exact result for the null distribution is not available.

From the general minimum Hellinger distance estimation theorem given above one can see that, for certain cases, the Hellinger distance cannot be minimized any further by adding a new component. This means that the HDT statistic is equal to 0; similar is the case for the LRT statistic. The identification of these cases can substantially reduce the computational effort required for applying the HDT via the bootstrap proposed in this section.

For the LRT, we found after extensive simulations that the LRT is 0 in all cases where the sample variance was less than the sample mean, i.e. if

$$s^2 < \bar{x}. \tag{12}$$

This is also the necessary condition derived from the *General Mixture Maximum Likelihood Theorem.* Unfortunately, the above result has not been proved analytically.

For the HDT statistic to be equal to zero the necessary condition can be found by checking if the second derivative of the Hellinger gradient function is negative. From the definition of the Hellinger gradient function it can be shown that

$$H'(P,\theta) = \sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x \mid \theta) \left(\frac{x-\theta}{\theta} \right)$$

and

$$H''(P,\theta) = \sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x \mid \theta) \left[\left(\frac{x-\theta}{\theta} \right)^2 - \frac{x}{\theta^2} \right].$$

Using Theorem 1, if the Hellinger distance in (2) has been minimized (or equivalently the distance in (3) has been maximized) by a k-support points mixing distribution P, the second derivative of the Hellinger Gradient function ought to be negative for all the support points. Hence, the following relation must hold for all the support points:

$$\sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x \mid \theta)(x-\theta)^2 < \sum_{x=0}^{\infty} \frac{x\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x \mid \theta) x,$$

for all θ in the support of P. In particular, for if $k = 1$, i.e. for testing the simple Poisson distribution against a 2-finite Poisson mixture, the above condition reduces to

$$\frac{\sum_{x=0}^{\infty} \sqrt{d(x) f(x \mid \theta_H)}(x-\theta_H)^2}{\sum_{x=0}^{\infty} \sqrt{d(x) f(x \mid \theta_H)}} < \theta_H, \tag{13}$$

where θ_H is the MHD estimator of the simple Poisson distribution which is the solution of the equation $\theta_H = \sum_{x=0}^{\infty} \sqrt{d(x) f(x \mid \theta_H)} x / \sum_{x=0}^{\infty} \sqrt{d(x) f(x \mid \theta_H)}$. Relation (13) is analogous to the one given above for the ML case in (12). The left-hand side is a

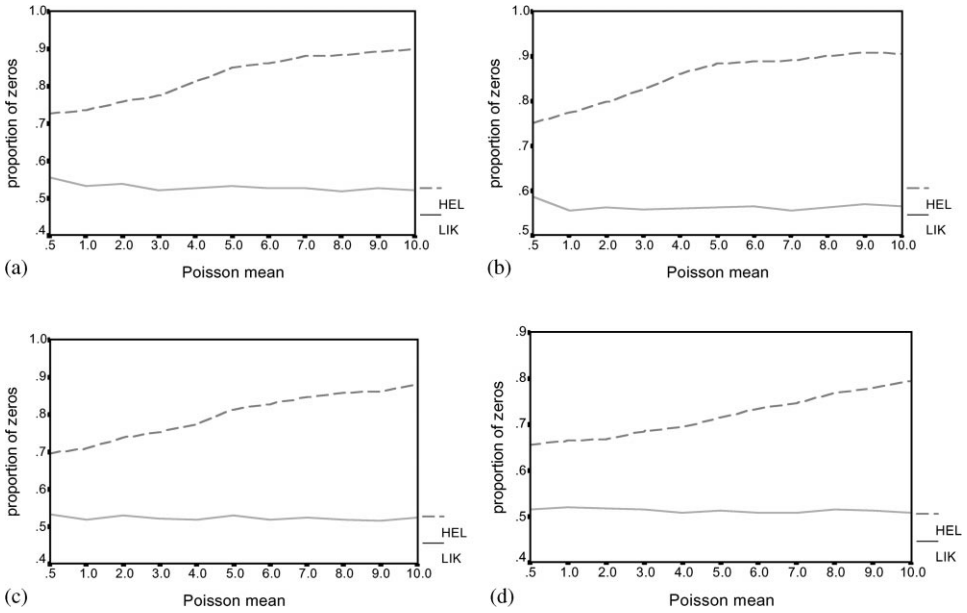

Fig. 2. The proportion of zeros calculated via 10 000 simulations for each value of Poisson mean and sample size.

weighted second moment around θ_H, corresponding to the sample variance in expression (12) for the ML case where the estimate of θ is the sample mean.

Again extensive simulation revealed that in all cases where (13) was satisfied, the HDT had a zero value.

For the LRT statistic, the proportion of zero values is near 0.5, as has been shown in Bohning et al. (1994). The proportion of 0 values of the HDT statistic is much higher. The proportion of 0 values was estimated via simulation for several sample sizes and values of the Poisson parameter. For each combination, 10 000 values were obtained and the probability of a 0 value was estimated as the proportion of 0 values among these 10 000 replications. It can be seen from Figs. 2a–d that the HDT statistic has a larger proportion of zero values. This proportion tends to increase with the value of the parameter of the Poisson distribution and to decrease with respect the sample size.

Another interesting question would be to prove the following conjecture:

Conjecture. *If the LRT statistic is* 0 *then the HDT statistic is also* 0. *The opposite is not necessarily true.*

We have not succeeded in proving the above conjecture, but we have run more than 100 million simulations with varying configurations of sampling distributions and sample sizes and there has not been any case with a 0 value for the LRT statistic and a nonzero value for the HDT statistic. We hope to be able to report a formal proof of the above conjecture soon.

Table 2
The 90th percentiles of the distribution of the HDT statistic generated via simulation

Mean	Sample size					
	20	50	100	200	250	500
0.3	0.69	0.76	0.83	1.04	1.07	1.16
0.5	0.72	0.88	1.05	1.17	1.19	1.22
0.75	0.85	0.99	1.14	1.26	1.29	1.29
1	0.92	1.06	1.19	1.33	1.32	1.28
1.5	0.94	1.09	1.25	1.36	1.38	1.28
2	0.89	1.08	1.22	1.34	1.34	1.16
2.5	0.83	1.00	1.17	1.28	1.26	1.12
3	0.74	0.94	1.10	1.18	1.19	1.07
5	0.11	0.35	0.60	0.79	0.80	0.83
7	0.01	0.07	0.22	0.37	0.42	0.56

Table 3
The 95th percentiles of the distribution of the HDT statistic generated via simulation

Mean	Sample size					
	20	50	100	200	250	500
0.3	1.33	1.43	1.65	1.89	1.95	2.09
0.5	1.36	1.66	1.85	2.06	2.11	2.23
0.75	1.56	1.85	2.04	2.22	2.25	2.33
1	1.76	1.93	2.13	2.35	2.32	2.42
1.5	1.85	2.07	2.26	2.41	2.46	2.58
2	1.88	2.09	2.28	2.46	2.50	2.48
2.5	1.83	2.03	2.25	2.46	2.46	2.42
3	1.75	1.96	2.20	2.36	2.42	2.39
5	0.94	1.32	1.66	1.93	1.97	1.95
7	0.49	0.69	0.99	1.26	1.34	1.43

4. Critical values for the HDT statistic

As has been previously mentioned, the distribution of the test statistic of the HDT is not known in closed form. The asymptotic results of Simpson (1989) do not apply and thus we need to estimate the null distribution via the parametric bootstrap. To do so we used the following procedure: 10 000 samples of size n were simulated from a Poisson distribution with mean λ. For each sample the value of the test statistic, say H_j, $j=1,\ldots,10\,000$, was calculated. The 10 000 values H_j were subsequently ordered leading to an ordered sample $H_{(j)}$, $j=1,\ldots,10\,000$, where $H_{(j)}$ is the jth-order statistic. Then the $100a\%$ critical point was estimated as $H_{(d)}$ where $d=[a^*10\,000]$, ($[a]$ is the integer part of a). This procedure was repeated 50 times. The entries of Tables 2–4 are the averages of these 50 repetitions. In Fig. 3 one can see the appropriate boxplots for the critical values.

It becomes obvious from Tables 2–4 that the critical values are not pivotal and they depend on both the sample size and the value of the Poisson parameter. This

Table 4
The 99th percentiles of the distribution of the HDT statistic generated via simulation

Mean	Sample size					
	20	50	100	200	250	500
0.3	2.66	3.35	3.67	4.06	4.19	4.52
0.5	3.31	3.84	4.02	4.39	4.45	4.75
0.75	3.80	4.08	4.35	4.62	4.71	4.83
1	4.02	4.29	4.52	4.81	4.81	5.02
1.5	4.40	4.48	4.77	4.97	5.06	5.31
2	4.42	4.60	4.86	5.07	5.11	5.38
2.5	4.39	4.54	4.77	5.10	5.15	5.31
3	4.33	4.46	4.75	5.01	5.09	5.42
5	3.42	3.71	4.18	4.55	4.62	4.86
7	2.54	2.70	3.18	3.75	3.83	4.33

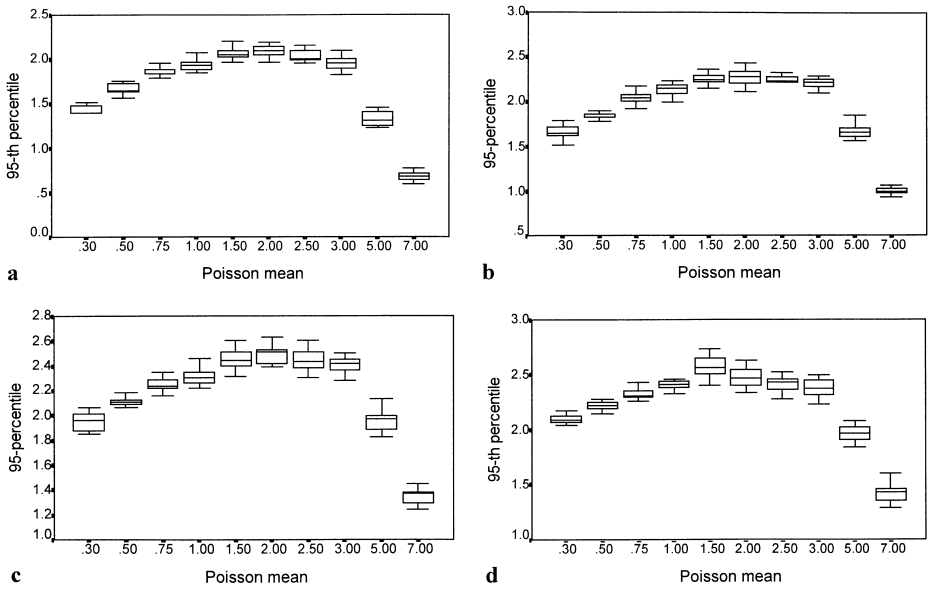

Fig. 3. Boxplots of the simulated percentiles of the HDT statistic for selected sample sizes ($n = 50$, 100, 250, 500) respectively for a to d.

makes full tabulation of the test statistic distribution impossible. We only report here these few critical values mainly as an initial indication. In practice, it would be better to proceed as follows: The researcher should calculate the observed value of the test statistic and reject the null hypothesis if this value is much greater than the reported values (employing the necessary interpolation for values not reported in the Tables 2–4). However, if the observed value is close to the reported values, the researcher should use the bootstrap for obtaining an estimate of the corresponding critical value for the specific value of λ. The use of a large number of bootstrap samples is recommended because the large proportion of 0 values leads to a poor estimation of the percentiles at the right tail of the distribution.

5. Power comparison for the HDT and the LRT

In order to study the performance of the HDT, the empirical power of the test is examined. Again the distribution of the test statistic under the alternative hypothesis is not known. To overcome this difficulty, a simulation-based approach is adopted again. The empirical power of the test is defined to be the proportion of times the null hypothesis is rejected when the data were generated from the alternative distribution. As critical values for the rejection of the null distribution, we used the results of the extensive simulation of the previous section. In order to compare the HDT to the LRT, the empirical power of the LRT was also calculated. To ensure comparability of the two tests the same bootstrap approach, as the one described in the previous section, was used for obtaining the critical values of the LRT.

Six different alternative distributions were chosen to represent the alternative hypothesis. All these alternatives have the same mean as the null distribution. Lindsay (1981) showed that the ML estimate of the mean of a k-finite Poisson mixture always coincides with the value of the sample mean. The six alternatives were 2-finite Poisson mixtures with parameter vectors:

(A) 0.5, 0.95λ, 1.05λ,
(B) 0.5, 0.5λ, 1.5λ,
(C) 0.8, 0.9λ, 1.4λ,
(D) 0.8, 0.5λ, 3λ,
(E) 0.2, 0.5λ, 1.125λ,
(F) 0.2, 0.9λ, 1.025λ,

where λ is the Poisson parameter of the simple Poisson model.

The alternatives were chosen to represent specific kinds of departure from the null distribution. For example, alternative A departs very little from a Poisson distribution. The same is true for alternative F, but now the resulting distribution is more skew. From these alternatives 50 000 samples were drawn and the empirical power is reported in Table 5. The values of the Poisson parameter, $\lambda = 1, 3, 5$ were used to generate the samples. The sample sizes were $n = 20$, 50, 100, 200, 250, 500. The significance level was set to $\alpha = 5\%$ for all the tests.

The entries of Table 5 reveal the nice performance of the HDT. The HDT seldom performs worse than the LRT; for several cases (especially for small sample sizes) the difference is substantial. This leads to the conclusion that the HDT is at least as efficient as the LRT, and its use is thus preferable because of its robustness. This issue is further discussed in the next section.

6. Robustness of the HDT

Assessing the robustness of a test statistic is not a straightforward task. The main problem is that there does not exist a global definition of the notion of robustness.

Table 5
The power of the HDT and the LRT

n	Alternatives											
	A		B		C		D		E		F	
	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT
$\lambda = 1$												
20	0.053	0.051	0.167	0.182	0.063	0.065	0.452	0.616	0.058	0.054	0.075	0.072
50	0.051	0.047	0.290	0.286	0.066	0.069	0.832	0.902	0.053	0.050	0.088	0.085
100	0.050	0.045	0.453	0.442	0.076	0.071	0.978	0.990	0.050	0.044	0.114	0.102
200	0.051	0.038	0.702	0.684	0.102	0.088	1.000	1.000	0.058	0.047	0.155	0.128
250	0.055	0.040	0.786	0.765	0.109	0.088	1.000	1.000	0.057	0.041	0.180	0.141
500	0.051	0.041	0.960	0.959	0.131	0.119	1.000	1.000	0.052	0.039	0.242	0.208
$\lambda = 3$												
20	0.054	0.058	0.546	0.581	0.087	0.103	0.544	0.962	0.052	0.055	0.151	0.152
50	0.052	0.050	0.872	0.874	0.110	0.128	0.901	1.000	0.056	0.052	0.239	0.223
100	0.054	0.045	0.992	0.993	0.157	0.164	0.993	1.000	0.055	0.045	0.375	0.331
200	0.053	0.040	1.000	1.000	0.229	0.242	1.000	1.000	0.052	0.036	0.576	0.520
250	0.054	0.039	1.000	1.000	0.273	0.275	1.000	1.000	0.053	0.037	0.662	0.594
500	0.057	0.056	1.000	1.000	0.473	0.538	1.000	1.000	0.057	0.057	0.894	0.873
$\lambda = 5$												
20	0.051	0.049	0.835	0.862	0.111	0.138	0.299	0.985	0.061	0.059	0.247	0.246
50	0.053	0.054	0.994	0.996	0.175	0.205	0.506	1.000	0.051	0.050	0.451	0.418
100	0.057	0.041	1.000	1.000	0.268	0.292	0.763	1.000	0.060	0.044	0.681	0.621
200	0.061	0.040	1.000	1.000	0.454	0.485	0.943	1.000	0.063	0.044	0.917	0.879
250	0.053	0.043	1.000	1.000	0.510	0.562	0.974	1.000	0.057	0.043	0.951	0.932
500	0.068	0.070	1.000	1.000	0.820	0.866	0.999	1.000	0.072	0.076	0.999	0.999

Usually, a procedure is said to be robust if a departure from the assumptions does not destroy the performance of the procedure. Two approaches are commonly considered in assessing robustness. The first is termed *Data Contamination* and refers to the case where some observations not belonging to the assumed model are included in the data set thus destroying the underlying assumptions. Such a case is the presence of some outliers at the tails of a distribution. The second approach is termed as *Model Deviation* and refers to the case where the assumed model is not correct but is a little different from the true model. For mixture models, model deviation is much more complicated as either the component distributions or the number of components can be incorrectly specified (see Lindsay, 1995).

The above two notions, however, have a common element. The usual way to describe data contamination is through mixture models, namely one assumes that the observations come from a model $(1-\xi)P + \xi G$, where P is the assumed distribution, G is the contaminant which causes the departure from the assumed model and ξ is the proportion of contaminated values. With this representation, data contamination implies model deviation. However, this representation can help us to examine the effect of a few observations, usually at the tails of the assumed distribution, where the model deviation implies more general intrinsic departures from the assumed model.

For a goodness of fit test, like the HDT or the LRT, one seeks a test which can detect the assumed model from the data. Clearly, the model deviation approach is misleading. If the test cannot discriminate between the true model and the assumed model, the test lacks power and thus it is not helpful. However, a goodness of fit test, which can ignore a few spurious observations, is very useful in practice since for some cases the rejection of a goodness of fit hypothesis is caused by a few observations. In the sequel, we will consider departure from assumptions through data contamination.

The robustness of tests has been examined for several tests and from several points of view. Ylvisaker (1977) examined the resistance of a test statistic which is defined as the smallest proportion of observations which can determine the decision ignoring the values of all the remaining observations. Lambert (1981) proposed the use of influence functions to examine the behavior of statistical tests. Hertier and Ronchetti (1994) have shown that the influence curves of both the level and the power of a test are proportional to the influence curves of the estimators used. Later, He et al. (1990) examined the power breakdown points of test statistics. The power breakdown point is the amount of contamination of each alternative distribution that can lead the test statistic to a null value. For a qualitative examination of test robustness the reader is referred to Lambert (1982). Simpson (1989) and Lindsay (1994) have shown that tests based on the Hellinger distance can be more robust than those based on the likelihood because of the robustness of the Hellinger estimators.

In our case, the fact that the null distribution of the test statistic is not known and has to be estimated via simulation, prohibits full adoption of the above mentioned approaches. However, in order to demonstrate the superiority of the HDT relative to LRT, some comparisons were made using the Influence function of the test statistic or by examining the performance of the tests when some contamination is present.

The influence function for a functional $T(F)$, where F is the empirical distribution, is defined as

$$\mathrm{IF}(x,T,F)=\lim_{t\to 0}\frac{T((1-t)F+t\Delta_x)-T(F)}{t} \tag{14}$$

and measures the change of the functional $T(F)$ if an infinitesimal contamination is added at point x (see Hampel et al., 1986). From (5), one can see that by their definition, the Hellinger gradient function and the gradient function can be regarded as influence functions for the corresponding distances when a new component is added. It would be interesting to examine the IF for the corresponding distances for the two methods defined in (3) and (4). From (14) the limit cannot be obtained. Thus, using de L'Hôpital's rule we obtain, for the ML method

$$\mathrm{IF}(z,L,F)=-\sum_{x=0}^{\infty} d(x)\ln f(x)+\ln f(z), \tag{15}$$

i.e. for a value of z such that $f(z)$ is near 0 (an outlier), the IF is very large. In other words, an outlier can dramatically change the loglikelihood.

On the other hand, the IF for the Hellinger distance is bounded, since the distance in (3) is bounded in [0,1]. The IF is given as

$$\mathrm{IF}(z,\phi,F)=\frac{1}{2}\left[-\sum_{x=0}^{\infty}\sqrt{d(x)f(x)}+\sqrt{f(z)}\right]. \tag{16}$$

To prove this note that, by definition, $\phi[(1-t)F+t\Delta_z]=\sqrt{1-t}\phi(F)+(1-\sqrt{1-t})\sqrt{f(z)}$. Then, differentiating we obtain (16). In both (15) and (16), the probability function $f(x)$ is calculated using the corresponding estimates. If z is an outlier we expect $f(z)$ to be very small, i.e. very close to 0. Since the logarithm near 0 decreases more sharply the influence function is also sharper. This indicates that the MHD is not influenced so much by an outlier. Note that the above influence function is based on the distances themselves and not on the maximized (minimized) distances.

On the other hand, the test statistics associated with the two methods will have influence functions which ignoring constants will depend on $\sqrt{f_1(z)}-\sqrt{f_0(z)}$, for the MHD method, and on $\ln\{f_1(z)/f_0(z)\}$, for the ML method, where the subscript in the probability function denotes the distribution used which is determined from the H_i, $i=0,1$. To see this result we can use the definition of the test statistics as the differences between the distances under the two hypotheses. Thus, the IF will be the difference of the two IF for the corresponding distances, and thus the first terms in (15) and (16) will lead to the statistics and the remaining of the above mentioned quantities.

Two facts support the superiority of the MHD method. The first is that if an outlier is present, the MHD estimates do not differ much between the two models whence an influence close to 0 is expected. For the ML method, on the contrary, the change of the estimates causes a positive influence. It is known that a mixed Poisson distribution has thicker tails than the simple Poisson distribution with the same mean (Shaked, 1980). For testing purposes the means of the two models are assumed to be equal (Lindsay, 1981) and thus the ratio f_1/f_0 is greater than 1; the influence is always positive.

An empirical result will also be given to further support the above-mentioned issue. Suppose that the functional $T(F)$ is the corresponding test statistic for the two methods. Since this statistic does not have a closed-form expression, it is not possible to compute the influence function. An alternative approach is the use of the Empirical Influence Function. According to Hampel et al. (1986, p. 93), the EIF of the estimator based on any sample is a plot of the values of the estimator, if one more observation (contaminant) is added at the point x.

So, this EIF was used to examine the behavior of the two tests. One thousand samples of size $n=25$, 100, 250, 1000 were sampled from a Poisson distribution with parameter $=1$. The EIF was then calculated by adding a $(n+1)$th observation at point x. The averaged influences for all the points $x=0,1,\ldots,20$ were then reported. By such an approach, results due to sampling errors were eliminated and a clearer and more reliable picture of the robustness of the test to contamination was obtained.

Fig. 4 can show the behavior of both the HDT and the LRT. The LRT is clearly influenced very much by the outlier observation, and even when the sample size is

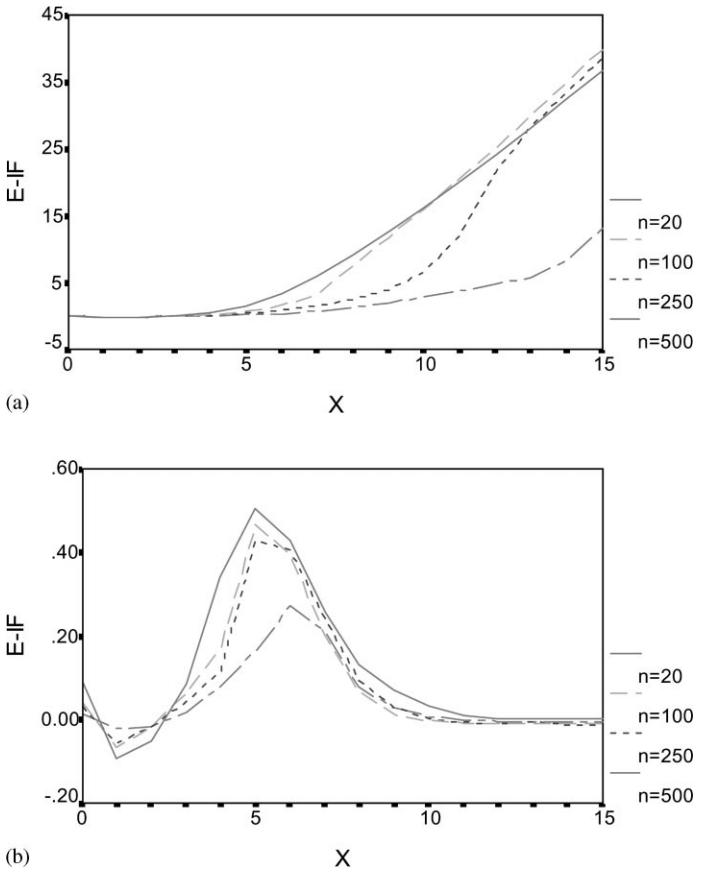

Fig. 4. (a) The E-If for the LRT statistic, when one more observation is added at point x. (b). The E-If for the HDT statistic, when one more observation is added at point x.

as large as 500, an outlier can lead to incorrect conclusions. Note also that an outlier can lead the test statistic to infinite values. This means that the resistance according to Ylvisaker (1977) and Simpson (1989) of the LRT is equal to 0, since an outlier is sufficient to make the test statistic very large. The behavior of the HDT is very different. On average, the test will not reject the null hypothesis, and an outlier cannot alter the decision. Note also that according to Simpson (1989), the resistance of the HDT can be calculated and it is always larger than $d^2(F)/(1+d^2(F))$, where $d(F)$ is the difference of the minimized distances of the two hypotheses.

In the sequel, the behavior of the test is examined when more outliers are present using samples from a Poisson distribution with mean 1, contaminated by a degenerate distribution at point $x=8, 12$. The proportions of contamination considered were $\xi=0.01$, 0.02 (i.e. ξ is the probability that an outlier observation is drawn at x). A robust test ought to cope with such a contamination, in the sense that the significance level of the test must not increase very much. The significance level was set at $\alpha = 5\%$. Table 6

Table 6
The calculated significance level of the test for contaminated models. The actual level is 5%

Models used in the comparison	$x = 8$,	$\xi = 0.01$	$x = 12$,	$\xi = 0.01$	$x = 8$,	$\xi = 0.02$	$x = 12$,	$\xi = 0.02$
Sample size n	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT
20	0.054	0.222	0.050	0.224	0.063	0.426	0.050	0.428
100	0.073	0.620	0.050	0.652	0.114	0.907	0.052	0.925
250	0.090	0.756	0.049	0.913	0.134	0.986	0.048	0.998
500	0.094	0.935	0.044	0.986	0.159	1.000	0.050	1.000

contains the true significance level when samples were taken from the 4 above described models for both tests. The entries of the table were based on 10 000 simulated samples.

From the entries of Table 6 one can see again that the HDT is far more robust. When the contamination is at $x = 12$ the HDT almost ignores this observation. Note that for $n = 500$ and $\xi = 2\%$ we have 10 outlier observations and the HDT ignores them. On the contrary, the LRT cannot cope with such a contamination, and as the sample size increases it almost surely rejects the null hypothesis.

It should be emphasized that robustness and power are rather conflicting issues for tests, especially when one aims at examining goodness of fit tests as is the case here. The reason is that a sensitive test is required, which can detect departures from the model under the null hypothesis. So, if a test is very sensitive, a few observations can destroy its performance. In this sense, it is preferable to find a test which is not so sensitive and can detect 'faults' which are caused by the alternative hypothesis but not from a contamination mechanism. HDT seems to be such a test, which combines high power when the data are not contaminated and robustness when the data have been contaminated.

Lindsay (1995) demonstrated that Neyman $C(a)$ tests can also have high power. In the mixture case, such a test does not require iterative algorithms and thus it is more practical. The HDT introduced in this paper has the added property of being robust and hence it is recommended when the data are suspected to be contaminated, at the cost of requiring some computing time.

7. The Hellinger gradient function as a diagnostic tool for the Poisson distribution

In this section we will examine the use of plots of the Hellinger gradient function as a diagnostic tool for detecting whether a k-finite mixture is appropriate and particularly to detect if the Poisson distribution is an adequate distribution for modeling the data. The idea is based on the use of the simple gradient function introduced by Lindsay and Roeder (1992). They proposed that the plot of the gradient function can reveal if the homogeneity model is more appropriate than the inhomogeneity model, i.e. if a simple Poisson distribution is more adequate than a finite Poisson mixture, and in

general if a k-finite mixture is more adequate than a $(k+1)$-finite mixture model. The key ingredient is that if the Poisson model is true, the gradient function should be a concave function with maximum at the point of the sample mean. Any deviation from this picture reveals departures from the simple Poisson model keeping in mind that small deviations can have been caused by sampling variability. We will follow such an approach by using the Hellinger gradient function instead of the gradient function.

The aim of a diagnostic plot is similar to the aim of a detector. It cannot say that something is surely true, but it can reveal if something is clearly false. Diagnostics can simply navigate through different choices. This is the case of the Hellinger gradient function as diagnostic tool. The concavity implies that the simple Poisson model is more adequate. However, a non-concave picture it is no proof of non-poissonity but an indication for this.

Let us now examine more thoroughly this issue. If the Poisson model is true, from the results of Section 3, we know that the Hellinger gradient function has zeroes only at the point of the MHD estimate of the Poisson parameter, and it is concave. Thus, a plot of the function suffices to provide a picture about the consistency of the assumed Poisson model. On the other hand, the resistance of the MHD method when some outliers are present makes the Hellinger gradient function a more promising diagnostic plot. Another interesting point is that, as Fig. 3 depicts, the gradient function will not be concave about 50% of the time when the data are generated from a Poisson distribution. For the Hellinger gradient function, the probability is lower. Therefore, the Hellinger gradient function can better detect if the data come from the Poisson distribution. Fig. 5 depicts some cases sampled from a Poisson with mean equal to 1 and sample size $n = 100$. Case 1 corresponds to the case when the two functions disagree. The gradient function shows that the Poisson model is inadequate, while the Hellinger gradient function supports the opposite as can be seen from the concavity of the function. In case 3, both functions support the simple Poisson model, while in cases 2 and 4 the simple Poisson model is judged as not adequate by both functions. One can see that the Hellinger gradient function can better detect the true model.

Expanding the use of the Hellinger gradient function, we may use it for more than one component. In each case the concavity of the Hellinger function supported the model with k-points of support, while any departure is evidence against this model in favor of a model with $(k+1)$ points of support.

Another important issue is the sampling error of the Hellinger gradient function. Lindsay and Roeder (1993) proposed the use of a confidence band, using component-wise asymptotic normality for all the points where the gradient function is evaluated. However, the asymptotic result is rather poor for small sample sizes. A truncated version of the gradient function was also used, because of the unlimited range of the Poisson distribution.

Finally, in the case where some outlier observations have contaminated the data, the Hellinger gradient function has a local maximum at the support point but it also has a sharp peak near the region where the outlier exists. On the contrary, the gradient function, does not have a local maximum. The above implies that both functions must

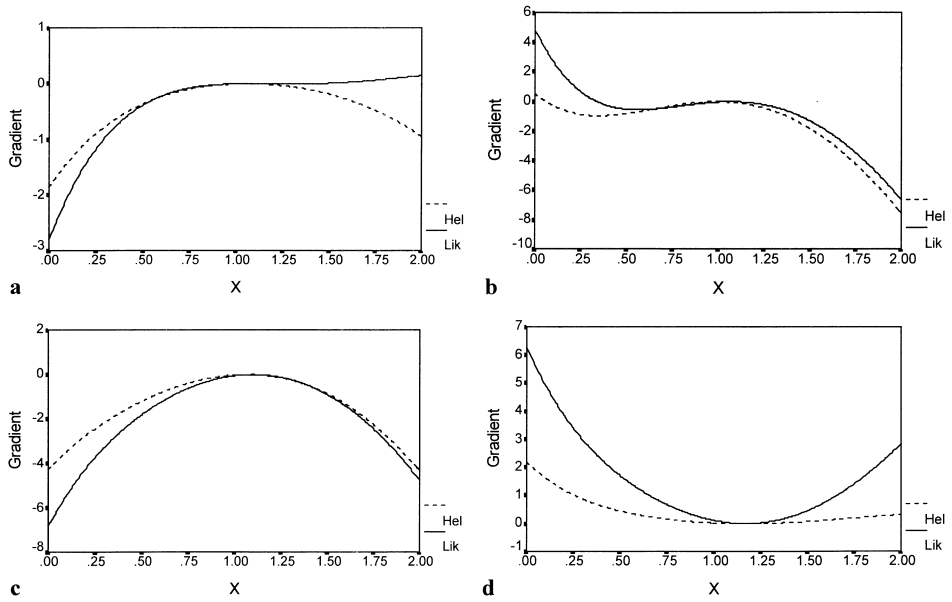

Fig. 5. The gradient function and the Hellinger Gradient function for samples of size $n = 100$ from a Poisson distribution with parameter 1.

be plotted in the entire range of the data. Sharp effects away from the main body of the data indicate outliers.

8. Conclusions

The MHD method for finite Poisson mixtures is both efficient and robust. It is also computationally feasible with low effort (at least the effort required is almost identical with the ML method). Since it combines these two potentially useful characteristics, its use is recommended. We have shown that we may use Hellinger distance-based methodologies for semiparametric estimation, resulting in hypothesis tests and diagnostic plotting which are very efficient and at the same time robust. The latter property is not true for likelihood based inferences when an outlier may cause inconsistencies. Therefore, MHD methodologies seem to be viable (if not better) alternatives which can cope with spurious data sets, and thus are highly recommended. Further research would be interesting in order to expand its use.

For example, consider the likelihood-based cluster analysis of rare events given in Symons et al. (1983). In such applications the presence of an outlier can cause problems if the ML estimates were used for obtaining the membership probabilities. An influenced ML estimate can lead to inconsistent results. A minimum Hellinger distance-based approach can be useful to cope with outliers in such applications.

The LRT was employed by Karlis and Xekalaki (1999) in a sequential manner to test for the number of components. The HDT described in Section 3 can be also used in a in a similar context.

Extensions of the Hellinger distance-based methodology to cases of finite mixtures of other distributions (like the normal or the exponential) are obvious. However, the effort is greater, since the MHD estimation is not so clear for continuous models. Cutler and Cordero-Brana (1996) have derived MHD estimators for finite normal mixtures. So, the MHD methodology presented in this paper could be extended to the case of k-finite normal mixtures.

Acknowledgements

The authors would like to thank an anonymous referee for constructive comments. Dimitris Karlis acknowledges support by a scholarship from the State Scholarships Foundation of Greece.

Appendix. The algorithm for deriving the MHD estimates

Karlis and Xekalaki (1998) developed a simple iterative algorithm for calculating the MHD estimates. This algorithm starts with initial values for the parameters, which are updated at each iteration until some kind of convergence is detected. Several initial values ought to be considered in order to be sure that the global minimum has been obtained. The steps of the algorithm are the following:

Step 1. Give the values obtained from the ith iteration $\lambda_j^{(i)}$, $j=1,\ldots,k$, and p, $j=1,\ldots,k-1$, calculate the weights w_{xj}, using $w_{xj}=f(x\mid\lambda_j^{(i)})/\sqrt{f_\theta(x)}$ where $f_\theta(x)$ is calculated using the current estimates.

Step 2. Calculate the parameter estimates using

$$\textit{Step 2a}\quad \lambda_j^{(i+1)}=\frac{\sum_{x=0}^{m} w_{xj}x\sqrt{d(x)}}{\sum_{x=0}^{m} w_{xj}\sqrt{d(x)}},\quad j=1,2,\ldots,k$$

and the mixing proportions using

$$\textit{Step 2b}\quad p_j^{(i+1)}=\frac{\sum_{x=0}^{m} p_j^{(i)}w_{xj}\sqrt{d(x)}}{\sum_{x=0}^{m}\sqrt{d(x)f_\theta(x)}},\quad j=1,\ldots,k-1 \text{ and } p_k=1-\sum_{i=1}^{k-1} p_i,$$

where m denotes the largest observed value, and with $f_\theta(x)$ represents the k-finite mixture given by (1).

Step 3. Check if some convergence criterion is satisfied, otherwise go back to step 1, using the current estimates as initial values to make the next iteration.

Clearly, we only need initial values for the estimates. If the initial values are within the acceptable range for the parameters, the estimated values are also within the range

of parameters. The algorithm is very similar to the well known EM algorithm for mixtures.

References

Beran, R.J., 1977. Minimum Hellinger distance estimates for parametric models. Ann. Statist. 5, 445–463.
Beran, R.J., 1988. Prepivoting test statistics: a bootstrap review of asymptotic refinements. J. Am. Statist. Assoc. 83, 687–697.
Bohning, D., 1995. A review of reliable maximum likelihood algorithms for semiparametric mixture models. J. Statist. Plann. Inference 47, 5–28.
Bohning, D., Dietz, E., Schaub, R., Schlattman, P., Lindsay, B., 1994. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann. Inst. Statist. Math. 46, 373–388.
Bohning, D., Hoffman, K.H., 1982. Numerical techniques for estimating probabilities. J. Statist. Comput. Simulation 14, 283–293.
Cutler, A., Cordero-Brana, O., 1996. Minimum Hellinger distance estimation for finite mixture models. J. Am. Statist. Assoc. 91, 1716–1724.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38.
Dionne, G., Artis, M., Guillen, M., 1996. Count data models for a credit scoring system. J. Empirical Finance 3, 303–325.
Eslinger, P.W., Woodward, W.A., 1991. Minimum Hellinger distance estimation for normal models. J. Statist. Comput. Simul. 39, 95–113.
Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986. Robust Statistics. Wiley, New York.
Hasselblad, V., 1969. Estimation of finite mixtures from the exponential family. J. Am. Statist. Assoc. 64, 1459–1471.
He, X., Simpson, D., Portnoy, S., 1990. Breakdown robustness of tests. J. Am. Statist. Assoc. 85, 446–452.
Hertier, S., Ronchetti, E., 1994. Robust bounded-influence tests in general parametric models. J. Am. Statist. Assoc. 89, 897–904.
Karlis, D., Xekalaki, E., 1998. Minimum Hellinger distance estimation for finite Poisson mixtures. Comput. Statist. Data Anal. 29, 81–103.
Karlis, D., Xekalaki, E., 1999. On testing for the number of components in finite Poisson mixtures. Ann. Ins. Statist. Math. 51, 149–161.
Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Statist. Assoc. 73, 805–811.
Lambert, D., 1981. Influence functions for testing. J. Am. Statist. Assoc. 76, 649–657.
Lambert, D., 1982. Qualitative robustness of tests. J. Am. Statist. Assoc. 77, 352–357.
Lesperance, M., Kalbfleisch, J., 1992. An algorithm for computing the nonparametric MLE of a mixing distribution. J. Am. Statist. Assoc. 87, 120–126.
Lindsay, B., 1981. Properties of the maximum likelihood estimator of a mixing distribution. In: Patil, G.P. (Ed.), Statistical Distributions in Scientific Work, Vol. 5, pp. 95–109, Reidel, Dordrecht, Holland.
Lindsay, B., 1983a. The geometry of mixture likelihood: a general theory. Ann. Statist. 11, 86–94.
Lindsay, B., 1983b. The geometry of mixture likelihood. Part II. The exponential family. Ann. Statist. 11, 783–792.
Lindsay, B., 1994. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann. Statist. 22, 1081–1114.
Lindsay, B., 1995. Mixture Models: Theory, Geometry and Applications, Regional Conference Series in Probability and Statistics, Vol. 5. Institute of Mathematical Statistics and Am. Statist. Association, Hayward, California.
Lindsay, B., Roeder, K., 1992. Residual diagnostics for mixture models. J. Am. Statist. Assoc. 87, 785–794.
Lindsay, B., Roeder, K., 1995. A review of semiparametric mixture models. J. Statist. Plann. Inference 47, 29–39.
Markatou, M., 2000. Mixture models, robustness and the weighted likelihood methodology. Biometrics 56, 483–486.

McLachlan, G., Basford, K., 1988. Mixture Models: Inference and Applications to Clustering. Marcel and Dekker Inc., New York.

Rider, P., 1962. Estimating the parameters of mixed Poisson, Binomial and Weibull distributions by the method of moments. Bulle. Internat. Statist. Inst. 39 (2), 225–232.

Self, S., Liang, K., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Statist. Assoc. 82, 605–610.

Shaked, M., 1980. On mixtures from exponential families. J. Roy. Statist. Soc. Ser. B 42, 192–198.

Simar, L., 1976. Maximum likelihood estimation of a compound Poisson process. Ann. Statist. 4, 1200–1209.

Simpson, D., 1987. Minimum Hellinger distance estimation for the analysis of count data. J. Am. Statist. Assoc. 82, 802–807.

Simpson, D., 1989. Hellinger deviance tests: efficiency, breakdown points and examples. J. Am. Statist. Assoc. 84, 107–113.

Symons, M., Grimson, R., Yuan, Y., 1983. Clustering of rare events. Biometrics 39, 193–205.

Whittle, P., 1973. Some general points in the theory of optimal experimental designs. J. Roy. Statist. Soc. Ser. B 35, 123–130.

Woodward, W., Whitney, P., Eslinger, P., 1995. Minimum Hellinger distance estimation of mixture proportions. J. Statist. Plann. Inference 48, 303–319.

Ylvisaker, D., 1977. Test resistance. J. Am. Statist. Assoc. 72, 551–557.