Open Journal of Statistics, 2012, 2, 369-382
http://dx.doi.org/10.4236/ojs.2012.24045 Published Online October 2012 (http://www.SciRP.org/journal/ojs)

Minimum Penalized Hellinger Distance for Model Selection in Small Samples

Papa Ngom*, Bertrand Ntep
Laboratoire de Mathematiques et Applications (LMA), Universite Cheikh Anta Diop, Dakar-Fann, Senegal
Email: *papa.ngom@ucad.edu.sn, ntepjojo@yahoo.fr

Received May 27, 2012; revised June 25, 2012; accepted July 10, 2012

ABSTRACT

In statistical modeling area, the Akaike information criterion AIC, is a widely known and extensively used tool for model choice. The φ-divergence test statistic is a recently developed tool for statistical model selection. The popularity of the divergence criterion is however tempered by their known lack of robustness in small sample. In this paper the penalized minimum Hellinger distance type statistics are considered and some properties are established. The limit laws of the estimates and test statistics are given under both the null and the alternative hypotheses, and approximations of the power functions are deduced. A model selection criterion relative to these divergence measures are developed for parametric inference. Our interest is in the problem to testing for choosing between two models using some informational type statistics, when independent sample are drawn from a discrete population. Here, we discuss the asymptotic properties and the performance of new procedure tests and investigate their small sample behavior.

Keywords: Generalized Information; Estimation; Hypothesis Test; Monte Carlo Simulation

1. Introduction

A comprehensive surveys on Pearson chi-square type statistics has been provided by many authors as Cochran [1], Watson [2] and Moore [3,4], in particular on quadratics forms in the cell frequencies. Recently, Andrews [5] has extended the Pearson chi-square testing method to non-dynamic parametric models, *i.e.*, to models with covariates. Because Pearson chi-square statistics provide natural measures for the discrepancy between the observed data and a specific parametric model, they have also been used for discriminating among competing models. Such a situation is frequent in Social Sciences where many competing models are proposed to fit a given sample. A well know difficulty is that each chi-square statistic tends to become large without an increase in its degrees of freedom as the sample size increases. As a consequence goodness-of-fit tests based on Pearson type chi-square statistics will generally reject the correct specification of every competing model.

To circumvent such a difficulty, a popular method for model selection, which is similar to use of Akaike [6] Information Criterion (AIC), consists in considering that the lower the chi-square statistic, the better is the model. The preceding selection rule, however, does not take into account random variations inherent in the values of the statistics.

We propose here a procedure for taking into account the stochastic nature of these differences so as to assess their significance. The main propose of this paper is to address this issue. We shall propose some convenient asymptotically standard normal tests for model selection based on φ-divergence type statistics. Following Vuong [7,8] the procedures considered here are testing the null hypothesis that the competing models are equally close to the data generating process (DGP) versus the alternative hypothesis that one model is closer to the DGP where closeness of a model is measured according to the discrepancy implicit in the φ-divergence type statistic used. Thus the outcomes of our tests provide information on the strength of the statistical evidence for the choice of a model based on its goodness-of-fit (see Ngom [9]; Diedhiou and Ngom [10]). The model selection approach roposed here differs from those of Cox [11], and Akaike [12] for non nested hypotheses. This difference is that the present approach is based on the discrepancy implicit in the divergence type statistics used, while these other approaches as Vuong's [7] tests for model selection rely on the Kullback-Leibler [13] information criterion (KLIC).

Beran [14] showed that by using the minimum Hellinger distance estimator, one can simultaneously obtain asymptotic efficiency and robustness properties in the presence of outliers. The works of Simpson [15] and

*Corresponding author.

Lindsay [16] have shown that, in the tests hypotheses, robust alternatives to the likelihood ratio test can be generated by using the Hellinger distance. We consider a general class of estimators that is very broad and contains most of estimators currently used in practice when forming divergence type statistics. This covers the case studies in Harris and Basu [17]; Basu *et al*. [18]; Basu and Basu [19] where the penalized Hellinger distance is used.

The remainder of this paper is organized as follows. Section 2 introduces the basic notations and definitions. Section 3 gives a short overview of divergence measures. Section 4 investigates the asymptotic distribution of the penalized Hellinger distance. In Section 5, some applications for testing hypotheses are proposed. Section 6 presents some simulation results. Section 7 concludes the paper.

2. Definitions and Notation

In this section, we briefly present the basic assumptions on the model and parameters estimators, and we define our generalized divergence type statistics. We consider a discrete statistical model, *i.e* $X_1, X_2, \cdots, X_n$ an independent random sample from a discrete population with support $\mathcal{X} = \{1, \cdots, m\}$. Let $\boldsymbol{P} = (p_1, \cdots, p_m)^{\mathrm{T}}$ be a probability vector *i.e.* $\boldsymbol{P} \in \Omega_m$ where Ω_m is the simplex of probability m-vectors, $m = (p_1, p_2, \cdots, p_m) \in \mathbb{R}^m$; $\sum_{i=1}^{m} p_i = 1, p_i \geq 0, i = 1, \cdots, m$.

We consider a parameter model

$$\mathcal{P} = \left\{ P_\theta = \left(p_1(\theta), \cdots, p_m(\theta) \right)^{\mathrm{T}} : \theta \in \Theta \right\}$$

which may or may not contain the true distribution $\mathcal{P}$, where Θ is a compact subset of k-dimensional Euclidean space (with $k < m - 1$). If $\mathcal{P}$contains $\mathcal{P}$, then there exists a $\theta_0 \in \Theta$ such that $P_{\theta_0} = P$ and the model $\mathcal{P}$is said to be correctly specified.

We are interested in testing $H_0 : P \in \mathcal{P}$ (with true parameter θ_0) versus H_1:$P \in \Omega_m - \mathcal{P}$.

By $\|\cdot\|$ we denote the usual Euclidean norm and we interpret probability distributions on $\mathcal{X}$ as row vectors from $\mathbb{R}^m$. For simplicity we restrict ourselves to unknown true parameters θ_0 satisfying the classical regularity conditions given by Birch [20]:

1) True θ_0 is an interior point of Θ and $p_{i\theta 0} > 0$ for $i = 1, \cdots, m$. Thus $P_{\theta_0} = (p_{i\theta 0}, \cdots, p_{m\theta 0})^{\mathrm{T}}$ is an interior point of the set Ω_m.

2) The mapping $P : \Theta \in \Omega_m$ is totally differentiable at θ_0 so that the partial derivatives of p_i with respect to each θ_j exist at θ_0 and $p_i(\theta)$ has a linear approximation at θ_0 given by

$$p_i(\theta) = p_i(\theta_0) + \sum_{j=1}^{k} (\theta_j - \theta_{0j}) \frac{\partial p_i(\theta_0)}{\partial(\theta_j)} + o\left(\|\theta - \theta_0\|\right)$$

where $o\left(\|\theta - \theta_0\|\right)$ denotes a function verifying

$$\lim_{\theta \to \theta_0} \frac{o\left(\|\theta - \theta_0\|\right)}{\|\theta - \theta_0\|} = 0.$$

3) The Jacobian matrix

$$\boldsymbol{J}(\theta_0) = \left(\frac{\partial P_\theta}{\partial \theta} \right)_{\theta = \theta_0} = \left(\frac{\partial p_i(\theta_0)}{\partial \theta_j} \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}}$$

is of full rank (*i.e*. of rank k and $k < m$).

4) The inverse mapping $P^{-1} : \mathcal{P} \to \Theta$ is continuous at P_{θ_0}.

5) The mapping $P : \Theta \to \Omega_m$ is continuous at every point $\theta \in \Theta$.

Under the hypothesis that $P \in \mathcal{P}$, there exists an unknown parameter θ_0 such that $P = P_{\theta_0}$ and the problem of point estimation appears in a natural way. Let n be sample size. We can estimate the distribution $P_{\theta_0} = (p_1(\theta), p_2(\theta), \cdots, p_m(\theta))^{\mathrm{T}}$ by the vector of observed frequencies $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_m)$ on $\mathcal{X}$ *i.e.* of measurable mapping $\mathcal{X}^n \to \Omega_m$.

This non parametric estimator $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_m)$ is defined by $\hat{p}_j = \frac{N_j}{n}$, $N_j = \sum_{i=1}^{n} T_j^i(X_i)$ where

$$T_j^i(X_i) = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

We can now define the class of φ-divergence type statistics considered in this paper.

3. A Brief Review of φ-Divergences

Many different measures quantifying the degree of discrimination between two probability distributions have been studied in the past. They are frequently called distance measures, although some of them are not strictly metrics. They have been applied to different areas, such as medical image registration (Josien P.W. Pluim [21], classification and retrieval, among others. This class of distances is referred, in the literature, as the class of φ, f or g-divergences (Csisza'r [11]; Vajda [22]; Morales *et al*. [23]; the class of disparities (Lindsay [16]). The divergence measures play an important role in statistical theory, especially in large theories of estimation and testing.

Later many papers have appeared in the literature, where divergence or entropy type measures of information have been used in testing statistical hypotheses. Among others we refer to Read and Cressie [24], Zografos *et al*. [25], Salicru' *et al*. [26], Bar-Hen and Daudin [27], Mene'ndez *et al*. [28]), Pardo *et al*. [29] and the references therein. A measure of discrimination between two probability distributions called φ-divergence, was

introduced by Csisza'r [30].

Recently, Broniatowski *et al*. [31] presented a new dual representation for divergences. Their aim was to introduce estimation and test procedures through divergence optimization for discrete or continuous parametric models. In the problem where independent samples are drawn from two different discrete populations, Basu *et al*. [32] developed some tests based on the Hellinger distance and penalized versions of it.

Consider two populations *X* and *Y*, according to classification criteria can be grouped into *m* classes species $x_1, x_2, \cdots, x_m$ and $y_1, y_2, \cdots, y_m$ with probabilities $P = (p_1, p_2, \cdots, p_m)$ and $Q = (q_1, q_2, \cdots, q_m)$ respectively. Then

$$D_\phi(P,Q) = \sum_{i=1}^{m} q_i \phi\left(\frac{p_i}{q_i}\right) \tag{3.2}$$

is the ϕ-divergence between *P* and *Q* (see Csisza'r, [30]) for every ϕ in the set Φ of real convex functions defined on $[0,\infty[$. The function $\phi(t)$ is assumed to verify the following regularity condition: $\phi : [0,\infty[\to \boldsymbol{R} \cup \{\infty\}$ is convex and continuous, where $0\phi\left(\frac{0}{0}\right) = 0$ and

$$0\phi\left(\frac{p}{0}\right) = \lim_{u\to\infty}\left(\phi(u)/\mu\right).$$

Its restriction on $[0,+\infty]$ is finite, twice continuously differentiable in a neighborhood of $u = 1$, with $\phi(1) = \phi'(1) = 0$ and $\phi''(1) = 1$ (cf. Liese and Vajda [33]).

We shall be interested also in parametric estimators

$$\hat{Q} = \hat{Q}_n = P_{\hat{\theta}} \tag{3.3}$$

of P_{θ_0} which can be obtained by means of various point estimators

$$\hat{\theta} = \hat{\theta}^{(n)} : \mathcal{X}^{(n)} \to \Theta$$

of the unknown parameter θ_0.

It is convenient to measure the difference between observed and expected frequencies P_{θ_0}. A minimum Divergence estimator of θ is a minimizer of $D_\phi\left(\hat{P}, P_{\theta_0}\right)$ where $\hat{P}$ is a nonparametric distribution estimate. In our case, where data come from a discrete distribution, the empirical distribution defined in (2.1) can be used.

In particular if we replace $\phi_1(x) = -4\left[\sqrt{x} - \frac{1}{2}(x+1)\right]$ in (3.2) we get the Hellinger distance between distribution $\hat{P}$ and P_θ given by

$$\begin{aligned} D_{\phi_1}\left(\hat{P}, P_\theta\right) &= HD_{\phi_1}\left(\hat{P}, P_\theta\right) \\ &= 2\sum_{i=1}^{m}\left(\left(\hat{p}_i^{1/2} - p_i^{1/2}\right)\right)^2 : \phi_1 \in \Phi \end{aligned} \tag{3.4}$$

Liese and Vajda [33], Lindsay [16] and Morales *et al*. [23] introduced the so-called *minimum φ-divergence estimate* defined by

$$D_\phi\left(\hat{P}, P_{\hat{\theta}}\right) = \min_{\theta\in\Theta} D_\phi\left(\hat{P}, P_\theta\right); \phi \in \Phi \tag{3.5}$$

$$\hat{\theta}_\phi = \arg\min_{\theta\in\Theta} D_\phi\left(\hat{P}, P_\theta\right); \phi \in \Phi \tag{3.6}$$

Remark 3.1. *The class of estimates* (3.4) *contains the maximum likelihood estimator* (*MLE*).

In particular if we replace $\phi = -\log x + x - 1$ *we get*

$$\begin{aligned} \hat{\theta}_{KL_m} &= \arg\min_{\theta\in\Theta} KL_m\left(P_\theta, \hat{P}\right) \\ &= \arg\min_{\theta\in\Theta} \sum_{i=1}^{m} -\log p_i(\theta)\hat{P}_i = MLE \end{aligned}$$

where KL_m *is the modified Kullback-Leibler divergence.*

Beran [14] first pointed out that the minimum Hellinger distance estimator (MHDE) of θ, defined by

$$\hat{\theta}_H = \arg\min_{\theta\in\Theta} HD_\phi\left(\hat{P}, P_\theta\right); \ \phi \in \Phi \tag{3.7}$$

has robustness proprieties.

Further results were given by Tamura and Boos [34], Simpson [15], and Basu *et al*. [35] for more details on this method of estimation. Simpson, however, noted that the small sample performance of the Hellinger deviance test at some discrete models such as the Poisson is somewhat unsatisfactory, in the sense that the test requires a very large sample size for the chi-square approximation to be useful (Simpson [15], **Table 3**). In order to avoid this problem, one possibility is to use the penalized Hellinger distance (see Harris and Basu, [36]; Basu, Basu and Basu, [19]; Basu *et al*. [32]). The penalized Hellinger distance family between the probability vectors $\hat{P}$ and P_θ is defined by:

$$\begin{aligned} &PHD^h\left(\hat{P}, P_\theta\right) \\ &= 2\left[\sum_{i\in\varpi}^{m}\left(\hat{p}_i^{\frac{1}{2}} - p_i^{\frac{1}{2}}(\theta)\right)^2 + h\sum_{i\in\varpi^C}^{m} p_i(\theta)\right] \end{aligned} \tag{3.8}$$

where *h* is a real positive number with

$$\varpi = \{i : \hat{p}_i \neq 0\} \text{ and } \varpi c = \{i : \hat{p}_i = 0\}$$

Note that when $h = 1$, this generates the ordinary Hellinger distance (Simpson, [15]).

Hence (3.7) can be written as follows

$$\hat{\theta}_{PH} = \arg\min_{\theta\in\Theta} PHD_\phi^h\left(\hat{P}, P_\theta\right) \tag{3.9}$$

One of the suggestions to use the penalized Hellinger is motivated by the fact that this suitable choice may lead to an estimate more robust than the MLE.

A model selection criterion can be designed to estimate an expected overall discrepancy, a quantity which reflects the degree of similarity between a fitted ap-

proximating model and the generating or true model. Estimation of Kullback's information (see Kullback-Leibler [13]) is the key to deriving the Akaike Information criterion AIC (Akaike [6]).

Motivated by the above developments, we propose by analogy with the approach introduced by Vuong [7,8], a new information criterion relating to the φ-divergences. In our test, the null hypothesis is that the competing models are as close to the data generating process (DGP) where closeness of a model is measured according to the discrepancy implicit in the penalized Hellinger divergence.

4. Asymptotic Distribution of the Penalized Hellinger Distance

Hereafter, we focus on asymptotic results. We assume that the true parameter θ_0 and mapping $P:\Theta\to\Omega_m$ satisfy conditions 1 - 6 of Birch [20].

We consider the m-vector $P_\theta=\left(p_{1\theta},\cdots,p_{m\theta}\right)^{\mathrm{T}}$, the $m\times k$ Jacobian matrix $\boldsymbol{J}_\theta=\left(J_{jl}(\theta)\right)_{j=1,\cdots,m\ ;\ l=1,\cdots,k}$ with $\left(J_{jl}(\theta)=\dfrac{\partial}{\partial\theta_l}p_{j\theta}\right)$ the $m\times k$ matrix

$\boldsymbol{D}_\theta=\operatorname{diag}\left(P_\theta^{-1/2}\right)J_\theta$ and the $k\times k$ Fisher information matrix

$$\boldsymbol{I}_\theta=\left(\sum_{j=1}^{m}\frac{1}{p_{j\theta}}\frac{\partial p_{j\theta}}{\partial\theta_r}\frac{\partial p_{j\theta}}{\partial\theta_s}\right)_{r,s=1,\cdots,k}=D_\theta(\theta)^{\mathrm{T}}D_\theta$$

where $\operatorname{diag}\left(P_\theta^{-1/2}\right)=\operatorname{diag}\left(\dfrac{1}{\sqrt{p_1(\theta)}},\cdots,\dfrac{1}{\sqrt{p_m(\theta)}}\right)$.

The above defined matrices are considered at the point $\theta\in\Theta$ where the derivatives exist and all the coordinates $p_j(\theta)$ are positive.

The stochastic convergences of random vectors $\boldsymbol{X}_n$ to a random vector $\boldsymbol{X}$ are denoted by $\boldsymbol{X}_n\xrightarrow{P}\boldsymbol{X}$ and $\boldsymbol{X}_n\xrightarrow{L}\boldsymbol{X}$ (convergences in probability and in law, respectively). Instead $c_n\boldsymbol{X}_n\xrightarrow{P}0$ for a sequence of positive numbers c_n we can write $\|\boldsymbol{X}\|=o_p c_n^{-1}$.

This relation means:

$$\lim x\to\infty\ \limsup x\to\infty\ \mathbb{P}\left(\|c_nX_n\|>x\right)=0.$$

An estimator $\hat{P}$ of P_{θ_0} is consistent if for every $\theta_0\in\Theta$ the random vector $\left(\hat{p}_1,\cdots,\hat{p}_m\right)$ tends in probability to $\left(p_{1\theta_0},\cdots,p_{m\theta_0}\right)$, *i.e.* if

$$\lim_{x\to\infty}\mathbb{P}\left(\left\|\hat{P}-P_{\theta_0}\right\|>\varepsilon\right)=0\text{ for all }\varepsilon>0$$

We need the following result to prove Theorem 4.3.

Proposition 4.1. (*Mandal et al.* [37])

Let $\phi\in\Phi$, let p:$\Theta\to\Omega m$ be twice continuously differentiable in a neighborhood of θ_0 and assume that conditions 1 - 5 *of Section* 2 *hold. Suppose that $\boldsymbol{I}_{\theta_0}$ is the $k\times k$ Fisher Information matrix and $\hat{\theta}_{PH}$ satisfying* (3.7).

Then the limiting distribution of $\sqrt{\boldsymbol{n}}\left(\hat{\theta}_{PH}-\theta_0\right)$ as $n\to+\infty$ is $N\left[0,I_{\theta_0}^{-1}\right]$.

Lemma 4.2. *We have*

$$\sqrt{\boldsymbol{n}}\left(\hat{\theta}_{PH}-\theta_0\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left[0,\sum\nolimits_{P_{\theta_0}}\right]$$

where $\hat{P}(\theta_0)=\left(\hat{P}_1\theta_0,\cdots,\hat{P}_m\theta_0\right)$ an estimator of $P_{\theta_0}=\left(p_{1\theta_0},\cdots,p_{m\theta_0}\right)$ defined in (2.1) *with*

$$\Sigma_{P_{\theta_0}}=\operatorname{diag}\left(P_{\theta_0}\right)-P_{\theta_0}P_{\theta_0}^{\mathrm{T}}$$

Proof. Denote

$$V=\left[\frac{N_1-np_{1\theta_0}}{\sqrt{n}},\cdots,\frac{N_m-np_{1m}}{\sqrt{n}}\right]$$

and $N_j=\sum_i^n T_j^i$ where

$$T_j^i(X_i)=\begin{cases}1 & \text{si } X_i=j\\ 0 & \text{otherwise}\end{cases}$$

$$V=\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}T_1^i-np_{1\theta_0}\right);\cdots;\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}T_m^i-np_{m\theta_0}\right)\right)$$

$$=\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}T_1^i-p_{1\theta_0}\right);\cdots;\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}T_m^i-p_{m\theta_0}\right)\right.$$

and applying the Central Limit Theorem we have

$$\left(\frac{\boldsymbol{N}_1-\boldsymbol{n}p_{1\theta_0}}{\sqrt{\boldsymbol{n}}},\cdots,\frac{\boldsymbol{N}_m-\boldsymbol{n}p_{m\theta_0}}{\sqrt{\boldsymbol{n}}}\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left[0,\Sigma_{P_{\theta_0}}\right]$$

where

$$\Sigma_{P_{\theta_0}}=\operatorname{diag}\left(P_{\theta_0}\right)-P_{\theta_0}P_{\theta_0}^{\mathrm{T}}.\tag{4.10}$$

For simplicity, we write $D_H^h\left(\hat{P},P_{\hat{\theta}PH}\right)$ instead $PHD^h\left(\hat{P},P_{\hat{\theta}PH}\right)$.

Theorem 4.3. *Under the assumptions of Proposition* (4.1), *we have*

$$\sqrt{n}\left(\hat{P},P_{\hat{\theta}PH}\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left[0,\Lambda_{\theta_0}\right].$$

where

$$\Lambda_{\theta_0}=\Sigma_{\theta_0}-\Sigma_{\theta_0}M_{\theta_0}^{\mathrm{T}}-M_{\theta_0}\Sigma_{\theta_0}+M_{\theta_0}\Sigma_{\theta_0}M_{\theta_0}^{\mathrm{T}}$$

$$M_{\theta_0}=J_\theta I_{\theta_0}^{-1}(\theta_0)^{\mathrm{T}}\operatorname{diag}\left(P_{\theta_0}^{1/2}\right)$$

$$\Sigma_{\theta_0}=\Sigma_{P_{\theta_0}}\tag{4.11}$$

Proof. A first order Taylor expansion gives

$$P_{\hat{\theta}PH} = P_{\theta_0} + J_{\theta_0}\left(\hat{\theta}_{PH} - \theta_0\right)^{\mathrm{T}} + o\left(\left\|\hat{\theta}_{PH} - \theta_0\right\|\right) \quad (4.12)$$

In the same way as in Morales *et al*. [28], it can be established that:

$$\hat{\theta}_{PH} = \theta_0 + I_{\theta_0}^{-1} D_{\theta_0}^{\mathrm{T}} \operatorname{diag}\left[P_{\theta_0}^{-1/2}\right]\left(\hat{P} - P_{\theta_0}\right)^{\mathrm{T}} + o\left(\left\|\hat{P} - P_{\theta_0}\right\|\right) \quad (4.13)$$

From (4.12) and (4.13) we obtain

$$P_{\hat{\theta}PH} = P_{\theta_0} + J_{\theta_0} I^{-1}\left(\theta_0\right) D_{\theta_0}^{\mathrm{T}} \operatorname{diag}\left[P_{\theta_0}^{-1/2}\right]\left(\hat{P} - P_{\theta_0}\right)^{\mathrm{T}} + o\left(\hat{P} - P_{\theta_0}\right)$$

therefore the random vectors

$$\begin{bmatrix} \hat{P} - P_{\theta_0} \\ P_{\hat{\theta}_{PH}} - P_{\theta_0} \end{bmatrix}_{2m\times 1} \text{ and } \begin{bmatrix} \boldsymbol{I} \\ M_{\theta_0} \end{bmatrix}_{2m\times m} \times \left(\hat{P} - P_{\theta_0}\right)_{m\times 1}$$

where $\boldsymbol{I}$ is the $m \times m$ unity matrix, have the same asymptotic distribution.

Furthermore it is clear (applying TCL) that

$$\sqrt{n}\left(\hat{P} - P_{\hat{\theta}PH}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \Sigma_{\theta_0}\right]$$

Being Σ_{θ_0} the $m \times m$ matrix $\operatorname{diag}\left(P_{\theta_0}\right) - P_{\theta_0} P_{\theta_0}^{\mathrm{T}}$. implies

$$\sqrt{n}\begin{bmatrix} \hat{P} - P_{\theta_0} \\ P_{\hat{\theta}_{PH}} - P_{\theta_0} \end{bmatrix}_{2m\times 1} \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \begin{pmatrix} I \\ M_{\theta_0} \end{pmatrix} \Sigma_{\theta_0}\left(I, M_{\theta_0}^{\mathrm{T}}\right)\right]$$

therefore, we get

$$\sqrt{n}\left(\hat{P} - P_{\hat{\theta}PH}\right) = \sqrt{n}\left(\hat{P} - P_{\theta_0}\right) + \sqrt{n}\left(P_{\theta_0} - P_{\hat{\theta}PH}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \Lambda\left(\theta_0\right)\right] \quad (4.14)$$

$$\Lambda_{\theta_0} = \Sigma_{\theta_0} - \Sigma_{\theta_0} M_{\theta_0}^{\mathrm{T}} - M_{\theta_0}\Sigma_{\theta_0} + M_{\theta_0}\Sigma_{\theta_0} M_{\theta_0}^{\mathrm{T}}$$ □

The case which is interest to us here is to test the hypothesis H_0:$P \in P$. Our proposal is based on the following penalized divergence test statistic $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right)$ where $\hat{P}$ and $P_{\hat{\theta}_{PH}}$ have been introduce in Theorem (4.3) and (3.7) respectively.

Using arguments similar to those developed by Basu [17], under the assumptions of (4.3) and the hypothesis H_0:$P = P_\theta$, the asymptotic distribution of $2nD_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right)$ is a chi-square when $h = 1$ with derees $m - k - 1$ degrees of freedom. Since the others members of penalized Hellinger distance tests differ from the ordinary Hellinger distance test only at the empty cells, they too have the same asymptotic distribution.

Considering now the case when the model is wrong *i.e.* H_1:$P \neq P_\theta$. We introduce the following regularity assumptions

(A_1) There exists $\theta_1 = \arg\inf_{\theta\in\Theta} PHD^h\left(P, P_\theta\right)$ such that:

$$P_{\hat{\theta}PH} \overset{as}{\mapsto} P_{\theta_1} \quad \text{when } n \to +\infty$$

(A_2) There exists $\theta_1 \in \Theta$; $\Lambda^* = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$, with $\boldsymbol{\Lambda}_{11} = \Sigma_p$ in (4.10) and $\boldsymbol{\Lambda}_{12} = \boldsymbol{\Lambda}_{21}$ such that

$$\sqrt{n}\begin{pmatrix} \hat{P} - P_{\theta_0} \\ P_{\hat{\theta}_{PH}} - P_{\theta_0} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \Lambda^*\right].$$

Theorem 4.4. *Under H*:$P \neq P_\theta$ *and assume that conditions* (A_1) *and* (A_2) *hold, we have*:

$$\sqrt{n}\left(D_H^h\left(\hat{P} - P_{\hat{\theta}PH}\right) - D_H^h\left(P, P_{\theta_1}\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \Omega_{(\theta,P)}^2\right]$$

where

$$\Omega_{(\theta,P)}^2 = H^{\mathrm{T}}\Lambda_{11}H + H^{\mathrm{T}}\Lambda_{12}J + J^{\mathrm{T}}\Lambda_{12}H + J^{\mathrm{T}}\Lambda_{22}J \quad (4.15)$$

$H^{\mathrm{T}} = \left(h_1, \cdots, h_m\right)$ *with*

$$h_i = \left(\frac{\partial}{\partial p_i^1} D_H^h\left(p^1, p^2\right)\right)_{p^1 = p, p^2 = p(\theta_1)}, i = 1, \cdots, m$$

And $J^{\mathrm{T}} = \left(j_1, \cdots, j_m\right)$ *with*

$$j_i = \left(\frac{\partial}{\partial p_i^2} D_H^h\left(p^1, p^2\right)\right)_{p^1 = p.p^2 = p(\theta_1)}, i = 1, \cdots, m$$

Proof. A first order Taylor expansion gives

$$D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) = D_H^h\left(P, P_{\theta_1}\right) + H^{\mathrm{T}}\left(\hat{P} - P\right) + J^{\mathrm{T}}\left(P_{\hat{\theta}_{PH}} - P_{\theta_1}\right) + o\left(\left\|\hat{P} - P\right\| + \left\|P_{\hat{\theta}_{PH}} - P_{\theta_1}\right\|\right) \quad (4.16)$$

From the assumed assumptions (A_1) and (A_2), the result follows. □

5. Applications for Testing Hypothesis

The estimate $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right)$ can be used to perform statistical tests.

5.1. Test of Goodness-Fit

For completeness, we look at $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right)$ in the usual way, *i.e.* as a goodness-of-fit statistic. Recall that here θ_{PH} is the minimum penalized Hellinger distance estimator of θ. Since $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right)$ is a consistent estimator

of $D_H^h\left(P,P_\theta\right)$, the null hypothesis when using the statistic $D_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)$ is $H_o: D_H^h\left(P,P_\theta\right)=0$ or equivalently, H_o:$P=P_\theta$.

Hence, if H_o is rejected so that one can infer that the parametric model P_θ is misspecified. Since $D_H^h\left(P,P_\theta\right)$ is non-negative and takes value zero only when $P=P_\theta$, the tests are defined through the critical region.

$$C_{\theta_{PH}}=\left\{2nD_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)>q_{\alpha,k}\right\}$$

where $q_{\alpha,k}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with $m-k-1$ degrees of freedom.

Remark 5.1. *Theorem* (4.4) *can be used to give the following approximation to the power of test*

$$H_o: D_H^h\left(P,P_\theta\right)=0.$$

Approximated power function is

$$\begin{aligned}\beta_{(P)}&=\mathbb{P}\left[2nD_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)>q_{\alpha,k}\right]\\&\approx 1-\mathcal{F}_n\left(\frac{q_{\alpha,k}-2nD_H^h\left(P,P_\theta\right)}{2\sqrt{n_{(\theta,P)}}}\right)\end{aligned}\qquad(5.17)$$

where $q_{\alpha,k}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with $m-k-1$ degrees of freedom and $\mathcal{F}_n$ is a sequence of distribution function tending uniformily to the standard normal distribution $\mathcal{F}(x)$. Note that if $H_o: D_H^h\left(P,P_\theta\right)\neq 0$, then for any fixed size α the probability of rejection $H_o: D_H^h\left(P,P_\theta\right)=0$ with the rejection rule $2nD_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)>q_{\alpha,k}$ tends to one as $n\to\infty$.

Obtaining the approximate sample n, guaranteeing a power β for a give alternative P, is an interesting application of Formula (5.17). If we wish the power to be equal to β^*, we must solve the equation

$$\beta^*=1-\mathcal{F}\left[\frac{\sqrt{n}}{\Omega_{(\theta,P)}}\left(\frac{1}{2n}q_{\alpha,k}-D_H^h\left(P,P_\theta\right)\right)\right].$$

It is not difficult to check that the sample size n^*, is the solution of the following equation

$$\begin{aligned}&n^2D_H^h\left(P,P_\theta\right)^2-nD_H^h\left(P,P_\theta\right)q_{\alpha,k}+\left(\frac{q_{\alpha,k}}{2}\right)^2\\&=n\Omega_{(\theta,P)}^2\left[\mathcal{F}^{-1}\left(1-\beta^*\right)\right]^2\end{aligned}$$

The solution is given by

$$n^*=\left(\frac{(a+b)-\sqrt{a(a+2b)}}{2D_H^h\left(P,P_\theta\right)^2}\right)$$

with $a=\Omega_{(\theta,P)}^2\left[\mathcal{F}^{-1}\left(1-\beta\right)\right]^2$ and $b=q_{\alpha,k}D_H^h\left(P,P_\theta\right)$ and the required size is $n_0=\left[n^*\right]+1$, where $[\cdot]$ denotes "integer part of".

5.2. Test for Model Selection

As we mentioned above, when one chooses a particular ϕ-divergence type statistic

$$D_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)=PHD_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)$$

with $\hat{\theta}_{PH}$ the corresponding minimum penalized Hellinger distance estimator of θ, one actually evaluates the goodness-of-fit of the parametric model P_θ according to the discrepancy $D_H^h\left(P,P_\theta\right)$ between the true distribution P and the specified model P_θ. Thus it is natural to define the best model among a collection of competing models to be the model that is closest to the true distribution according to the discrepancy $D_H^h\left(P,P_\theta\right)$.

In this paper we consider the problem of selecting between two models. Let $G_\mu=\{G(.|\mu;\mu\in\Gamma\}$ be another model, where Γ is a q-dimensional parametric models P_θ. In a similar way, we can define the minimum penalized Hellinger distance estimator of μ and the corresponding discrepancy $D_H^h\left(P,G_\mu\right)$ for the model G_μ.

Our special interest is the situation in which a researcher has two competing parametric models P_θ and G_μ, and he wishes to select the better of two models based on their discrimination statistic between the observations and models P_θ and G_μ, defined respectively by $D_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)$ and $D_H^h\left(\hat{P},P_{\hat{\mu}_{PH}}\right)$.

Let the two competing parametric models P_θ and G_μ with the given discrepancy $D_H^h\left(P,.\right)$.

Définition 5.2.

$H_0^{eq}: D_H^h\left(P,P_\theta\right)=D_H^h\left(P,P_\mu\right)$ *means that the two models are equivalent*,

$H_{p_\theta}: D_H^h\left(P,P_\theta\right)<D_H^h\left(P,P_\mu\right)$ *means that* P_θ *is better than* G_μ,

$H_{G_\mu}: D_H^h\left(P,P_\theta\right)<D_H^h\left(P,P_\mu\right)$ *means that* P_θ *is worse than* G_μ.

Remark 5.3. 1) *It does not require that the same divergence type statistics be used in forming* $D_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)$ *and* $D_H^h\left(\hat{P},P_{\hat{\mu}_{PH}}\right)$. *Choosing, however, different discrepancy for evaluating competing models is hardly jutified.*

2) *This definition does not require that either of the competing models be correctly specified. On the other hand, a correctly specified model must be at least as good as any other model.*

The following expression of the indicator $D_H^h\left(P,P_\theta\right)-D_H^h\left(P,P_\mu\right)$ is unknown, but from the previous section, it can be estimated by the difference

$$\sqrt{n}\left[D_H^h\left(\hat{P},P_{\hat{\theta}_{PH}}\right)-D_H^h\left(\hat{P},P_{\hat{\mu}_{PH}}\right)\right]$$

This difference converges to zero under the null hypo-

thesis H_0^{eq}, but converges to a strictly negative or positive constant when H_{P_θ} and H_{G_μ} holds.

These properties actually justify the use of $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right)$ as a model selection indicator and common procedure of selecting the model with highest goodness-of-fit.

As argued in the introduction, however, it is important to take into account the random nature of the difference $D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right)$ so as to assess its significance. To do so we consider the asymptotic distribution of $\sqrt{n}\left[D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right)\right]$ under H_0^{eq}.

Our major task is to propose some tests for model selection, *i.e*. for the null hypothesis H_0^{eq} against the alternative H_{P_θ} or H_{G_μ}. We use the next lemma with $\hat{\theta}_{PH}$ an d $\hat{\mu}_{PH}$ as the corresponding minimum penalized Hellinger distance estimator of θ and μ.

Using P and P_θ defined earlier, we consider the vector $K_\theta^{\mathrm{T}} = (k_1, \cdots, k_m)$ where

$$k_i = \left(\frac{\partial}{\partial p_i^1} D_H^h\left(P^1, P^2\right)\right)_{P^1=P, P^2=P_\theta} \quad \text{with } i = 1, \cdots, m$$

$$Q_\theta^{\mathrm{T}} = (q_1, \cdots, q_m)$$

where

$$q_i = \left(\frac{\partial}{\partial p_i^2} D_H^h\left(P^1, P^2\right)\right)_{P^1=P, P^2=P_\theta} \quad \text{with } i = 1, \cdots, m$$

Lemma 5.4. *Under the assumptions of the Theorem* (4.4), *we have*

(*i*) *for the model* P_θ

$$D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(P, P_\theta\right) + K_\theta^{\mathrm{T}}\left(\hat{P}, P\right) + Q_\theta^{\mathrm{T}}\left(P_{\hat{\theta}_{PH}} - P_\theta\right) + o_p(1)$$

(*ii*) *for model* G_μ

$$D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right) = D_H^h\left(P, G_\mu\right) + K_\theta^{\mathrm{T}}\left(\hat{P}, P\right) + Q_\theta^{\mathrm{T}}\left(G_{\hat{\mu}_{PH}} - G_\mu\right) + o_p(1)$$

Proof.

The results follow from a first order Taylor expansion.

We define

$$\Gamma^2 = \left(K_\theta - K_\mu; Q_\theta - Q_\mu\right)^{\mathrm{T}} \ \Lambda^* \ \left(K_\theta - K_\mu; Q_\theta - Q_\mu\right)$$

which is the variance of

$$\left(K_\theta - K_\mu; Q_\theta - Q_\mu\right)^{\mathrm{T}} \begin{pmatrix} \hat{P} & P \\ P_{\theta_{PH}} & P_{\theta_1} \end{pmatrix}.$$

Since K_θ, K_μ, Q_θ, Q_μ, and Λ^* are consistently estimated by their sample analogues $K_{\hat{\theta}}$, $K_{\hat{\mu}}$, $Q_{\hat{\theta}}$, $Q_{\hat{\mu}}$ and Λ^*, hence Γ^2 is consistently estimated by

$$\hat{\Gamma}^2 = \left(K_{\hat{\theta}} - K_{\hat{\mu}}; Q_{\hat{\theta}} - Q_{\hat{\mu}}\right)^{\mathrm{T}} \ \hat{\Lambda}^* \ \left(K_{\hat{\theta}} - K_{\hat{\mu}}; Q_{\hat{\theta}} - Q_{\hat{\mu}}\right)$$

Next we define the model selection statistic and its asymptotic distribution under the null and alternative hypothesis.

Let

$$HI^h = \frac{\sqrt{n}}{\hat{\Gamma}}\left\{D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right)\right\}$$

where HI^h stands for the penalized Hellinger Indicator.

The following theorem provides the limit distribution of HI^h under the null and alternatives hypothesis.

Theorem 5.5. *Under the assumptions of Theorem* (4.4), *suppose that* $\Gamma \neq 0$, *then*

1) *Under the null hypothesis* H_0^{eq}, $HI^h \xrightarrow{\mathcal{L}} \mathcal{N}[0,1]$.

2) *Under the null hypothesis* $H_{P_\theta} \to -\infty$ *in probability*.

3) *Under the null hypothesis* $H_{G_\mu} \to +\infty$ *in probability*.

Proof.

From the Lemma (5.4), it follows that

$$\begin{aligned} &D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right) \\ &= D_H^h\left(P, P_\theta\right) - D_H^h\left(P, G_\mu\right) + K_\theta^{\mathrm{T}}\left(\hat{P} - P\right) - K_\mu^{\mathrm{T}}\left(\hat{P} - P\right) \\ &\quad + Q_\theta^{\mathrm{T}}\left(P_{\hat{\theta}_{PH}} - P_\theta\right) - Q_\mu^{\mathrm{T}}\left(G_{\hat{\mu}_{PH}} - G_\mu\right) + o_p(1) \end{aligned}$$

Under $H_0^{eq}: P_\theta = G_\mu$ and $P_{\hat{\theta}_{PH}} = G_{\hat{\mu}_{PH}}$ we get

$$\begin{aligned} &D_H^h\left(\hat{P}, P_{\hat{\theta}_{PH}}\right) - D_H^h\left(\hat{P}, P_{\hat{\mu}_{PH}}\right) \\ &= K_\theta^{\mathrm{T}}\left(\hat{P} - P\right) - K_\mu^{\mathrm{T}}\left(\hat{P} - P\right) \\ &\quad + Q_\theta^{\mathrm{T}}\left(P_{\hat{\theta}_{PH}} - P_\theta\right) - Q_\mu^{\mathrm{T}}\left(P_{\hat{\theta}_{PH}} - P_\theta\right) + o_p(1) \\ &= \left(K_\theta - K_\mu; Q_\theta - Q_\mu\right)^{\mathrm{T}} \begin{pmatrix} \hat{P} & P \\ P_{\theta_{PH}} & P_{\theta_1} \end{pmatrix} + o_p(1) \end{aligned}$$

Finally, applying the Central Limit Theorem and assumptions (A_1) and (A_2), we can now immediately obtain $HI^h \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. □

6. Computational Results

6.1. Example

To illustrate the model procedure discussed in the preceding section, we consider an example. We need to define the competing models, the estimation method used for each competing model and the Hellinger penalized penalized type statistic to measure the departure of each proposed parametric model from the true data generating process.

For our competing models, we consider the problem of choosing between the family of Poisson distribution and

the family of Geometric distribution. The Poisson distribution $P(\lambda)$ is parameterized by λ and has density

$$f(x,\lambda)=\frac{\exp(-\lambda)\times\lambda^{x}}{x!} \text{ for } x\in\mathbb{N}$$

and zero otherwise.

The Geometric distribution $G(p)$ is parameterized by p and has density

$$G(x,p)=(1-p)^{x-1}\times p \text{ for } x\in\mathbb{N}^{*}$$

and zero otherwise. We use the minimum penalized Hellinger distance statistic to evaluate the discrepancy of the proposed model from the true data generating process.

We partition the real line into m intervals $\{[C_{i-1},C_i],i=1,\cdots,m\}$ where $C_0=0$ and $C_m=+\infty$. The choice of the cells is discussed below.

The corresponding minimum penalized Hellinger distance estimator of λ and p are:

$$\hat{\lambda}_{PH}=\arg\min_{\lambda\in\Theta} D_H^h\left(\hat{P},P_\lambda\right)$$
$$=\arg\min_{\lambda\in\Theta}\left[\sum_{i\in\varpi}^{m}\left(f_i^{1/2}-p_{i\lambda}^{1/2}\right)^2+\sum_{i\in\varpi^c}^{m}p_{i\lambda}\right]$$

$$\hat{p}_{PH}=\arg\min_{\lambda\in\Theta} D_H^h\left(\hat{P},P_p\right)$$
$$=\arg\min_{\lambda\in\Theta}\left[\sum_{i\in\varpi}^{m}\left(f_i^{1/2}-p_{i\lambda}^{1/2}\right)^2+h\sum_{i\in\varpi^c}^{m}p_{ip}\right]$$

$p_{i\lambda}$ and p_{ip} are probabilities of the cells $[C_{i-1},C_i]$ under the Poisson and Geometric true distribution respectively.

We consider various sets of experiments in which data are generated from the mixture of a Poisson and Geometric distribution. These two distributions are mixture of a Poisson and Geometric distribution. These two distributions are calibrated so that their two means are close (4 and 5 respectively). Hence the DGP (Data Generating Process) is generated from $M(\pi)$ with the density

$$m(\pi)=\pi\, Pois(4)+(1-\pi)\, Geom(0.2)$$

where π ($\pi\in[0,1]$ is specific value to each set of experiments. In each set of experiment several random sample are drawn from this mixture of distributions. The sample size varies from 20 to 300, and for each sample size the number of replication is 1000. In each set of experiment, we choose two values of the parameter h = 1 and $h=1/2$, where h = 1 corresponds to the classic Hellinger distance. The aim is to compare the accuracy of the selection model depending on the parameter setting chosen. In order a perfect fit by the proposed method, for the chosen parameters of these two distributions, we note that most of the mass is concentrated between 0 and 10. Therefore, the chosen partition has eight cells defined by $\{[C_{i-1},C_i]=[i-1,i],i=1,\cdots,7\}$ and $[C_7,C_8]=[7,+\infty]$ represents the last cell. We choose different values of π which are 0.00, 0.25, 0.535, 0.75, 1.00.

Although our proposed model selection procedure does not require that the data generating process belong to either of the competing models, we consider the two limiting cases π = 1.00 and π = 0.00 for they correspond to the correctly specified cases. To investigate the case where both competing models are misspecified but not at equal distance from the DGP, we consider the case π = 0.25, π = 0.75 and π = 0.5 second case is interpreted similarly as a Geometric slightly contaminated by a Poisson distribution. The former case correspond to a DGP which is Poisson but slightly contaminated by a Geometric distribution. In the last case, π = 0.535 is the value for which the Poisson $D_H^h\left(\hat{P},G_{\hat{\lambda}_{PH}}\right)$ and the Geometric $D_H^h\left(\hat{P},G_{\hat{p}_{PH}}\right)$ family are approximatively at equal distance to the mixture $m(\pi)$ according to the penalized Hellinger distance with the above cells.

Thus this set of experiments corresponds approximatively to the null hypothesis of our proposed model selection test $\mathcal{HI}^h$. The results of our different sets of experiments are presented in **Tables 1-5**. The first half of each table gives the average values of the minimum penalized Hellinger distance estimator $\hat{\lambda}_{PH}$ and $\hat{p}_{PH}$, the penalized Hellinger goodness-of-fit statistics $D_H^h\left(\hat{P},G_{\hat{\lambda}_{PH}}\right)$ and $D_H^h\left(\hat{P},G_{\hat{p}_{PH}}\right)$, and the Hellinger indicator statistics $\mathcal{HI}^h$. The values in parentheses are standard errors. The second half of each table gives in percentage the number of times our proposed model selection procedure based on $\mathcal{HI}^h$ favors the Poisson model, the Geometric model, and indecisive.

The tests are conducted at 5% nominal significance level. In the first two sets of experiments (π = 0.00 and π = 1.00) where one model is correctly specified, we use the labels "*correct*, *incorrect*" and "*indecisive*" when a choice is made. The first halves of **Tables 1-5** confirm our asymptotic results.

They all show that the minimum penalized Hellinger estimators $\hat{\lambda}_{PH}$ and $\hat{p}_{PH}$ converge to their pseudo-true values in the misspecified cases and to their true values in the correctly specified cases as the sample size increases. With respect to our $\mathcal{HI}^h$, its diverges to $-\infty$ or $+\infty$ at the approximate rate of $\sqrt{n}$ except in the **Table 5**. In the latter case the $\mathcal{HI}^h$ statistic converges, as expected, to zero which is the mean of the asymptotic $N(0,1)$ distribution under our null hypothesis of equivalence.

With the exception of **Tables 1** and **2**, we observed a large percentage of incorrect decisions. This is because both models are now incorrectly specified. In contrast, turning to the second halves of the **Tables 1** and **2**, we first note that the percentage of correct choices using $\mathcal{HI}^h$ statistic steadily increases and ultimately converges to 100%.

Table 1. DGP = Pois(4).

n		20	30	40	50	300
$\hat{P}$		**0.210(0.03)**	**0.195(0.03)**	**0.197(0.02)**	**0.205(0.02)**	**0.201(0.01)**
λ		3.950(0.40)	4.090(0.4)	4.015(0.31)	4.015(0.28)	4.0115(0.13)
DHP (Pois)	$h = 1$	0.133(0.07)	0.081(0.05)	0.059(0.03)	0.042(0.03)	0.037(0.01)
	$h = 1/2$	0.096(0.04)	0.064(0.03)	0.048(0.02)	0.034(0.02)	0.03(0.01)
DHP (Geom)	$h = 1$	0.391(0.28)	0.348(0.12)	0.208(0.09)	0.282(0.10)	0.271(0.05)
	$h = 1/2$	0.278(0.07)	0.262(0.08)	0.242(0.06)	0.236(0.06)	0.231(0.03)
$\mathcal{HI}^h$	$h = 1/2$	–3.67(2.14)	–4.32(2.69)	–4.34(2.38)	–4.83(2.52)	–4.97(2.18)
	Correct	77%	87%	92%	96%	100%
	Indecisive	23%	13%	08%	04%	00%
	Incorrect	00%	00%	00%	00%	00%
$\mathcal{HI}^h$	$h = 1$	–3.61(3.03)	–3.98(2.48)	–3.73(2.29)	–4.16(2.35)	–4.25(1.87)
	Correct	70%	79%	83%	86%	93%
	Indecisive	30%	21%	17%	17%	07%
	Incorrect	00%	00%	00%	00%	00%

Table 2. DGP = Geom(0.2).

n		20	30	40	50	300
$\hat{P}$		0.196(0.04)	0.213(0.03)	0.203(0.02)	0.203(0.02)	0.201(0.01)
λ		3.920(1.0)	4.206(0.89)	4.109(0.67)	4.009(0.58)	4.035(0.34)
DHP (Pois)	$h = 1.0$	0.356(0.14)	0.309(0.10)	0.271(0.09)	0.253(0.08)	0.244(0.07)
	$h = 0.5$	0.281(0.1)	0.273(0.07)	0.254(0.07)	0.246(0.07)	0.237(0.02)
DHP (Geom)	$h = 1$	0.150(0.06)	0.089(0.05)	0.053(0.03)	0.039(0.02)	0.033(0.01)
	$h = 1/2$	0.103(0.04)	0.267(0.03)	0.044(0.02)	0.035(0.02)	0.027(0.98)
$\mathcal{HI}^h$	$h = 1/2$	1.880(1.43)	2.560(1.37)	3.020(1.25)	3.340(1.14)	3.40(1.03)
	Correct	36%	62%	77%	84%	92%
	Indecisive	64%	38%	23%	16%	08%
	Incorrect	00%	00%	00%	00%	00%
$\mathcal{HI}^h$	h = 1	1.710(1.07)	2.260(1.05)	2.760(0.96)	3.01(0.65)	4.19(0.32)
	Correct	36%	62%	77%	84%	92%
	Indecisive	64%	38%	23%	16%	08%
	Incorrect	00%	00%	00%	00%	00%

Table 3. DGP = 0.75 × Geom(0.2) + 0.25 × Pois(4).

n		20	30	40	50	300
$\hat{P}$		0.213(0.13)	0.197(0.12)	0.208(0.08)	0.202(0.05)	0.202(0.01)
λ		4.160(0.72)	3.910(0.55)	4.180(0.55)	3.970(0.43)	4.022(0.21)
DHP (Pois)	$h = 1$	0.546(0.13)	0.472(0.1)	0.412(0.09)	0.402(0.08)	0.367(0.06)
	$h = 1/2$	0.344(0.07)	0.340(0.05)	0.320(0.05)	0.311(0.05)	0.304(0.03)
DHP (Geom)	$h = 1$	0.150(0.06)	0.089(0.05)	0.053(0.03)	0.039(0.02)	0.033(0.01)
	$h = 1/2$	–3.67(2.62)	–4.32(2.53)	–4.34(2.47)	–4.83(2.27)	–5.37(2.01)
$\mathcal{HI}^h$	$h = 1/2$	1.220(1.02)	1.820(0.89)	2.080(1.12)	2.370(0.99)	3.102(0.84)
	Geom	23%	40%	50%	64%	81%
	Indecisive	77%	60%	50%	36%	19%
	Pois	00%	00%	00%	00%	00%
$\mathcal{HI}^h$	$h = 1$	0.840(1.29)	0.831(1.27)	0.845(1.16)	0.967(1.05)	1.131(0.78)
	Geom	17%	15%	19%	22%	33%
	Indecisive	80%	83%	89%	77%	66%
	Pois	03%	02%	02%	01%	01%

Table 4. DGP = 0.75 × Pois(4) + 0.25 × Geom(0.2).

n		20	30	40	50	300
$\hat{P}$		0.213(0.03)	0.212(0.03)	0.210(0.02)	0.206(0.02)	0.203(0.01)
λ		4.110(0.43)	4.090(0.31)	3.970(0.28)	4.020(0.26)	4.019(0.17)
DHP (Pois)	$h = 1$	1.779(0.45)	1.634(0.30)	1.650(0.28)	1.570(0.24)	1.520(0.21)
	$h = 1/2$	1.443(0.24)	1.473(0.21)	1.520(0.20)	1.500(0.18)	1.483(0.14)
DHP (Geom)	$h = 1$	2.055(0.35)	1.870(0.25)	0.053(0.03)	0.039(0.02)	0.033(0.01)
	$h = 1/2$	1.640(0.15)	1.660(0.15)	1.700(0.14)	1.690(0.13)	1.632(0.10)
$\mathcal{HI}^h$	$h = 1/2$	–2.40(1.27)	–2.44(1.1)	–2.49(1.08)	–2.77(1.01)	–2.89(0.92)
	Geom	00%	00%	00%	00%	00%
	Indecisive	38%	37%	32%	27%	21%
	Pois	62%	63%	68%	83%	79%
$\mathcal{HI}^h$	$h = 1$	–2.18(1.37)	–2.37(1.33)	2.31(1.16)	–2.66(1.18)	–2.83(1.06)
	Geom	00%	00%	00%	00%	00%
	Indecisive	48%	45%	46%	30%	24%
	Pois	52%	55%	54%	70%	76%

Table 5. DGP = 0.535 × Pois(4) + 0.465 × Geom(0.2).

n		20	30	40	50	300
$\hat{P}$		0.196(0.06)	0.204(0.05)	0.211(0.03)	0.213(0.207)	0.204(0.01)
λ		3.968(0.61)	3.962(0.46)	3.981(0.374)	4.023(0.309)	4.011(0.11)
DHP (Pois)	$h = 1$	2.869(0.63)	2.600(0.46)	2.582(0.36)	2.525(0.38)	2.311(0.25)
	$h = 1/2$	2.633(0.30)	2.492(0.28)	2.369(0.27)	2.302(0.26)	21.142(0.17)
DHP (Geom)	$h = 1$	2.867(0.52)	2.682(0.37)	2.553(0.30)	2.495(0.26)	2.237(0.12)
	$h = 1/2$	2.157(0.21)	2.200(0.20)	2.263(0.20)	2.287(0.19)	2.237(0.12)
$\mathcal{HI}^h$	$h = 1/2$	–0.079(1.04)	0.038(1.05)	0.182(0.99)	0.334(1.10)	0.442(0.67)
	Geom	03%	04%	05%	10%	13%
	Indecisive	92%	92%	93%	88%	88%
	Pois	05%	04%	02%	02%	01%
$\mathcal{HI}^h$	$h = 1$	0.186(1.14)	0.248(1.64)	0.378(0.90)	0.452(0.86)	0.617(0.73)
	Geom	05%	06%	04%	09%	11%
	Indecisive	92%	90%	95%	90%	88%
	Pois	03%	04%	01%	01%	01%

The preceding comments for the second halves of **Tables 1** and **2** also apply to the second halves of **Tables 3** and **4**.

In all **Tables 1-4**, the results confirm, in small samples, the relative domination of the model selection procedure based on the penalized Hellinger statistic test ($h = 1/2$) than the other corresponding to the choice of classical Hellinger statistic test ($h = 1$), in percentages of correct decisions. **Table 5** also confirms our asymptotics results: as sample size increases, the percentage of rejection of both models converges, as it should, to 100%.

In **Figures 1**, **3**, **5**, **7** and **9** we plot the histogram of datasets and overlay the curves for Geometric and Poisson distribution. When the DGP is correctly specified **Figure 1**, the Poisson distribution has reasonable chance of being distinguished from geometric distribution.

Similarly, in **Figure 3**, as can be seen, the Geometric distribution closely approximates the data sets. In **Figures 5** and **7** two distributions are close but the Geometric (**Figure 5**) and the Poisson distributions (**Figure 7**) does appear to be much closer to the data sets. When $\pi = 0.535$, the distribution for both (**Figure 9**) Poisson distribution and Geometric distribution are similar, while being slightly corresponding to the ordinary Hellinger distance. As expected, our statistic divergence $\mathcal{HI}^h$ diverges to $-\infty$ (**Figures 2** and **8**) and to $+\infty$ (**Figures 4** and **8**) more rapidly symmetrical about the axis that passes through the mode of data distribution. This follows from

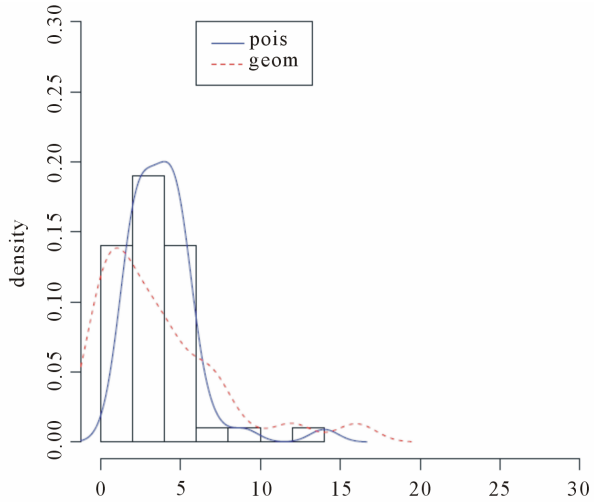

Figure 1. Histogram of DGP Pois(4) with $n = 50$.

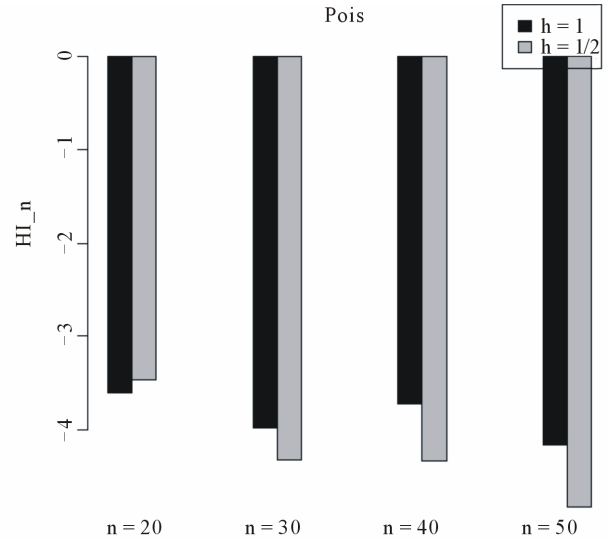

Figure 2. Comparative barplot of $\mathcal{HI}^h$ depending n.

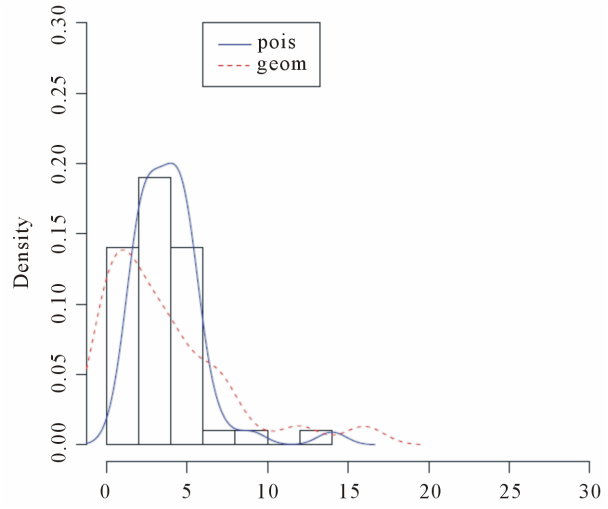

Figure 3. Histogram of DGP-Geom(0.2) with $n = 50$.

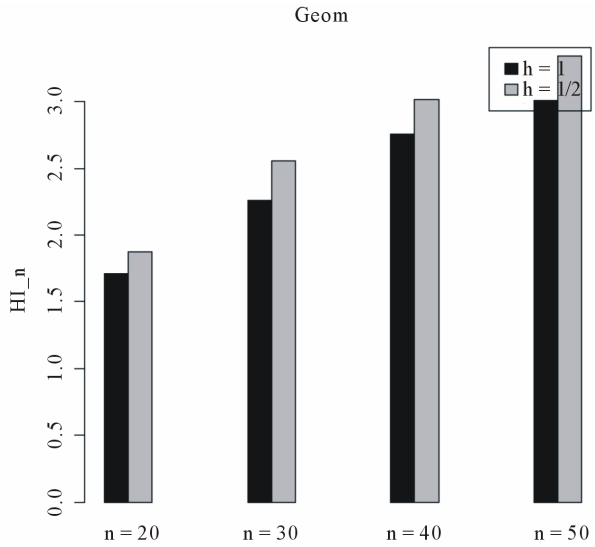

Figure 4. Comparaison barplot of $\mathcal{HI}^h$ depending n.

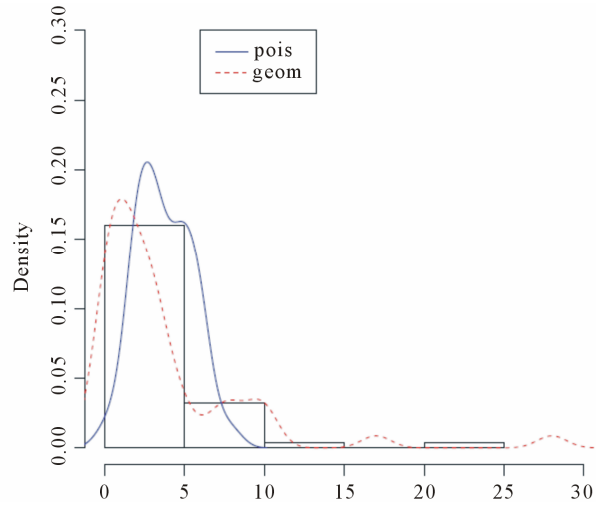

Figure 5. Histogram of DGP = 0.75 "Geom + 0.25" Pois with $n = 50$.

the fact that these two distributions are equidistant from the fact that these two distributions are equidistant from the DGP and would be difficult to distinguish from data in practice.

The preceding results in tables and the Theorem (5.5) confirm, in **Figures 2**, **4**, **6** and **8**, that the Hellinger indicator for the model selection procedure based on panelized hellinger divergence statistic with $h = 0.5$ (light bars) dominates the procedure obtained with $h = 1$ (dark bars) when we use the penalized Hellinger distance test than the classical Hellinger distance test. Hence, **Figure 10** allows a comparison with the asymptotic $\mathcal{N}(0, 1)$ approximation under our null hypothesis of equivalence. Hence the indicator $\mathcal{HI}^{1/2}$, based on the penalized Hellinger distance is closer to the mean of $\mathcal{N}(0, 1)$ than is the indicator $\mathcal{HI}^1$.

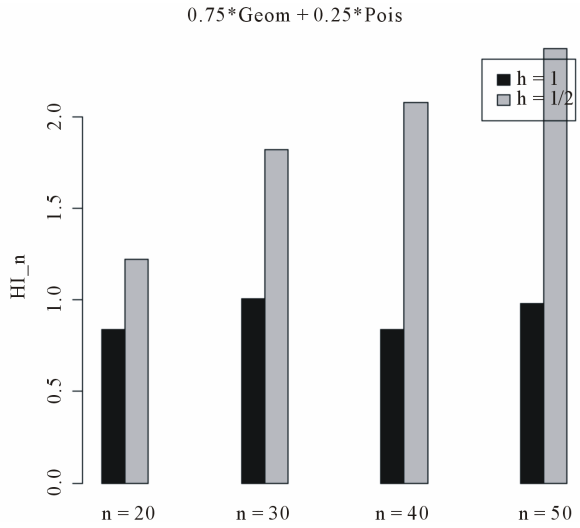

Figure 6. Comparative barplot of $\mathcal{HI}^h$ depending n.

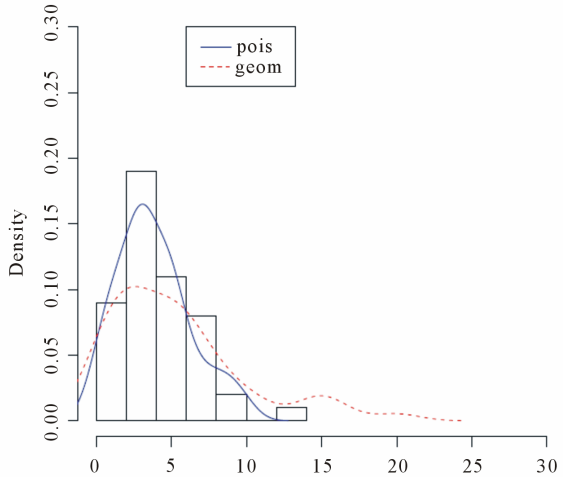

Figure 7. Histogram of DGP = 0.25 × "Geom + 0.75" Pois with n = 50.

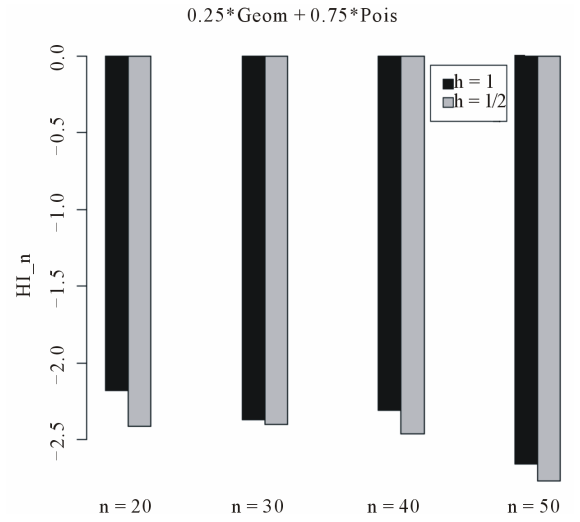

Figure 8. Comparative barplot of $\mathcal{HI}^h$.

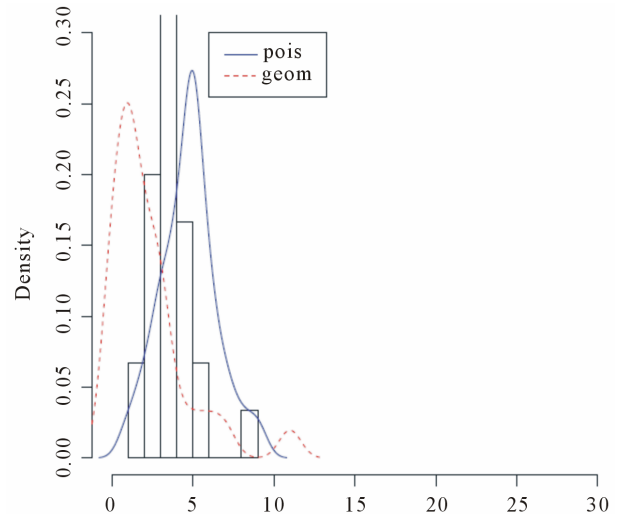

Figure 9. Histogram of DGP = 0.465 "Geom + 0.535" Pois with n = 50.

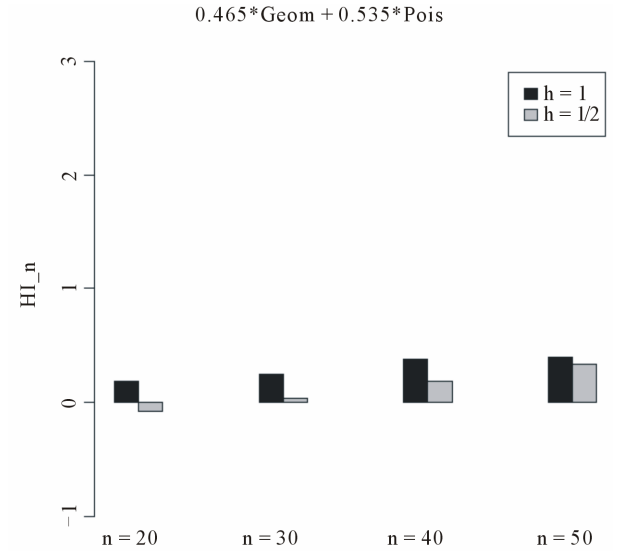

Figure 10. Comparative barplot of $\mathcal{HI}^h$ depending n.

7. Conclusion

In this paper we investigated the problems of model selection using divergence type statistics. Specifically, we proposed some asymptotically standard normal and chi-square tests for model selection based on divergence type statistics that use the corresponding minimum penalized Hellinger estimator. Our tests are based on testing whether the competing models are equally close to the true distribution against the alternative hypotheses that one model is closer than the other where closeness of a model is measured according to the discrepancy implicit in the divergence type statistics used. The penalized Hellinger divergence criterion outperforms classical criteria for model selection based on the ordinary Hellinger distance, especially in small sample, the difference is

expected to be minimal for large sample size. Our work can be extended in several directions. One extension is to use random instead of fixed cells. Random cells arise when the boundaries of each cell c_i depend on some unknown parameter vector γ, which are estimated. For various examples, see e.g., Andrews [37]. For instance, with appropriate random cells, the asymptotic distribution of a Pearson type statistic may become independent of the true parameter θ_o under correct specification. In view of this latter result, it is expected that our model selection test based on penalized Hellinger divergence measures will remain asymptotically normally or chi-square distributed.

8. Acknowledgements

This research was supported, in part, by grants from AIMS (African Institute for Mathematical Sciences) 6 Melrose Road, Muizenberg-Cape Town 7945 South Africa.

REFERENCES

[1] W. G. Cochran, "The χ^2 Test of Goodness of Fit," *The Annals of Mathematical Statistics*, Vol. 23, No. 3, 1952, pp. 315-345. doi:10.1214/aoms/1177729380

[2] G. S. Watson, "On the Construction of Significance Tests on the Circle and the Sphere," *Biometrika*, Vol. 43, No. 3-4, 1956, pp. 344-352. doi:10.2307/2332913

[3] D. S. Moore, "Chi-Square Tests in Studies in Statistics," 1978.

[4] D. S. Moore, "Tests of Chi-Squared Type Goodness of Fit Techniques," 1986.

[5] D. W. K. Andrews, "Chi-Square Diagnostic Tests for Econometric Models: Theory," *Econometrica*, Vol. 56, No. 6, 1988, pp. 1419-1453. doi:10.2307/1913105

[6] H. A. Kaike, "Information Theory and Extension of the Likelihood Ratio Principle," *Proceedings of the Second International Symposium of Information Theory*, 1973, pp. 257-281.

[7] Q. H. Vuong, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrika*, Vol. 57, No. 2, 1989, pp. 257-306. doi:10.2307/1912557

[8] Q. H. Vuong and W. Wang, "Minimum Chi-Square Estimation and Tests for Model Selection," *Journal of Econometrics*, Vol. 57, No. 1-2, 1993, pp. 141-168. doi:10.1016/0304-4076(93)90104-D

[9] P. Ngom, "Selected Estimated Models with *Á*-Divergence Statistics Global," *Journal of Pure and Applied Mathematics*, Vol. 3, No. 1, 2007, pp. 47-61.

[10] A. Diédhiou and P. Ngom, "Cutoff Time Based on Generalized Divergence Measure," *Statistics and Probability Letters*, Vol. 79, No. 10, 2009, pp. 1343-1350. doi:10.1016/j.spl.2009.02.006

[11] D. R. Cox, "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Los Angeles, 20-30 June 1961, pp. 105-123.

[12] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transaction on Information Theory*, Vol. 19, No. 6, 1974, pp. 716-723.

[13] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, No. 1, 1951, pp. 79-86. doi:10.1214/aoms/1177729694

[14] R. J. Bearn, "Minimum Hellinger Distance Estimates for Parametric Models," *The Annals of Mathematical Statistics*, Vol. 5, No. 3, 1977, pp. 445-463.

[15] D. G. Simpson, "Hellinger Deviance Test: Efficiency, Breakdown Points and Examples," *Journal of American Statistical Association*, Vol. 84, No. 405, 1989, pp. 107-113. doi:10.1080/01621459.1989.10478744

[16] B. G. Lindsay, "Efficiency versus Robustness: The Case for Minimum Distance Hellinger Distance and Related Methods," *Annals of Statistics*, Vol. 22, No. 2, 1994, pp. 1081-1114. doi:10.1214/aos/1176325512

[17] A. Basu and B. G. Lindsay, "Minimum Disparity Estimation for Continuous Models: Efficiency, Distributions and Robustness," *The Annals of Mathematical Statistics*, Vol. 46, No. 4, 1994, pp. 683-705. doi:10.1007/BF00773476

[18] A. Basu, I. R. Harris and S. Basu, "Tests of Hypotheses in Discrete Models Based on the Penalized Hellinger Distance," *Statistics and Probability Letters*, Vol. 27, No. 4, 1996, pp. 367-373. doi:10.1016/0167-7152(95)00101-8

[19] A. Basu and S. Basu, "Penalized Minimum Disparity Methods for Multinomial Models," *Statistica Sinica*, Vol. 8, 1998, pp. 841-860.

[20] M. W. Birch, "The Detection of Partial Association, II: The General Case," *Journal of the Royal Statistical Society*, Vol. 27, No. 1, 1965, pp. 111-124.

[21] J. P. W. Pluim, J. B. A. Maintz and A. M. Viergever, "f-Information Measures to Medical Image Registration," *IEEE Transactions on Medical Imaging*, Vol. 23, No. 12, 2004, pp. 1508-1516. doi:10.1109/TMI.2004.836872

[22] I. Vajda, "Theory of Statistical Evidence and Information," Kluwe Academic Plubisher, Dordrecht, 1989.

[23] D. Morales, L. Pardo and I. Vajda, "Asymptotic Divergence of Estimates of Discrete Distribution," *Journal of Statistical Planning and Inference*, Vol. 483, No. 3, 1995, pp. 347-369. doi:10.1016/0378-3758(95)00013-Y

[24] N. Cressie and T. R. C. Read, "Multinomial Goodness of Fit Test," *Journal of the Royal Statistical Society*, Vol. 463, No. 3, 1984, pp. 440-464.

[25] K. Zografos and K. Ferentinos, "Divergence Statistics Sampling Properties and Multinomial Goodness of fit and Divergence Tests," *Communications in Statistics—Theory and Methods*, Vol. 19, No. 5, 1990, pp. 1785-1802. doi:10.1080/03610929008830290

[26] M. Salicru, D. Morales, M. L. Menendez, *et al*., "On the Applications of Divergence Type Measures in Testing Statistical Hypotheses," *Journal of Multivariate Analysis*, Vol. 51, No. 2, 1994, pp. 372-391. doi:10.1006/jmva.1994.1068

[27] A. Bar-Hen and J. J. Dandin, "Generalisation of the Ma-

halanobis Distance in the Mixed Case," *Journal of Multivariate Analysis*, Vol. 532, No. 2, 1995, pp. 332-342. doi:10.1006/jmva.1995.1040

[2] L. Pardo, D. Mmorales, M. Salicrù and M. L. Menendez, "Generalized Divergences Measures: Amount of Information, Asymptotic-Distribution and Its Applications to Test Statistical Hypotheses," *International Sciences*, Vol. 84, No. 3-4, 1995, pp. 181-198.

[3] M. L. Menendez, L. Pardo, M. Salicrù and D. Morales, "Divergence Measures, Based on Entropy Functions and Statistical Inference," *Sankyã*: *The Indian Journal of Statistics*, Vol. 57, No. 3, 1995, pp. 315-337.

[4] I. Csiszár, "Information-Type Measure of Difference of Probability Distribution and Indirect Observations," *Studia Scientiarum Mathematicarum Hungarica*, Vol. 2, 1967, pp. 299-318.

[5] M. Broniatowski and A. Toma, "Dual Divergence Estimators and Tests: Robustness Results," *Journal of Multivariate Analysis*, Vol. 102, No. 1, 2011, pp. 20-36.

[6] A. Basu, A. Mandal and L. Pardo, "Hypothesis Testing for Two Discrete Populations Based on the Hellinger Distance," *Statistics and Probability Letters*, Vol. 80, No. 3-4, 2010, pp. 206-214. doi:10.1016/j.spl.2009.10.008

[7] F. Liese and I. Vajda, "Convex Statistical Distance, vol. 95 of Teubner-Texte zur Mathematik," 1987.

[8] R. Tamura and D. D. Boos, "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance," *Journal of American Statistical Association*, Vol. 81, No. 333, 1989, pp. 223-229.

[9] A. Basu, S. Sarkar and A. N. Vidyashankar, "Minimum Negative Exponential Disparity Estimation in Parametric Models," *Journal of Statistical Planning and Inference*, Vol. 582, No. 2, 1997, pp. 349-370. doi:10.1016/S0378-3758(96)00078-X

[10] I. R. Harris and A. Basu, "Hellinger Distance as Penalized Loglikelihood," *Communications in Statistics—Theory and Methods*, Vol. 21, No. 3, 1994, pp. 637-646. doi:10.1080/03610929208830804

[11] A. Mandal, R. K. Patra and A. Basu, "Minimum Hellinger Distance Estimation with Inlier Modification," *Sankhya*, Vol. 70, 2008, pp. 310-322.

