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 \( \varphi \)-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 \( \varphi \)-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 \( \varphi \)-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 proposed 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
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 $X = \{1, \cdots, m\}$. Let $P = (p_1, \cdots, p_m)^T$ be a probability vector i.e. $P \in \Omega_m$ where $\Omega_m$ 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

$$P = \{ P = (p_1(\theta), \cdots, p_m(\theta))^T : \theta \in \Theta \}$$

which may or may not contain the true distribution $P_0$ where $\Theta$ is a compact subset of k-dimensional Euclidean space (with $k < m - 1$). If $P$ contains $P_0$ then there exists a $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$ and the model $P$ is said to be correctly specified.

We are interested in testing $H_0 : P \in P$ (with true parameter $\theta_0$) versus $H_1 : P \in \Omega_m \cap \neg P$.

By $\| \|$ we denote the usual Euclidean norm and we interpret probability distributions on $X$ as row vectors from $\mathbb{R}^m$. For simplicity we restrict ourselves to unknown true parameters $\theta_0$ satisfying the classical regularity conditions by Birch [20]:

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

2) The mapping $P : \Theta \in \Omega_m$ is totally differentiable at $\theta_0$ so that the partial derivatives of $p_i$ with respect to each $\theta_j$ exist at $\theta_0$ and $p_i(\theta)$ has a linear approximation at $\theta_0$ given by

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

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

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

3) The Jacobian matrix

$$J(\theta_0) = \left( \frac{\partial p_i(\theta_0)}{\partial \theta_j} \right)_{j=1}^k_{i=1}$$

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

4) The inverse mapping $P^{-1} : P \to \Theta$ is continuous at $P_{\theta_0}$.

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

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

This non parametric estimator $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_m)$ is defined by

$$\hat{p}_i = \frac{N_i}{n}, \quad N_j = \sum_{i=1}^m T_{ij}^n(X_i) \quad \text{where}$$

$$T_{ij}^n(X_i) = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

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

3. A Brief Review of $\phi$-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 $\phi, f$ 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 $\phi$-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 $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_m$ with probabilities $P=(p_1, p_2, \ldots, p_m)$ and $Q=(q_1, q_2, \ldots, q_m)$ respectively. Then

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

is the $\phi$-divergence between $P$ and $Q$ (see Csisza'r, [30]) for every $\phi$ in the set $\Phi$ of real convex functions defined on $[0, \infty)$. The function $\phi(u)$ is assumed to verify the following regularity condition: $\phi:[0, \infty[ \to R \cup \{\infty\}$ is convex and continuous, where $0\phi(0) = 0$ and

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

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

$$\hat{Q} = \hat{D}_\theta = P_{\hat{\theta}}$$  \hspace{1cm} (3.3)

of $P_{\theta}$, which can be obtained by means of various point estimators

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

of the unknown parameter $\theta$.

It is convenient to measure the difference between observed and expected frequencies $P_{\theta}$ . A minimum divergence estimator of $\theta$ is a minimizer of $D_\phi(\hat{P}, P_{\theta})$ where $\hat{P}$ is a nonparametric distribution estimate. In our case, where data come from a discrete distribution, the empirical distribution defined in (2.2) can be used.

In particular if we replace $\phi(1) = -4 \sqrt{x - \frac{1}{2}} (x+1)$ in (3.2) we get the Hellinger distance between distribution $\hat{P}$ and $P_{\theta}$ given by

$$D_\phi(\hat{P}, P_{\theta}) = HD_\phi(\hat{P}, P_{\theta})$$

$$= 2 \sum_{i=1}^{m} \left( \frac{\hat{p}_i^{\frac{1}{2}}}{\hat{p}_i^{\frac{1}{2}}} - \frac{1}{\hat{p}_i^{\frac{1}{2}}} \right)^2 : \phi \in \Phi$$  \hspace{1cm} (3.4)

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

$$D_{\phi}(\hat{P}, P_{\theta}) = \min_{\phi \in \Theta} D_{\phi}(\hat{P}, P_{\theta}) : \phi \in \Phi$$  \hspace{1cm} (3.5)

$$\hat{\theta} = \arg \min_{\phi \in \Theta} D_{\phi}(\hat{P}, P_{\theta}) : \phi \in \Phi$$  \hspace{1cm} (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

$$\hat{\theta}_{KL} = \arg \min_{\phi \in \Theta} KL_{\phi}(P_{\theta}, \hat{P})$$

$$= \arg \min_{\phi \in \Theta} \sum_{i=1}^{m} -\log p_i(\theta) \hat{p}_i = MLE$$

where $KL_{\theta}$ is the modified Kullback-Leibler divergence.

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

$$\hat{\theta}_{H} = \arg \min_{\phi \in \Theta} HD_{\phi}(\hat{P}, P_{\theta}) : \phi \in \Phi$$  \hspace{1cm} (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 $P_{\theta}$ is defined by:

$$PHD_{\phi}(\hat{P}, P_{\theta})$$

$$= 2 \left[ \sum_{i=1}^{m} \left( \frac{\hat{p}_i^{\frac{1}{2}}}{\hat{p}_i^{\frac{1}{2}}} - \frac{1}{\hat{p}_i^{\frac{1}{2}}} \right)^2 + h \sum_{i=1}^{n} p_i(\theta) \left( \frac{\hat{p}_i^{\frac{1}{2}}}{\hat{p}_i^{\frac{1}{2}}} - \frac{1}{\hat{p}_i^{\frac{1}{2}}} \right)^2 \right]$$  \hspace{1cm} (3.8)

where $h$ is a real positive number with

$$\sigma = \{ i : \hat{p}_i \neq 0 \}$$

and $\sigma c = \{ i : \hat{p}_i = 0 \}$

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

Hence (3.7) can be written as follows

$$\hat{\theta}_{PH} = \arg \min_{\phi \in \Theta} PHD_{\phi}(\hat{P}, P_{\theta})$$  \hspace{1cm} (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 \( \theta_0 \) and mapping \( P : \Theta \rightarrow \Omega_m \) satisfy conditions 1 - 5 of Section 2 hold. Suppose that \( I_{\theta_0} \) is the k \( \times \) k Fisher Information matrix and \( \hat{\theta}_{PH} \) satisfying (3.7).

Then the limiting distribution of \( \sqrt{n}\left(\hat{\theta}_{PH} - \theta_0\right) \) as \( n \rightarrow +\infty \) is \( N\left[\theta_0, \Sigma_{\theta_0}\right] \).

**Lemma 4.2.** We have \( \sqrt{n}\left(\hat{\theta}_{PH} - \theta_0\right) \xrightarrow{D} N\left[\theta_0, \Sigma_{\theta_0}\right] \)

where \( \hat{P}(\theta_0) = \left(\hat{p}_{1\theta_0}, \ldots, \hat{p}_{m\theta_0}\right) \) an estimator of \( P_{\theta_0} = \left(p_{1\theta_0}, \ldots, p_{m\theta_0}\right) \) defined in (2.1) with

\[
\Sigma_{\theta_0} = \text{diag} \left(\hat{P}(\theta_0) - p_{\theta_0} P_{\theta_0}^T\right)
\]

**Proof.** Denote

\[
V = \left[\frac{N_i - np_{i\theta_0}}{\sqrt{n}}, \ldots, \frac{N_m - np_{m\theta_0}}{\sqrt{n}}\right]
\]

and \( N_j = \sum T_j^i \) where

\[
T_j^i = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise} \end{cases} \]

and applying the Central Limit Theorem we have

\[
\left[\frac{N_1 - np_{1\theta_0}}{\sqrt{n}}, \ldots, \frac{N_m - np_{m\theta_0}}{\sqrt{n}}\right] \xrightarrow{D} N\left[0, \Sigma_{\theta_0}\right]
\]

where

\[
\Sigma_{\theta_0} = \text{diag} \left(\hat{P}(\theta_0) - p_{\theta_0} P_{\theta_0}^T\right).
\]

For simplicity, we write \( D^*_H \left(\hat{P}, P_{\theta_0}\right) \) instead \( \text{PHD}^*_H \left(\hat{P}, P_{\theta_0}\right) \).

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

\[
\sqrt{n}\left(\hat{P}, P_{\theta_0}\right) \xrightarrow{D} N\left[0, \Lambda_{\theta_0}\right]
\]

where

\[
\Lambda_{\theta_0} = \Sigma_{\theta_0} - \Sigma_{\theta_0} M_{\theta_0}^T - M_{\theta_0} \Sigma_{\theta_0} + M_{\theta_0} \Sigma_{\theta_0} M_{\theta_0}^T
\]

\[
M_{\theta_0} = J_{\theta_0} \hat{I}_{\theta_0}^{-1} \left(\theta_0\right)^T \text{diag} \left(P_{\theta_0}^2\right)
\]

\[
\Sigma_{\theta_0} = \Sigma_{\theta_0}
\]

**Proof.** A first order Taylor expansion gives
\[ P_{\hat{\theta}_{PH}} = P_{\theta_0} + J_{\theta_0} \left( \hat{\theta}_{PH} - \theta_0 \right)^T + o \left( \| \hat{\theta}_{PH} - \theta_0 \| \right) \] (4.12)

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

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

From (4.12) and (4.13) we obtain

\[ P_{\hat{\theta}_{PH}} = P_{\theta_0} + J_{\theta_0}^{-1} \left( \theta_0 \right) D_{\theta_0}^T \text{diag} \left[ P_{\theta_0}^{1/2} \right] \left( \hat{P} - P_{\theta_0} \right)^T + o \left( \left\| \hat{P} - P_{\theta_0} \right\| \right) \]

therefore the random vectors

\[ \begin{bmatrix} \hat{P} - P_{\theta_0} \\ P_{\theta_0} - P_{\theta_0} \end{bmatrix}_{2\times m+1} \text{ and } \begin{bmatrix} I \\ M_{\theta_0} \end{bmatrix}_{2\times m} \times \begin{bmatrix} \hat{P} - P_{\theta_0} \end{bmatrix}_{m+1} \]

where \( I \) is the \( m \times m \) unity matrix, have the same asymptotic distribution.

Furthermore it is clear (applying TCL) that

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

Being \( \Sigma_{\theta_0} \) the \( m \times m \) matrix \( \text{diag}(P_{\theta_0}) - P_{\theta_0} P_{\theta_0}^T \) implies

\[ \sqrt{n} \left[ \begin{bmatrix} \hat{P} - P_{\theta_0} \\ P_{\theta_0} - P_{\theta_0} \end{bmatrix}_{2\times m+1} \right] \overset{\mathcal{L}}{\rightarrow} N \left[ \begin{bmatrix} I \\ M_{\theta_0} \end{bmatrix} \Sigma_{\theta_0} \left( I, M_{\theta_0}^T \right) \right] \]

therefore we get

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

\[ \Lambda_{\theta_0} = \Sigma_{\theta_0} - \Sigma_{\theta_0} M_{\theta_0} M_{\theta_0}^T - M_{\theta_0} \Sigma_{\theta_0} + M_{\theta_0} \Sigma_{\theta_0} M_{\theta_0}^T \]

The case which is interesting to us here is to test the hypothesis \( H_0: P \in \mathcal{P} \). Our proposal is based on the following penalized divergence test statistic \( D^h_R \left( \hat{P}, P_{\theta PH} \right) \) where \( \hat{P} \) and \( P_{\theta PH} \) have been introduced 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 \( H_0: P = P_{\theta_0} \) the asymptotic distribution of

\[ 2nD^h_R \left( \hat{P}, P_{\theta PH} \right) \]

is a chi-square when \( h = 1 \) with degrees \( m - k - 1 \) 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 \( i.e. \ H_1: P \neq P_{\theta_0} \). We introduce the following regularity assumptions

(A1) There exists \( \theta_1 = \arg \inf_{\theta \neq \theta_0} PHD^h \left( P, P_{\theta} \right) \) such that:

\[ P_{\hat{\theta}_{PH}} \overset{w}{\rightarrow} P_{\theta_1} \quad \text{when } n \rightarrow +\infty \]

(A2) There exists \( \theta_1 \in \Theta \); \( \Lambda^* = \begin{bmatrix} \Lambda_{11} \\ \Lambda_{12} \end{bmatrix} \) with \( \Lambda_{11} = \Sigma_{\theta_1} \) in (4.10) and \( \Lambda_{12} = \Lambda_{21} \) such that

\[ \sqrt{n} \left( \hat{P} - P_{\theta_0} \right) \overset{\mathcal{L}}{\rightarrow} N \left[ 0, \Lambda^* \right] \]

**Theorem 4.4.** Under \( H_1: P \neq P_{\theta_0} \) and assume that conditions (A1) and (A2) hold, we have:

\[ \sqrt{n} \left( \hat{P} - P_{\theta PH} \right) - D^h_R \left( P, P_{\theta_1} \right) \overset{\mathcal{L}}{\rightarrow} N \left[ 0, \Omega^2_{(\theta_1, \theta_0)} \right] \]

where \( \Omega^2_{(\theta_1, \theta_0)} = H^T \Lambda_{11} H + H^T \Lambda_{12} J + J^T \Lambda_{21} H + J^T \Lambda_{22} J \) (4.15)

\[ H^T = \begin{bmatrix} h_1 & \cdots & h_m \end{bmatrix} \]

with

\[ H^T = \begin{bmatrix} j_1 & \cdots & j_m \end{bmatrix} \]

And \( J^T = \begin{bmatrix} j_1 & \cdots & j_m \end{bmatrix} \)

\[ D^h_R \left( \hat{P}, P_{\theta PH} \right) = D^h_R \left( P, P_{\theta_1} \right) + H^T \left( \hat{P} - P \right) \]

\[ + J^T \left( P_{\theta PH} - P_{\theta_1} \right) + o \left( \| \hat{P} - P \| + \| P_{\theta PH} - P_{\theta_1} \| \right) \] (4.16)

From the assumed assumptions (A1) and (A2), the result follows.

5. Applications for Testing Hypothesis

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

5.1. Test of Goodness-Fit

For completeness, we look at \( D^h_R \left( \hat{P}, P_{\theta PH} \right) \) in the usual way, \( i.e. \) as a goodness-of-fit statistic. Recall that here \( \theta_{PH} \) is the minimum penalized Hellinger distance estimator of \( \theta \). Since \( D^h_R \left( \hat{P}, P_{\theta PH} \right) \) is a consistent estimator
of \( D^h_l \left( P, P_0 \right) \), the null hypothesis when using the statistic \( D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) \) is \( H_0 : D^h_l \left( P, P_0 \right) = 0 \) or equivalently, \( H_{\alpha} : P = P_0 \).

Hence, if \( H_0 \) is rejected so that one can infer that the parametric model \( P_0 \) is misspecified. Since \( D^h_l \left( P, P_0 \right) \) is non-negative and takes value zero only when \( P = P_0 \), the tests are defined through the critical region.

\[
C_{\hat{\theta}_{11}} \left( \frac{2nD^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right)}{\sigma} > q_{a, k} \right)
\]

where \( q_{a, k} \) is the \((1 - \alpha)\)-quantile of the \( \chi^2 \)-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 \( H_{\alpha} : D^h_l \left( P, P_0 \right) = 0 \).

Approximated power function is

\[
\begin{align*}
\beta_{(P)} &= \mathbb{P} \left[ 2nD^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) > q_{a, k} \right] \\
&\approx 1 - \mathcal{F}_q \left( q_{a, k} - \frac{2nD^h_l \left( P, P_0 \right)}{\sigma} \right)^2
\end{align*}
\]  

(5.17)

where \( q_{a, k} \) is the \((1 - \alpha)\)-quantile of the \( \chi^2 \)-distribution with \( m - k - 1 \) degrees of freedom and \( \mathcal{F}_q \) is a sequence of distribution function tending uniformly to the standard normal distribution \( \mathcal{F}(x) \). Note that if \( H_{\alpha} : D^h_l \left( P, P_0 \right) \neq 0 \), then for any fixed size \( \alpha \) the probability of rejection \( H_{\alpha} : D^h_l \left( P, P_0 \right) = 0 \) with the rejection rule \( 2nD^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) > q_{a, k} \) tends to one as \( n \rightarrow \infty \).

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

\[
\beta = 1 - \mathcal{F} \left( \frac{\sqrt{n} D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right)}{\sigma} \right)^2
\]

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

\[
n^* \left[ n^* \left( a + b \right) - n D^h_l \left( P, P_0 \right) q_{a, k} \right] = n \Omega_2 \left[ \mathcal{F}^{-1} \left( 1 - \beta \right) \right]^2
\]

The solution is given by

\[
n^* = \frac{\left( a + b \right) - n D^h_l \left( P, P_0 \right) q_{a, k}}{2 D^h_l \left( P, P_0 \right)^2}
\]

with \( a = \Omega_2 \left[ \mathcal{F}^{-1} \left( 1 - \beta \right) \right]^2 \) and \( b = q_{a, k} D^h_l \left( P, P_0 \right) \) and the required size is \( n_0 = \left[ n^* \right] + 1 \), where \([ ] \) denotes “integer part”.

### 5.2. Test for Model Selection

As we mentioned above, when one chooses a particular \( \phi \)-divergence type statistic

\[
D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) = PH^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right)
\]

with \( \hat{\theta}_{ij} \) the corresponding minimum penalized Hellinger distance estimator of \( \theta \), one actually evaluates the goodness-of-fit of the parametric model \( P_\theta \) according to the discrepancy \( D^h_l \left( P, P_\theta \right) \) between the true distribution \( P \) and the specified model \( P_\theta \). 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 \( D^h_l \left( P, P_\theta \right) \).

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

Our special interest is the situation in which a researcher has two competing parametric models \( P_\mu \) and \( G_\mu \) and he wishes to select the better of two models based on their discrimination statistic between the observations and models \( P_\mu \) and \( G_\mu \) defined respectively by

\[
D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) \quad \text{and} \quad D^h_l \left( \hat{P}, P_{\hat{\theta}_{12}} \right)
\]

Let the two competing parametric models \( P_\mu \) and \( G_\mu \) with the given discrepancy \( D^h_l \left( P, . \right) \).

**Définition 5.2.**

\( H_\mu^G : D^h_l \left( P, P_\mu \right) = D^h_l \left( P, P_\mu \right) \) means that the two models are equivalent,

\( H_\mu^G : D^h_l \left( P, P_\mu \right) < D^h_l \left( P, P_\mu \right) \) means that \( P_\mu \) is better than \( G_\mu \),

\( H_\mu^G : D^h_l \left( P, P_\mu \right) < D^h_l \left( P, P_\mu \right) \) means that \( P_\mu \) is worse than \( G_\mu \).

**Remark 5.3. 1.** It does not require that the same divergence type statistics be used in forming \( D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) \) and \( D^h_l \left( \hat{P}, P_{\hat{\theta}_{12}} \right) \). Choosing, however, different discrepancy for evaluating competing models is hardly justified.

**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

\[
D^h_l \left( P, P_\mu \right) - D^h_l \left( P, P_\mu \right)
\]

is unknown, but from the previous section, it can be estimated by the difference

\[
\sqrt{n} \left[ D^h_l \left( \hat{P}, P_{\hat{\theta}_{11}} \right) - D^h_l \left( \hat{P}, P_{\hat{\theta}_{12}} \right) \right]
\]

This difference converges to zero under the null hypo-
thesis $H^0_H$, but converges to a strictly negative or positive constant when $H^0_H$ and $H^0_{G\mu}$ holds.

These properties actually justify the use of $D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right)$ 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 $D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right)$ so as to assess its significance. To do so we consider the asymptotic distribution of $\sqrt{n} \left( D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right) \right)$ under $H^0_H$.

Our major task is to propose some tests for model selection, i.e. for the null hypothesis $H^0_H$ against the alternative $H_{\theta_{\mu}}$ or $H_{\mu}$. We use the next lemma with $\hat{\theta}_{\mu_{\mu}}$ and $\hat{\mu}_{\mu_{\mu}}$ as the corresponding minimum penalized Hellinger distance estimator of $\theta$ and $\mu$.

Using $P$ and $P_{\mu}$ defined earlier, we consider the vector $K^T \left( k_1, \ldots, k_m \right)$ where

$$ k_i = \left( \frac{\partial}{\partial P_i} D^h_{H} \left( P^1, P^2 \right) \right)_{P^1=P_i, P^2=P_0} \quad \text{with} \quad i = 1, \ldots, m $$

and $Q^T \left( q_1, \ldots, q_m \right)$ where

$$ q_i = \left( \frac{\partial}{\partial P_i} D^h_{H} \left( P^1, P^2 \right) \right)_{P^1=P_0, P^2=P_i} \quad \text{with} \quad i = 1, \ldots, m $$

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

(i) for the model $P_{\mu}$

$$ D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( P, P_{\mu} \right) = K^T \left( \hat{P}, P \right) + Q^T \left( \hat{P}, P \right) + o_p \left( 1 \right) $$

(ii) for model $G_{\mu}$

$$ D^h_{H} \left( \hat{P}, P_{\mu} \right) = D^h_{H} \left( P, G_{\mu} \right) + K^T \left( \hat{P}, P \right) + Q^T \left( G_{\mu} - G_{\mu} \right) + o_p \left( 1 \right) $$

**Proof.**

The results follow from a first order Taylor expansion.

We define

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

which is the variance of

$$ \left( K_{\theta} - K_{\mu} ; Q_{\theta} - Q_{\mu} \right)^T \left( \hat{P} \atop P_{\theta_{\mu}} \right) \left( P \atop P_{\mu} \right) . $$

Since $K_{\theta}$, $K_{\mu}$, $Q_{\theta}$, $Q_{\mu}$, and $\Lambda^*$ are consistently estimated by their sample analogues $K_{\hat{\theta}}$, $K_{\hat{\mu}}$, $Q_{\hat{\theta}}$, $Q_{\hat{\mu}}$ and $\Lambda^*$, hence $\Gamma^2$ is consistently estimated by

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

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

Let

$$ H^h_I = \sqrt{n} \left( D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right) \right) $$

where $H^h_I$ stands for the penalized Hellinger Indicator.

The following theorem provides the limit distribution of $H^h_I$ 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 $H^0_H$, $H^h_I \overset{D}{\rightarrow} \mathcal{N} \left( 0, 1 \right)$.

2) Under the null hypothesis $H_{\theta_{\mu}}$, $H^h_I \overset{P}{\rightarrow} -\infty$ in probability.

3) Under the null hypothesis $H_{\mu}$, $H^h_I \overset{P}{\rightarrow} +\infty$ in probability.

**Proof.**

From the Lemma (5.4), it follows that

$$ D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right) \overset{D}{=} D^h_{H} \left( P, P_{\theta_{\mu}} \right) - D^h_{H} \left( P, P_{\mu} \right) + K^T \left( \hat{P}, P \right) \overset{D}{=} \left( P_{\theta_{\mu}} - P_{\mu} \right) - K^T \left( \hat{P}, P \right) + o_p \left( 1 \right) $$

Under $H^0_H$, $P_{\theta_{\mu}} = P_{\mu}$ and $P_{\theta_{\mu}} = P_{\mu}$, we get

$$ D^h_{H} \left( \hat{P}, P_{\theta_{\mu}} \right) - D^h_{H} \left( \hat{P}, P_{\mu} \right) \overset{D}{=} K^T \left( \hat{P}, P \right) = \left( P_{\theta_{\mu}} - P \right) - K^T \left( \hat{P}, P \right) \overset{D}{=} \left( P_{\theta_{\mu}} - P_{\mu} \right) - K^T \left( P_{\theta_{\mu}} - P_{\mu} \right) + o_p \left( 1 \right) $$

Finally, applying the Central Limit Theorem and assumptions (A1) and (A2), we can now immediately obtain $H^h_I \overset{D}{\rightarrow} \mathcal{N} \left( 0, 1 \right)$.

**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 $\lambda$ and has density
\[ f(x, \lambda) = \frac{\exp(-\lambda) \times \lambda^x}{x!} \quad \text{for} \quad x \in \mathbb{N} \]
and zero otherwise.

The Geometric distribution $G(p)$ is parameterized by $p$ and has density
\[ G(x, p) = (1 - p) \times x^{x-1} \times p \quad \text{for} \quad x \in \mathbb{N}^+ \]
and zero otherwise. We use the minimum penalized Hel-
linger 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, \ldots, 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 $\lambda$ and $p$ are:
\[
\hat{\lambda}_{PH} = \arg\min_{\lambda \in \Theta} D^h_{PH}(\hat{P}, P_{\lambda}) = \arg\min_{\lambda \in \Theta} \left[ \sum_{i=0}^{m} \left( f_i^2 - \hat{p}_i^2 \right) + \sum_{i=0}^{m} p_i \right]
\]
\[
\hat{p}_{PH} = \arg\min_{p \in \Theta} D^h_{PH}(\hat{P}, P_{p}) = \arg\min_{p \in \Theta} \left[ \sum_{i=0}^{m} \left( f_i^2 - \hat{p}_i^2 \right) + h \sum_{i=0}^{m} p_i \right]
\]
$P_{\lambda}$ and $P_{p}$ 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
\[ m(\pi) = \pi \text{Pois}(4) + (1 - \pi) \text{Geom}(0.2) \]
where $\pi$ ($\pi \in [0, 1]$) 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 Hel-
linger 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
\[ [C_{i-1}, C_i] = [i-1, i], i = 1, \ldots, 7 \]
and
\[ [C_7, C_8] = [7, +\infty] \]
represents the last cell. We choose different values of $\pi$
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 $\pi = 1.00$ and $\pi = 0.00$ 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 $\pi = 0.25$, $\pi = 0.75$ and $\pi = 0.5$ second case is interpreted similarly as a Geometric slightly contaminated by a Pois-
sion distribution. The former case correspond to a DGP which is Poisson but slightly contaminated by a Geometric distribution. In the last case, $\pi = 0.535$ is the value for which the Poisson $D^h_{PH}(\hat{P}; G_{\pi = 0.535})$ and the Geometric $D^h_{PH}(\hat{P}; G_{\pi = 0.535})$ family are approximatively at equal dis-
tance to the mixture $m(\pi)$ according to the penalized Hel-
linger distance with the above cells.

Thus this set of experiments corresponds approximatively to the null hypothesis of our proposed model selec-
tion test $\mathcal{H}^h$. 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 pe-
nalized Hellinger distance estimator $\hat{\lambda}_{PH}$ and $\hat{p}_{PH}$, the penalized Hellinger goodness-of-fit statistics
\[ D^h_{PH}(\hat{P}, G_{\pi = 0.535}) \text{ and } D^h_{PH}(\hat{P}, G_{\pi = 0.535}) \text{, and the Hellinger indicator statistics } \mathcal{H}^h. \]
The values in parentheses are standard errors. The second half of each table gives in percentage the number of times our proposed model selec-
tion procedure based on $\mathcal{H}^h$ 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 ($\pi = 0.00$ and $\pi = 1.00$) 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 $\mathcal{H}^h$, its diverges to $-\infty$ or $+\infty$ at the approximate rate of $\sqrt{n}$ except in the Table 5.

In the latter case the $\mathcal{H}^h$ statistic converges, as ex-
pected, 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 $\mathcal{H}^h$ statistic steadily increases and ultimately converges to 100%.
Table 1. DGP = Pois(4).

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>0.210(0.03)</td>
<td>0.195(0.03)</td>
<td>0.197(0.02)</td>
<td>0.205(0.02)</td>
<td>0.201(0.01)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>3.950(0.40)</td>
<td>4.090(0.4)</td>
<td>4.015(0.31)</td>
<td>4.015(0.28)</td>
<td>4.0115(0.13)</td>
</tr>
<tr>
<td>DHP (Pois)</td>
<td>( h = 1 )</td>
<td>0.133(0.07)</td>
<td>0.081(0.05)</td>
<td>0.059(0.03)</td>
<td>0.042(0.03)</td>
</tr>
<tr>
<td></td>
<td>( h = 1/2 )</td>
<td>0.096(0.04)</td>
<td>0.064(0.03)</td>
<td>0.048(0.02)</td>
<td>0.034(0.02)</td>
</tr>
<tr>
<td>DHP (Geom)</td>
<td>( h = 1 )</td>
<td>0.391(0.28)</td>
<td>0.348(0.12)</td>
<td>0.208(0.09)</td>
<td>0.282(0.10)</td>
</tr>
<tr>
<td></td>
<td>( h = 1/2 )</td>
<td>0.278(0.07)</td>
<td>0.262(0.08)</td>
<td>0.242(0.06)</td>
<td>0.236(0.06)</td>
</tr>
<tr>
<td>( H^A )</td>
<td>( h = 1/2 )</td>
<td>-3.67(2.14)</td>
<td>-4.32(2.69)</td>
<td>-4.34(2.38)</td>
<td>-4.83(2.52)</td>
</tr>
<tr>
<td>Correct</td>
<td>77%</td>
<td>77%</td>
<td>77%</td>
<td>77%</td>
<td>77%</td>
</tr>
<tr>
<td>Indecisive</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 2. DGP = Geom(0.2).

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>0.196(0.04)</td>
<td>0.213(0.03)</td>
<td>0.203(0.02)</td>
<td>0.203(0.02)</td>
<td>0.201(0.01)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>3.920(1.0)</td>
<td>4.206(0.89)</td>
<td>4.109(0.67)</td>
<td>4.009(0.58)</td>
<td>4.053(0.34)</td>
</tr>
<tr>
<td>DHP (Pois)</td>
<td>( h = 1.0 )</td>
<td>0.356(0.14)</td>
<td>0.309(0.10)</td>
<td>0.271(0.09)</td>
<td>0.253(0.08)</td>
</tr>
<tr>
<td></td>
<td>( h = 0.5 )</td>
<td>0.281(0.1)</td>
<td>0.273(0.07)</td>
<td>0.254(0.07)</td>
<td>0.246(0.07)</td>
</tr>
<tr>
<td>DHP (Geom)</td>
<td>( h = 1 )</td>
<td>0.150(0.06)</td>
<td>0.089(0.05)</td>
<td>0.053(0.03)</td>
<td>0.039(0.02)</td>
</tr>
<tr>
<td></td>
<td>( h = 1/2 )</td>
<td>0.103(0.04)</td>
<td>0.267(0.03)</td>
<td>0.044(0.02)</td>
<td>0.035(0.02)</td>
</tr>
<tr>
<td>( H^A )</td>
<td>( h = 1/2 )</td>
<td>1.880(1.43)</td>
<td>2.560(1.37)</td>
<td>3.020(1.25)</td>
<td>3.340(1.14)</td>
</tr>
<tr>
<td>Correct</td>
<td>36%</td>
<td>62%</td>
<td>77%</td>
<td>84%</td>
<td>92%</td>
</tr>
<tr>
<td>Indecisive</td>
<td>64%</td>
<td>38%</td>
<td>23%</td>
<td>16%</td>
<td>08%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0%</td>
<td>00%</td>
<td>00%</td>
<td>00%</td>
<td>00%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>0.213(0.13)</td>
<td>0.197(0.12)</td>
<td>0.208(0.08)</td>
<td>0.202(0.05)</td>
<td>0.202(0.01)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>4.160(0.72)</td>
<td>3.910(0.55)</td>
<td>4.180(0.55)</td>
<td>3.970(0.43)</td>
<td>4.022(0.21)</td>
</tr>
<tr>
<td>DHP (Pois)</td>
<td>( h = 1 )</td>
<td>0.546(0.13)</td>
<td>0.472(0.1)</td>
<td>0.412(0.09)</td>
<td>0.402(0.08)</td>
</tr>
<tr>
<td></td>
<td>( h = 1/2 )</td>
<td>0.344(0.07)</td>
<td>0.340(0.05)</td>
<td>0.320(0.05)</td>
<td>0.311(0.05)</td>
</tr>
<tr>
<td>DHP (Geom)</td>
<td>( h = 1 )</td>
<td>0.150(0.06)</td>
<td>0.089(0.05)</td>
<td>0.053(0.03)</td>
<td>0.039(0.02)</td>
</tr>
<tr>
<td></td>
<td>( h = 1/2 )</td>
<td>-3.67(2.62)</td>
<td>-4.32(2.53)</td>
<td>-4.34(2.47)</td>
<td>-4.83(2.27)</td>
</tr>
<tr>
<td>( H^A )</td>
<td>Geom</td>
<td>23%</td>
<td>40%</td>
<td>50%</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>Indecisive</td>
<td>77%</td>
<td>60%</td>
<td>50%</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>Pois</td>
<td>00%</td>
<td>00%</td>
<td>00%</td>
<td>00%</td>
</tr>
<tr>
<td>( H^A )</td>
<td>Geom</td>
<td>84%</td>
<td>83%</td>
<td>89%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>Indecisive</td>
<td>17%</td>
<td>15%</td>
<td>19%</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td>Pois</td>
<td>03%</td>
<td>02%</td>
<td>02%</td>
<td>01%</td>
</tr>
</tbody>
</table>
Table 4. DGP = 0.75 × Pois(4) + 0.25 × Geom(0.2).

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\lambda} )</td>
<td>4.110(0.43)</td>
<td>4.090(0.31)</td>
<td>3.970(0.28)</td>
<td>4.020(0.26)</td>
<td>4.019(0.17)</td>
</tr>
<tr>
<td>DHP (Pois)</td>
<td>( h = 1 )</td>
<td>1.779(0.45)</td>
<td>1.634(0.30)</td>
<td>1.650(0.28)</td>
<td>1.570(0.24)</td>
</tr>
<tr>
<td>DHP (Geom)</td>
<td>( h = 1 )</td>
<td>2.055(0.35)</td>
<td>1.870(0.25)</td>
<td>0.053(0.03)</td>
<td>0.039(0.02)</td>
</tr>
</tbody>
</table>

\[ \mathcal{H}_I^a \]

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\hat{\lambda}} )</td>
<td>0.213(0.03)</td>
<td>0.212(0.03)</td>
<td>0.210(0.02)</td>
<td>0.206(0.02)</td>
<td>0.203(0.01)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>4.110(0.43)</td>
<td>4.090(0.31)</td>
<td>3.970(0.28)</td>
<td>4.020(0.26)</td>
<td>4.019(0.17)</td>
</tr>
<tr>
<td>DHP (Pois)</td>
<td>( h = 1 )</td>
<td>1.779(0.45)</td>
<td>1.634(0.30)</td>
<td>1.650(0.28)</td>
<td>1.570(0.24)</td>
</tr>
<tr>
<td>DHP (Geom)</td>
<td>( h = 1 )</td>
<td>2.055(0.35)</td>
<td>1.870(0.25)</td>
<td>0.053(0.03)</td>
<td>0.039(0.02)</td>
</tr>
</tbody>
</table>

\[ \mathcal{H}_I^a \]

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 \( \mathcal{H}_I^a \) 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
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 penalized 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 $\mathcal{H}/2$, based on the penalized Hellinger distance is closer to the mean of $\mathcal{N}(0, 1)$ than is the indicator $\mathcal{H}_1^1$.
Figure 6. Comparative barplot of $\mathcal{H}^h$ depending on $n$.

Figure 7. Histogram of DGP = $0.25 \times \text{Geom} + 0.75 \times \text{Pois}$ with $n = 50$.

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

Figure 9. Histogram of DGP = $0.465 \times \text{Geom} + 0.535 \times \text{Pois}$ with $n = 50$.

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

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 $\gamma$, 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 $\theta_0$, 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


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


