entropy

MDPI

Article

Minimum Penalized ϕ-Divergence Estimation under Model Misspecification

M. Virtudes Alba-Fernández [1,*], M. Dolores Jiménez-Gamero [2] and F. Javier Ariza-López [3]

[1] Departamento de Estadística e Investigación Operativa, Universidad de Jaén, 23071, Jaén, Spain
[2] Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, 41012, Sevilla, Spain; dolores@us.es
[3] Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría, Universidad de Jaén, 23071, Jaén, Spain; fjariza@ujaen.es
* Correspondence: mvalba@ujaen.es; Tel.: +34-953212142

Received: 8 March 2018; Accepted: 27 April 2018; Published: 30 April 2018

Abstract: This paper focuses on the consequences of assuming a wrong model for multinomial data when using minimum penalized ϕ-divergence, also known as minimum penalized disparity estimators, to estimate the model parameters. These estimators are shown to converge to a well-defined limit. An application of the results obtained shows that a parametric bootstrap consistently estimates the null distribution of a certain class of test statistics for model misspecification detection. An illustrative application to the accuracy assessment of the thematic quality in a global land cover map is included.

Keywords: minimum penalized ϕ-divergence estimator; consistency; asymptotic normality; goodness-of-fit; bootstrap distribution estimator; thematic quality assessment

1. Introduction

In many practical settings, individuals are classified into a finite number of unique nonoverlapping categories, and the experimenter collects the number of observations falling in each of such categories. In statistics, that sort data is called multinomial data. Examples arise in many scientific disciplines: in economics, when dealing with the number of different types of industries observed in a geographical area; in biology, when counting the number of individuals belonging to one of k species (see, for example, Pardo [1], pp. 94–95); in sports, when considering the number of injured players in soccer matches (see, for example, Pardo [1], p. 146); and many others.

When dealing with multinomial data, one often finds zero cell frequencies, even for large samples. Although many examples can be given, we will center on the following one, since two related data sets will be analyzed in Section 4. Zero cell frequencies are usually observed when the quality of the geographic information data is assessed, and specifically, when we pay attention to the thematic component of this quality. Roughly speaking, the thematic quality refers to the correctness of the qualitative aspect of an element (pixel, feature, etc.). To give an assessment of the thematic accuracy, a comparison is needed between the label considered as true of a feature and the label assigned to the same feature after a classification (among a number of labels previously stated). This way, each element/feature, which really belongs to a particular category, can be classified as belonging to the same category (correct assignment), or as belonging to another one (incorrect assignment). Given a sample of n elements belonging to a particular category, after collecting the number of elements correctly classified, X_1, and the number of incorrect classifications in a set of $k-1$ possible categories, X_i, $i = 2, \ldots, k$, we obtain a multinomial vector $(X_1, X_2, \ldots, X_k)^t$, for which small or zero cell frequencies are often observed associated with the incorrect classifications, X_i, $i = 2, \ldots, k$.

Motivated by this example in the geographic information data context, as well as many others, along this paper, it will be assumed that the available information can be summarized by means of a random vector $X = (X_1, \ldots, X_k)^t$ having a k-cell multinomial distribution with parameters n and $\pi = (\pi_1, \ldots, \pi_k)^t \in \Delta_{0k} = \{(\pi_1, \ldots, \pi_k)^t : \ \pi_i \geq 0, \ 1 \leq i \leq k, \ \sum_{i=1}^k \pi_i = 1\}$, $X \sim \mathcal{M}_k(n; \pi)$ in short. Notice that, if $\pi \in \Delta_{0k}$, then some components of π may equal 0, implying that some cell frequencies can be equal to zero, even for large samples. In many instances, it is assumed that π belongs to a parametric family $\pi \in \mathcal{P} = \{P(\theta) = (p_1(\theta), \ldots, p_k(\theta))^t, \ \theta \in \Theta\} \subset \Delta_k = \{(\pi_1, \ldots, \pi_k)^t : \ \pi_i > 0, \ 1 \leq i \leq k, \ \sum_{i=1}^k \pi_i = 1\}$, where $\Theta \subseteq \mathbb{R}^s$, $k - s - 1 > 0$ and $p_1(\cdot), \ldots, p_k(\cdot)$ are known real functions.

When it is assumed that $\pi \in \mathcal{P}$, π is usually estimated through $P(\hat{\theta}) = (p_1(\hat{\theta}), \ldots, p_k(\hat{\theta}))^t$ for some estimator $\hat{\theta}$ of θ. A common choice for $\hat{\theta}$ is the maximum likelihood estimator (MLE), which is known to have good asymptotic properties. Basu and Sarkar [2] and Morales et al. [3] have shown that these properties are shared by a larger class of estimators: the minimum ϕ-divergence estimators (MϕE). This class includes MLEs as a particular case. However, as illustrated in Mandal et al. [4], the finite sample performance of these estimators can be improved by modifying the weight that each ϕ-divergence assigns to the empty cells. The resulting estimator is called the minimum penalized ϕ-divergence estimator (MPϕE). Moreover, Mandal et al. [4] have shown that such estimators have the same asymptotic properties as the MϕEs. Specifically, they are strongly consistent and, conveniently normalized, asymptotically normal. To derive these asymptotic properties, it is assumed that the probability model is correctly specified, that is to say, that we are sure about $\pi \in \mathcal{P}$.

If the parametric model is not correctly specified, Jiménez-Gamero et al. [5] have shown that, under certain assumptions, the MϕEs still have a well defined limit, and, conveniently normalized, they are asymptotically normal. For the MLE, these results were known from those in [6]. Because, as argued before, the use of penalized ϕ-divergences may lead to better performance of the resulting estimators, the aim of this piece of research is to investigate the asymptotic properties of the MPϕEs under model misspecification. If the model considered is true, we obtain as a particular case the results in [4].

The usefulness of the results obtained is illustrated by applying them to the problem of testing goodness-of-fit to the parametric family $\mathcal{P}$,

$$H_0 : \ \pi \in \mathcal{P},$$

against the alternative

$$H_1 : \ \pi \notin \mathcal{P},$$

using as a test statistic a penalized ϕ_1-divergence between a nonparametric estimator of π, the relative frequencies, and a parametric estimator of π, obtained by assuming that the null hypothesis is true, $P(\hat{\theta})$, $\hat{\theta}$ being an MPϕ_2E. Here, ϕ_1 and ϕ_2 may differ. The convenience of using this type of test statistics is justified in Mandal et al. [7]. Although these authors show that, under H_0, such test statistics are asymptotically distribution free, the asymptotic approximation to the null distribution of the test statistics in this class is rather poor. Some numerical examples illustrate this unsatisfactory behavior of the asymptotic approximation. By using the fact that the MPϕE always converges to a well-defined limit, whether the model in H_0 is true or not, we prove that the bootstrap consistently estimates the null distribution of these test statistics. We then retake the previously cited numerical examples to exemplify the usefulness of the bootstrap approximation which, despite the demand for more computing time, is more accurate than that yielded by the asymptotic null distribution for small and moderate sample sizes.

The rest of the paper is organized as follows. Section 2 studies certain asymptotic properties of MPϕ_2Es; specifically, conditions are given for the strong consistency and asymptotic normality. Section 3 uses such results to prove that a parametric bootstrap provides a consistent estimator to the null distribution of test statistics based on penalized ϕ-divergences for testing H_0. Section 4 displays an application of the results obtained in the context of a classification work in a cover land map.

Before ending this section we introduce some notation: all limits in this paper are taken when $n \to \infty$; $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution; $\xrightarrow{P}$ denotes convergence in probability; $\xrightarrow{a.s.}$ denotes the almost sure convergence; let $\{A_n\}$ be a sequence of random variables and let $\epsilon \in \mathbb{R}$, then $A_n = O_P(n^{-\epsilon})$ means that $n^{\epsilon} A_n$ is bounded in probability, $A_n = o_P(n^{-\epsilon})$ means that $n^{\epsilon} A_n \xrightarrow{P} 0$, and $A_n = o(n^{-\epsilon})$ means that $n^{\epsilon} A_n \xrightarrow{a.s.} 0$; $N_k(\mu, \Sigma)$ denotes the k-variate normal law with mean μ and variance matrix Σ; all vectors are column vectors; the superscript ^{t} denotes transpose; if $x \in \mathbb{R}^k$, with $x^t = (x_1, \ldots, x_k)$, then $Diag(x)$ is the $k \times k$ diagonal matrix whose (i,i) entry is x_i, $1 \leq i \leq k$, and

$$\Sigma_x = Diag(x) - xx^t;$$

I_k denotes the $k \times k$ identity matrix; to simplify notation, all 0s appearing in the paper represent vectors of the appropriate dimension.

2. Some Asymptotic Properties of MPϕEs

Let $X \sim \mathcal{M}_k(n; \pi)$, with $\pi \in \Delta_{0k}$, and let $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)^t$ be the vector of relative frequencies,

$$\hat{\pi}_i = \frac{X_i}{n}, \quad 1 \leq i \leq k. \tag{1}$$

Let $\mathcal{P}$ be a parametric model satisfying Assumption 1 below.

Assumption 1. $\mathcal{P} = \{P(\theta) = (p_1(\theta), \ldots, p_k(\theta))^t, \ \theta \in \Theta\} \subset \Delta_k$, *where* $\Theta \subseteq \mathbb{R}^s$, $k - s - 1 > 0$ *and* $p_1(.), \ldots, p_k(.) : \Theta \longrightarrow \mathbb{R}$ *are known twice continuously differentiable in int*Θ *functions.*

Let $\phi : [0, \infty) \to \mathbb{R} \cup \{\infty\}$ be a continuous convex function. For arbitrary $Q = (q_1, \ldots, q_k)^t \in \Delta_{0k}$ and $P = (p_1, \ldots, p_k)^t \in \Delta_k$, the ϕ-divergence between Q and P is defined by (Csiszár [8])

$$D_{\phi}(Q, P) = \sum_{i=1}^{k} p_i \phi(q_i / p_i).$$

Note that

$$D_{\phi}(Q, P) = \sum_{i/q_i > 0} p_i \phi(q_i / p_i) + \phi(0) \sum_{i/q_i = 0} p_i.$$

The penalized ϕ-divergence for the tuning parameter h between Q and P is defined from the above expression by replacing $\phi(0)$ with h as follows (see Mandal et al. [4]):

$$D_{\phi,h}(Q, P) = \sum_{i/q_i > 0} p_i \phi(q_i / p_i) + h \sum_{i/q_i = 0} p_i.$$

If

$$\hat{\theta}_{\phi,h} = \arg\min_{\theta} D_{\phi,h}(\hat{\pi}, P(\theta)),$$

then $\hat{\theta}_{\phi,h}$ is called the MPϕE of θ.

In order to study some of the properties of $\hat{\theta}_{\phi,h}$, we will assume that ϕ satisfies Assumption 2 below.

Assumption 2. $\phi : [0, \infty) \to \mathbb{R}$ *is a strictly convex function, twice continuously differentiable in* $(0, \infty)$.

Assumption 2 is assumed when dealing with estimators based on minimum divergence, since it lets us take Taylor series expansions of $D_{\phi}(\hat{\pi}, P(\theta))$, which is useful to derive asymptotic properties of the MϕEs. For example, Section 3 of Lindsay [9] assumes that the function ϕ (he calls G what we call ϕ) is a thrice differentiable function (which is stronger than Assumption 2); Theorem 3 in Morales et al. [3]

requires, among other conditions, ϕ to meet Assumption 2 to derive the consistency and asymptotic normality of MϕEs.

Assumption 2 is also assumed in Mandal et al. [4] (they call G what we call ϕ) to study the consistency and asymptotic normality of MPϕEs. Specifically, these authors show that, if $\pi \in \mathcal{P}$ and θ_0 is the true parameter value, then, under suitable regularity conditions including Assumption 2, the MPϕE is consistent for θ_0, and $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$ is asymptotically normal with a mean of 0 and a variance matrix equal to the inverse of the information matrix.

Next we will only assume that $\pi \in \Delta_{0k}$, that is, the assumption that $\pi \in \mathcal{P}$ is dropped. In this context, we prove that the MPϕE is consistent for θ_0, where now θ_0 is the parameter vector that minimizes $D_{\phi,h}(\pi, P(\theta))$, that is to say, $\theta_0 = \arg\min_\theta D_{\phi,h}(\pi, P(\theta))$. Note that θ_0 also depends on ϕ and h, so to be rigorous we should denote it by $\theta_{0,\phi,h}$, but to simplify notation we will simply denote it as θ_0. We also show that $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$ is asymptotically normal with a mean of 0. With this aim, we will also assume the following.

Assumption 3. *$D_{\phi,h}(\pi, P(\theta))$ has a unique minimum at $\theta_0 \in int\Theta$.*

Assumption 3 is assumed in papers on estimators based on minimum divergence estimation. For example, it is Assumption A3(b) in [6], which states, that it is the fundamental identification condition for quasi-maximum likelihood estimators to have a well-defined limit; and it is contained in Assumptions 7 and 9 in [10], required for minimum chi-square estimators to have a well-defined limit; it also coincides with Assumption 30 in [9], imposed for the same reason.

Let θ_0 be as defined in Assumption 3. Then $P(\theta_0)$ is the (ϕ, h)-projection of π on $\mathcal{P}$. Section 3 in [11] shows that Assumption 3 holds for two-way tables when $\mathcal{P}$ is the uniform association model, so the (ϕ, h)-projection always exists for such model. Nevertheless, this projection may not exist, or may not be defined uniquely. See Example 2 in [12] for an instance where there is no unique minimum (because although Θ is that example is convex, the family $\{P(\theta), \theta \in \Theta\}$ is not convex, so the uniqueness of the projection is not guaranteed). Let $\Delta_k(\phi, \mathcal{P}, h) = \{\pi \in \Delta_{0k}$ such that Assumption 3 holds$\}$.

From now on, we will assume that the components of π are sorted so that $\pi_1, \ldots, \pi_m > 0$, and $\pi_{m+1} = \ldots = \pi_k = 0$, for some $1 < m \leq k$, where, if $m = k$, then it is understood that all components of π are positive. We will write $\pi^+ = (\pi_1, \ldots, \pi_m)^t$ and $\hat{\pi}^+ = (\hat{\pi}_1, \ldots, \hat{\pi}_m)^t$. The next result shows the strong consistency and asymptotic normality of the MPϕE.

Theorem 1. *Let $\mathcal{P}$ be a parametric family satisfying Assumption 1. Let ϕ be a real function satisfying Assumption 2. Let $X \sim \mathcal{M}_k(n; \pi)$ with $\pi \in \Delta_k(\phi, \mathcal{P}, h)$. Then*

(a) $\hat{\theta}_{\phi,h} \xrightarrow{a.s.} \theta_0$.

(b) $\sqrt{n}\begin{pmatrix} \hat{\pi}^+ - \pi^+ \\ \hat{\theta}_{\phi,h} - \theta_0 \end{pmatrix} \xrightarrow{\mathcal{L}} N_{m+s}(0, A\Sigma_{\pi^+}A^t)$, *where* $A^t = (I_m, \ G^t)$ *and G is defined in Equation (7). In particular,*

$$\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0) \xrightarrow{\mathcal{L}} N_s(0, G\Sigma_{\pi^+}G^t) \tag{2}$$

(c) $\sqrt{n}\begin{pmatrix} \hat{\pi}^+ - \pi^+ \\ P(\hat{\theta}_{\phi,h}) - P(\theta_0) \end{pmatrix} \xrightarrow{\mathcal{L}} N_{2m}(0, B\Sigma_{\pi^+}B^t)$, *where* $B^t = (I_m, \ G^tD_1(P(\theta_0)))$, *with* $D_1(P(\theta))$ *defined in Equation (8).*

Remark 1. *Observe that, if $m = k$, then the penalization has no effect asymptotically; by contrast, if $m < k$, then the presence of the tuning parameter h influences the covariance matrix of the asymptotic law of $\sqrt{n}(\hat{\theta}_{\phi,h} - \theta_0)$ and $\sqrt{n}(P(\hat{\theta}_{\phi,h}) - P(\theta_0))$.*

Remark 2. *If $\pi \in \mathcal{P}$, we obtain as a particular case the results in Mandal et al. [4]. Our conditions are weaker than those in [4]. The reason is that they allow an infinite number of categories, while we are assuming that such a number is finite, k. Therefore, when the number of categories is finite, the assumptions in [4] for the consistency and asymptotic normality of the MPϕE can be weakened.*

As a consequence of Theorem 1, the following corollary gives the asymptotic behavior of $D_{\phi_1,h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2,h_2}))$, for arbitrary ϕ_1, ϕ_2, and h_1, h_2, that may or may not coincide. Part (a) of Corollary 1, which assumes that the model $\mathcal{P}$ is correctly specified, has been previously proven in [7]. It is included here for the sake of completeness. Part (b), which describes the limit in law under alternatives is, to the best of our knowledge, new.

Corollary 1. *Let $\mathcal{P}$ be a parametric family satisfying Assumption 1. Let ϕ_1 and ϕ_2 be two real functions satisfying Assumption 2. Let $X \sim \mathcal{M}_k(n;\pi)$ with $\pi \in \Delta_k(\phi, \mathcal{P}, h)$.*

(a) For $\pi \in \mathcal{P}$,

$$T = \frac{2n}{\phi_1''(1)}\{D_{\phi_1,h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2,h_2})) - \phi_1(1)\} \xrightarrow{\mathcal{L}} \chi^2_{k-s-1}.$$

(b) For $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$, let $\theta_0 = \arg\min_\theta D_{\phi_2,h_2}(\pi, P(\theta))$. Then

$$W = \sqrt{n}\{D_{\phi_1,h_1}(\hat{\pi}, P(\hat{\theta}_{\phi_2,h_2})) - D_{\phi_1,h_1}(\pi, P(\theta_0))\} \xrightarrow{\mathcal{L}} N(0, \varrho^2)$$

where $\varrho^2 = a^t B \Sigma_\pi B^t a$, with B, as defined in Theorem 1 with $\phi = \phi_2$ and $h = h_2$,

$$a^t = \left(\phi_1'\left(\frac{\pi_1}{p_1(\theta_0)}\right), \ldots, \phi_1'\left(\frac{\pi_m}{p_m(\theta_0)}\right), v_1, \ldots, v_m, \underbrace{h_1, \ldots, h_1}_{k-m \text{ times}}\right),$$

and v_i, $1 \le i \le m$, are as defined in Equation (5) with $\phi = \phi_1$ and $h = h_1$.

Remark 3. *If $\pi \in \mathcal{P}$, the asymptotic behavior of the statistic T does not depend either on ϕ_1, ϕ_2, or on h_1, h_2. In fact, the asymptotic law of T is the same as if non-penalized divergences were used.*

Remark 4. *When $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$, if $m = k$, then the asymptotic distribution of W does not depend on h_1, h_2; by contrast, if $m < k$, then the asymptotic distribution of W does depend on h_1 and h_2.*

Remark 5. *(Properties of the asymptotic test) As a consequence of Corollary 1(a), we have that for testing H_0 vs. H_1, the test that rejects the null hypothesis when $T \ge \chi^2_{k-s-1,1-\alpha}$ is asymptotically correct, in the sense that $P_0(T \ge \chi^2_{k-s-1,1-\alpha}) \to \alpha$, where $\chi^2_{k-s-1,1-\alpha}$ stands for the $1-\alpha$ percentile of the χ^2_{k-s-1} distribution and P_0 stands for the probability when the null hypothesis is true. From Corollary 1(b), it follows that such a test is consistent against fixed alternatives $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$, in the sense that $P(T \ge \chi^2_{k-s-1,1-\alpha}) \to 1$.*

3. Application to Bootstrapping Goodness-Of-Fit Tests

As observed in Remark 5, the test that rejects H_0 when $T \ge \chi^2_{k-s-1,1-\alpha}$ is asymptotically correct and consistent against fixed alternatives. Nevertheless, the χ^2 approximation to the null distribution of the test statistic is rather poor. Next we illustrate this fact with three examples. The last one is motivated by a real data set application in Section 4. All computations have been performed using programs written in the R language [13].

Example 1. *Let $X \sim \mathcal{M}_3(n;\pi)$, with $\pi \in \mathcal{P}$ so that*

$$p_1(\theta) = \frac{1}{3} - \theta,\ p_2(\theta) = \frac{2}{3} - \theta,\ p_3(\theta) = 2\theta,\ 0 < \theta < 1/3.$$

The problem of testing goodness-of-fit to this family is dealt with by considering as test statistic a penalized ϕ_1-divergence and an MPϕ_2E, with ϕ_1 and ϕ_2, two members of the power-divergence family, defined as follows:

$$PD_\lambda(x) = \frac{1}{\lambda(\lambda+1)}\left(x^{(\lambda+1)} - x - \lambda(x-1)\right), \lambda \ne 0, -1,$$

$PD_0(x) = x\log(x) - x + 1$, for $\lambda = 0$, and $PD_{-1}(x) = -\log(x) + x - 1$, for $\lambda = -1$. We thank an anonymous referee for pointing out that the power divergence family is also known as the α-divergence family (see, for example, Section 4 of Amari [14]).

In order to evaluate the performance of the χ^2 approximation to the null distribution of T, we carried out an extensive simulation experiment. As a previous part of the simulation experiment, we evaluated the possible effect of the tuning parameter h_2 on the accuracy of the MPϕ_2E. For this goal, we generated 10,000 samples of size 200 from the parametric family with $\theta = 0.3333$, and calculated the MPϕ_2E with $h_2 = 0.5, 1, 2, 5, 10$ and $\phi_2 = PD_{-2}$, which correspond to the modified chi-square test statistic (see, for example, [1], p. 114). We calculated the root mean square deviation (RMSD) of the resulting estimations,

$$RMSD = \sqrt{\frac{\sum_{i=1}^{10,000} \left(\hat{\theta}_{-2,h_2} - \theta\right)^2}{10,000}},$$

obtaining 0.00156, 0.00128, 0.00128, 0.00128, and 0.00128, respectively. According to these results, there are rather small differences in the performance of the MPϕ_2E for the values of h_2 considered. Because of this, we fixed $\phi_2 = PD_{-2}$ and $h_2 = 0.5, 1, 2$.

Next, to study the goodness of the asymptotic approximation, we generated 10,000 samples of size $n = 100$ from the parametric family with $\theta = 0.3333$, and calculated the test statistic T with $h_1 = h_2 = 0.5$ and $\phi_1(x) = \phi_2(x) = PD_{-2}(x)$, as well as the associated p-values corresponding to the asymptotic null distribution. We then computed the fraction of these p-values, which are less than or equal to the nominal values $\alpha = 0.05, 0.10$ (top and below in tables). This experiment was repeated for $n = 150$, 200, $h_1 = h_2 = 1, 2$, $\phi_1 = PD_1$ (which corresponds to the chi-square test statistic) and $\phi_1 = PD_2$. Table 1 shows the results obtained. We also considered the case $h_1 \neq h_2$, obtaining quite close outcomes. Table 2 displays the results obtained for $n = 200$ and $\phi_1 = \phi_2 = PD_{-2}$. Looking at these tables, we conclude that the asymptotic null distribution does not provide an accurate estimation of the null distribution of T since the type I error probabilities are much greater than the nominal values, 0.05 and 0.10. Therefore, other approximations of the null distribution should be studied.

Table 1. Type I error probabilities obtained using asymptotic approximation for Example 1 with $\theta = 0.3333$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, and $h_1 = h_2 \in \{0.5, 1, 2\}$.

	$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$		
	$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$		
n	**0.5**	**1**	**2**	**0.5**	**1**	**2**	**0.5**	**1**	**2**
100	0.996	0.996	0.998	0.995	0.997	0.996	0.995	0.997	0.997
	0.996	0.996	0.998	0.995	0.997	0.996	0.995	0.997	0.997
150	0.995	0.995	0.996	0.994	0.995	0.996	0.994	0.994	0.995
	0.995	0.995	0.996	0.994	0.995	0.996	0.994	0.994	0.995
200	0.992	0.993	0.994	0.992	0.994	0.991	0.993	0.993	0.994
	0.992	0.994	0.994	0.992	0.994	0.991	0.993	0.993	0.994

Table 2. Type I error probabilities obtained using asymptotic approximation for Example 1 with $n = 200$, $\theta = 0.3333$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	**(0.5, 1)**	**(1, 0.5)**	**(0.5, 2)**	**(2, 0.5)**	**(1, 2)**	**(2, 1)**
	0.989	0.997	0.998	0.998	0.994	0.998
	0.999	0.997	0.998	0.998	0.994	0.999

Example 2. *Let $X \sim \mathcal{M}_3(n; \pi)$, with $\pi \in \mathcal{P}$ so that*

$$p_1(\theta) = 0.5 - 2\theta, \;\; p_2(\theta) = 0.5 + \theta, \;\; p_3(\theta) = \theta, \;\; 0 < \theta < 1/4.$$

We repeated the simulation schedule described in Example 1 for this law with $\theta = 0.24$. Tables 3 and 4 report the obtained results. In contrast to the results for Example 1, where the asymptotic approximation gives a rather liberal test, in this case the resulting test is very conservative. Therefore, we again conclude that the asymptotic null distribution does not provide an accurate estimation of the null distribution of T.

Table 3. Type I error probabilities obtained using asymptotic approximation for Example 2 with $\theta = 0.24$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, and $h_1 = h_2 \in \{0.5, 1, 2\}$.

	$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$		
	$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$		
n	**0.5**	**1**	**2**	**0.5**	**1**	**2**	**0.5**	**1**	**2**
100	0.016	0.017	0.017	0.013	0.013	0.014	0.013	0.014	0.015
	0.034	0.036	0.036	0.031	0.030	0.031	0.030	0.033	0.033
150	0.018	0.019	0.017	0.014	0.014	0.014	0.013	0.015	0.016
	0.035	0.039	0.037	0.031	0.033	0.032	0.035	0.033	0.032
200	0.024	0.022	0.022	0.014	0.016	0.016	0.014	0.015	0.016
	0.043	0.042	0.040	0.032	0.034	0.032	0.032	0.035	0.033

Table 4. Type I error probabilities obtained using asymptotic approximation for Example 2 with $n = 200$, $\theta = 0.24$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	**(0.5, 1)**	**(1, 0.5)**	**(0.5, 2)**	**(2, 0.5)**	**(1, 2)**	**(2, 1)**
	0.017	0.017	0.018	0.019	0.018	0.016
	0.035	0.033	0.035	0.040	0.036	0.034

Example 3. *Let $X \sim \mathcal{M}_4(n; \pi)$, with $\pi \in \mathcal{P}$ so that*

$$p_1(\theta) = \theta^2,\ p_2(\theta) = \theta(1-\theta),\ p_3(\theta) = \theta(1-\theta),\ p_4(\theta) = (1-\theta)^2,\ 0 < \theta < 1. \tag{3}$$

We repeated the simulation schedule described in Example 1 for this law with $\theta = 0.8$. Tables 5 and 6 report the results obtained. Looking at these tables, we see that the test based on asymptotic approximation is liberal, and conclude, as in the previous examples, that other approximations of the null distribution should be considered.

Table 5. Type I error probabilities obtained using asymptotic approximation for Example 3 with $\theta = 0.8$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, and $h_1 = h_2 \in \{0.5, 1, 2\}$.

	$\phi_1 = PD_{-2}$			$\phi_1 = PD_1$			$\phi_1 = PD_2$		
	$h_1 = h_2$			$h_1 = h_2$			$h_1 = h_2$		
n	**0.5**	**1**	**2**	**0.5**	**1**	**2**	**0.5**	**1**	**2**
100	0.063	0.066	0.074	0.095	0.107	0.111	0.122	0.136	0.131
	0.122	0.120	0.125	0.157	0.165	0.161	0.181	0.190	0.182
150	0.063	0.064	0.066	0.083	0.082	0.084	0.099	0.105	0.100
	0.114	0.118	0.113	0.137	0.134	0.136	0.153	0.159	0.152
200	0.062	0.061	0.061	0.075	0.079	0.074	0.086	0.091	0.086
	0.111	0.111	0.115	0.129	0.137	0.123	0.145	0.148	0.144

Table 6. Type I error probabilities obtained using asymptotic approximation for Example 3 with $n = 200$, $\theta = 0.8$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	(0.5, 1)	(1, 0.5)	(0.5, 2)	(2, 0.5)	(1, 2)	(2, 1)
	0.060	0.062	0.063	0.062	0.063	0.058
	0.108	0.114	0.113	0.112	0.113	0.109

The reason for the unsatisfactory results in the three examples is that the asymptotic approximation requires unaffordably large sample sizes when some cells have extremely small probabilities, which provoke the presence of zero cell frequencies. To appreciate this fact, notice that Example 1 requires $n > 30,000$ to obtain expected cell frequencies greater than 10.

Motivated by these examples, the aim of this section is to study another way of approximating the null distribution of T, the bootstrap. The null bootstrap distribution of T is the conditional distribution of

$$T^* = \frac{2n}{\phi_1''(1)}\{D_{\phi_1,h_1}(\hat{\pi}^*, P(\hat{\theta}^*_{\phi_2,h_2})) - \phi_1(1)\},$$

given $(X_1, \ldots, X_k)$, where $\hat{\pi}^*$ is defined as $\hat{\pi}$ with $(X_1, \ldots, X_k)$ replaced by $(X_1^*, \ldots, X_k^*) \sim \mathcal{M}_k(n; P(\hat{\theta}_{\phi_2,h_2}))$, and $\hat{\theta}^*_{\phi_2,h_2} = \arg\min_\theta D_{\phi_2,h_2}(\hat{\pi}^*, P(\theta))$.

Let P_* denote the bootstrap conditional probability law, given $(X_1, \ldots, X_k)$. The next theorem gives the weak limit of T^*.

Theorem 2. *Let $\mathcal{P}$ be a parametric family satisfying Assumption 1. Let ϕ_1 and ϕ_2 be two real functions satisfying Assumption 2. Let $X \sim \mathcal{M}_k(n; \pi)$ with $\pi \in \Delta_k(\phi, \mathcal{P}, h)$. Then*

$$\sup_x |P_*(T^* \leq x) - P(Y \leq x)| \xrightarrow{P} 0$$

where $Y \sim \chi^2_{k-s-1}$.

Recall that, from Corollary 1(a), when H_0 is true, the test statistic T converges in law to a χ^2_{k-s-1} law. Thus, the result in Theorem 2 implies the consistency of the null bootstrap distribution of T as an estimator of the null distribution of T. It is important to remark that the result in Theorem 2 holds whether H_0 is true or not, that is, the bootstrap properly estimates the null distribution, even if the available data does not obey the law in the null hypothesis. This is due to the fact that, under the assumed conditions, the MPϕE always converges to a well-defined limit.

Remark 6. *Properties of the Bootstrap Test. Similarly to Remark 5, as a consequence of Corollary 1(a) and Theorem 2, we have that, for testing H_0 vs. H_1, the test that rejects the null hypothesis when $T \geq T^*_{1-\alpha}$ is asymptotically correct, in the sense that $P_0(T \geq T^*_{1-\alpha}) \to \alpha$, where $T^*_{1-\alpha}$ stands for the $1-\alpha$ percentile of the bootstrap distribution of T. From Corollary 1(b) and Theorem 2, it follows that such a test is consistent against fixed alternatives $\pi \in \Delta_k(\phi_2, \mathcal{P}, h_2) - \mathcal{P}$, in the sense that $P(T \geq T^*_{1-\alpha}) \to 1$.*

In practice, the bootstrap p-value must be approximated by simulation as follows:

1. Calculate the observed value of the test statistic for the available data $(X_1, \ldots, X_k)$, T_{obs}.
2. Generate B bootstrap samples $(X_1^{b*}, \ldots, X_k^{b*}) \sim \mathcal{M}_k(n; P(\hat{\theta}_{\phi_2,h_2}))$, $b = 1, \ldots, B$, and calculate the test statistic for each bootstrap sample obtaining T^{*b}, $b = 1, \ldots, B$.
3. Approximate the p-value by means of the expression

$$\hat{p}_{boot} = \frac{\text{card}\{b : T_b^{*b} \geq T_{obs}\}}{B}.$$

For the numerical experiments previously described, whose results are displayed in Tables 1–6, we also calculated the bootstrap p-values. This was done by generating $B = 1000$ bootstrap samples to approximate each p-value, and calculating the fraction of these p-values, which are less than or equal to 0.05 and 0.10 (top and bottom in the tables). Tables 7–12 display the estimated type I error probabilities obtained by using the bootstrap approximation as well as those obtained with the asymptotic approximation (bootstrap, B, and asymptotic, A, in the tables) taken from Tables 1–6 in order to facilitate the comparison between them. Looking at Tables 7–12, we conclude that the bootstrap approximation is superior to the asymptotic one for small and moderate sample sizes, since in all cases the bootstrap type I error probabilities were closer to the nominal values than those obtained using the asymptotic null distribution. This superior performance of the bootstrap null distribution estimator has been noticed in other inferential problems, where ϕ-divergences are used as test statistics (see, for example, [5,12,15,16]).

Table 7. Asymptotic and bootstrap type I error probabilities for Example 1 with $\theta = 0.3333$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, $h_1 = h_2 \in \{0.5, 1, 2\}$.

	$h_1 = h_2$	0.5		1		2	
ϕ_1	n	B	A	B	A	B	A
PD_{-2}	100	0.051	0.996	0.048	0.996	0.048	0.998
		0.110	0.996	0.103	0.996	0.109	0.998
	150	0.055	0.995	0.050	0.995	0.056	0.996
		0.106	0.995	0.101	0.995	0.109	0.996
	200	0.053	0.992	0.053	0.993	0.056	0.994
		0.103	0.992	0.106	0.994	0.108	0.994
PD_1	100	0.057	0.995	0.056	0.997	0.055	0.996
		0.110	0.995	0.110	0.997	0.107	0.996
	150	0.054	0.994	0.052	0.995	0.055	0.996
		0.110	0.994	0.104	0.995	0.114	0.996
	200	0.055	0.992	0.051	0.994	0.052	0.991
		0.106	0.992	0.103	0.994	0.106	0.991
PD_2	100	0.055	0.995	0.056	0.997	0.054	0.997
		0.110	0.995	0.109	0.997	0.107	0.997
	150	0.054	0.994	0.055	0.994	0.056	0.995
		0.107	0.994	0.106	0.994	0.110	0.995
	200	0.054	0.993	0.053	0.993	0.055	0.994
		0.107	0.993	0.105	0.993	0.108	0.994

Table 8. Asymptotic and bootstrap type I error probabilities for Example 1 with $n = 200$, $\theta = 0.3333$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	A	B	A	B	A	B	A	B	A	B	A
	0.061	0.989	0.050	0.997	0.059	0.996	0.042	0.998	0.044	0.994	0.063	0.998
	0.107	0.999	0.113	0.997	0.106	0.996	0.095	0.998	0.105	0.994	0.115	0.999

Table 9. Asymptotic and bootstrap type I error probabilities for Example 2 with $\theta = 0.24$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, and $h_1 = h_2 \in \{0.5, 1, 2\}$.

$h_1 = h_2$		0.5		1		2	
ϕ_1	n	B	A	B	A	B	A
PD_{-2}	100	0.057	0.016	0.055	0.017	0.051	0.017
		0.111	0.034	0.110	0.036	0.102	0.036
	150	0.049	0.018	0.048	0.019	0.051	0.017
		0.097	0.035	0.103	0.039	0.101	0.036
	200	0.051	0.024	0.055	0.022	0.051	0.022
		0.099	0.043	0.102	0.042	0.099	0.040
PD_1	100	0.058	0.013	0.054	0.013	0.051	0.014
		0.114	0.031	0.113	0.030	0.106	0.031
	150	0.050	0.014	0.051	0.014	0.052	0.014
		0.098	0.031	0.103	0.031	0.100	0.032
	200	0.049	0.014	0.054	0.016	0.052	0.016
		0.099	0.032	0.104	0.034	0.099	0.032
PD_2	100	0.055	0.013	0.053	0.014	0.050	0.015
		0.110	0.030	0.108	0.033	0.104	0.033
	150	0.050	0.013	0.052	0.015	0.051	0.016
		0.097	0.032	0.103	0.033	0.098	0.032
	200	0.049	0.014	0.051	0.015	0.051	0.016
		0.100	0.032	0.102	0.035	0.098	0.033

Table 10. Asymptotic and bootstrap type I error probabilities for Example 2 with $n = 200$, $\theta = 0.24$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	A	B	A	B	A	B	A	B	A	B	A
	0.048	0.017	0.051	0.017	0.052	0.018	0.053	0.019	0.050	0.018	0.049	0.016
	0.101	0.035	0.099	0.033	0.100	0.035	0.105	0.040	0.103	0.036	0.101	0.034

Table 11. Asymptotic and bootstrap type I error probabilities for Example 3 with $\theta = 0.8$, $\phi_1 = PD_\lambda$, $\lambda \in \{-2, 1, 2\}$, $\phi_2 = PD_{-2}$, and $h_1 = h_2 \in \{0.5, 1, 2\}$.

$h_1 = h_2$		0.5		1		2	
ϕ_1	n	B	A	B	A	B	A
PD_{-2}	100	0.066	0.063	0.058	0.066	0.044	0.074
		0.119	0.122	0.101	0.120	0.086	0.125
	150	0.053	0.063	0.050	0.064	0.045	0.066
		0.098	0.114	0.095	0.118	0.093	0.113
	200	0.051	0.062	0.047	0.061	0.046	0.061
		0.099	0.111	0.096	0.111	0.100	0.115
PD_1	100	0.049	0.095	0.049	0.107	0.041	0.111
		0.103	0.157	0.098	0.065	0.084	0.161
	150	0.050	0.083	0.040	0.082	0.040	0.084
		0.098	0.137	0.090	0.134	0.087	0.136
	200	0.046	0.075	0.048	0.079	0.044	0.074
		0.095	0.129	0.102	0.137	0.092	0.123
PD_2	100	0.043	0.122	0.045	0.136	0.037	0.131
		0.099	0.181	0.046	0.190	0.077	0.182
	150	0.040	0.099	0.047	0.105	0.035	0.100
		0.041	0.153	0.093	0.159	0.081	0.152
	200	0.043	0.086	0.048	0.091	0.043	0.086
		0.092	0.145	0.097	0.148	0.090	0.144

Table 12. Asymptotic and bootstrap type I error probabilities for Example 3 with $n = 200$, $\theta = 0.8$, $\phi_1 = \phi_2 = PD_{-2}$, $h_1 \neq h_2$, and $h_1, h_2 \in \{0.5, 1, 2\}$.

(h_1, h_2)	(0.5, 1)		(1, 0.5)		(0.5, 2)		(2, 0.5)		(1, 2)		(2, 1)	
	B	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**
	0.047	0.060	0.048	0.062	0.051	0.063	0.049	0.062	0.048	0.063	0.044	0.058
	0.095	0.108	0.099	0.114	0.099	0.113	0.097	0.112	0.099	0.113	0.092	0.109

4. Application to the Evaluation of the Thematic Classification in Global Land Cover Maps

This section displays the results of an application of our proposal to two real data sets related to the thematic quality assessment of a global land cover (GLC) map. The data comprise the results of two thematic classifications of the land cover category "Evergreen Broadleaf Trees" (EBL) and summarize the number of sample units correctly classified in this class, and the number of confusions with other land cover classes: "Deciduous Broadleaf Trees" (DBL), "Evergreeen Needleleaf Trees" (ENL), and "Urban/Built Up" (U). The results of these two classifications were collected from two different global land cover maps: the Globcover map and the LC-CCI map (see Tsendbazar et al. [17] for additional details) and they are displayed in Table 13.

Table 13. Thematic classification of the Evergreen Broadleaf Trees (EBL) class.

		Globcover Map	**LC-CCI Map**
Classified Data	EBL	165	172
	DBL	13	5
	ENL	7	5
	U	0	0

Parametric specifications of the multinomial vector of probabilities are quite attractive since they describe in a concise way the classification pattern. Because of this, given the similarity between the two observed classifications in Table 13, we are interested in the search of a parametric model suitable to depict the thematic accuracy of this class in both GLC maps. For this purpose, we consider the parametric family in Equation (3) of Example 3. The presence of a zero cell frequency in each data set leads us to consider a penalized ϕ-divergence as a test statistic for testing goodness-of-fit to such a parametric family.

Table 14 displays the observed values of the test statistic T and the associated bootstrap p-values for the goodness-of-fit test with respect to the parametric family in Equation (3) for the two observed classifications of the EBL class in Table 13. Looking at this table, it can be concluded that the null hypothesis cannot be rejected in both cases. Therefore, the parametric model in Equation (3) provides an adequate description of the thematic classification of the EBL class.

Table 14. Results of the goodness-of-fit test applied to the thematic classification of the EBL class.

	Globcover Map			**LC-CCI Map**		
	$\hat{\theta}_{-2,0.5} = 0.9490$			$\hat{\theta}_{-2,0.5} = 0.9721$		
ϕ_1	PD_{-2}	PD_1	PD_2	PD_{-2}	PD_1	PD_2
T_{obs}	2.3015	2.7618	3.0111	0.1432	0.1432	0.1433
$\hat{p}_{boot}$	0.1700	0.2253	0.2926	0.9283	0.9200	0.9148
	$\hat{\theta}_{-2,1} = 0.9503$			$\hat{\theta}_{-2,1} = 0.9725$		
T_{obs}	2.7686	3.3752	3.6962	0.2821	0.2823	0.2826
$\hat{p}_{boot}$	0.1801	0.2325	0.2671	0.8431	0.9162	0.9182
	$\hat{\theta}_{-2,2} = 0.9527$			$\hat{\theta}_{-2,2} = 0.9732$		
T_{obs}	3.6352	4.5400	5.0219	0.5492	0.5508	0.5514
$\hat{p}_{boot}$	0.1300	0.2492	0.2584	0.7526	0.8144	0.8291

5. Proofs

Notice that

$$\begin{aligned} D_{\phi,h}(\pi, P(\theta)) &= \sum_{i=1}^{m} p_i(\theta)\phi\left(\frac{\pi_i}{p_i(\theta)}\right) + h\sum_{i=m+1}^{k} p_i(\theta) \\ &= h\mathrm{I}(m<k) + \sum_{i=1}^{m} p_i(\theta)\phi_h\left(\frac{\pi_i}{p_i(\theta)}\right) \end{aligned}$$

where I stands for the indicator function, $\phi_h(x) = \phi(x) - h$, if $m < k$, and $\phi_h(x) = \phi(x)$, if $m = k$. Let

$$D^{+}_{\phi,h}(\pi, P(\theta)) = \sum_{i=1}^{m} p_i(\theta)\phi_h\left(\frac{\pi_i}{p_i(\theta)}\right).$$

Clearly,

$$\arg\min_{\theta} D_{\phi,h}(\hat{\pi}, P(\theta)) = \arg\min_{\theta} D^{+}_{\phi,h}(\hat{\pi}, P(\theta)).$$

Note that, if Assumptions 1 and 2 hold, then Assumption 3 implies that

$$\frac{\partial}{\partial\theta} D^{+}_{\phi}(\pi, P(\theta_0)) = \sum_{i=1}^{m} \frac{\partial}{\partial\theta} p_i(\theta_0) v_i = 0 \tag{4}$$

where

$$v_i = \phi\left(\frac{\pi_i}{p_i(\theta_0)}\right) - \frac{\pi_i}{p_i(\theta_0)}\phi'\left(\frac{\pi_1}{p_i(\theta_0)}\right) - h\mathrm{I}(m<k) \tag{5}$$

$1 \le i \le m$, and $\phi'(x) = \frac{\partial}{\partial x}\phi(x)$. The $s \times s$ matrix

$$\mathbb{D}_2 = \frac{\partial^2}{\partial\theta\partial\theta^t} D^{+}_{\phi}(\pi, P(\theta_0)) = \sum_{i=1}^{m} \frac{\partial^2}{\partial\theta\partial\theta^t} p_i(\theta_0) v_i + \sum_{i=1}^{m} \frac{\partial}{\partial\theta} p_i(\theta_0) \frac{\partial}{\partial\theta} p_i(\theta_0)^t w_i \tag{6}$$

is positive definite, where

$$w_i = \frac{\pi_i^2}{p_i^3(\theta_0)}\phi''\left(\frac{\pi_i}{p_i(\theta_0)}\right),$$

$1 \le i \le m$, and $\phi''(x) = \frac{\partial^2}{\partial x^2}\phi(x)$. Therefore, by the Implicit Function Theorem (see, for example, Dieudonne [18], p. 272), there is an open neighborhood $U \subseteq (0,1)^m$ of π^+ and s unique functions, $g_i : U \to \mathbb{R}$, $1 \le i \le s$, so that

(i) $\hat{\theta}_{\phi} = (g_1(\hat{\pi}^+), \ldots, g_s(\hat{\pi}^+))^t$, $\forall\ n \ge n_0$, for some $n_0 \in \mathbb{N}$;

(ii) $\theta_0 = (g_1(\pi^+), \ldots, g_s(\pi^+))^t$;
(iii) $g = (g_1, \ldots, g_s)^t$ is continuously differentiable in U and the $s \times m$ Jacobian matrix of g at $(\pi_1, \ldots, \pi_m)$ is given by

$$G = \mathbb{D}_2^{-1} D_1(P(\theta_0)) Diag(\varpi) \tag{7}$$

where

$$D_1(P(\theta)) = \left(\frac{\partial}{\partial \theta} p_1(\theta), \ldots, \frac{\partial}{\partial \theta} p_m(\theta) \right), \tag{8}$$

$\varpi = (\varpi_1, \ldots, \varpi_m)^t$,

$$\varpi_i = \frac{\pi_i}{p_i^2(\theta_0)} \phi'' \left(\frac{\pi_i}{p_i(\theta_0)} \right),$$

and $1 \leq i \leq m$.

Proof of Theorem 1. Part (a) follows from (i) and (ii) above and the fact that $\hat{\pi}^+ \to \pi^+$ a.s. From (i)–(ii), and taking into account that $\sqrt{n}(\hat{\pi}^+ - \pi^+)$ is asymptotically normal, it follows that

$$\hat{\theta}_\phi = \theta_0 + G(\pi, P(\theta_0), \phi)(\hat{\pi} - \pi) + o_P(n^{-1/2}). \tag{9}$$

Parts (b) and (c) follow from Equation (9) and the asymptotic normality of $\sqrt{n}(\hat{\pi}^+ - \pi^+)$. □

Proof of Corollary 1. Part (a) was shown in Theorem 5.1 in [7]. To prove (b), we first demonstrate that

$$W = W_0 + r_n \tag{10}$$

where

$$W_0 = \sqrt{n} \left\{ \sum_{j=1}^{m} p_j(\hat{\theta}_{\phi_2,h_2}) \phi_1 \left(\frac{\hat{\pi}_j}{p_j(\hat{\theta}_{\phi_2,h_2})} \right) + h_1 \sum_{j=m+1}^{k} p_j(\hat{\theta}_{\phi_2,h_2}) - D_{\phi_1,h_1}(\pi, P(\theta_0)) \right\} + r_n,$$

and $r_n = o_P(1)$. Notice that

$$\begin{aligned} r_n &= \sqrt{n}\{h_1 - \phi_1(0)\} \sum_{j:\hat{\pi}_j=0,\, \pi_j>0} p_j(\hat{\theta}_{\phi_2,h_2}) \\ &= \sqrt{n}\{h_1 - \phi_1(0)\} \sum_{j=1}^{m} p_j(\hat{\theta}_{\phi_2,h_2}) \mathrm{I}(\hat{\pi}_j = 0). \end{aligned}$$

Therefore,

$$0 \leq E|r_n| \leq \sqrt{n}|h_1 - \phi_1(0)| \sum_{j=1}^{m} P(\hat{\pi}_j = 0) = \sqrt{n}|h_1 - \phi_1(0)| \sum_{j=1}^{m} (1 - \pi_j)^n \to 0,$$

which implies $r_n = o_P(1)$. From Theorem 1 and Taylor expansion, it follows that $W_0 \xrightarrow{\mathcal{L}} N(0, \varrho^2)$; hence, the result in part (b) is proven. □

Proof of Theorem 2. The proof of Theorem 2 is parallel to that of Theorem 2 in [5], so we omit it. □

Author Contributions: M.V. Alba-Fernández and M.D. Jiméz-Gamero conceived and designed the experiments; M.V. Alba-Fernández performed the experiments; M.V. Alba-Fernández and F.J. Ariza-López analyzed the data; F.J. Ariza-López contributed materials; M.V. Alba-Fernández and M.D. Jiméz-Gamero wrote the paper.

Acknowledgments: The authors thank the anonymous referees for their valuable time and careful comments, which improved the presentation of this paper. The research in this paper has been partially funded by grants: CTM2015–68276–R of the Spanish Ministry of Economy and Competitiveness (M.V. Alba-Fernández and F.J. Ariza-López) and MTM2017-89422-P of the Spanish Ministry of Economy, Industry and Competitiveness, ERDF support included (M.D. Jiménez-Gamero).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLE	maximum likelihood estimator
MϕE	minimum ϕ-divergence estimator
MPϕE	minimum penalized ϕ-divergence estimator
RMSD	root mean square deviation
B	bootstrap
A	asymptotic
GLC	global land cover
EBL	evergreen broadleaf trees
DBL	deciduous broadleaf trees
ENL	evergreeen needleleaf trees
U	urban/built up

References

1. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall: London, UK; CRC Press: Boca Raton, FL, USA, 2006.
2. Basu, A.; Sarkar, S. On disparity based goodness-of-fit tests for multinomial models. *Stat. Probab. Lett.* **1994**, *19*, 307–312. [CrossRef]
3. Morales, D.; Pardo, L.; Vajda, I. Asymptotic divergence of estimates of discrete distributions. *J. Stat. Plann. Inference* **1995**, *48*, 347–369. [CrossRef]
4. Mandal, A.; Basu, A.; Pardo, L. Minimum disparuty inference and the empty cell penalty: Asymptotic results. *Sankhya Ser. A* **2010**, *72*, 376–406. [CrossRef]
5. Jiménez-Gamero, M.D.; Pino-Mejías, R.; Alba-Fernández, M.V.; Moreno-Rebollo, J.L. Minimum ϕ-divergence estimation in misspecified multinomial models. *Comput. Stat. Data Anal.* **2011**, *55*, 3365–3378. [CrossRef]
6. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
7. Mandal, A.; Basu, A. Minimum disparity inference and the empty cell penalty: Asymptotic results. *Electron. J. Stat.* **2011**, *5*, 1846–1875. [CrossRef]
8. Csiszár, I. Information type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
9. Lindsay, B.G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [CrossRef]
10. Vuong, Q.H.; Wang, W. Minimum χ-square estimation and tests for model selection. *J. Econom.* **1993**, *56*, 141–168. [CrossRef]
11. Alba-Fernández, M.V.; Jiménez-Gamero, M.D.; Lagos-Álvarez, B. Divergence statistics for testing uniform association in cross-classifications. *Inf. Sci.* **2010**, *180*, 4557–4571. [CrossRef]
12. Jiménez-Gamero, M.D.; Pino-Mejías, R.; Rufián-Lizana, A. Minimum K_ϕ-divergence estimators for multinomial models and applications. *Comput. Stat.* **2014**, *29*, 363–401. [CrossRef]
13. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: https://www.R-project.org/ (accessed on 29 April 2018).
14. Amari, S. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796. [CrossRef] [PubMed]
15. Alba-Fernández, M.V.; Jiménez-Gamero, M.D. Bootstrapping divergence statistics for testing homogeneity in multinomial populations. *Math. Comput. Simul.* **2009**, *79*, 3375–3384. [CrossRef]
16. Jiménez-Gamero, M.D.; Alba-Fernández, M.V.; Barranco-Chamorro, I.; Muñoz-García, J. Two classes of divergence statistics for testing uniform association. *Statistics* **2014**, *48*, 367–387. [CrossRef]

17. Tsendbazar, N.E.; de Bruina, S.; Mora, B.; Schoutenc, L.; Herolda, M. Comparative assessment of thematic accuracy of GLC maps for specific applications using existing reference data. *Int. J. Appl. Earth. Obs. Geoinf.* **2016**, *44*, 124–135. [CrossRef]
18. Dieudonne, J. *Foundations of Modern Analysis*; Academic Press: New York, NY, USA; London, UK, 1969.