Asymptotic Properties of a Statistical Estimator of the Jeffreys Divergence: The Case of Discrete Distributions

Vladimir Glinskiy [1,2], Artem Logachov [1,3], Olga Logachova [4], Helder Rojas [5,6], Lyudmila Serga [1,2,*] and Anatoly Yambartsev [7]

1   Department of Business Analytics, Siberian Institute of Management—Branch of the Russian Presidential Academy of National Economy and Public Administration, Novosibirsk State University of Economics and Management, 630102 Novosibirsk, Russia; glinskiy-vv@ranepa.ru (V.G.); logachev-av@ranepa.ru (A.L.)
2   Department of Statistics, Novosibirsk State University of Economics and Management, 630099 Novosibirsk, Russia
3   Department of Computer Science in Economics, Novosibirsk State Technical University (NSTU), 630087 Novosibirsk, Russia
4   Department of Higher Mathematics, Siberian State University of Geosystems and Technologies (SSUGT), 630108 Novosibirsk, Russia; o.m.logacheva@sgugit.ru
5   Escuela Profesional de Ingeniería Estadística, Universidad Nacional de Ingeniería, Lima 00051, Peru; h.rojas-molina23@imperial.ac.uk
6   Department of Mathematics, Imperial College London, London SW7 2AZ, UK
7   Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo (USP), São Paulo 05508-220, Brazil; yambar@usp.br
*   Correspondence: serga-lk@ranepa.ru

Abstract: We investigate the asymptotic properties of the plug-in estimator for the Jeffreys divergence, the symmetric variant of the Kullback–Leibler (KL) divergence. This study focuses specifically on the divergence between discrete distributions. Traditionally, estimators rely on two independent samples corresponding to two distinct conditions. However, we propose a one-sample estimator where the condition results from a random event. We establish the estimator's asymptotic unbiasedness (law of large numbers) and asymptotic normality (central limit theorem). Although the results are expected, the proofs require additional technical work due to the randomness of the conditions.

1. Introduction

The Kullback–Leibler (KL) divergence was developed as an extension of Shannon's information; see [1]. This measure is widely used in various research fields where there is a need to compare two (probability) measures. Significant research has been focused on generalizing this concept. KL divergence is considered nowadays as a special case of f-divergence measures, with $f(\cdot) = \ln(\cdot)$. In turn, this class of divergence is a part of a broader class of divergences known as information divergences. For a literature review of various classes of information divergence measures, we refer the reader to [2]. For the estimation of f-divergence, refer to [3] and the references therein.

In mathematical statistics, the Kullback–Leibler (KL) divergence introduced in [1], also called relative entropy in information theory, is a type of statistical measure used to quantify the dissimilarity between two probability measures. This measure of divergence has been widely used in various fields, such as variational inference [4,5], Bayesian inference [6,7], metric learning [8–10], machine learning [11,12], computer vision [13,14], physics [15], biology [16], and information geometry [17], among many other application fields. It is worth mentioning works related to the application of various types of divergences for

goodness-of-fit problems, where theoretical and empirical distributions are compared. For instance, see [18–20] and the references therein.

We consider an estimation problem of the symmetrized version of KL divergence known as Jeffreys divergence. There is a vast body of literature dedicated to the statistical estimation of information-type divergence. Most of the works are dedicated to continuous-type distributions, because, in general, it is hard to calculate the definite integral in the definition of the convergence. See [21,22] and references within. However, here, we focused on discrete (or categorical) random variables.

Let $\mathcal{X} = \{a_1, a_2, \ldots, a_r\}$ be a finite set, where $r \geq 2$. A random variable with values from $\mathcal{X}$ is characterized by the mass distribution $\mathbf{p} = (p_a)_{a \in \mathcal{X}}$. Let $\mathcal{P}(\mathcal{X})$ be the set of all (positive) distributions on $\mathcal{X}$,

$$\mathcal{P}(\mathcal{X}) = \left\{ \mathbf{p} = (p_a)_{a \in \mathcal{X}} \in \mathbb{R}^r : p_a > 0 \quad \text{and} \quad \sum_{a \in \mathcal{X}} p_a = 1 \right\}.$$

For any two $\mathbf{p}, \mathbf{q} \in \mathcal{P}(\mathcal{X})$, the KL divergence between $\mathbf{p}$ and $\mathbf{q}$ is defined as follows

$$\mathcal{D}_{KL}(\mathbf{p}||\mathbf{q}) = \sum_{a \in \mathcal{X}} p_a \ln \left(\frac{p_a}{q_a} \right).$$

KL divergence is asymmetric, and the values of $\mathcal{D}_{KL}(\mathbf{p}||\mathbf{q})$ and $\mathcal{D}_{KL}(\mathbf{q}||\mathbf{p})$ are different in general. This lack of symmetry, in some specific contexts, can be a disadvantage when measuring similarities between probability measures; see, e.g., [17,23–28]. Thus, it is often quite useful to work with a symmetrization of the KL divergence which is defined as

$$\begin{aligned} \mathcal{D}_{KL}^{sym}(\mathbf{p}||\mathbf{q}) \quad &:= \quad \mathcal{D}_{KL}(\mathbf{p}||\mathbf{q}) + \mathcal{D}_{KL}(\mathbf{q}||\mathbf{p}), \\ &= \quad \sum_{a \in \mathcal{X}} (p_a - q_a) \ln \left(\frac{p_a}{q_a} \right). \end{aligned}$$

In [1], Kullback and Leibler consider this symmetrized version and refer to it as the divergence between the two measures. This concept was earlier introduced and studied by Harold Jeffreys in 1948, with some papers and Wikipedia citing the second edition of [29]. Today, this symmetrization is known as Jeffreys divergence. For some advantages and applications of including symmetrization in KL divergence, see, for example, [30–35]. Additionally, the Jeffreys divergence is related to the Population Stability Index (PSI) used in finance and serves as a foundation of the Cluster Validity Index (CVI), as discussed in [36].

The problem of estimating information divergence (and entropy) for discrete random variables is well established in the literature; see, for example, [37–43]. It is important to note that these references do not represent a comprehensive list of all relevant works; they are merely examples that illustrate the diversity of approaches in this field. In these works, the authors examine the convergence properties of estimators derived through various methods for discrete distributions. They analyze the asymptotic behavior and provide theoretical insights into the performance of these estimators, contributing to a deeper understanding of various types of divergence and related measures. All such studies known to us generally operate within the following framework. Given two independent samples $X_1^{(1)}, X_2^{(1)}, \ldots, X_l^{(1)}$ and $X_1^{(2)}, X_2^{(2)}, \ldots, X_m^{(2)}$ drawn from two discrete distributions $\mathbf{p}$ and $\mathbf{q}$, respectively, the goal is to develop an estimator for the divergence based on the sample sizes l and m and to analyze their convergence.

However, in practice, the following scenario arises. It involves paired sample observations: $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where $Y_i \in \{0, 1\}$ indicates a binary condition, and $\mathbf{p}$ and $\mathbf{q}$ represent the conditional distributions $\mathbf{P}(\cdot \mid Y = 0)$ and $\mathbf{P}(\cdot \mid Y = 1)$, respectively. In other words, the sample sizes l and m mentioned above are realizations of a random experiment and should be treated as random variables, where l has a binomial

distribution with success probability $\mathbf{P}(Y = 0)$ in a sequence of n trials, and $m = n - l$. Below, in Section 2.1, we consider an example of this scenario.

This paper specifically focuses on this scenario, establishing the asymptotic properties of the plug-in estimator within this framework. To the best of our knowledge, there are few theoretical works on the asymptotic properties of estimating the Jeffreys divergence, and none directly address the framework we consider. While the standard δ-method can be employed to derive the central limit theorem (CLT), it typically yields a general and cumbersome variance formula involving the product of large matrices. In contrast, our direct and probabilistic approach provides a more explicit and straightforward variance formula. An additional benefit of our approach is that the results obtained can be effectively integrated into undergraduate statistics courses.

The paper is organized as follows. In the next section, Section 2, we introduce definitions and formulate the asymptotic results with proofs. Auxiliary results are provided in Section 3.

2. Notations and Main Results

Let (X, Y) be a bivariate random vector with values from $\mathcal{X} \times \{0, 1\}$ defined on the probability space $(\Omega, \mathcal{A}, \mathbf{P})$. Recall that $\mathcal{X} = \{a_1, a_2, \dots, a_r\}$ is a finite set, where $r \geq 2$. We denote by $\mathbf{E}$ the expectation with respect to $\mathbf{P}$. The (marginal) distribution of Y is a Bernoulli distribution with success probability $p := \mathbf{P}(Y = 1) \in (0, 1)$, and let $q := 1 - p = \mathbf{P}(Y = 0)$. Let $\mathbf{p} = (p_j)_{1 \leq j \leq r}$ and $\mathbf{q} = (q_j)_{1 \leq j \leq r}$ two conditional distributions of X,

$$p_j := \mathbf{P}(X = a_j \mid Y = 1) \quad \text{and} \quad q_j := \mathbf{P}(Y = a_j \mid Y = 0),$$

for any $j = 1, \dots, r$. We will assume that all probabilities above are positive,

$$\min_{1 \leq j \leq r} \{p_j, q_j\} > 0,$$

and $\mathbf{p} \neq \mathbf{q}$, i.e. there exists at least one j, such that $p_j \neq q_j$. Moreover, we do not assume that the number r is known. Consider the empirical probability measures $\hat{\mathbf{p}}_n = (\hat{p}_{j,n})_{1 \leq j \leq r}$ and $\hat{\mathbf{q}}_n = (\hat{q}_{j,n})_{1 \leq j \leq r}$, generated by the sequence i.i.d. random variables $(X_1, Y_1) \dots, (X_n, Y_n)$ from the bivariate distribution, $(\mathbf{p}, \mathbf{q})$, defined from

$$\hat{p}_{j,n} := \frac{\sum\limits_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)}{\sum\limits_{i=1}^{n} \mathbf{I}(Y_i = 1)}, \qquad \hat{q}_{j,n} := \frac{\sum\limits_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 0)}{\sum\limits_{i=1}^{n} \mathbf{I}(Y_i = 0)}, \tag{1}$$

where $\mathbf{I}(A)$ is an indicator of an event A. We assume that the above fractions are zero when the corresponding denominator takes the zero value. Additionally, consider the following notations

$$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(Y_i = 1), \qquad \hat{q}_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(Y_i = 0). \tag{2}$$

Based on (1), we define the plug-in estimator for the Jeffreys divergence $\mathcal{D}_{KL}^{sym}(\mathbf{p}||\mathbf{q})$ as follows,

$$\mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n || \hat{\mathbf{q}}_n) := \sum_{j=1}^{r} (\hat{p}_{j,n} - \hat{q}_{j,n}) \ln \frac{\hat{p}_{j,n}}{\hat{q}_{j,n}}. \tag{3}$$

where the standard convention is adopted: if $x \geq 0$, then $0 \log(0/x) = 0$, and if $x > 0$, then $x \log(x/0) = \infty$. Note that the estimator appears to depend on r. However, for any symbol $a_j \in \mathcal{X}$ with $\hat{p}_{j,n} = 0$ or $\hat{q}_{j,n} = 0$, the corresponding term is not included in the sum. Thus, the sum contains only the terms corresponding to the observed symbols from $\mathcal{X}$, and we do not need to know r.

In this paper, we study the convergence properties, specifically, the asymptotic unbiasedness, or law of large numbers (LLN), and asymptotic normality, or central limit theorem (CLT), of the plug-in estimator (3). To this end, denote

$$
\begin{aligned}
\eta_n \quad &:= \quad \mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n || \hat{\mathbf{q}}_n) - \mathcal{D}_{KL}^{sym}(\mathbf{p} || \mathbf{q}) \\
&= \quad \sum_{j=1}^{r} \left((\hat{p}_{j,n} - \hat{q}_{j,n}) \ln \frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} - (p_j - q_j) \ln \frac{p_j}{q_j} \right).
\end{aligned}
$$

Theorem 1. *(Law of Large Numbers) The following equality holds*

$$
\lim_{n \to \infty} \eta_n = 0 \quad a.s.
$$

Proof. It follows directly from the strong law of large numbers for sequences $\hat{p}_n$, $\hat{q}_n$, $\hat{p}_{j,n}$, $\hat{q}_{j,n}$, $1 \le j \le r$, defined in (1) and (2). $\quad\square$

Theorem 2. *(Central Limit Theorem) The following convergence holds true*

$$
\sqrt{n}\eta_n \xrightarrow{d} \xi \sim N(0, \sigma^2), \text{ as } n \to \infty,
$$

where $\xrightarrow{d}$ means the convergence in distribution,

$$
\begin{aligned}
\sigma^2 = \mathbf{E} \Bigg(\sum_{j=1}^{r} \Bigg(&\left(\frac{1}{p} (\mathbf{I}(X = a_j, Y = 1) - pp_j) - p_j (\mathbf{I}(Y = 1) - p) \right) \left(1 + \ln \frac{p_j}{q_j} - \frac{q_j}{p_j} \right) \\
&+ \left(\frac{1}{q} (\mathbf{I}(X = a_j, Y = 0) - qq_j) - q_j (\mathbf{I}(Y = 0) - q) \right) \left(1 + \ln \frac{q_j}{p_j} - \frac{p_j}{q_j} \right) \Bigg) \Bigg)^2.
\end{aligned}
$$

The plug-in estimator of variance yields the following corollary.

Corollary 1. *The following convergence holds true*

$$
\sqrt{n} \frac{\eta_n}{\hat{\sigma}_n} \xrightarrow{d} \tilde{\xi} \sim N(0, 1), \text{ as } n \to \infty,
$$

where

$$
\begin{aligned}
\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\sum_{j=1}^{r} \Bigg[&\left(\frac{1}{\hat{p}_n} (\mathbf{I}(X_i = a_j, Y_i = 1) - \hat{p}_n \hat{p}_{j,n}) - \hat{p}_{j,n} (\mathbf{I}(Y_i = 1) - \hat{p}_n) \right) \left(1 + \ln \frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} - \frac{\hat{q}_{j,n}}{\hat{p}_{j,n}} \right) \\
&+ \left(\frac{1}{\hat{q}_n} (\mathbf{I}(X_i = a_j, Y_i = 0) - \hat{q}_n \hat{q}_{j,n}) - \hat{q}_{j,n} (\mathbf{I}(Y_i = 0) - \hat{q}_n) \right) \left(1 + \ln \frac{\hat{q}_{j,n}}{\hat{p}_{j,n}} - \frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} \right) \Bigg] \Bigg)^2.
\end{aligned}
$$

Indeed, the sequence $\hat{\sigma}_n^2$, by the law of large numbers, converges in probability to σ^2, and hence the proof of the corollary directly follows from Theorem 2 and Slutsky's theorem. Recall that Slutsky's theorem states that if a sequence of random variables converges in distribution and another sequence converges in probability to a constant, then their product (or sum) also converges in distribution, which is applicable in this context to complete the proof of this corollary.

Let us now proceed with the proof of Theorem 2.

Proof. To prove the theorem, we will show that the asymptotics of the sequence of interest matches those of the sequence in Lemma 3, involving the differences $\hat{p}_{j,n} - p_j$ and $\hat{q}_{j,n} - q_j$ for $j = 1, \ldots, r$. To achieve this, we need some technical work, isolating the relevant term

from Lemma 3 and the term that converges to zero in probability. We rewrite the $\sqrt{n}\eta_n$ in the following way:

$$
\begin{aligned}
\sqrt{n}\eta_n =& \sqrt{n}\sum_{j=1}^{r}\left((\hat{p}_{j,n}-\hat{q}_{j,n})\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} - (p_j-q_j)\ln\frac{p_j}{q_j}\right)\\
=& \sqrt{n}\sum_{j=1}^{r}\left((\hat{p}_{j,n}-\hat{q}_{j,n})\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} \pm (p_j-q_j)\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} - (p_j-q_j)\ln\frac{p_j}{q_j}\right)\\
=& \sqrt{n}\sum_{j=1}^{r}((\hat{p}_{j,n}-p_j)-(\hat{q}_{j,n}-q_j))\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} + \sqrt{n}\sum_{j=1}^{r}(p_j-q_j)\ln\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\\
=& \sqrt{n}\sum_{j=1}^{r}((\hat{p}_{j,n}-p_j)-(\hat{q}_{j,n}-q_j))\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} + \sqrt{n}\sum_{j=1}^{r}(p_j-q_j)\ln\left(1+\frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right).
\end{aligned}
\tag{4}
$$

Denote

$$
y_{j,n} := \ln\left(1+\frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right).
$$

We will use the Taylor expansion for a function $f(x)$ with a remainder term in Lagrange form (see e.g., [44] (p. 880)): for any x with $|x| < 1$, there exists a positive $K_x \in (0,1)$ such that

$$
f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(K_x x)}{2!}x^2.
$$

Applying this to the function $f(x) = \ln(1+x)$, we have

$$
\ln(1+x) = x - \frac{x^2}{2(1+K_x x)^2},
\tag{5}
$$

for $|x| < 1$, where $K_x \in (0,1)$. It is easy to see that the following upper bound holds when $|x| \le \frac{1}{2}$ and $K_x \in (0,1)$:

$$
\frac{x^2}{2(1+K_x x)^2} \le 2x^2.
\tag{6}
$$

Applying (5) and (6) with $x = \frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}$, we obtain

$$
\begin{aligned}
\left|y_{j,n} - \frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right| \le& 2\left(\frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right)^2 \mathbf{I}\left(\left|\frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right| \le \frac{1}{2}\right)\\
&+ \left|y_{j,n} - \frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right|\mathbf{I}\left(\left|\frac{\hat{p}_{j,n}q_j-p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}}\right| > \frac{1}{2}\right) \quad \text{a.s.}
\end{aligned}
\tag{7}
$$

for any $1 \leq j \leq r$. Using (7) and upper bounds (17) and (18) from Lemma 1 in Section 3 (Auxiliary Results), we obtain the following: for any $\varepsilon > 0$, $1 \leq j \leq r$ and for some constant $C > 0$,

$$
\begin{aligned}
&\mathbf{P}\left(\sqrt{n} \left| y_{j,n} - \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| > \varepsilon \right) \\
&\leq \mathbf{P}\left(2 \left(\frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right)^2 \mathbf{I}\left(\left| \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| \leq \frac{1}{2} \right) > \frac{\varepsilon}{2\sqrt{n}} \right) \\
&\quad + \mathbf{P}\left(\sqrt{n} \left| y_{j,n} - \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| \mathbf{I}\left(\left| \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| > \frac{1}{2} \right) > \frac{\varepsilon}{2} \right) \\
&\leq \mathbf{P}\left(\left(\frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right)^2 > \frac{\varepsilon}{4\sqrt{n}} \right) + \mathbf{P}\left(\left| \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| > \frac{1}{2} \right) \\
&\leq 2\mathbf{P}\left(\left| \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}} \right| > \frac{\min(\sqrt{\varepsilon}, 1)}{2 n^{\frac{1}{4}}} \right) \\
&\leq 2\mathbf{P}\left(q_j |\hat{p}_{j,n} - p_j| + p_j |\hat{q}_{j,n} - q_j| > \frac{\min(\sqrt{\varepsilon}, 1) p_j q_j}{4 n^{\frac{1}{4}}}, \hat{q}_{j,n} \geq \frac{q_j}{2} \right) + 2\mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2} \right) \\
&\leq 2\mathbf{P}\left(|\hat{p}_{j,n} - p_j| > \frac{\min(\sqrt{\varepsilon}, 1) p_j}{8 n^{\frac{1}{4}}} \right) + 2\mathbf{P}\left(|\hat{q}_{j,n} - q_j| > \frac{\min(\sqrt{\varepsilon}, 1) q_j}{8 n^{\frac{1}{4}}} \right) \\
&\quad + 2\mathbf{P}\left(|q_j - \hat{q}_{j,n}| > \frac{q_j}{2} \right) \leq 4 \exp\{-\sqrt{n} C\} + 2 \exp\{-n C\}.
\end{aligned}
\tag{8}
$$

Denote

$$
z_{j,n} := \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j \hat{q}_{j,n}}, \quad v_{j,n} := \frac{\hat{q}_{j,n} - q_j}{q_j}.
$$

For any $1 \leq j \leq r$, we have

$$
\begin{aligned}
\left| z_{j,n} - \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j q_j} \right| &= \frac{1}{p_j q_j} \left| \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{1 + \frac{\hat{q}_{j,n} - q_j}{q_j}} - (\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}) \right| \\
&= \frac{1}{p_j q_j} \left| (\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}) \left(-v_{j,n} + \sum_{k=2}^{\infty} (-v_{j,n})^k \right) \right| \mathbf{I}\left(|v_{j,n}| \leq \frac{1}{2} \right) \\
&\quad + \left| z_{j,n} - \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j q_j} \right| \mathbf{I}\left(|v_{j,n}| > \frac{1}{2} \right) \\
&\leq \frac{2}{p_j} |\hat{p}_{j,n} - p_j| \cdot |v_{j,n}| \mathbf{I}\left(|v_{j,n}| \leq \frac{1}{2} \right) + \frac{2}{q_j} |\hat{q}_{j,n} - q_j| \cdot |v_{j,n}| \mathbf{I}\left(|v_{j,n}| \leq \frac{1}{2} \right) \\
&\quad + \left| z_{j,n} - \frac{\hat{p}_{j,n} q_j - p_j \hat{q}_{j,n}}{p_j q_j} \right| \mathbf{I}\left(|v_{j,n}| > \frac{1}{2} \right) \quad \text{a.s.}
\end{aligned}
\tag{9}
$$

Using the last bound (9) together with upper bounds (17) and (18) from Lemma 1, we obtain the following: for any $\varepsilon > 0$, $1 \leq j \leq r$ and for some constant $C > 0$,

$$
\begin{aligned}
&\mathbf{P}\left(\sqrt{n}\left|z_{j,n} - \frac{\hat{p}_{j,n}q_j - p_j\hat{q}_{j,n}}{p_jq_j}\right| > \varepsilon\right) \\
&\leq \mathbf{P}\left(\sqrt{n}\left|z_{j,n} - \frac{\hat{p}_{j,n}q_j - p_j\hat{q}_{j,n}}{p_jq_j}\right|\mathbf{I}\left(|v_{j,n}| > \frac{1}{2}\right) > \frac{\varepsilon}{3}\right) \\
&\quad + \mathbf{P}\left(|\hat{p}_{j,n} - p_j| \cdot |v_{j,n}|\mathbf{I}\left(|v_{j,n}| \leq \frac{1}{2}\right) > \frac{\varepsilon p_j}{6\sqrt{n}}\right) \\
&\quad + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| \cdot |v_{j,n}|\mathbf{I}\left(|v_{j,n}| \leq \frac{1}{2}\right) > \frac{\varepsilon q_j}{6\sqrt{n}}\right) \\
&\leq \mathbf{P}\left(|\hat{p}_{j,n} - p_j| > \frac{\sqrt{\varepsilon p_j}}{\sqrt{6}n^{\frac{1}{4}}}\right) + \mathbf{P}\left(|v_{j,n}| > \frac{\sqrt{\varepsilon p_j}}{\sqrt{6}n^{\frac{1}{4}}}\right) \\
&\quad + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| > \frac{\sqrt{\varepsilon q_j}}{\sqrt{6}n^{\frac{1}{4}}}\right) + \mathbf{P}\left(|v_{j,n}| > \frac{\sqrt{\varepsilon q_j}}{\sqrt{6}n^{\frac{1}{4}}}\right) + \mathbf{P}\left(|v_{j,n}| > \frac{1}{2}\right) \\
&\leq 4\exp\{-C\sqrt{n}\} + \exp\{-Cn\}.
\end{aligned}
\tag{10}
$$

Once more, the inequalities (17) and (18) from Lemma 1 and Lemma 2 (see Section 3) give us the following: for any $\varepsilon > 0$, $1 \leq j \leq r$ and some constant $C > 0$,

$$
\begin{aligned}
&\mathbf{P}\left(\sqrt{n}\left|((\hat{p}_{j,n} - p_j) - (\hat{q}_{j,n} - q_j))\ln\frac{\hat{p}_{j,n}}{\hat{q}_{j,n}} - ((\hat{p}_{j,n} - p_j) - (\hat{q}_{j,n} - q_j))\ln\frac{p_j}{q_j}\right| \geq \varepsilon\right) \\
&\leq \mathbf{P}\left(|\hat{p}_{j,n} - p_j| \cdot \left|\ln\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right| \geq \frac{\varepsilon}{2\sqrt{n}}\right) + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| \cdot \left|\ln\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right| \geq \frac{\varepsilon}{2\sqrt{n}}\right) \\
&\leq \mathbf{P}\left(|\hat{p}_{j,n} - p_j| \geq \frac{\sqrt{\varepsilon}}{\sqrt{2}n^{\frac{1}{4}}}\right) + 2\mathbf{P}\left(\left|\ln\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right| \geq \frac{\sqrt{\varepsilon}}{\sqrt{2}n^{\frac{1}{4}}}\right) + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| \geq \frac{\sqrt{\varepsilon}}{\sqrt{2}n^{\frac{1}{4}}}\right) \\
&\leq 4\exp\{-C\sqrt{n}\}.
\end{aligned}
\tag{11}
$$

Denote

$$
\xi_n := \sqrt{n}\sum_{j=1}^{r}\left((\hat{p}_{j,n} - p_j)\left(1 + \ln\frac{p_j}{q_j} - \frac{q_j}{p_j}\right) + (\hat{q}_{j,n} - q_j)\left(1 + \ln\frac{q_j}{p_j} - \frac{p_j}{q_j}\right)\right).
$$

Relations (4), (8), (10), and (11) imply that for any $\varepsilon > 0$ there exist constants $C_1 > 0$ and $C > 0$ such that

$$
\mathbf{P}(|\sqrt{n}\eta_n - \xi_n| > \varepsilon) \leq C_1\exp\{-C\sqrt{n}\}.
$$

Thus, thanks to Slutsky's theorem, the random sequences η_n and ξ_n have the same limit by distribution. Finally, Lemma 3 with

$$
b_j = 1 + \ln\frac{p_j}{q_j} - \frac{q_j}{p_j}, \quad c_j = 1 + \ln\frac{q_j}{p_j} - \frac{p_j}{q_j},
$$

concludes the proof. □

2.1. Example

We show how to cluster catastrophic processes (see, [45,46]), using Jeffreys Divergence for their characteristics. Suppose that we have two types of insurance claims. Let $Y \in \{0, 1\}$ be a type of claim with $\mathbf{P}(Y = 1) = p$ and $\mathbf{P}(Y = 0) = q$, and let X be the size of damage, or payment according to the claim. The conditional distribution of X is $\mathbf{P}(X = a_j|Y = 1) = p_j$,

$\mathbf{P}(X = a_j | Y = 0) = q_j$, where a_j, $1 \le j \le r$. We assume that there is no difference between the conditional distribution of X if

$$\mathcal{D}_{KL}^{sym}(\mathbf{p}||\mathbf{q}) = \sum_{j=1}^{r} (p_j - q_j) \ln \left(\frac{p_j}{q_j} \right) < \alpha, \tag{12}$$

where α is a small critical value that distinguishes the types of claims. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be the sample of size n, and let $\mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n||\hat{\mathbf{q}}_n)$ and $\hat{\sigma}_n$ be the respective estimations. For sufficiently large n, we can estimate the probability (see Corollary 1)

$$\mathbf{P}\left(\mathcal{D}_{KL}^{sym} < \alpha \right) = \mathbf{P}\left(\eta_n > \mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n||\hat{\mathbf{q}}_n) - \alpha \right) = \mathbf{P}\left(\eta_n > \mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n||\hat{\mathbf{q}}_n) - \alpha \right)$$

$$= \mathbf{P}\left(\sqrt{n} \frac{\eta_n}{\hat{\sigma}_n} > \sqrt{n} \frac{\mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n||\hat{\mathbf{q}}_n) - \alpha}{\hat{\sigma}_n} \right) \approx \frac{1}{\sqrt{2\pi}} \int_{b_n}^{\infty} e^{-\frac{t^2}{2}} dt,$$

where

$$b_n = \sqrt{n} \frac{\mathcal{D}_{KL}^{sym}(\hat{\mathbf{p}}_n||\hat{\mathbf{q}}_n) - \alpha}{\hat{\sigma}_n}.$$

These relations propose a calculation of the probability that inequality (12) is true.

3. Auxiliary Results

The following lemma contains the proofs of inequalities (17) and (18) used in the proof of Theorem 2. The other inequalities of the lemma, (13)–(16), are used in the proofs in this section.

Lemma 1. *For any $g > 0$, the following inequalities hold*

$$\mathbf{P}(|\hat{p}_n - p)| > g) \le 2 \exp \left\{ -\frac{ng^2}{2(\max(p,q))^2} \right\}, \tag{13}$$

$$\mathbf{P}(|\hat{q}_n - q)| > g) \le 2 \exp \left\{ -\frac{ng^2}{2(\max(p,q))^2} \right\}, \tag{14}$$

$$\max_{1 \le j \le r} \mathbf{P}\left(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(X_i = a_j, Y_i = 1) - pp_j) > g \right) \le 2 \exp \left\{ -\frac{ng^2}{2(\max(pp_{\max}, 1 - pp_{\min}))^2} \right\}, \tag{15}$$

$$\max_{1 \le j \le r} \mathbf{P}\left(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(X_i = a_j, Y_i = 0) - qq_j) > g \right) \le 2 \exp \left\{ -\frac{ng^2}{2(\max(qq_{\max}, 1 - qq_{\min}))^2} \right\}, \tag{16}$$

$$\max_{1 \le j \le r} \mathbf{P}\left(|\hat{p}_{j,n} - p_j| > g \right)$$

$$\le 2 \exp \left\{ -\frac{ng^2 p^2}{2^7 (\max(p,q))^2 p_{\max}^2} \right\} + 2 \exp \left\{ -\frac{ng^2 p^2}{8(\max(pp_{\max}, 1 - pp_{\min}))^2} \right\} \tag{17}$$

$$+ \exp \left\{ -\frac{np^2 p_{\min}^2}{2(\max(pp_{\max}, 1 - pp_{\min}))^2} \right\} + \exp \left\{ -\frac{np^2}{8(\max(p,q))^2} \right\},$$

$$\max_{1 \le j \le r} \mathbf{P}\left(\left|\hat{q}_{j,n} - q_j\right| > g\right)$$

$$\le 2 \exp\left\{-\frac{ng^2 q^2}{2^7 (\max(p,q))^2 q_{\max}^2}\right\} + 2 \exp\left\{-\frac{ng^2 q^2}{8 (\max(qq_{\max}, 1 - qq_{\min}))^2}\right\} \quad (18)$$

$$+ \exp\left\{-\frac{nq^2 q_{\min}^2}{2 (\max(qq_{\max}, 1 - qq_{\min}))^2}\right\} + \exp\left\{-\frac{nq^2}{8 (\max(p,q))^2}\right\},$$

where $p_{\min} := \min\limits_{1 \le j \le r} p_j$, $p_{\max} := \max\limits_{1 \le j \le r} p_j$, $q_{\min} := \min\limits_{1 \le j \le r} p_j$, $q_{\max} := \max\limits_{1 \le j \le r} p_j$.

Proof. The Hoeffding inequality proves (13). Indeed,

$$\mathbf{P}(|\hat{p}_n - p)| > g) = \mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{I}(Y_i = 1) - p)\right| > g\right)$$

$$\le \mathbf{P}\left(\sum_{i=1}^n (\mathbf{I}(Y_i = 1) - p) > ng\right) + \mathbf{P}\left(-\sum_{i=1}^n (\mathbf{I}(Y_i = 1) - p) > ng\right)$$

$$\le 2 \exp\left\{-\frac{ng^2}{2 (\max(p,q))^2}\right\}.$$

Inequalities (14)–(16) are obtained in the same way, and therefore we omit their proofs. Let us prove (17). We have

$$\mathbf{P}\left(\left|\hat{p}_{j,n} - p_j\right| > g\right) = \mathbf{P}\left(\left|\frac{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)} \pm \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{p} - p_j\right| > g\right)$$

$$\le \mathbf{P}\left(\left|\frac{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)} - \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{p}\right| > \frac{g}{2}\right) \quad (19)$$

$$+ \mathbf{P}\left(\left|\frac{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{p} - p_j\right| > \frac{g}{2}\right) =: \mathbf{P}_1 + \mathbf{P}_2.$$

Denote

$$A := \left\{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1) > 2pp_j\right\}, \quad B := \left\{\frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1) < \frac{p}{2}\right\}.$$

Let us find the upper bound for the probability $\mathbf{P}_1$. We have

$$\mathbf{P}_1 = \mathbf{P}\left(\left|\frac{\left(p - \frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)\right)\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}{p\frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)}\right| > \frac{g}{2}\right)$$

$$= \mathbf{P}\left(\left|p - \frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)\right| > \frac{gp\frac{1}{n}\sum_{i=1}^n \mathbf{I}(Y_i = 1)}{2\frac{1}{n}\sum_{i=1}^n \mathbf{I}(X_i = a_j, Y_i = 1)}, \overline{A}, \overline{B}\right) + \mathbf{P}(A) + \mathbf{P}(B) \quad (20)$$

$$\le \mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{I}(Y_i = 1) - p)\right| > \frac{gp}{8p_j}\right) + \mathbf{P}(A) + \mathbf{P}(B).$$

Next, for any $1 \leq j \leq r$, inequality (13) provides the following upper bound for the first term from the righthand side of the last inequality above (20),

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 1) - p)\right| > \frac{gp}{8p_j}\right) \leq 2\exp\left\{-\frac{ng^2p^2}{2^7(\max(p,q))^2\left(\max\limits_{1\leq j\leq r} p_j\right)^2}\right\}. \quad (21)$$

The upper bound for the probabilities $\mathbf{P}(A)$, $\mathbf{P}(B)$, and $\mathbf{P}_2$ are obtained by the Hoeffding inequality:

$$\mathbf{P}(A) = \mathbf{P}\left(\sum_{i=1}^{n}(\mathbf{I}(X_i = a_j, Y_i = 1) - pp_j) > npp_j\right)$$

$$\leq \exp\left\{-\frac{np^2\left(\min\limits_{1\leq j\leq r} p_j\right)^2}{2\left(\max\limits_{1\leq j\leq r}(\max(pp_j, 1 - pp_j))\right)^2}\right\},$$

$$\mathbf{P}(B) = \mathbf{P}\left(\sum_{i=1}^{n}(p - \mathbf{I}(Y_i = 1)) > \frac{pn}{2}\right) \leq \exp\left\{-\frac{np^2}{8(\max(p,q))^2}\right\}, \quad (22)$$

$$\mathbf{P}_2 \leq 2\exp\left\{-\frac{ng^2p^2}{8\left(\max\limits_{1\leq j\leq r}(\max(pp_j, 1 - pp_j))\right)^2}\right\}.$$

Finally, inequalities (19)–(22) yield relation (17).
The proof of inequality (18) has a similar structure as the proof of (17). $\square$

The next lemma contains a huge, multilevel upper bound, which was used in the proof of Theorem 2.

Lemma 2. *For any $g > 0$, the following inequality holds*

$$\max_{1\leq j\leq r}\mathbf{P}\left(\left|\ln\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right)\right| > g\right)$$

$$\leq 4\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^{11}(\max(p,q))^2p_{\max}^2}\right\} + 4\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^7(\max(pp_{\max}, 1 - pp_{\min}))^2}\right\}$$

$$+ 2\exp\left\{-\frac{np^2p_{\min}^2}{2^9(\max(p,q))^2p_{\max}^2}\right\} + 2\exp\left\{-\frac{np^2p_{\min}^2}{2^5(\max(pp_{\max}, 1 - pp_{\min}))^2}\right\}$$

$$+ 3\exp\left\{-\frac{np^2p_{\min}^2}{2(\max(pp_{\max}, 1 - pp_{\min}))^2}\right\} + 3\exp\left\{-\frac{np^2}{8(\max(p,q))^2}\right\} \quad (23)$$

$$+ 4\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^{11}(\max(p,q))^2q_{\max}^2}\right\} + 4\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^7(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\}$$

$$+ 2\exp\left\{-\frac{nq^2q_{\min}^2}{2^9(\max(p,q))^2q_{\max}^2}\right\} + 2\exp\left\{-\frac{nq^2q_{\min}^2}{2^5(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\}$$

$$+ 3\exp\left\{-\frac{nq^2q_{\min}^2}{2(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\} + 3\exp\left\{-\frac{nq^2}{8(\max(p,q))^2}\right\}.$$

Proof. For any $g > 0$ and $1 \leq j \leq r$, we obtain

$$\mathbf{P}\left(\left|\ln\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right)\right| > g\right) = \mathbf{P}\left(\ln\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right) > g\right) + \mathbf{P}\left(-\ln\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}}\right) > g\right) =: \mathbf{P}_1 + \mathbf{P}_2. \tag{24}$$

Let us bound $\mathbf{P}_1$ from above. Using inequality $e^g - 1 \geq g$ and using Lemma 1, for any $1 \leq j \leq r$, we obtain

$$\mathbf{P}_1 = \mathbf{P}\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}} > e^g\right) = \mathbf{P}\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}} - 1 > e^g - 1\right) \leq \mathbf{P}\left(\frac{\hat{p}_{j,n}q_j}{p_j\hat{q}_{j,n}} - 1 > g\right)$$

$$\leq \mathbf{P}\left(\frac{\hat{p}_{j,n}q_j - p_j\hat{q}_{j,n}}{p_j\hat{q}_{j,n}} > g, \hat{q}_{j,n} \geq \frac{q_j}{2}\right) + \mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2}\right)$$

$$\leq \mathbf{P}\left(\hat{p}_{j,n}q_j - p_j\hat{q}_{j,n} > \frac{g_1 p_j q_j}{2}\right) + \mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2}\right)$$

$$= \mathbf{P}\left(q_j(\hat{p}_{j,n} - p_j) - p_j(\hat{q}_{j,n} - q_j) > \frac{g p_j q_j}{2}\right) + \mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2}\right)$$

$$\leq \mathbf{P}\left(q_j|\hat{p}_{j,n} - p_j| + p_j|\hat{q}_{j,n} - q_j| > \frac{g p_j q_j}{2}\right) + \mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2}\right)$$

$$\leq \mathbf{P}\left(|\hat{p}_{j,n} - p_j| > \frac{g p_j}{4}\right) + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| > \frac{g q_j}{4}\right) + \mathbf{P}\left(\hat{q}_{j,n} < \frac{q_j}{2}\right)$$

$$= \mathbf{P}\left(|\hat{p}_{j,n} - p_j| > \frac{g p_j}{4}\right) + \mathbf{P}\left(|\hat{q}_{j,n} - q_j| > \frac{g q_j}{4}\right) + \mathbf{P}\left(q_j - \hat{q}_{j,n} > \frac{q_j}{2}\right)$$

Next, we apply the upper bounds (17) and (18) for the probabilities in the last line from the inequality above. Thus,

$$\mathbf{P}_1 \leq 2\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^{11}(\max(p,q))^2p_{\max}^2}\right\} + 2\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^7(\max(pp_{\max}, 1 - pp_{\min}))^2}\right\}$$

$$+ \exp\left\{-\frac{np^2p_{\min}^2}{2(\max(pp_{\max}, 1 - pp_{\min}))^2}\right\} + \exp\left\{-\frac{np^2}{8(\max(p,q))^2}\right\}$$

$$+ 2\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^{11}(\max(p,q))^2q_{\max}^2}\right\} + 2\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^7(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\} \tag{25}$$

$$+ 2\exp\left\{-\frac{nq^2q_{\min}^2}{2^9(\max(p,q))^2q_{\max}^2}\right\} + 2\exp\left\{-\frac{nq^2q_{\min}^2}{2^5(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\}$$

$$+ 2\exp\left\{-\frac{nq^2q_{\min}^2}{2(\max(qq_{\max}, 1 - qq_{\min}))^2}\right\} + 2\exp\left\{-\frac{nq^2}{8(\max(p,q))^2}\right\}.$$

We obtain the upper bound for $\mathbf{P}_2$ in the same way:

$$
\begin{aligned}
\mathbf{P}_2 \leq{} & 2\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^{11}(\max(p,q))^2p_{\max}^2}\right\} + 2\exp\left\{-\frac{ng^2p^2p_{\min}^2}{2^7(\max(pp_{\max},1-pp_{\min}))^2}\right\} \\
& + 2\exp\left\{-\frac{np^2p_{\min}^2}{2^9(\max(p,q))^2p_{\max}^2}\right\} + 2\exp\left\{-\frac{np^2p_{\min}^2}{2^5(\max(pp_{\max},1-pp_{\min}))^2}\right\} \\
& + 2\exp\left\{-\frac{np^2p_{\min}^2}{2(\max(pp_{\max},1-pp_{\min}))^2}\right\} + 2\exp\left\{-\frac{np^2}{8(\max(p,q))^2}\right\} \\
& + 2\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^{11}(\max(p,q))^2q_{\max}^2}\right\} + 2\exp\left\{-\frac{ng^2q^2q_{\min}^2}{2^7(\max(qq_{\max},1-qq_{\min}))^2}\right\} \\
& + \exp\left\{-\frac{nq^2q_{\min}^2}{2(\max(qq_{\max},1-qq_{\min}))^2}\right\} + \exp\left\{-\frac{nq^2}{8(\max(p,q))^2}\right\}.
\end{aligned}
\tag{26}
$$

Inequalities (24)–(26) imply (23). □

Lemma 3. *For any given constants b_j, c_j, $1 \leq j \leq r$, such that $\sum_{j=1}^{r}(|b_j| + |c_j|) > 0$, the following convergence takes place*

$$
\sqrt{n}\sum_{j=1}^{r}\left((\hat{p}_{j,n} - p_j)b_j + (\hat{q}_{j,n} - q_j)c_j\right) \xrightarrow{d} \xi \sim N(0,\sigma^2), \text{ as } n \to \infty,
$$

with

$$
\begin{aligned}
\sigma^2 := \mathbf{E}\Bigg(\sum_{j=1}^{r}\Bigg(& \left(\frac{1}{p}(\mathbf{I}(X=a_j,Y=1) - pp_j) - p_j(\mathbf{I}(Y=1) - p)\right)b_j \\
& + \left(\frac{1}{q}(\mathbf{I}(X=a_j,Y=0) - qq_j) - q_j(\mathbf{I}(Y=0) - q)\right)c_j\Bigg)\Bigg)^2.
\end{aligned}
$$

Proof. Denote

$$
y_{j,n} := \frac{1}{\sqrt{n}p}\sum_{i=1}^{n}(\mathbf{I}(X_i=a_j,Y_i=1) - pp_j) - \frac{p_j}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{I}(Y_i=1) - p),
$$

$$
z_n := \frac{1}{np}\sum_{i=1}^{n}(\mathbf{I}(Y_i=1) - p), \quad v_{j,n} := \frac{1}{\sqrt{n}p}\sum_{i=1}^{n}(\mathbf{I}(X_i=a_j,Y_i=1) - pp_j).
$$

We have

$$
|\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| = \left| \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)}{p\left(1 + \frac{1}{np} \sum_{i=1}^{n} (\mathbf{I}(Y_i = 1) - p)\right)} - \sqrt{n}p_j - y_{j,n} \right|
$$

$$
= \left| \frac{1}{\sqrt{n}p} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)\left(1 - z_n + \sum_{k=2}^{\infty} (-z_n)^k\right) - \sqrt{n}p_j - y_{j,n}\right| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right)
$$

$$
+ |\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| \mathbf{I}\left(|z_n| > \frac{1}{2}\right)
$$

$$
= \left| (v_{j,n} + \sqrt{n}p_j)\left(1 - z_n + \sum_{k=2}^{\infty} (-z_n)^k\right) - \sqrt{n}p_j - y_{j,n}\right| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right)
$$

$$
+ |\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| \mathbf{I}\left(|z_n| > \frac{1}{2}\right) \tag{27}
$$

$$
= \left| -v_{j,n}z_n + \frac{1}{\sqrt{n}p} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1) \sum_{k=2}^{\infty} (-z_n)^k \right| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right)
$$

$$
+ |\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| \mathbf{I}\left(|z_n| > \frac{1}{2}\right)
$$

$$
\leq |v_{j,n}z_n| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right) + \left| \frac{2}{\sqrt{n}p} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)z_n^2\right| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right)
$$

$$
+ |\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| \mathbf{I}\left(|z_n| > \frac{1}{2}\right) \quad \text{a.s.}
$$

Utilizing (27), for any $1 \leq j \leq r$, $\varepsilon > 0$, we obtain

$$
\mathbf{P}(|\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| > \varepsilon)
$$

$$
\leq \mathbf{P}\left(|v_{j,n}z_n| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right) > \frac{\varepsilon}{3}\right)
$$

$$
+ \mathbf{P}\left(\left| \frac{2}{\sqrt{n}p} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)z_n^2\right| \mathbf{I}\left(|z_n| \leq \frac{1}{2}\right) > \frac{\varepsilon}{3}\right)
$$

$$
+ \mathbf{P}\left(|\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| \mathbf{I}\left(|z_n| > \frac{1}{2}\right) > \frac{\varepsilon}{3}\right) \tag{28}
$$

$$
\leq \mathbf{P}\left(|v_{j,n}z_n| > \frac{\varepsilon}{3}\right) + \mathbf{P}\left(\left| \frac{2}{\sqrt{n}p} \sum_{i=1}^{n} \mathbf{I}(X_i = a_j, Y_i = 1)z_n^2\right| > \frac{\varepsilon}{3}\right) + \mathbf{P}\left(|z_n| > \frac{1}{2}\right)
$$

$$
=: \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3.
$$

We obtain the upper bound for $\mathbf{P}_1$ using inequalities from Lemma 1. Indeed, inequalities (13) and (15)) imply that for any $\varepsilon > 0$, there exists $C > 0$ such that

$$
\mathbf{P}_1 = \mathbf{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(X_i = a_j, Y_i = 1) - pp_j)\right| \cdot \left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(Y_i = 1) - p)\right| > \frac{p^2\varepsilon}{3\sqrt{n}}\right)
$$

$$
\leq \mathbf{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(X_i = a_j, Y_i = 1) - pp_j)\right| > \frac{p\sqrt{\varepsilon}}{\sqrt{3}n^{\frac{1}{4}}}\right) \tag{29}
$$

$$
+ \mathbf{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{I}(Y_i = 1) - p)\right| > \frac{p\sqrt{\varepsilon}}{\sqrt{3}n^{\frac{1}{4}}}\right) \leq 2\exp\{-C\sqrt{n}\}.
$$

We obtain the upper bound for the probability $\mathbf{P}_2$ in the same way. Indeed, inequalities (13) and (15) imply that for any $\varepsilon > 0$, there exists $C > 0$ such that

$$
\begin{aligned}
\mathbf{P}_2 =&\mathbf{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(X_i = a_j, Y_i = 1)\right)\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 1) - p)\right|^2 > \frac{p^3\varepsilon}{6\sqrt{n}}\right)\\
\leq&\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 1) - p)\right|^2 > \frac{p^2\varepsilon}{3\sqrt{n}p_j}, \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(X_i = a_j, Y_i = 1) \geq \frac{pp_j}{2}\right)\\
&+\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(X_i = a_j, Y_i = 1) < \frac{pp_j}{2}\right)\\
\leq&\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 1) - p)\right| > \frac{p\sqrt{\varepsilon}}{\sqrt{3p_j}n^{\frac{1}{4}}}\right) + \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(pp_j - \mathbf{I}(X_i = a_j, Y_i = 1))\right| > \frac{pp_j}{2}\right)\\
\leq&\exp\{-C\sqrt{n}\} + \exp\{-Cn\}.
\end{aligned}
\tag{30}
$$

Finally, the bound for the probability $\mathbf{P}_3$ is provided by inequality (13):

$$
\mathbf{P}_3 = \mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 1) - p)\right| > \frac{1}{2}\right) \leq \exp\{-Cn\},
\tag{31}
$$

for some $C > 0$. Relations (28)–(31) imply that for any $\varepsilon > 0$, there exists $C > 0$ such that

$$
\max_{1 \leq j \leq r} \mathbf{P}(|\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n}| > \varepsilon) \leq 4\exp\{-C\sqrt{n}\}.
\tag{32}
$$

Denote

$$
y'_{j,n} := \frac{1}{\sqrt{nq}}\sum_{i=1}^{n}(\mathbf{I}(X_i = a_j, Y_i = 0) - qq_j) - \frac{q_j}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{I}(Y_i = 0) - q).
$$

In a completely similar manner to the above, for any $\varepsilon > 0$ and some $C > 0$, we obtain

$$
\max_{1 \leq j \leq r} \mathbf{P}(|\sqrt{n}(\hat{q}_{j,n} - q_j) - y'_{j,n}| > \varepsilon) \leq 4\exp\{-C\sqrt{n}\}.
\tag{33}
$$

It is easy to see that

$$
\begin{aligned}
\sqrt{n}&\sum_{j=1}^{r}\left((\hat{p}_{j,n} - p_j)b_j + (\hat{q}_{j,n} - q_j)c_j\right)\\
&= \sum_{j=1}^{r}\left(y_{j,n}b_j + y'_{j,n}c_j\right) + \sum_{j=1}^{r}(\sqrt{n}(\hat{p}_{j,n} - p_j) - y_{j,n})b_j + \sum_{j=1}^{r}(\sqrt{n}(\hat{q}_{j,n} - q_j) - y'_{j,n})c_j\\
&=: \xi_n + \phi_n + \zeta_n.
\end{aligned}
$$

By (32) and (33), for any $\varepsilon > 0$, we obtain

$$
\lim_{n \to \infty} \mathbf{P}(|\phi_n + \zeta_n| > \varepsilon) = 0.
$$

Therefore, by Slutsky's theorem, the weak limit of the original sequence coincides with the limit sequences ξ_n. It remains to be noted that

$$
\begin{aligned}
\xi_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{r}\Bigg(&\left(\frac{1}{p}(\mathbf{I}(X_i = a_j, Y_i = 1) - pp_j) - p_j(\mathbf{I}(Y_i = 1) - p)\right)b_j\\
&+ \left(\frac{1}{q}(\mathbf{I}(X_i = a_j, Y_i = 0) - qq_j) - q_j(\mathbf{I}(Y_i = 0) - q)\right)c_j\Bigg)
\end{aligned}
$$

and to apply the central limit theorem. $\quad\square$

4. Conclusions

In this paper, we studied the symmetrized version of the KL divergence, $\mathcal{D}_{KL}^{sym}(\mathbf{p}||\mathbf{q}) := \mathcal{D}_{KL}(\mathbf{p}||\mathbf{q}) + \mathcal{D}_{KL}(\mathbf{q}||\mathbf{p})$. Today, it is known as Jeffreys divergence and is popular for classification problems. In this context, the distributions $\mathbf{p}$ and $\mathbf{q}$ represent the conditional probability distributions of a characteristic of interest under two different classes or conditions.

We established the asymptotic unbiasedness and normality of the plug-in estimator for the Jeffreys divergence. We considered a one-sample estimator, where, for a given sample size n, the number of observations in one condition is random and follows a binomial distribution. This differs from the traditional approach, where the properties of estimators are studied as the given sample sizes n and m of the two classes increase (see [38]). The results were expected, but additional technical work was required due to the randomness of the number of observations in one class. Moreover, we did not find detailed proofs for the Jeffreys divergence in the literature.

In this paper, we avoided referencing some known methods for proving normality, such as the δ-method, and provided detailed proofs instead. We believe that such proofs are accessible to undergraduate students.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request.

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

References

1. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
2. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
3. Rubenstein, P.; Bousquet, O.; Djolonga, J.; Riquelme, C.; Tolstikhin, I.O. Practical and consistent estimation of f-divergences. *Adv. Neural Inf. Process. Syst.* **2019**, 32, 4072–4082.
4. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
5. Zhang, C.; Bütepage, J.; Kjellström, H.; Mandt, S. Advances in variational inference. *Jieee Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2008–2026. [CrossRef]
6. Tzikas, D.G.; Likas, A.C.; Galatsanos, N.P. The variational approximation for Bayesian inference. *IEEE Signal Process. Mag.* **2008**, *25*, 131–146. [CrossRef]
7. Jewson, J.; Smith, J.Q.; Holmes, C. Principles of Bayesian inference using general divergence criteria, *Entropy* **2018**, 20, 442. [CrossRef]
8. Ji, S.; Zhang, Z.; Ying, S.; Wang, L.; Zhao, X.; Gao, Y. Kullback–Leibler divergence metric learning. *IEEE Trans. Cybern.* **2020**, *52*, 2047–2058. [CrossRef]
9. Noh, Y.K.; Sugiyama, M.; Liu, S.; Plessis, M.C.; Park, F.C.; Lee, D.D. Bias reduction and metric learning for nearest-neighbor estimation of Kullback-Leibler divergence. *Artif. Intell. Stat.* **2014**, *1*, 669–677. [CrossRef]
10. Suárez, J.L.; García, S.; Herrera, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing* **2021**, *425*, 300–322. [CrossRef]
11. Claici, S.; Yurochkin, M.; Ghosh, S.; Solomon, J. Model fusion with Kullback-Leibler divergence. *Int. Conf. Mach. Learn.* **2020**, *1*, 2038–2047.
12. Póczos, B.; Xiong, L.; Schneider, J. Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv* **2012**, arXiv:1202.3758.
13. Cui, S.; Luo, C. Feature-based non-parametric estimation of Kullback–Leibler divergence for SAR image change detection. *Remote. Sens. Lett.* **2016**, *7*, 1102–1111. [CrossRef]
14. Deledalle, C.A. Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Electron. J. Stat.* **2017**, *11*, 3141–3164. [CrossRef]
15. Granero-Belinchón, C.; Roux, S.G.; Garnier, N.B. Kullback-Leibler divergence measure of intermittency: Application to turbulence. *Phys. Rev.* **2018**, *97*, 013107. [CrossRef]

16. Charzyńska, A.; Gambin, A. Improvement of the k-NN entropy estimator with applications in systems biology. *Entropy* **2015**, *18*, 13. [CrossRef]
17. Belavkin, R.V. Asymmetric topologies on statistical manifolds. *Int. Conf. Geom. Sci. Inf.* **2015**, *1*, 203–210.
18. Jager, L.; Wellner, J.A. Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **2007**, *35*, 2018–2053. [CrossRef]
19. Vexler, A.; Gurevich, G. Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Comput. Stat. Data Anal.* **2010**, *54*, 531–545. [CrossRef]
20. Evren, A.; Tuna, E. On some properties of goodness of fit measures based on statistical entropy. *Int. J. Res. Rev. Appl. Sci.* **2012**, *13*, 192–205.
21. Bulinski, A.; Dimitrov, D. Statistical estimation of the Kullback–Leibler divergence. *Mathematics* **2021**, *9*, 544. [CrossRef]
22. Broniatowski, M. Estimation of the Kullback-Leibler divergence. *Math. Methods Stat.* **2003**, *12*, 391–409.
23. Seghouane, A.K.; Amari, S.I.Variants of the Kullback-Leibler divergence and their role in model selection. *Ifac Proc. Vol.* **2006**, *39*, 826–831. [CrossRef]
24. Audenaert, K.M. On the asymmetry of the relative entropy. *J. Math. Phys.* **2013**, *54*, 073506. [CrossRef]
25. Pinski, F.J.; Simpson, G.; Stuart, A.M.; Weber, H. Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *Siam J. Math. Anal.* **2015**, *27*, 4091–4122. [CrossRef]
26. Zeng, J.; Xiao, F. A fractal belief KL divergence for decision fusion. *Eng. Appl. Artif. Intell.* **2023**, *121*, 106027. [CrossRef]
27. Kamiński, M. On the Symmetry Importance in a Relative Entropy Analysis for Some Engineering Problems. *Symmetry* **2022**, *14*, 1945. [CrossRef]
28. Johnson, D.H.; Sinanovic, S. Symmetrizing the kullback-leibler distance. *IEEE Trans. Inf. Theory* **2001**, *1*, 1–10.
29. Jeffreys, H. *The Theory of Probability*; Oxford Classic Texts in the Physical Sciences: Oxford, UK, 1998.
30. Chen, J.; Matzinger, H.; Zhai, H.; Zhou, M. Centroid estimation based on symmetric kl divergence for multinomial text classification problem. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 1174–1177.
31. Andriamanalimanana, B.; Tekeoglu, A.; Bekiroglu, K.; Sengupta, S.; Chiang, C.F.; Reale, M.; Novillo, J. Symmetric kullback-leibler divergence of softmaxed distributions for anomaly scores. In Proceedings of the 2019 IEEE Conference on Communications and Network Security (CNS), Washington, DC, USA , 10–12 June 2019; Volume 1, pp. 1–6.
32. Domke, J. An easy to interpret diagnostic for approximate inference: Symmetric divergence over simulations. *arXiv* **2021**, arXiv:2103.01030.
33. Nguyen, B.; Morell, C.; De Baets, B. Supervised distance metric learning through maximization of the Jeffrey divergence. *Pattern Recognit.* **2017**, *64*, 215–225. [CrossRef]
34. Moreno, P.; Ho, P.; Vasconcelos, N.A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *Adv. Neural Inf. Process. Syst.* **2003**, *6*.
35. Yao, Z.; Lai, Z.; Liu, W. A symmetric KL divergence based spatiogram similarity measure. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; Volume 1, 193–196.
36. Said, A.B.; Hadjidj, R.; Foufou, S. Cluster validity index based on Jeffrey divergence. *Pattern Anal Appl.* **2017**, *20*, 21–31. [CrossRef]
37. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [CrossRef]
38. Zhang, Z.; Grabchak, M. Nonparametric estimation of Küllback-Leibler divergence. *Neural Comput.* **2014**, *26*, 2570–2593. [CrossRef]
39. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885. [CrossRef]
40. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2017**, *63*, 6774–6798. [CrossRef]
41. Bulinski, A.; Dimitrov, D. Divergence Measures Estimation and Its Asymptotic Normality Theory in the discrete case. *Eur. J. Pure Appl. Math.* **2019**, *12*, 790–820.
42. Yao, L.Q.; Liu, S.H. Symmetric KL-divergence by Stein's Method. *arXiv* **2024**, arXiv:2401.11381.
43. Bobkov, S.G.; Chistyakov, G.P.; Götze, F. Rényi divergence and the central limit theorem. *Ann. Probab.* **2019**, *47*, 270–323. [CrossRef]
44. Abramowitz, M.; Stegun, I.A. (Eds). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed; Dover: New York, NY, USA, 1972.
45. Rojas, H.; Logachov, A.; Yambartsev, A. Order Book Dynamics with Liquidity Fluctuations: Asymptotic Analysis of Highly Competitive Regime. *Mathematics* **2023**, *11*, 4235. [CrossRef]
46. Logachov, A.; Logachova, O.; Yambartsev, A. Processes with catastrophes: Large deviation point of view *Stoch. Process. Their Appl.* **2024**, *176*, 104447. [CrossRef]