Asymptotic and finite-sample distributions of one- and two-sample empirical relative entropy, with application to change-point detection

Matthieu Garcin[a],[*] Louis Perot[b]

December 18, 2025

Abstract

Relative entropy, as a divergence metric between two distributions, can be used for offline change-point detection and extends classical methods that mainly rely on moment-based discrepancies. To build a statistical test suitable for this context, we study the distribution of empirical relative entropy and derive several types of approximations: concentration inequalities for finite samples, asymptotic distributions, and Berry-Esseen bounds in a pre-asymptotic regime. For the latter, we introduce a new approach to obtain Berry-Esseen inequalities for nonlinear functions of sum statistics under some convexity assumptions. Our theoretical contributions cover both one- and two-sample empirical relative entropies. We then detail a change-point detection procedure built on relative entropy and compare it, through extensive simulations, with classical methods based on moments or on information criteria. Finally, we illustrate its practical relevance on two real datasets involving temperature series and volatility of stock indices.

Keywords − Berry-Esseen bounds, concentration inequalities, information theory, Kullback-Leibler divergence, structural break detection, two-sample divergence testing

1   Introduction

Detecting abrupt changes in time series is crucial in many fields, from climatology [61, 27] to finance [46, 5, 11]. It enables one to assess the validity of a model over a given time interval and to specify models that appropriately describe or predict time series.

One of the most widespread online methods for change-point detection is the cumulative sum (CUSUM) procedure [56]. It tracks down significant deviations from the mean and is thus mainly based on moments. In this paper, we focus on offline methods. Beyond mean shifts, structural breaks may occur within parametric models when some parameters suddenly move to new values. Offline approaches in this setting may be based on moments or on information criteria [67, 29].

The latter case consists in finding the time partition that minimizes the information criterion for a model built as a succession of parameter regimes, each associated with a segment of the partition.

In an offline non-parametric setting, one aims at detecting variations in a non-parametric probability distribution [43, 33]. A distribution may indeed change without affecting the mean or the variance, which limits the relevance of moment-based approaches such as CUSUM [26, 69]. In this distribution-based framework, change-point detection amounts to assessing the statistical significance of the deviation between two empirical distributions. This is precisely the practical objective of the present paper.

Among the existing divergence metrics that could be used in this context, one can cite Wasserstein distance, Hellinger distance, Kolmogorov-Smirnov statistic, or relative entropy [36, 33, 49]. We will focus on the latter metric, which is the expectation of the log-likelihood ratio. This ratio is the statistic leading to the uniformly highest power among the statistical tests of probability divergence, under the assumptions of Neyman-Pearson lemma [24].

The use of relative entropy in the context of change-point detection was already mentioned sporadically in literature [52, 42]. But the challenge of knowing the exact distribution of empirical relative entropy often leads to the construction of statistical tests with a threshold based on simulated quantiles [59]. The theoretical objective of the present paper is to introduce approximations of this distribution. We focus on three types of them. The most natural is the asymptotic distribution, which may however not always be relevant in the context of change-point detection, where we may be interested in small samples, for example to rapidly draw an alert after a break. We obtain as well pre-asymptotic and finite-sample bounds of the distribution, based either on a Berry-Esseen approach or on concentration inequalities. One can then use these bounds, instead of the asymptotic approximation, to build conservative statistical tests. Our Berry-Esseen bounds are obtained for a nonlinear function of a sum statistic, whose limit distribution is non-Gaussian. Our inequality controls two effects: the classical speed of convergence for an approximation of our statistic and the error related to this approximation. This method can easily be replicated to other kinds of nonlinear statistics. We propose as well extensions to two samples, that is approximations of the distribution of the relative entropy between two empirical probabilities, which is particularly useful in the context of change-point detection. This question is challenging because relative entropy does not satisfy a triangle inequality and because its empirical version may become unbounded when the reference probability is estimated from few observations.

In a simulation study, we show the benefit of using change-point detection methods based on relative entropy, compared to methods based on moments or on information criteria. Two applications to a climate dataset and to financial time series highlight the practical relevance of the method. We study a daily time series of temperatures at Embrun, France, during more than 25 years, as well as six daily volatility series of stock indices, during about 20 years.

The paper is organized as follows. Section 2 introduces theoretical results about the distribution of empirical relative entropy. In Section 3, we present the change-point detection method based on relative entropy along with baseline approaches. Section 4 contains a simulation study and Section 5 the application to temperature and volatility series. Section 6 concludes.

2 Distribution of empirical relative entropy

We want to compare to each other two discrete probability distributions, with a finite number of possible states, in order to build a statistical test of equality of the distributions. This can be done either by testing one after the other the equality of probabilities for each possible state of the random variable, or by aggregating all these probabilities in a single statistic, thus leading to

a single test. It is the purpose of relative entropy.

After a brief introduction on the concept, we detail the asymptotic distribution of the empirical version of relative entropy, with either one or two samples. We also propose pre-asymptotic and finite-sample bounds of its distribution, deriving both a Berry-Esseen inequality and various concentration inequalities.

2.1 Relative entropy

We consider a discrete probability, with a finite number $k \geq 2$ of possible categories: $p = (p_1, ..., p_k)^t \in (0, 1)^k$, Z^t standing for the transposed vector of Z. The Shannon entropy related to this categorical distribution is

$$H(p) = -\sum_{i=1}^{k} p_i \log(p_i),$$

where we use the convention $0 \log(0) = 0$ [21]. The entropy quantifies the uncertainty of the distribution [21, 31]. The minimum uncertainty corresponds to a concentration in a single state, leading to the minimum entropy, $H(p) = 0$. The maximum entropy is reached by a uniform distribution, for which we get $H(p) = \log(k)$.

When working with data, we can calculate an empirical entropy, based on empirical probabilities. We observe $X_1, ..., X_n$, iid random variables, which may be either discrete or continuous. We discretize these variables by defining k possible states $\Omega_1, ..., \Omega_k$, which may for example be intervals. We have $\mathbb{P}(X_j \in \Omega_i) = p_i$ for all $j \in [\![1, n]\!]$ and $i \in [\![1, k]\!]$. We also define the empirical probability $\widehat{p}_n = (\widehat{p}_{n,1}, ..., \widehat{p}_{n,k})^t \in [0, 1]^k$, such that

$$\widehat{p}_{n,i} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{X_j \in \Omega_i}. \tag{1}$$

The quantity $H(\widehat{p}_n)$ is then the empirical entropy. The asymptotic distribution of the empirical entropy is either a chi-squared or a Gaussian distribution, depending on the nature of p [7, 78].

One can also replace the probability p by a conditional probability. It leads to the evaluation of the complexity of the dependence structure between two variables, what has been shown to be useful for time series, the presence of serial dependence being a useful asset for forecasting purposes [16]. It has been shown that the distribution of conditional Shannon entropy and of the close concept of mutual information is similar to the one of the non-conditional entropy [53, 66, 15, 55].

Relative entropy, sometimes also called Kullback-Leibler divergence, uses the concept of entropy to compare to each other two probability distributions, $p, q \in (0, 1)^k$:

$$D_{\mathrm{KL}}(p\|q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}.$$

Relative entropy is non-negative and not bounded. But, with a finite number of states, the infinity of the relative entropy is equivalent to the existence of a state i for which $p_i \neq 0$ and $q_i = 0$. Relative entropy is not symmetric in p and q. When q is uniform, relative entropy more simply writes $D_{\mathrm{KL}}(p\|q) = \log(k) - H(p)$.

Again, replacing p by its empirical counterpart $\widehat{p}_n$ leads to an empirical relative entropy. But one can also use relative entropy to compare to each other two empirical probabilities. We deal with the two-sample framework in this paper, considering that the two datasets are generated in the same distribution p. We are thus given iid observations $X_1, ..., X_{n+m}$ with $\mathbb{P}(X_j \in \Omega_i) = p_i$ for all $j \in$

3

$[\![1, n+m]\!]$ and $i \in [\![1, k]\!]$. We define a first empirical probability based on n observations, following equation (1), and a second empirical probability based on m other independent observations,

$$\widehat{q}_{m,i} = \frac{1}{m} \sum_{j=n+1}^{n+m} \mathbb{1}_{X_j \in \Omega_i}.$$

In what follows, we study both $D_{\mathrm{KL}}\left(\widehat{p}_n \| p\right)$ and $D_{\mathrm{KL}}\left(\widehat{p}_n \| \widehat{q}_m\right)$. It is worth mentioning a specific challenge in the two-sample relative entropy: even though $p_i > 0$ whatever i, one cannot guarantee that $\widehat{q}_{m,i} \neq 0$. Beyond this trivial situation which leads to an infinite estimate, $\widehat{q}_{m,i}$ can be lower than the true value p_i and amplify much the empirical relative entropy.

2.2   Asymptotic and pre-asymptotic distributions of empirical relative entropy

We first focus on the asymptotic distribution of empirical relative entropy, with one or two samples, thanks to the central limit theorem. Then, an extension of Berry-Esseen bounds to a nonlinear function of a sum statistic provides us with a non-asymptotic expression converging in distribution toward the limit of the central limit theorem. We are assuming that the data are generated according to the probability p, so that the theoretical relative entropy should be equal to zero. Because of a well-known bias, the empirical relative entropy is positive.

The main challenge, when studying the statistical properties of the empirical relative entropy, is that we have a nonlinear function of the observations. A Taylor expansion can however make the problem feasible. Unlike what is done in the classical delta method, the first-order term of the expansion is equal to zero, so we need a second-order expansion and thus a convergence toward a chi-squared distribution [8, 54]. Another possibility consists in using Wilks' theorem [70].

Regarding the speed of convergence toward the chi-squared distribution, we would like to use a Berry-Esseen approach. However, the literature dedicated to Berry-Esseen pre-asymptotic bounds in the case of a nonlinear statistic is very recent and still narrow. The purpose of Berry-Esseen inequality is to provide an upper bound to the Kolmogorov-Smirnov statistic between the distribution of a finite-sample statistic and its limit according to the central limit theorem. When the statistic is a nonlinear function of the observations, it may be possible to linearise it and thus to express the divergence with respect to a Gaussian distribution [58, 64]. This solution is not relevant in our case because relative entropy requires at least a quadratic approximation and has a non-Gaussian limit. Divergence with respect to non-Gaussian distributions have been scarcely explored in the literature. One can cite a first attempt with a statistic equal to the square of the sum of the observations, the limit being χ_1^2 [38]. It is a first step but it is not enough in our case for which the limit is χ_{k-1}^2. Promising results have been obtained in multidimensional extensions, with a chi-square limit, but with a number of degrees of freedom higher than 9 [9] or 5 [40, 41], and a constant in the bound not explicitly specified or obtained by an indirect numerical procedure, requiring for example the number of integer vectors in a given ellipsoid [39, Theorem 1]. A very recent article also puts forward a solution which is valid whatever the number of degrees of freedom of the chi-squared distribution, but with unspecified constants and a limited domain of validity that excludes the right tail of the distribution [25]. Unfortunately, the right tail is quite important for our application to a statistical test for change-point detection.

Theorem 1 gives the central limit of the one-sample relative entropy along with pre-asymptotic bounds in a Berry-Esseen approach. These bounds constitute a new result. We think it is one of the very rare attempts to obtain Berry-Esseen bounds for a statistic defined by a non-trivial nonlinear function of observations which does not converge to a Gaussian. It takes into account both the error of the quadratic approximation, known as the relative Pearson divergence, and the speed

of convergence of this approximation. It exploits Raič's theorem, which is a recent multivariate extension of Berry-Esseen inequality, with well-specified constants [60].

Theorem 1. *Let $X_1, ..., X_n$ be iid variables such that $\mathbb{P}(X_j \in \Omega_i) = p_i$ with $p = (p_1, ..., p_k)^t \in (0,1)^k$. Then, when $n \to \infty$, we have*

$$2nD_{KL}\left(\widehat{p}_n \| p\right) \xrightarrow{d} \chi^2_{k-1},$$ (2)

where $\xrightarrow{d}$ stands for the convergence in distribution. Let $x > 0$. We have

$$F_{\chi^2_{k-1}}\left(\kappa^{down}_{n,k}(x)\right) - \mathcal{E}_{n,k} \leq \mathbb{P}\left(2nD_{KL}\left(\widehat{p}_n \| p\right) \leq x\right) \leq F_{\chi^2_{k-1}}\left(\kappa^{up}_{n,k}(x)\right) + \mathcal{E}_{n,k},$$ (3)

where $F_{\chi^2_{k-1}}$ is the cdf of the χ^2_{k-1} distribution,

$$\mathcal{E}_{n,k} = \left(42(k-1)^{1/4} + 16\right) \sum_{i=1}^{k} \frac{(1-p_i)^{3/2}}{(np_i)^{1/2}},$$ (4)

and where, for $\eta \in \{up, down\}$, we have

$$\sqrt{\kappa^{\eta}_{n,k}(x)} = \begin{cases} \min\{(-1)^{\mathbb{1}_{\eta=up}}\kappa_{n,k,r}(x) | r \in \{0,1,2\}, (-1)^{\mathbb{1}_{\eta=up}}\kappa_{n,k,r}(x) > 0\} & \text{if } 27x \leq 4\mu n \\ \min\{(-1)^{\mathbb{1}_{\eta=up}}\kappa_{n,k,>}(x) | (-1)^{\mathbb{1}_{\eta=up}}\kappa_{n,k,>}(x) > 0\} & \text{else,} \end{cases}$$

with the convention $\min(\emptyset) = +\infty$, the notation $\mu = \min_{i \in [\![1,k]\!]} p_i$, as well as

$$\kappa_{n,k,r}(x) = \frac{\sqrt{\mu n}}{3}\left[2\cos\left(\frac{1}{3}\arccos\left(\frac{27x}{2\mu n} - 1\right) - \frac{2r\pi}{3}\right) - 1\right]$$

and

$$\kappa_{n,k,>}(x) = \sqrt[3]{\frac{\sqrt{\mu n}}{3}}\left[\sqrt[3]{-\frac{\mu n}{9} + \frac{3x}{2} + \sqrt{-\frac{\mu nx}{3} + \frac{9}{4}x^2}} + \sqrt[3]{-\frac{\mu n}{9} + \frac{3x}{2} - \sqrt{-\frac{\mu nx}{3} + \frac{9}{4}x^2}}\right] - \frac{\sqrt{\mu n}}{3}.$$

Moreover, if $n \to \infty$, we have

$$\begin{cases} \kappa^{up}_{n,k}(x) & = & x + \frac{x^{3/2}}{\sqrt{\mu n}} + \mathcal{O}\left(\frac{1}{n}\right) \\ \kappa^{down}_{n,k}(x) & = & x - \frac{x^{3/2}}{\sqrt{\mu n}} + \mathcal{O}\left(\frac{1}{n}\right). \end{cases}$$ (5)

The proof of Theorem 1 is postponed in Appendix B.

Formula (3) gives pre-asymptotic bounds for the cdf of the empirical relative entropy. The bounds deal with two approximations. First, $\mathcal{E}_{n,k}$ is the Berry-Esseen component, related to the speed of convergence toward the asymptotic distribution. Second, the chi-squared cdf are not simple functions of x as it would be the case if relative entropy was a simple quadratic function. The variable x is to be replaced by $\kappa^{\text{up}}_{n,k}(x)$ and $\kappa^{\text{down}}_{n,k}(x)$, which take into account the error of the quadratic approximation of relative entropy. These quantities are defined as the smallest positive solutions of a cubic equation. When n increases, $\kappa^{\text{up}}_{n,k}(x)$ and $\kappa^{\text{down}}_{n,k}(x)$ tend toward x, as one can see in equation (5).

We think our approach can be reproduced for obtaining pre-asymptotic bounds for the cdf of some nonlinear function of sum statistics, under some convexity condition: considering a Taylor expansion of the nonlinear function, using Raič's theorem to get the speed of convergence, taking

into account a bound of the residual, which in our case is more subtle than the maximum third derivative, which is infinite.

The Berry-Esseen part of the bounds is uniform in x. There is a recent effort in the literature to obtain non-uniform bounds in the linear framework [57]. Our problem could certainly benefit in the future from potential extensions of these non-uniform bounds to nonlinear functions of sum statistics.

For the two-sample problem, Theorem 2 proposes an asymptotic distribution. We haven't found such a result in the literature but it is worth mentioning a close contribution with the asymptotic distribution of the two-sample Jeffreys divergence, which is a symmetric version of relative entropy [37].

Theorem 2. *Let $X_1, ..., X_{n+m}$ be iid variables such that $\mathbb{P}(X_j \in \Omega_i) = p_i$ with $p = (p_1, ..., p_k)^t \in (0,1)^k$. Then, when $n \to \infty$, $m \to \infty$, and $\frac{n}{n+m} \to \lambda \in (0,1)$, we have*

$$2\frac{nm}{n+m}D_{KL}(\widehat{p}_n \| \widehat{q}_m) \xrightarrow{d} \chi^2_{k-1}.$$

The proof of Theorem 2 is postponed in Appendix C.

When the two samples have the same size, that is $n = m$, $D_{KL}(\widehat{p}_n \| \widehat{q}_m)$ is asymptotically distributed like χ^2_{k-1}/n. It is to be compared to the more concentrated asymptotic distribution of $D_{KL}(\widehat{p}_n \| p)$, which is $\chi^2_{k-1}/2n$.

In the two-sample case, we do not propose Berry-Esseen bounds. Indeed, in the one-sample case, we were able to find an upper bound of the rest expressed as a simple function of the two probabilities. But in the two-sample case the rest of the quadratic approximation of relative entropy depends on the divergence between each empirical probability and the true probability, that is $\widehat{p}_{n,i} - p_i$ and $\widehat{q}_{m,i} - p_i$, and not on the difference between the two empirical probabilities, $\widehat{p}_{n,i} - \widehat{q}_{m,i}$. We can however propose a Berry-Esseen bound for the quadratic approximation instead of the relative entropy itself, as shown in Proposition 1. In Section 2.3, we will present finite-sample results in the two-sample case, directly applied to relative entropy.

Proposition 1. *With the assumptions of Theorem 2 and $x > 0$, we have*

$$\left| \mathbb{P}\left(\frac{nm}{n+m} \sum_{i=1}^{k} \frac{(\widehat{p}_i - \widehat{q}_i)^2}{p_i} \leq x \right) - F_{\chi^2_{k-1}}(x) \right| \leq \frac{n^2 + m^2}{(nm)^{1/2}(n+m)} \mathcal{E}_{n+m,k},$$

with $\mathcal{E}_{\cdot,\cdot}$ defined in equation (4).

The proof of Proposition 1 is postponed in Appendix C.

When $n = m$, the bound in Proposition 1 is $\mathcal{E}_{2n,k}$, that is $\mathcal{E}_{n,k}/2^{1/2}$.

2.3 Concentration inequalities for empirical relative entropy

Beside the Berry-Esseen bounds, one can obtain finite-sample bounds of the distribution of the empirical relative entropy by the mean of concentration inequalities. They in general offer simpler expressions than Berry-Esseen bounds, without referring to the limit distribution. Instead, they exploit various methods such as the method of types for the famous Sanov inequality, and a recursion technique or the moment-generating function in the two promising alternatives we present below. In addition, as we will see in Section 4.1, concentration inequalities give tighter bounds than the Berry-Esseen approach when n is small.

6

In what follows, we expose three existing inequalities, with two among the most recent and promising ones, along with a small refinement for the last one. We propose as well a new concentration inequalities in the two-sample case.

Sanov inequality is the most well-known concentration inequality for relative entropy. It writes

$$\mathbb{P}\left(D_{\mathrm{KL}}\left(\widehat{p}_n \| p\right) \geq x\right) \leq \frac{(n+k-1)!}{n!(k-1)!} e^{-nx} \leq (n+1)^k e^{-nx}, \tag{6}$$

some works focusing on the first inequality [22, 54], while others prefer the second one [21, Theorem 11.2.1], which is a simpler bound, in particular when k is large.

Based on a recursion technique, Mardia's bounds improve Sanov inequality, in particular when one increases k. There are several Mardia's bounds, each of which apply to a specific range of values for k. Among them, we focus on the one that demonstrated superior performance for the values of k considered in our tests:

$$\mathbb{P}\left(D_{\mathrm{KL}}\left(\widehat{p}_n \| p\right) \geq x\right) \leq \frac{6e^2}{\pi^{3/2}}\left(\frac{ne^3}{2\pi k}\right)^{k/2} e^{-nx}. \tag{7}$$

It holds when $3 \leq k \leq 2 + \sqrt{ne^3/2\pi}$ [54].

Agrawal proposed another concentration inequality exploiting the moment-generating function of the empirical relative entropy [2]. We expose it in the following proposition, along with a slightly improved version.

Proposition 2. *If $x > (k-1)/n$, then*

$$\mathbb{P}(D_{KL}\left(\widehat{p}_n \| p\right) \geq x) \leq \mathcal{M}_{k,n}^1(x) \leq \mathcal{M}_{k,n}^2(x) \leq \mathcal{M}_{k,n}^3(x),$$

with

$$\begin{cases} \mathcal{M}_{k,n}^1(x) &= \inf_{t\in[0,n)} e^{-tx}\left(\sum_{j=0}^n \frac{n!}{n^{2j}(n-j)!}t^j\right)^{k-1} \\ \mathcal{M}_{k,n}^2(x) &= e^{-nx}\left(\sum_{j=0}^n \frac{n!e}{n^j(n-j)!}\left(1-\frac{k-1}{nx}\right)^j\right)^{k-1} \\ \mathcal{M}_{k,n}^3(x) &= e^{-nx}\left(\frac{enx}{k-1}\right)^{k-1}. \end{cases}$$

The proof of Proposition 2 is postponed in Appendix D.

The third bound in Proposition 2 is the one put forward by Agrawal. As we will see in Section 4.1, it is both simple and very performing, compared to the ones of equations (6) and (7), at least when k is small. The second bound, $\mathcal{M}_{k,n}^2(x)$, slightly improves $\mathcal{M}_{k,n}^3(x)$, but it is more appropriate when n is small because it requires the calculation of a sum of n terms. The algorithmic complexity for obtaining the first bound, $\mathcal{M}_{k,n}^1(x)$, is even worse. Indeed, in addition to the sum of n terms, it requires a numerical optimization.

We also remark, beyond the traditional e^{-nx} of the concentration inequalities, that x appears in other parts of the expression of the bounds of Proposition 2, whereas neither equation (6) nor equation (7) exhibits this. The consequence is that Agrawal's bounds are closer to the true probability when x is small, compared to other methods. When one looks for a quantile at a given probability, k small or n not too small lead to a small quantile and thus Agrawal's formula provides us with a tighter upper bound of the quantile, compared to the alternatives of equations (6) and (7). This will be confirmed in the study presented in Section 4.1.

For building a concentration inequality in the two-sample framework, we can split relative entropy in two parts so that we can directly sum the upper bounds of the one-sample case. But such a

decomposition does not naturally arise because there is no triangle inequality for relative entropy. Pinsker's inequality shows however a correspondence between relative entropy and total variation, which could be used to obtain the desired decomposition. For discrete probabilities, it writes

$$\left(\sum_{i=1}^{k}|p_i - q_i|\right)^2 \leq 2D_{\mathrm{KL}}\left(p\|q\right) \leq \left(\frac{1}{\min_{i\in[\![1,k]\!]} q_i} - \min_{i\in[\![1,k]\!]}\frac{p_i}{q_i}\right)\left(\sum_{i=1}^{k}|p_i - q_i|\right)^2, \qquad (8)$$

the left bound being the traditional Pinsker's inequality [68, 17] and the right bound one of the reverse Pinsker's inequalities [62], among several other versions [14, 63]. We note that $(1/\min_i q_i) - (\min_i p_i/q_i) \geq k - 1$. We thus get, as an interesting side result, the following decomposition of relative entropy.

Proposition 3. *Let p, q, and r be categorical distributions with k categories. If $\min_{i\in[\![1,k]\!]} q_i > 0$, we have:*

$$D_{KL}\left(p\|q\right) \leq \left(\frac{2}{\min_{i\in[\![1,k]\!]} q_i} - 2\min_{i\in[\![1,k]\!]}\frac{p_i}{q_i}\right)\left(D_{KL}\left(p\|r\right) + D_{KL}\left(q\|r\right)\right).$$

The proof of Proposition 3 is postponed in Appendix E.1.

In the particular case $r = p$, it simplifies to

$$D_{\mathrm{KL}}\left(p\|q\right) \leq \left(\frac{2}{\min_{i\in[\![1,k]\!]} q_i} - 2\min_{i\in[\![1,k]\!]}\frac{p_i}{q_i}\right)D_{\mathrm{KL}}\left(q\|p\right).$$

The scalar in front of the relative entropy on the right-hand side of the above equation can be very large when one considers a probability q that is far from being uniform and a probability p that largely diverges from q. We also note that, in Proposition 3, the probabilities can be exchanged in each relative entropy of the right-hand side of the inequality. Therefore, $D_{\mathrm{KL}}\left(p\|r\right) + D_{\mathrm{KL}}\left(q\|r\right)$ can for example be replaced by $D_{\mathrm{KL}}\left(p\|r\right) + D_{\mathrm{KL}}\left(r\|q\right)$.

Using the one-sample Agrawal's concentration inequality introduced in Proposition 2 along with the decomposition put forward in Proposition 3, we get a concentration inequality for the two-sample relative entropy, when the two samples are assumed to follow the same theoretical distribution. It is the purpose of Theorem 3.

Theorem 3. *Let $m, n > 0$ and $X_1, ..., X_{n+m}$ be iid variables such that $\mathbb{P}(X_j \in \Omega_i) = p_i$ with $p = (p_1, ..., p_k)^t \in (0,1)^k$. We note*

$$\beta_{m,n} = \frac{2}{\min_{i\in[\![1,k]\!]} \widehat{q}_{m,i}} - 2\min_{i\in[\![1,k]\!]}\frac{\widehat{p}_{n,i}}{\widehat{q}_{m,i}}$$

and we assume that $\min_{i\in[\![1,k]\!]} \widehat{q}_{m,i} > 0$ and that $x \geq \beta_{m,n}(k-1)(m+n)/mn$. Then,

$$\boxed{\mathbb{P}\left(D_{KL}\left(\widehat{p}_n\|\widehat{q}_m\right) \geq x\right) \leq \widetilde{\mathcal{M}}_{k,n,m}^1(x) \leq \widetilde{\mathcal{M}}_{k,n,m}^2(x) \leq \widetilde{\mathcal{M}}_{k,n,m}^3(x),}$$

with

$$\begin{cases} \widetilde{\mathcal{M}}_{k,n,m}^1(x) = \inf_{s\in[0,\min(m,n))} e^{-sx/\beta_{m,n}}\left(\sum_{i=0}^{m}\frac{m!}{m^{2i}(m-i)!}s^i \sum_{j=0}^{n}\frac{n!}{n^{2j}(n-j)!}s^j\right)^{k-1} \\ \widetilde{\mathcal{M}}_{k,n,m}^2(x) = e^{-\sigma_{m,n,x}x/\beta_{m,n}}\left(\sum_{i=0}^{m}\frac{m!}{m^{2i}(m-i)!}\sigma_{m,n,x}^i \sum_{j=0}^{n}\frac{n!}{n^{2j}(n-j)!}\sigma_{m,n,x}^j\right)^{k-1} \\ \widetilde{\mathcal{M}}_{k,n,m}^3(x) = e^{-\sigma_{m,n,x}x/\beta_{m,n}}\left(\left(1-\frac{\sigma_{m,n,x}}{m}\right)\left(1-\frac{\sigma_{m,n,x}}{n}\right)\right)^{1-k} \end{cases}$$

and

$$\sigma_{m,n,x} = \frac{n+m}{2} - \frac{\beta_{m,n}(k-1)}{x} - \sqrt{\frac{\beta_{m,n}^2(k-1)^2}{x^2} + \frac{(m-n)^2}{4}}.$$

8

The proof of Theorem 3 is postponed in Appendix E.2.

When the two samples have the same size, that is $m = n$, $\sigma_{m,n,x}$ becomes $n - 2\beta_{n,n}(k-1)/x$, the bounds simplify, and we get, for example,

$$\widetilde{\mathcal{M}}^3_{k,n,n}(x) = e^{-nx/\beta_{n,n}} \left(\frac{enx}{2\beta_{n,n}(k-1)} \right)^{2(k-1)}$$

or

$$\widetilde{\mathcal{M}}^2_{k,n,n}(x) = e^{-nx/\beta_{n,n}} \left(e \sum_{i=0}^{n} \left(1 - \frac{2\beta_{n,n}(k-1)}{nx} \right)^i \left[\prod_{j=0}^{i-1} \left(1 - \frac{j}{n} \right) \right] \right)^{2(k-1)},$$

expression in which we have replace the ratio of factorials by a product that is equal but easier to compute for large values of n. When $m \to \infty$, we have

$$\beta_{m,n} \to \beta_n = \frac{2}{\min_{i \in [\![1,k]\!]} q_i} - 2 \min_{i \in [\![1,k]\!]} \frac{\widehat{p}_{n,i}}{q_i},$$

as well as $\sigma_{m,n,x} \to \sigma_{n,x} = n - \beta_n(k-1)/x$ and

$$\widetilde{\mathcal{M}}^3_{k,n,m}(x) \to e^{-nx/\beta_n} \left(\frac{enx}{\beta_n(k-1)} \right)^{k-1}.$$

We remark that this limit is different from the expression of $\mathcal{M}^3_{k,n}(x)$ provided for the one-sample case in Proposition 2. The reason is the presence of β_n, which should be 1 in order for the limit to match the one-sample case, but which in reality is higher than $2(k-1)$. It thus appears that the reverse Pinsker's inequality is quite pessimistic and that the $\beta_{m,n}$ of Theorem 3 is too large. In Section 4.1, we show that replacing $\beta_{m,n}$ by 1 leads to upper bounds that are numerically satisfying. We note $\widetilde{\mathcal{M}}^{2,\star}_{k,n,m}(x)$ and $\widetilde{\mathcal{M}}^{3,\star}_{k,n,m}(x)$ these new quantities. Even though we cannot prove it, we conjecture that Theorem 3 may apply for a large range of probabilities when one replaces $\beta_{m,n}$ by 1.

3 Change-point detection

Since relative entropy measures the divergence between two distributions, we can use it in the framework of change-point detection. The various bounds for the distribution of its empirical counterpart thus provide possible thresholds for a statistical test of change-point. We introduce the test along with some other traditional offline alternatives. One of them, based on a difference of AIC is close in some ways to our method. We will thus compare the two approaches.

3.1 Relative entropy and statistical tests for change-point detection

We adopt the classical formalism of change-point detection [76]. We consider a sequence $X_1, ..., X_{2n}$. Under the null hypothesis, $X_1, ..., X_{2n}$ are identically distributed. Under the alternative hypothesis, there exists $t^\star \in [\![2, 2n-1]\!]$ such that $X_1, ..., X_{t^\star}$ are identically distributed and follow a discrete probability q, whereas $X_{t^\star+1}, ..., X_{2n}$ are also identically distributed but follow another discrete probability p.

Two families of tests exist, based either on an online statistic or on an offline statistic. In the online case, $t^\star$ is close to $2n$ and the sequential update of the statistic is supposed to lead to a rapid

detection of the change-point [72]. CUSUM is a widespread statistic used in this framework [56] and online change-point detection often consists in detecting a change in the cumulative mean or in another moment. In the offline case, one generally focuses on $t^\star = n$ [47], so that people consider a divergence statistic between the probabilities of the two sub-samples. The offline approach makes it possible to compare probabilities, parametric or nonparametric ones, instead of only moments.

In this article, we are interested in offline change-point detection, specifically when n is not very large. Our test thus consists in $X_1, ..., X_n \sim q$ and $X_{n+1}, ..., X_{2n} \sim p$, with H_0 corresponding to $p = q$ and H_1 to $p \neq q$. Translating the spirit of CUSUM in the offline approach, we propose two baseline change-point tests based on the comparison of the mean (respectively the variance) of the sub-samples $X_1, ..., X_n$ and $X_{n+1}, ..., X_{2n}$, using thus a t-test (resp. F-test), for which the asymptotic distribution is well known.

The method we put forward here is based on the empirical relative entropy $D_{\mathrm{KL}}(\widehat{p}_n \| \widehat{q}_n)$, with the two distributions $\widehat{p}_n$ and $\widehat{q}_n$ estimated on the two sub-samples. Since the exact cdf $F_{H_0, n}$ of $D_{\mathrm{KL}}(\widehat{p}_n \| \widehat{q}_n)$ under H_0 is unknown, both in the one- and in the two-sample cases, simulations are often used to select a threshold corresponding to a given significance level [59]. Instead, using the two-sample asymptotic distribution of $D_{\mathrm{KL}}(\widehat{p}_n \| \widehat{q}_n)$, as described in Theorem 2, we have a good approximation of its true cdf under H_0, provided that n is not too small. For a more conservative approximation of $F_{H_0, n}$, specifically for small values of n, we can use the two-sample concentration inequalities introduced in Theorem 3. Whatever the approximation $\widehat{F}_{H_0, n}$ of $F_{H_0, n}$, the quantile $\widehat{F}_{H_0, n}^{-1}(1 - \alpha)$ provides us with a threshold that can be used to define a test of nominal significance level α. If $D_{\mathrm{KL}}(\widehat{p}_n \| \widehat{q}_n)$ is above $\widehat{F}_{H_0, n}^{-1}(1 - \alpha)$, we reject H_0.

The simulation study of Section 4 evaluates the different bounds of the cdf $F_{H_0, n}$ and provides as well the actual significance level and the power of the statistical tests introduced above.

3.2 Another approach using AIC

Another possible offline method to detect a change-point consists in determining whether a single model for describing $X_1, ..., X_{2n}$ is more relevant than using a model for $X_1, ..., X_n$ and another one for $X_{n+1}, ..., X_{2n}$. This can be done by comparing information criteria in the two settings. This approach has been explored both for parametric models [48, 29] and nonparametric ones [77, 44].

Like in Section 3.1, we consider the distributions $\widehat{q}_n$ and $\widehat{p}_n$, empirical counterparts of the k-categorical distributions q and p describing respectively the subsequence $X_1, ..., X_n$ and $X_{n+1}, ..., X_{2n}$. Thus, the likelihood of $(X_1, ..., X_n)$ is $\prod_{j=1}^{n} q_{X_j} = \prod_{i=1}^{k} q_i^{n\widehat{q}_{n,i}}$. Plugging the estimator $\widehat{q}_n$ of q in this expression and doing the same for $(X_{n+1}, ..., X_{2n})$ leads to the following log-likelihood of $(X_1, ..., X_{2n})$ in the two-model case, which corresponds to H_1:

$$\ell_{H_1} = n \sum_{i=1}^{k} \left\{ \widehat{p}_{n,i} \log(\widehat{p}_{n,i}) + \widehat{q}_{n,i} \log(\widehat{q}_{n,i}) \right\}.$$

If a single model describes the sequence $X_1, ..., X_{2n}$, that is under H_0, the empirical probability of the category i is $(\widehat{p}_{n,i} + \widehat{q}_{n,i})/2$, so that the log-likelihood of $(X_1, ..., X_{2n})$ is now

$$\ell_{H_0} = n \sum_{i=1}^{k} (\widehat{p}_{n,i} + \widehat{q}_{n,i}) \log\left(\frac{\widehat{p}_{n,i} + \widehat{q}_{n,i}}{2} \right).$$

Noting that we have $k - 1$ parameters in the one-model case and $2(k - 1)$ otherwise, we obtain the following criterion of difference of AICs between the approaches:

$$\Delta \mathrm{AIC}(\widehat{p}_n, \widehat{q}_n) = -2(k - 1) - 2(\ell_{H_0} - \ell_{H_1}).$$

If $\Delta \mathrm{AIC}(\widehat{p}_n, \widehat{q}_n) > 0$ then the two-model approach is informationally more relevant and we validate the presence of a change-point.

3.3 Link between relative entropy and the AIC-based method

Writing, for all $i \in [\![1, k]\!]$, that $\widehat{p}_{n,i} = \widehat{q}_{n,i} + \varepsilon_i$, a second-order Taylor expansion provides us with

$$(\widehat{p}_{n,i} + \widehat{q}_{n,i}) \log\left(\frac{\widehat{p}_{n,i} + \widehat{q}_{n,i}}{2}\right) - \widehat{p}_{n,i} \log(\widehat{p}_{n,i}) - \widehat{q}_{n,i} \log(\widehat{q}_{n,i}) = -\frac{\varepsilon_i^2}{4\widehat{q}_{n,i}} + o\left(\varepsilon_i^2\right)$$

and with

$$D_{\mathrm{KL}}\left(\widehat{p}_n \| \widehat{q}_n\right) = \sum_{i=1}^{k} \varepsilon_i + \frac{\varepsilon_i^2}{2\widehat{q}_{n,i}} + o\left(\varepsilon_i^2\right).$$

Since $\widehat{p}_{n,i}$ and $\widehat{q}_{n,i}$ are probabilities, we have $\sum_{i=1}^{k} \varepsilon_i = 0$ and finally

$$\Delta \mathrm{AIC}(\widehat{q}_n + \varepsilon, \widehat{q}_n) = -2(k-1) + nD_{\mathrm{KL}}\left(\widehat{q}_n + \varepsilon \| \widehat{q}_n\right) + o\left(\|\varepsilon\|_2^2\right). \tag{9}$$

Therefore, both approaches, based either on relative entropy or on a difference of AICs, are very close to each other, up to an affine transformation, as soon as the two subsequences have close probabilities. The method based on $\Delta \mathrm{AIC}$ is not a statistical test but equation (9) gives a natural threshold on the relative entropy for determining whether there is a change-point in the sequence. This threshold is $2(k-1)/n$ and is not based on any of the approximations presented in Section 2 of the distribution of the relative entropy.

The proximity between relative entropy and $\Delta \mathrm{AIC}$ is however not a surprise. Indeed, the justification of AIC comes from the expectation of the relative entropy between the data-generating distribution and an estimated parametric distribution, the number of parameters appearing because of this expectation [3, 19]. However, the fact that the first distribution is in practice replaced by an empirical distribution is overlooked in the AIC method, which thus clearly states an asymptotic framework. Extensions of AIC, like the corrected AIC [45, 18], may be used instead in order to better deal with small samples.

Change-point detection methods based either on the difference of AICs or on the relative entropy both depend on the discretization parameter k. As we will see in the simulations of Section 4.2 and in the empirical study of Section 5, one detects a bigger number of change-points when k is larger. Indeed, the differences between two distributions are more apparent with finer discretizations. We might therefore be inclined to increase k but a limitation appears when we estimate a zero probability in some categories, because of a too large k with respect to n. Using continuous probabilities and the differential relative entropy would thus be an interesting extension of this work.

4 A simulation study

We want to compare all the bounds of the distribution of empirical relative entropy, provided in Section 2, with the distribution obtained by simulation, for various sample sizes. Then, we propose a comparison of the change-point detection methods introduced in Section 3, also based on simulations.

4.1 A simulation-based evaluation of the bounds of the distribution of relative entropy

In this paragraph, we consider a uniform distribution for both p and q. Indeed, numerical experiments indicate that the distribution of empirical relative entropy obtained from simulations does not seem to be significantly affected by changes in the probabilities, provided that the number of categories remains unchanged and that we stay under H_0: $p = q$.

We first evaluate the distributions coming from the central limit theorem and Berry-Esseen-like bounds, introduced in Section 2.2. We display in Figure 1 the cdf of the empirical relative entropy obtained by simulations, along with the asymptotic distribution and Berry-Esseen bounds, for n equal to $500,000$ or $2,000,000$ and $k \in \{2,4\}$.
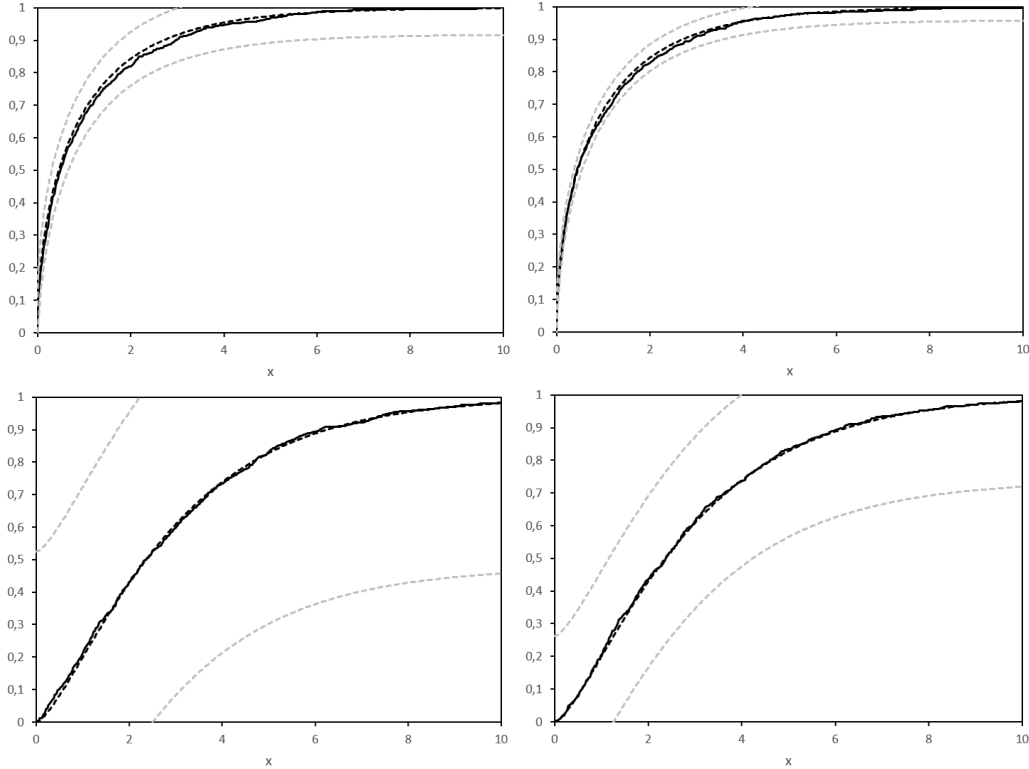

Figure 1: For $\widehat{p}_n$ estimator of the uniform probability p, asymptotic cdf of $2nD_{\mathrm{KL}}(\widehat{p}_n \| p)$ in x (black dotted line) along with the Berry-Esseen bounds (grey dotted lines), as in Theorem 1, and cdf obtained by 1,000 simulations (black solid line). The size of the sample n is 500,000 (left graphs) and 2,000,000 (right graphs), and the number of categories k is 2 (top graphs) and 4 (bottom graphs).

In this pre-asymptotic setting, we observe that the asymptotic cdf is very close to to simulated one and that the Berry-Esseen bounds seem pessimistic when $k = 4$. This is not surprising since our bound $\mathcal{E}_{n,k}$ is based on a uniform multivariate extension of Berry-Esseen inequality. Indeed, there is currently no consensus on an optimal choice of the constant that appears there, unlike the univariate case. In order to have an idea of the minimum n making the bounds relevant, we can stress that $\mathcal{E}_{n,k} < 1$ when $n > 1.37 \times 10^5$ (respectively $n > 3.36 \times 10^3$) in the case $k = 4$

(resp. $k = 2$). One may also wonder about the role of $\kappa_{n,k}^{\text{down}}(x)$ and $\kappa_{n,k}^{\text{up}}(x)$ in inequality 3. Their importance is very limited because they converge rapidly toward x as n goes to infinity. Specifically, in Figure 1, the differences $x - \kappa_{n,k}^{\text{down}}(x)$ and $\kappa_{n,k}^{\text{up}}(x) - x$ are increasing in x and reach 0.32% (respectively 0.45%, 0.64%, 0.91%) of x, for $x = 10$, $n = 2,000,000$, and $k = 2$ (resp. $n = 2,000,000$ and $k = 4$, $n = 500,000$ and $k = 2$, $n = 500,000$ and $k = 4$).

We now focus on finite-sample distributions and concentration inequalities. We will let k and n vary around the baseline case $n = 100$, $k = 4$. We compare quantiles of the empirical relative entropy at 75% and 95%, in the one- and two-sample cases, following various methods: the empirical quantiles obtained from 10,000 simulations, the quantile in the asymptotic distribution of Theorems 1 and 2, and the quantiles obtained from concentration inequalities. Specifically, in the one-sample case, the considered bounds are the two Sanov's bounds of equation (6), Mardia's bound provided in equation (7), and Agrawal's second and third bounds, $\mathcal{M}_{k,n}^2$ and $\mathcal{M}_{k,n}^3$, defined in Proposition 2. With two samples, the only bounds considered are the extensions of the above Agrawal's bounds, $\widetilde{\mathcal{M}}_{k,n,n}^{2,\star}$ and $\widetilde{\mathcal{M}}_{k,n,n}^{3,\star}$, defined after Theorem 3, with the same size for the two subsamples.

Some of the concentration inequalities are valid under some condition. Agrawal's condition $x > (k-1)/n$ is always verified in our results and we have to discard the case $k = 2$ for Mardia's bound. In the two-sample case, we also constrain k to be less than or equal to 8; otherwise, simulations sometimes lead to empty categories, and consequently to zero estimated probabilities $\widehat{q}_n$, when $n = 100$. This would cause the relative entropy to diverge to infinity.

The obtained quantiles, as functions of n, are displayed in Figure 2. We observe that the quantile of the asymptotic distribution is very accurate: it only slightly underestimates the true quantile when n is small (less than 50) and this underestimation becomes more pronounced the further we move into the tail of the distribution. Regarding concentration inequalities, Sanov's bounds largely overestimate the quantile. Mardia's bound is only slightly better than the best Sanov's bound, whereas the two Agrawal's bounds, which are very close to each other both in the one- and the two-sample cases, are significantly better than the other bounds, though not as accurate as the quantile of the asymptotic distribution.

Figure 3 shows the quantiles as functions of k. The quantile at 95% of the asymptotic distribution more clearly underestimates the true quantile when k is larger. Once again, Sanov's bounds are the least accurate. Mardia's bound comes next but it increases more slowly with k than the other concentration bounds do. This is consistent with its known reliability for large values of k. Nevertheless, for the range of values considered for k, Agrawal's bounds are our best concentration bounds. The two versions are again very close to each other, both for one and two samples. For this reason, we will now only consider $\mathcal{M}_{k,n}^3$ and $\widetilde{\mathcal{M}}_{k,n,m}^{3,\star}$, whose expression is simpler than $\mathcal{M}_{k,n}^2$ and $\widetilde{\mathcal{M}}_{k,n,m}^{2,\star}$.

4.2 Detection of change-point assessed by simulations

We now evaluate the propensity of the methods introduced in Section 3 to detect an existing change-point and not to trigger an alert when there is no change-point. The method put forward in this work is a statistical test on the empirical relative entropy, with various approximations of its distribution, but always in a two-sample context. The benchmark methods are a t-test, an F-test, and the Δ AIC method.

Among all the possible models depicting a change-point, we use a simple parametric categorical model, with k categories $\{1, 2, ..., k\}$ of ranked probability. For $i \in [\![1, k]\!]$, it is defined by

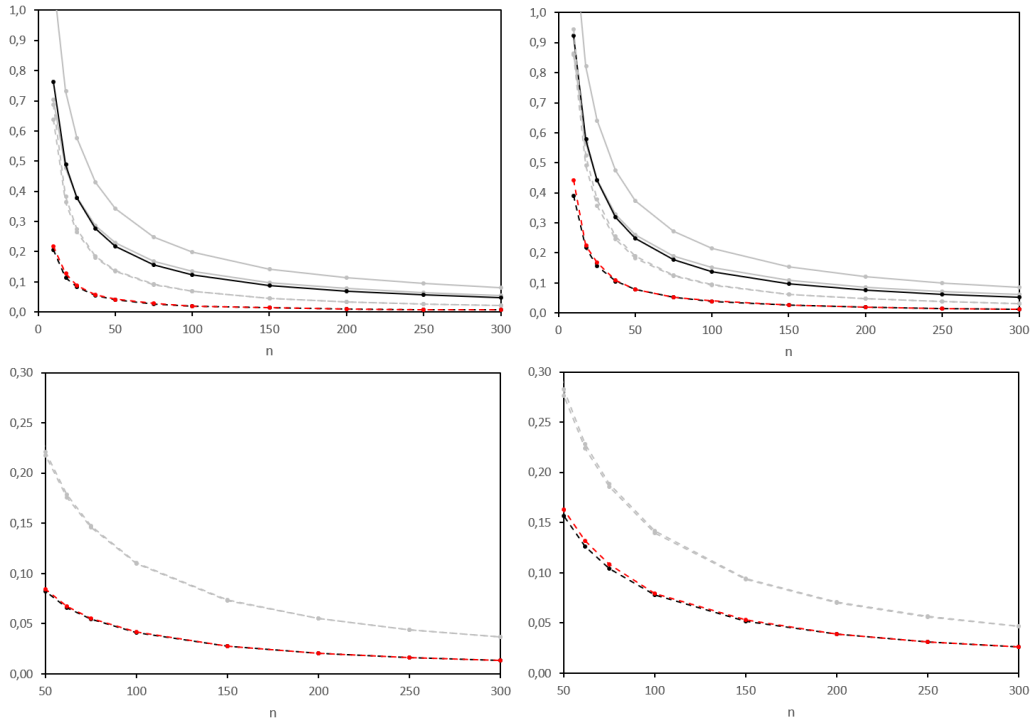$$\pi_i^{\phi,k} = \frac{e^{-\phi i}}{\sum_{j=1}^k e^{-\phi j}},$$

Figure 2: Quantile of the relative entropy as a function of n, at 75% (left graphs) and 95% (right graphs), for one (top graphs) or two samples (bottom graphs), according to the 10,000 simulations (red line), to the asymptotic distribution (black dashed line), to the second and third Agrawal's bounds (grey dashed lines), to the two Sanov's bounds (grey solid lines), and to Mardia's bound (black solid line). Other parameters are $k = 4$ and p_i constant in i.
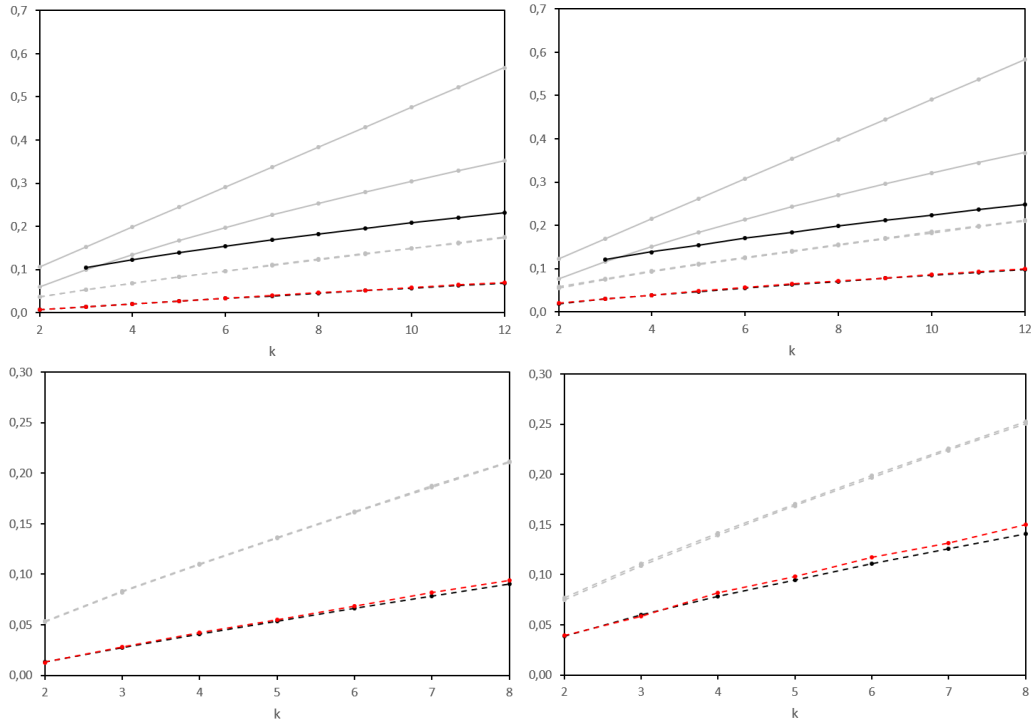
Figure 3: Quantile of the relative entropy as a function of k, at 75% (left graphs) and 95% (right graphs), for one (top graphs) or two samples (bottom graphs), according to the 10,000 simulations (red line), to the asymptotic distribution (black dashed line), to the second and third Agrawal's bounds (grey dashed lines), to the two Sanov's bounds (grey solid lines), and to Mardia's bound (black solid line). Other parameters are $n = 100$ and p_i constant in i.

which can be seen as either a finite version of a discrete exponential distribution [6] or a variation of the finite geometric distribution [20]. It is a simple way of modelling with a single parameter both the uniform distribution, when $\phi = 0$, and disparity between categories, when $\phi \neq 0$. In our change-point framework, we generate n first i.i.d. variables in the distribution $\pi^{\phi,k}$ then n others in the distribution $\pi^{\phi+\psi,k}$. The presence of a change-point is thus characterized by $\psi \neq 0$. The closer ψ is to zero, the less significant is the change-point. We use the same baseline case as in Section 4.1, that is $n = 100$, $k = 4$ and $\phi = 0$, and we consider the following standalone variations: $n = 50$, $k = 6$, and $\phi = 0.3$. In all the experiments, we consider a large range of values for ψ: $[-0.8, 0.8]$.

For a simulated trajectory of length $2n$, with or without a change-point, depending on the value of ψ, we carry out the statistical tests of Section 3. For the tests based on relative entropy, we estimate an empirical probability for each category rather than the parameter of the distribution used to generate the data. For each test, we record its outcome and average it over 10,000 trajectories simulated with the same parameters. We then consider the proportion of times each test rejects the null hypothesis of no change-point.

Figure 4 gathers the results. The nominal confidence level is set at 95%. Each curve intersects the Y-axis ($\psi = 0$) at a probability equal to one minus the actual confidence level. The size of each test can thus be read at this specific point of the corresponding graph. The Δ AIC approach is liberal, producing more inappropriate rejections of H_0 than expected. The other tests are at or above the nominal confidence level, with two of them being even conservative: the F-test and the relative entropy test based on the $\widetilde{\mathcal{M}}^{3,\star}_{k,n,n}$ approximation. The power of each test can be read directly from the graph when $\psi \neq 0$. It shows that the t-test is the most powerful, followed by the Δ AIC method, the relative entropy test based on the asymptotic distribution, the one based on $\widetilde{\mathcal{M}}^{3,\star}_{k,n,n}$, and finally the F-test, which has some difficulties detecting change-points.

Overall, the t-test seems the most appropriate for this example. However, this analysis, based on the sole average of random variables, cannot be as rich as an approach based on the full distribution, like tests on relative entropy. We thus consider another generating model, with states $\{1, 2, 3, 4\}$ of probability $q = (0.25, 0.25, 0.25, 0.25)$ for the first sub-sample and $p = (p_1, 0.5 - p_1, 0.5 - p_1, p_1)$ for the second one. The variables have the same expectation in the first and in the second sub-samples, but not the same distribution. Figure 5 shows that the t-test is now unable to detect change-points when they exist. The Δ AIC method is liberal again, whereas all the other methods have an appropriate actual confidence level.

The conclusions of the two examples indicate that change-point tests based on relative entropy are particularly relevant and do not seem to strongly depend on the way the data are generated.

5   Application to real data

This section presents an application of the method based on relative entropy to detect change-points. We study two datasets, one of temperature series and another one of volatility series.

5.1   Study of a climate dataset

Change-point detection in temperature time series is a well-established topic in the climatology literature, typically conducted on yearly or monthly averaged data. While CUSUM-type procedures are widespread [28], many existing approaches are designed for offline analysis. Most methods aim at detecting changes in the mean [61], or in linear or quadratic trends [27, 65]. Less frequently, nonparametric trend changes are considered, using for example wavelet-based techniques [27].
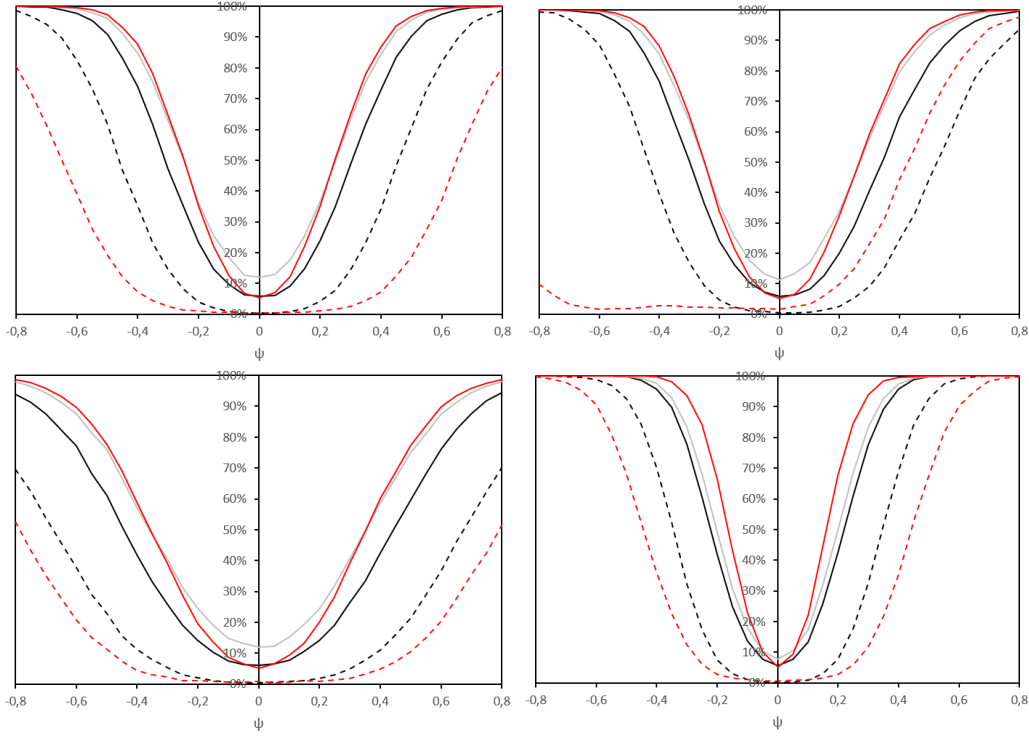
Figure 4: Proportion of the 10,000 simulated time series for which a change-point is detected, as a function of ψ. The tests used are based i/ on the relative entropy with a threshold defined by the 95% quantile in either the asymptotic two-sample distribution (solid black line) or the $\widetilde{\mathcal{M}}_{k,n,n}^{3,\star}$ bound (dashed black line); ii/ on the Δ AIC method (solid grey line); iii/ on the t-test with 95% confidence (solid red line); iv/ on the F-test with 95% confidence (dashed red line). The parameters are $(k, n, \phi) = (4, 100, 0)$, except in the top right graph ($\phi = 0.3$), in the bottom left graph ($n = 50$), and in the bottom right graph ($k = 6$).
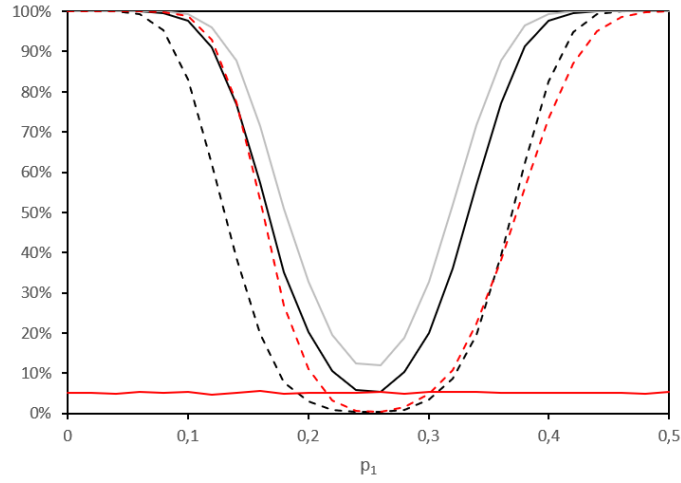
17

Figure 5: Proportion of the 10,000 simulated time series for which a change-point is detected, as a function of p_1, where the generating probability of the $n = 100$ first observations is $q = (0.25, 0.25, 0.25, 0.25)$ and, for the next n observations, $p = (p_1, 0.5 - p_1, 0.5 - p_1, p_1)$. The tests used are based i/ on the relative entropy with a threshold defined by the 95% quantile in either the asymptotic two-sample distribution (solid black line) or the $\widetilde{\mathcal{M}}_{k,n,n}^{3,\star}$ bound (dashed black line); ii/ on the Δ AIC method (solid grey line); iii/ on the t-test with 95% confidence (solid red line); iv/ on the F-test with 95% confidence (dashed red line).

Depending on the specification, the significance of the change-point is assessed using information criteria such as $\Delta\,\mathrm{AIC}$ [61, 65], classical statistical tests [28, 61], or statistical tests in a simulated distribution under a specific noise assumption, for example a white Gaussian noise, a long-range noise [13, 73], or even an alpha-stable noise [27]. Besides, standard normal homogeneity tests (SNHT) are also widely used in climatology [4, 28]. The main purpose of SNHT is to detect artificial changes in a station's time series, such as those caused by instrument replacement, by a comparison to the mean value of surrounding stations.

We use a public dataset of temperatures at Embrun (Hautes-Alpes, France), between January 1999 and December 2024, obtained from the website of Météo-France.[1] The dataset is sampled every 3 hours. We average it so that we get the daily and the weekly time series displayed in Figure 6. Of course, the seasonality is an important feature of such time series. We thus consider the distribution of daily or weekly temperatures in a one-year window. Then, we compare the distribution of two distinct years using relative entropy, like in the method described in Section 3. Constructed from high-resolution temporal data, this distribution-based approach therefore differs from the more commonly used climatological methods, which rely on yearly averages.
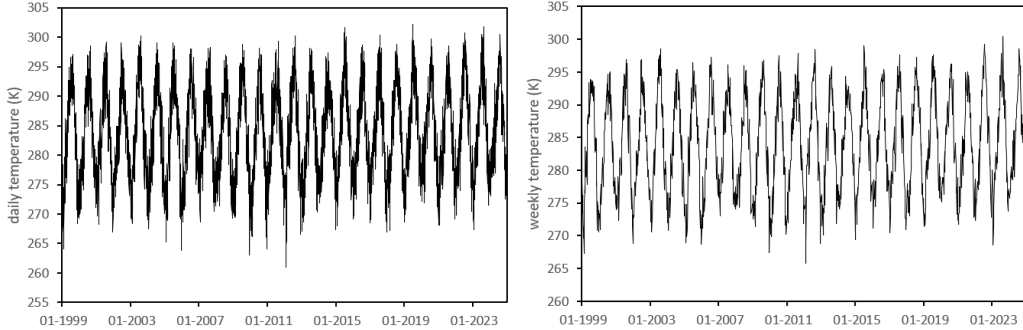

Figure 6: Daily (left) and weekly (right) average temperature at Embrun between January 1999 and December 2024.

Figure 7 shows this relative entropy, along with the bounds corresponding to the quantile at 99% of various distributions: the one- and two-sample asymptotic distribution and Agrawal's type bounds $\mathcal{M}_{k,n}^3$ and $\widetilde{\mathcal{M}}_{k,n,n}^{3,\star}$. The $\Delta\,\mathrm{AIC}$ is very close to the affine transformation of the relative entropy, as expressed in equation (9), so that we also display the threshold $2(k-1)/n$ as a relevant approximation of the significance bound of the relative entropy. The two important parameters of our change-point detection method are n and k. The first one simply corresponds to the number of days or week in a year. Regarding k, we consider a small value, 4, and a bigger one, 10 for daily and only 6 for weekly data. Indeed, when we simply consider 52 weekly observations, we get several probabilities equal to zero if we work with a fine discretization of 10 categories. We plot two curves showing the relative entropy between the year ending at the x-axis date and either the previous year or a reference year taken as the first year in the dataset.

The results shown in Figure 7 reveal several change-points. The greater k, the finer the discretization of the probability and the higher the resolution of change-point detection. Increasing n, by considering daily instead of weekly observations, also leads to the detection of more change-points. Although the weekly graphs show an upward trend in relative entropy, a large peak in three of the four graphs also indicates a strong change-point in 2007. It correspond to a particularly hot winter in 2006-2007. This is confirmed by a kernel density representation of the temperatures this year, compared to both the density the year before and the density during the first year of the

---

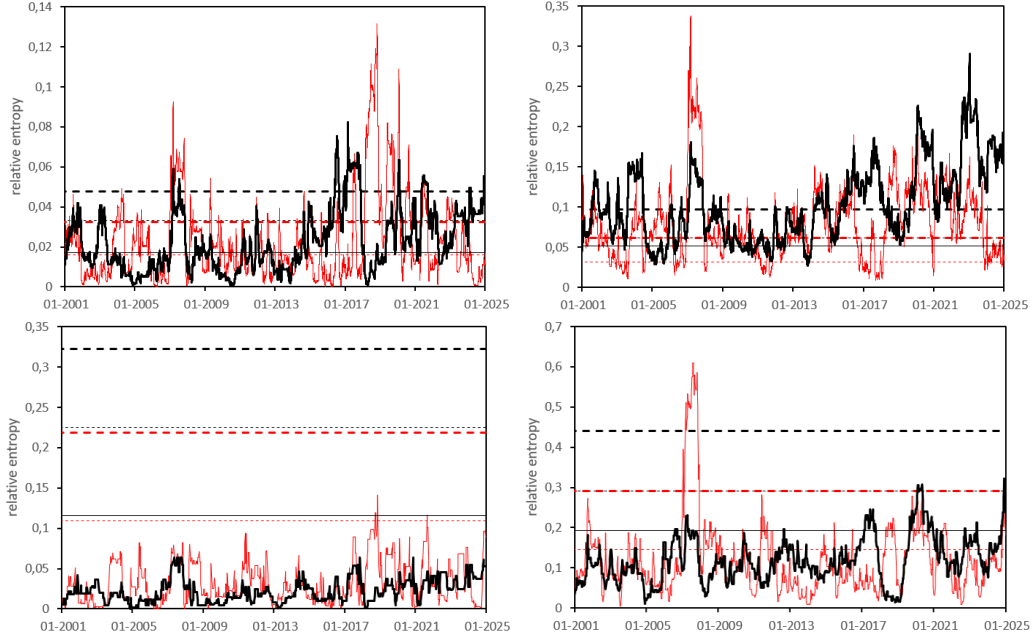[1] https://donneespubliques.meteofrance.fr/, station 7591.

Figure 7: Relative entropy between empirical probabilities estimated on a year of data finishing at the date indicated in abscissa and on either the previous year (red curve) or the first year of data (black curve). The categories are delimited by empirical quantiles on the whole dataset, of probabilities $\{1/k, 2/k, ..., (k-1)/k\}$, where k is either 4 (left graphs), 6 (bottom right graph), or 10 (top right graph). Dashed horizontal lines are confidence bounds at 99%, obtained using the one- and two-sample asymptotic distributions (thin and thick red), as well as $\mathcal{M}_{k,n}^3$ and $\widetilde{\mathcal{M}}_{k,n,n}^{3,\star}$ (thin and thick black). The solid horizontal line corresponds to the value $2(k-1)/n$, where n is the number of observations each year, for the daily (top graphs) or weekly sampling (bottom graphs).

dataset. See Figure 8. We see that the right tails coincide, whereas a large divergence appears in the left tails, in addition to the fluctuations in the bulk of the density. The bandwidth used for each density ranges between 1.3 and 1.7. It was selected so as to maximize a complexity criterion, thereby ensuring a good balance between underfitting and overfitting [31].
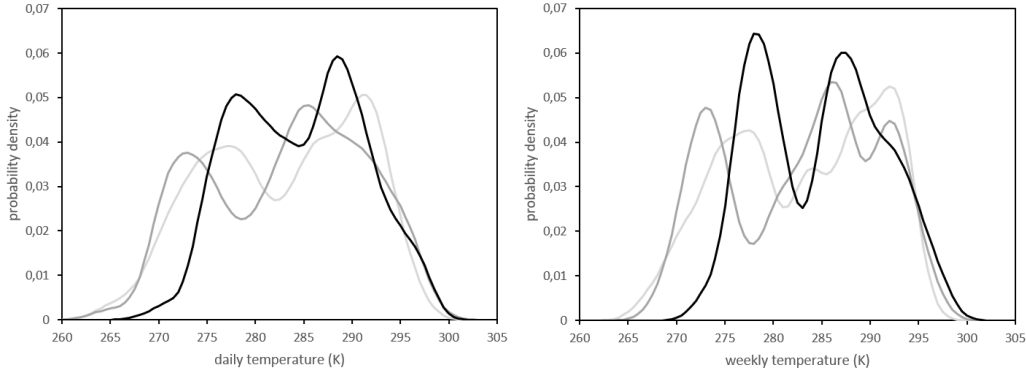

Figure 8: Kernel density estimates of the daily (left graph) and weekly (right graph) average temperature, with a complexity-based bandwidth. The data used correspond to a year of observations in 1999 (light grey), from July 2005 to June 2006 (dark grey), from July 2006 to June 2007 (black).

5.2 Study of a financial dataset

Volatility is a key quantity in financial markets, which leads investment decisions. Starting from a model in which volatility is a fixed parameter, a CUSUM method applied to squares of price returns was first proposed to detect a change-point in volatility [46]. The heteroscedastic nature of financial time series led to the introduction of stochastic processes of volatility, with serial dependence. However, the traditional CUSUM method too frequently misses the detection of change-points in this stochastic framework [23], requiring some adjustments [50]. Besides these online approaches, offline methods have been proposed to detect multiple change-points in the parameters of a stochastic volatility model like GARCH [51, 5]: they consist in finding the partition of time optimizing a least-square-based accuracy criterion along with a penalization, in the spirit of the methods based on information criteria described in Section 3.2.

While econometric models are usually defined in discrete time under regular sampling, quantitative finance research often turns to continuous-time volatility models, which can be discretized under irregular sampling schemes. Among these models, the rough volatility model, which is based on a geometric fractional Brownian motion (fBm), is very popular [35]. It makes it possible to depict serial dependence in volatility thanks to a single parameter, the Hurst exponent [16]. The propensity of this kind of model to forecast future volatility is promising [30, 74, 12]. However, some works suggest that the Hurst exponent is not enough to describe the serial dependence of volatility series, and that multifractal extensions [32, 71] or jumps [1, 34] are to be considered. The detection of change-points in volatility series is therefore crucial for identifying the time intervals in which the fBm framework remains valid. The solutions proposed in the literature consist in detecting jumps [11] or a change-point in the Hurst exponent using CUSUM [10] or using the local absolute variation [75].

We propose here to apply the change-point tests described in Section 3. The data are five daily

time series of realized volatility computed with a five-minute discretization of prices of stock indices, imported from the formerly available Oxford-Man Institute of Quantitative Finance Realized Library: the AEX index, the CAC 40 index, the Nikkei 225 index (N225), the Oslo Exchange All-share index (OSEAX), and the S&P 500 index (SPX). The series starts on January 2000, except N225, which starts in February 2000 and OSEAX in September 2001. The end date of our sample is on the 12th April 2021. We import as well from Bloomberg the VIX, which is obtained from option prices on the SPX, in the same time period 2000-2021. Figure 9 shows the six time series.
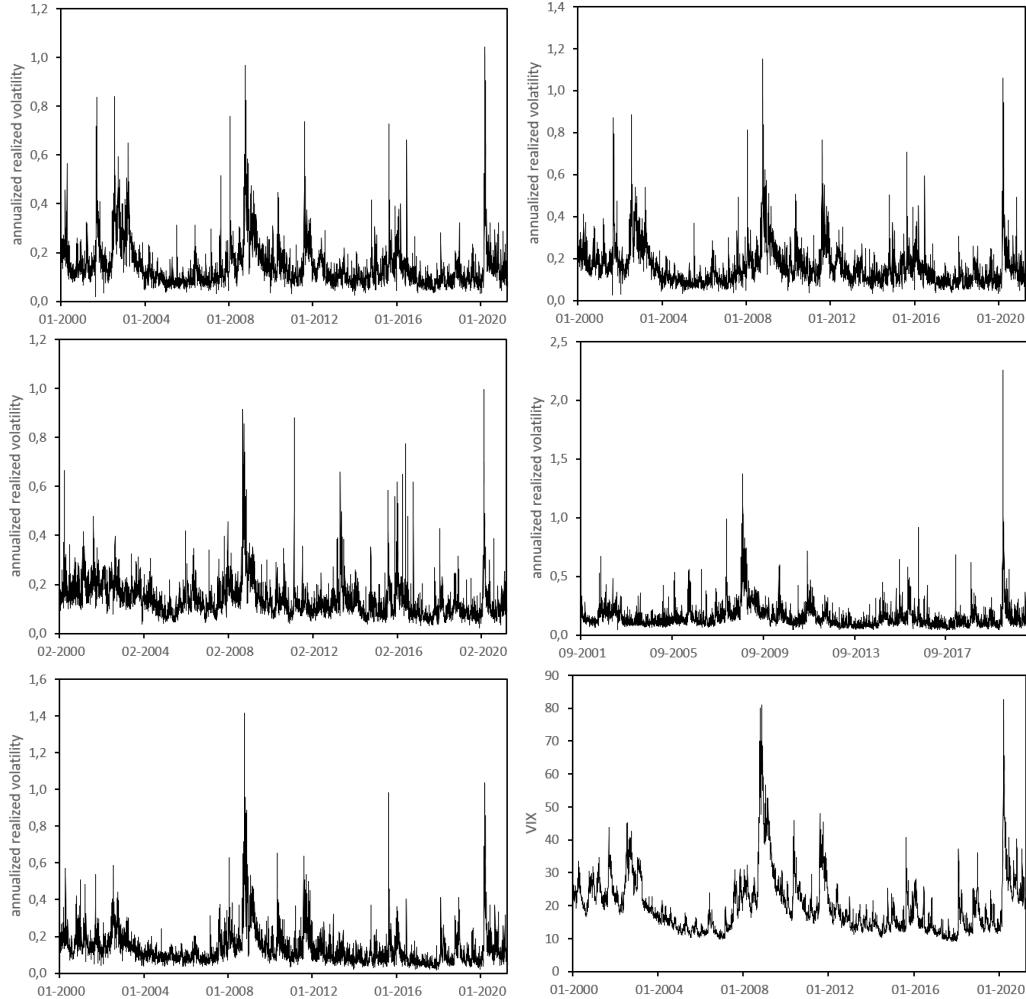

Figure 9: Daily time series of realized volatility for AEX (top left graph), CAC 40 (top right), N225 (middle left), OSEAX (middle right), SPX (bottom left), followed by the VIX (bottom right).

For forecasting applications, we're interested in the dependence structure of a volatility series. We therefore discretize each increment of the volatility process, depending on whether it is positive or negative, resulting in an indicator equal to 1 or 0 [15, 16]. We then consider a vector of three consecutive indicators, which has thus eight possible categories. We consider a window size n, corresponding to one year of data, for estimating the categorical probabilities. We will also consider a smaller n, with a three-month period. In the latter case, in order not to have zero probabilities, we reduce the number of categories to six, merging the two categories corresponding

22

to a trend, that is $(0,0,0)'$ and $(1,1,1)'$, and merging those defined by a strict alternation of 0s and 1s, that is $(0,1,0)'$ and $(1,0,1)'$.

Depending on the series and the threshold selected, the relative entropy between two probabilities built with one year of data indicates a few change-points, as one can see in Figure 10. A clear change-point appears during the global financial crisis (around 2008-2009) for the VIX. Change-points are less obvious for realized volatilities in this period, even though the relative entropy of two consecutive years seems to be significant for SPX and N225, but not if the reference year is the first year. Other important change-points include the interval 2014-2015 for OSEAX. This can be explained by the Norwegian economy's dependence on oil prices: the price of Brent crude fell by 50% in less than a year starting in June 2014, which reshaped both the index's sectoral composition and the serial dependence pattern of its volatility. The peaks in relative entropy for the N225 starting in March 2011 correspond to the major event of Fukushima accident. Interestingly, this appears as a single isolated peak in the realized volatility series, whereas the change-point indicator remains elevated for several months, more realistically reflecting the prolonged uncertainty following the event.

Figure 11 displays the results for three-month windows for the realized volatility of SPX and VIX. Relative entropy is much more fluctuating. For the two series, a change-point is detected during the global financial crisis. The COVID-19 crisis (around 2020) also generates a change-point, whereas it was not obvious with one-year windows.

6   Conclusion

In this paper, we have been interested in the distribution of relative entropy, in the one- and two-sample cases. We have thus presented the asymptotic distribution along with Berry-Esseen bounds. For finite samples, we also have provided concentration bounds. All these approximations of the true distribution are useful for building change-point statistical tests. We have thus described a method based on relative entropy and compared it to more classical approaches thanks to extensive simulations. It highlights the limitations of moments-based tests in some situations, where modifications in the probability distribution do not make the first moments deviate, thus making the test based on relative entropy more general. Two applications to real data, namely climate and finance datasets, emphasize the practical relevance of this method. It makes the climate change visible and unveils modifications in the serial dependence of volatility that we can relate to macroeconomic events. Our theoretical findings rely on recent advances, including results on the multivariate Berry-Esseen inequality and on reverse Pinsker's inequality. These tools do not yet exhibit the same level of theoretical maturity as classical results such as the univariate Berry-Esseen inequality or Pinsker's inequality. Further advances on these two inequalities, for example regarding a non-uniform version of multivariate Berry-Esseen bounds, would directly strengthen our theoretical bounds and enhance the accuracy of our change-point application.

References

[1] E. Abi Jaber and N. De Carvalho. Reconciling rough volatility with jumps. *SIAM journal on financial mathematics*, 15(3):785–823, 2024.

[2] R. Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE transactions on information theory*, 66(10):6297–6302, 2020.
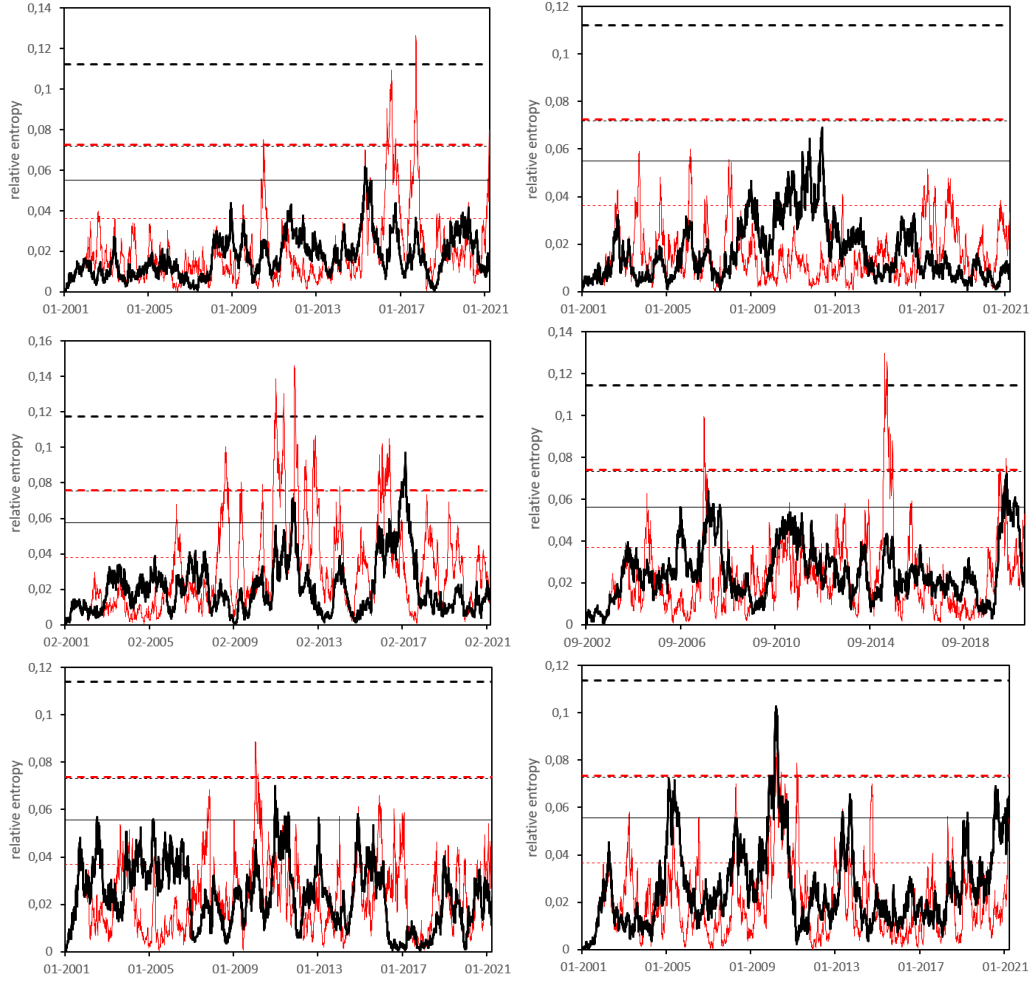
Figure 10: For each daily series of realized volatility for AEX (top left graph), CAC 40 (top right), N225 (middle left), OSEAX (middle right), SPX (bottom left), followed by the VIX (bottom right), relative entropy between empirical discrete probabilities estimated on a year of data finishing at the date indicated in abscissa and on either the previous year (red curve) or the first year of data (black curve). Dashed horizontal lines are confidence bounds at 99%, obtained using the one- and two-sample asymptotic distributions (thin and thick red), as well as $\mathcal{M}^3_{k,n}$ and $\widetilde{\mathcal{M}}^{3,\star}_{k,n,n}$ (thin and thick black). The solid horizontal line corresponds to the value $2(k-1)/n$, where n is the number of observations each year.

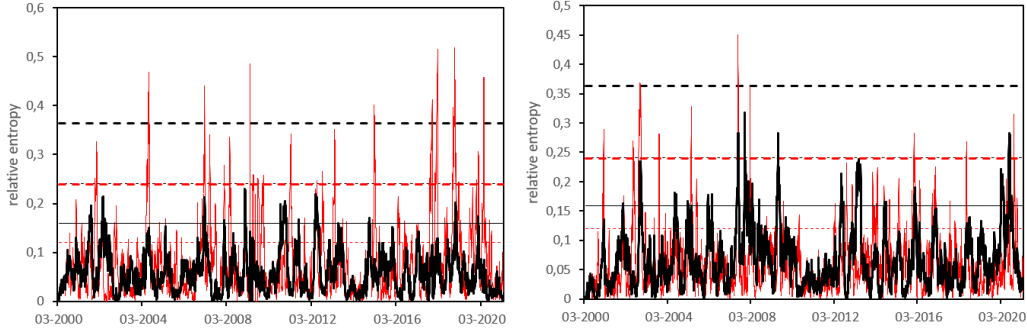Figure 11: For the daily series of the realized volatility of SPX (left graph) and the series of VIX (right graph) relative entropy between empirical discrete probabilities estimated on three months of data finishing at the date indicated in abscissa and on either the previous year (red curve) or the first year of data (black curve). Dashed horizontal lines are confidence bounds at 99%, obtained using the one- and two-sample asymptotic distributions (thin and thick red), as well as $\mathcal{M}^3_{k,n}$ and $\widetilde{\mathcal{M}}^{3,\star}_{k,n,n}$ (thin and thick black). The solid horizontal line corresponds to the value $2(k-1)/n$, where n is the number of observations in the three-month window.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[4] H. Alexandersson. A homogeneity test applied to precipitation data. *Journal of climatology*, 6(6):661–675, 1986.

[5] E. Andreou and E. Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of applied econometrics*, 17(5):579–600, 2002.

[6] A. Barbiero and A. Hitaj. A new discrete exponential distribution: properties and applications. *Journal of statistical theory and practice*, 19(3):39, 2025.

[7] G.P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of probability & its applications*, 4(3):333–336, 1959.

[8] F. Bavaud. Information theory, relative entropy and statistics. In *Formal theories of information: From Shannon to semantic information theory and general concepts of information*, pages 54–78. Springer, 2009.

[9] V. Bentkus and F. Götze. Uniform rates of convergence in the CLT for quadratic forms in multidimensional spaces. *Probability theory and related fields*, 109:367–416, 1997.

[10] M. Bibinger. Cusum tests for changes in the Hurst exponent and volatility of fractional Brownian motion. *Statistics & probability letters*, 161:108725, 2020.

[11] M. Bibinger, M. Jirak, and M. Vetter. Nonparametric change-point analysis of volatility. *Annals of statistics*, 45(4):1542–1578, 2017.

[12] M. Bibinger, J. Yu, and C. Zhang. Modeling and forecasting realized volatility with multivariate fractional Brownian motion. *Preprint*, 2025.

[13] R. Blender and K. Fraedrich. Long time memory in global warming simulations. *Geophysical research letters*, 30(14):1769, 2003.

[14] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

[15] X. Brouty and M. Garcin. A statistical test of market efficiency based on information theory. *Quantitative finance*, 23(6):1003–1018, 2023.

[16] X. Brouty and M. Garcin. Fractal properties, information theory, and market efficiency. *Chaos, solitons & fractals*, 180:114543, 2024.

[17] C.L. Canonne. A short note on an inequality between KL and TV. *Preprint*, 2022.

[18] J.E. Cavanaugh. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & probability letters*, 33(2):201–208, 1997.

[19] J.E. Cavanaugh and A.A. Neath. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley interdisciplinary reviews: computational statistics*, 11(3):e1460, 2019.

[20] R. Chattamvelli and R. Shanmugam. Geometric distribution. In *Discrete distributions in engineering and the applied sciences*, pages 65–82. Springer, 2020.

[21] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & sons, second edition, 2006.

[22] I. Csiszár. The method of types. *IEEE transactions on information theory*, 44(6):2505–2523, 2002.

[23] M. De Pooter and D. Van Dijk. Testing for changes in volatility in heteroskedastic time series – a further examination. Technical report, 2004.

[24] S. Eguchi and J. Copas. Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of multivariate analysis*, 97(9):2034–2040, 2006.

[25] X. Fang, S.-H. Liu, and Q.-M. Shao. Cramér-type moderate deviation for quadratic forms with a fast rate. *Bernoulli*, 29(3):2466–2491, 2023.

[26] T. Flynn and S. Yoo. Change detection with the kernel cumulative sum algorithm. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 6092–6099. IEEE, 2019.

[27] C. Franzke. Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of climate*, 25(12):4172–4183, 2012.

[28] C. Gallagher, R. Lund, and M. Robbins. Changepoint detection in climate time series with long-term trends. *Journal of climate*, 26(14):4994–5006, 2013.

[29] Z. Gao, X. Xiao, Y.-P. Fang, J. Rao, and H. Mo. A selective review on information criteria in multiple change point detection. *Entropy*, 26(1):50, 2024.

[30] M. Garcin. Forecasting with fractional Brownian motion: a financial perspective. *Quantitative finance*, 22(8):1495–1512, 2022.

[31] M. Garcin. Complexity measure, kernel density estimation, bandwidth selection, and the efficient market hypothesis. In A. Sinha, editor, *Select topics of econophysics*. De Gruyter, 2024.

[32] M. Garcin and M. Grasselli. Long versus short time scales: the rough dilemma and beyond. *Decisions in economics and finance*, 45(1):257–278, 2022.

[33] M. Garcin, J. Klein, and S. Laaribi. Estimation of time-varying kernel densities and chronology of the impact of COVID-19 on financial markets. *Journal of applied statistics*, 51(11):2157–2177, 2024.

[34] M. Garcin, K. Sawaya, and T. Valade. Prediction of linear fractional stable motions using codifference, with application to non-Gaussian rough volatility. *Preprint*, 2025.

[35] J. Gatheral, T. Jaisson, and M. Rosenbaum. Volatility is rough. *Quantitative finance*, 18(6):933–949, 2018.

[36] A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[37] V. Glinskiy, A. Logachov, O. Logachova, H. Rojas, L. Serga, and A. Yambartsev. Asymptotic properties of a statistical estimator of the Jeffreys divergence: the case of discrete distributions. *Mathematics*, 12(21):3319, 2024.

[38] F. Götze and A.N. Tikhomirov. Asymptotic expansions in non-central limit theorems for quadratic forms. *Journal of theoretical probability*, 18(4):757–811, 2005.

[39] F. Götze and V.V. Ulyanov. Asymptotic distribution of χ^2-type statistics. *Preprint*, 2003.

[40] F. Götze and A.Y. Zaitsev. Uniform rates of approximation by short asymptotic expansions in the CLT for quadratic forms. *Journal of mathematical sciences*, 176(2):162–189, 2011.

[41] F. Götze and A.Y. Zaitsev. Explicit rates of approximation in the CLT for quadratic forms. *Annals of probability*, 42(1):354–397, 2014.

[42] A. Hamadouche, A. Kouadri, and A. Bakdi. A modified Kullback divergence for direct fault detection in large scale systems. *Journal of process control*, 59:28–36, 2017.

[43] A. Harvey and V. Oryshchenko. Kernel density estimation for time series data. *International journal of forecasting*, 28(1):3–14, 2012.

[44] K. Haynes, P. Fearnhead, and I.A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and computing*, 27(5):1293–1305, 2017.

[45] C.M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[46] C. Inclan and G.C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American statistical association*, 89(427):913–923, 1994.

[47] B. James, K.L. James, and D. Siegmund. Tests for a change-point. *Biometrika*, 74(1):71–83, 1987.

[48] R.H. Jones and I. Dey. Determining one or more change points. *Chemistry and physics of lipids*, 76(1):1–6, 1995.

[49] M. Kelbert. Survey of distances between the most popular distributions. *Analytics*, 2(1):225–245, 2023.

[50] P. Kokoszka and R. Leipus. Change-point estimation in ARCH models. *Bernoulli*, 6(6):513–539, 2000.

[51] M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

[52] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural networks*, 43:72–83, 2013.

[53] Z.A. Lomnicki and S.K. Zaremba. The asymptotic distributions of estimators of the amount of transmitted information. *Information and control*, 2(3):260–284, 1959.

[54] J. Mardia, J. Jiao, E. Tánczos, R.D. Nowak, and T. Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and inference: a journal of the IMA*, 9(4):813–850, 2020.

[55] M. Marinescu and C. Balcau. On the use of mutual information for testing independence. *Preprint*, 2025.

[56] E.S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[57] I. Pinelis. On the nonuniform Berry–Esseen bound. In *Inequalities and extremal problems in probability and statistics*, pages 103–138. Elsevier, 2017.

[58] I. Pinelis and R. Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic journal of statistics*, 10:1001–1063, 2016.

[59] J. Plasse and N.M. Adams. Multiple changepoint detection in categorical data streams. *Statistics and computing*, 29(5):1109–1125, 2019.

[60] M. Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824–2853, 2019.

[61] J. Reeves, J. Chen, X.L. Wang, R. Lund, and Q.Q. Lu. A review and comparison of change-point detection techniques for climate data. *Journal of applied meteorology and climatology*, 46(6):900–915, 2007.

[62] I. Sason and S. Verdú. Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets. In *2015 IEEE information theory workshop-fall*, pages 214–218. IEEE, 2015.

[63] I. Sason and S. Verdú. f-divergence inequalities. *IEEE transactions on information theory*, 62(11):5973–6006, 2016.

[64] Q.-M. Shao and Z.-S. Zhang. Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

[65] X. Shi, C. Beaulieu, R. Killick, and R. Lund. Changepoint detection: An analysis of the Central England temperature series. *Journal of climate*, 35(19):6329–6342, 2022.

[66] A. Shternshis, P. Mazzarisi, and S. Marmi. Measuring market efficiency: The Shannon entropy of high-frequency financial time series. *Chaos, solitons & fractals*, 162:112403, 2022.

[67] C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

[68] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer science & business media, 2008.

[69] P. Wang and W. Ning. Nonparametric CUSUM change-point detection procedures based on modified empirical likelihood. *Computational statistics*, 40:4991–5021, 2025.

[70] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of mathematical statistics*, 9(1):60–62, 1938.

[71] P. Wu, J.-F. Muzy, and E. Bacry. From rough to multifractal volatility: The log S-fBm model. *Physica A: statistical mechanics and its applications*, 604:127919, 2022.

[72] Y. Yu, O.H. Madrid Padilla, D. Wang, and A. Rinaldo. A note on online change point detection. *Sequential analysis*, 42(4):438–471, 2023.

[73] N. Yuan, M. Ding, Y. Huang, Z. Fu, E. Xoplaki, and J. Luterbacher. On the long-term climate memory in the surface air temperature records over Antarctica: A nonnegligible factor for trend evaluation. *Journal of climate*, 28(15):5922–5934, 2015.

[74] Q. Zhu, X. Diao, and C. Wu. Volatility forecast with the regularity modifications. *Finance research letters*, 58:104008, 2023.

[75] Q. Zhu, X. Diao, and C Wu. A test for change points under the roughness of stochastic volatility: the case of the VIX index. *Applied economics letters*, 32(7):951–959, 2025.

[76] S. Zhu, B. Chen, Z. Chen, and P. Yang. Asymptotically optimal one-and two-sample testing with kernels. *IEEE transactions on information theory*, 67(4):2074–2092, 2021.

[77] C. Zou, G. Yin, L. Feng, and Z. Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of statistics*, 42(3):970–1002, 2014.

[78] A.M. Zubkov. Limit distributions for a statistical estimate of the entropy. *Theory of probability & its applications*, 18(3):611–618, 1974.

A   A useful lemma for proving Theorem 1

Lemma 1. *The real solutions of the cubic equation $ax^3 + x^2 - d = 0$ are*

$$\frac{1}{3a}\left[2\cos\left(\frac{1}{3}\arccos\left(\frac{27a^2d - 2}{2}\right) - \frac{2r\pi}{3}\right) - 1\right], \tag{10}$$

with $r \in \{0, 1, 2\}$, when $d \leq 4/27a^2$, and

$$\sqrt[3]{-\frac{1}{27a^3} + \frac{d}{2a} + \sqrt{\frac{-4d + 27d^2a^2}{108a^4}}} + \sqrt[3]{-\frac{1}{27a^3} + \frac{d}{2a} - \sqrt{\frac{-4d + 27d^2a^2}{108a^4}}} - \frac{1}{3a}$$

when $d > 4/27a^2$. Moreover, when $d > 0$ and $a \to 0$, we get three real roots. One diverges to $+\infty$ if $a < 0$ and to $-\infty$ if $a > 0$. The other two behave like:

$$\begin{cases} x_0 & = & \sqrt{d} - \frac{d}{2}a + \mathcal{O}(a^2) \\ x_1 & = & -\sqrt{d} - \frac{d}{2}a + \mathcal{O}(a^2). \end{cases}$$

Proof. Defining y by $x + 1/3a$, we get the following equation in y: $y^3 + py + q = 0$, where $p = -1/3a^2$ and $q = (2/27a^3) - (d/a)$. The discriminant $-(p/3)^3 - (q/2)^2$ is equal to $(4d - 27d^2a^2)/108a^4$.

29

There are two cases, depending on the value of d. First, if $d \leq 4/27a^2$, the discriminant is non-negative and there are three real roots (given below for the equation in x), among which one is double when the discriminant is zero, after Cardano's formula,

$$\frac{1}{3a}\left[2\cos\left(\frac{1}{3}\arccos\left(\frac{27a^2d-2}{2}\right)-\frac{2r\pi}{3}\right)-1\right],$$

where $r \in \{0,1,2\}$. If $d > 4/27a^2$, the discriminant is non-negative and there is only one real root for the equation in x,

$$\sqrt[3]{-\frac{q}{2}+\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}+\sqrt[3]{-\frac{q}{2}-\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}-\frac{1}{3a},$$

that is, when replacing p and q by their expression in a and d,

$$\sqrt[3]{-\frac{1}{27a^3}+\frac{d}{2a}+\sqrt{\frac{-4d+27d^2a^2}{108a^4}}}+\sqrt[3]{-\frac{1}{27a^3}+\frac{d}{2a}-\sqrt{\frac{-4d+27d^2a^2}{108a^4}}}-\frac{1}{3a}.$$

When $a \to 0$, the leading term in the discriminant is $d/27a^4$, which is positive if $d > 0$. Therefore, the solutions are like in equation (10). We easily get the limit

$$\lim_{a\to 0}2\cos\left(\frac{1}{3}\arccos\left(\frac{27a^2d-2}{2}\right)-\frac{2r\pi}{3}\right)-1=\begin{cases}0 & \text{if } r \in \{0,1\}\\-3 & \text{if } r = 2.\end{cases}$$

In this asymptotic case, the root with $r = 2$ will diverge to $+\infty$ if $a < 0$ and to $-\infty$ if $a > 0$. For the two other roots, we apply a perturbative expansion. We start with the simplified problem corresponding to $a = 0$, that is $x^2 - d = 0$. The two solutions are $\sqrt{d}$ and $-\sqrt{d}$. Next, we consider the perturbation of the two roots:

$$\begin{cases}x_0 & = & \sqrt{d}+\beta_0 a+\mathcal{O}(a^2)\\x_1 & = & -\sqrt{d}+\beta_1 a+\mathcal{O}(a^2).\end{cases}$$

Plugging x_0 in the cubic equation, we get

$$ad^{3/2}+2d^{1/2}\beta_0 a=\mathcal{O}(a^2),$$

so that $\beta_0 = -d/2$. Using the same approach for x_1, we find as well that $\beta_1 = -d/2$. We therefore get the result displayed in the lemma. $\qquad\square$

B Proof of Theorem 1

The proof of Theorem 1 uses Lemma 1, which is provided in Appendix A.

Proof. The observations X_j being independent of each other, the random vector $n\widehat{p}_n$ is a multinomial variable of parameter p. The multivariate central limit theorem thus gives

$$\sqrt{n}\left(\widehat{p}_n-p\right)\xrightarrow{\text{d}}\mathcal{N}\left(0,\text{diag}(p)-pp^t\right),$$

where $\mathcal{N}(\mu,\sigma^2)$ is the Gaussian distribution of mean μ and variance σ^2. As a consequence,

$$\sqrt{n}\begin{pmatrix}\frac{\widehat{p}_{n,1}-p_1}{\sqrt{p_1}}\\\vdots\\\frac{\widehat{p}_{n,k}-p_k}{\sqrt{p_k}}\end{pmatrix}\xrightarrow{\text{d}}\mathcal{N}\left(0,\Gamma_k\right),\tag{11}$$

where $\Gamma_k = I_k - uu^t$, with I_k the identity matrix and $u = (\sqrt{p_1}, ..., \sqrt{p_k})^t$. We now need the eigenspaces of Γ_k. Since Γ_k is real and symmetric, its eigenvectors form an orthogonal basis. First, u is an eigenvector associated with the eigenvalue 0:

$$\Gamma_k u = u - uu^t u = (1 - u^t u)u = \left(1 - \sum_{i=1}^{k} p_i\right) u = 0.$$

Let's consider any vector v orthogonal to u, that is $u^t v = 0$. The space of all possible vectors v is thus of dimension $k - 1$. Then, from the orthogonality condition, we easily get $\Gamma_k v = v$. Therefore, the eigenspace of Γ_k associated with the eigenvalue 1 is of dimension $k - 1$, which is also the rank of the matrix Γ_k. The nonzero eigenvalues of Γ_k being all equal to 1, we have the reduced spectral decomposition $\Gamma_k = V I_{k-1} V^t = VV^t$, where $V \in \mathbb{R}^{k \times (k-1)}$ is a matrix whose columns are orthonormal unit-eigenvalue eigenvectors, so that $V^t V = I_{k-1}$.

Let us now focus on the empirical relative entropy, $f(\widehat{p}_n)$, written as a function of $\widehat{p}_n$, where $f(q) = D_{\mathrm{KL}}(q\|p)$. The function f can be decomposed in a sum of k univariate functions: $f(q) = \sum_{i=1}^{k} f_i(q_i)$, where $f_i(q_i) = q_i \log(q_i/p_i)$, $f_i'(q_i) = 1 + \log(q_i/p_i)$, and $f_i^{(j)}(q_i) = (-1)^j (j - 2)! q_i^{1-j}$ for $j \geq 2$. This decomposition eases the second-order Taylor expansion of f, which can be seen as the sum of k distinct expansions, and where we also note that $f_i(p_i) = 0$ and $f_i'(p_i) = 1$:

$$f(\widehat{p}_n) = \frac{1}{2} \sum_{i=1}^{k} \frac{(\widehat{p}_{n,i} - p_i)^2}{p_i} + \sum_{i=1}^{k} R_i(\widehat{p}_{n,i}), \tag{12}$$

with $R_i(\widehat{p}_{n,i}) = f_i^{(3)}(\xi_{n,i})(\widehat{p}_{n,i} - p_i)^3/6 = (p_i - \widehat{p}_{n,i})^3/6\xi_{n,i}^2$, where $\xi_{n,i}$ is in the interval delimited by $\widehat{p}_{n,i}$ and p_i.

From the decomposition $\Gamma_k = VV^t$, we can write the limit appearing in formula (11) as VG, where $G \in \mathbb{R}^{k-1}$ is a standard Gaussian vector. Then,

$$\sum_{i=1}^{k} \frac{n(\widehat{p}_{n,i} - p_i)^2}{p_i} \xrightarrow{\mathrm{d}} G^t V^t V G = G^t G,$$

using the orthonormal property of V. This limit follows a chi-square distribution χ_{k-1}^2. Moreover, since $\widehat{p}_{n,i} \xrightarrow{\mathbb{P}} p_i$, we have $\xi_{n,i} \xrightarrow{\mathbb{P}} p_i$ and, by the continuous mapping theorem, $(p_i - \widehat{p}_{n,i})/6\xi_{n,i}^2 \xrightarrow{\mathbb{P}} 0$, because $p_i \neq 0$. Therefore, $nR_i(\widehat{p}_{n,i}) = n(p_i - \widehat{p}_{n,i})^2(p_i - \widehat{p}_{n,i})/6\xi_{n,i}^2 \xrightarrow{\mathbb{P}} 0$, since it is a product of a sequence converging in distribution with a sequence converging in probability to zero. Finally, starting from equation (12), we see that $2nf(\widehat{p}_n)$ is a sum of a sequence converging in distribution to χ_{k-1}^2 with k sequences converging in probability to zero. Slutsky's theorem thus leads to equation (2).

We now have to prove the second part of the theorem. We're going to use a multivariate extension of Berry-Esseen theorem [60] and apply it first to the quadratic part of f: the rest, namely the R_i functions, will be considered in a second stage. We define $Y_1, ..., Y_n$, iid vectors of $\mathbb{R}^k$, by

$$Y_j = \left(\frac{\mathbb{1}_{X_j \in \Omega_1} - p_1}{\sqrt{np_1}}, ..., \frac{\mathbb{1}_{X_j \in \Omega_k} - p_k}{\sqrt{np_k}}\right)^t.$$

It is easy to see that $\mathbb{E}(Y_j) = 0$. We will need the expression of the third moment of the Euclidean

norm of Y_j:

$$
\begin{aligned}
\mathbb{E}\left[\left(Y_j^t Y_j\right)^{3/2}\right] & = \mathbb{E}\left[\left(\sum_{i=1}^k \frac{(\mathbb{1}_{X_j \in \Omega_i} - p_i)^2}{n p_i}\right)^{3/2}\right] \\
& = \sum_{u=1}^k \mathbb{P}(X_j \in \Omega_u)\left(\frac{(1-p_u)^2}{n p_u} + \sum_{i \neq u} \frac{p_i^2}{n p_i}\right)^{3/2} \quad (13)\\
& = \frac{1}{n^{3/2}} \sum_{u=1}^k p_u\left(\frac{(1-p_u)^2}{p_u} + 1 - p_u\right)^{3/2} \\
& = \frac{1}{n^{3/2}} \sum_{u=1}^k \frac{(1-p_u)^{3/2}}{p_u^{1/2}}.
\end{aligned}
$$

We also define $W = \sum_{j=1}^n Y_j$, whose i-th component, for $i \in [\![1,k]\!]$, is $(\widehat{p}_{n,i} - p_i)(n/p_i)^{1/2}$, and

$$
\Theta = W^t W = \sum_{i=1}^k \frac{n(\widehat{p}_{n,i} - p_i)^2}{p_i}.
$$

The properties of multinomial variables indicate that the covariance matrix of the vector W is $\Gamma_k = V V^t$, whose rank is $k-1$ as explained above. Therefore, we can decompose W in an orthonormal basis of dimension $k-1$: we define a vector $U \in \mathbb{R}^{k-1}$ as $V^t W$, so that $W = VU$ and $\Theta = U^t U$. Then, it appears that $U = \sum_{j=1}^n T_j$, where $T_j = V^t Y_j$. By linearity, $\mathbb{E}(T_j) = 0$ and, using the independence of Y_j and Y_ℓ for $\ell \neq j$, we have as well

$$
\begin{aligned}
\sum_{j=1}^n \mathbb{E}(T_j T_j^t) & = \sum_{j=1}^n V^t \mathbb{E}(Y_j Y_j^t) V \\
& = V^t \mathbb{E}(W W^t) V \\
& = V^t \Gamma_k V \\
& = V^t V V^t V = I_{k-1}.
\end{aligned}
$$

These conditions on U and T_j are the ones that are required for Raič's multivariate Berry-Esseen theorem [60, Theorem 1.1]. For a proper application of this theorem, two other assumptions are still to be verified. First, because $Y_j = V T_j$, Y_j and T_j have the same Euclidean norm, $Y_j^t Y_j = T_j^t V^t V T_j = T_j^t T_j$, so that $\mathbb{E}\left[\left(T_j^t T_j\right)^{3/2}\right] = \sum_{u=1}^k (1-p_u)^{3/2}/n^{3/2} p_u^{1/2}$ after equation (13). Second, the set $\mathcal{A}_{k-1,x} = \{Z \in \mathbb{R}^{k-1} | Z^t Z \leq x\}$ is convex because it is the sublevel set of a convex function. We now have all the conditions for applying Raič's theorem [60, Theorem 1.1]:

$$
\begin{aligned}
\left|\mathbb{P}\left(\Theta \leq x\right) - F_{\chi_{k-1}^2}(x)\right| & = \left|\mathbb{P}\left(U \in \mathcal{A}_{k-1,x}\right) - \mathbb{P}\left(G \in \mathcal{A}_{k-1,x}\right)\right| \\
& \leq \left(42(k-1)^{1/4} + 16\right) \sum_{j=1}^n \mathbb{E}\left[\left(T_j^t T_j\right)^{3/2}\right] \quad (14)\\
& \leq \left(42(k-1)^{1/4} + 16\right) \sum_{u=1}^k \frac{(1-p_u)^{3/2}}{(n p_u)^{1/2}},
\end{aligned}
$$

where $G \in \mathbb{R}^{k-1}$ is a standard Gaussian vector.

We now adapt the above Berry-Esseen inequality to include the residual $\mathcal{R} = 2n \sum_{i=1}^k R_i(\widehat{p}_{n,i})$ of $2n f(\widehat{p}_n)$, given by the Taylor expansion displayed in equation 12. If $\widehat{p}_{n,i} \geq p_i$, then $\xi_{n,i} \geq p_i$ and we simply have $|R_i(\widehat{p}_{n,i})| \leq |p_i - \widehat{p}_{n,i}|^3/6 p_i^2$. If $\widehat{p}_{n,i} < p_i$, we cannot properly bound the residual of the second-order Taylor expansion with the same method and we need a more precise analysis. The non-truncated Taylor series provides, for $x \in [0, p]$,

$$
R_i(x) = (x - p_i)^3 \sum_{j=3}^\infty \frac{(-1)^j (x - p_i)^{j-3}}{j(j-1) p_i^{j-1}}.
$$

Noting $h_i(x) = R_i(x)/(x-p_i)^3$, we have

$$
h_i'(x) = \sum_{j=3}^\infty \frac{(p_i - x)^{j-2}(j-3)}{j(j-1) p_i^{j-1}},
$$

32

which is non-negative whatever $x \le p_i$. Since we also have $h_i(p_i) = -1/6p_i^2$ and $h_i(0) = -1/2p_i^2$, we finally get $\max_{x \in [0,p_i]} |h_i(x)| = 1/2p_i^2$. Combining this result with our analysis for $x \ge p_i$, we find that $1/2p_i^2$ is an upper bound for $|h(x)|$ whatever $x \in [0,1]$. Then, noting that L^q norms are non-increasing functions of q, and writing $\mu = \min_{i \in [\![1,k]\!]} p_i$, we get

$$|\mathcal{R}| \le 2n \sum_{i=1}^{k} \frac{|p_i - \widehat{p}_{n,i}|^3}{2p_i^2} \le \frac{n}{\sqrt{\mu}} \left(\sum_{i=1}^{k} \left| \frac{p_i - \widehat{p}_{n,i}}{\sqrt{p_i}} \right|^2 \right)^{3/2} = \frac{\Theta^{3/2}}{\sqrt{\mu n}}.$$

Since $2nf(\widehat{p}_n) = \Theta + \mathcal{R}$ and $(\Theta \le x - \rho) \Rightarrow (\Theta + \mathcal{R} \le x) \Rightarrow (\Theta \le x + \rho)$ for $\rho \ge |\mathcal{R}|$, we have

$$\mathbb{P}\left(K_{n,k}^{\mathrm{down}}(\Theta) \le x \right) \le \mathbb{P}\left(2nf(\widehat{p}_n) \le x \right) \le \mathbb{P}\left(K_{n,k}^{\mathrm{up}}(\Theta) \le x \right), \tag{15}$$

where $K_{n,k}^{\mathrm{up}} : z \ge 0 \mapsto z - (\mu n)^{-1/2} z^{3/2}$ and $K_{n,k}^{\mathrm{down}} : z \ge 0 \mapsto z + (\mu n)^{-1/2} z^{3/2}$. Also, because of the convexity of f [21, Theorem 2.7.2], we have the convexity of the set $\{q | 2nf(q) \le x\}$. So we can refine formula (15) by taking into account a constraint of convexity, that is the three sets in this formula have to be convex. Noting that $K_{n,k}^{\mathrm{up}}(0) = K_{n,k}^{\mathrm{down}}(0) = 0 < x$ and that $(\Theta = 0) \Rightarrow (\forall i, \widehat{p}_{n,i} = p_i) \Rightarrow (2nf(\widehat{p}_n) = 0 < x)$, 0 belongs to the three intervals of admissible values of Θ and the convexity constraint modifies inequalities (15) in

$$\mathbb{P}\left(\Theta \in [0, \kappa_{n,k}^{\mathrm{down}}(x)] \right) \le \mathbb{P}\left(2nf(\widehat{p}_n) \le x \right) \le \mathbb{P}\left(\Theta \in [0, \kappa_{n,k}^{\mathrm{up}}(x)] \right),$$

where $\kappa_{n,k}^{\mathrm{up}}(x)$ (respectively $\kappa_{n,k}^{\mathrm{down}}(x)$) is the smallest positive root of $z \mapsto K_{n,k}^{\mathrm{up}}(z) - x$ (resp. $z \mapsto K_{n,k}^{\mathrm{down}}(z) - x$) if it exists, otherwise the probabilities are trivially equal to 1. An explicit expression for $\kappa_{n,k}^{\mathrm{up}}(x)$ and $\kappa_{n,k}^{\mathrm{down}}(x)$ is obtained as the solution of a cubic equation, solved with Cardano's formula. Indeed, using Lemma 1 applied to $\sqrt{z}$, with $d = x$ and $a = -(\mu n)^{-1/2}$ for $\kappa_{n,k}^{\mathrm{up}}(x)$ or $a = (\mu n)^{-1/2}$ for $\kappa_{n,k}^{\mathrm{down}}(x)$, we directly obtain the expression displayed in Theorem 1. We also have $(\mu n)^{-1/2} \to 0$ and Lemma 1 provides us with the asymptotic behaviour of $\kappa_{n,k}^{\mathrm{up}}(x)$, namely

$$\sqrt{\kappa_{n,k}^{\mathrm{up}}(x)} = \sqrt{x} + \frac{x}{2\sqrt{\mu n}} + \mathcal{O}\left(\frac{1}{n} \right),$$

which gives

$$\kappa_{n,k}^{\mathrm{up}}(x) = x + \frac{x^{3/2}}{\sqrt{\mu n}} + \mathcal{O}\left(\frac{1}{n} \right).$$

A similar idea makes it possible to lead to the result displayed in Theorem 1 for $\kappa_{n,k}^{\mathrm{down}}(x)$.

Finally, we can use Raič's theorem again, using directly formula (14) with a new value for x:

$$
\begin{aligned}
\mathbb{P}\left(2nf(\widehat{p}_n) \le x \right) &\le \mathbb{P}\left(\Theta \le \kappa_{n,k}^{\mathrm{up}}(x) \right) \\
&\le F_{\chi_{k-1}^2}(\kappa_{n,k}^{\mathrm{up}}(x)) + \left| \mathbb{P}\left(\Theta \le \kappa_{n,k}^{\mathrm{up}}(x) \right) - F_{\chi_{k-1}^2}(\kappa_{n,k}^{\mathrm{up}}(x)) \right| \\
&\le F_{\chi_{k-1}^2}(\kappa_{n,k}^{\mathrm{up}}(x)) + \left(42(k-1)^{1/4} + 16 \right) \sum_{u=1}^{k} \frac{(1-p_u)^{3/2}}{(np_u)^{1/2}}.
\end{aligned}
$$

We similarly get

$$
\begin{aligned}
\mathbb{P}\left(2nf(\widehat{p}_n) \le x \right) &\ge \mathbb{P}\left(\Theta \le \kappa_{n,k}^{\mathrm{down}}(x) \right) \\
&\ge F_{\chi_{k-1}^2}(\kappa_{n,k}^{\mathrm{down}}(x)) - \left| \mathbb{P}\left(\Theta \le \kappa_{n,k}^{\mathrm{down}}(x) \right) - F_{\chi_{k-1}^2}(\kappa_{n,k}^{\mathrm{up}}(x)) \right|
\end{aligned}
$$

and we can now conclude with the statement of the theorem. $\qquad \square$

C   Proof of Theorem 2 and Proposition 1

We start by the proof of Theorem 2.

Proof. As seen in the proof of Theorem 1, the vector whose i-th component, for $i \in [\![1,k]\!]$, is $\sqrt{n}(\widehat{p}_{n,i} - p_i)/\sqrt{p_i}$ converges toward a Gaussian vector G_1 of covariance matrix Γ_k. We have the same result when considering $\sqrt{m}(\widehat{q}_{m,i} - p_i)/\sqrt{p_i}$, with the convergence toward the Gaussian vector G_2 of covariance matrix Γ_k and independent of G_1. Therefore, we get, for $n, m \to \infty$,

$$\sqrt{\frac{nm}{n+m}} \begin{pmatrix} \frac{\widehat{p}_{n,1} - \widehat{q}_{m,1}}{\sqrt{p_1}} \\ \vdots \\ \frac{\widehat{p}_{n,k} - \widehat{q}_{m,k}}{\sqrt{p_k}} \end{pmatrix} = \sqrt{n}\sqrt{\frac{m}{n+m}} \begin{pmatrix} \frac{\widehat{p}_{n,1} - p_1}{\sqrt{p_1}} \\ \vdots \\ \frac{\widehat{p}_{n,k} - p_k}{\sqrt{p_k}} \end{pmatrix} - \sqrt{m}\sqrt{\frac{n}{n+m}} \begin{pmatrix} \frac{\widehat{q}_{m,1} - p_1}{\sqrt{p_1}} \\ \vdots \\ \frac{\widehat{q}_{m,k} - p_k}{\sqrt{p_k}} \end{pmatrix}$$

$$\xrightarrow{d} \sqrt{1-\lambda}G_1 - \sqrt{\lambda}G_2,$$

which, by independence, is a Gaussian vector of covariance matrix Γ_k.

The successive derivatives of the relative entropy are $\partial_{p_i} D_{KL}(p\|q) = 1 + \log(p_i/q_i)$, $\partial_{q_i} D_{KL}(p\|q) = -p_i/q_i$, $\partial^2_{p_i p_i} D_{KL}(p\|q) = 1/p_i$, $\partial^2_{q_i q_i} D_{KL}(p\|q) = p_i/q_i^2$, and $\partial^2_{p_i q_i} D_{KL}(p\|q) = -1/q_i$. Therefore, noting $\delta_{n,i} = \widehat{p}_{n,i} - p_i$ and $\gamma_{m,i} = \widehat{q}_{m,i} - p_i$, a second-order Taylor expansion gives

$$\begin{aligned} D_{KL}(\widehat{p}_n\|\widehat{q}_m) &= \sum_{i=1}^k \left[\delta_{n,i} - \gamma_{m,i} + \frac{\delta_{n,i}^2}{2p_i} + \frac{\gamma_{m,i}^2}{2p_i} - \frac{\delta_{n,i}\gamma_{m,i}}{p_i} + o\left(\delta_{n,i}^2 + \gamma_{m,i}^2\right) \right] \\ &= \sum_{i=1}^k \frac{(\delta_{n,i} - \gamma_{m,i})^2}{2p_i} + o\left(\delta_{n,i}^2 + \gamma_{m,i}^2\right) \\ &= \sum_{i=1}^k \frac{(\widehat{p}_{n,i} - \widehat{q}_{m,i})^2}{2p_i} + o\left(\delta_{n,i}^2 + \gamma_{m,i}^2\right). \end{aligned}$$

The leading term of this expansion is proportional to the Euclidean norm of the vector whose limit it the above Gaussian of covariance Γ_k. The framework is so exactly the same as in the proof of Theorem 1 and we can conclude that

$$2\frac{nm}{n+m} D_{KL}(\widehat{p}_n\|\widehat{q}_m) \xrightarrow{d} \chi^2_{k-1}.$$

$\square$

We now prove Proposition 1.

Proof. We define, for $j \in [\![1, n+m]\!]$,

$$Y_j = \left(\frac{\mathbb{1}_{X_j \in \Omega_1} - p_1}{\sqrt{(n+m)p_1}}, ..., \frac{\mathbb{1}_{X_j \in \Omega_k} - p_k}{\sqrt{(n+m)p_k}} \right)^t$$

and

$$\widetilde{Y}_j = \begin{cases} \sqrt{\frac{m}{n}} Y_j & \text{if } j \in [\![1, n]\!] \\ -\sqrt{\frac{n}{m}} Y_j & \text{if } j \in [\![n+1, n+m]\!]. \end{cases}$$

Then, $\widetilde{Y}_1, ..., \widetilde{Y}_{n+m}$ are independent of each other and such that, noting $\widetilde{W} = \sum_{j=1}^{n+m} \widetilde{Y}_j$, we have $\mathbb{E}(\widetilde{Y}_j) = 0$ and, after Theorem 2, $\sum_{j=1}^{n+m} \mathbb{E}(\widetilde{Y}_j \widetilde{Y}_j^t) = \mathbb{E}(\widetilde{W}\widetilde{W}^t) = \Gamma_k$, with the same $\Gamma_k = VV^t$ as in the proof of Theorem 1. Therefore, we can use Raič's theorem again [60, Theorem 1.1], applied to $\widetilde{U} = V^t\widetilde{W}$ and $\widetilde{T}_j = V^t\widetilde{Y}_j$, $\widetilde{T}_j$ having the same Euclidean norm as $\widetilde{Y}_j$. Noting

$$\widetilde{\Theta} = \widetilde{W}^t\widetilde{W} = \frac{nm}{n+m} \sum_{i=1}^k \frac{(\widehat{p}_i - \widehat{q}_i)^2}{p_i},$$

we can thus write, like in the proof of Theorem 1,

$$\left| \mathbb{P}\left(\widetilde{\Theta} \le x \right) - F_{\chi^2_{k-1}}(x) \right| \le \left(42(k-1)^{1/4} + 16 \right) \sum_{j=1}^{n+m} \mathbb{E}\left[\left(\widetilde{Y}_j^t \widetilde{Y}_j \right)^{3/2} \right]. \tag{16}$$

We know, from equation (13) the expression of $\mathbb{E}\left[\left(Y_j^t Y_j \right)^{3/2} \right]$, from which we deduce that

$$\mathbb{E}\left[\left(\widetilde{Y}_j^t \widetilde{Y}_j \right)^{3/2} \right] = \begin{cases} \frac{m^{3/2}}{n^{3/2}(n+m)^{3/2}} \sum_{u=1}^{k} \frac{(1-p_u)^{3/2}}{p_u^{1/2}} & \text{if } j \in [\![1,n]\!] \\ \frac{n^{3/2}}{m^{3/2}(n+m)^{3/2}} \sum_{u=1}^{k} \frac{(1-p_u)^{3/2}}{p_u^{1/2}} & \text{if } j \in [\![n+1, n+m]\!]. \end{cases}$$

Therefore

$$\begin{aligned} \sum_{j=1}^{n+m} \mathbb{E}\left[\left(\widetilde{Y}_j^t \widetilde{Y}_j \right)^{3/2} \right] &= \left(\frac{m^{3/2}}{n^{1/2}(n+m)^{3/2}} + \frac{n^{3/2}}{m^{1/2}(n+m)^{3/2}} \right) \sum_{u=1}^{k} \frac{(1-p_u)^{3/2}}{p_u^{1/2}} \\ &= \frac{m^2 + n^2}{(nm)^{1/2}(n+m)^{3/2}} \sum_{u=1}^{k} \frac{(1-p_u)^{3/2}}{p_u^{1/2}} \end{aligned}$$

and, combining it with equation (16), we finally obtain the result displayed in Proposition 1. □

D Proof of Proposition 2

Proof. We know that the moment-generating function of $D_{\mathrm{KL}}\left(\widehat{p}_n \| p \right)$ is upper bounded by

$$M(t) = \left(\sum_{j=0}^{n} \frac{n!}{n^{2j}(n-j)!} t^j \right)^{k-1},$$

for $t \in [0, n)$ [2, Proposition II.2, Lemmas II.4 and II.5]. So, using Chernoff inequality, we get the first bound displayed in Proposition 2, $\mathcal{M}_{k,n}^1(x)$. The last bound, $\mathcal{M}_{k,n}^3(x)$, is Agrawal's bound, which relies on the fact that $M(t) \le M_{\mathrm{Agrawal}}(t) = (1 - t/n)^{-k+1}$ and on the optimisation in t of the Chernoff bound $\inf_{t \in [0,n)} e^{-tx} M_{\mathrm{Agrawal}}(t)$, the minimum being reached for t equal to $t^\star = n - (k-1)/x \in [0, n)$ [2, Theorem I.2]. Using this same value $t^\star$ in $M(t)$, we get the middle bound, $\mathcal{M}_{k,n}^2(x)$, which is higher than $\mathcal{M}_{k,n}^1(x)$ because $t^\star \in [0,n)$, and lower than $\mathcal{M}_{k,n}^3(x)$ because $M(t) \le M_{\mathrm{Agrawal}}(t)$ for all $t \in [0, n)$, including $t^\star$. □

E Proofs for two-sample concentration inequalities

E.1 Proof of Proposition 3

Proof. Noting $\|x\|_1 = \sum_{i=1}^{k} |x_i|$, we have, by triangular inequality $\|x - y\|_1 \le \|x\|_1 + \|y\|_1$, so, combining this with Young's inequality, we get

$$\|x - y\|_1^2 \le 2\|x\|_1^2 + 2\|y\|_1^2. \tag{17}$$

Using successively the right part of formula (8), formula (17), and the left part of formula (8), we get

$$\begin{aligned} D_{\mathrm{KL}}\left(p \| q \right) &\le \left(\frac{1}{2\min_i q_i} - \min_i \frac{p_i}{2q_i} \right) \|p - q\|_1^2 \\ &\le \left(\frac{1}{\min_i q_i} - \min_i \frac{p_i}{q_i} \right) \left(\|p - r\|_1^2 + \|q - r\|_1^2 \right) \\ &\le \left(\frac{2}{\min_i q_i} - 2\min_i \frac{p_i}{q_i} \right) \left(D_{\mathrm{KL}}\left(p \| r \right) + D_{\mathrm{KL}}\left(q \| r \right) \right). \end{aligned}$$

□

E.2 Proof of Theorem 3

Proof. Using Proposition 3 along with the independence of $\widehat{p}_n$ and $\widehat{q}_m$, the moment-generating function of the two-sample relative entropy follows, for $t \geq 0$:

$$\mathbb{E}\left[\exp\left(tD_{\mathrm{KL}}\left(\widehat{p}_n \| \widehat{q}_m\right)\right)\right] \leq \mathbb{E}\left[\exp\left(t\beta_{m,n}D_{\mathrm{KL}}\left(\widehat{p}_n \| p\right)\right]\mathbb{E}\left[t\beta_{m,n}D_{\mathrm{KL}}\left(\widehat{q}_m \| p\right)\right)\right].$$

According to Agrawal's inequality, we thus have, for $t \in [0, \min(m,n)/\beta_{m,n})$ [2, Theorem 1.3]:

$$\mathbb{E}\left[\exp\left(tD_{\mathrm{KL}}\left(\widehat{p}_n \| \widehat{q}_m\right)\right)\right] \leq \left(\frac{1}{(1 - t\beta_{m,n}/m)(1 - t\beta_{m,n}/n)}\right)^{k-1}. \tag{18}$$

Therefore, by a substitution $s = t\beta_{m,n}$ and Chernoff inequality, we obtain, for $x > 0$, $\mathbb{P}\left(D_{\mathrm{KL}}\left(\widehat{p}_n \| \widehat{q}_n\right) \geq x\right) \leq \inf_{s \in [0, \min(m,n))} \exp(f(s))$, with

$$f(s) = -\frac{sx}{\beta_{m,n}} - (k-1)\log\left(1 - \frac{s}{m}\right) - (k-1)\log\left(1 - \frac{s}{n}\right).$$

We have $f(0) = 0$ and $\lim_{s \to \min(m,n)} f(s) = +\infty$. The derivative of f is

$$f'(s) = -\frac{x}{\beta_{m,n}} - (k-1)\left(\frac{1}{s-m} + \frac{1}{s-n}\right).$$

Noting $\lambda = x/\beta_{m,n}(k-1)$, s is a zero of f' iff

$$\lambda s^2 + (2 - \lambda(n+m))s + \lambda mn - m - n = 0,$$

equation whose discriminant is $4 + \lambda^2(m-n)^2 > 0$, leading to the two possible roots

$$\begin{cases} s^\star &= \frac{\lambda(n+m) - 2 - \sqrt{4 + \lambda^2(m-n)^2}}{2\lambda} \\ s^\bullet &= \frac{\lambda(n+m) - 2 + \sqrt{4 + \lambda^2(m-n)^2}}{2\lambda}. \end{cases}$$

By symmetry, one can assume $m \leq n$. The condition $s^\star \in [0, \min(m,n))$ gives, for the upper bound,

$$\lambda(n+m) - 2 - \sqrt{4 + \lambda^2(m-n)^2} < 2\lambda m,$$

that is

$$\lambda(n-m) - 2 < \sqrt{4 + \lambda^2(m-n)^2}. \tag{19}$$

When $\lambda(n-m) < 2$, formula (19) always holds. If $\lambda(n-m) \geq 2$, then, after considering the square, formula (19) simplifies to $-4\lambda(n-m) < 0$, which always holds since $\lambda(n-m) \geq 2$. Now, the lower constraint $s^\star \geq 0$ leads to

$$\lambda(m+n) - 2 \geq \sqrt{4 + \lambda^2(m-n)^2},$$

that is $\lambda(m+n) \geq 2$ and $\lambda \geq (m+n)/mn$. A direct calculation also shows that $(m+n)/mn > 2/(m+n)$. On the other hand, the root $s^\bullet$ never falls in the interval $[0, \min(m,n))$. Indeed, still assuming $m \leq n$, the condition $s^\bullet < m$ requires

$$\lambda(n-m) - 2 < -\sqrt{4 + \lambda^2(m-n)^2}, \tag{20}$$

which is never verified because, by the elementary inequality, $2(4 + \lambda^2(m-n)^2) \geq (2 + \lambda|m-n|)^2$, so that condition (20) writes

$$\lambda(n-m) - 2 < -\frac{1}{\sqrt{2}}(2 + \lambda|m-n|)$$

and $\lambda(n - m) < \sqrt{2}(2 - \sqrt{2})/(1 + \sqrt{2}) < 0$, which contradicts the assumptions.

Therefore, when $\lambda \geq \max(2/(m+n), (m+n)/mn) = (m+n)/mn$, then $f'(0) \leq 0$ and the minimum of f in the interval $[0, \min(m, n))$ is reached in $s^\star$. If $\lambda < (m + n)/mn$, then $f'(0) > 0$, with no root of f' in $[0, \min(m, n))$, so that the minimum of f in this interval is reached in 0. This provides us with the bound $\widetilde{\mathcal{M}}^3_{k,n,m}(x)$, in which $\sigma_{m,n,x}$ is simply equal to $s^\star$.

Like in Proposition 2, one can also modify formula (18) to use the true moment-generating function, so we get $\widetilde{\mathcal{M}}^1_{k,n,m}(x)$ and $\widetilde{\mathcal{M}}^2_{k,n,m}(x)$. $\qquad\square$