On the sampling distribution of an ℓ^2 distance between Empirical Distribution Functions with applications to nonparametric testing

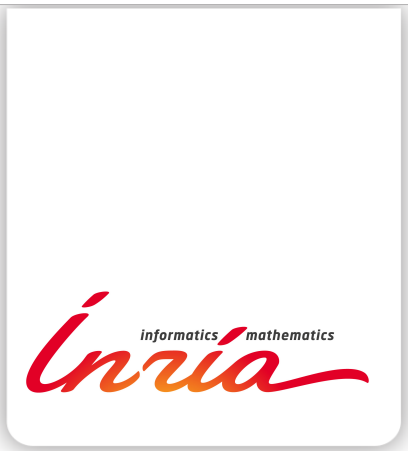
Francois Caron, Chris Holmes, Emmanuel Rio

On the sampling distribution of an ℓ^2 distance between Empirical Distribution Functions with applications to nonparametric testing

François Caron, Chris Holmes, Emmanuel Rio

RESEARCH REPORT
N° 7931
April 16, 2012
Project-Team ALEA

ISSN 0249-6399 ISRN INRIA/RR--7931--FR+ENG



On the sampling distribution of an ℓ^2 distance between Empirical Distribution Functions with applications to nonparametric testing

François Caron*, Chris Holmes†, Emmanuel Rio‡

Project-Team ALEA

Research Report n° 7931 — version 2 — initial version April 16, 2012 — revised version November 22, 2012 — 15 pages

Abstract: We consider a situation where two sample sets of independent real valued observations are obtained from unknown distributions. Under a null hypothesis that the distributions are equal, it is well known that the sample variation of the infinity norm, maximum, distance between the two empirical distribution functions has as asymptotic density of standard form independent of the unknown distribution. This result underpins the popular two-sample Kolmogorov-Smirnov test. In this article we show that other distance metrics exist for which the asymptotic sampling distribution is also available in standard form. In particular we describe a weighted squared-distance metric derived from a binary recursion of the real line which is shown to follow a sum of chi-squared random variables. This motivates a nonparametric test based on the average divergence rather than the maximum, which we demonstrate exhibits greater sensitivity to changes in scale and tail characteristics when the distributions are unequal, while maintaining power for changes in central location.

Key-words: two-sample test, nonparametric test, binary tree

* INRIA, Institut de Mathématiques de Bordeaux, University of Bordeaux, France
† Department of Statistics, Oxford University
‡ UMR 8100 CNRS, Université de Versailles Saint-Quentin en Yvelines, Laboratoire de Mathématiques de Versailles

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération

Bâtiment A 29

33405 Talence Cedex

Sur la distribution d'une distance ℓ^2 entre fonctions de distributions empiriques, avec application au test non paramétrique d'adéquation entre les distributions de deux échantillons.

Résumé : On considère la situation où un ensemble de deux échantillons $\{X_1, \ldots, X_{n^{(1)}}\}$, $\{Y_1, \ldots, Y_{n^{(2)}}\}$, d'observations réelles indépendantes est obtenu à partir de distributions inconnues $\{F_X, F_Y\}$, avec les fonctions de répartition empiriques $\{\hat{F}_X, \hat{F}_Y\}$. Dans ce cas, il est connu que sous l'hypothèse nulle $F_X \equiv F_Y$ la distance maximum ℓ^∞ $||\hat{F}_X - \hat{F}_Y||_\infty$, a une distribution asymptotique de forme standard indépendantes de F. Ce résultat forme la base du test d'adéquation de Kolmogorov-Smirnov. Dans cet article, nous montrons que d'autres distances entre fonctions de répartitions existent pour lesquelles la distribution d'échantillonnage asymptotique est connue. En particulier, nous proposons une distance ℓ^2 pondérée dérivée d'une récursion binaire de $\mathbb{R}$ qui suit asymptotiquement la même distribution qu'une somme de variables aléatoires χ^2. Ce résultat suggère un test non paramétrique basé sur la divergence moyenne plutôt que le maximum pour lequel nous montrons une plus grande sensibilité à des changements de variance et de queues de distribution, tout en conservant des résultats similaires pour détecter des changements du paramètre de location.

Mots-clés : test d'adéquation non paramétrique, récursion binaire

1 Introduction

The sampling distribution of the empirical distribution function defined as $\hat{F}_n(x) = k/n$ for real-valued $x_{(k)} < x < x_{(k+1)}$, where $x_{(i)}$ denotes the realised ith order statistic, is an important quantity in statistics in part as it allows for nonparametric hypothesis testing. For instance when two sets of samples have been obtained under different conditions $\{X_1, \ldots, X_{n^{(1)}}\}$, $\{Y_1, \ldots, Y_{n^{(2)}}\}$ it is well known from the work of Kolmogorov (1933) that for a particular ℓ^∞ norm the distance $d_\infty(\hat{F}_X, \hat{F}_Y) = ||\hat{F}_X - \hat{F}_Y||_\infty = \max_{x\in\mathbb{R}} |\hat{F}_X(x) - \hat{F}_Y(x)|$ has an asymptotic distribution of standard form independent of F_X, F_Y under the null hypothesis $X_i, Y_i \sim F_X$. This result underpins the popular two-sample Kolmogorov-Smirnov test. In this paper we explore the sample distribution of other distance metrics $d(\hat{F}_X, \hat{F}_Y)$. In particular we derive the distribution of an ℓ^2 distance calculated from a recursive binary partition of the real line $\mathbb{R}$. We show that this statistic is independent of F_X, F_Y, and follows a weighted sum of χ^2 under the null hypothesis $X_i, Y_i \sim F_X$. This motivates a nonparametric test based on the weighted average divergence $d_w(\hat{F}_X, \hat{F}_Y)$ rather than the maximum. We demonstrate numerically that this test appears to show greater sensitivity in detecting departures in the scale or tails of F_X, F_Y, while maintaining power to detect shifts in central location.

The rest of the paper is as follows. In Section 2 we outline our partitioning scheme of $\mathbb{R}$ and derive the asymptotic sampling distribution of the ℓ^2 distance under the null hypothesis. In Section 3 we provide some illustrations that the test can be more powerful than other conventional non-parametric test. In Section 4 we derive a similar test for k-sample test. Finally in Section 5 we provide a discussion of our work.

2 An ℓ^2 distance for two-sample test

Suppose one has two sets of samples, $\{X, Y\}$ of sample size $\{n^{(1)}, n^{(2)}\}$ obtained say under different treatments. We write $n = n^{(1)} + n^{(2)}$. Let $\hat{F}_0(x)$ be the empirical distribution of the joint sample $\{X, Y\}$.

We shall first consider a dyadic (binary) tree that recursively partitions $\mathbb{R}$ into disjoint measurable sets such that at the mth level of the tree we find $\mathbb{R} = \cup_{j=0}^{2^m-1} B_j^{(m)}$ where $B_i^{(m)} \cap B_j^{(m)} = \varnothing$ for all $i \neq j$. The jth junction in the tree at level i has associated set $B_j^{(i)}$ and clearly $B_j^{(i)} = (B_{2j}^{(i+1)}, B_{2j+1}^{(i+1)})$ for all i, j. It will be convenient in what follows to simply index the sets using base 2 subscript and drop the superscript so that, for example, B_{000} indicates the first set in level 3, B_{0011} the forth set in level 4 and so on. Such a recursive binary tree is represented in Figure 1. Let Π denote the partition structure defined by the collection of sets $\Pi = (B_0, B_1, B_{00}, \ldots)$ and in particular consider a partition centered on the quantiles of $\hat{F}_0$. That is, at level m,

$$B_j = [\hat{F}_0^{-1}(j^*/2^m), \hat{F}_0^{-1}([j^* + 1]/2^m))], \tag{1}$$

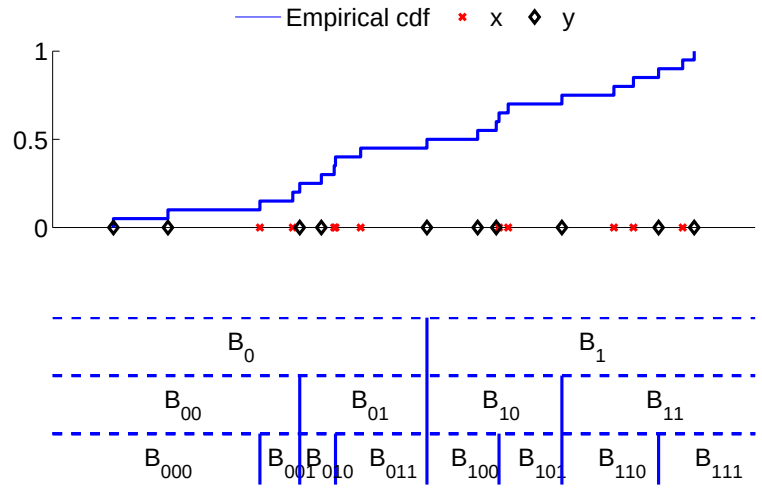


Figure 1: Recursive binary tree constructed from the empirical cdf.

where j^* is the decimal representation of the binary number j and m is the level of the tree.

We write resp. $n_j^{(1)}$ and $n_j^{(2)}$ the number of observations in X and Y that lie in the subset B_j of Π. We note $n_j^{(12)} = n_j^{(1)} + n_j^{(2)}$. Hence $n_0^{(1)}$ is the number of items from X in the first 50% percentiles of $\hat{F}_0$, $n_{00}^{(1)}$ is the number of items from X in the first 25% percentiles, etc. We also have $n_j^{(k)} = n_{j0}^{(k)} + n_{j1}^{(k)}$, for $k = 1, 2, 12$, where 12 denotes the combined sample set.

Consider the level j with given $n_j^{(12)}, n_j^{(1)}$ and $n_{j0}^{(12)}$. Under H_0, all the permutations of the $n_j^{(12)}$ are equiprobable, and $n_{j0}^{(1)}$ is distributed from the hypergeometric distribution $n_{j0}^{(1)} \sim \text{HypGeo}(n_j^{(12)}, n_{j0}^{(12)}, n_j^{(1)})$ whose probability function is given by

$$\Pr(n_{j0}^{(1)} = k | \Pi, n_j^{(1)}, n_j^{(12)}, n_{j0}^{(12)}, H_0) = \frac{\binom{n_j^{(1)}}{k}\binom{n_j^{(12)} - n_j^{(1)}}{n_{j0}^{(12)} - k}}{\binom{n_j^{(12)}}{n_{j0}^{(12)}}} \tag{2}$$

if $\max(0, n_j^{(1)} - n_j^{(12)} + n_{j0}^{(12)}) \leq k \leq \min(n_{j0}^{(12)}, n_j^{(1)})$ and 0 otherwise. $\binom{n}{p}$ is the usual binomial coefficient. The hypergeometric distribution is obtained when sampling without replacement $n_{j0}^{(12)}$ balls from an urn containing $n_j^{(1)}$ red balls and $n_j^{(12)} - n_j^{(1)}$ white balls. The conditional

moments are then given by

$$\mu_j = \mathbb{E}[n_{j0}^{(1)}|n_j^{(1)}] = n_j^{(1)}\frac{n_{j0}^{(12)}}{n_j^{(12)}}$$
$$\sigma_j^2 = \mathbb{E}\left[\left(n_{j0}^{(1)} - \mu_j\right)^2 |n_j^{(1)}\right]$$
$$= \frac{n_j^{(1)}(n_j^{(12)} - n_j^{(1)})n_{j0}^{(12)}(n_j^{(12)} - n_{j0}^{(12)})}{\left(n_j^{(12)}\right)^2(n_j^{(12)} - 1)}$$

Definition 1 *Let T be the weighted ℓ^2 distance defined by*

$$T = \sum_j \lambda_j z_j \tag{3}$$

where $z_j = \left(n_{j0}^{(1)} - \mu_j\right)^2$ and the sum goes over all elements j such that $n_j^{(12)} > 0$. Let $\log_2(a) = \log(a)/\log 2$ and $l(j)$ be the level of the tree associated to the binary number j. For all j with $l(j) \leq [\log_2(n)] - 1$, the weights λ_j are defined by

$$\lambda_j = \frac{n^{(12)}(n^{(12)} - 1)}{2^{l(j)}p_j(1-p_j)n_j^{(12)}n^{(1)}n^{(2)}} \tag{4}$$

where $p_j = n_{j0}^{(12)}/n_j^{(12)}$, $l(j)$ is the level of section j, and $\lambda_j = 0$ for $l(j) \geq [\log_2(n)]$.

The value z_j measures the departure from the expected value at each level j, and there are 2^m such contributions at each level m of the binary tree. The realisation of T can then be used as a test statistic to quantify departures from the null hypothesis.

We now provide a result on the asymptotic distribution of the test statistic T under the null.

Proposition 1 *Let*

$$W = \sum_j 2^{-l(j)}Y_j^2$$

where $Y_j \sim \mathcal{N}(0,1)$ for all j. Then

$$\inf\{\|T - V\|_2 : V \stackrel{Law}{=} W\} \leq c_1(n^{(1)}n^{(2)})^{-1}n^{3/2}\log_2(n), \tag{5}$$

where c_1 is some positive constant. It provides the rate of approximation $\mathcal{O}(n^{-1/2}\log n)$ with respect to the minimal L^2-distance as n tends to infinity, and consequently the rate of approximation $\mathcal{O}(n^{-1/3}(\log n)^{2/3})$ with respect to the Prokhorov distance. The proof is given in Appendix A.1.

Hence we can use T to test for the hypothesis $F_X \equiv F_Y$. Note we do not have the asymptotic distribution of the test statistic T under the alternative. Nonetheless, the next theorem provides a lower bound for T under H_1.

Proposition 2 *Assume that the ratio n_1/n converges to $\alpha \in]0,1[$ when n tends to ∞. Then under the alternative $F_X \neq F_Y$, we have*

$$\liminf_{n\to\infty} n^{-1}T \geq c_2 \frac{\alpha}{1-\alpha} \text{ almost surely} \tag{6}$$

where $c_2 > 0$ is some strictly positive constant. The proof is given in Appendix A.2

3 Simulations

In Section 2 we showed how the test statistic T can be used to measure departures from the null hypothesis that the underlying distribution functions are the same. To examine the operating performance of the method we consider the following experiments designed to explore various canonical departures from the null.

a) Mean shift: $Y^{(1)} \sim \mathcal{N}(0,1)$, $Y^{(2)} \sim \mathcal{N}(\theta,1)$, $\theta = 0,\ldots,3$

b) Variance shift: $Y^{(1)} \sim \mathcal{N}(0,1)$, $Y^{(2)} \sim \mathcal{N}(0,\theta^2)$, $\theta = 1,\ldots,3$

c) Mixture: $Y^{(1)} \sim \mathcal{N}(0,1)$, $Y^{(2)} \sim \frac{1}{2}\mathcal{N}(\theta,1) + \frac{1}{2}\mathcal{N}(-\theta,1)$, $\theta = 0,\ldots,3$

d) Tails: $Y^{(1)} \sim \mathcal{N}(0,1)$, $Y^{(2)} \sim t(\theta^{-1})$, $\theta = 10^{-3},\ldots,10$

e) Skew: $Y^{(1)} \sim \mathcal{N}(0,1)$, $Y^{(2)} \sim S\mathcal{N}(0,1,\theta)$, $\theta = 1,\ldots,10$

f) Lognormal mean shift: $\log Y^{(1)} \sim \mathcal{N}(0,1)$, $\log Y^{(2)} \sim \mathcal{N}(\theta,1)$, $\theta = 0,\ldots,3$

g) Lognormal variance shift: $\log Y^{(1)} \sim \mathcal{N}(0,1)$, $\log Y^{(2)} \sim \mathcal{N}(0,\theta^2)$, $\theta = 1,\ldots,3$

where $S\mathcal{N}(0,1,\lambda)$ is the skew normal distribution of skewness parameter λ. Comparisons are performed with $n^{(1)} = n^{(2)} = 50$ against the two-sample Kolmogorov-Smirnov and Wilcoxon rank test. To compare the models we explore the "power to detect the alternative". We evaluate numerically the value θ such that the statistical power is at least 80% for each of the three tests. The results, obtained from 10^6 independent simulations, are reported in Table 1. Globally, our approach gives similar results when detecting a shift in the median, but outperforms other methods for detecting shifts in variance or tails.

4 Extension to multiple sample test

Our partitioning approach can be readily extended to deal with multiple tests for multiple treatments or conditions. Consider now that we are given p samples $y^{(1)} \underset{i.i.d.}{\sim} F^{(1)}$, $y^{(2)} \underset{i.i.d.}{\sim} F^{(2)},\ldots,y^{(p)} \underset{i.i.d.}{\sim} F^{(p)}$, $F^{(1)},\ldots,F^{(p)}$ unknown, and we wish to test the following hypothesis

$$H_0 : F^{(1)} = F^{(2)} = \ldots = F^{(p)}$$

	Tree Test	K-S	Wilcoxon
Gaussian: mean	0.67	0.67	**0.58**
Gaussian: variance	**2.19**	2.73	—
Gaussian: mixture	**1.42**	1.64	—
Gaussian: skewness	1.13	1.14	**0.94**
Gaussian:tail	**2.02**	2.97	—
Lognormal: mean	0.69	0.68	**0.58**
Lognormal: variance	**2.20**	2.78	—

Table 1: Value θ needed to achieve 80% power for the binary tree, the Kolmogorov-Smirnov and the Wilcoxon two-sample tests. Lower values indicate better performances.

against the alternative that at least one of the distributions is different from the others. We consider the binary tree $\Pi = (B_0, B_1, B_{00}, \ldots)$ constructed from the empirical distribution of $y = (y^{(1)}, \ldots, y^{(k)})$. We note $n_j^{(k)}$, $j = \{\}, 0, 1, \ldots$, $k = 1, \ldots, p$ the number of items from $y^{(k)}$ in interval B_j, and $n_j = \sum_{k=1}^{p} n_j^{(k)}$. Assuming $n_j^{(k)} > 0$ for all k, then the vector

$$z_{j0} = (n_{j0}^{(1)}, n_{j0}^{(2)}, \ldots, n_{j0}^{(p-1)})$$

follows the multivariate hypergeometric distribution $\text{MvHypGeo}((n_j^{(1)}, n_j^{(2)}, \ldots, n_j^{(p)}), n_{j0})$ whose pdf is given by

$$\frac{\Pi_{k=1}^{p} \binom{n_j^{(k)}}{n_{j0}^{(k)}}}{\binom{n_j}{n_{j0}}}$$

This distribution is obtained when sampling without replacement from an urn with $n_j^{(k)}$ balls of color k, $k = 1, \ldots, p$. The two first moments μ_j and Σ_j are defined by

$$\mu_j = \mathbb{E}\left[z_{j0}|z_j\right] = \frac{n_{j0}^{(12)}}{n_j^{(12)}} \left(n_j^{(1)}, n_j^{(2)}, \ldots, n_j^{(p-1)}\right)$$

$$\Sigma_j(k,l) = cov(n_{j0}^{(k)}, n_{j0}^{(l)}|z_j) = -\frac{n_{j0} n_j^{(k)} n_j^{(l)}}{n_j^2} \frac{n_j - n_{j0}}{n_j - 1}$$

$$\Sigma_j(k,k) = var(n_{j0}^{(k)}|z_j) = \frac{n_j^{(k)}}{n_j} \left(1 - \frac{n_j^{(k)}}{n_j}\right) n_j \frac{n_j - n_{j0}}{n_j - 1}$$

and we can show that z_{j0} is asymptotically multivariate normal Fraser (1956) and so

$$\begin{aligned} S_j &= (z_{j0} - \mu_j)^T \Sigma_j^{-1} (z_{j0} - \mu_j) \\ &= \frac{n_j - 1}{n_j - n_{j0}} \sum_{k=1}^{p} \frac{\left(n_{j0}^{(k)} - n_j^{(k)} \frac{n_{j0}}{n_j}\right)^2}{n_j^{(k)} \frac{n_{j0}}{n_j}} \end{aligned}$$

	Tree Test	K-S	Kruskal-Wallis
Gaussian: mean	0.67	0.63	**0.57**
Gaussian: variance	**2.25**	3	—
Gaussian: mixture	**1.44**	1.70	—
Gaussian: skewness	1.01	1.05	**0.83**
Gaussian:tail	**2.02**	3.16	—
Lognormal: mean	0.67	0.63	**0.56**
Lognormal: variance	**2.27**	3.12	—

Table 2: Value θ needed to achieve 80% power for the binary tree, the Kolmogorov-Smirnov and the Kruskal-Wallis four-sample tests. Lower values indicate better performances.

is asymptotically $\mathcal{X}^2(p-1)$ distributed. We extend the test statistic T introduced in the above section by considering the similar test statistic T_2 defined by

$$T_2 = \sum_{j|\#\{k|n_j^{(k)}>0\}\geq 2} 2^{-l(j)} S_j \tag{7}$$

The same experiments as in Section 3 are performed. We consider a 4-sample test where we define $F_3 = F_4 = F_1$. Our test is compared to the Kolmogorov-Smirnov test with pairwise comparisons, where the threshold value is set so that to have overall size $\alpha = 0.05$. We also compare the test to the Kruskall-Wallis test. Results are given in Table 2. Similar to the two-sample test, our test performs similarly to detect a shift in the median but outperforms other methods to detect a shift in the variance or tails.

5 Discussion

We have derived the sampling distribution for a weighted ℓ^2 distance between empirical sampling distributions drawn from the same probability law. This allowed us to derive a simple nonparametric test statistic for departures from the null hypothesis of no treatment effect. This method offers an alternative nonparametric procedure to the usual Kolmogorov-Smirnov two sample test. The Kolmogorov-Smirnov test is known to have relatively low power for against alternative which differ in scale (Capon, 1965; Klotz, 1967). Simulations conducted in Section 3 and 4 show that the new binary tree test has similar power against alternatives which differ in location while largely outperforming Kolmogorov-Smirnov test in scale/tail alternatives.

A Proofs

A.1 Upper bounds for the minimal L^2-distance of T to the limiting law under H_0

In this subsection we prove Proposition 1. The first step is to approximate the test statistic T by a weighted sum of independent $\chi^2(1)$-distributed random variables. This will be done using the Tusnády's type lemma proved in Castelle and Laurent-Bonvalot (1998), which we now recall.

Lemma 1 *(Lemma 2.5 in Castelle and Laurent-Bonvalot (1998)) - Let Y be a standard normal random variable and Φ be the distribution function of Y. Let G_{n,n_1,n_2} be the distribution function of the hypergeometric distribution $\mathcal{H}(n,n_1,n_2)$. We denote by m the mean of $\mathcal{H}(n,n_1,n_2)$ (recall $m=n_1n_2/n$). Let $p=n_1/n$, $p'=n_2/n$ and $\delta=2p-1$, $\delta'=2p'-1$. Set*

$$\sigma=\sqrt{npp'(1-p)(1-p')}. \tag{8}$$

Then, for each positive η, there exists positive constants c and d such that, if $|\delta\delta'|\leq 1-\eta$, then

$$|G^{-1}_{n,n_1,n_2}\circ\Phi(Y)-m-\sigma Y|\leq c+dY^2. \tag{9}$$

Starting from Lemma 1, we now approximate the statistic T in L^2 by a weighted sum of squares of standard normal random variables. To achieve this approximation, we construct a random variable with the same distribution as T from a family of independent standard normal random variables indexed by the binary tree, in the same way as in Komlós et al. (1975). So let $\mathcal{B}$ denote the binary tree and let $(Y_j)_{j\in\mathcal{B}}$ be a collection of independent standard normal random variables. The random variables $(n_j^{(12)})_{j\in\mathcal{B}}$ are given (and the collection of Gaussian r.v.'s is independent of this collection). Suppose that the random variables $n_j^{(1)}$, with the adequate multinomial distribution at each level (and consequently $n_j^{(2)}$)), have already been defined up to level k from the Gaussian random variables $(Y_j)_{l(j)<k}$. We want to define the random variables at the level $k+1$. By definition (1),

$$n_{j0}^{(12)}=[n_j^{(12)}/2]. \tag{10}$$

At level $k+1$, we set

$$n_{j0}^{(1)}=G^{-1}_{n_j^{12},n_{j0}^{(12)},n_j^{(1)}}\circ\Phi(Y_j)$$

Let m_j denote the mean of $n_{j0}^{(1)}$,

$$p_j=n_{j0}^{(12)}/n_j^{(12)} \quad\text{and}\quad p'_j=n_j^{(1)}/n_j^{(12)}.$$

Set

$$s_j=\sqrt{n_j^{(12)}p_j(1-p_j)p'_j(1-p'_j)}.$$

By Eq. (10), $\delta_j = 2p_j - 1 = 0$ if $n_j^{(12)}$ is even, and

$$|\delta_j| = |2p_j - 1| = 1/n_j^{(12)}$$

otherwise. It follows that $|\delta_j| \leq 1/3$, provided that $n_j^{(12)} \geq 2$. In that case, Lemma 1 applies with $\eta = 1/3$, and

$$|n_{j0}^{(1)} - m_j - s_j Y_j| \leq c + dY_j^2. \tag{11}$$

Let us now denote by $n = n^{(1)} + n^{(2)}$ the global size of the sample, and write n in basis 2, that is $n = a_1 a_2 \cdots a_l$ with $a_1 = 1$. We may without loss of generality assume that $n^{(1)} \leq n^{(2)}$. At the level zero,

$$n = n_1^{(12)} = a_1 a_2 \cdots a_l \text{ and, at level } l-2,\ n_j^{(12)} \geq 2.$$

Hence we will stop the construction at level $l-2$ (note that $[\log_2(n)] = l-1$), so that all the numbers $n_j^{(12)}$ satisfy $n_j^{(12)} \geq 2$. The test statistic T is defined by

$$T = \sum_{k=0}^{l-2} T'_k, \text{ where } T'_k = \sum_{\{j:l(j)=k\}} \lambda_j |n_{j0}^{(1)} - m_j|^2$$

for positive weights λ_j to be specified later.

Let

$$T'' = \sum_{k=0}^{l-2} T''_k, \text{ where } T''_k = \sum_{\{j:l(j)=k\}} \lambda_j \sigma_j^2 Y_j^2.$$

Clearly

$$\|T - T''\|_2 \leq \sum_{k=0}^{l-2} \|T''_k - T'_k\|_2.$$

Now

$$T''_k - T'_k = \sum_{\{j:l(j)=k\}} \lambda_j (|n_{j0}^{(1)} - m_j|^2 - \sigma_j^2 Y_j^2).$$

Recall that we have to bound up the $\mathbb{L}_2$-norm of this random variable. Now

$$\mathbb{E}(|n_{j0}^{(1)} - m_j|^2 \mid \mathcal{F}_k) = \sigma_j^2.$$

Consequently

$$M_k = \mathbb{E}(T''_k - T'_k \mid \mathcal{F}_k) = 0.$$

Next we bound up the conditional variance of $(T''_k - T'_k)$ given $\mathcal{F}_k$. From our construction the random variables $(|n_{j0}^{(1)} - m_j|^2 - \sigma_j^2 Y_j^2)_j$ are independent at the scale $l(j) = k$ conditionally to the σ-field $\mathcal{F}_k$ generated by the random variables $(n_j^{(1)})$ with $l(j) = k$. Hence the conditional variance is the sum over j of individual conditional variances. By the inequality (11), we have

$$(|n_{j0}^{(1)} - m_j|^2 - \sigma_j^2 Y_j^2)^2 \leq \|(c + dY_j^2 + (\sigma_j - s_j)|Y_j|)^2 (c + dY_j^2 + (\sigma_j + s_j)|Y_j|)^2.$$

Now

$$\sigma_j - s_j = s_j\Big(\sqrt{\frac{n_j^{(12)}}{n_j^{(12)}-1}} - 1\Big) \leq \frac{s_j}{2(n_j^{(12)}-1)} \leq \frac{s_j}{n_j^{(12)}} \leq \frac{1}{4}.$$

Hence the conditional variance of $T_k'' - T_k'$ given $\mathcal{F}_k$, which we denote by V_k, satisfies

$$V_k \leq c_0^2 \sum_{\{j:l(j)=k\}} \lambda_j^2(1+\sigma_j^2),$$

for some positive universal constant c_0. Note that $\sigma_j^2 \leq 2^{-3} n_j^{(12)}$. Also, from the definition of $n_j^{(12)}$,

$$n2^{-l(j)-1} \leq n_j^{(12)} \leq n2^{1-l(j)},$$

which ensures that $\sigma_j^2 \leq n2^{-1-l(j)}$. Since $n2^{-1-l(j)} \geq 1$ for $l(j) \leq l-2$, we get the deterministic upper bound

$$V_k \leq c_0^2 n \sum_{\{j:l(j)=k\}} \lambda_j^2 2^{-k}.$$

From the above bounds and the fact that $\|T_k'' - T_k'\|_2^2 = \mathbb{E}(V_k)$, we now have:

$$\|T - T''\|_2 \leq c_0\sqrt{n}\sum_{k=0}^{l-2}\Big(2^{-k}\sum_{\{j:l(j)=k\}}\lambda_j^2\Big)^{1/2}. \tag{12}$$

Now, let

$$W = \sum_{\{j:l(j)<l-1\}} \lambda_j \mathbb{E}(\sigma_j^2) Y_j^2.$$

Clearly W is a weighted sum of $\chi^2(1)$-distributed independent random variables. We first give an explicit formula for W.

By definition, $n_j^{(1)}$ has the hypergeometric distribution $\mathcal{H}(n, n_1, n_j^{(12)})$. Hence

$$\mathbb{E}(\sigma_j^2) = \frac{n_{j0}^{(12)} n_{j1}^{(12)}}{n_j^{(12)}(n_j^{(12)}-1)} \mathbb{E}(n_j^{(1)} n_j^{(2)}) = \frac{n_1 n_2}{n(n-1)} p_j(1-p_j) n_j^{(12)},$$

which leads to the definition formula

$$W = \frac{n_1 n_2}{n(n-1)} \sum_{\{j:l(j)<l-1\}} \lambda_j p_j(1-p_j) n_j^{(12)} Y_j^2 \tag{13}$$

Now we bound up $\|T'' - W\|_2$. Clearly

$$\|T'' - W\|_2 \leq \sum_{k=0}^{l-2} \| \sum_{\{j:l(j)=k\}} \lambda_j(\sigma_j^2 - \mathbb{E}(\sigma_j^2)) Y_j^2 \|_2.$$

Now

$$\sigma_j^2 - \mathbb{E}(\sigma_j^2) = p_j(1-p_j)\frac{n_j^{(1)} n_j^{(2)} - \mathbb{E}(n_j^{(1)} n_j^{(2)})}{n_j^{(12)} - 1}$$

Recall $n_j^{(2)} = n_j^{(12)} - n_j^{(1)}$. Therefrom

$$\|T'' - W\|_2 \leq \sum_{k=0}^{l-2} (\|A_k\|_2 + \|B_k\|_2),$$

with

$$A_k = \sum_{\{j:l(j)=k\}} \lambda_j \frac{p_j(1-p_j)n_j^{(12)}}{n_j^{(12)}-1} (n_j^{(1)} - \mathbb{E}(n_j^{(1)}) Y_j^2$$

and

$$B_k = \sum_{\{j:l(j)=k\}} \lambda_j \frac{p_j(1-p_j)n_j^{(12)}}{n_j^{(12)}-1} (n_j^{(1)} p'_j - \mathbb{E}(n_j^{(1)} p'_j) Y_j^2.$$

Now $(n_j^{(1)})_{j:l(j)=k}$ depends on the random variables $(Y_k)_{l(k)<l(j)}$, which ensures that this random vector is independent of the Gaussian random vector $(Y_j)_{j:l(j)=k}$. Furthermore, it can be proven that $(n_j^{(1)})_{l(j)=k}$ is a negatively associated random vector. It follows that, for distincts j and j' with $l(j) = k$,

$$\mathrm{cov}\,(n_j^{(1)}, n_{j'}^{(1)}) \leq 0 \;\text{ and }\; \mathrm{cov}\,(n_j^{(1)} p_j, n_{j'}^{(1)} p'_j) \leq 0.$$

Consequently

$$\|A_k\|_2^2 \leq 3 \sum_{\{j:l(j)=k\}} \lambda_j^2 \Big(\frac{p_j(1-p_j)n_j^{(12)}}{n_j^{(12)}-1} \Big)^2 \mathrm{Var}\, n_j^{(1)} \leq \frac{3}{4} \sum_{\{j:l(j)=k\}} \lambda_j^2 \mathrm{Var}\, n_j^{(1)}$$

and, in a similar way

$$\|B_k\|_2^2 \leq \frac{3}{4} \sum_{\{j:l(j)=k\}} \lambda_j^2 \mathrm{Var}\,(n_j^{(1)} p'_j).$$

Now $(p'_j)^2$ is a 2-Lipschitz function of p'_j. Consequently

$$\mathrm{Var}\,((p'_j)^2) \leq 4\mathrm{Var}\,(p'_j)$$

and therefrom $\mathrm{Var}\,(n_j^{(1)} p'_j) \leq 4\mathrm{Var}\, n_j^{(1)}$. It follows that

$$\|T'' - W\|_2 \leq \frac{3}{2}\sqrt{3} \sum_{k=0}^{l-2} \Big(\sum_{\{j:l(j)=k\}} \lambda_j^2 \mathrm{Var}\, n_j^{(1)} \Big)^{1/2}.$$

Since $n_j^{(1)}$ has the hypergeometric distribution $\mathcal{H}(n, n_1, n_j^{(12)}$, we have:

$$\mathrm{Var}\, n_j^{(1)} = \frac{n_1 n_2 n_j^{(12)} (n - n_j^{(12)}}{n^2(n-1)} \leq \frac{n_j^{(12)} n_1}{n} \leq 2^{1-k} n_1 \leq 2^{-k} n,$$

since $n_1 \leq (n/2)$. Hence

$$\|T'' - W\|_2 \leq 3\sqrt{n} \sum_{k=0}^{l-2} \Big(2^{-k} \sum_{\{j:l(j)=k\}} \lambda_j^2 \Big)^{1/2}.$$

Both the above bounds and (12) then imply that

$$\|T - W\|_2 \leq (c_0 + 3)\sqrt{n} \sum_{k=0}^{l-2} \Big(2^{-k} \sum_{j:l(j)=k} \lambda_j^2 \Big)^{1/2}. \tag{14}$$

We now choose the coefficients λ_j in such a way that it will be feasible to approximate the density of W. To this aim, we define λ_j by the equation

$$\lambda_j p_j (1 - p_j) = \frac{n(n-1)}{2^{l(j)} n_j^{(12)} n_1 n_2}$$

For this choice of λ_j, W has the same distribution as

$$W' = \sum_{k=0}^{l-2} 2^{-k} Z_k,$$

where $(Z_k)_{k \geq 0}$ is a sequence of independent random variables with respective laws $\chi^2(2^k)$.

Now p_j belongs to $[1/3, 2/3]$, from the definition of $n_{j0}^{(12)}$. Hence $p_j(1 - p_j) \geq (2/9)$, and consequently

$$\lambda_j \leq 9n/(n_1 n_2).$$

Setting $c_1 = 9(c_0 + 3)$, applying (14) and noticing that $l - 1 \leq \log_2(n) = \log(n)/\log 2$, we get that

$$\|T - W\|_2 \leq c_1 (n_1 n_2)^{-1} n^{3/2} \log_2(n), \tag{15}$$

which provides the rate of approximation $\mathcal{O}(n^{-1/2} \log n)$ with respect to the minimal L^2-distance as n tends to infinity, and consequently the rate of approximation $\mathcal{O}(n^{-1/3} (\log n)^{2/3})$ with respect to the Prokhorov distance.

A.2 Lower bounds for T under H_1

In this subsection, we give some lower bounds for the orders of magnitude of T under H_1. Throughout we assume that n tends to ∞ and (n_1/n) converges to α in $]0, 1[$. Let $F = \alpha F_X + (1 - \alpha) F_Y$. We set

$$G_1 = F_X \circ F^{-1} \quad \text{and} \quad G_2 = F_Y \circ F^{-1}.$$

By the strong law of large numbers, for any fixed j at a fixed level m,

$$\lim_{n \to \infty} n^{-1} (n_{j0}^1 - \mu_j) = \alpha (G_1(2^{-m-1}(2j+1)) - (G_1(2^{-m} j) + G_1(2^{-m}(j+1))/2).$$

Now $\lim_{n \to \infty} p_j = 1/2$. Moreover, it can be proven that

$$\lim_{n \to \infty} n^{-1} 2^{l(j)} n_j^{(12)} = 1.$$

The above facts ensure that the weights λ_j satisfy

$$\lim_{n\to\infty} n\lambda_j = 4/(\alpha-\alpha^2).$$

Consequently, a.s.

$$\lim_{n\to\infty} n^{-1}\lambda_j z_j = \frac{4\alpha}{1-\alpha}(G_1(2^{-m-1}(2j+1)) - (G_1(2^{-m}j) + G_1(2^{-m}(j+1))/2)^2$$

In order to write the statistic T in a different way, we now assume that $F^{(1)}$ and $F^{(2)}$ have a density w.r.t. the Lebesgue measure. Then G_1 has a locally integrable density g_1 on $[0,1]$.

We now recall the definition of the Haar system in $L^2([0,1])$. Let $e_0 = 1$, and, at, for $m \geq 0$, at scale m, for $j = 0, 1, \ldots 2^m - 1$, the functions $\tilde{e}_{m,j}$ be defined by

$$\tilde{e}_{m,j}(x) = 1 \ \text{ for } \ x \in]j2^{-m}, (j+1/2)2^{-m}], \ \tilde{e}_{m,j}(x) = -1 \ \text{ for } \ x \in](j+1/2)2^{-m}, (j+1)2^{-m}],$$

and $\tilde{e}_{m,j}(x) = 0$ otherwise.

With this definition,

$$\lim_{n\to\infty} n^{-1}\lambda_j z_j = \frac{\alpha}{1-\alpha}\Big(\int_0^1 (g_1(u)-1)\tilde{e}_{m,j}(u)du\Big)^2.$$

Let then

$$\gamma_{j,k}(g) = \int_0^1 g(u)\tilde{e}_{m,j}(u)du.$$

Since the Haar system is a total system in $L^2([0,1])$, for any function g with null integral over $[0,1]$,

$$\int_0^1 g^2(u)du = \sum_{m=0}^{\infty}\sum_{j=0}^{2^m-1} 2^m\gamma_{m,j}^2(g)$$

Consequently

$$\sum_{m=0}^{\infty} 2^m \sum_{j=0}^{2^m-1}\Big(\int_0^1 (g_1(u)-1)\tilde{e}_{m,j}(u)du\Big)^2 = \int_0^1 (g_1(u)-1)^2du.$$

Now, under H_1, $G_1(x) \neq x$ on some subset of $[0,1]$ with positive Lebesgue measure, which ensures that

$$I := \int_0^1 (g_1(u)-1)^2du > 0.$$

Hence, for M large enough,

$$A_M := \sum_{m=0}^{M} 2^m \sum_{j=0}^{2^m-1}\Big(\int_0^1 (g_1(u)-1)\tilde{e}_{m,j}(u)du\Big)^2 > 0, \ \text{since} \ \lim_{M\to\infty} A_M = I.$$

The above facts ensure that

$$\liminf_{n\to\infty} n^{-1}T \geq \frac{\alpha}{1-\alpha}2^{-M}A_M > 0 \text{ almost surely} \tag{16}$$

(here M is a fixed integer such that $A_M > 0$ and consequently the term on right hand is positive and does not depend on n).

References

Capon, J. (1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test. *Journal of the American Statistical Association*, 60:843–853.

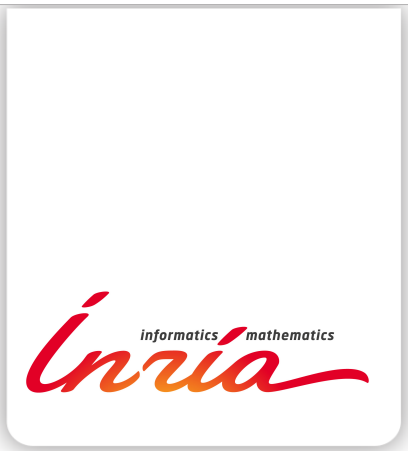
Castelle, N. and Laurent-Bonvalot, F. (1998). Strong approximations of bivariate uniform empirical processes. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 34(4):425–480.

Fraser, D. (1956). A vector form of the Wald-Wolfowitz-Hoeffding theorem. *The Annals of Mathematical Statistics*, 27:540–543.

Klotz, J. (1967). Asymptotic efficiency of the two sample Kolmogorov-Smirnov. *Journal of the American Statistical Association*, 62:932–938.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari*, 4(1):83–91.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF i. *Z. Wahrsch. Verw. Gebiete*, 32:111–131.



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération

Bâtiment A 29

33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399