On a L_1-Test Statistic of Homogeneity

Gérard BIAU [a] and László GYÖRFI [b]

[a] Institut de Mathématiques et de Modélisation de Montpellier
Equipe de Probabilités et Statistique
Université Montpellier II, UMR CNRS 5149, CC 51
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
e-mail: biau@math.univ-montp2.fr

[b] Department of Computer Science and Information Theory
Budapest University of Technology and Economics, H-1521 Stoczek u. 2
Budapest, Hungary
email: gyorfi@szit.bme.hu

Abstract

We present a simple and explicit multivariate procedure for testing homogeneity of two independent samples of size n. The test statistic T_n is the L_1 distance between the two empirical distributions restricted to a finite partition. We first discuss Chernoff-type large deviation properties of T_n. This results in a distribution-free strongly consistent test of homogeneity, which rejects the null if T_n becomes large. Then the asymptotic null distribution of the test statistic is obtained, leading to a new consistent test procedure.

AMS Classification: 62G10.
Key words and phrases: homogeneity testing, partitions, large deviations, consistent testing, central limit theorem, Poissonization.

1 Introduction

Consider two mutually independent samples of $\mathbb{R}^d$-valued random vectors $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$ with *i.i.d.* components distributed according to unknown probability measures μ and μ'. We are interested in testing the homogeneity null hypothesis that the two samples are drawn according to the same distribution, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

Such tests have been extensively studied in the statistical literature for special parametrized models, *e.g.* for linear or loglinear models. For example, the analysis of variance provides standard tests of homogeneity when μ and μ' belong to a normal family on the real line. For multinomial models these tests are discussed in common statistical textbooks, together with the related problem of testing independence in contingency tables. For testing homogeneity in more general parametric models, we refer the reader to the monograph of Greenwood and Nikulin [8] and further references therein.

In the present note, we discuss a simple approach based on a L_1 distance test statistic associated with some adequate partition. The advantage of our test procedure is that, besides being explicit and relatively easy to carry out, it requires very few assumptions on the partition sequence, and it is consistent. Let us now describe our test statistic.

Denote by μ_n and μ'_n the empirical measures associated with the samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, respectively, so that

$$\mu_n(A) = \frac{\#\{i : X_i \in A, i = 1, \ldots, n\}}{n}$$

for any Borel subset A, and, similarly,

$$\mu'_n(A) = \frac{\#\{i : X'_i \in A, i = 1, \ldots, n\}}{n}.$$

Based on a finite partition $\mathcal{P}_n = \{A_{n1}, \ldots, A_{nm_n}\}$ of $\mathbb{R}^d$ ($m_n \in \mathbb{N}^*$), we let the test statistic comparing μ_n and μ'_n be defined as

$$T_n = \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu'_n(A_{nj})|.$$

The note is organized as follows. We first discuss in Section 2 Chernoff-type large deviation properties of T_n. This results in a distribution-free strongly consistent test of homogeneity, which rejects the null hypothesis if T_n becomes large, i.e., T_n is larger than a critical value. In Section 3, we derive the asymptotic null distribution of T_n. This yields another – in fact, a smaller – critical value resulting in a consistent asymptotically α-level test procedure. Proofs of the presented results will appear elsewhere.

2 Large deviation properties

For testing a simple hypothesis versus a composite alternative, Györfi and van der Meulen [9] introduced a related goodness of fit test statistic L_n defined as

$$L_n = \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu(A_{nj})|.$$

Beirlant, Devroye, Györfi and Vajda [2], and Devroye and Györfi [5] proved that if

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{1}$$

and

$$\lim_{n\to\infty} \max_{j=1,\dots,m_n} \mu(A_{nj}) = 0, \tag{2}$$

then, for all $0 < \varepsilon < 2$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}\{L_n > \varepsilon\} = -g_L(\varepsilon),$$

where

$$g_L(\varepsilon) = \inf_{0<p<1-\varepsilon/2} D(p \,\|\, p+\varepsilon/2) \quad \text{and} \quad D(\alpha \,\|\, \beta) = \alpha \ln \frac{\alpha}{\beta} + (1-\alpha) \ln \frac{1-\alpha}{1-\beta}.$$

This means that

$$\mathbf{P}\{L_n > \varepsilon\} = e^{-n(g_L(\varepsilon)+o(1))} \quad \text{as } n \to \infty.$$

The following theorem extends the results of Beirlant, Devroye, Györfi and Vajda [2], and Devroye and Györfi [5] to the statistic T_n.

Theorem 1 *Assume that conditions (1) and (2) are satisfied. Then, under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,*

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}\{T_n > \varepsilon\} = -g_T(\varepsilon),$$

where

$$g_T(\varepsilon) = (1+\varepsilon/2)\ln(1+\varepsilon/2) + (1-\varepsilon/2)\ln(1-\varepsilon/2).$$

Observe that for small ε, $g_T(\varepsilon) \approx \varepsilon^2/4$, and that $\lim_{\varepsilon\uparrow 2} g_T(\varepsilon) = 2\ln 2$. According to Beirlant, Devroye, Györfi and Vajda [2], for small ε, $g_L(\varepsilon) \approx \varepsilon^2/2$. Moreover, in contrast to $g_T(\varepsilon)$, the rate function $g_L(\varepsilon)$ is unbounded as $\varepsilon \uparrow 2$, so T_n and L_n have different large deviation properties.

The technique of Theorem 1 yields a distribution-free strongly consistent test of homogeneity, which rejects the null hypothesis if T_n becomes large. The concept of strongly consistent test is quite unusual, it means that for each point in the null hypothesis, with probability one, the test accepts $\mathcal{H}_0$ for all sufficiently large n, and for each point in the alternative, with probability one, the test rejects $\mathcal{H}_0$ for all sufficiently large n. In other words, denoting by $\mathbf{P}_0$ (*resp.* $\mathbf{P}_1$) the probability under the null hypothesis (*resp.* under the alternative), we have

$$\mathbf{P}_0\{\text{rejecting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1$$

and

$$\mathbf{P}_1\{\text{accepting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1.$$

In a real life problem, for example, when we get the data sequentially, one gets data just once, and should make good inference for these data. Strong consistency means that the single sequence of inference is almost surely perfect if the sample size is large enough. This concept is close to the definition of discernability introduced by Dembo and Peres [4]. For an example and references we refer the reader to Devroye and Lugosi [6]. We insist on the fact that the test presented in Theorem 2 below is entirely distribution-free, i.e., the measures μ and μ' are completely arbitrary.

Theorem 2 *Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_1\sqrt{\frac{m_n}{n}},$$

where $c_1 > 2\sqrt{\ln 2} \approx 1.6651$. Assume that condition (1) is satisfied and

$$\lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

Then, under $\mathcal{H}_0$, almost surely, from some (random) n onwards the test always accepts. Moreover, if $\mu \neq \mu'$, and for any sphere S centered at the origin

$$\lim_{n\to\infty} \max_{A_{nj}\cap S\neq 0} \operatorname{diam}(A_{nj}) = 0, \tag{3}$$

then, almost surely, from some (random) n onwards the test always rejects.

3 Asymptotic normality

Beirlant, Györfi and Lugosi [3] proved that, under conditions (1) and (2),

$$\sqrt{n}\,(L_n - \mathbf{E}\{L_n\})\,/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\xrightarrow{\mathcal{D}}$ stands for the convergence in distribution and $\sigma^2 = 1 - 2/\pi$. The technique of Beirlant, Györfi and Lugosi [3] involves a Poisson representation of the empirical process in conjunction with Bartlett's [1] idea of partial inversion for obtaining characteristic functions of conditional distributions. Using the method of these authors, we can prove the following:

Theorem 3 *Assume that conditions (1) and (2) are satisfied. Then, under $\mathcal{H}_0$, with a centering sequence $(C_n)_{n\geq 1}$,*

$$\sqrt{n}\,(T_n - C_n)\,/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\sigma^2 = 2(1 - 2/\pi)$.

The main difficulty in proving Theorem 3 is that it states asymptotic normality of sums of *dependent* random variables. To overcome this problem, we use a 'Poissonization' argument originating from the fact that an empirical process is equal in distribution to the conditional distribution of a Poisson process given the sample size (for more on Poissonization techniques, we refer the reader to Giné, Mason and Zaitsev [7]).

Theorem 3 yields the asymptotic null distribution of a consistent homogeneity test, which rejects the null hypothesis if T_n becomes large. In contrast to Theorem 2, and because of condition (2), this new test is *not* distribution-free. In particular, the measures μ and μ' have to be nonatomic.

Theorem 4 *Put $\alpha \in (0,1)$, and let C denote some positive universal constant. Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_2\sqrt{\frac{m_n}{n}} + C\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha),$$

where $c_2 = 2/\sqrt{\pi} \approx 1.1284$, and where Φ denotes the standard normal distribution function. Then, under the conditions of Theorem 3, the test has asymptotic significance level α. Moreover, under the additional condition (3), the test is consistent.

References

[1] Bartlett, M. S. (1938). The characteristic function of a conditional statistic, *J. London Math. Soc.*, 13, pp. 62-67.

[2] Beirlant, J., Devroye, L., Györfi, L. and Vajda, I. (2001). Large deviations of divergence measures on partitions, *J. Statist. Plann. Inference*, 93, pp. 1–16.

[3] Beirlant, J., Györfi, L. and Lugosi, G. (1994). On the asymptotic normality of the L_1- and L_2-errors in histogram density estimation, *Canad. J. Statist.*, 22, pp. 309-318.

[4] Dembo, A. and Peres, Y. (1994). A topological criterion for hypothesis testing, *Ann. Statist.*, 22, pp. 106–117.

[5] Devroye, L. and Györfi, L. (2002). Distribution and density estimation, in *Principles of Nonparametric Learning*, (L. Györfi, Ed.), Springer-Verlag, Wien, pp. 223-286.

[6] Devroye, L. and Lugosi, G. (2002). Almost sure classification of densities, *J. Nonparametr. Stat.*, 14, pp. 675-698.

[7] Giné, E., Mason, D. M. and Zaitsev, A. Yu (2003). The L_1-norm density estimator process, *Ann. Probab.*, 31, pp. 719-768.

[8] Greenwood, P. E. and Nikulin, M. S. (1996). *A Guide to Chi-Squared Testing*, Wiley, New York.

[9] Györfi, L. and van der Meulen, E. C. (1990). A consistent goodness of fit test based on the total variation distance, in *Nonparametric Functional Estimation and Related Topics*, (G. Roussas, Ed.), Kluwer, Dordrecht, pp. 631-645.