Applied Mathematics Letters 22 (2009) 1029–1032

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

Using data compressors to construct order tests for homogeneity and component independence

Daniil Ryabko [a,b,*], Jürgen Schmidhuber [b]

[a] *INRIA Lille-Nord Europe, France*

[b] *IDSIA, Switzerland*

ARTICLE INFO

Article history:
Received 6 December 2007
Received in revised form 31 January 2008
Accepted 31 January 2008

Keywords:
Data compressors
Kolmogorov complexity
Component independence
Homogeneity testing
Information theory
Nonparametric statistical tests
Distribution-free statistical tests

ABSTRACT

Nonparametric order tests for homogeneity and component independence are proposed, which are based on data compressors. For homogeneity testing the idea is to compress the word obtained by ordering the combined samples and writing the number of the sample in the place of each element. H_0 should be rejected if the string is compressed to a certain degree and accepted otherwise. We show that such a test obtained from an ideal data compressor is valid against all alternatives. Component independence is reduced to homogeneity testing.

1. Introduction

We consider two classical problems of mathematical statistics. The first one is homogeneity testing: r samples $\mathbf{X}^j = \{X_1^j, \dots X_{m_j}^j\}$, $1 \le j \le r$ with elements in $\mathbb{R}^d$ are given. The elements are drawn independently and within each sample the distribution is the same. The hypothesis H_0 is that $\mathbf{X}^j$ for all values of j are distributed according to the same distribution and the alternative H_1 is that at least two distributions are different. No assumptions are made on the form of the distributions. The second problem is component independence: a sample $Z_1, \dots, Z_n$ is given, generated i.i.d. according to some distribution F_Z. Each Z_i consists of two components Z_i^1 and Z_i^2. We wish to test whether the components are independent of each other. That is, H_0 means that the marginal distributions are independent whereas H_1 means that there is some dependency. Again, no assumption is made on the distribution F_Z.

The literature on these statistical problems is vast, in particular there are many classic nonparametric tests (e.g. Kolmogorov–Smirnov test for homogeneity), some of which use ranks (e.g. Wilcoxon's test); we refer to [1] for an overview. The idea to use real-life data compressors for testing some classical statistical hypotheses was suggested in [2,3]. The hypotheses considered there mostly concern data samples drawn from discrete spaces. Some tests for continuous spaces are also proposed based on partitioning.

Here we extend this approach to *order* (rank) tests, allowing testing for homogeneity and component independence without the need of partitioning the sample spaces and making them finite. The main advantage of the suggested tests is that they are absolutely distribution free and provide a simple tool for applying real data compressors (and thus all the achievements of that field) to solving statistical problems.

* Corresponding address: INRIA Lille-Nord Europe, Parc Scientifique de la Haute Borne, Park Plaza - Bât A, 40 avenue Halley, 59650 Villeneuve d'Ascq, France. Tel.: +33 3 59 57 79 23; fax: +33 3 59 57 78 50.

E-mail addresses: daniil.ryabko@inria.fr, daniil@ryabko.net (D. Ryabko), juergen@idsia.ch (J. Schmidhuber).

0893-9659/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2008.01.008

The idea of using data compressors for tasks other than actual data compression was suggested in [4–6], where data compressors are applied to such problems as classification and clustering. These works were largely inspired by Kolmogorov complexity, which is also an important tool for the present work. An "ideal" data compressor is one which compresses its input up to its Kolmogorov complexity. This is intuitively obvious since, informally, Kolmogorov complexity of a string is the length of the shortest program that outputs this string. Such data compressors do not exist; in particular, Kolmogorov complexity itself is incomputable. Real data compressors, however, can be considered as approximations of ideal ones. Intuitively, an ideal data compressor can find all regularities in data, while real data compressors can find regularities (in sequential data) of those types that were found practically useful and simple enough.

The suggested tests provide a simple way to use these regularities to find inhomogeneities (homogeneity testing) and dependencies (component independence) in the data. Thus in this work we provide a simple empirical procedure for testing homogeneity and component independence with data compressors; we show that for an ideal data compressor this procedure provides a statistical test which is valid against all alternatives (Type II error goes to zero); while Type I error is guaranteed to be below a pre-defined significance level for all data compressors, not only for ideal ones. Based on the results of [2], it may be conjectured that ideal codes in this work can be replaced by arbitrary universal codes (e.g. [9,10]).

2. Main results

Homogeneity testing. We are given r samples $\mathbf{X}^j = \{X_1^j, \dots \mathbf{X}_{m_j}^j\}$, $1 \le j \le r$. The elements X_i^j, $1 \le i \le m_j$ are i.i.d. with distribution F_j on $\mathbb{R}^d$ ($d \in \mathbb{N}$), $1 \le j \le r$. We wish to test whether $F_1 = F_2 = \cdots = F_r$. No assumptions are made on the distributions F_j. Thus $H_0 = \{(F_1, \dots, F_r) : F_1 = \cdots = F_r\}$ and $H_1 = \{(F_1, \dots, F_r) : F_i \ne F_j \text{ for some } i, j\}$.

Denote $m = \sum_{j=1}^r m_j$. We use $|K|$ for the length of a string K. A *code* ϕ is a function $\phi : B^* \to \{1, 2\}^*$ from finite words over $B = \{1, 2, \dots, r\}$ to finite binary words, such that ϕ is an injection ($a \ne b$ implies $\phi(a) \ne \phi(b)$). For $r = 2$ a trivial example of a code is the identity $\phi_{id}(a) = a$. Less trivial examples that we have in mind are data compressors, such as `zip, rar, arj,` or others. We will construct (reasonable) tests for homogeneity from (good) data compressors.

First assume that $\mathbf{d} = \mathbf{1}$ (that is, $X_i^j \in \mathbb{R}$). Let $Z_1 \le Z_2 \le \cdots \le Z_m$ denote the sample constructed by ordering jointly all r samples $\mathbf{X}^1, \dots, \mathbf{X}^r$. Construct the word $A = A_1 \dots, A_m$ as follows: for each i write $A_i = t$ if Z_i is taken from the sample $\mathbf{X}^t$ ($Z_i \in \mathbf{X}^t$) where ties are broken by randomization: if $Z_j = Z_{j+1} = \cdots Z_{j'}$ and there are m'_j elements of the sample $\mathbf{X}^j$ which are equal to Z_j then the word $A_j \dots A_{j'}$ is chosen randomly from all $\frac{(\sum_{j=1}^r m'_j)!}{\prod_{j=1}^r (m'_j!)}$ r-letter words which have m'_j elements j, $1 \le j \le r$, assigning equal probabilities to all words. Now consider the case $\mathbf{d} > \mathbf{1}$. Construct samples $\bar{\mathbf{X}}^{\mathbf{j}} = \bar{X}_1^j, \dots, \bar{X}_{m_j}^j$ as follows: $\bar{X}_t^j := x_{jt}^{11}, x_{jt}^{21}, \dots, x_{jt}^{d1}, x_{jt}^{12}, x_{jt}^{22}, \dots, x_{jt}^{d2}, \dots$ where x_{jt}^{uv} is the vth element in the binary expansion of the uth component of X_t^j (in case the expansion is ambiguous always take the one with more zeros). Denote the described function which converts $\mathbf{X}^j$ to $\bar{\mathbf{X}}^j$ by τ. Construct the string A applying the (single-dimensional) procedure described above to the samples $\bar{\mathbf{X}}^j$, $1 \le j \le r$.

For any code ϕ the *test for homogeneity* G_ϕ rejects the hypothesis H_0 at the level of significance α if $|\phi(A)| \le \log \alpha N$, where $N = \frac{(\sum_{i=1}^r m_i)!}{\prod_{i=1}^r m_i}$ and log is base 2, and accepts H_0 otherwise.

The intuition is as follows. Let $r = 2$ (just two samples). Under H_0 (the distributions F_1 and F_2 are equal) the string A is a random binary string with m_1 1s and m_2 2s; all such strings are equiprobable. Thus a good data compressor should be able to compress A to about $\log N$ bits, but no code can compress many such strings to less than $\log N - t$ bits ($t > 0$), since there are N such strings and only $2^{-t}N$ binary strings of length $\log N - t$.

Theorem 1 (*Type I Error*). *For any code ϕ and any $\alpha \in [0, 1]$ the Type I error of the test G_ϕ with level of significance α is not greater than α:*

$$P\{(\mathbf{X}^1, \dots, \mathbf{X}^r) : G_\phi(\mathbf{X}^1, \dots, \mathbf{X}^r) = \textit{reject}\} \le \alpha \tag{1}$$

for all $P = (F_1, \dots, F_r) \in H_0$.

Proof. As was noted, under H_0 for every string $a \in B^m$ such that a has m_j elements j, for $1 \le j \le r$, we have $P(A = a) = 1/N$. Since there are only αN binary strings of length $\log \alpha N$ and ϕ is an injection, we get $P\{(\mathbf{X}^1, \dots, \mathbf{X}^r) : |\phi(A)| \le \log \alpha N\} \le \frac{1}{N} N\alpha = \alpha$ which together with the definition of G_ϕ implies (1). □

Clearly, for some codes the test is of little use (for example if ϕ is the identity mapping) and Theorem 1 is only useful when the Type II error is small too. Next we will define "ideal" codes and show that for them the probability of accepting H_0 goes to zero under any distribution in H_1. Codes used in real data compressors can then be thought of as practical approximations of ideal codes.

The complexity of a string $A \in B^*$ with respect to a Turing machine ζ is defined as $C_\zeta(A) = \min_p\{l(p) : \zeta(p) = A\}$, where p ranges over all binary strings (interpreted as programs for ζ; minimum over empty set is defined as ∞). There exists a Turing machine ζ such that $C_\zeta(A) \le C_{\zeta'}(A) + c_{\zeta'}$ for any A and any Turing machine ζ' (the constant $c_{\zeta'}$ depends on ζ' but

not on A). Fix any such ζ and define the *Kolmogorov complexity* of a string $A \in \{0,1\}^\infty$ as $C(A) := C_\zeta(A)$. For more details see e.g. [7,8]. Clearly, if A is a binary string then $C(A) \le |A| + b$ for some b depending only on ζ.

We say that a code ϕ is *ideal* if for some constant c the equality $|\phi(A)| \le C(A) + c$ holds for every A. Clearly such codes exist.

Theorem 2 (*Type II Error: Universal Validity*). *For any ideal code ϕ Type II error of the test G_ϕ with any fixed significance level $\alpha > 0$ goes to zero: $P(G_\phi(\mathbf{X}^1, \dots, \mathbf{X}^r) = accept) \to 0$ for any P in H_1 if $m_j \to \infty$, $1 \le j \le r$, in such a way that $0 < a < \frac{m_j}{m} < b < 1$ for all j and for some a, b.*

Proof. First observe that if $\mathbf{X}^i$ and $\mathbf{X}^j$ are distributed according to different distributions then $\bar{\mathbf{X}}^{\mathbf{i}}$ and $\bar{\mathbf{X}}^{\mathbf{j}}$ are also distributed according to different distributions. Indeed, the function τ is one-to-one, and transforms cylinder sets (sets of the form $\{x \in \mathbb{R}^d : x^{u_1 v_1} = b_1, \dots, x^{u_t v_t} = b_t; b_l \in \{0,1\}, t, u_l, v_l \in \mathbb{N}(1 \le l \le t)\}$) to cylinder sets. So together with F_j it defines some distribution $F_{\bar{j}}$ on $\mathbb{R}$. If distributions $F_{\bar{j}}$ and $F_{\bar{i}}$ are different then they are different on some cylinder set T, but then $F_{\bar{i}}(\tau(T)) \ne F_{\bar{j}}(\tau(T))$. Thus further in the proof we will assume that $d = 1$. We will consider the case of two samples $r = 2$, the general case is analogous. We have to show that the Kolmogorov complexity $C(A) = |\phi(A)|$ of the string A is less than $\log \alpha N \ge mh(\frac{m_2}{m}) + \log \frac{\alpha}{m}$ for any fixed α from some m onwards. To show this, we have to find a sufficiently short description $s(A)$ of the string A; then the Kolmogorov complexity $|\phi(A)|$ is not greater than $|s(A)| + c$ for some constant c.

If H_1 is true then there exist some interval $T = (-\infty, t]$ and some $\delta > 0$ such that $|F_1(T) - F_2(T)| > 2\delta$. Hence from some m_1, m_2 onwards with probability 1 we have

$$\left| \frac{\#\{x \in \mathbf{X}^1 \cap T\}}{m_1} - \frac{\#\{y \in \mathbf{X}^2 \cap T\}}{m_2} \right| > \delta. \tag{2}$$

Let A' be the starting part of A that consists of all elements that belong to T and let $m_1' := \#\{x \in \mathbf{X}^1 \cap T\}$ and $m_2' := \#\{y \in \mathbf{X}^2 \cap T\}$. A description of A' can be constructed as the index of A' in the set (ordered, say, lexicographically) of all binary strings of length $m_1' + m_2'$ that have m_1' zeros and m_2' ones plus the description of m_1' and m_2'. The length of such a description is bounded by $\log \frac{(m_1'+m_2')!}{m_1'!m_2'!} + \log m_1' + \log m_2' + const \le (m_1' + m_2')h(\frac{m_2'}{m_1'+m_2'}) + \log m_1' + \log m_2' + const$. Let $\bar{A}$ denote the remaining part of A (that is, what goes after A'). The length of the description of $\bar{A}$ is bounded by $(\bar{m}_1 + \bar{m}_2)h(\frac{\bar{m}_2}{\bar{m}_1+\bar{m}_2}) + \log \bar{m}_2 + \log \bar{m}_1 + const$ where $\bar{m}_1 = m_1 - m_1'$ and $\bar{m}_2 = m_2 - m_2'$. Since h is concave and $\frac{m_2}{m_1+m_2}$ is between $\frac{m_2'}{m_1'+m_2'}$ and $\frac{\bar{m}_2}{\bar{m}_1+\bar{m}_2}$, from Jensen's inequality we obtain

$$h\left(\frac{m_2}{m_1+m_2}\right) - \left(\frac{m_1'+m_2'}{m_1+m_2} h\left(\frac{m_2'}{m_1'+m_2'}\right) + \frac{\bar{m}_1+\bar{m}_2}{m_1+m_2} h\left(\frac{\bar{m}_2}{\bar{m}_1+\bar{m}_2}\right)\right) > 0.$$

Denote this difference by $\gamma(m_1, m_2, m_1', m_2')$. Let $\gamma = \inf \gamma(m_1, m_2, m_1', m_2')$ where the infimum is taken over all pairs m_1, m_2 that satisfy the condition of the theorem $0 < a < \frac{m_j}{m} < b < 1$ and m_1', m_2' that satisfy (2). It follows that $\inf |\frac{m_2'}{m_1'} - \frac{m_2}{m_1}| > 0$ and $\inf |\frac{\bar{m}_2}{\bar{m}_1} - \frac{m_2}{m_1}| > 0$. Thus, γ is positive and depends only on a, b and δ. To uniquely describe A we need the description of A' and $\bar{A}$ and also m_1 and m_2; these have to be encoded in a self-delimiting way; the length of such a description $s(A)$ is bounded by the lengths of description of $A', \bar{A}$ plus $\log m$ and some constant. Thus

$$\begin{aligned} mh\left(\frac{m_2}{m}\right) + \log\alpha - \log m - |\phi(A)| &\ge mh\left(\frac{m_2}{m}\right) + \log\alpha - 6\log m \\ &\quad - m\left(\frac{m_1'+m_2'}{m} h\left(\frac{m_2'}{m_1'+m_2'}\right) + \frac{\bar{m}_1+\bar{m}_2}{m} h\left(\frac{\bar{m}_2}{\bar{m}_1+\bar{m}_2}\right)\right) - c \\ &\ge m\gamma - 6\log m - c \end{aligned}$$

for some constant c; clearly, this expression is positive from some m onwards. □

Component independence testing. A sample $Z = Z_1, \dots, Z_n$ is given where each Z_i consists of r components $Z_i^1, Z_i^2, \dots, Z_i^r$, $Z_i^j \in \mathbb{R}^{d_j}$. The sample is generated i.i.d. with distribution F_Z on $\mathbb{R}^d$, where $d := \sum_{j=1}^r d_j$. The goal is to test whether the components are distributed independently. That is, H_0 is that $F_Z(Z_1^1 \in T_1, \dots, Z_1^r \in T_r) = \prod_{j=1}^r F_Z(Z_1^j \in T_j)$ for all measurable $T_j \subset \mathbb{R}^{d_j}$, $1 \le j \le r$. H_1 is the negation of H_0 (the equality is false for some selection of the sets T_j, $1 \le j \le r$). Again, no assumption is made on the form of the distribution F_Z.

Fix any code ϕ and construct the test for component independence I_ϕ as follows. Assume that $n = 2m$ for some m and define the samples X and $\bar{Y}$ as the first and the second halves of the sample Z: $X_1 = Z_1, \dots, X_m = Z_m$ and $\bar{Y}_1 = Z_{m+1}, \dots, \bar{Y}_m = Z_{2m}$ (if n is odd then make samples X and $\bar{Y}$ of sizes $[n/2]$ and $n - [n/2]$). Construct the sample Y from $\bar{Y}$ by permuting the components independently: $Y_i^j = \bar{Y}_{\pi_j(i)}^j$, $1 \le i \le m$, $1 \le j \le r$ where π_j are permutations of $\{1 \dots m\}$, selected at random with equal probabilities from the set of all such permutations, independently of each other.

The *component independence test* I_ϕ (with level of significance α) consists in application of the test for homogeneity G_ϕ to the samples X and Y (with level of significance α). Indeed, it is easy to check that H_0 is true if and only if X and Y are distributed according to the same distribution (i.e. only in this case permuting one of the components independently does not change the distribution). So we get the following statement.

Theorem 3. *For any code ϕ and any $\alpha \in (0, 1]$ the Type I error of the test I_ϕ with level of significance α is not greater than α. If, in addition, the code ϕ is ideal then the Type II error of I_ϕ error tends to 0 as the sample size n approaches infinity.*

Acknowledgement

This research was supported by the Swiss NSF grant 200021-113364.

References

[1] E. Lehmann, Testing Statistical Hypotheses, 2nd ed., John Wiley & Sons, New York, 1986.
[2] B. Ryabko, J. Astola, A. Gammerman, Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series, Theoret. Comput. Sci. 359 (2006) 440–448.
[3] B. Ryabko, V. Monarev, Using information theory approach to randomness testing, J. Statist. Plann. Inference 133 (1) (2005) 95–110.
[4] R. Cilibrasi, P. Vitányi, Clustering by compression, IEEE Trans. Inform. Theory 51 (4) (2005).
[5] R. Cilibrasi, R. de Wolf, P. Vitányi, Algorithmic clustering of music, Comput. Music J. 28 (4) (2004) 49–67.
[6] M. Li, X. Chen, X. Li, B. Ma, P. Vitányi, The similarity metric, IEEE Trans. Inform Theory 50 (12) (2004) 3250–3264.
[7] N. Vereshchagin, A. Shen, V. Uspensky, in: Lecture Notes on Kolmogorov Complexity, 2004, unpublished. http://lpcs.math.msu.su/~ver/kolm-book.
[8] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, Springer, 1997.
[9] B. Ryabko, Prediction of random sequences and universal coding, Probl. Inf. Transm. 24 (2) (1988) 87–96.
[10] J. Ziv, A. Lempel, Compression of individual sequences via variable-rate coding, IEEE Trans. Inform Theory IT-24 (5) (1978) 530–536.