Control

and Cybernetics

VOL. 18 (1989) No. 1

Empirical tests of multidimensional uniformity

by

JÓZEF POCIECHA
ANDRZEJ SOKOŁOWSKI

Institute of Statistics, Econometrics
and Computer Science
Academy of Economics
Rakowicka 27
31-510 Kraków, Poland

Five test statistics are proposed to test the multidimensional uniformity. They are based on normalized spanning tree which is supposed to have statistically equal links if null hypothesis holds. Critical values have been obtained through Monte Carlo simulations.

The classical cluster analysis methods do not usually involve statistical inference. They often do not assume the existence of the so called general population (or generating model). The new approach assumes that the set of operational taxonomic units (OTU's) has been generated by the mixture of random variables. Such a mixture can be described by three parameters: s — the number of subpopulations (subgroups, clusters) in the generating model, M — the matrix of parameters describing subpopulations, p — vector which defines the structure of the mixture (p_j is the fraction of OTU's generated by j-th contributing variable). It is convenient to assume each subpopulation to be modelled by the probability distribution of the same type (e.g. normal). Each pair of these distributions should be different at least in one of their parameters (mean, median, mode).

The aim of statistical cluster analysis to estimate s, M and p which is equivalent to the establishment of the number of subpopulations (clusters), the structure of the mixture and statistical characteristics of subgroups.

At the beginning of the analysis which may include the application of cluster analysis procedure we should check whether the set of observations is „suitable" for grouping. In many clustering methods it is not tested if the set of OTU's has been generated by „mixture" with $s \rangle 1$ or only by one population variable with $s = 1$, when it should not be artificially clustered.

There are three kinds of hypotheses which can be formulated, within our area of interest, concerning the population which generates the analysed set of OTU's:

— hypothesis of homogenity

— hypothesis of unimodality

— hypothesis of uniformity

The first one assumes that there is only one population variable which creates the mixture ($s = 1$). This type of approach is discussed by Bock [1]. The second hypothesis tells that this variable is unimodal. Some tests for unimodality were proposed by Hartigan [2] or Hartigan and Hartigan [3]. The third hypothesis assumes that the underlying distribution is uniform.

The set of operational taxonomic units can be considered as the set of points in the k-dimensional space called the classification space, which is generated by attributes describing analysed objects. The classification space is usually multidimensional ($k \rangle 1$).

In this paper we have concentrated on the hypothesis of multidimensional uniformity:

$$H_o\colon F \in F_u$$
$$H_1\colon F \notin F_u$$

where F_u is a multidimensional uniform distribution. If the uniformity hypothesis holds, the OTU's should be distributed „equally" in the classification space. No concentration of points and no „holes" are expected. The structure of points can be described by the spanning tree presented as a graph containing n nodes connected by $(n-1)$ links. First, the nearest neighbours are linked using information derived from the distance matrix and then the resulting subgraphs are joined to get the connected graph of n nodes (OTU's). Such a graph with no loops is a part of multidimensional net with observations located at some of its links. If the underlying distribution is uniform this net should be regular, i.e. links connecting the closest objects should be statistically equal. This does not depend on the dimension of classification space.

Let's take a_i as the i-th link in the minimal graph. The transformation $a_i^* = a_i / \sum a_i$ leads to the normalized graph with links a_i^* and with total length equal to unity.

We have considered five test statistics which compare the actual structure of the normalized minimal graph with the theoretical graph with equal links.

1. Dissimilarity of structure

$$A_1 = 1 - \sum_{i=1}^{n-1} \min\{a_i^*;\ 1/(n-1)\} \qquad (1)$$

This statistic takes values from [0,1].

At this point „the structure" is understood as the set of values a_i^* satisfying $a_i^* \geqslant 0$ and $\sum a_i^* = 1$. The statistic A_1 given by formula (1) measures the difference between the structure of a sample graph and the theoretical graph.

2. Minimal link

$$A_2 = \min_i \{a_i^*\} \tag{2}$$

3. Maximal link

$$A_3 = \max_i \{a_i^*\} \tag{3}$$

4. Range

$$A_4 = A_3 - A_2 \tag{4}$$

5. Variance

$$A_5 = (n-1)^{-1} \sum_{i=1}^{n-1} [a_i^* - 1/(n-1)]^2 \tag{5}$$

The analytical way of finding the distributions of the above statistics is very difficult. For this reason we propose the empirical way to define the distributions under consideration. The distributions of the above statistics have been investigated by means of Monte Carlo analysis. The number of objects varied from 3 up to 50. The consideration of rather low number of objects is usually sufficient for economic and geographical investigations (the number of provinces — voivodships — in Poland is 49). And also because of normalization of links in the minimal graph, the critical values of statistics rather quickly tend to zero. Publication of critical values of test statistics for higher number of objects would widely extend the size of tables which are here presented. In this paper we present only quantiles suitable for testing the uniformity (the significance levels $\alpha = 01$ or $\alpha = 0.05$). Initial estimates have been obtained with 100 simulation runs and then they were smoothed by polynomials. Polynomials of order 4 or 5 had given very good approximations.

All statistics should be used in the one-sided versions. Statistic A_2 is left-sided (too small links are „not allowed"). The null hypothesis is rejected when $A_2 \leqslant A_{2\alpha}$, because:

$$P\{A_2 \leqslant A_{2\alpha}\} \cong \alpha$$

Other statistics are right-sided. The choice of the particular test statistic depends on the kind of alternative distribution (hypothesis) we can expect. Statistic A_2 is good against unimodal alternatives while A_3 reacts on the multimodality or „gaps" in probability. A_4 combines both previous, so that compact distant clusters are the most suitable alternative. Statistics A_1 and A_5 take into account all links in the graph. The latter one tends to zero rather quickly, with the increasing number of OTU's, and this can require good accuracy of calculations.

It should be noticed that the distance measure which is used at the stage of creation of distance matrix can produce some undesirable effects. The variance of distance measure ought to be independent of the number of attributes to prevent the artificial uniformity of the graph which can occur for a classification space with a higher order dimension. Unfortunately, standardization of the data also transforms the structure but cannot be avoided when attributes have different measure units.

We have made many experiments with artificial data sets (from random number generators) and with sets of real objects but at the time we are not able to settle one particular alternative distribution to estimate the power of proposed statistics.

Table 1 presents the critical values of structure dissimilarity measure (A_1). An

Table 1

Critical values of structure dissimilarity measure (A_1)

α n	0.10	0.05	α n	0.10	0.05
3	0.4607	0.4962	27	0.3251	0.3395
4	0.4482	0.4801	28	0.3233	0.3368
5	0.4365	0.4658	29	0.3217	0.3342
6	0.4258	0.4531	30	0.3201	0.3318
7	0.4159	0.4418	31	0.3186	0.3296
8	0.4068	0.4318	32	0.3172	0.3275
9	0.3983	0.4228	33	0.3158	0.3257
10	0.3906	0.4147	34	0.3144	0.3241
11	0.3835	0.4074	35	0.3131	0.3227
12	0.3770	0.4008	36	0.3118	0.3215
13	0.3710	0.3947	37	0.3105	0.3206
14	0.3656	0.3892	38	0.3093	0.3198
15	0.3606	0.3841	39	0.3081	0.3192
16	0.3560	0.3793	40	0.3069	0.3189
17	0.3519	0.3748	41	0.3058	0.3186
18	0.3481	0.3705	42	0.3047	0.3185
19	0.3446	0.3665	43	0.3036	0.3184
20	0.3414	0.3626	44	0.3027	0.3183
21	0.3384	0.3589	45	0.3017	0.3182
22	0.3358	0.3554	46	0.3009	0.3180
23	0.3333	0.3519	47	0.3002	0.3179
24	0.3310	0.3486	48	0.2996	0.3177
25	0.3289	0.3455	49	0.2991	0.3175
26	0.3269	0.3424	50	0.2987	0.3174

Critical values of structure dissimilarity measure (A_1)

Table 2

Critical values for statistics A_2 , A_3 , A_4 , A_5 at $\alpha = 0.05$

n	A_2	A_3	A_4	A_5
3	0.0326	0.8977	0.8552	0.2075
4	0.0181	0.8192	0.7832	0.1319
5	0.0133	0.7479	0.7176	0.0835
6	0.0082	0.6833	0.6581	0.0538
7	0.0050	0.6251	0.6042	0.0364
8	0.0044	0.5727	0.5557	0.0269
9	0.0039	0.5259	0.5121	0.0219
10	0.0015	0.4841	0.4731	0.0193
11	0.0008	0.4470	0.4384	0.0177
12	0.0008	0.4142	0.4076	0.0162
13	0.0008	0.3855	0.3805	0.0146
14	0.0007	0.3604	0.3567	0.0127
15	0.0007	0.3386	0.3359	0.0106
16	0.0005	0.3199	0.3180	0.0084
17	0.0005	0.3038	0.3025	0.0063
18	0.0005	0.2902	0.2893	0.0045
19	0.0005	0.2788	0.2781	0.0044
20	0.0005	0.2693	0.2687	0.0043
21	0.0004	0.2614	0.2609	0.0042
22	0.0004	0.2550	0.2545	0.0040
23	0.0004	0.2498	0.2492	0.0039
24	0.0003	0.2456	0.2448	0.0038
25	0.0003	0.2423	0.2414	0.0037
26	0.0003	0.2396	0.2386	0.0036
27	0.0003	0.2374	0.2362	0.0034
28	0.0003	0.2356	0.2343	0.0033
29	0.0003	0.2340	0.2326	0.0032
30	0.0003	0.2325	0.2311	0.0031
31	0.0003	0.2310	0.2296	0.0030
32	0.0003	0.2294	0.2280	0.0029
33	0.0003	0.2277	0.2265	0.0027
34	0.0003	0.2257	0.2245	0.0026
35	0.0003	0.2234	0.2223	0.0025
36	0.0003	0.2208	0.2199	0.0024
37	0.0003	0.2178	0.2172	0.0023
38	0.0003	0.2145	0.2141	0.0021
39	0.0003	0.2108	0.2107	0.0020
40	0.0003	0.2068	0.2070	0.0019
41	0.0003	0.2025	0.2029	0.0018
42	0.0003	0.1980	0.1985	0.0016
43	0.0003	0.1932	0.1935	0.0015
44	0.0003	0.1883	0.1890	0.0014
45	0.0003	0.1834	0.1841	0.0013
46	0.0003	0.1786	0.1790	0.0011
47	0.0003	0.1739	0.1740	0.0010
48	0.0003	0.1695	0.1690	0.0009
49	0.0003	0.1656	0.1643	0.0008
50	0.0003	0.1622	0.1600	0.0007

Critical values for statistics A_2 , A_3 ,A_4 , A_5 at $\alpha = 0.05$

empirical value of A_1 greater than $A_{1\alpha}$ indicates that hypothesis of multidimensional uniformity ought to be rejected. In authors' experience, this statistic looks the most promising. In Table 2 critical values of remaining statistics are given, for the significance level $\alpha = 0.05$.

References

[1] BOCK H.H. On Some Significance Tests in Cluster Analysis. *Journal of Classification*, 2 (1985).

[2] HARTIGAN J.A.The Span Test for Unimodality, in: Classification and Related Methods of Data Analysis, H.H. Bock (ed.) North-Holland, Amsterdam, 1988.

[3] HARTIGAN J.A., HARTIGAN P.M. The Dip Test of Unimodality. *The Annals of Statistics*, 13 (1985).

Received, May 1988.

Testy empiryczne wielowymiarowej jednorodności

Zaproponowano pięć statystyk testowych przeznaczonych do sprawdzania hipotezy wielowymiarowej jednorodności. Oparte są one na znormalizowanym dendrycie minimalnym, o którym można przypuścić, że w rozważanym przypadku będzie miał statystycznie równe krawędzie, jeśli prawdziwa jest hipoteza zerowa. Wartości krytyczne dla odpowiednich testów zostały otrzymane przy pomocy symulacji metodą Monte Carlo.

Эмпирические тесты многомерной однородности

Предлагается пять тестовых статистик, предназначенных для проверки гипотезы многомерной однородности. Они основаны на нормализованном минимальном дереве, о котором можно предположить, что в рассматриваемом случае будет иметь статистически равные грани, если правомерна нулевая гипотеза. Критические значения для соответствующих тестов были получены с помощью иммитации посредством метода Монте Карло.

Instructions to Authors

„Control and Cybernetics" publishes original papers which have not been published and will not be simultaneously submitted elsewhere. The preferred language of the papers is English.

No paper should exceed in length 20 typewritten pages (210–297 mm) of the text, double spaced and with 50 mm margin on the left-hand side. Manuscripts should be submitted in triplicate, typed only on one side of the sheet of paper.

The plan and form of the submitted manuscripts is as follows:

1. The heading should include the title, full names and surnames of the authors in the alphabetic order, as well as the names and addresses of the institutions they represent. The heading should be followed by a concise ummary (of approximately 15 typewritten lines).

2. Figures, photographs, tables, diagrams etc. should be enclosed to the manuscript. The texts related should be typed on a separate page.

3. All elements of mathematical formulae should be typewritten whenever possible. A special attention is to be paid towards differentiating between capital and smaller letters. All the Greek letters appearing in the text should be defined. Indices and exponents should be written with special care. Round brackets should not be replaced by the inclined fraction line.

In general, elements easily confused are to be identified by the appropriate previously discusse measures or by a circled word or words explaining the element.

4. References should be listed in alphabetical order on a separate sheet. For journals the following information should appear: names (including initials or first names) of all authors, full title of paper, and journal name, volume, issue, pages, year of publication. Books cited should list author(s), full title, edition, place of publication, publisher, and year. Examples are:

Lukes D. Optimal regulation of nonlinear dynamical systems. SIAM J. Control 7 (1969) 1, 75-100.

Athans M., Falb P. Optimal Control. New York, Mc Graw-Hill 1966.

Wskazówki dla autorów

W wydawnictwie Control and Cybernetics drukuje się prace oryginalne nie publikowane w innych czasopismach. Zalecane jest nadsyłanie artykułów w języku angielskim. W przypadku nadesłania artykułu w języku polskim Redakcja może zalecić przetłumaczenie na język angielski. Objętość artykułu nie powinna przekraczać 1 arkusza wydawniczego, czyli ok. 20 stron maszynopisu formatu A4 z zachowaniem interlinii i marginesu 5 cm z lewej strony. Prace należy składać w 3 egzemplarzach. Układ pracy i forma powinny być dostowane do niżej podanych wskazówek.

1. W nagłówku należy podać tytuł pracy, następnie imię (imiona) i nazwisko (nazwiska) autora (autorów) w porządku alfabetycznym oraz nazwę reprezentowanej instytucji i nazwę miasta. Po tytule należy umieścić krótkie streszczenie pracy (do 15 wierszy maszynopisu).

2. Materiał ilustracyjny powinien być dołączony na oddzielnych stronach. Podpisy pod rysunki należy podać oddzielnie.

3. Wzory i symbole powinny być wpisane na maszynie bardzo starannie.

Szczególną uwagę należy zwrócić na wyraźne zróżnicowanie małych i dużych liter. Litery greckie powinny być objaśnione na marginesie. Szczególnie dokładnie powinny być opisane indeksy (wskaźniki) i oznaczenia potęgowe. Należy stosować nawiasy okrągłe.

4. Spis literatury powinien być podany na końcu artykułu. Numery pozycji literatury w tekście zaopatruje się w nawiasy kwadratowe. Pozycje literatury powinny zawierać nazwisko autora (autorów) i pierwsze litery imion oraz dokładny tytuł pracy (w języku oryginału), a ponadto:

a) przy wydawnictwach zwartych (książki) — miejsce i rok wydania oraz wydawcę;

b) przy artykułach z czasopism: nazwę czasopisma, numer tomu, rok wydania i numer bieżący.

Pozycje literatury radzieckiej należy pisać alfabetem oryginalnym, czyli tzw. grażdanką.