QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 123-130, 1995

EMPIRICAL SIGNIFICANCE TEST OF THE GOODNESS-OF-FIT FOR SOME PYRAMIDAL CLUSTERING PROCEDURES

CARLES CAPDEVILA MARQUÈS* and ANTONI ARCAS PONS‡

Through a series of simulation tests by Monte Carlo methods, some aspects related to the inference concerning pyramidal trees build by the maximum and minimum methods are considered. In this sense, the quantiles of the γ-Goodman-Kruskal statistic allow us to tabulate a significance test of the goodness-of-fit of a pyramidal clustering procedure. On the other side, the pyramidal method of maximum is observed to be clearly better (more efficient) than that of the minimum in terms of the expected value for the gamma statistic. Both for the maximum and minimum methods, a relation between the number of objects to classify and the gamma distribution is observed.

Key words: Goodnes-of-fit, Pyramidal Clustering, Sample distribution of the γ-statistic, Simulation test.

1. INTRODUCTION

Ultrametric trees are the most studied representations using discrete models. The aim of this model is to achieve a family of partitions that can be interpreted as a set of "natural" classifications of the population to classify, Ω.

Pyramidal trees, introduced by E. Diday, are a logical generalization of ultrametric trees. They are less restrictive structures where recovering replaces the concept of

*Carles Capdevila Marquès. Departament de Matemàtica. Universitat de Lleida. C/. Bisbe Messeguer, s/n. 25003. Lleida.

‡Antoni Arcas Pons. Departament d'Estadística. Universitat de Barcelona. Avda. Diagonal, 645. 08028 Barcelona.

-Article rebut el juny de 1994.

-Acceptat el novembre de 1994.

partition, obtaining a representation which bears more information and is closer to the initial dissimilarities. Diday (1986), Fichet (1984), Durand (1986, 1988) have studied some interesting topics about this model.

The pyramidal–representation models intend to detect the presence of a pyramidal structure of the population, starting from a dissimilarity matrix concerning the population. The process having this aim consist in transforming the initial dissimilarity into a pyramidal dissimilarity by means of some known pyramidal clustering procedure algorithm; this pyramidal dissimilarity is equivalent to an indexed pyramid.

In applied problems it is necessary to measure the fitting between the pyramidal tree obtained from some algorithm and the initial structure. In this sense, the most used parameters are the γ–Goodman–Kruskal coefficient (1954) and the cophenetic correlation coefficient ρ (Farris J.S., 1969).

In spite of having these coefficients as a measure of the fitting between the initial structure and the pyramidal tree obtained, it is difficult to determine exactly up to which point these coefficients are really significant for some particular case.

For example, let $\Omega = \{\omega_1, \dots, \omega_6\}$ and let δ be a dissimilarity on Ω given by the matrix

$$\delta = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ & 0 & 2 & 1 & 3 & 4 \\ & & 0 & 3 & 2 & 1 \\ & & & 0 & 2 & 3 \\ & & & & 0 & 4 \\ & & & & & 0 \end{pmatrix}$$

If we now carry out a pyramidal clustering procedure by the methods of the minimum and the maximum and calculate the value of the gamma and rho coefficients, we obtain: $\gamma = 0.92$ and $\rho = 0.89$ in the case of the maximum, and $\gamma = 0.80$ and $\rho = 0.59$ in the case of the minimum. Although these values are near to unity, there is nothing allowing us to assure whether they are really significant, i.e. up to which point they reflect the fitting between the initial structure and the pyramidal tree obtained. Nevertheless, it is necessary to find an objective criterion showing if the fitting is good. This becomes a characteristic problem in Statistical Inference.

On the other side, it would be also convenient to be able to evaluate the power of the pyramidal representation methods.

Generally, it is very difficult to find the exact distribution function for γ and ρ. Therefore, we shall develop our study from an empirical point of view, using some simulation techniques by means of Monte Carlo methods. For this purpose, it was necessary to programme a pyramidal clustering procedure algorithm and to create a simulation programme which made possible to obtain the sample distribution of gamma and to tabulate a goodness-of-fit test of the pyramidal representation with regard to the γ statistic, using the methods of the maximum and the minimum.

Also, the γ distribution as a function of the number of objects to classify is being studied. Finally, some results refering to the power–efficiency of the methods of the maximum and the minimum, and no commented in this paper, are obtained.

2. SIMULATION TESTS

Two simulation tests have been carried out, which we have called S1 and S2, by means of the programmes SIMULU and SIMULN respectively, made up for this purpose.

Basing on a value n, which represents the number of objects in the population to classify, $\Omega = \{\omega_1, \ldots, \omega_n\}$, the S1 test consists in generating N random dissimilarities U(0,1), δ_i^n with i = 1, ..., N. Starting from each one of them, a pyramidal representation is carried out using the methods of the maximum and the minimum. By this means, we shall obtain N pyramidal dissimilarities $d_{M,i}^n$ and $d_{m,i}^n$ for each one of both methods; we shall then compare each pyramidal dissimilarity with the respective initial dissimilarity through the gamma coefficient and obtain N gamma values $\gamma_{M,i}^n$ in the case of the maximum method, and other N gamma values $\gamma_{m,i}^n$ in the case of the minimum method. From these N values the mean M_γ, the standard deviation S_γ and the quantiles Q_α ($\alpha = 0.05,\ 0.10,\ 0.50,\ 0.75,\ 0.90,\ 0.95$) of the γ statistic are calculated. This results are shown in Table 1 and Table 3.

In our study we have considered populations with $n = 4, 5, \ldots, 18, 10, 25$ objects. The number of simulations carried out was N $= 200$ for $n = 25$, and N $=$ 1000 for the other values of n.

The S2 test was set out in the same terms as the S1 test, but replacing the random dissimilarities U(0,1) by values with a distribution N(0,1) adding the constant 10 in order to avoid negative values in the dissimilarity matrix. The results obtained in this test are shown in Table 2 and Table 4.

The aim of setting out a second test was mainly to see whether the distribution of the dissimilarities generated randomly had an effect anyhow on the results. As it can be seen from Tab. 1 and Tab. 2 as well as in Tab. 3 and Tab. 4, the results obtained in both cases (uniform or normal distribution) for the means, standard deviations and quantiles of gamma are virtually the same. As a conclusion, these results seem to point out that in case of a random assignation of dissimilarities, the relationship between the number of objects and the expected value of gamma does not depend on the distribution used for generating the random dissimilarities. Anyway, in order to confirm this result, which we are just suggesting here, it would be convenient to carry out further tests with other distributions for the initial dissimilarities, which we leave for a later work.

3. GOODNESS-OF-FIT TEST IN A PYRAMIDAL CLUSTERING PROCEDURE

Table 1 and Table 2, show that the sample means of gamma and its standard deviations decrease as n increases. In addition, the results obtained by the minimum pyramidal clustering procedure coincide with those obtained by L. Hubert (1974), where a relationship between n and the sample mean of gamma, as well as between n and the standard deviation of gamma, in the case of the minimum hierarchical clustering procedure.

From theoretical hypothesis about Ω and Π (algorithm), it is very difficult to find the exact distribution function for γ and ρ. We have found an approximation of these functions by means of the quantiles obtained in the simulation (Tab. 3 and Tab. 4).

In a more concrete way, if $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $\Delta = \{\delta_\Omega\}$, i.e. the dissimilarities family defined on Ω, we can consider Π the algorithm that transforms a dissimilarity defined on Ω into a pyramidal form.

Let the function $$\begin{array}{rccc} \gamma_\Pi & : & \Delta & \longrightarrow & \mathbb{R} \\ & & \delta_\Omega & \longrightarrow & \gamma_\Pi(\delta_\Omega) \end{array}$$

where $\gamma_\Pi(\delta_\Omega)$ is the Goodman–Kruskal coefficient between δ_Ω and the pyramidal dissimilarity obtained from Π.

If δ_Ω is pyramidal, then $\Pi(\delta_\Omega)$ is pyramidal too and coincides with δ_Ω, so $\gamma_\Pi(\delta_\Omega) = 1$. On the contrary, if δ_Ω is obtained by random generation, $\gamma_\Pi(\delta_\Omega)$ would be close to zero. In this way, if we consider a population Δ with random distances generated by some method, it could be possible to tabulate the distribution function for γ.

If we want to know if some dissimilarity obtained could be represented by a pyramidal structure, we are forced to consider a null hypothesis $\mathbf{H}_0$ that represents randomness in the sense that Δ contains pyramidal dissimilarities obtained using the process Π from random distances generated by some method. In this way, from the quantiles table of γ (n= $4, \ldots, 18, 20, 25$ objects), we obtain a goodness–of–fit test of the pyramidal representation using the minimum method and the maximum method. We also obtain that maximum method works better than minimum method in the above sense.

In practice, if the gamma value after applying a pyramidal clustering procedure is greater than the quantile Q_α, we can reject the randomness hypothesis for the initial dissimilarity at a significance level of $1 - \alpha$.

Relationship between the number n of objects of Ω and the sample mean (M_γ) and sample standard deviation (S_γ) for gamma, for the methods of the minimum and the maximum. Mean and Standard Deviation based on a sample of N = 200 for n = 25 and N=1000 for $n = 4, \ldots, 18, 20$.

Table 1

Initial random dissimilarity U(0,1)

	Maximum		Minimum	
n	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.74	0.10	0.54	0.17
7	0.66	0.09	0.47	0.14
8	0.59	0.09	0.40	0.13
9	0.53	0.08	0.36	0.11
10	0.49	0.07	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.06	0.25	0.08
14	0.36	0.06	0.23	0.08
15	0.34	0.06	0.22	0.07
16	0.32	0.05	0.21	0.06
17	0.30	0.05	0.20	0.06
18	0.28	0.05	0.18	0.06
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Table 2

Initial random dissimilarity N(0,1)+10

	Maximum		Minimum	
n	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.75	0.11	0.54	0.16
7	0.66	0.10	0.46	0.15
8	0.59	0.09	0.41	0.13
9	0.54	0.09	0.36	0.12
10	0.49	0.08	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.07	0.25	0.08
14	0.36	0.06	0.24	0.07
15	0.34	0.06	0.21	0.07
16	0.32	0.06	0.20	0.06
17	0.30	0.05	0.19	0.06
18	0.28	0.05	0.18	0.05
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Table 3

Relationship between n and the quantiles of the γ statistic. Initial random dissimilarity U(0,1)

	Maximum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.86	0.86	0.86	1.00	1.00	1.00	1.00
5	0.68	0.71	0.78	0.85	0.91	1.00	1.00
6	0.56	0.61	0.68	0.75	0.81	0.86	0.90
7	0.49	0.53	0.60	0.66	0.72	0.78	0.82
8	0.43	0.47	0.52	0.60	0.65	0.71	0.74
9	0.39	0.43	0.47	0.53	0.58	0.64	0.67
10	0.37	0.39	0.44	0.49	0.53	0.58	0.61
11	0.33	0.35	0.40	0.45	0.50	0.55	0.58
12	0.29	0.32	0.37	0.42	0.47	0.51	0.54
13	0.28	0.30	0.34	0.39	0.43	0.46	0.48
14	0.26	0.28	0.32	0.36	0.40	0.44	0.46
15	0.24	0.26	0.30	0.34	0.38	0.42	0.44
16	0.23	0.25	0.28	0.32	0.36	0.39	0.41
17	0.21	0.23	0.26	0.30	0.34	0.37	0.39
18	0.21	0.22	0.25	0.28	0.32	0.35	0.37
20	0.17	0.19	0.22	0.26	0.29	0.32	0.33
25	0.13	0.15	0.18	0.21	0.23	0.25	0.26

	Minimum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.45	0.45	0.64	0.82	1.00	1.00	1.00
5	0.31	0.37	0.48	0.64	0.77	0.88	0.94
6	0.24	0.32	0.43	0.54	0.64	0.76	0.81
7	0.24	0.28	0.37	0.46	0.56	0.66	0.72
8	0.19	0.23	0.31	0.41	0.49	0.57	0.61
9	0.17	0.21	0.28	0.36	0.44	0.51	0.55
10	0.16	0.19	0.26	0.32	0.40	0.46	0.50
11	0.14	0.17	0.23	0.30	0.36	0.41	0.46
12	0.13	0.16	0.22	0.27	0.33	0.38	0.41
13	0.12	0.15	0.20	0.25	0.30	0.35	0.38
14	0.12	0.14	0.18	0.23	0.29	0.33	0.36
15	0.10	0.13	0.17	0.21	0.26	0.31	0.34
16	0.10	0.12	0.16	0.20	0.25	0.29	0.31
17	0.10	0.12	0.16	0.20	0.24	0.28	0.30
18	0.08	0.11	0.14	0.18	0.22	0.26	0.28
20	0.08	0.10	0.13	0.16	0.20	0.23	0.25
25	0.06	0.09	0.10	0.13	0.16	0.19	0.20

Table 4

Relationship between n and the quantiles of the γ statistic. Initial random dissimilarity N(0,1) + 10

	Maximum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.86	0.86	0.86	1.00	1.00	1.00	1.00
5	0.67	0.71	0.79	0.86	1.00	1.00	1.00
6	0.56	0.60	0.67	0.75	0.83	0.88	0.91
7	0.50	0.53	0.59	0.66	0.73	0.79	0.82
8	0.44	0.48	0.54	0.60	0.65	0.71	0.74
9	0.40	0.42	0.48	0.53	0.60	0.64	0.68
10	0.36	0.39	0.44	0.49	0.55	0.60	0.62
11	0.32	0.35	0.41	0.45	0.51	0.55	0.57
12	0.30	0.33	0.37	0.42	0.46	0.50	0.53
13	0.28	0.30	0.34	0.39	0.43	0.47	0.50
14	0.25	0.27	0.32	0.36	0.40	0.45	0.47
15	0.24	0.26	0.30	0.33	0.37	0.41	0.43
16	0.22	0.25	0.28	0.32	0.35	0.39	0.41
17	0.21	0.23	0.26	0.30	0.34	0.37	0.39
18	0.20	0.21	0.24	0.28	0.31	0.35	0.37
20	0.18	0.19	0.22	0.26	0.29	0.32	0.33
25	0.13	0.15	0.18	0.21	0.23	0.25	0.26

	Minimum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.45	0.45	0.64	0.82	1.00	1.00	1.00
5	0.31	0.37	0.48	0.60	0.77	0.88	0.93
6	0.27	0.32	0.43	0.55	0.67	0.75	0.81
7	0.21	0.27	0.36	0.45	0.56	0.66	0.70
8	0.19	0.24	0.33	0.41	0.50	0.58	0.63
9	0.17	0.22	0.29	0.36	0.44	0.52	0.56
10	0.16	0.20	0.26	0.33	0.40	0.46	0.50
11	0.15	0.19	0.24	0.30	0.36	0.42	0.45
12	0.14	0.17	0.22	0.27	0.34	0.39	0.42
13	0.12	0.15	0.20	0.25	0.31	0.36	0.38
14	0.12	0.14	0.18	0.23	0.28	0.33	0.37
15	0.10	0.13	0.16	0.21	0.26	0.30	0.32
16	0.10	0.12	0.16	0.20	0.25	0.29	0.31
17	0.10	0.12	0.15	0.20	0.23	0.27	0.30
18	0.09	0.11	0.14	0.18	0.21	0.25	0.27
20	0.08	0.10	0.13	0.16	0.20	0.23	0.25
25	0.06	0.07	0.10	0.13	0.16	0.18	0.20

Finally, we have studied the efficiency of the maximum method and the minimum method through other simulation tests, in the sense of establishing which one best recovers a possible pyramidal structure underlying the initial data. In this sense, we would just point out that the results obtained in these tests show that the pyramidal method of the maximum generally is more efficient than the pyramidal method of the minimum.

REFERENCES

[1] **Baker, F.B., Hubert L.J.** (1975). "Measuring the power of hierarchical cluster analysis". *Journal of the American Statistical Association*, **Vol. 70 ,349**.

[2] **Bertrand P., Diday E.** (1985). "A visual representation of the compatibility between an order and a dissimilarity index: the pyramids". *Computational Statistics Quarterly*, **Vol. 2, Issue 1**, 1985, 31–41.

[3] **Bock, H.H.** (1984). "Statistical testing and evaluation methods in cluster analysis". *Proceedings of the Indian Statistical Institute Golden Jubilee*, International Conference on Statistics: Applications and New Directions.

[4] **Diday, E.** (1986). "Une représentation visuelle des classes empietantes: Les pyramides". *Rairo. Analyse des données*. **52,** 475–526.

[5] **Durand, C.** (1986). "Sur la représentation pyramidale en analyse des données". *Mémoire de DEA en Mathématiques Appliquées*. Université de Provence. Marseille.

[6] **Durand, C.** (1988). "Une approximation de Robinson inférieur maximale". *Rapport de Recherche. Laboratoire de Mathématiques Appliquées et Informatique*. **N88-02**. Université de Provence. Marseille.

[7] **Farris, J.S.** (1969). "On the cophenetic correlation coefficient". *Syst. Zoology*, **18(3)**, 279–285.

[8] **Fichet, B.** (1984). "Sur une extension de la notion de hiérarchie et son equivalence avec certaines matrices de Robinson". *Journées de Statistique*. Montpellier.

[9] **Goodman, L.A., Kruskal, W.H.** (1954). "Measures of association for cross-classification". *Journal of the American Statistical Association*. **Vol. 49, Dec.1954**, 732–764.

[10] **Hubert, L.J.** (1974). "Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures". *Journal of the American Statistical Association*, **Vol.69, 347.**

[11] **Jardine, N., Sibson, R.** (1971). *Mathematical Taxonomy*. Wiley, New York.