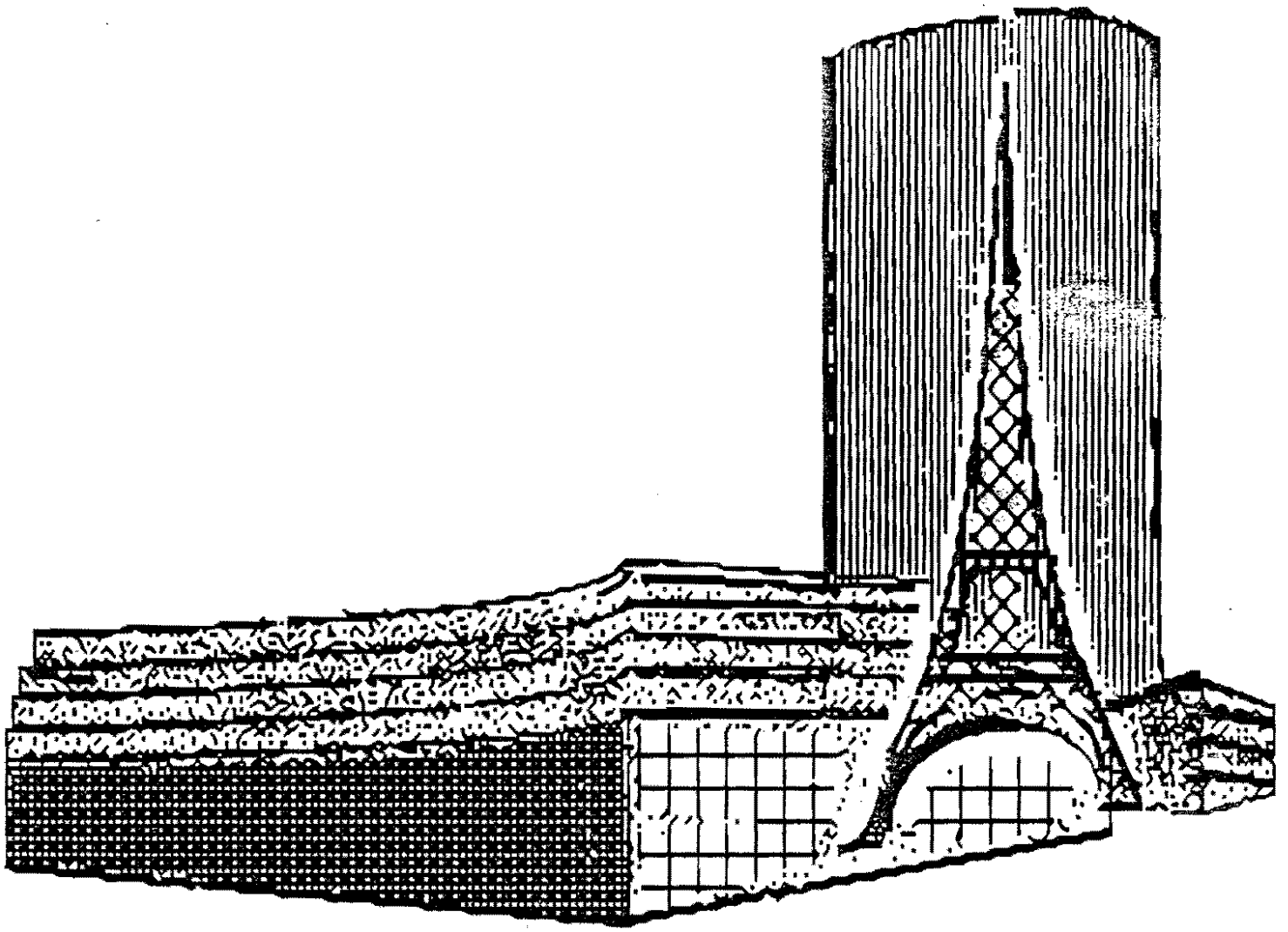
CLUB SAS® 92

Comptes rendus du club francophone d'utilisateurs du Système SAS®



Paris

7-9 octobre 1992

CHOIX DU NOMBRE D'AXES EN ANALYSE DES CORRESPONDANCES

Gilbert SAPORTA-CNAM

RESUME

Le choix du nombre d'axes en analyse factorielle et en particulier en analyse des correspondances est le plus souvent effectué en se servant de critères empiriques ou en faisant appel à l'expérience du praticien: détection d'un coude dans le diagramme de décroissance des valeurs propres, obtention d'un pourcentage d'inertie cumulée fixé plus ou moins arbitrairement à 80 %, interprétabilité des axes...
Or dans le cas de l'analyse factorielle des correspondances d'une table de contingence, il existe une procédure statistique encore méconnue qui permet de résoudre élégamment et efficacement ce problème: il s'agit du test d'ajustement du chi-deux entre le tableau analysé et sa reconstitution à l'aide de k axes proposé par E.Malinvaud, repris ultérieurement par E.Andersen et l'auteur.

I RAPPEL SUR LA FORMULE DE RECONSTITUTION

Etant donné un tableau de contingence **N** comprenant p lignes et q colonnes d'éléments n_{ij}, l'analyse des correspondances fournit un nombre de valeurs propres non triviales égal à min (p-1, q-1). On supposera pour la suite $p<q$. On notera a_{ik} et b_{jk} les coordonnées respectives des lignes et des colonnes sur l'axe n°k associé à la valeur propre μ_k.

Ces coordonnées sont normalisées par la relation $\Sigma\,(a_{ik})^2 = \Sigma\,(b_{jk})^2 = \mu_k$

On a alors la formule de reconstitution :

$$n_{ij} = (n_i n_j / n) (1 + \Sigma\, a_{ik}\, b_{jk} / \sqrt{\mu_k})$$

en sommant sur tous les axes.

On remarque que k=0 correspond à la situation d'indépendance; on a une reconstitution approchée $\tilde{n}_{ij}$ en se limitant à k axes.

II LES TESTS D'AJUSTEMENT

II.1 Le test usuel du chi-deux

Il consiste à comparer les n_{ij} observés sur un échantillon de n observations aux effectifs espérés sous l'hypothèse Hk de l'existence de k axes dans la population. Les estimateurs des moindres carrés pondérés de ces espérances sont alors donnés par les $\tilde{n}ij$ de la formule de reconstitution limitée à l'ordre k.

Le test usuel du chi-deux conduit alors à calculer la quantité:

$$Q_k = \sum_{i,j} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

Si k= 0, hypothèse d'indépendance, cette quantité Q_k est comparée à un chi-deux à (p-1)(q-1) degrés de liberté.

Si k=1 on la compare à un chi-deux à (p-2)(q-2) degrés de liberté. D'une façon générale on montre que ,sous l'hypothèse H_k, Q_k suit asymptotiquement une loi du chi-deux à (p-k-1)(q-k-1) degrés de liberté.

On effectuera donc une suite de tests du chi-deux en augmentant k jusqu'à ce que l'hypothèse H_k soit acceptée avec un risque d'erreur fixé. Cela revient concrétement à considérer l'écart entre le tableau et sa reconstitution comme un bruit aléatoire.

II.2 Le test modifié

Le test précédent implique de calculer les estimations $\tilde{n}_{ij}$ pour chaque valeur de k ce qui n'est pas une sortie usuelle de la PROC CORRESP ni d'autres logiciels d'analyse de données.

Si comme le suggére E.Malinvaud, on modifie les dénominateurs de Q_k , en remplaçant $\tilde{n}_{ij}$ par $n_i n_j / n$, les calculs ne nécessitent aucun investissement supplémentaire car la quantité modifiée :

$$Q'_k = \sum_{i,j} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\frac{n_{i.} n_{.j}}{n}}$$

est égale à n fois la somme des valeurs propres négligées:

$$Q'_k = = n(I - \mu_1 - \mu_2 - \ldots - \mu_k) = n(\mu_{k+1} + \mu_{k+2} + \ldots + \mu_{p-1}),$$

la démonstration de cette propriété est un exercice recommandé pour les statisticiens!

En pratique il n'y a que peu de différences entre Q et Q' et on procédera de la même façon qu'en II.1.

Des simulations extensives effectuées par L.Zater ont montré que ce test permet de reconnaitre la bonne dimension d'un tableau de données plus souvent que toute les méthodes empiriques.

III LE PROGRAMME CORRAXE ET UN EXEMPLE

Un programme SAS permettant d'effectuer le test Q à partir d'une table de contingence a été écrit par B.Dang et F.Tico , lors d'un projet de fin de cycle C au CNAM en 1991.En voici une illustration sur un exemple qui n'est d'ailleurs pas un tableau de contingence au sens strict puisqu'il s'agit d'un tableau donnant le nombre d'évocations de 19 items de qualité concernant 13 marques de beurres allégés, par un millier d'interviewés, le total n vaut 21900.

Tableau ASTRA.DAT:

```
269  70 69 223 14 21 153 118 165 168 23 36  89
178  74 46 138 12 13 128  90 158 131 20 23  82
124  22 25  84  6  7  70  46  86  61  6  7  22
184  95 74 184 12 26 158  96 162 229 20 31 138
214  80 59 192 18 25 168 114 177 172 21 31 102
201  65 32 153 15 17 115  90 138 130 13 22  76
110  58 30 105  8 13  98  55 114 105 12 15  55
243 115 68 217 20 21 231 138 227 247 33 43 113
303 137 95 286 24 39 271 165 251 327 36 51 146
253 117 77 244 20 31 210 132 217 282 26 43 124
121  60 35 117  8 18  98  65 101 134 15 21  95
 73  20 12  61 11  5  88  31  44  54  6  2  23
 86  46 29  88  9 12 146  38  82 112 11 15  49
158  74 39 127 10 13 121  85 149 175 18 19  84
240 113 98 216 21 33 196 134 197 276 26 45 124
 76  38 20  92  7 13  60  46  70  75  9 13  54
215  93 55 193 17 26 173 110 173 194 27 34  92
167  76 49 162 16 22 130  93 142 155 17 29  82
 85  51 27  82  7 10  77  43  87  83 12 13  49
```

Voici les valeurs propres et les pourcentages d'inertie associés:

```
μ1  = 0.0064   39.37%
μ2  = 0.0045   27.93%
μ3  = 0.0017   10.24%
μ4  = 0.0014    8.32%
μ5  = 0.0008    4.65%
μ6  = 0.0006    3.45%
μ7  = 0.0004    2.21%
μ8  = 0.0003    1.82%
μ9  = 0.0001    0.80%
μ10 = 0.0001    0.73%
μ11 = 0.0001    0.44%
μ12 = 0.0000    0.03%
```

n fois l'inertie est égal à 356.28 ce qui correspond à une valeur inacceptable pour un chi-deux à 12x18 = 216 degrés de liberté; on repousse l'hypothèse H_0, au moins un axe est nécessaire.

Les résultats du test Q sont fournis par le programme CORRAXE SAS:

TEST DE MALINVAUD
prendre le plus petit nombre d'axes tel que
le niveau soit superieur a 10%

nb d'axes a retenir	stat du test	nb de degres de liberte	niveau du test
1	215.357	187	0.07604
2	116.935	160	0.99569
3	82.249	135	0.99990
4	51.564	112	1.00000
5	35.017	91	1.00000
6	22.867	72	1.00000
7	14.476	55	1.00000
8	7.567	40	1.00000
9	4.586	27	1.00000
10	1.691	16	1.00000
11	0.121	7	1.00000

On choisit alors de ne retenir que deux axes.

Le test Q' donne des résultats comparables:

TEST Q'

nb d'axes a retenir	stat du test	nb de degres de liberte	niveau du test
1	214.84	187	0.08
2	115.33	160	0.9969
3	78.85	135	0.9999
4	49.21	112	1.0000

Les différences sont minimes entre les deux statistiques de test et conduisent à la même réponse.

Le programme permet également de calculer les reconstitutions successives du tableau de données avec 0 axe (hypothèse d'indépendance), 1 axe , 2 axes etc.
Voici la reconstitution avec les deux premiers axes que l'on comparera au tableau initial:

tableau de contingence approxime par 2 axe(s)

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	total
l1	264.0	79.0	58.6	209.6	16.6	21.5	147.6	122.6	189.1	167.5	21.4	30.9	89.7	1418.0
l2	179.9	66.5	46.2	153.3	12.7	17.6	125.9	88.3	140.3	145.0	17.1	24.3	75.7	1093.0
l3	121.5	25.6	20.2	87.6	8.3	6.8	70.5	50.4	78.3	54.4	8.1	9.9	24.4	566.0
l4	175.1	103.0	66.7	180.8	13.3	27.2	154.6	104.2	169.4	228.2	23.6	36.5	126.3	1409.0
l5	221.5	84.3	57.9	190.7	16.3	22.1	164.0	109.3	175.0	184.9	21.6	30.3	95.2	1373.0
l6	194.0	60.3	44.0	155.8	12.7	16.2	116.2	90.7	141.1	129.0	16.2	23.0	67.8	1067.0
l7	115.4	50.0	33.1	104.5	9.3	12.9	99.2	59.2	96.8	111.5	12.5	17.3	56.4	778.0
l8	251.4	109.9	71.7	228.3	21.4	27.9	232.3	128.0	212.3	247.1	27.7	37.0	121.2	1716.0
l9	303.7	140.5	91.9	282.2	24.9	36.1	273.0	159.5	262.5	314.2	34.6	48.1	159.7	2131.0
l10	253.1	118.3	78.0	236.1	19.9	30.8	216.0	134.5	219.3	262.8	28.9	41.3	137.1	1776.0
l11	114.8	64.0	42.0	115.6	8.3	17.0	93.6	67.1	107.8	141.0	14.8	23.0	78.9	888.0
l12	71.1	22.1	13.3	57.2	8.1	4.7	88.8	29.3	53.5	53.7	6.7	5.5	16.1	430.0
l13	83.4	48.4	27.3	85.7	11.5	10.9	141.8	43.5	82.8	116.6	12.2	12.7	46.2	723.0
l14	153.0	71.7	47.5	142.9	11.7	18.8	126.6	81.8	132.6	158.8	17.4	25.3	83.9	1072.0
l15	235.2	118.4	77.8	226.3	18.0	31.0	199.0	129.8	210.7	262.2	28.2	41.8	140.7	1719.0
l16	83.8	38.7	26.3	77.7	5.6	10.4	58.6	45.4	71.7	84.2	9.3	14.3	47.1	573.0
l17	216.5	88.1	59.3	191.2	16.7	22.9	174.6	108.8	176.4	195.2	22.3	30.9	99.2	1402.0
l18	174.6	73.2	49.7	155.9	12.7	19.3	131.0	89.7	143.6	160.6	18.2	26.4	85.1	1140.0
l19	88.0	41.9	27.5	82.7	7.1	10.9	77.5	47.0	77.0	93.4	10.2	14.5	48.4	626.0
total	3300.0	1404.0	939.0	2964.0	255.0	365.0	2691.0	1689.0	2740.0	3110.0	351.0	493.0	1599.0	21900

CONCLUSION

Les tests proposés, s'ils ne sont pas absolus, sont plus surs que les méthodes pifomètriques usuelles et peuvent être utilisés facilement. Le choix du risque d'erreur renferme certes une part d'arbitraire comme tout emploi de test.
Cependant ces tests ne marchent que pour des tableaux de comptage, et leur extension aux tableaux disjonctifs de l'analyse des correspondances multiples est encore à l'étude.

REFERENCES

E. Andersen, "*Statistical analysis of categorical data*", Springer Verlag, 1990

E. Malinvaud, "Data analysis in applied socio-economic statistics with special consideration of correspondence analysis ", *Marketing Science Conference*, Jouy en Josas, 1987

G. Saporta, " *Probabilités, analyse des données et statistique*" Technip, 1990

L. Zater, "*Contribution à l'étude de la variabilité des valeurs propres et au choix de la dimension en analyse des correspondances* " Thése de doctorat, Université Paris-Dauphine, 1989