6th International Symposium on Applied Stochastic Models
and Data Analysis
Hania, Crete, Greece, May 3-6 , 1993

ABOUT THE SELECTION OF THE NUMBER OF COMPONENTS IN CORRESPONDENCE ANALYSIS

Gilbert SAPORTA **Narcisa TAMBREA**

CEDRIC
Conservatoire National des Arts et Métiers
292 rue Saint Martin,
75141 Paris Cedex 03,France

ABSTRACT
Selecting the right number of axes in correspondence analysis is usually done by using empirical criteria such as :
- detection of an inflexion in the diagram of eigenvalues
- getting an arbitrary amount of the cumulated percentage of inertia

We examine the application of a chi-square goodness of fit test between the data table and its reconstitution with k eigenvalues. This test which has been proposed by E.Malinvaud, then by E.Andersen and G.Saporta has a good behaviour for frequency tables but fails to apply to multiple correspondence analysis. This failure, however enlightens some properties of this test and of correspondence analysis.

Keywords: correspondence analysis, eigenvalues, dimensionality

I THE RECONSTITUTION FORMULA FOR A CONTINGENCY TABLE

Let **N** be a contingency table with p rows and n columns of frequencies n_{ij}, coreespondence analysis provides r = min (p-1, q-1) non trivial eigenvalues. We will denote by a_{ik} et b_{jk} the coordinates of the rows and of the columns along the k th axis normalised by the relationship:

$$\sum_i \left(a_{ik}\right)^2 = \sum_j \left(b_{jk}\right)^2 = m_k$$

We then get the reconstitution formula, which is a weighted singular value decomposition of **N**:

$$n_{ij} = \left(n_{i.} n_{.j} / n\right)\left(1 + \sum_{i,j} a_{ik} b_{jk} / \sqrt{m_k}\right)$$

We may notice that k = 0 gives the independence table; we get the best approximation of rank k, $\tilde{n}_{ij}$, when using only the first k terms of the sum.

II GOODNESS OF FIT TESTS

II.1 The usual chi-square test

It consists in comparing the observed n_{ij} from a sample of size n to the expected frequencies under the hypothesis H_k of only k axes for the whole population(p_{ij} table) .Weighted least squares estimates ot these expectations are precisely the $\tilde{n}_{ij}$ of the reconstitution formula with its first k terms.
We then compute the classical chi-square statistic:

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

If k= 0, i.e the independence case, this quantity Q_0 is compared to a chi-square with (p-1)(q-1) degrees of freedom .

If k=1, Q_1 is compared to a chi-square with (p-2)(q-2) degrees of freedom. In the general case it is easy to proove that under hypothesis H_k, Q_k is asymptotically distributed like a chi-square with (p-k-1)(q-k-1) degrees of freedom.

So we perform a sequence of chi-square tests beginning with k = 0 until hypothesis H_k be accepted with a specified significance level. In other words we accept H_k if the difference between the data table and its reconstitution is not significantly different from a random noise.

II.2 A modified version

For the previous test, we need to compute the estimates $\tilde{n}_{ij}$ for each value of k which is not a standard output of CA software
If following E.Malinvaud, we use for the denominators of Qk ,$n_{i.}n_{.j}/n$ instead of ñij, les no special computations ares required since the modified test statistic

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

is equal to n times the sum of the discarded eigenvalues:

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

For tables with reasonably high frequencies there is only a slight difference between Q and Q' and the same sequence of chi-square tests than in II.1 may be applied.

Extensive Monte-Carlo experiments by L.Zater have shown that this test recovers the right dimension of a table more often than the other empirical techniques

II.3 example

The analyzed data table, which was not actually a real contingency table, gives the number of times where each of a thousand respondents associates an item (among 19) to 13 brands of diet butters. Due to multiple answers n=21900.

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

Here are the eigenvalues and the percentages of inertia

$\mu_1 = 0.0064$ 39.37%
$\mu_2 = 0.0045$ 27.93%
$\mu_3 = 0.0017$ 10.24%
$\mu_4 = 0.0014$ 8.32%
$\mu_5 = 0.0008$ 4.65%
$\mu_6 = 0.0006$ 3.45%
$\mu_7 = 0.0004$ 2.21%
$\mu_8 = 0.0003$ 1.82%
$\mu_9 = 0.0001$ 0.80%
$\mu_{10} = 0.0001$ 0.73%
$\mu_{11} = 0.0001$ 0.44%
$\mu_{12} = 0.0000$ 0.03%

n times the inertia is equal to 356.28 which is a too high value for a chi-square with 12x18 = 216 degrees of freedom; so the hypothesis H_0 is rejected, and at least one axis is necessary.

The following results lead clearly to keep 2 axes, which perfectly fits to the habits of marketing people!

k	Qk	Degrees of freedom	significance level
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

Q' gives similar results:

k	Q'_k	Degrees of freedom	significance level
1	214.84	187	0.08
2	115.33	160	0.9969
3	78.85	135	0.9999
4	49.21	112	1.0000

The computer program written with the SAS language by two students (B.Dang Tran et F.Tico) gives also the sequence of the approximations of **N**. Here is the approximation with two axes:

													total
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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.

Notice that all the approximations have the same margins than the data matrix.

III. SOME TRIALS FOR MULTIPLE CORRESPONDENCE ANALYSIS

Multiple correspondence analysis of p categorical variables with m_1, m_2,.. m_p categories is nothing else than usual correspondence analysis applied to the (n, Σ m_i) matrix of indicator variables (the so-called disjunctive table) **X** or to the Burt's table **B** = **X'X**.
Burt's table being a concatenation of all cross-tabulations, and the sum of its eigenvalues being related to all the chi-square measures of departure from independence, the first idea was to apply the chi-square test presented here to **B** rather than to **X** since the approximation of a matrix filled with 0 and 1 leads to special problems.

We used for our experiments a real-life data set of 11 variables with respectively 2,4,3,4,4,4,2,4,5,6,3 categories (41 in the whole) observed upon 308 units.
The number of non-trivial eigenvalues is thus equal to 30.

At eye a jump may be detected after the first two axes.

k	eigenvalue	inertia %	cumulative inertia	diagram of eigenvalues
1	0.036053	12.43	12.43	_____________________________________
2	0.029648	10.22	22.66	_______________________________
3	0.020160	6.95	29.61	_____________________
4	0.018235	6.28	35.90	___________________
5	0.016864	5.81	41.72	_________________
6	0.014471	4.99	46.71	_______________
7	0.014132	4.87	51.58	_______________
8	0.012439	4.29	55.87	_____________
9	0.012310	4.24	60.12	_____________
10	0.011316	3.90	64.02	____________
11	0.010244	3.53	67.56	___________
12	0.009832	3.39	70.95	__________
13	0.009451	3.25	74.21	__________
14	0.007957	2.74	76.95	________
15	0.007768	2.67	79.63	________
16	0.007222	2.49	82.12	_______
17	0.006763	2.33	84.46	_______
18	0.006058	2.08	86.55	______
19	0.005566	1.91	88.47	______
20	0.004858	1.67	90.14	_____
21	0.004523	1.56	91.70	_____
22	0.004267	1.47	93.17	____
23	0.003774	1.30	94.48	____
24	0.003286	1.13	95.61	___
25	0.002802	0.96	96.58	___
26	0.002592	0.89	97.47	___
27	0.002150	0.74	98.21	__
28	0.001877	0.64	98.86	__
29	0.001773	0.61	99.47	__

30	0.001523	0.52	100.00	__

III.1 Approximations of the complete Burt's table

Here is the list of values of the test statististics Q_k and Q'_k:

k	Q_k	Q'_k
0	10804.52	10804.52
1	7898.63	9460.91
2	5326.73	8356.00
3	4808.80	7604.69
4	5057.26	6925.12
5	4031.73	6296.64
6	4073.94	5757.34
7	2868.33	5230.66
8	4370.22	4767.10
9	11460.66	4308.33
10	2444.09	3886.62
11	5367.80	3504.85
12	485.04	3138.42
13	547.68	2786.18
14	2046.96	2489.62
15	969.23	2200.12
16	1241.42	1930.99
17	942.12	1678.93
18	577.63	1453.14
19	2037.66	1245.69
20	-2351.46	1064.66
21	-1567.51	896.10
22	548.17	737.07
23	623.76	596.42
24	720.79	473.97
25	435.80	369.56
26	2382.90	272.95
27	93.80	192.83
28	98.84	122.86
29	37.54	56.78
30	0.00	0.00

The remarkable and disappointing feature is that the behaviour of Q_k is not monotonic and even takes negative values. This is due to the diagonal blocks of **B**. Since they are diagonal and contain the marginal frequencies of the variables, the approximations of the zeros are in some respects difficult and give some time negative values. The consequence is that the denominators of Q_k may be very small or negative giving inappropriate values for a chi-square.

The values of Q'_k are more satisfactory but they decrease very slowly. The comparison with a chi-square is not relevant however ,because the Burt's table being symmetric, the subarrays are counted twice . Problems with small values may also occurr in contingency tables and since the modified chi-square Q'_k is less sensitive to this phenomenon, it is certainly preferable to Q_k

III.2 Approximations of a half Burt's table

The second attempt to evaluate the approximation of **B** by k axes was to consider only the p(p-1) upper blocks of **B.** Here are the values of both statistics Q_k and Q'_k:

k	Q_k	Q'_k
0	782.26	782.262
1	698.41	672.562
2	143.96	581.456
3	334.41	590.556
4	709.02	596.225
5	522.91	615.386
6	740.67	618.754
7	284.11	636.605
8	1182.12	648.825
9	4845.17	648.632
10	452.43	655.125
11	2009.92	655.822
12	-356.07	632.389
13	-245.42	599.383
14	556.24	578.680
15	80.98	533.695
16	267.84	505.081
17	162.43	461.973
18	7.92	415.608
19	774.75	377.015
20	-1390.42	326.520
21	-971.49	284.893
22	112.54	237.191
23	183.20	196.155
24	250.52	161.617
25	132.12	131.376
26	1124.84	99.648
27	-4.22	70.648
28	21.93	47.585
29	5.31	22.292
30	0.00	0.000

It is still impossible to interpret the values of Q_k, since they are not decreasing nor positive. Q'_k suffers also from a slight non monotonicity. and has in the average a very low rate of decrease. The explanation of the non monotonicity here is that there are cells with small frequencies : the approximation for all cells is not monotonic and this time there no compensation due to the diagonal blocks.
The degree of freedom for Q'_0 is easy to calculate : it is equal to :

$$\sum_{i>j}(m_i-1)(m_j-1)=396$$

Despite the fact that it is not clear which degree of freedom we have to use when k is greater than zero, we may use the 5 % percentile of a chi-square with 396 df as an indicator of the goodness of fit of the approximation of **B.** Since this percentile is equal to 442 , we may see that at least 19 axes are necessary which shows how difficult it is to approximate **B** and that this kind of approach might be irrelevant.

*III.3 Approximation of the disjunctive table **X***

Since a direct approximation of $\mathbf{X}=\left(\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_p\right)$ by k axes is meaningless we transformed the approximated table $\mathbf{X}^{(k)}$ into the closest disjunctive table $\hat{\mathbf{X}}^{(k)}$ as follows:
for each variable s=1,...,p and for $\sum_{t=1}^{s-1} m_t+1\le j_0\le\sum_{t=1}^{s} m_t$, we put

$$\hat{x}_{ij_0}^{(k)}=\begin{cases}1; \text{ if } & x_{ij_0}^{(k)}=\max\limits_{\sum_{t=1}^{s-1} m_t+1\le j\le\sum_{t=1}^{s} m_t} x_{ij}^{(k)}\\ 0; & \text{otherwise}\end{cases}$$

where $1\le i\le n$.
To compare the two tables $\mathbf{x}$ and $\hat{\mathbf{X}}^{(k)}$ we counted the differences:

$$D^k=\frac{1}{2}\cdot\sum_{i,j}\left|x_{ij}-\hat{x}_{ij}^{(k)}\right|$$

For k=0 the upper relationship is:

$$D^0=\sum_{s=1}^{p}\left(n-\hat{n}_s\right)$$

where $\hat{n}_s$ is the maximal marginal frequencies of the s variable since the 0-order approximation of each cell is equal to the marginal frequency of the corresponding category.
We can, also, compute the differences for each variable s=1,...,p if we count only for $\sum_{t=1}^{s-1} m_t + 1 \le j \le \sum_{t=1}^{s} m_t$.

Here is the list of the differences $D_1^k + D_2^k + \ldots + D_p^k = D^k$:

k	D_1^k	D_2^k	.	.	.	.	.	.	.	.	D_{11}^k	D^k
0	82 +	195 +	99 +	161 +	131 +	177 +	50 +	122 +	221 +	216 +	94 =	1548
1	77 +	180 +	98 +	105 +	125 +	100 +	50 +	122 +	219 +	189 +	89 =	1354
2	76 +	168 +	89 +	97 +	120 +	89 +	47 +	122 +	172 +	161 +	92 =	1233
3	48 +	128 +	85 +	95 +	121 +	89 +	38 +	120 +	146 +	158 +	86 =	1114
4	36 +	115 +	90 +	66 +	89 +	91 +	35 +	121 +	142 +	132 +	79 =	996
5	36 +	90 +	75 +	62 +	91 +	91 +	40 +	107 +	122 +	106 +	75 =	895
6	36 +	78 +	67 +	57 +	77 +	91 +	40 +	106 +	109 +	100 +	75 =	836
7	35 +	74 +	51 +	43 +	77 +	90 +	32 +	104 +	98 +	93 +	74 =	771
8	36 +	70 +	48 +	37 +	67 +	89 +	32 +	67 +	100 +	83 +	70 =	699
9	36 +	64 +	45 +	34 +	66 +	80 +	32 +	40 +	93 +	75 +	71 =	636
10	31 +	51 +	33 +	31 +	64 +	69 +	31 +	38 +	66 +	70 +	55 =	539
11	35 +	33 +	27 +	29 +	60 +	59 +	27 +	27 +	63 +	55 +	34 =	449
12	23 +	26 +	30 +	32 +	51 +	33 +	29 +	15 +	61 +	41 +	26 =	367
13	22 +	26 +	16 +	27 +	39 +	20 +	29 +	16 +	49 +	29 +	10 =	283
14	19 +	22 +	14 +	21 +	38 +	19 +	11 +	16 +	37 +	28 +	10 =	235
15	19 +	24 +	10 +	15 +	19 +	18 +	10 +	5 +	27 +	21 +	11 =	179
16	12 +	26 +	12 +	14 +	12 +	16 +	1 +	6 +	17 +	10 +	9 =	135
17	10 +	14 +	6 +	14 +	4 +	15 +	1 +	7 +	7 +	8 +	5 =	91
18	10 +	17 +	6 +	13 +	3 +	15 +	1 +	3 +	4 +	6 +	4 =	82
19	7 +	15 +	4 +	15 +	3 +	8 +	1 +	1 +	1 +	4 +	2 =	61
20	7 +	9 +	3 +	12 +	3 +	8 +	0 +	0 +	1 +	3 +	2 =	48
21	5 +	6 +	2 +	10 +	1 +	6 +	0 +	0 +	0 +	1 +	1 =	32
22	5 +	6 +	1 +	6 +	0 +	5 +	0 +	0 +	0 +	0 +	0 =	23
23	2 +	1 +	0 +	5 +	0 +	6 +	0 +	0 +	0 +	0 +	0 =	14
24	3 +	1 +	1 +	6 +	0 +	3 +	0 +	0 +	0 +	0 +	0 =	14
25	0 +	0 +	0 +	5 +	0 +	4 +	0 +	0 +	0 +	0 +	0 =	9
26	0 +	0 +	0 +	5 +	0 +	0 +	0 +	0 +	0 +	0 +	0 =	5
27	0 +	0 +	0 +	4 +	0 +	0 +	0 +	0 +	0 +	0 +	0 =	4
28	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 =	0
29	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 =	0
30	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 =	0

The values of D^k decrease very slowly and the empirical criteria about the detection of an inflexion in the diagram of D^k does not give conclusive results. If we apply the same criteria for each diagram D_s^k and consider the maximal number of the axes, we need at least 10 axes.

CONCLUSION

The modified chi-square statistic Q'_K has a good behaviour for contingency tables . However one has to be careful when some frequencies are low. On the other hand, the application to multiple correspondence analysis is disappointing.

A possible interpretation is that MCA is not an adequate method to approximate either Burt's table (see Greenacre 1991) or a disjunctive table, but should be considered from an other point of view.

REFERENCES

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

M.Greenacre, "Interpreting multiple correspondence analysis", *Applied Stochastic Models and Data Analysis*, 7, 195-210, 1991

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