Vol.2021-CH-125 No.9
2021/2/13

対応分析を巡る諸問題について

矢野　環 [1]

概要：　文化事象の分析には、様々な手法が使われている。特に、非線形な関係の解明に役立つ対応分析は有効と考えられている。即ち、線形な関係を解析する主成分分析系に比較して、ということである。しかし、（多重）対応分析にはまた様々な問題点も指摘されており、その検討も十分ではないと思われる。そこで、実例を中心に、取り扱い方を提起したい。

キーワード：対応分析、多重対応分析、カイ２乗検定、調整済残差、TIA（全情報解析）、クラメールの V

1. はじめに

対応分析・多重対応分析は、様々な変形や提案がなされている。R には ca, cocoresp, CAinterprTools, CAvariants や MCAvariants などのパッケージが提供されている。しかし、最も利用されている FactoMineR においてすら、結果の表示を変更した方が良いものも認められる。更には、以前からの西里静彦氏の Total Information Analysis（TIA）の指摘もきちんと検討されているとは思えない。剰え、不適切なグラフとなる biplot(corresp(data))まで未だに使われている。biplot のグラフを採用するにしても、その配置の安定性も検討すべきであろう。MCA の biplot を使用するのであれば、必要に応じて適切な手法を採用するのがよいであろう。さらに、西里により提案された Cramer's V の利用も検討すべきであると思われる。

2. 対応分析、配置の信頼度

対応分析は、分割表（Contingency Table）の分析に使われる。手法としては、分割表の残差を用いるものであり、カイ２乗検定と共通する。

分割表は、母集団からの（ランダム）標本と見做せる場合もあるが、一方ではある現象の記述であり母集団を想像し難い場合もある。対応分析を行った結果は、低次元のグラフで表示する場合が多い。その手法の問題点は夙に西里静彦等によって指摘されており、後に触れる。まずは、典型的実例について、配置の信頼度の表示を検討する。これは ellipseCA として FactoMineR に実装されているのだが、大学院生の研究などでもあまり実用に供されていない。

実例のデータとして、HairEyeColor の男女 529 人を合併した haireye を用いる。目の色（Hazel 等）髪の色（Black 等）の 4×4 行列である。図１は、ellipseCA(CA(haireye))の結果である。信頼楕円は、髪の Brown と Red 以外はよく分離している。

データによっては、10000 件程度無いと安定しない場合もある。表２は化粧品と使用年代の 8000 件であり、斎藤・豊田[9]が古くにこの問題を取り上げた例である。元々 24798 件であり、約三分の一に縮小したデータである。図２に見えるように、楕円の大きさは列和（行和）の大きさの逆順に成っていることを斎藤・豊田が指摘している。なお、行変数と列変数の見かけの近さと、正の大きな残差が対応していることを見やすいように、網掛した。Hair-Eye で特に Blond-Blue, Black-Brown がグラフ上でも近接している。

表１　haireye　data と 残差

Hair\Eye	Brown	Blue	Hazel	Green	Brown	Blue	Hazel	Green
Brown	119	84	54	29	1.23	-1.95	1.35	-0.35
Blond	7	94	10	16	-5.85	7.05	-2.23	0.61
Black	68	20	15	5	4.4	-3.07	-0.48	-1.95
Red	26	17	14	14	-0.07	-1.73	0.85	2.28

図１　ellipseCA(CA(haireye))

表２　cosme8

cosme8	u24	f25.39	f40.54	f55.69	f70	sum
Fancl	1999	522	373	217	59	3170
Shiseido	1655	245	251	358	206	2715
Kose	426	126	102	75	24	753
Kanebo	390	73	78	114	67	722
MaxFactor	330	89	103	90	28	640
sum	4800	1055	907	854	384	8000

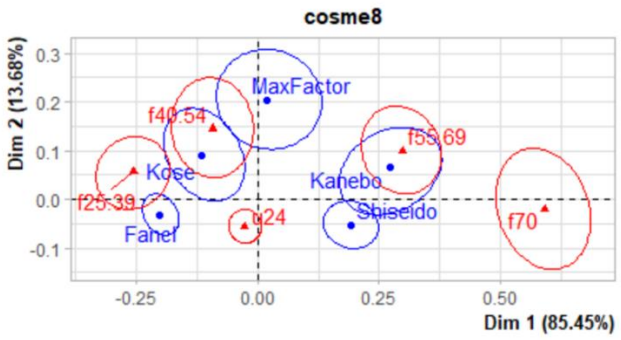

図２　cosme8

1 同志社大学名誉教授、埼玉大学名誉教授
同志社大学人文科学研究所　嘱託研究員（外部）

3. 対応分析、調整済残差

現在は、独立性のカイ２乗検定が有意である時に、元データのどのセルが独立性から乖離しているかを見ると解釈されている。原論文は、検定が有意かどうかはともかく、調整済残差を見ることにより意味のあるセルを特定できるという趣旨である。実際、例となっているデータ hab は、４つのコンプレッサー cp1~4 のどの部分(North, Center, South)で故障が発生するかの集計表であり、表３にデータと２通りの残差を表示した。独立性検定のカイ２乗値は 0.068 であり、有意ではない。

表 3 Haberman の例[1]。残差、調整済残差

	N	C	S	N	C	S	N	C	S
cp1	17	17	12	0.6	1.67	-1.78	0.86	2.27	-2.78
cp2	11	9	13	0.14	0.3	-0.35	0.19	0.38	-0.52
cp3	11	8	19	-0.33	-0.45	0.62	-0.45	-0.59	0.94
cp4	14	7	28	-0.42	-1.47	1.46	-0.6	-2.01	2.32

残差では、絶対値が 1.96 を超える有意なセルはない。しかし、調整済残差を見れば、４つのセルで有意となる。

対応分析 CA は残差を元にしている。では、調整済残差を元にして同様の分析 CAadj を行うと、より有用な結果が得られるであろうか。実際に２通りの処理を行った結果を同時に表示したのが図３上である。この場合は左程の相違はない。実際、プロクラステス変換を行ってみると、下図の通りほぼ重なっている。つまり、配置としてはほぼ同じである。実際この例は殆ど１次元的である。

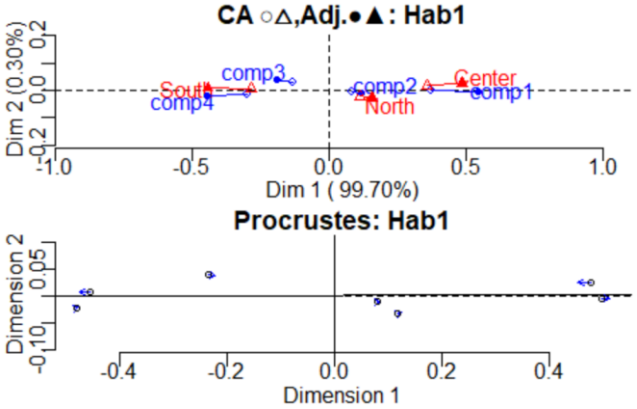

図 3　CA, CAadj 比較と、プロクラステス変換

4. TIA 全情報解析　Total Information Analysis

西里等は、対応分析において第 j 次元での行変数空間と列変数空間は、 ρ を第 j 次特異値とするとき、arccos(ρ)の角度を成しており、決して同一空間に biplot すべきものではないことを注意した。特異値が 1 に近ければほぼ同じ方向であるが、0 に近ければ直交する程の差異がある。これに伴い、行変数あるいは列変数同志の距離を”within”, 行変数と列変数の間の距離を”between” として、それを総合した距離を super-distance として扱うことを提唱した[2~8]。対応分析で有効なのが k 次元であるとすれば、この距離は 2k 次元空間の距離である。それを用いる解析を、TIA（全情報解析）と呼ぶ。距離は MDS（多次元尺度法）や、クラスター分解に使用することが推奨されている。

実際の例で見てみよう。先の haireye に適用して、距離を求め、MDS の結果を通常の biplot と比較すれば、次の通り。

表 4 haireye の super-distance

	Black	Brown	Red	Blond	Brown	Blue	Hazel	Green
Black	0	0.45	0.65	1.35	0.57	0.93	0.55	0.74
Brown	0.45	0	0.32	0.99	0.46	0.64	0.28	0.44
Red	0.65	0.32	0	1.04	0.57	0.71	0.41	0.51
Blond	1.35	0.99	1.04	0	1.15	0.77	0.98	0.86
Brown	0.57	0.46	0.57	1.15	0	1.04	0.4	0.78
Blue	0.93	0.64	0.71	0.77	1.04	0	0.81	0.58
Hazel	0.55	0.28	0.41	0.98	0.4	0.81	0	0.45
Green	0.74	0.44	0.51	0.86	0.78	0.58	0.45	0

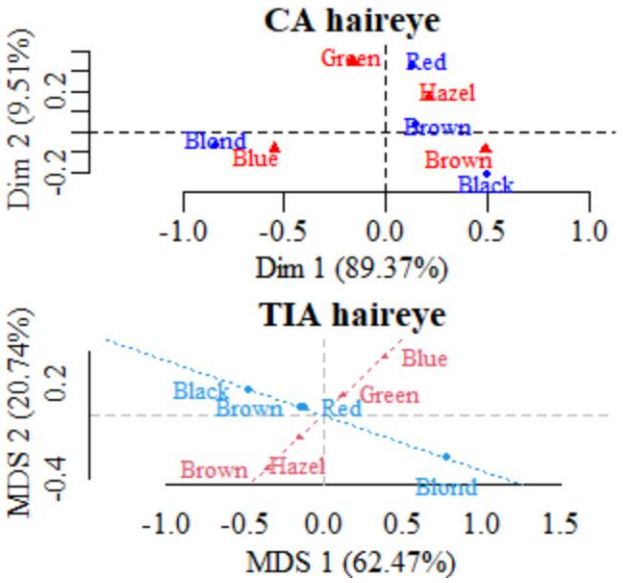

図 4　通常の biplot と super-distance の MDS の 1-2 軸

ここで図４では、行変数と列変数の１次元表示が角度を成して現れている。本来の角度は 62.81° であり、ほぼその程度に見えている。そして、最大３次元（実質２次元）が６次元（実質４次元）に展開されたが、その 1-3 軸でのグラフが、元の biplot にほぼ等しいことが解る（図５）。

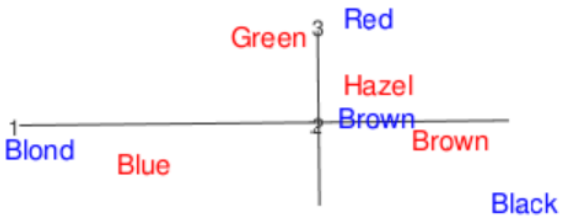

図 5　6 次元空間内の、1-3 軸での配置。

しかし、先の結果での Hair-Eye で特に残差に応じてグラフ上で近接していた Blond-Blue, Black-Brown を見ると、Blond-Blue は確かに最小距離であるが、Black-Brown よりもむしろ Black-Hazel の方が短い距離となっている。このように微妙な差異があり、全体像が単純な biplot よりも複雑化する。しかし、TIA は常に考慮すべきであると考える。

勿論、TIA-MDS における 1-2 次元に、元の第１次元の状況がいつでも現れるというわけではない。また、行変数の配置と列変数の配置が図５のようにどこかに簡単に現れるというわけでもない。

5. MCA での配置例

西里も低次元 biplot を用いることがある。図 6 は[3]の p.14 Fig.1.5 の元図である。但し、行変数の配置を変更した。

表 5　西里の良く使うデータ

	血圧	偏頭痛	年齢	不安度	体重	身長
1	a	c	c	c	a	a
2	a	c	a	c	b	c
3	c	c	c	c	a	c
4	c	c	c	c	a	a
5	b	a	b	b	c	b
6	b	a	b	c	c	a
7	b	b	b	a	a	c
8	a	c	a	c	a	c
9	b	b	b	a	a	b
10	a	c	b	b	a	c
11	b	a	a	c	b	b
12	b	b	c	c	b	b
13	c	c	c	c	c	a
14	a	c	a	b	a	a
15	c	c	c	c	a	b

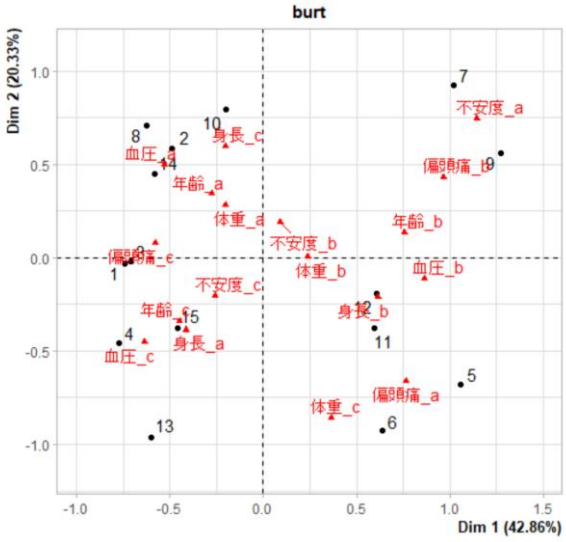

図 6　plot(MCA(nishia, method="Burt", graph=F))

変数「不安度」が a であるのは 7,9 の 2 件であり、図 5 でのその 2 点の平均に「不安度_a」がある。同様に、「偏頭痛_a」は、5,6,11 の 3 点の平均となっている。「年齢_a」についても同様。これは CA 全般で良く知られている一種の duality である。

6. Cramer's V

名義尺度の変数は MCA で取り扱える。さらに Likert scale も、事前に間隔が指定された数値として扱うのではなく、また事前に順序の指定された順序尺度でもなく、名義尺度として取り扱うべきであるというのが、西里の意見である。

そして、2 つの名義尺度変数における「相関係数」に相当するものとして、Cramer's V, 即ちクラメールの連関係数がある。その有効性を見るために、典型的なデータとして psych の bfi を用いて検証する。bfi は、BigFive 性格特性 ACENO に関わる 5×5=25 変数と付加的な変数からなっている。各変数は 6 段階の Likert scale 1,2,..,6 からなる。通常はその相関係数、あるいはポリコリック相関係数によって取り扱われる。しかし、名義尺度として Cramer's V を計算することができる。その結果を相関係数として取扱い、因子分析を行うこともできる。実際、図 6 のように 5 因子が明確に現れる。斜交変換 oblimin でも勿論明確に出る。

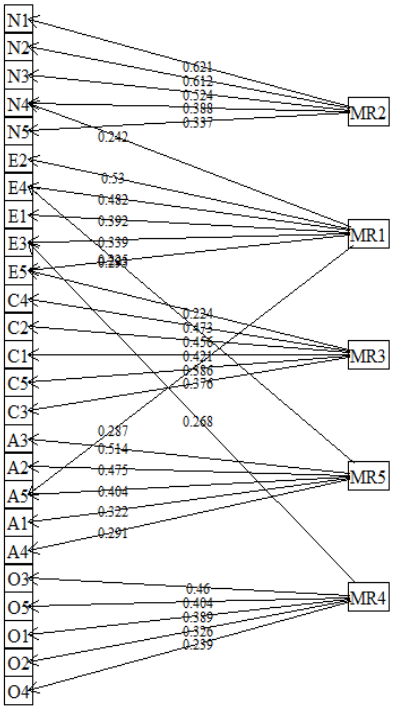

図 7　Cramer's V の行列からの因子分析(varimax)

参考文献

[1] Haberman, S. J.: The Analysis of Residuals in Cross-Classified Tables, Biometrics, 29, 205-220, 1973.
[2] Nishisato, S.: Analysis of categorical data: Dual scaling and its applications. University of Toronto Press. 1980.
[3] Nishisato, S.: Dual Scaling, in Chap.1 of D.Kaplan (eds). The Sage Handbook of Qualitative Methodology for the Social Sciences. Sage Publ., 3-23. 2004.
[4] Nishisato, S.; A New Framework for Multidimensional Data Analysis, in Weihs, C. and Gaul. W. (eds.), Classification –the Ubiquitous Challenge, Heidelberg, Springer, 280-287. 2005.
[5] **西里静彦**　データ解析への洞察　数量化の存在理由 K.G.リブレット No.18 関西大学出版会, 2007
[6] Nishisato, S., & Clavel, J.C.: Total information analysis: Comprehensive dual scaling. Behaviormetrika, 37, 15-32, 2010.
[7] 西里静彦: 測度のデータへの回帰による最適データの生成、データ分析の理論と応用, 1(1), 1-10, 2011.
[8] 西里静彦: 行動科学への数理の応用：探索的データ解析と測度の関係の理解、行動計量学 41(2), 89-102, 2014.
[9] 齋藤朗宏,豊田秀樹: コレスポンデンス分析における布置の精度, オペレーションズ・リサーチ:経営の科学 49(3), 168-173, 2004.