7. 分析

データ行列全体をさまざまな視点から分析します。

7.1. 統計量の分析

7.1.1. 範囲の分析

四分位と範囲を使ってデータの範囲の状態を調べます。

X1	v1	v2	v3
h1	38	18	5
h2	35	10	6
h3	28	44	48
h4	22	30	62
h5	24	29	89

X1	v1	v2	v3
最小値	22	10	5
第 1 四分位値(25%)	24	18	6
中央値(50%)	28	29	48
第 3 四分位値(75%)	35	30	62
最大値	38	44	89
範囲	16	34	84

7.1.2. 中心の分析

データの中心を示す平均値、中央値、最頻値と大数平均値、大数最頻値と、平均値・中間値との関係を差・両側相対値・範囲内位置で調べます。たとえば、平均値(Me)と中央値(Md)の両側相対値(Contr)は

$$Contr(Me, Md) = (Me - Md) / (Me + Md)$$

また、たとえば平均値(Me)の「範囲内位置」(Pos.Rg)は次のようにして求めます(Min:最小値; Max:最大値)。

$$Pos.Rg(Me) = (Me – Min) / Max$$

X1	v1	v2	v3
平均値	29.40	26.20	42.00
中央値	28.00	29.00	48.00
差(-)	1.40	-2.80	-6.00
両側相対値	.02	-.05	-.07
中間値	30.00	27.00	47.00
差(-)	-.60	-.80	-5.00
両側相対値	-.01	-.02	-.06
範囲内位置	.46	.48	.44

X1	v1	v2	v3
中央値	28.00	29.00	48.00
平均値	29.40	26.20	42.00
差(-)	-1.40	2.80	6.00
両側相対値	-.02	.05	.07
中間値	30.00	27.00	47.00
差(-)	-2.00	2.00	1.00
両側相対値	-.03	.04	.01
範囲内位置	.38	.56	.51

X1	v1	v2	v3
大数平均値	29.11	26.33	41.56
中央値	28.00	29.00	48.00
差(-)	1.11	-2.67	-6.44
両側相対値	.02	-.05	-.07
中間値	30.00	27.00	47.00
差(-)	-.89	-.67	-5.44
両側相対値	-.02	-.01	-.06
範囲内位置	.44	.48	.44

X1	v1	v2	v3
大数最頻値	24.67	25.67	66.33
中央値	28.00	29.00	48.00
差(-)	-3.33	-3.33	18.33
両側相対値	-.06	-.06	.16
中間値	30.00	27.00	47.00
差(-)	-5.33	-1.33	19.33
両側相対値	-.10	-.03	.17
範囲内位置	.17	.46	.73

7.1.3. 変動の分析

平均を中心にした変動を示す各種の統計量を比較します。

X1	v1	v2	v3
h1	38	18	5
h2	35	10	6
h3	28	44	48
h4	22	30	62
h5	24	29	89

X1	v1	v2	v3
分散	38.240	133.760	1062.000
標準偏差	6.184	11.565	32.588
変動係数	.210	.441	.776
規定標準偏差	.105	.221	.388
均等度	.895	.779	.612

7.1.4. 平衡の分析

データ X1 の縦列の数値の偏りを示す指標として、以下のような両側相対値を考えます。たとえば、v1 {38, 35, 28, 22, 24)の中間値[（最大値+最小値）/2]は(38 + 22) / 2 = 30 ですが、この中間値より大きな数値(Positive: Ps)は 38, 35 の 2 数です。また、中間値より小さな数値(Negative: Ng)は 28, 22, 24 の 3 数です。そこで、**中間値平衡度数**(Mid Balance Count: Mid.BC)は

$$Mid.BC = (Ps - Ng) / (Ps + Ng) = (2 - 3) / (2 + 3) = -.20$$

となり、ややデータ数が中間値より下に多いことがわかります。

次に、データの数ではなく、次のような数値を計算し、その結果を**中間値平衡値**(Mid Balalance Value: Mid.BV)とします。たとえば、v1 {38, 35, 28, 22, 24)のなかで中間値 30 より大きな数値 38, 35 の差は、8, 5 なので、Ps=8+5=13 になります。また、中間値 30 より小さな数値 28, 22, 24 の差は、2, 8, 6 なので、Ng=2+8+6=16 になります。そこで

$$Mid.BC = (Ps - Ng) / (Ps + Ng) = (13－16) / (13+16) = -.10$$

となります。

X	v1	v2	v3
中間値	30	27	47
正数	2	3	3
負数	3	2	2
平衡度	-.200	.200	.200
規定歪度	.139	.079	.057
規定尖度	.434	.599	.469

X	v1	v2	v3
中間値	30	27	47
正値	13	22	58
負値	16	26	83
平衡度	-.103	-.083	-.177
規定歪度	.139	.079	.057
規定尖度	.434	.599	.469

平衡度数も平衡値も、Ps と Ng が同じ数値であればゼロになり、Ps>Ng のときは正値になり、Ps<Ng のときは負値になります。どちらもは–1 < Ccm / Cvm< +1 の両端を含まない範囲をとります。

このようにそれぞれのデータと比較する参照値は、中間値だけでなく、平均値(Mean)や中央値(Median)を使うことができます。平均値を使うと平衡値がかならずゼロ(0)になるので平衡度数を使います。逆に、中央値を使うと平衡度数がかならずゼロ(0)になるので平衡値を使います。

X	v1	v2	v3
平均値	29	26	42
正数	2	3	3
負数	3	2	2
平衡度	-.200	.200	.200
規定歪度	.139	.079	.057
規定尖度	.434	.599	.469

X	v1	v2	v3
中央値	28	29	48
正値	17	16	55
負値	10	30	85
平衡度	.259	-.304	-.214
規定歪度	.139	.079	.057
規定尖度	.434	.599	.469

この中央値平衡値は先述の平衡指数です（→「統計量」「歪度」）。

7.1.5. 推移の分析

データの並びの推移の様子を、振動性・平衡性・単峰性・正規性・連続性・平滑性・定常性を示す係数で数量化します。

X1	v1	v2	v3
h1	38	18	5
h2	35	10	6
h3	28	44	48
h4	22	30	62
h5	24	29	89

X1	v1	v2	v3
振動性	-.778	.193	1.000
平衡性	.259	-.304	-.214
単峰性	.778	.719	1.000
正規性	.965	.738	.930
連続性	.438	.496	.426
平滑性	.875	.763	1.000
定常性	.691	.446	.691

7.2. 集中分析

データ行列の横行と縦列に与えられる数値情報に従って、拡散した行列の分布パタンを再編成し、行列の対角線に近い位置に高い数値を集中化することによって、データ全体の最適な分布形態を探る技法を「集中分析」(Concentration Analysis)と名づけます。私たちが開発した「原点偏差法」(Zero Deviation: ZD)と一般の多変数解析を利用する方法を比較します。

7.2.1. 外的基準

はじめに「原点偏差法」(ZD)の「外的基準による集中化」(Concentration by exterior criterion)を説明します。この方法の目的は、たとえば下左表のようなデータ行列があり、これの横行(h1, h2, ..., h5)を一定の基準に従って並べ替えてプラス(+)で示した反応を対角線に近い位置に並べることです。

Lv	v1	v2	v3	v4
h1	+	+		
h2			+	
h3		+		
h4			+	+
h5	+	+	+	

→

Lv	v1	v2	v3	v4
h1	+	+		
h3		+		
h5	+	+	+	
h2			+	
h4			+	+

このように集中化すると横行に関しては(h1, h3, h5)と(h2, h4)がそれぞれ集中し、縦列に関しては(v1, v2)と(v3, v4)がそれぞれ集中化されていることがわかります。ここで言う「集中化」(concentration)とは反応の分布が互いに近接して全体で一定の傾向を示すことを意味します。上右図の左上のグループは横行(h1, h3, h5)が縦列(v1, v2)を選択していることを示します。右下のグループ(h2, h4; v3, v4)も同様です。そして、h1 - h3 - h5 - h2 - h4 という行の順序が v1 - v2 - v3 - v4 という列の順序に対応していることをデータ«+»の分布が裏付けています。このようにすれば、原データ行列では見えにくい行と列の関係を集中化した行列全体を見ながら観察できるようになります。

全体の横行を並べ替えるために、それぞれの行の原点からの遠隔度(farness)を次のようにして探ります。たとえば h1 は v1 と v2 に反応しているので(1, 2)の位置に反応点がある、と考えられるので、その遠隔度（＝原点偏差 ZD）は[1]

$$ZD(h1) = [(1^2 + 2^2) / 2]^{1/2} = 1.581 \quad (...1)$$

[1] 統計値の「分散」を参照。分散を計算するときはそれぞれのデータと、データの平均からの偏差を計算しますが、原点偏差法では平均の代わりに原点(0)から偏差を計算します。

となります。他の列については、それぞれ(3), (2), (3, 4), (1, 2, 3)の位置に反応点(+)があるので

$$ZD(h2) = [(3^2) / 1]^{1/2} = 3.000 \quad (\ldots4)$$
$$ZD(h3) = [(2^2) / 1]^{1/2} = 2.000 \quad (\ldots2)$$
$$ZD(h4) = [(3^2 + 4^2) / 2]^{1/2} = 3.535 \quad (\ldots5)$$
$$ZD(h5) = [(1^2 + 2^2 + 3^2) / 3]^{1/2} = 2.160 \quad (\ldots3)$$

この横行の原点偏差ベクトル(3.000, 2.000, 3.535, 3.160))を行列で示すと

$$ZD_{n1} = [SumH(X_{np} * S_{p1}^2) / SumH(X_{np})]^{1/2}$$

ここで X_{np} は原データ行列、S_{p1} は連番{1, 2, ..., P}を成分にする縦ベクトル、SumH は行列の行和縦ベクトルを返す関数です。上の最終列の(…1), (…4), …という番号は原点偏差の昇順の順位を示します。この順位に従って行を並べ替えると上右表の分布（集中化行列）が得られます。

●異分布同偏差問題

次のように分布のパタンが異なるにもかかわらず原点偏差が等しい 2 つのデータが存在します。

P2	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
h1				+		+					+
h2			+					+		+	

P2	偏差
h1	7.594
h2	7.594

これは原点偏差を 2 乗和の平均の 2 乗根で計算した結果です(「2 乗原点偏差」$ZD^{\{2\}}$と呼びます；上表を参照）

$$ZD^{\{2\}}(4, 6, 11) = [(4^2 + 6^2 + 11^2) / 3]^{1/2} = 7.594$$
$$ZD^{\{2\}}(3, 8, 10) = [(3^2 + 8^2 + 10^2) / 3]^{1/2} = 7.594$$

このとき原点偏差を 3 乗和の平均の 3 乗根で計算すると($ZD^{\{3\}}$)、結果が弁別されるので異分布同偏差問題を解決できます。

$$ZD^{\{3\}}(4, 6, 11) = [(4^3 + 6^3 + 11^3) / 3]^{1/3} = 8.128$$
$$ZD^{\{3\}}(3, 8, 10) = [(3^3 + 8^3 + 10^3) / 3]^{1/3} = 8.005$$

逆に 1 乗の偏差を適用した原点偏差($ZD^{\{1\}}$)は単に反応した連番の平均となり、さらに高頻度で原点偏差の異分布同偏差問題が生じます。たとえば、(2), (1, 3), (1, 2, 3)の原点距離（＝平均）はどれも 2 になります。

$$ZD^{\{1\}}(2) = 2 / 1 = 2$$
$$ZD^{\{1\}}(1, 3) = (1 + 3) / 2 = 2$$

$ZD^{\{1\}}(1, 2, 3) = (1 + 2 + 3) / 3 = 2$

このようなケースが非常に多くなることは簡単に予想できるので、行列を集中化するためには役立ちません。これに 2 乗原点偏差 $ZD^{\{2\}}$を使えば次のように弁別できます。

$ZD^{\{2\}}(2) = (2^2 / 1^2)^{1/2} = (4 / 1)^{1/2} = 2$
$ZD^{\{2\}}(1, 3) = [(1^2 + 3^2) / 2]^{1/2} = (10 / 2)^{1/2} = 2.236$
$ZD^{\{2\}}(1, 2, 3) = [(1^2 + 2^2 + 3^2) / 3]^{1/2} = (14 / 3)^{1/2} = 2.160$

ここで、$ZD^{\{2\}}(1, 3)$の原点偏差が $ZD^{\{2\}}(2)$の原点偏差よりも少し遠くなっています。反応の分布で見れば、{-, +, -}よりも{+, -, +}の方が全体的に原点点から少し離れている、と解釈することになります。一方、$ZD^{\{2\}}(1, 2, 3)$の原点偏差(2.160)は ZD2(2) = 2 と ZD2(1, 3) = 2.236 の間にあります。これは、$ZD^{\{2\}}(1, 3)$の反応分布{+, -, +}と、$ZD^{\{2\}}(1, 2, 3)$の反応分布{v, v, v}を比較すると、中間の反応点があるほうが原点に近くなることがわかります。一般化すると、最大反応点が右にあるほど原点偏差が大きくなり、最大反応点が同じ条件であれば中間点があると原点偏差が小さくなります。

最後に 3 乗原点偏差を適用した $ZD^{\{3\}}$を比較しましょう。

$ZD^{\{3\}}(2) = (2^3 / 1)^{1/3} = (8 / 1)^{1/3} = 2.000$
$ZD^{\{3\}}(1, 3) = [(1^3 + 3^3) / 2]^{1/3} = (28 / 2)^{1/3} = 2.410$
$ZD^{\{3\}}(1, 2, 3) = [(1^3 + 2^3 + 3^3) / 3]^{1/3} = (36 / 3)^{1/3} = 2.289$

ここでも先の最大反応点と中間反応点についての一般化が適用できることがわかります。最大反応点は原点からの離れ方を示し、中間反応点は、その離れ方を引き戻す力がある、と考えれば、私たちの直感に沿った観察になるでしょう。この原点偏差法ではベクトルが原点点にどれほど偏っているのか、を計算します。このとき、乗数が大きくなるほど、ベクトルの右に位置する数値が強調されます。

■ラテンアメリカスペイン語の語彙変異 (1): 外的基準

次のデータはラテンアメリカスペイン語の「農夫」を示す言語形式とその地理的分布です(Cahuzac: 1980)。語形はアルファベット順に並べ、国名は北から南に配置されています。MX: México, CU: Cuba, RD: República Dominicana, PR: Puerto Rico, C5: (Guatemala, El Salvador, Honduras, Nicaragua, Costa Rica, Panamá), VE: Venezuela, CO: Colombia, EC: Ecuador, PE: Perú, BO: Bolivia, CH: Chile, PA: Paraguay, UR: Uruguay, AR: Argentina.

Cahuzac (1980)	MX	CU	RD	PR	C5	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
01 cacahuero							+	+							
02 cafetalista	+	+		+											
03 camilucho													+	+	+
04 campero													+	+	+
05 camperuso							+	+							
06 campirano					+	+	+	+							
07 campiruso					+	+									
08 campista	+			+	+	+	+								
09 campusano						+								+	+
10 campuso					+										
11 colono			+	+											
12 comparsa													+	+	+
13 conuquero		+	+	+			+	+							
14 coquero									+	+	+				
15 chagrero								+	+						
16 changador													+	+	+
17 chilero	+				+	+									
18 chuncano													+	+	+
19 enmaniguado		+	+	+											
20 estanciero													+	+	+
21 gaucho											+		+	+	+
22 guajiro		+	+												
23 guanaco					+	+									
24 guaso		+							+	+	+	+			+
25 huasicama								+	+						
26 huertero	+									+		+			+
27 hulero	+				+	+									
28 invernador										+		+	+	+	+
29 jíbaro			+	+											
30 lampero										+	+				+
31 lanudo							+	+	+						
32 llanero							+	+							
33 macanero	+														
34 manuto			+			+									
35 montero		+	+												
36 montubio		+	+					+	+	+					
37 paisano			+						+	+					
38 pajuerano		+							+	+					
39 partidario											+			+	+
40 payazo							+	+							
41 piona													+	+	+
42 ranchero	+	+	+			+									
43 rondín											+				
44 sabanero							+	+							
45 veguero			+			+									
46 viñatero										+		+	+	+	+
47 yanacón										+	+	+			+

このデータ行列の列を外的基準にして固定し、原点偏差法によって行を並べ替えると全体の分布は次のように集中化されます。

ZD.R	MX	CU	RD	PR	C5	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
33 macanero	+														
22 guajiro		+	+												
35 montero		+	+												
02 cafetalista	+	+		+											
19 enmaniguado		+	+	+											
11 colono			+	+											
29 jíbaro			+	+											
42 ranchero	+	+	+			+									
17 chilero	+				+	+									
27 hulero	+				+	+									
34 manuto			+			+									
45 veguero			+			+									
10 campuso					+										
08 campista	+			+	+	+	+								
07 campiruso					+	+									
23 guanaco					+	+									
13 conuquero		+	+	+			+	+							
06 campirano					+	+	+	+							
01 cacahuero							+	+							
05 camperuso							+	+							
32 llanero							+	+							
40 payazo							+	+							
44 sabanero							+	+							
36 montubio		+	+					+	+	+					
31 lanudo							+	+	+						
38 pajuerano		+							+	+					
37 paisano			+						+	+					
15 chagrero								+	+						
25 huasicama								+	+						
14 coquero									+	+	+				
43 rondín											+				
24 guaso		+							+	+	+	+			+
26 huertero	+									+		+			+
47 yanacón										+	+	+			+
30 lampero										+	+				+
09 campusano						+								+	+
28 invernador										+		+	+	+	+
46 viñatero										+		+	+	+	+
21 gaucho											+		+	+	+
39 partidario											+			+	+
03 camilucho													+	+	+
04 campero													+	+	+
12 comparsa													+	+	+
16 changador													+	+	+
18 chuncano													+	+	+
20 estanciero													+	+	+
41 piona													+	+	+

このようにデータ行列全体が対角化された反応パタンによって、一定の地域に集中する一定の語形の集まりを観察することができます。

7.2.2. 内的基準

先の外的基準による集中化では縦列を固定して、つまり外的基準として縦列を選択し、横行を原点偏差の大きさに従って並べ替えました。ここでは列も固定せずに、つまり外的基準を設定しないで行列全体を集中化する方法を考えます。

先のサンプルデータは、たとえば 5 つの地域(h1, h2, ..., h5)について 4 つの言語特徴(v1, v2, v3, v4)がどのように反応しているかを示していること

を想定しましょう。ここでは言語地理区分をする上で外部的な基準がなく、あるのは特徴を共有する地域のベクトルと(横因子：h1, h2, …, h5)、地域に共通する特徴のベクトルだけです(縦因子：v1, v2, v3, v4)。そこで、特徴(v)がどの地域(h)にあるかを調べ、該当するときに+印をつけたのが下左表です。このままでは地域(h)についても特徴(v)についてもどのような分布パタンがあるのかわからないので、地域(h)については特徴(v)の選択が近いものを並べ、特徴(v)については地域(h)の選択の仕方が近いものを並べるという操作をすると下右表が得られます。

Lv	v1	v2	v3	v4
h1	+	+		
h2			+	
h3		+		
h4			+	+
h5	+	+	+	

→

Lv	v2	v1	v3	v4
h3	+			
h1	+	+		
h5	+	+	+	
h2			+	
h4			+	+

「内的基準による集中化」(Concentration by interior criterion)とは上左表のようなデータから上右表のようなパタンを得る方法です。「最良のパタン」とは「反応するデータ(+印)がなるべく左上から右下に伸びる対角線に近い位置に集まるようなパタン」と決めて、このようなパタンを得る方法を考えます。先と同様の計算を何度か繰り返しますが方法は簡単です。

はじめに「外的基準のある集中化」と同様に横行の中で反応した＋印の原点偏差を計算します。このとき縦因子ベクトル(v)の初期値を(Yp: 1, 2, 3, 4)としておきます。

[1]
$$ZD^{\{3\}}(h1) = [(1^3 + 2^3) / 2]^{1/3} = 1.651$$
$$ZD^{\{3\}}(h2) = [(3^3) / 1]^{1/3} = 3.000$$
$$ZD^{\{3\}}(h3) = [(2^3) / 1]^{1/3} = 2.000$$
$$ZD^{\{3\}}(h4) = [(3^3 + 4^3) / 2]^{1/3} = 3.570$$
$$ZD^{\{3\}}(h5) = [(1^3 + 2^3 + 3^3) / 3]^{1/3} = 2.289$$

ここで求められた数値の連続(ベクトル Xn: 1.651, 3.000, 2,000, 3.570, 2.289)を最大値(Max: 3.570)と最小値(Min: 1.651)を使って、次のように限定化します（→「得点」「限定得点」）。

$$Xn' = (Xn - Min) / (Max - Min)$$

その結果、ベクトル Xn'の成分は

Xn' : (.000, .703, .182, 1.000, .333)

になります。次に今度は各縦列の原点偏差を計算します。たとえば縦列 v1 は縦に見た(1, 5)の位置で反応しているので ZD(v1)は、Xn'の 1 番の成分

(.000)と 5 番の成分(.333)を使って

[2] $ZD^{\{3\}}(v1) = \{[(.000)^3 + (.333)^3] / 2\}^{1/3} = .264$

同様にして,

$$ZD^{\{3\}}(v2:1,3,5) = \{[((.000)^3 + (.182)^3 + (.333)^3] / 3\}^{1/3} = .243$$
$$ZD^{\{3\}}(v3:2,4,5) = \{[(.703)^3 + (1.000)^3 + (.333)^3] / 3\}^{1/3} = .773$$
$$ZD^{\{3\}}(v4:4) = \{(1.000)^3 / 1\}^{1/3} = 1.000$$

このベクトル(Yp: .264, .243, .773, 1.000)を最大値(Max: 1.000)と最小値(Min: 0.243)を使って限定化します(Yp')。

Yp': (.028, .000, .700, 1.000)

これで第 1 回目の縦と横のベクトルの更新作業が終わりました。次に、この Yp'を使って、この[1]と[2]の作業を繰り返します。何度も繰り返すうちに、Yp のベクトルの成分に変化がなくなれば、そこで更新作業を終了し、Xn と Yp を昇順でソートして次を出力します。

ZD.A	v2	v1	v3	v4
h3	+			
h1	+	+		
h5	+	+	+	
h2			+	
h4			+	+

Row	Xn
h3	.000
h1	.085
h5	.569
h2	.820
h4	1.000

Column	v2	v1	v3	v4
Yp	.000	.094	.724	1.000

並べ替えの手段とした横行(h)と縦列(v)の原点偏差(ZD)はパタン化が集中したときの横行間の近さと各縦列間の近さをそれぞれ示しています。そこで、原点偏差の値によって個体のグルーピングと属性のグルーピングができます。

＊原点偏差による集中分析法の開発にあたって後述する荻野綱男(1980)の「交互平均法」を参考にしました。

プログラム (R)

```
CONCENT=function(D,s=3){
  if(is.character(D[1,1])){
    E=ReplDF(D,'¥¥+',1); E=Empty2zero(E); E=Cha2Num(E); E=Concent(E,s)$df
    E=ReplDF(E,1,'+'); E=ReplDF(E,'0',''); E
```

```
  } else Concent(D,s)}
Concent=function(D,s=3){
  # Diagonal concentration D:df, s=0:none/1:row/2:col/3:alll
  nr=NR(D); nc=NC(D); CR=CubeRoot; RS=RowSums; CS=ColSums
  #nrow,ncol,rename
  if(s==3) D=BestGK(D) # best Goodman&Kruskal
  Wr=matrix(1:nc,nr,nc,byrow=T) #weight-row
  Wc=matrix(1:nr,nr,nc,byrow=F) #weight-col
  for(i in 1:500) {
    W=D
    if(s==1|s==3){R=CR(Dv(RS((D*Wr)^3),RS(abs(D))));D=D[order(R),]} #row|all
    if(s==2|s==3){C=CR(Dv(CS((D*Wc)^3),CS(abs(D))));D=D[,order(C)]} #col|all
    if(s<3|all(D==W)) break
  }; list(sel=s,df=D,row=Rescale(R),col=Rescale(C), dg=GK(D), iter=i-1)} #op
```

R では, 負値の三乗根(-27)^(1/3)は計算できないので, 自作関数 CubeRoot を使用しました。

●量的データの距離

原点偏差集中分析を次のような量的データに適用します。

P1	v1	v2	v3	v4
d1	1	1	5	3
d2	3	4	4	4
d3	1	2	4	3
d4	7	2	2	2
d5	5	6	2	4

Dst.a	v1	v2	v4	v3
d4	7	2	2	2
d5	5	6	4	2
d2	3	4	4	4
d3	1	2	3	4
d1	1	1	3	5

Row	Xn
d4	
d5	.194
d2	.816
d3	.911
d1	1.000

Column	v1	v2	v4	v3
Yp		.311	.793	1.000

それぞれの原点偏差計算の最初のプロセスで先と同様の式を使います。

$$ZD^{\{3\}} = [SumH(X_{np} * A_p)^3 / SumH(X_{np})]^{1/3}$$

ここで Ap は先のように Sp = [1, 2, …, p]とするのではなく、中間点を 0 として、左側をマイナスにし、右側をプラスにします[2]。Ap = [-1.5, -0,5, 0,5, 1.5]。よって

[2] これは左側の数値と中間値との差を考慮するためです。左側は 3 乗されるのでマイナス(-)の符号は保持されます。

$$Ap = Sp - (p+1)/2$$

たとえば行 d1 の原点偏差は

$$ZD^{\{3\}}(d1) = \{[(1*(-1.5))^3 + (1*(-0.5))^3 + (5*(0.5))^3 + (3*(1.5))^3]\} / (1+1+5+3)^{1/3}$$

先の行列(X_{np})の成分は 0 または 1 でしたが、この行列(P1)では行列成分の数値になります。上の式は成分が非負であれば、小数点のある行列でも適用できます。

●初期状態に依存

次の入力データ（Xnp:「メガネ」の語彙変異）の縦因子（国名）はラテンアメリカの歴史・地理的な順番で並んでいます。

Xnp	ES	MX	GU	EL	CR	PN	CU	RD	PR	EC	CO	VE	PE	BO	PA	UR	CH	AR
ant		+	+	+	+	+			+		+	+	+		+		+	+
esp						+	+	+	+									
gaf	+	+				+		+		+	+							+
lent		+	+			+		+				+	+	+	+	+	+	+

この行列を原点偏差法で集中化すると以下になります(ZD1)。

ZD1	CU	ES	EC	PR	CO	RD	PN	EL	CR	MX	AR	GU	VE	PE	PA	CH	BO	UR
esp	+			+		+	+											
gaf		+	+		+	+	+			+	+							
ant				+	+		+	+	+	+	+	+	+	+	+	+		
lent						+	+			+	+	+	+	+	+	+	+	+

入力データを列で ABC 順にソートした行列が次の Anp です。

Anp	AR	BO	CH	CO	CR	CU	EC	EL	ES	GU	MX	PA	PE	PN	PR	RD	UR	VE
ant	+		+	+	+			+		+	+	+	+	+	+			+
esp						+								+	+	+		
gaf	+			+			+		+		+			+		+		
lent	+	+	+							+	+	+	+	+		+	+	+

これを原点偏差法で集中化すると次のように先の結果と異なります(ZD2)。

ZD2	EC	ES	CO	CR	EL	AR	MX	CH	GU	PA	PE	VE	PN	BO	UR	RD	PR	CU
gaf	+	+	+			+	+						+			+		
ant			+	+	+	+	+	+	+	+	+	+	+				+	
lent						+	+	+	+	+	+	+	+	+	+	+		
esp													+			+	+	+

このように原点偏差法による集中分析の結果は入力データの初期状態に依存します。原点偏差集中分析は入力行列の状態をできるだけ保ちながら、最大の集中化を目指す方法です。よって入力データの状態が適切でないと適切でない出力になります。この実験の Anp のように、国名を ABC 順に並べる理由はありません。そこで最適な初期状態として行と列をそれぞれ逆順に並べ替えて最も集中係数(Goodman&Kruskal)が高い行列を用意し、これを集中化させます。結果は下の ZD3 になります。

Xnp	ES	EC	CO	AR	RD	MX	PN	UR	BO	CU	CR	EL	CH	PA	PE	VE	PR	GU
ant			+	+		+	+				+	+	+	+	+	+	+	+
esp					+		+			+							+	
gaf	+	+	+	+	+	+	+											
lent				+	+	+	+	+	+				+	+	+	+		+

ZD3	ES	EC	CU	RD	PN	BO	UR	MX	AR	CO	PR	GU	VE	PE	PA	CH	EL	CR
gaf	+	+		+	+			+	+	+								
esp			+	+	+						+							
lent				+	+	+	+	+	+			+	+	+	+	+		
ant					+			+	+	+	+	+	+	+	+	+	+	+

●最大分割

(a) 最大分割数

次は集中分析の結果を分布の分割対称数が最大になるように分割した図です。

Cor.A	v2	v1	v3
d1	v		
d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

Cor.A	v2 v1	v3
d1	A	B
d4		
d5		
d2		
d3	C	D

ここで、左上の区画と右下の区画になるべく多く反応点が集まり、右上と左下の反応点が少なくなるような位置を探すと、その分割点が(4, 2)の位

置(d2:v2)であることがわかります。ここで、分割点に左上、右上、左下、右下の区分をそれぞれ A, B, C, D として、A, D の反応点の平均(Positive:Ps)と B, C の反応点の平均(Negative:Ng)を計算し、さらに次の対照値(Z)を計算します。

$$Z = (Ps - Ng) / (Ps + Ng)$$

上の表で計算すると

$$Ps = (7 + 1) / (8 + 1) = .889,\ Ng = (1 + 1) / (4 + 2) = .333$$
$$Z = (.889 - .333) / (.889 + .333) = .456$$

実際にはプログラムを使ってすべての可能な分割点をくまなく探し、それぞれの Z 係数を計算して、その最大値を求めます。その最大値を「最大分割数」(Maximal Division Count: MDC)とよびます。

なお、次の分割も Z 値は同じになります。

Dst.A	v1	v2	v3
d1	v		
d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

しかし、この場合 AD 領域の反応数(3+4=7)と BC 領域の反応数(2+1=3)を先の場合(7+1=8; i1+1=2 と比べると、区画が最大化されていないようです。（一方、AD 領域が完全に反応点でおおわれています。）そこで、それぞれの領域の反応数を全領域を 10 倍した数(N*P*10)で割って、これを重みとして、Ps と Ng にそれぞれ加えて Z 値を計算します。

(b) 最大分割値

区画内の反応数だけでなく、それぞれの反応点の位置を次のようにして考慮します。たとえば、先の表で d1:v1 の反応点の座標(1, 1)と分割点の座標(4, 2)のユークリッド距離(E)を計算すると

$$E(1, 1, 4, 2) = [(1 - 4)^2 + (1 - 2)^2]^{1/2} = (9 + 1)^{1/2} = 3.162$$

同様にして、すべての反応点について分割点との距離を測り、その距離を、先の「最大分割数」と同様に、A, B, C, D の区画に割り振って求めた量を「最大分割値」(Maximal Division Value: MDV)とよびます。

Dst.A	v1 v2	v3
d1	v	

Contr.coef.	値
MaxDiv.C: 4-2	.600

d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

MaxDiv.V: 4-2	.770

●交互平均法

荻野綱男(1980)によって提案された「交互平均法」による集中化法を紹介します。

P1	v1	v2	v3	v4	SHn
h1	1	1	5	3	10
h2	3	4	4	4	15
h3	1	2	4	3	10
h4	7	2	2	2	13
h5	5	6	2	4	17
SVp	17	15	17	16	65

→

Ogi.A	v3	v4	v2	v1
h1	5	3	1	1
h3	4	3	2	1
h2	4	4	4	3
h5	2	4	6	5
h4	2	2	2	7

Row	Xn
h1	.000
h3	.143
h2	.449
h5	.771
h4	1.000

Colum	v3	v4	v2	v1
Yp	.000	.288	.615	1.000

はじめに縦軸の h1 ~ h5 に(1, 2, ..., 5)という成分をもった縦ベクトル Xn を用意し、これを入力行列 Xnp に左積し、その結果を横ベクトル Yp とし、この Yp を列和横ベクトル SVp で割って相対化します。

$$Yp \leftarrow Xn^{T}\ Xnp\ /\ SVp$$

たとえば行列 P1 の 1 列では次のように計算します。

$$(1*1 + 3*2 + 1*3 + 7*4 + 5*5)\ /\ 17 = 63\ /\ 17 \fallingdotseq 3.71$$

Yp の成分全体は(3.71, 3.53, 2.53, 3.00)になります。

このベクトル Yp 成分の最大値 Max(=3.71)と最小値 Min(= 2.53)を使って限定得点とします[3]。

$$Yp \leftarrow (Yp - Min)\ /\ (Max - Min)$$
$$Yp = (1.00, .85, .00, .40)$$

次にこの Yp を行列 Xnp に右積して、行について同様の計算をし、新たな縦ベクトル Xn を得ます。

[3] 荻野はこの場合の最大値を 4、最小値を 1 としていますが、ここでは規定化するために最大値を 1、最小値を 0 とします。また計算も少し簡単になります。

$Xn \leftarrow Xnp\ Yp^T / SHn$

$Xn \leftarrow (Xn - Min) / (Max - Min)$

$Xn = (.00, .54, .20, 1.00, .90)$

これで新たな Xn が求められ、これを使って再度 Yp を計算し、その Yp を使って Xn を計算します。次がその結果です。

$Xn = (.00, .48, .16, 1.00, .81)$

さらに何度も同じ計算を繰り返すと次第に変化が少なくなるので、そのときに計算を終了します。

荻野綱男(1980)「敬語における丁寧さの数量化：札幌における敬語調査から(2)」『国語学』vol. 120, pp. 13-24.

■ラテンアメリカスペイン語の語彙変異 (2): 内的基準

先に行（南北の配置）を外的基準にした分析をしましたが、今回は外的基準を設定しないでデータ行列(Cahuzac: 1980)の内的基準にしたがって同じデータ行列を分析してみましょう。次の表を見ると先の分析と比べて、さらに強く集中化されていることがわかります。

ZD.A	PA	UR	AR	CH	BO	PE	MX	CU	PN	RD	PR	C5	EC	CO	VE
03 camilucho	+	+	+												
04 campero	+	+	+												
12 comparsa	+	+	+												
16 changador	+	+	+												
18 chuncano	+	+	+												
20 estanciero	+	+	+												
41 piona	+	+	+												
21 gaucho	+	+	+		+										
39 partidario		+	+		+										
43 rondín					+										
28 invernador	+	+	+	+		+									
46 viñatero	+	+	+	+		+									
47 yanacón			+	+	+	+									
30 lampero			+		+	+									
26 huertero			+	+		+	+								
09 campusano		+	+						+						
24 guaso			+	+	+	+		+					+		
33 macanero							+								
14 coquero					+	+							+		
42 ranchero							+	+	+	+					
02 cafetalista							+	+			+				
22 guajiro								+		+					
35 montero								+		+					
17 chilero							+		+			+			
27 hulero							+		+			+			
34 manuto									+	+					
45 veguero									+	+					
38 pajuerano						+		+					+		
37 paisano						+				+			+		
19 enmaniguado								+		+	+				
11 colono										+	+				
29 jíbaro										+	+				
07 campiruso									+			+			
23 guanaco									+			+			
10 campuso												+			
36 montubio						+		+		+			+	+	
08 campista							+		+		+	+			+
13 conuquero								+		+	+			+	+
06 campirano									+			+		+	+
15 chagrero													+	+	
25 huasicama													+	+	
31 lanudo													+	+	+
01 cacahuero														+	+
05 camperuso														+	+
32 llanero														+	+
40 payazo														+	+
44 sabanero														+	+

一般にデータを扱うときは分析者が先に一定の基準を設けて、それにしたがって分析をすることが多いのですが、それではデータの構造が本来有している内的基準が考慮されていません。このような方法を「前範疇化」(precategorization)と呼ぶことにします。本当はさらに良い結果が得られるにもかかわらず、分析者が先に基準を縛ることによって、その結果に自らが縛られていることがあります。いつも先に決めた基準で同じような分析をするよりも、むしろ内的基準による分析の結果によって範疇化をすれば方法が柔軟になり、新しい発見に出会う可能性が高まります。このような方法を「後範疇化」(postcategorization)と呼びます。どちらの方法も可能ですが、言語研究で後者の方法はあまり行われていないようです。

■中世スペイン語の 4 つの二重文字

次は中世スペイン語の二重文字-ss-, #ss, -ff-, #ff と、地域、年代、書体で分類した相対頻度表（千語率）です。

Factores	-ss-	#ss-	-ff-	#ff-
León	4.8	5.8	1.4	6.1
C. la Vieja	9.7	9.6	1.5	10.1
C. la Nueva	1.8	2.4	0.5	1.9
Aragón	11	0.9	2.1	3
A1200	16.8	3.1	1.2	6.2
A1225	5.7	0.2	0.5	0
A1250	12.2	3.1	1.5	6.8
A1275	13.1	13.8	1.5	12.4
A1300	13.4	14.8	4	22.1
A1325	14.3	20.5	3.6	23.8
A1350	8.9	10.6	3.1	9.4
A1375	3.7	3.7	1.9	3.3
A1400	6.6	1	0.9	1.2
A1425	3.9	0.5	0.7	1.1
A1450	7.8	0.5	1.1	2.3
A1475	7.3	0	0.9	0.8
cortesana	0.6	0.5	0.7	0.8
albalaes	19.2	19.4	3	12.9
privilegios	7.9	1.6	3.1	7.9
gótica	2.8	0.6	0.6	2.2
gótica cursiva	10.7	15.8	2.8	19.6
gótica cursiva [albalaes]	9.2	19.2	1.8	20.3
gótica cursiva [precortesana]	0.2	2	0.3	1.7
gótica libraria	10.3	9	3.5	11
gótica redonda	12.8	1.1	0.7	1.6
humanística	3.3	0	3.3	0
humanística redonda	0.5	3.7	0	1
precortesana	3.9	3.9	0.7	2.7

次は標準原点偏差集中分析の結果です。

Dst.a	-ss-	#ss-	#ff-	-ff-
A1200		3.1		1.2
albalaes				3.0
gótica redonda		1.1	1.6	.7

Aragón		.9	3.0	2.1
A1250		3.1		1.5
A1275				1.5
A1325				3.6
A1475			.8	.9
A1450		.5	2.3	1.1
A1300				4.0
gótica libraria				3.5
A1400		1.0	1.2	.9
C. la Vieja				1.5
A1225		.2		.5
gótica cursiva				2.8
privilegios		1.6		3.1
A1350				3.1
gótica cursiva [albalaes]				1.8
A1425	3.9	.5	1.1	.7
León	4.8			1.4
precortesana	3.9	3.9	2.7	.7
A1375	3.7	3.7	3.3	1.9
gótica	2.8	.6	2.2	.6
C. la Nueva	1.8	2.4	1.9	.5
humanística redonda	.5	3.7	1.0	
gótica cursiva [precortesana]	.2	2.0	1.7	.3
humanística	3.3			3.3
cortesana	.6	.5	.8	.7

行列の集中化によって、横行因子（地域、年代、文字種）が、縦列因子の2文字の並び方(-ss- > #ss- > #ff- > -ff-)に沿って配列されています。この集中分析によって、スペイン語史でしばしば取り上げられる語頭の#ff-の分布が書体と年代に関わること、そしてその関係が段階的に解釈することが可能になりました。#ff-, -ff-が-ss-, #ss-よりも比較的遅く現れ、#ff-が#ss-と近似していることから、仮説として#ss→#ff という影響（類推作用）があったことが考えられます。そこで、頻度が低い-ff-を外して再度集中分析にかけると、先の仮説を支持する傾向が次のように明示化される結果が得られました。

Dst.a	-ss-	#ss-	#ff-
A1200	16.8	3.1	6.2
gótica redonda	12.8	1.1	1.6
albalaes	19.2	19.4	12.9
Aragón	11.0	.9	3.0

A1250	12.2	3.1	6.8
A1475	7.3		.8
A1450	7.8	.5	2.3
A1400	6.6	1.0	1.2
A1225	5.7	.2	
humanística	3.3		
A1425	3.9	.5	1.1
A1275	13.1	13.8	12.4
precortesana	3.9	3.9	2.7
gótica	2.8	.6	2.2
A1375	3.7	3.7	3.3
privilegios	7.9	1.6	7.9
cortesana	.6	.5	.8
humanística redonda	.5	3.7	1.0
C. la Nueva	1.8	2.4	1.9
gótica cursiva [precortesana]	.2	2.0	1.7
C. la Vieja	9.7	9.6	10.1
A1350	8.9	10.6	9.4
León	4.8	5.8	6.1
gótica libraria	10.3	9.0	11.0
gótica cursiva	10.7	15.8	19.6
gótica cursiva [albalaes]	9.2	19.2	20.3
A1300	13.4	14.8	22.1
A1325	14.3	20.5	23.8

■アンダルシア方言の開母音

スペイン語アンダルシア方言では語末子音が消失し、それとともに先行する母音が開く現象が地域によって見られます。次の表は『アンダルシア言語民俗地図』(Manuel Alvar y Antonio Llorente: *Atlas lingüístico y etnográfico de Andalucía*, 1973)の資料をもとに作成した各県の出現頻度表です（調査地点数によって標準化）。+が開母音化、++が大きな開母音化を示します。

R	CA	SE	H	MA	CO	GR	J	AL
1533B:miel:el>e+	9	10	17	15	20	46	29	30
1533C:miel:el>e:	4	6	11	16	12	16	11	3
1615A:caracol:-ól>ó+(:)	3	3	2	5	15	19	14	11
1615B:caracol:-ól>ó(:)	15	27	18	16	3	6	1	2
1616A:árbol:-ol>o+		1	2		6	6	8	6
1616B:árbol:-ol>o	17	30	23	26	18	23	11	11
1618A:sol:-ól>ó+(:)	3	9	7	13	13	19	12	11
1618B:sol:-ól>ó(:)	15	21	15	13	1	6	1	1
1623A:beber:-ér>é+l			2	1	10	19	11	20
1623B:beber:-ér>é+	3	7	4	6	13	15	17	8
1623C:beber:-ér>é	15	24	19	19	2	4		
1626C:tos:o++				2	7	18	10	12
1626C:tos:-ós>ó+	5	7	7	13	18	27	17	19
1626D:tos:-ós>ó	11	22	10	9	2	2	1	
1627B:nuez:-éθ>é+	5	13	7	17	20	39	25	26
1627C:nuez:e++				5	14	26	18	18
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1629B:voz:-óθ>ó+	3	5	3	12	22	44	30	30
1629C:voz:-óθ>ó	14	23	18	13	2	2	1	1
1689A:niños:-os>-o+		2	1	4	22	44	31	30
1689B:niños:-os>oh[os)		1	4		2	8	3	8
1690A:pared:-éd>é+		8	6	10	17	24	19	11
1693B:redes:redes>re+	4	6	14	12	3	16	6	6
1694B:clavel:-él>é+,	3	6		15	20	40	24	29
1694C:clavel:-él>ér						5	1	1
1695A:claveles:e-es>-e-e+		2		4	2	4	2	3
1695B:claveles:e-es>-e+-e+		1		7	18	33	24	21
1695C:claveles:-e-es>-e-e:		1		3	1	1	2	1

これを集中分析した結果の最大分割数(MDC)と最大分割値(MDV)を示します。

Contr.coef.	Valuer
MaxDiv.C: 7-4	.581
MaxDiv.V: 7-4	.755

次に集中化した分布表を分割点(28, 4)で区分しました。+は開母音、#は最大開母音を示します。

Dst.A	CA	SE	H	MA	CO	GR	J	AL
1626D:tos:-ós>ó	11	22	10	9	2	2	1	
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1623C:beber:-ér>é	15	24	19	19	2	4		
1629C:voz:-óθ>ó	14	23	18	13	2	2	1	1
1618B:sol:-ól>ó(:)	15	21	15	13	1	6	1	1
1615B:caracol:-ól>ó(:)	15	27	18	16	3	6	1	2
1616B:árbol:-ol>o	17	30	23	26	18	23	11	11
1533C:miel:el>e:	4	6	11	16	12	16	11	3
1693B:redes:redes>re+	4	6	14	12	3	16	6	6
1618A:sol:-ól>ó+(:)	3	9	7	13	13	19	12	11
1695C:claveles:-e-es>-e-e:		1		3	1	1	2	1
1623B:beber:-ér>é+	3	7	4	6	13	15	17	8
1695A:claveles:e-es>-e-e+		2		4	2	4	2	3
1690A:pared:-éd>é+		8	6	10	17	24	19	11
1626C:tos:-ós>ó+	5	7	7	13	18	27	17	19
1533B:miel:el>e+	9	10	17	15	20	46	29	30
1627B:nuez:-éθ>é+	5	13	7	17	20	39	25	26
1615A:caracol:-ól>ó+(:)	3	3	2	5	15	19	14	11
1694B:clavel:-él>é+,	3	6		15	20	40	24	29
1629B:voz:-óθ>ó+	3	5	3	12	22	44	30	30
1616A:árbol:-ol>o+		1	2		6	6	8	6
1694C:clavel:-él>ér						5	1	1
1695B:claveles:e-es>-e+-e+		1		7	18	33	24	21
1689B:niños:-os>oh[os)		1	4		2	8	3	8
1627C:nuez:e++				5	14	26	18	18
1689A:niños:-os>-o+		2	1	4	22	44	31	30
1626C:tos:o++				2	7	18	10	12
1623A:beber:-ér>é+l			2	1	10	19	11	20

アンダルシア方言の開母音化の現象は東地域、すなわちコルドバ(CO), コルドバ(CO), ハエン(J), グラナダ(GR), アルメリア(AL) で優勢であることがわかります。一方、西部のカディス(CA), セビリア(SE)、ウエルバ(H), マラガ(MA)では上表の左上部(A 領域）の数値が高く、これは開母音化が比較的少ない領域です。

＊参照：Manuel Alvar. 1973. *Estructuralismo, geogafía lingüística y dialectología actual*, p.203.

7.2.3. 集中係数

データが集中する状態を測る係数「集中係数」(Coefficient of concentration)を設定し、これらを集中化の効果を示す指標とします。

(1) 相関係数

(a) 連番相関係数

以下に先の表を再掲します。

Lv	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Lv	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

縦と横の軸のデータ行列からなる表を散布図と見て、これから次のような X と Y の軸のデータ行列を作り、そこから「連番相関値」(Sequential Correlation Coefficient: SCC)を次のように計算します。

データ : (X, Y) = (1, 1) (2, 1) (2, 2) (3, 1) (3, 2) (3, 3) (4, 3) (5, 3) (5, 4)
SCC = 0.820

(b) 参照相関係数

実は､それぞれの反応点は連番のように等間隔で並んでいるのではなく、次のように列と行の係数が対応しているので、次にそれぞれの係数を参照した数直線を軸にすべきでしょう。

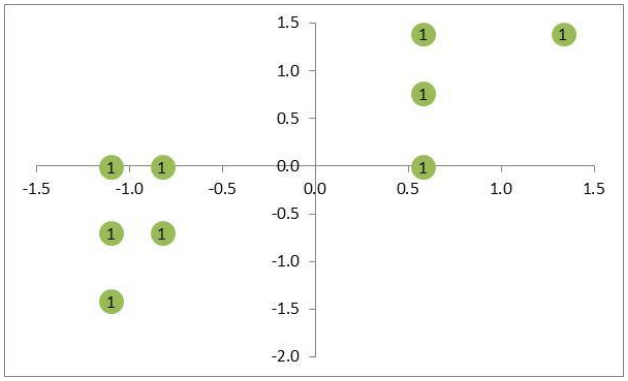
Lv	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Lv	係数
d3	1.42
d1	0.71
d5	0.01
d2	0.76
d4	1.38

Lv	v2	v1	v3	v4
係数	1.10	0.82	0.58	1.34

次の「集中バブル図」はそれぞれの反応点を X 軸と Y 軸の標準化された係数の位置によってプロットしています。X 軸は 4 座標あり、Y 軸は 5 座

標です。



「参照相関係数」(Referential Correlation Coefficient: RCC)はこの座標にもとづいて計算した相関係数です。

データ：(X, Y) = (-1.10, -1.42) (-1.10, -0.71) … (1.34, 1.38)
RCC = 0.835

原点偏差集中行列	集中前	集中後	**差**
連番相関係数	.226	.820	.594
参照相関係数	.563	.835	.273

それぞれのセルにある値を反応の頻度と見なします。

(2) 近接係数

下の集中行列(Prox)の「近接度」(proximity)は次のように.671 になります(→「関係」「近接」）。

Prox.	v1	v2	v3	v5	v4
h2	11	7	10	1	0
h1	10	19	14	12	7
h4	0	1	2	3	3
h3	0	0	1	1	12

Prox	Value
Ref. cor.	.791
Union	.846
Tightness	.942
Proximity	.671

行列全体の近接度を計算するためには、行列全体の相関係数の計算と同様に、行列のそれぞれの成分に横因子の値と縦因子の値を加重します。たとえば、h2:v1 の横因子は行番号/行数、その縦因子は列番号/列数を掛け、それぞれをベクトルにして、最後にこの 2 つのベクトル間の近接度を計算します。

プログラム

```
Function proximityC(Xnp, Sel)   '近接係数(2016.6.28) Sel=0: 連番; 1: 参照
  Dim i&, j&, k&, Xn, Yn, N&, P&, Nn&, V: N = nR(Xnp): P = nC(Xnp)
  For i = 1 To N: For j = 1 To P
    If Xnp(i, j) <> 0 Then Nn = Nn + 1
  Next: Next
  ReDim Xn(Nn, 1), Yn(Nn, 1)
  For i = 1 To N: For j = 1 To P
    If Xnp(i, j) <> 0 Then
      k = k + 1
      If Sel = 0 Then Xn(k, 1) = Xnp(i, j) * i / N: Yn(k, 1) = Xnp(i, j) * j / P
      If Sel = 1 Then Xn(k, 1) = Xnp(i, j) * Vn(i, 2):   Yn(k, 1) = Xnp(i, j) * Hp(j, 2)
    End If
  Next: Next
  proximityC = proximity(Xn, Yn) '近接度
End Function
```

(3) 順序連関係数

縦因子と横因子の順序がそれぞれ決まっているとき、その順序に従った数値の配列を示す数(Positive:Ps)と、逆の順序に従った数値の配列を示す数(Negative:Ng)の両側相対値は「グッドマンとクラスカルの順序連関係数」(Goodman and Kruskal's Rank Measure of Association: GK)とよばれます。この GK を集中係数の一つとして使います。集中化した行列の個体と変数の並び方が GK の算出に向いているからです。GK の計算の具体例は→「関係」「順序連関係数」を参照してください。

(4) 緊密係数

次表は「メガネ」を意味するスペイン語の地理変異を示します(ant.: anteojos, esp.: espejuelos, gaf.: gafas, len.:lentes)。

Xnp	ES	MX	GU	EL	CR	PN	CU	RD	PR	EC	CO	VE	PE	BO	PA	UR	CH	AR
ant.		+	+	+	+	+			+		+	+	+		+		+	+
esp.						+	+	+	+									
gaf.	+	+				+		+		+	+							+
len.		+	+			+		+				+	+	+	+	+	+	+

上表を原点偏差法(Dst)と対応分析法(Cor)を使って集中化すると下の 2 つの表になります。

Dst	CU	ES	EC	PR	CO	RD	PN	EL	CR	MX	AR	GU	VE	PE	PA	CH	BO	UR
esp..	+	(1)	(2)	+	(3)	+	+											
gaf.		+	+	(4)	+	+	+	(5)	(6)	+	+							
ant.				+	+	(7)	+	+	+	+	+	+	+	+	+	+		
len.						+	+	(8)	(9)	+	+	+	+	+	+	+	+	+

Cor	CU	PR	RD	ES	EC	PN	CO	MX	AR	EL	CR	GU	VE	PE	PA	CH	BO	UR
esp..	+	+	+	(1)	(2)	+												
gaf.		(3)	+	+	+	+	+	+	+									
ant.		+	(4)	(5)	(6)	+	+	+	+	+	+	+	+	+	+	+		
len.			+	(7)	(8)	+	(9)	+	+	(10)	(11)	+	+	+	+	+	+	+

上の 2 つの表を比べると、反応点(R: +)の数は当然同じですが(R=34)、その分布の緊密度が異なっていることがわかります。緊密度を示す数値として、それぞれの表の赤で記した、反応点に囲まれた無反応点(N)の数を数えると、Dst では 9 個, Cor では 11 個です。「反応点に囲まれた」という状態は、上下と左右のいずれかの反応点に囲まれていることを意味します。

そこで次の式で緊密係数(Tightness: T)を計算します。分母の(R + N)は、上のそれぞれの表の赤の領域にも反応点(+)を認めたと仮定したときの反応点全体の数を示します。そのときに完全に緊密な(tight)分布を示すことになるからです。

$$T = R / (R + N)$$

たとえば上の Dst と Cor の緊密度は、それぞれ

$$T(Dst) = 34 / (34 + 9) = .791$$
$$T(Cor) = 34 / (34 + 11) = .756$$

対応分析(Cor.)は分布の参照相関係数(RCC)の最大化を目指すために、当然その結果の相関係数は原点偏差法(Dst.)より大きくなっています(RCC(Cor) =.691 > RCC(Dst) = .644)。一方、原点偏差法は分布の形状の類似度を重視するので、上で見たように緊密度が高くなっています[4]。どちらの方が優れた方法である、と決めることはできないので、適宜、それぞれの相関係数と緊密度などを比較しながら利用するとよいでしょう。

次に、0-1 データではなく、頻度データの緊密度の計算法を説明します。下左表(Dst)が入力データです。これを見ると、i1:v4 の 7 が行の数値のまとまりを崩しているので、行の緊密性を阻害していることがわかります。これが 12 であれば行は緊密になります。同様にして、赤で記したセルの値が

[4] なお、このケースでは、連番相関係数(SCC)もゼロ偏差法のほうが高くなりました。SCC(Dst) = .637 > SCC(Cor) = .627.

それぞれ下右表のようになれば行列全体の緊密性が完全になります[5]。そこで、下左表(Dst.i)の緊密係数 T(Dst.i)は

T(Dst.i) = Dst.i の総和 / Dst.t の総和 = 114 / 142 = .803

Dst.i	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10		1
h3			1	12	1
h4		1	2	3	3

→

Dst.t	v1	v2	v3	v4	v5
h1	10	19	14	12	12
h2	11	11	11	12	3
h3		1	2	12	3
h4		1	2	3	3

上左表(Dst.i.)を集中化すると下左表(Dst.c.)になります。集中化すると、緊密を阻害するセルが少なくなり、実際に i2:v2(7)と i4:v4(3)だけになりました。下右表でそれぞれの値を 10, 7 にすると、完全に緊密な行列になるので、下左表(Dst.c.)の緊密係数 T(Dst.c.)は

T(Dst.c) = Dst.c の総和 / Dst.c.t の総和 = 114 / 121 = .942

Dst.c.	v1	v2	v3	v5	v4
h2	11	7	10	1	
h1	10	19	14	12	7
h4		1	2	3	3
h3			1	1	12

→

Dst.c.t	v1	v2	v3	v5	v4
h2	11	10	10	1	
h1	10	19	14	12	7
h4		1	2	3	7
h3			1	1	12

言語の変異･変化の研究において作成された行列の分布状態については、その相関係数の値が高いほど、縦因子と横因子の関係が強いことを示すので、相関係数の数値はたしかに重要です。一方、私たちはデータの分布の連続性や緊密性も重視します。上左表(Dst.c.)は、ほとんど完全な緊密性をもつ上右表(Dst.c.t.)に近い状態を示しています(94.2%)。そこで、上右表から縦因子の連続性と横因子の連続性がそれぞれ関連していることが確認できます。このことは入力データ(Dst.i.)の状態ではほとんど不可能でした。

(5) クラメア連関係数

クロス集計表の列と行の連関度の指数として「クラメアの連関係数」(Cramer's Measure of Association: CMA)が使われます。CMA は期待値から計算されるカイ 2 乗値を、その理論的な最大値で割ることによって求められます。次はデータ例（下左表: X_{np}）と、その期待値（下右表: E_{np}）です。

[5] たとえば h1:v4 は 13 でも 14 でも緊密度は変わりがありません。そこで、緊密化のための数値は可能な数値のうちの最小値とします。具体的には、当該のセルを囲む 2 つの数値のうち小さい方の値を使います。行列の端にあるために当該セルを囲むことができないときは、行または列だけで計算します。

$$E_{np} = (Sh * Sv) / S$$

ここで Sh, Sv, S はそれぞれ引数行列(X_{np})の横和列、縦和行、総和を示します。

X_{np}	v1	v2	v3	和 Sh
d1	45	48	66	159
d2	56	59	54	169
d3	58	51	78	187
d4	77	72	20	169
和 Sv	236	230	218	S: 684

E_{np}	v1	v2	v3
d1	54.860	53.465	50.675
d2	58.310	56.827	53.863
d3	64.520	62.880	59.599
d4	58.310	56.827	53.863

次は、その χ2 乗値(C_{np})と、クラメアの連関係数(CMA)を示します。

$$C_{np} = (X_{np} - E_{np})^2 / E_{np}$$

C_{np}	v1	v2	v3
d1	1.772	0.559	4.634
d2	0.092	0.083	0.000
d3	0.659	2.245	5.681
d4	5.991	4.051	21.289

CMA　0.185

χ^2 は次のように C_{np} の総和として定義されます。

$$\chi^2 = \Sigma_i \Sigma_j [(X_{ij} - E_{ij})^2 / E_{ij}]$$

ここで X_{ij} はデータの実測値を示し、E_{ij} はその期待値を示します。

クラメアの連関係数(CMA)の式は χ^2 を、χ^2 の最大値 χ^2_{max} で割って相対化した値です。

$$CMA = (\chi^2 / \chi^2_{max})^{1/2}$$

χ^2 の最大値 χ^2_{max} を次のようにして求めます。はじめに期待値を求めるために、横和(Sh)と縦和(Sv)と総和(S)を使います。

$$E_{ij} = Sh_i Sv_j / S$$

よって

$$\begin{aligned}
\chi^2 &= \Sigma_i \Sigma_j [(X_{ij} - E_{ij})^2 / E_{ij}] && \leftarrow \chi^2 \text{の定義} \\
&= \Sigma_i \Sigma_j [(X_{ij} - Sh_i Sv_j / S)^2 / (Sh_i Sv_j / S)] && \leftarrow E_{ij} = Sh_i Sv_j / S \\
&= \Sigma_i \Sigma_j [(S X_{ij} - Sh_i Sv_j) / S]^2 / (Sh_i Sv_j / S) && \leftarrow S \text{で通分} \\
&= \Sigma_i \Sigma_j \{[(S^2 X_{ij}^2 - 2 S X_{ij} Sh_i Sv_j + Sh_i^2 Sv_j^2) / S^2] (S / Sh_i Sv_j)\} \\
&= \Sigma_i \Sigma_j [(S^2 X_{ij}^2 / Sh_i Sv_j - 2 S X_{ij} + Sh_i Sv_j) / S]
\end{aligned}$$

$$= \Sigma_i \Sigma_j (S\ X_{ij}^2 / Sh_i\ Sv_j) - 2 \Sigma_i \Sigma_j X_{ij} + \Sigma_i \Sigma_j (Sh_i\ Sv_j / S)$$

ここで、第 2 項の $\Sigma_i \Sigma_j X_{ij}$ は総和(S)を示します。また、第 3 項の $\Sigma_i \Sigma_j Sh_i\ Sv_j$ は総和の 2 乗(S^2)を示します。

よって

$$\chi^2 = S\ \Sigma_i \Sigma_j (X_{ij}^2 / Sh_i\ Sv_j) - 2S + S$$

[1] $$= S\ [\Sigma_i \Sigma_j (X_{ij}^2 / Sh_i\ Sv_j) - 1]$$

さて、χ^2 が最大の χ^2_{max} となるのは、次のようにセルの縦和と横和がセルの値と同じ、というケースです。一般に、次のような最大の連関度を示すデータ行列の χ^2 を求めてみましょう。このとき、行と列のそれぞれの選択において一方が他方と完全に連関しています。

Xmax	v1	v2	…		vP	和 Sh
d1	X_1	0	0	0	0	X_1
d2	0	X_2	0	0	0	X_2
:	0	0	…	0	0	…
dN	0	0	0	X_M	0	X_M
和 Sv	X_1	X_2	…	X_M	0	S

ここで M は行数 N と列数 P を比べて小さなほうの値を示します。

[2] $$M = \min(N, P)$$

このように長方形のデータ行列の場合、その中の最大の正方形の中で縦も横も他と共有しない、というケースを考えるからです。この場合、先の χ^2 式を使うと χ^2_{max} は次のように計算されます。

$$\chi^2_{max} = S\ [\Sigma_i \Sigma_j (X_{ij}^2 / Sh_i\ Sv_j) - 1] \quad \leftarrow [1]$$
$$= S\ (X_1^2 / X_1 X_1 + X_2^2 / X_2 X_2 + \ldots + X_M^2 / X_M X_M - 1) \quad \leftarrow \text{上表}$$
$$= S\ (M - 1) \quad \leftarrow \text{すべて分子=分母、これらが M 個}$$
$$= S\ [\min(N, P) - 1] \quad \leftarrow [2]$$

よってクラメア(Cramer)連関係数(CMA)は次の式になります。

$$CMA = \{\chi^2 / \chi^2_{max}\}^{1/2} = \{\chi^2 / [S\ (\min(N, P) - 1)]^{1/2}$$

この式で根をとるのは、分子と分母の計算で次数が 2 になっているためです。なお、クラメア連関係数は、軸の順序を変えても全体の期待値は変化しないので、集中化の前後で変化しません。よって、クラメア連関係数これを、集中化の有無にかかわらず、データ行列がもつ列と行の連関度を示すものとして使用します。

(6) 相補係数

次の 2 つのデータ例(1), (2)を比べると、(1)では反応点(v)の縦横の重なりがなく、(2)では d2:v4 に反応点があって、その行と列の反応点と重なっていることがわかります。(1)のような分布は、反応点が行列の中で互いに補い合っているため「相補分布」(complementary distribution)とよばれます。この相補分布を示すデータでは v1, v4 ならば必ず d1 を選択し、v2, v3 ならば d2 を選択しています。相補分布を示す行列では行または列を反応点を失うことなく重ね合わせることができます。一方、(2)の v4 は d1 と d2 を同時に選択しているので、この状態は「自由変異」(free variation)とよばれます。自由変異がある行列で行または列を重なると、一部の反応点（ここでは d2:v4）が失われます。

(1)	v1	v2	v3	v4
d1	+			+
d2		+	+	

(2)	v1	v2	v3	v4
d1	+			+
d2		+	+	+

相補分布と自由変異の間にはさまざまな程度の差があります。(2)は v1, v2, v3 までは相補分布になっていますが、d1:v3 に反応点があればその相補分布性は低くなります。そこで、このような相補性を示す指標として次のような「相補係数」(complement coefficient: CC)を考えます。

相補分布ではそれぞれの反応点の縦または横に競合する反応点がありません。(1)では横に競合反応点がありますが、縦に競合する反応点がありません。一方、(2)の d2:v4 は横と縦にそれぞれ 2 個(v2, v3)と 1 個(v4)があります。そこで d2:v4 の変異数を 2 x 1 = 2 とします。完全な自由変異は行列のすべての成分に反応点がある場合です。よって、各反応点には最大で 3 x 1 = 3 の変異数が考えられます。

相補係数(CC)は次の式で求めます。

$$CC = 1 - (\Sigma_i \Sigma_j D_{ij} X_i Y_i) / [\Sigma_i \Sigma_j D_{ij} (P - 1) (N - 1)]$$

ここで D はデータ、X は縦の競合反応点数、Y は横の競合反応点数、N は行数、P は列数を示します。この式の P - 1 は同じ行で競合するデータ数の考えられる最大値を示します。N-1 も同様です。

上の(2)の行列では d1:v1 で X=0, Y=1; d1:v4 で X=1, Y=1; d2:v2 で X=0, Y=2, d2:v3 で X=0, Y=2; d2:v4 で X=1, Y=2 なので分子は

$$(1*1*0 + 0*2*1 + 0*2*1 + \underline{1*1*1}) + (0*3*1 + 1*2*0 + 1*2*0 + \underline{1*2*1}) = 3$$

このようにデータが原点であるときや、行和または列和が原点になるとき、すなわち相補分布を示すときはカウントされなくなります。分母は

$$(\underline{1*3*1} + 0*3*1 + 0*3*1 + \underline{1*3*1}) + (0*3*1 + \underline{1*3*1} + \underline{1*3*1} + \underline{1*3*1}) = 15$$

この商 3 / 15 = .200 は自由変異の大きさを示し、競合する反応点が大きければ大きいほど、この商は大きくなります。そこで、相補係数(CC)はその逆数にします。

CC = 1 - .200 = .800

次のような定量的データについては、積の項の行と列の最大値を考慮して相補係数(CC)を計算します。

(3)	v1	v2	v3	v4
d1	2			3
d2		4	5	6

$$CC = 1 - (\Sigma_i \Sigma_j D_{ij} * X_i * Y_i) / (\Sigma_i \Sigma_j D_{ij} * Mr_i * Mc_j)$$

ここで、D はデータ、X は縦の競合反応点数、Y は横の競合反応点数、Mr は該当行の最大値、Mc は該当列の最大値を示します。

上のデータ例で相補係数(CC)を計算します。分数の分子は

[2*3*0 + 0*(2+3)*4 + 0*(2+3)*5 + 3*2*6] +
[0*(4+5+6)*2 + 4*(5+6)*0 + 5*(4+6)*0 + 6*(4+5)*3] = 36 + 162 = 198

分数の分母は

[2*(3+3+3)*2 + 0*(3+3+3)*4 + 0*(3+3+3)*5 + 3*(3+3+3)*6] +
[0*(6+6+6)*2 + 4*(6+6+6)*4 + 5*(6+6+6)*5 + 6*(6+6+6)*6]
= (36 + 162) + (288 + 450 + 648) = 1584

よって相補係数(CC)は

CC = 1 – 198 / 1585 = .875

行列(1)では分子の項がどれも 0 になるので CC = 1.000 になります。このセクションで扱ってきた行列 Lv の相補係数は.861 になります。なお、行列を集中化しても相補係数は変化しません。

(7) 対角性係数

集中分析の目的は数値の二次元で展開される分布を，行と列，行または列の順番を変えることによって，可能な限り，分布の対角線の近傍に集めることです。そこで，その対角線の近傍に集中する度合いを「対角性係数」(diagonality)として計算する方法を考えます。

次の例を使って，対角性係数を説明します。

```
> D=DF(A=c(10,11,0,0),B=c(19,7,0,1),
```

```
C=c(14,10,1,2),D=c(7,0,12,3),E=c(12,1,1,3))
> rownames(D)=VT('w1,w2,w3,w4')
> E=Concent(D)
> B2(D,E)
      A  B  C  D  E :     A  C  B  E  D
1 w1 10 19 14  7 12 : w2 11 10  7  1  0
2 w2 11  7 10  0  1 : w1 10 14 19 12  7
3 w3  0  0  1 12  1 : w4  0  2  1  3  3
4 w4  0  1  2  3  3 : w3  0  1  0  1 12
> Diag(D); Diag(E) #0.6319613; 0.7064935
```

上の左の行列は集中化前の状態(D)を示し，右の行列は集中化後の状態(E)を示します。この 2 つの行列の対角性係数を関数 Diag()を使って求めると，それぞれ 0.632, 0.704 になります。次が関数 Diag()です。

```
Diag=function(D){
  if(any(D<0)) D=D+abs(Min(D))
  nr=NR(D); nc=NC(D); Rs=RowSums(D); E=MATRIX(0,nr,nc); d=e=0
  r=Rnd(nr/2); E[1:r,nc]=Rs[1:r]; E[(r+1):nr,1]=Rs[(r+1):nr]
  P=Rnd(seq(1,nc,length.out=nr))
  for(i in 1:nr) {p=P[i]; for(j in 1:nc) {
    if(is.na(D[i,j])) next
    else {d=d+abs(p-j)*D[i,j]; e=e+abs(p-j)*E[i,j]}
  }}; unname(1-d/e)
} #Diagonality [0,1]
```

以下では，集中化前の状態(D)を使います。この D の行和(Rs)を求め，その前半をゼロ行列(E)の最終列に，後半にその開始列に配置します。この E は最大の非対角性を示します。そして，1 から行数(nr)までの数値(整数)を列数分(nc), P(1, 2, 4, 5)に用意します。この 4 個の整数がそれぞれの行の対角成分の位置になります。たとえば, 1 行(r1)では 1=c1 が対角位置であり，2 行(r2)では 2=c2 が対角位置です。4 行(r4)では，5=c5 が対角位置になります。

```
> Rs
w1 w2 w3 w4
62 29 14  9
> E
   c1 c2 c3 c4 c5
r1  0  0  0  0 62
r2  0  0  0  0 29
r3 14  0  0  0  0
r4  9  0  0  0  0
> C
[1] 1 2 4 5
```

関数の for(i in 1:nr)は行の変化を示します。それぞれの行に対応する対

角位置 p を P[i]から取り出し，列の変化を示す for(j in 1:nc)に移ります。この繰り返し演算では p と j の差(距離=abs(p-j))に x=D[i,j]を掛けて加算し，対角位置からの距離の和 d を足し上げていきます。同じことを最大非対角行列 E についても行い，距離の和 e を求めます。行列全体の計算を終えた後に求められた d, e を使った d/e は範囲[0,1]の「非対角性」を示すので，逆に，その補数 1-d/e は「対角性」を示します。なお D が負値を含むときは，最初に D の最小値(負値)の絶対値を加算して全体を正値にします。対角性係数を含めた各種の集中係数を比較するために次の実験をします。

```
> D1=DF(a=c(9,0,0,0),b=c(0,9,0,0),c=c(0,0,9,0),d=c(0,0,0,9))
> D2=DF(a=c(0,9,0,0),b=c(9,0,0,0),c=c(0,0,9,0),d=c(0,0,0,9))
> D3=DF(a=c(9,0,0,0),b=c(1,9,0,0),c=c(0,1,9,0),d=c(0,0,1,9))
> D4=DF(a=c(1,9,0,0),b=c(0,0,9,0),c=c(0,0,0,9),d=c(0,0,0,0))

> B4(D1,D2,D3,D4)
  [D1] a b c d : [D2] a b c d : [D3] a b c d : [D4] a b c d
1    1 9 0 0 0 :    1 0 9 0 0 :    1 9 1 0 0 :    1 1 0 0 0
2    2 0 9 0 0 :    2 9 0 0 0 :    2 0 9 1 0 :    2 9 0 0 0
3    3 0 0 9 0 :    3 0 0 9 0 :    3 0 0 9 1 :    3 0 9 0 0
4    4 0 0 0 9 :    4 0 0 0 9 :    4 0 0 0 9 :    4 0 0 9 0
```

行列[D1]は完全な対角化を示し，[D2]は 1 行と 2 行が入れ替わり，一部で対角化が崩れています。[D3]では対角線から右に少し延長しています。[D4]では対角線にほとんど数値がありません。次は，これらの行列を使って，クラメア連関係数(Cramer)，相関係数(CorCross)，グッドマン-クラスカル順序連関係数(GK)，対角性係数(Diag)を計算した結果です。

```
Cramer(D1); CorCross(D1); GK(D1); Diag(D1) # 1,1,1,1
Cramer(D2); CorCross(D2); GK(D2); Diag(D2) # 1, .800, .666, .800
Cramer(D3); CorCross(D3); GK(D3); Diag(D3) # .904, .971, 1, .970
Cramer(D4); CorCross(D4); GK(D4); Diag(D4) # .816, .979, 1, .591
```

最初の行(D1)では，どの係数を使っても確かに最大値 1 を示しています。2 番目の行(D2)では，クラメア連関係数(Cramer)では 1 になっています。クラメア連関係数の分子はカイ二乗値を使って，全体で期待値と比較しているので，個々の数値の位置が変わっても影響がありません。よって，集中係数としてクラメア連関係数を使うことはできません。3 番目の行(D3)ではグッドマン-クラスカル順序連関係数(GK)が最大の 1 を示しています。D3 のような分布では，逆方向の順はないためにこのような結果になります。最後の行(D4)では，クラメア連関係数(Cramer)，相関係数(CorCross)，グッドマン-クラスカル順序連関係数(GK)は高い数値を算出しますが，対角性係数(Diag)の数値は対角成分からの距離があるために，低い数値となっています。ここで D4 の分布を見ると，全体的にあまり対角化されていないことがわかります。よって，集中係数として対角性係数(Diag)が最適だと言えるでしょう。

次に，集中化前(D)と集中化後(E)の行列の各種の集中係数を比較します。

```
> D=DF(A=c(10,11,0,0),B=c(19,7,0,1),
+      C=c(14,10,1,2),D=c(7,0,12,3),E=c(12,1,1,3))
> rownames(D)=VT('w1,w2,w3,w4')

> E=Concent(D); B2(D,E)
  [D]  A  B  C  D  E : [E]  A  C  B  E  D
1  w1 10 19 14  7 12 :  w2 11 10  7  1  0
2  w2 11  7 10  0  1 :  w1 10 14 19 12  7
3  w3  0  0  1 12  1 :  w4  0  2  1  3  3
4  w4  0  1  2  3  3 :  w3  0  1  0  1 12

> Cramer(D); Cramer(E) #0.4367797; 0.4367797 (!)
> CorCross(D); CorCross(E) #0.2399302; 0.6100977
> GK(D); GK(E) #0.1972339; 0.6951293
> Diag(D); Diag(E) #0.6319613; 0.7064935
```

はじめのクラメア連関係数(Cramer)は 2 つの行列で同じ結果を示しています。次の相関係数(CorCross)とグッドマン-クラスカル順序連関係数(GK)は，どちらも集中化前(D)は低い数値ですが(0.240, 0.197)，集中化後(E)は大きく増加しています(0.610, 0.695)。最後の対角性係数は集中化前(D)でも 0.5 以上あり，集中化後(E)はさらに上昇し，0.7 を超えています。頻度分布を見ると，どちらにも一定の集中化が見られるので，対角性係数が信頼できます。

次はここで使用したクラメア連関係数(Cramer)，相関係数(CorCross)，グッドマン-クラスカル順序連関係数(GK)の関数です。

```
Cramer=function(D) {
  if(any(D<0)) D=D+abs(Min(D))
  sqrt(Chi(D)/Sum(D)/min(NC(D)-1, NR(D)-1))}
#Cramer's measure of association Cramer's V [0, 1]

CorCross=function(D,R=NULL,C=NULL){
  if(any(D<0)) D=D+abs(Min(D))
  nr=NR(D); nc=NC(D); E=NULL
  if(is.null(R)) R=1:nr; if(is.null(C)) C=1:nc
  for(i in 1:nr){for(j in 1:nc){
    n=D[i,j]; if(is.na(D[i,j])|n==0) next
    E=rbind(E,W=matrix(c(R[i],C[j]),n,2,byrow=T))
  }}; cor(E[,1],E[,2])
} #correlation of cross table (D:df, R:row weight, C:col weight) [-1, 1]

GK=function(D){
```

```
  if(any(D<0)) D=D+abs(Min(D))
  nr=NR(D); nc=NC(D); p=0; n=0 #p:positive; n:negative
  for(i in 1:(nr-1)) {for(j in 1:(nc-1)) {
    if(is.na(D[i,j])) next; p=p+D[i,j]*Sum(D[(i+1):nr,(j+1):nc])}}
  for(i in 1:(nr-1)) {for(j in 2:nc){
    if(is.na(D[i,j])) next; n=n+D[i,j]*Sum(D[(i+1):nr,1:(j-1)])}}
  unname((p-n)/(p+n))
}# Goodman and Kruskal's. rank measure of association [-1, 1]
```

■多変数集中分析の比較

同じデータ(Cahuzac 1980)を使って集中化の 4 つの方法の分析結果を比較します。次が集中化された分布パタンです。

語形	AR	BO	CH	CO	CR	CU	EC	EL	GU	HO	MX	NI	PA	PE	PN	PR	RD	UR	VE
1 cacahuero				v															v
2 cafetalista						v					v					v			
3 camilucho	v												v					v	
4 campero	v												v					v	
5 camperuso				v															v
6 campirano				v	v			v	v	v		v			v				v
7 campiruso					v			v	v	v		v			v				
8 campista					v			v	v	v	v	v			v	v			v
9 campusano	v														v			v	
10 campuso					v			v	v	v		v							
11 colono																v	v		
12 comparsa	v												v					v	
13 conuquero				v		v										v	v		v
14 coquero		v					v							v					
15 chagrero				v			v												
16 changador	v												v					v	
17 chilero					v			v	v	v	v	v			v				
18 chuncano	v												v					v	
19 enmariguado						v										v	v		
20 estanciero	v												v					v	
21 gaucho	v	v											v					v	
22 guajiro						v											v		
23 guanaco				v				v	v	v		v			v				
24 guaso	v	v	v			v	v							v					
25 huasicama				v			v												
26 huertero	v		v								v			v					
27 hulero				v				v	v	v	v	v			v				
28 invernador	v		v										v	v				v	
29 jibaro																v	v		
30 lampero	v	v												v					
31 lanudo				v			v												v
32 llanero				v															v
33 macanero					v						v								
34 manuto																v		v	
35 montero						v												v	
36 montubio				v		v	v							v				v	
37 paisano							v							v				v	
38 pajuerano						v	v							v					
39 partidario	v	v																v	
40 payazo				v															v
41 piona	v												v					v	
42 ranchero						v					v					v		v	
43 rondin		v																	
44 sabanero				v															v
45 veguero																v		v	
46 viñatero	v		v										v	v				v	
47 yanacón	v	v	v											v					

(1) データ行列

語形	PA	UR	AR	BO	CH	GU	HO	NI	EL	CR	PE	PN	EC	MX	CO	VE	CU	RD	PR
16 changador	v	v	v																
12 comparsa	v	v	v																
41 piona	v	v	v																
18 chuncano	v	v	v																
20 estanciero	v	v	v																
3 camilucho	v	v	v																
4 campero	v	v	v																
21 gaucho	v	v	v	v															
39 partidario		v	v	v															
43 rondin				v															
10 campuso						v	v	v	v	v									
46 viñatero	v	v	v		v						v								
28 invernador	v	v	v		v						v								
30 lampero			v	v							v								
9 campusano		v	v									v							
47 yanacón			v	v	v						v								
23 guanaco						v	v	v	v	v		v							
7 campiruso						v	v	v	v	v		v							
17 chilero						v	v	v	v	v		v		v					
27 hulero						v	v	v	v	v		v		v					
26 huertero			v		v							v		v					
14 coquero				v								v	v						
6 campirano						v	v	v	v	v		v			v	v			
33 macanero											v			v					
24 guaso			v	v	v							v	v				v		
8 campista						v	v	v	v	v		v		v		v			v
38 pajuerano												v	v				v		
15 chagrero													v		v				
25 huasicama													v		v				
37 paisano												v	v					v	
31 lanudo													v		v	v			
36 montubio												v	v		v		v	v	
42 ranchero													v	v			v	v	
44 sabanero															v	v			
40 payazo															v	v			
1 cacahuero															v	v			
5 camperuso															v	v			
32 llanero															v	v			
45 veguero													v					v	
34 manuto													v					v	
2 cafetalista														v			v		v
13 conuquero															v	v	v	v	v
22 guajiro																	v	v	
35 montero																	v	v	
19 enmariguado																	v	v	v
11 colono																		v	v
29 jibaro																		v	v

(2) 原点偏差集中分析(N=3)

語形	AR	UR	PA	PE	CH	BO	EC	CU	RD	CO	PR	VE	MX	PN	CR	NI	GU	HO	EL
28 invernador	v	v	v	v	v														
46 viñatero	v	v	v	v	v														
24 guaso	v			v	v	v	v	v											
21 gaucho	v	v	v			v													
47 yanacón	v			v	v	v													
20 estanciero	v	v	v																
3 camilucho	v	v	v																
4 campero	v	v	v																
12 comparsa	v	v	v																
41 piona	v	v	v																
16 changador	v	v	v																
18 chuncano	v	v	v																
39 partidario	v	v					v												
30 lampero	v				v		v												
14 coquero					v		v	v											
36 montubio					v			v	v	v	v								
38 pajuerano					v			v	v										
37 paisano					v			v		v									
26 huertero	v				v	v								v					
43 rondin							v												
22 guajiro								v	v										
35 montero								v	v										
15 chagrero								v			v								
25 huasicama								v			v								
19 enmaniguado									v	v		v							
9 campusano	v	v												v					
31 lanudo								v			v		v						
29 jibaro										v		v							
11 colono										v		v							
13 conuquero									v	v	v	v	v						
32 llanero											v		v						
5 camperuso											v		v						
44 sabanero											v		v						
1 cacahuero											v		v						
40 payazo											v		v						
2 cafetalista									v			v		v					
34 manuto										v					v				
45 veguero										v					v				
42 ranchero									v	v					v				
33 macanero														v		v			
10 campuso															v		v	v	v
23 guanaco															v	v	v	v	v
7 campiruso															v	v	v	v	v
6 campirano											v		v		v	v	v	v	v
17 chilero														v	v	v	v	v	v
27 hulero														v	v	v	v	v	v
8 campista												v	v	v	v	v	v	v	v

(3) 主成分集中分析

語形	AR	RD	UR	CU	PA	PE	EC	BO	CH	CO	PR	VE	MX	PN	CR	NI	GU	HO	EL
2 cafetalista				v							v		v						
1 cacahuero										v		v							
5 camperuso										v		v							
32 llanero										v		v							
40 payazo										v		v							
44 sabanero										v		v							
43 rondin									v										
11 colono		v									v								
29 jibaro		v									v								
31 lanudo							v			v		v							
15 chagrero							v			v									
25 huasicama							v			v									
26 huertero	v					v			v					v					
35 montero		v		v															
22 guajiro		v		v															
19 enmaniguado		v		v															
13 conuquero		v		v						v	v	v							
42 ranchero		v		v										v	v				
14 coquero						v	v	v											
34 manuto		v													v				
45 veguero		v													v				
38 pajuerano				v		v	v												
30 lampero	v					v		v											
37 paisano		v				v	v												
39 partidario	v		v					v											
20 estanciero	v		v		v														
4 campero	v		v		v														
41 piona	v		v		v														
12 comparsa	v		v		v														
16 changador	v		v		v														
18 chuncano	v		v		v														
3 camilucho	v		v		v														
47 yanacón	v					v		v	v										
9 campusano	v		v											v					
21 gaucho	v		v		v				v										
36 montubio		v		v		v	v			v									
46 viñatero	v		v		v	v				v									
28 invernador	v		v		v	v				v									
24 guaso	v			v		v	v	v	v	v									
33 macanero															v		v		
10 campuso																	v	v	v
8 campista											v	v	v	v	v	v	v	v	v
17 chilero															v	v	v	v	v
27 hulero															v	v	v	v	v
7 campiruso																v	v	v	v
23 guanaco																v	v	v	v
6 campirano											v		v				v	v	v

(4) 因子集中分析

語形	PA	UR	AR	BO	CH	PE	EC	CU	RD	CO	PR	MX	VE	PN	CR	NI	GU	HO	EL
3 camilucho	v	v	v																
4 campero	v	v	v																
12 comparsa	v	v	v																
16 changador	v	v	v																
18 chuncano	v	v	v																
20 estanciero	v	v	v																
41 piona	v	v	v																
21 gaucho	v	v	v	v	v														
39 partidario		v	v	v															
28 invernador	v	v	v			v	v												
46 viñatero	v	v	v			v	v												
43 rondin				v															
30 lampero				v	v		v												
47 yanacón				v	v	v	v												
24 guaso				v	v	v	v	v	v										
14 coquero					v		v	v											
9 campusano			v	v										v					
26 huertero				v		v	v					v							
38 pajuerano							v	v	v										
37 paisano							v	v		v									
36 montubio							v	v	v	v	v								
25 huasicama								v			v								
15 chagrero								v			v								
22 guajiro									v	v									
35 montero									v	v									
31 lanudo								v			v		v						
19 enmaniguado									v	v		v							
2 cafetalista											v	v							
13 conuquero									v	v	v	v		v					
42 ranchero									v	v			v		v				
11 colono										v		v							
29 jibaro										v		v							
34 manuto										v					v				
45 veguero										v					v				
32 llanero											v			v					
44 sabanero											v			v					
40 payazo											v			v					
5 camperuso											v			v					
1 cacahuero											v			v					
33 macanero													v			v			
8 campista											v	v	v	v	v	v	v	v	v
6 campirano											v			v					
17 chilero													v	v	v	v	v	v	v
27 hulero													v	v	v	v	v	v	v
7 campiruso														v	v	v	v	v	v
23 guanaco														v	v	v	v	v	v
10 campuso															v	v	v	v	v

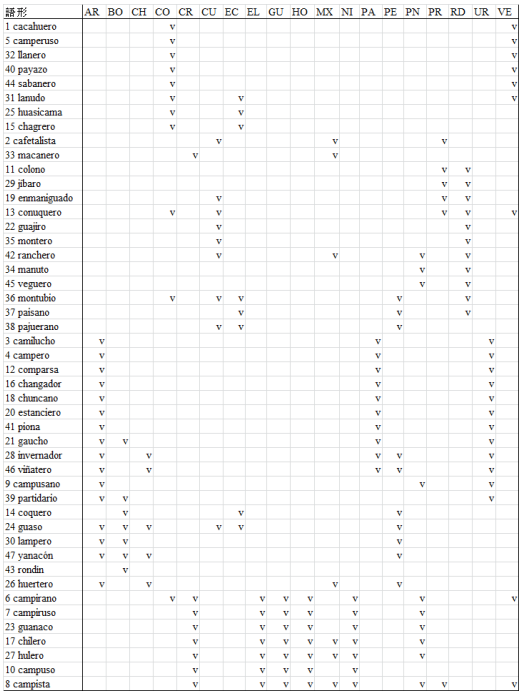
(5) 対応集中分析

語形	AR	PA	UR	BO	CH	PE	EC	CO	VE	CR	EL	GU	HO	NI	PN	MX	CU	RD	PR
1 cacahuero								v	v										
5 camperuso								v	v										
32 llanero								v	v										
40 payazo								v	v										
44 sabanero								v	v										
31 lanudo							v	v	v										
25 huasicama							v	v											
15 chagrero							v	v											
2 cafetalista																v	v		v
33 macanero										v						v			
11 colono																		v	v
29 jibaro																		v	v
19 enmaniguado																	v	v	v
13 conuquero								v	v									v	v
22 guajiro																	v	v	
35 montero																		v	v
42 ranchero															v	v	v	v	
34 manuto															v			v	
45 veguero															v			v	
36 montubio					v	v	v										v	v	
37 paisano					v	v												v	
38 pajuerano					v	v											v		
3 camilucho	v	v	v																
4 campero	v	v	v																
12 comparsa	v	v	v																
16 changador	v	v	v																
18 chuncano	v	v	v																
20 estanciero	v	v	v																
41 piona	v	v	v																
21 gaucho	v	v	v	v															
28 invernador	v	v	v		v	v													
46 viñatero	v	v	v		v	v													
9 campusano	v		v												v				
39 partidario	v		v	v															
14 coquero				v		v	v												
24 guaso	v			v	v	v	v											v	
30 lampero	v			v		v													
47 yanacón	v			v	v	v													
43 rondin				v															
26 huertero	v				v	v										v			
6 campirano								v	v	v	v	v	v	v	v				
7 campiruso										v	v	v	v	v	v				
23 guanaco										v	v	v	v	v	v				
17 chilero										v	v	v	v	v	v	v			
27 hulero										v	v	v	v	v	v	v			
10 campuso										v	v	v	v	v					
8 campista										v	v	v	v	v	v	v			v

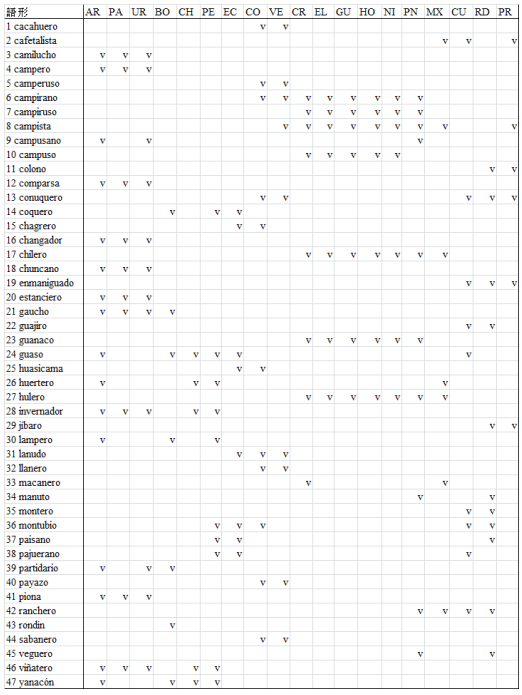
(6) クラスター集中分析

これらの図を見ると、(2) 原点偏差法、(3)主成分分析、(5) 対応分析は分布の対角化を示し、 (4)因子分析と(6)クラスター分析では分布の局在化を示していることがわかります。

主成分分析と因子分析は両軸で分析しますが、原点偏差法、対応分析、クラスター分析では列または行だけでも分析可能です。しかし、クラスター分析では 1 軸（列または行）だけでは次のようにあまり良い局在化の結果が得られません。



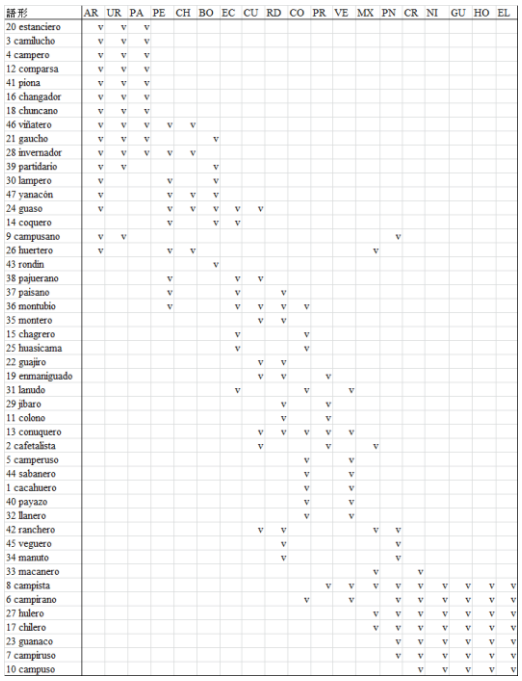
語形	AR	BO	CH	CO	CR	CU	EC	EL	GU	HO	MX	NI	PA	PE	PN	PR	RD	UR	VE
1 cacahuero				v															v
5 camperuso				v															v
32 llanero				v															v
40 payazo				v															v
44 sabanero				v															v
31 lanudo				v			v												v
25 huasicama				v			v												
15 chagrero				v			v												
2 cafetalista						v					v					v			
33 macanero					v						v								
11 colono																v	v		
29 jibaro																v	v		
19 enmariguado						v										v	v		
13 conuquero				v		v										v	v		v
22 guajiro						v											v		
35 montero						v											v		
42 ranchero						v					v				v		v		
34 manuto															v		v		
45 veguero															v		v		
36 montubio				v		v	v							v			v		
37 paisano							v							v			v		
38 pajuerano						v	v							v					
3 camilucho	v												v					v	
4 campero	v												v					v	
12 comparsa	v												v					v	
16 changador	v												v					v	
18 chuncano	v												v					v	
20 estanciero	v												v					v	
41 piona	v												v					v	
21 gaucho	v	v											v					v	
28 invernador	v		v										v	v				v	
46 viñatero	v		v										v	v				v	
9 campusano	v														v			v	
39 partidario	v	v																v	
14 coquero		v						v						v					
24 guaso	v	v	v			v	v							v					
30 lampero	v	v												v					
47 yanacón	v	v	v											v					
43 rondin		v																	
26 huertero	v		v								v			v					
6 campirano				v	v			v	v	v		v			v				v
7 campiruso					v			v	v	v		v			v				
23 guanaco					v			v	v	v		v			v				
17 chilero					v			v	v	v	v	v			v				
27 hulero					v			v	v	v	v	v			v				
10 campuso					v			v	v	v		v							
8 campista					v			v	v	v	v	v			v	v			v



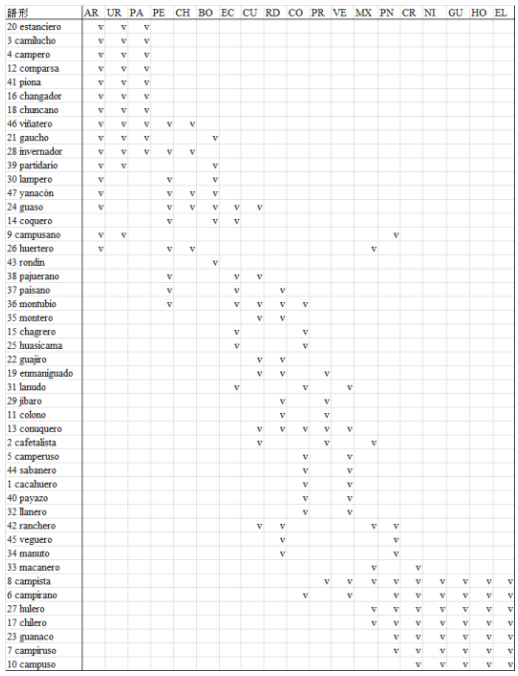
語形	AR	PA	UR	BO	CH	PE	EC	CO	VE	CR	EL	GU	HO	NI	PN	MX	CU	RD	PR
1 cacahuero								v	v										
2 cafetalista																v	v		v
3 camilucho	v	v	v																
4 campero	v	v	v																
5 camperuso								v	v										
6 campirano								v	v	v	v	v	v	v	v				
7 campiruso										v	v	v	v	v	v				
8 campista									v	v	v	v	v	v	v	v			v
9 campusano	v		v												v				
10 campuso										v	v	v	v	v					
11 colono																		v	v
12 comparsa	v	v	v																
13 conuquero								v	v								v	v	v
14 coquero					v		v												
15 chagrero							v		v										
16 changador	v	v	v																
17 chilero										v	v	v	v	v	v	v			
18 chuncano	v	v	v																
19 enmariguado																	v	v	v
20 estanciero	v	v	v																
21 gaucho	v	v	v	v													v	v	
23 guanaco										v	v	v	v	v	v				
24 guaso	v			v	v	v	v										v		
25 huasicama							v	v											
26 huertero	v				v	v										v			
27 hulero										v	v	v	v	v	v	v			
28 invernador	v	v	v		v	v													
29 jibaro																		v	v
30 lampero	v			v		v													
31 lanudo							v	v	v										
32 llanero								v	v										
33 macanero										v						v			
34 manuto																v		v	
35 montero																	v	v	
36 montubio							v	v	v								v	v	
37 paisano							v	v										v	
38 pajuerano							v	v									v		
39 partidario	v		v	v															
40 payazo								v	v										
41 piona	v	v	v																
42 ranchero																v	v	v	v
43 rondin				v															
44 sabanero								v	v										
45 veguero																v		v	
46 viñatero	v	v	v		v	v													
47 yanacón	v			v	v	v													

(8) クラスター分析：列　(9) クラスター分析：行

原点偏差による集中化は、たとえば行を外的基準として固定し、列を集中化することができます。そのとき、行の状態によって結果が変わるので、対応分析など他の方法で適した配列を見つけ、それを外的基準にする、という方法が考えられます。次は先の主成分分析と対応分析の結果として得られた行（各国の地理的な配置）を固定し、列を集中化した結果です。



語形	AR	UR	PA	PE	CH	BO	EC	CU	RD	CO	PR	VE	MX	PN	CR	NI	GU	HO	EL
20 estanciero	v	v	v																
3 camilucho	v	v	v																
4 campero	v	v	v																
12 comparsa	v	v	v																
41 piona	v	v	v																
16 changador	v	v	v																
18 chuncano	v	v	v																
46 viñatero	v	v	v	v	v														
21 gaucho	v	v	v				v												
28 invernador	v	v	v	v	v														
39 partidario	v	v					v												
30 lampero	v			v			v												
47 yanacón	v			v	v		v												
24 guaso	v			v	v		v	v	v										
14 coquero				v			v	v											
9 campusano	v	v																	
26 huertero	v			v	v														
43 rondin							v												
38 pajuerano				v				v											
37 paisano				v				v											
36 montubio				v				v	v	v	v								
35 montero								v	v										
15 chagrero								v			v								
25 huasicama								v			v								
22 guajiro									v	v									
19 enmariguado									v	v		v							
31 lanudo								v			v		v						
29 jibaro										v		v							
11 colono										v		v							
13 conuquero									v	v	v	v							
2 cafetalista									v			v		v					
5 camperuso											v		v						
44 sabanero											v		v						
1 cacahuero											v		v						
40 payazo											v		v						
32 llanero											v		v						
42 ranchero												v	v		v				
45 veguero													v		v				
34 manuto													v		v				
33 macanero																v			
8 campista											v	v	v	v	v	v	v	v	v
6 campirano											v		v						
27 hulero														v	v	v	v	v	v
17 chilero														v	v	v	v	v	v
23 guanaco															v	v	v	v	v
7 campiruso															v	v	v	v	v
10 campuso																v	v	v	v

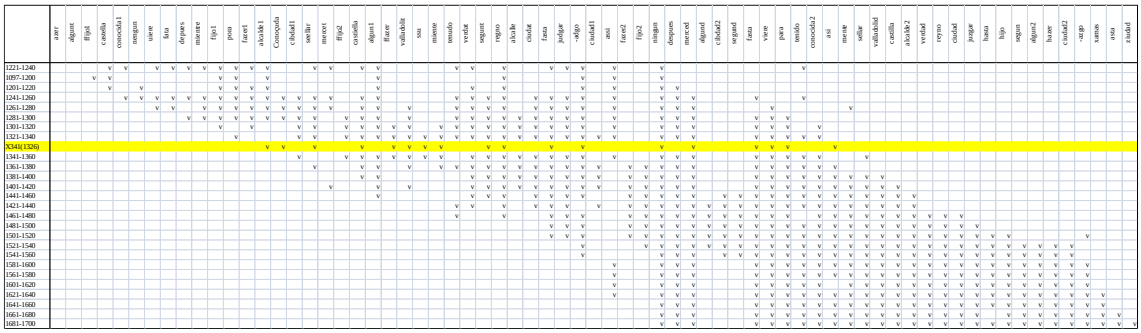


語形	AR	UR	PA	PE	CH	BO	EC	CU	RD	CO	PR	VE	MX	PN	CR	NI	GU	HO	EL
20 estanciero	v	v	v																
3 camilucho	v	v	v																
4 campero	v	v	v																
12 comparsa	v	v	v																
41 piona	v	v	v																
16 changador	v	v	v																
18 chuncano	v	v	v																
46 viñatero	v	v	v	v	v														
21 gaucho	v	v	v				v												
28 invernador	v	v	v	v	v														
39 partidario	v	v					v												
30 lampero	v			v			v												
47 yanacón	v			v	v		v												
24 guaso	v			v	v		v	v	v										
14 coquero				v			v	v											
9 campusano	v	v												v					
26 huertero	v			v	v										v				
43 rondin							v												
38 pajuerano				v				v											
37 paisano				v				v											
36 montubio				v				v	v	v	v								
35 montero								v	v										
15 chagrero								v			v								
25 huasicama								v			v								
22 guajiro									v	v									
19 enmariguado									v	v		v							
31 lanudo								v			v		v						
29 jibaro										v		v							
11 colono										v		v							
13 conuquero									v	v	v	v							
2 cafetalista									v			v		v					
5 camperuso											v		v						
44 sabanero											v		v						
1 cacahuero											v		v						
40 payazo											v		v						
32 llanero											v		v						
42 ranchero												v	v		v				
45 veguero													v		v				
34 manuto													v		v				
33 macanero																v			
8 campista											v	v	v	v	v	v	v	v	v
6 campirano											v		v						
27 hulero														v	v	v	v	v	v
17 chilero														v	v	v	v	v	v
23 guanaco															v	v	v	v	v
7 campiruso															v	v	v	v	v
10 campuso																v	v	v	v

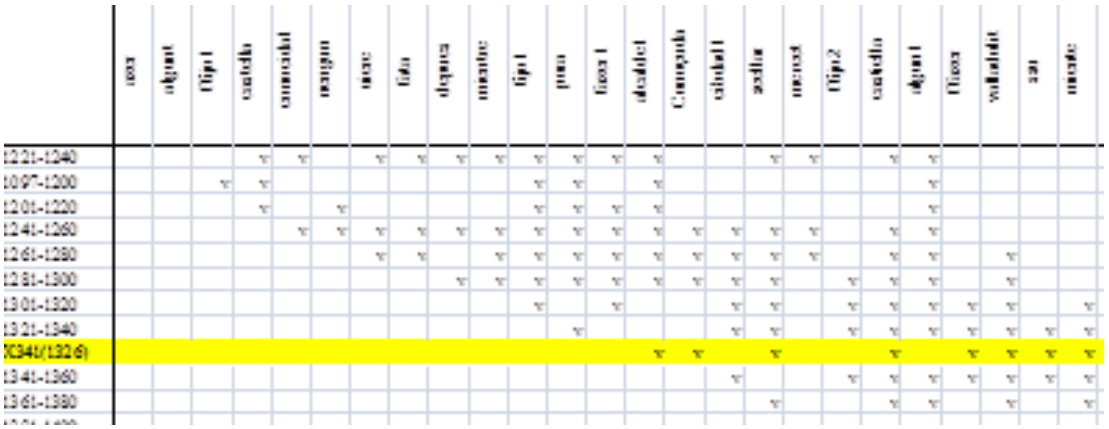
(10) 主成分分析→原点偏差：列　(11) 対応分析→原点偏差：列

■古文献の成立年代推定

次の研究（Kawasaki 2012）は発行日が記されていない中世スペイン語公証文書の年代を、発行日が記されている他の文書の言語特徴の集中化された分布から推定しています。



この図では列に年代を入れ固定し、行（言語特徴）を集中化しています。次に左上の部分を拡大しました。



この黄色の部分の横行が年代不詳の文献です。これを含めて全体を集中化すると、この行が一定の年代に位置づけられるので、その年代を推定することができます。そのためには適切な言語特徴（年代差を示す、頻度が高い、地域差が少ない、など）を選択し、実験を繰り返さなければなりません。

7.2.4. 順列集中分析

これまでに見たように、集中分析にはさまざまな方法があり、当然その分析結果も異なります。各種の分析結果から最終的な結論として選択をするとき、選択した方法の正しい理解と集中行列と集中係数の比較が必要です。

一つの集中分析で、たとえば近接係数が高い数値を示したとしても、それがはたしてすべての可能な集中行列の中で最大の近接係数を示している

のか不明なこともあります[6]。そこで逆に考えて、近接係数が最大になるような配列を探ります。これまでの方法では集中係数は結果として出力されましたが、ここでは逆に分析の条件とします。順列を使って可能な行列をすべて出力し、その中で最大の緊密係数を示す集中行列を「順列近接集中行列」とします。

分析対象とする配列が、たとえば 5 行 4 列ならば、行と列のそれぞれの順列による配列の数は 5! * 4! = 120 * 24 = 2880 個になります。しかし、N 個の順列の数は N が 1 増えるごとに、6! = 720, 7! = 5040, ...というように非常に多くなります。実際には行数の階乗と列数の階乗の積になるので、膨大な数になります。そこで行の並べ方は原点偏差を使って一義的に求められるので、順列は列の並べ方を求めるときにだけ使います。

列の順列の数については、実験によれば 7! = 5040 が実用の範囲内で可能です。これ以上多くなると実行時間が非常に長くなるためです。たとえば 5 行 20 列などの横長の行列の場合は、行列を転置して 20 行 5 列に変えてから分析します。その場合、必要ならば出力行列を転置して入力行列と同じ行数と列数に戻します。行数と列数のどちらも 7 個を超えるときは、どちらかをクラスター分析などで分類し 7 個以内にしなければなりません。

下表(D)は集中化していない入力行列と各集中係数を示します。

D	AR	BO	CO	CU	EL	ES	GU	PN	PR	RD
esp.				+				+	+	+
gaf.	+		+			+		+		+
ant.	+		+		+		+	+	+	
len.	+	+					+	+		+

Cor.n	Value
連番相関係数	-.208
参照相関係数	-.208
連番近接係数	.546
参照近接係数	-.157
GK 順序係数	-.209
緊密係数	.541

上のデータを対応分析と原点偏差法で集中化した結果が下表(Z)です。

Z	BO	ES	AR	CO	GU	EL	PN	RD	PR	CU
gaf.		+	+	+			+	+		
len.	+		+		+		+	+		
ant.			+	+	+	+	+		+	
esp.							+	+	+	+

nZD.a	Value
連番相関係数	.478
参照相関係数	.571
連番近接係数	.682
参照近接係数	.425
GK 順序係数	.463
緊密係数	.769

上の結果と、次の順列連番近接集中分析の結果(P)を比較すると、参照相関係数は下降していますが(.571 > .559)、連番係数が上昇しています(.682 > .702)。上表では左上部分が欠けていますが、下表では満たされていて、

[6] 対応分析が出力した参照相関係数は理論的に最大になるので例外です。

両軸値の近接性が高くなります。つまり対角部分に強く集中しています。連番相関係数(.565)や順序係数(.537)も高いことにも注目すべきでしょう。

P	CU	ES	PR	CO	RD	PN	EL	AR	GU	BO
esp.	+		+		+	+				
gaf.		+		+	+	+		+		
ant.			+	+		+	+	+	+	
len.					+	+		+	+	+

P	Value
連番相関係数	.565
参照相関係数	.559
連番近接係数	.702
参照近接係数	.604
GK 順序係数	.537
緊密係数	.769

次の 2 個の表は GK 順序係数 G)と緊密係数(T)を最大化した集中行列です。

G	ES	CO	EL	AR	GU	PN	BO	RD	PR	CU
gaf.	+	+		+		+		+		
ant.		+	+	+	+	+			+	
len.				+	+	+	+	+		
esp.						+		+	+	+

G.	Value
連番相関係数	.559
参照相関係数	.573
連番近接係数	.696
参照近接係数	.416
GK 順序係数	.552
緊密係数	.741

T	ES	BO	AR	CO	GU	RD	PN	EL	PR	CU
gaf.	+		+	+		+	+			
len.		+	+		+	+	+			
ant.			+	+	+		+	+	+	
esp.						+	+		+	+

T	Value
連番相関係数	.552
参照相関係数	.577
連番近接係数	.684
参照近接係数	.447
GK 順序係数	.522
緊密係数	.800

以上のすべての集中行列で、それぞれの行(Row)と列(Col.)の参照値を計算すると、どれも単調に上昇しています。つまり、上→下と、左→右の原点偏差の連続性は完全に保たれているのです。次の表(T)は緊密係数による集中行列の行と列の参照値を示しています。

T	ES	BO	AR	CO	GU	RD	PN	EL	PR	CU
gaf.	+		+	+		+	+			
len.		+	+		+	+	+			
ant.			+	+	+		+	+	+	
esp.						+	+		+	+

Row	Xn
gaf.	.000
len.	.068
ant.	.388
esp.	1.000

Col.	ES	BO	AR	CO	GU	RD	PN	EL	PR	CU
Yp	.0	.333	.430	.470	.532	.633	.641	.667	.857	1.000

このように、順列集中分析を行うことで、考えられる係数の数と同じだけ多くの完全な集中行列があり得ることがわかりました。そこで、内的な基準は、縦因子と横因子が循環的に依存しているので、それにしたがって出力される集中行列も、その両者に依存しています。因子と行列が循環的に依存しあう状況の中で唯一決定権があるのが集中係数です。つまり、集中係数を決めることによって、それが最大となる集中行列と縦と横の因子ベクトルが（参照値）が決まります。

それでは、どの集中行列を選択したらよいのでしょうか？この問題は、分析の目的に従う、と考えます。データの分布の相関性を重視するならば、当然相関係数を最大にする集中行列を探すべきです。そのとき、集中行列の縦因子と横因子の参照値ベクトルによって計算した相関係数を使えば、集中行列の本質的な相関を示す値になります。一方、単純に行や列の連番を使うと、行列が同じ間隔で区切られた升目の位置による相関を見ることになるので表面的な観察になります。参照近接係数と連番近接係数の違いについても同様です。近接係数を使った順列集中分析では、表の対角線上にできるだけ近接させるデータの配列を探します。データの反応点の流れができるだけ左上から右下に流れるようにしたいときには、順序係数を使います。データの緊密度を重視するときには、つまりデータの連続性を損なうギャップが少なくなるような集中化をするときには、緊密係数を使った順列集中分析を行います。

●順列の数と並べ方

たとえば(1, 2)のような 2 個の数字の並べ方は(1, 2), (2, 1)という 2 通りあり、(1, 2, 3)のような 3 個の数字の並べ方は(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)という 6 通りあります。このような並べ方は「順列」(permutation)とよばれます。順列の数を求めるには、成分が 2 個ならば 2 の階乗(2! = 2 * 1 =1), 成分が 3 個ならば 3 の階乗(3! = 3 * 2 * 1 = 6)という計算をします。よって、成分が N 個ならば N の階乗 N!で求めることができます。次が N の階乗を求めるプログラムです。

```
Function factorial(N) 'N の階乗
```

```
    Dim i%: factorial = 1 '初期値
    For i = 1 To N
        factorial = factorial * i
    Next
End Function
```

そして、たとえば(1, 2, 3)のような 3 個の数字の並べ方の中で、2 番目の(1, 3, 2)を求めるには、次々に列挙する代わりに、次のような計算をします。N 個の数字の並べ方を、(1, 2, …, N)から(N, N-1, …, 1)にしたとき、その番数 P に当たる数字の並べ方（ベクトル）を求めます。たとえば、(1,2,3,4)のような 4 個の数字な並べ方は 4*3*2*1 = 4! = 24 通りあり、それらは末尾の数字の位置を変換しながら列挙すると次のようになります。

P 番：
 0 番：(1,2,3,4)
 1 番：(1,2,4,3)
 2 番：(1,3,2,4)
 3 番：(1,3,4,2)
 4 番：(1,4,2,3)
 5 番：(1,4,3,2)
 6 番：(2,1,3,4)
 7 番：(2,1,4,3) <==
        (…)
23 番：(4,3,2,1) <==

上の P「番」はゼロ(0)番から 23 番までとします。(1, 2, 3, 4)のように並んだ最初の数字を N(1)とすると、N(1)が 1 であるのが 0 番から 5 番までで、6 番からは N(1)は 2 になっています。N(1)には(1, 2, 3, 4)のいずれかが入りますが、それは番の数を(4 - 1)! = 3! = 6 で割ったときの整数部で決まります。たとえば P=7 番については、7/6 = 1…1 (1 余り 1)ですから、ベクトル(1, 2, 3, 4)の 1 番(=ベクトルの 2 番目)の数字=2 になります。このように成分の「番」の開始もゼロ(0)にします。ここで確認することは、最初の成分 2 は、P=7 を 3!=6 で割った商(1)として求めたことです。上の図を見ると、わかるように、最初の成分が 6 回ずつ繰り返されるからです。次の成分は、2(=2!)回ずつ繰り返されること、そして、その次の成分が 1(1!)回ずつ繰り返されることも確認しておきましょう。そして最後の成分も 0!(=1)回ずつ繰り返されています。以下では、この規則性を利用します。

次の成分 N(2)は、N(1)に入った数字を除いた 3 個の数字(1, 3, 4)のいずれかが入ります。P=7 番の N(2)は先の N(1)のときの割り算で余った数=1 を 2!で割ります。1 / 2! = 0…1、よって(1, 3, 4)の 0 番(1 番目)の数字=1 になります。

次の成分 N(3)を求めるために、先の割り算の余り(1)を 1!で割ります。1/

1! = 1...0、よって N(3)は(3, 4)の 1 番(2 番目)の数字=4 になります。

最後は(3)に決まっていますが、操作を一般化すると、先の余り 0 を 0! で割ります。0 / 0! = 0...0、よって(3)という 1 成分のベクトルの 0 番の数字 3 が選択されます。

このようにして、(2, 1, 4, 3)という順列のベクトル(P=7)が得られます。

以上の操作をまとめると

$$7 = 3!*(1) + 2!*(0) + 1!*(1) + 0!*(0) \leftarrow 6*1 + 2*0 + 1*1 + 1*0 = 7$$

どのような整数であっても上のように階乗と整数の積和に分解できます。整数を階乗と整数の積和にすることを「階乗分解」(factorial decomposition: FD)と呼ぶと、p=4 成分の b=7 番の順列の成分 FD(p, b)は

$$FD(4, 7) = (1, 0, 1, 0) \leftarrow 7 = 3!*(1) + 2!*(0) + 1!*(1) + 0!*(0)$$

N(1): 3!が 1 個→(1, 2, 3, 4)の 1 番(2 番目)=2
N(2): 2!が 0 個→(1, 3, 4)の 0 番(1 番目)=1
N(3): 1!が 1 個→(3, 4)で 1 番(2 番目)=4
N(4): 0!が 0 個→(3)の 0 番(1 番目)=3

よって求めるベクトルは (2, 1, 4, 3) になります。階乗分解 FD(N, P)の最初の階乗は P を割った商が整数になる最大の階乗です。先の例では、7 を割って商が整数になるのは、1!, 2!, 3!までで、4!=24 では商が小数になってしまいます。よって、最初の階乗を 3!として、これから割り算の余りを次々に階乗分解していきます。順列集中分析では次のプログラムを使います。

```
Function factDecompos(P, B) '階乗分解: P=成分数、B=番数
    Dim Xp, Zp, i&, Sp&, Ft&, Fp&, Ix&
    Fp = factorial(P): ReDim Xp(Fp), Zp(1, P): Sp = B '余り
    For i = 1 To Fp: Xp(i) = i: Next '連番
    For i = P - 1 To 0 Step -1
      Ft = factorial(i): Ix = Sp ¥ Ft '階乗 : 商:
      Zp(1, P - i) = Xp(Ix + 1) '返し配列
      Sp = B Mod Ft:   Xp = delElement(Xp, Ix + 1) '余り : 要素削減配列
    Next
    factDecompos = Zp '(2, 1, 4, 3)などの横ベクトル
End Function
```

＊いたかなや(2014)「順列を一瞬で取得するプログラム」を参照しました。
http://itakanaya9.hatenablog.com/entry/2014/02/17/121428 (参照 : 2016/6/25)

7.2.5. 集中バブルチャート

プログラムは下中表の集中行列の縦因子(Xn)と横因子(Yp)の参照値を考慮したバブルチャートを出力します。

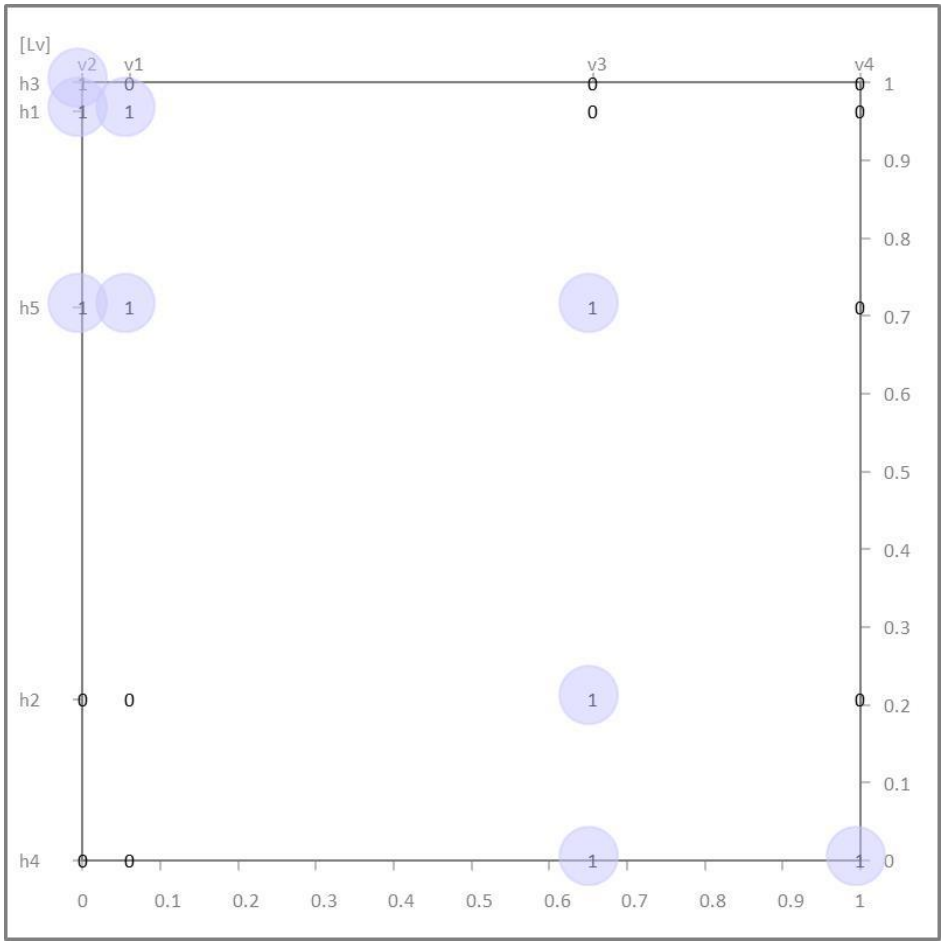
Lv	v1	v2	v3	v4
h1	+	+		
h2			+	
h3		+		
h4			+	+
h5	+	+	+	

Cor.a	v2	v1	v3	v4
h3	+			
h1	+	+		
h5	+	+	+	
h2			+	
h4			+	+

Row	Xn
h3	
h1	.037
h5	.289
h2	.793
h4	1.000

Column	v2	v1	v3	v4
Yp		.061	.657	1.000

上の集中行列では、すべての反応点の間隔と位置は連番によって均一に扱われていますが、次のチャートでは、原点からの偏差を参照値として、軸を設定しているので、それぞれの反応点の本質的な間隔と位置がわかります。



7.3. クラスター分析

関係行列（相関行列、連関行列、距離行列、文字行列）や、多変数分析の結果を見ると互いに関係の深い成分とそうでない成分があることがわかります。こうした関係にもとづいて全体がどのようなグループ（群）に分類されるのかを見る手法の１つが「クラスター分析」(Cluster analysis)です。「樹形図」(Dendrogram)というグラフを出力します。

(1) 最近隣法

クラスター分析には多くの方法があります。はじめに一番簡単な「最近隣法」(Nearest neighour method)を取り上げましょう。スペイン語圏の語彙バリエーション研究から得られた相関係数行列を用いて説明します。データの規模を小さくして 6 カ国だけにしたサンプルデータを使います。それぞれ ES:スペイン, GE:赤道ギニア, CU:キューバ, RD:ドミニカ共和国, PR:プエルトリコ, MX:メキシコを示します。

6 か国	1. ES	2. GE	3. CU	4. RD	5. PR	6. MX
1. ES	1.00					
2. GE	0.61	1.00				
3. CU	0.51	0.45	1.00			
4. RD	0.54	0.45	0.54	1.00		
5. PR	0.58	0.49	0.56	0.68	1.00	
6. MX	0.45	0.34	0.39	0.45	0.50	1.00

これを距離行列に変換します。→3.6.4.

6 か国	1. ES	2. GE	3. CU	4. RD	5. PR	6. MX
1. ES	0.00	0.20	0.25	0.23	0.21	0.28
2. GE	0.20	0.00	0.28	0.27	0.25	0.33
3. CU	0.25	0.28	0.00	0.23	0.22	0.30
4. RD	0.23	0.27	0.23	0.00	0.16	0.28
5. PR	0.21	0.25	0.22	0.16	0.00	0.25
6. MX	0.28	0.33	0.30	0.28	0.25	0.00

最初のクラスタリングで距離の最小値(0.16)をもつ組み合わせである 4:RD と 5:PR が合体します。

6 か国	1. ES	2. GE	3. CU	4. RD:5. PR	6. MX
1. ES	0.00	0.20	0.25	0.21	0.28
2. GE	0.20	0.00	0.28	0.25	0.33
3. CU	0.25	0.28	0.00	0.22	0.30

4. RD: PR	0.21	0.25	0.22	0.00	0.25
6. MX	0.28	0.33	0.30	0.25	0.00

2 番目のクラスタリングで次に距離が近い値(.20)をもつ成分 1 と成分 2 が合体します。

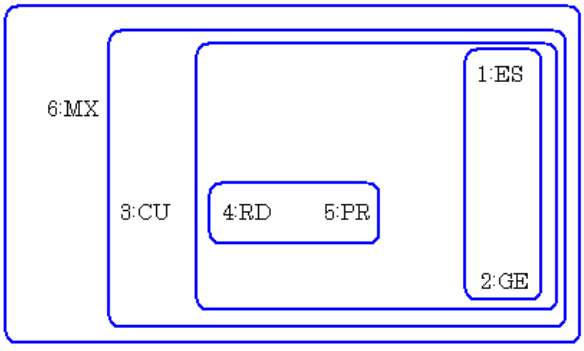
6 か国	1. ES: 2. GE	3. CU	4. RD: .16;5. PR	6. MX
1. ES: 2. GE	0.00	0.25	0.21	0.28
3. CU	0.25	0.00	0.22	0.30
4. RD: 5. PR	0.21	0.22	0.00	0.25
6. MX	0.28	0.30	0.25	0.00

3 番目のクラスタリングではすでに存在する(1+2)のグループと(4+5)のグループが合体します。このとき、(1+2) と 3 の距離は、1-3, 2-3 の間のそれぞれの距離のうち小さなほうの値とします。同様に(1+2)と(4+5)の距離は 1-4, 1-5, 2-4, 2-5 の中で一番小さな値をとります。以下同様にして最後の 5 番目のクラスタリングで成分 6 が全体に組み込まれます。

6 か国	1. ES: 2. GE:4. RD:5. PR	3. CU	6. MX
1. ES: 2. GE:4. RD:5. PR	0.00	0.22	0.25
3. CU	0.22	0.00	0.30
6. MX	0.25	0.30	0.00

6 か国	1. ES: 2. GE: 4. RD:5. PR: ;3. CU	6. MX
1. ES: 2. GE: 4. RD:5. PR: ;3. CU	0.00	0.25
6. MX	0.25	0.00

各国を空間に配置しそのグルーピングを行うと次のようになります。

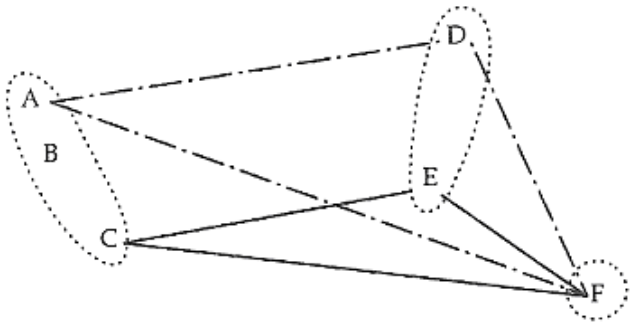


Nearest	R.	Max. 1.00 + Min. 0.00
1. ES	-1.00	
2. GE	0.61	
4. RD	0.58	
5. PR	0.68	
3. CU	0.56	
6. MX	0.50	

＊この例では相関係数行列を一度距離に置き換えてからクラスター分析にかけていますが、上の図（樹形図）にはクラスターの合流点として入力の数値（相関係数）が出力されています。

(2) 最遠隣法

最近隣法ではグループと 1 つの成分またはグループ間の距離をグループを構成する成分のあらゆる組み合わせのペアで一番距離の近い数値を示すものとして定義しました。たとえば次の図で



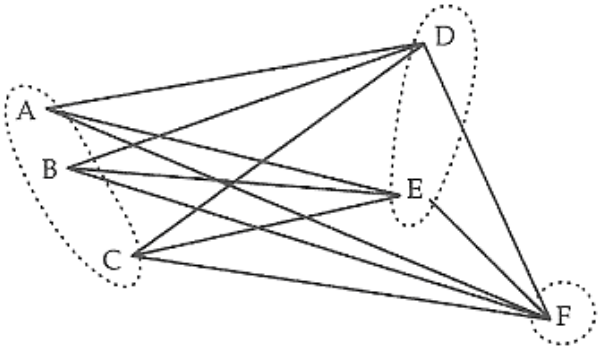
[A+B+C]というグループと[D+E]というグループの間の距離を A-D, A-E, B-D, B-E, C-D, C-E というペアの中から一番近いものを選んで、この場合、C-E によって、2 つのグループ間の距離と見なしています。F も含めた 3 つのグループの距離は、C-E, C-F, E-F によって計測されます。

次に取り上げる「最遠隣法」(Furthest neighbour method)はグループ間の最も遠い成分の間の距離を採用します。つまり、上図の A-D, A-F, D-F の距離で 3 つのグループの距離と見なすのです。最近隣法では近くにデータがあれば、それを結びつけるという手法をとりますが、最遠隣法では一番遠くにあるデータを見つけ、これの距離を 2 つのグループの距離とする点が違います。グループ間の距離が決定されたならば、あとの操作は同じです。

Farthest	R.	Max. 1.00 + Min. 0.00
1. ES	-1.00	
2. GE	0.61	
3. CU	0.45	
4. RD	0.54	
5. PR	0.68	
6. MX	0.34	

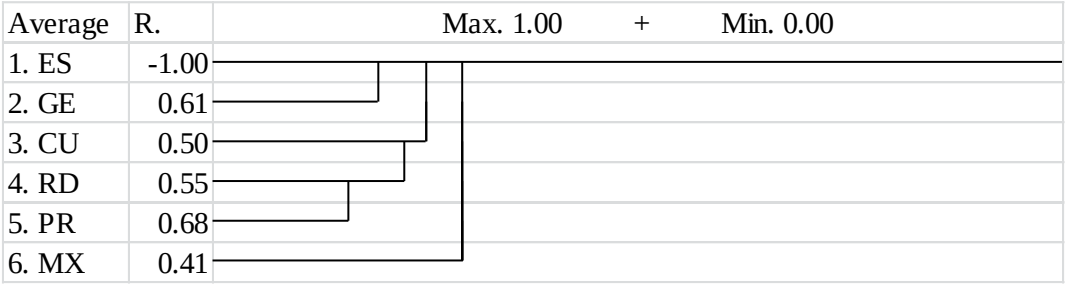
(3) 平均結合法

最近隣法と最遠隣法はグループ間の距離を決定するのに正反対の考え方をしています。しかし、グループ間の距離を 1 つの代表値で計算していることではどちらも同じです。ここで取り上げる「平均結合法」(Mean linkage method)はグループ間の距離を前二者のように単純にそれぞれのグループの 1 成分に代表させず、すべての組み合わせのペアの距離の平均値をもって 2 つのグループの距離と見なす手法です。たとえば、次の図で



[A+B+C]というグループと[D+E]というグループ間の距離を A-D, A-E, B-D, B-E, C-D, C-E というペアのすべての距離を足して、ペアの数(6)で割った値を 2 つのグループ間の距離と見なします。

先の距離行列の中で、すべての距離の中で最小値は 4:RD と 5:PR の間の.16 です。最初にこの 2 国を 1 つのグループをなすと見なすのは最近隣法や最遠隣法と同じです。新しいグループ名を(4+5)と名付けて、新たに相関行列を作成します。このときグループ (4+5)と 1, 2, 3, 6 との相関係数は、それぞれの組み合わせの平均値とします。これが群間平均法の要点です。たとえば、1 と(4+5)では、1-4 の.23 と 1-5 の.21 を足して 2 で割ります。以下、2, 3, 6 についても同様に比較します。最後に次の図が得られます。



(4) 過程平均結合法

クラスター分析法にはほかにも多くの手法があります。これまで扱ってきた 3 つの手法は代表的なものですが、どれも原初の対称行列の成分をもとに距離を測っています。ここで提案する「過程平均法」は平均結合法に連関しますが、クラスタリングの各ステップで、原初の対称行列の成分に戻るのではなく、ステップを踏むときの対称行列の状態をもとに、新しく距

離を平均して求めます。

次は「成績 1」のデータ行列（下左）からマハラノビス距離（下右）を計算した結果です（平均化、最大値比：→3.6.4 (3)）。

項目	a.**役立つ**	b.楽しい
A.文法解説	86	29
B.ビデオ	53	78
C.活動	48	53
D.映画	43	96
E.音読	110	42
F.筆写	93	11
G.観察	37	50
H.小テスト	89	15

S	A	B	C	D	E	F	G	H
A	0.00	0.48	0.40	0.66	0.50	0.19	0.56	0.17
B	0.48	0.00	0.40	0.18	0.60	0.66	0.58	0.62
C	0.40	0.40	0.00	0.54	0.81	0.48	0.19	0.44
D	0.66	0.18	0.54	0.00	0.70	0.84	0.70	0.80
E	0.50	0.60	0.81	0.70	0.00	0.63	1.00	0.63
F	0.19	0.66	0.48	0.84	0.63	0.00	0.59	0.04
G	0.56	0.58	0.19	0.70	1.00	0.59	0.00	0.55
H	0.17	0.62	0.44	0.80	0.63	0.04	0.55	0.00

はじめに F+H が全体の最短距離(.04)によって結合します。

S	A	B	C	D	E	[F+H]	G
A	0.00	0.48	0.40	0.66	0.50	0.18	0.56
B	0.48	0.00	0.40	0.18	0.60	0.64	0.58
C	0.40	0.40	0.00	0.54	0.81	0.46	0.19
D	0.66	0.18	0.54	0.00	0.70	0.82	0.70
E	0.50	0.60	0.81	0.70	0.00	0.63	1.00
[F+H]	0.18	0.64	0.46	0.82	0.63	0.02	0.57
G	0.56	0.58	0.19	0.70	1.00	0.57	0.00

結合した[F+H]と他の成分、たとえば A との距離 $D_{([F+H]:A)}$は次のように計算されています。

$$D_{([F+H]:A)} = [D_{(F:A)} + D_{(H:A)}] / 2 = (.19 + .17) / 2 = .18$$

他も同様です。これは平均結合法と同じです。次に上の表の中での最短距離(.18)をもつ[B+D]が結合します。

D	A	[B+D]	C	E	[F+H]	G
A	0.00	0.57	0.40	0.50	0.18	0.56
[B+D]	0.57	0.09	0.47	0.65	0.73	0.64
C	0.40	0.47	0.00	0.81	0.46	0.19
E	0.50	0.65	0.81	0.00	0.63	1.00
[F+H]	0.18	0.73	0.46	0.63	0.02	0.57
G	0.56	0.64	0.19	1.00	0.57	0.00

上と同様に[B+D]に関わる距離が再計算されています。次のステップで

[A+[F+H]]という群が形成されます（最短距離：.18）。

D	[A+[F+H]]	[B+D]	C	E	G
[A+[F+H]]	0.09	0.65	0.43	0.57	0.57
[B+D]	0.65	0.09	0.47	0.65	0.64
C	0.43	0.47	0.00	0.81	0.19
E	0.57	0.65	0.81	0.00	1.00
G	0.57	0.64	0.19	1.00	0.00

このとき過程平均法では、たとえば[A+[F+H]]と[B+D]の距離を次の式で計算します。上の表ではなく直前のステップの表から $D_{(A:[B+D])}$ と $D_{([F+H]:[B+D])}$ に該当する値を求めます。

$D_{([A+[F+H]]:[B+D])} = [D_{(A:[B+D])} + D_{([F+H]:[B+D])}] / 2 = (.57 + .73) / 2 = .65$

＊一方、群平均法では、この計算を原初の対称行列に戻って次の式を適用しました。

$D_{([A+F+H]:[B+D])} = [D_{(A:B)} + D_{(A:D)} + D_{(F:B)} + D_{(F:D)} + D_{(H:B)} + D_{(H:D)}] / 6 = 67.7$

過程平均法における距離の再計算法として幾何平均を使用する次を提案します。先の最初のステップの例で示すと次のようになります。

$D_{([F+H]:A)} = [D_{(F:A)}\ D_{(H:A)}]^{1/2} = (.19 \times .17)^{1/2} = .18$

この結果は先とほとんど変わりませんが、多くの計算では結果にかなりの影響が出ます。先の算術平均をとる方法を「過程算術平均結合法」とよび、今回の幾何平均をとる方法を「過程幾何結合平均法」とよぶことにします。

次は、これまで扱った 5 つの方法を同じデータに適用して比較した結果です。

(1) 最近隣法

Nearest	D.	Min. 0.000 + Max. 1.000
A.文法解説	1.000	
F.筆写	0.165	
H.小テスト	0.044	
B.ビデオ	0.402	
D.映画	0.179	
C.活動	0.400	
G.観察	0.193	
E.音読	0.496	

(2) 最遠隣法

Furthest	D.	Min. 0.000 + Max. 1.000
A.文法解説	1.000	
F.筆写	0.190	
H.小テスト	0.044	
C.活動	0.587	
G.観察	0.193	
B.ビデオ	1.000	
D.映画	0.179	
E.音読	0.703	

(3) 平均結合法

Average	D.	Min. 0.000 + Max. 1.000
A.文法解説	1.000	
F.筆写	0.178	
H.小テスト	0.044	
C.活動	0.503	
G.観察	0.193	
B.ビデオ	0.628	
D.映画	0.179	
E.音読	0.696	

(4) 過程算術平均結合法

P. A. Av.	D.	Min. 0.000 + Max. 1.000
A.文法解説	1.000	
F.筆写	0.178	
H.小テスト	0.044	
C.活動	0.497	
G.観察	0.193	
B.ビデオ	0.602	
D.映画	0.179	
E.音読	0.694	

(5) 過程幾何平均結合法

P. G. Av.	D.	Min. 0.000 + Max. 1.000
A.文法解説	1.000	
F.筆写	0.177	
H.小テスト	0.044	
C.活動	0.492	
G.観察	0.193	
B.ビデオ	0.590	
D.映画	0.179	
E.音読	0.680	

5 つの方法を比較すると結果は連関していますが、最近隣法が他の方法に比べて分類する力が弱いことがわかります。他の 4 つの方法ではそれぞれの結合点が異なっています。一般に結合点が最小値に近いほどクラスターが原点に近い位置で形成されているので分類能力があると解釈できます。上の例では過程幾何平均法が全体的に結合点が小さな値になっています。

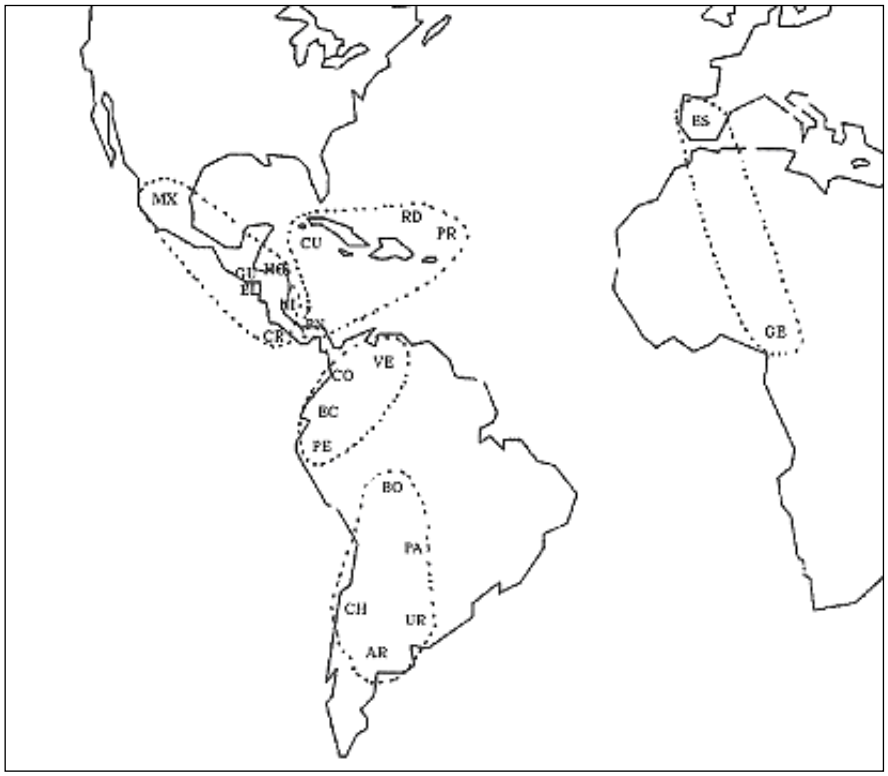
■地域語彙変異によるクラスター分析

クラスター分析はさまざまな分野で使われています。その理由のひとつとして他の多変数解析法と比べて理解しやすく、また結果も明示的でわかりやすいことが挙げられるでしょう。

連関度係数として何を使うか、また、クラスタリングアルゴリズムをどれにするかで、さまざまな組み合わせが可能です。それぞれの性質をよく理解しデータの特徴や先行研究を踏まえたうえで納得できる結論を導くようにしたいと思います。

コンピュータは一定の条件さえ与えれば、それなりの答えを出してくれますが、これは可能な分析法の一つにすぎません。他の方法による結果と比較しながら総合的に判断すべきです。

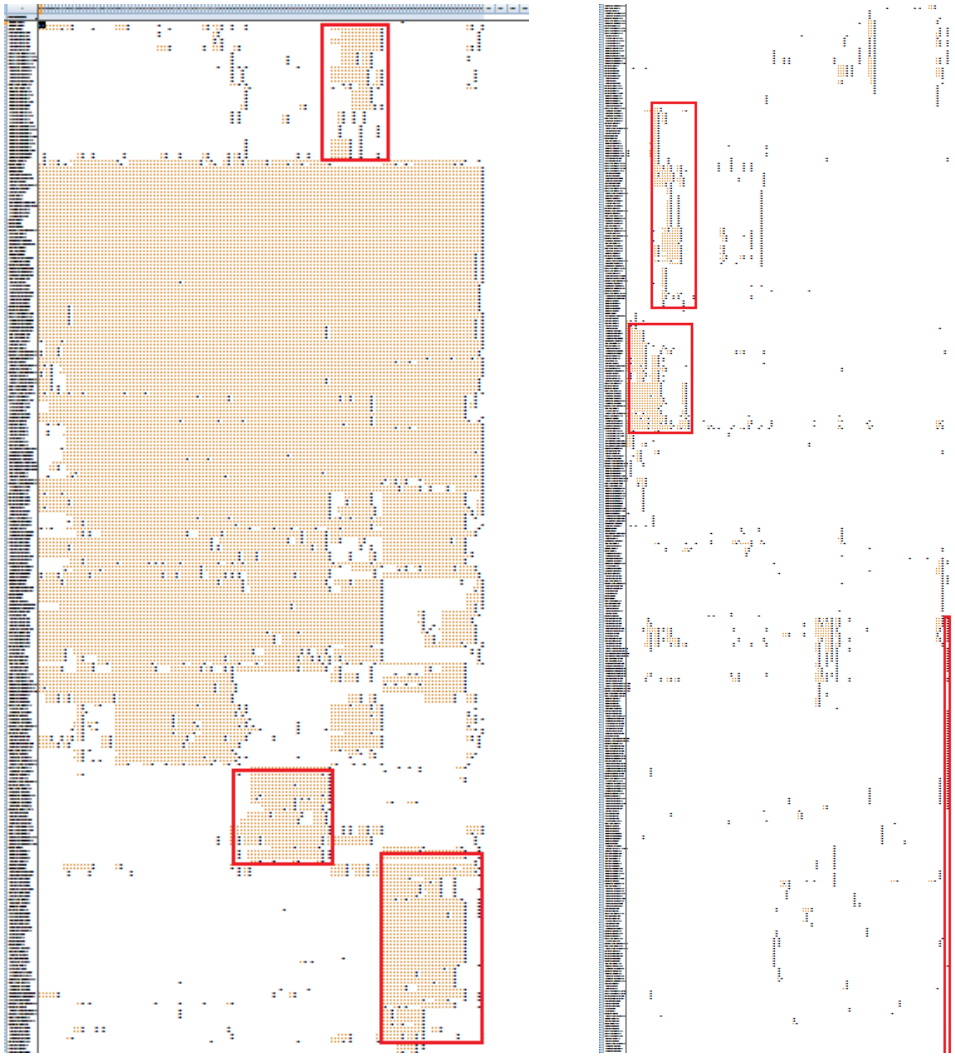
次は、スペイン語の語彙変異によるスペイン語圏地域をクラスター分類したものです。全体はスペイン・アフリカ、カリブ海地域、メキシコ・中米、南米北部、アンデス・ラプラタに分類されました。このような分類は、異なる言語特徴を選択しても、しばしば同じ結果になります。



クラスター分析：スペイン語圏の語彙バリエーション

■大規模データのクラスター集中分析

次は、カタルニア語の動詞形態の地理分布を列（動詞形態）と行（地点）でクラスター分析し、それぞれを集中化した結果です。大きな分布の塊の他に、一定の語形と地点で収集した部分（赤い線で囲みました）が観察されます。その部分についての語形の特徴を探ると、地理的な基準ではなく言語的な基準から地域を確定することができます。また、逆に、そのように確定された地域の言語特徴を抽出することができます。



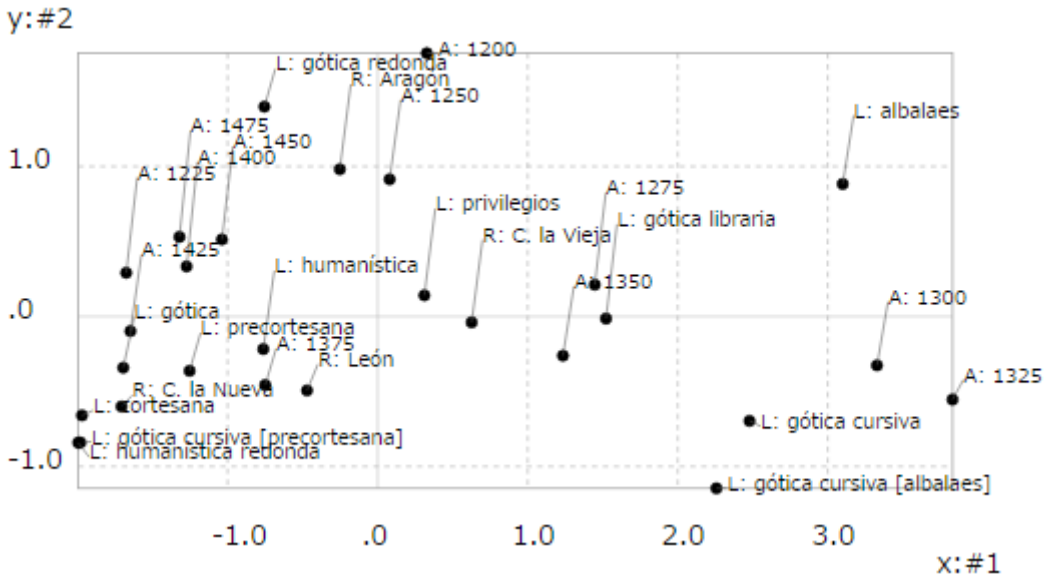
＊プログラムは奥村(1986:170-180)を参照しました。

●グループ分析

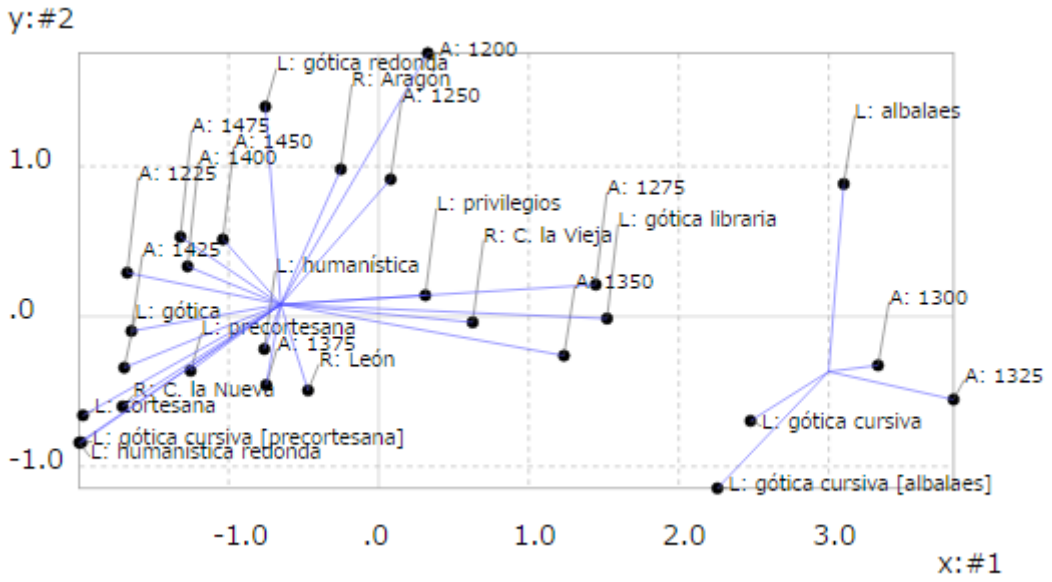
次のデータはスペイン語公証文書の地域(R)、年代(A)、書体(L)と二重字 ss, ff の語頭(#ss, #ff)と語中(-ss-, -ff-)に現れた正規化頻度（千語率）を示します。

行	N.D4.ss-ff	-ss-	#ss-	-ff-	#ff-
1	R: León	4.8	5.8	1.4	6.1
2	R: C. la Vieja	9.7	9.6	1.5	10.1
3	R: C. la Nueva	1.8	2.4	0.5	1.9
4	R: Aragón	11	0.9	2.1	3
5	A: 1200	16.8	3.1	1.2	6.2
6	A: 1225	5.7	0.2	0.5	0
7	A: 1250	12.2	3.1	1.5	6.8
8	A: 1275	13.1	13.8	1.5	12.4
9	A: 1300	13.4	14.8	4	22.1
10	A: 1325	14.3	20.5	3.6	23.8
11	A: 1350	8.9	10.6	3.1	9.4
12	A: 1375	3.7	3.7	1.9	3.3
13	A: 1400	6.6	1	0.9	1.2
14	A: 1425	3.9	0.5	0.7	1.1
15	A: 1450	7.8	0.5	1.1	2.3
16	A: 1475	7.3	0	0.9	0.8
17	L: cortesana	0.6	0.5	0.7	0.8
18	L: albalaes	19.2	19.4	3	12.9
19	L: privilegios	7.9	1.6	3.1	7.9
20	L: gótica	2.8	0.6	0.6	2.2
21	L: gótica cursiva	10.7	15.8	2.8	19.6
22	L: gótica cursiva [albalaes]	9.2	19.2	1.8	20.3
23	L: gótica cursiva [precortesana]	0.2	2	0.3	1.7
24	L: gótica libraria	10.3	9	3.5	11
25	L: gótica redonda	12.8	1.1	0.7	1.6
26	L: humanística	3.3	0	3.3	0
27	L: humanística redonda	0.5	3.7	0	1
28	L: precortesana	3.9	3.9	0.7	2.7

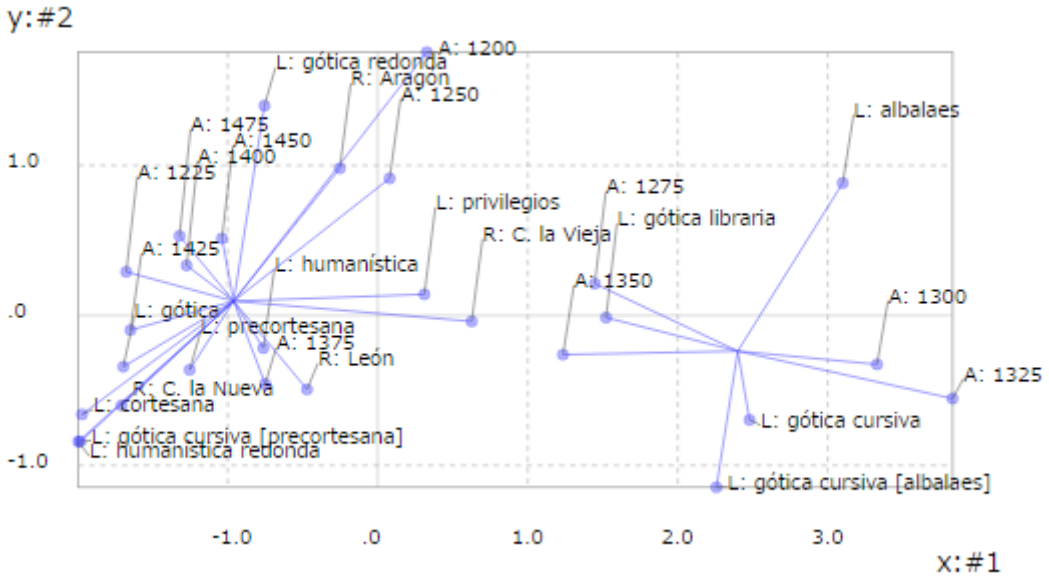
このデータを主成分分析し、個体(R, A, L)の第 1 主成分と第 2 成分の得点をそれぞれ X 軸と Y 軸にしてプロットすると次の図になります。



各座標間の距離を使ってクラスター分析し（平均法）、2 グループに分けると次のように結果になりました。



上の図を見ると、中央にある A:1275, L:gótica libraria, A:1350 のポイントは左のグループの中心よりも、むしろ右のグループの中心に近いようです。これは平均法で次々にグルーピングした結果ですが、上の出力からすべてのプロットとそれぞれの中心との距離を測り、一番近い中心をもつグループに集合させると、次のように、より適切なグルーピングができます。



実際には、こうして出来上がった新しいグループはメンバーが変わることによって中心が動くので、再度すべてのポイントについて一番近い中心を探さなければなりません。プログラムはこの作業を繰り返し、すべてのグループの構成に変化がなくなったときに終了します。この方法は James MacQueen の k-means 法とよばれます。k-means 法では求めるグループの数だけ仮の中心点を置くことから始めますが、ランダムな配置では効果的な

グルーピングができないことがあるので、ここでは最初に平均法でもとめたクラスターの中心点から始めることにしました[7]。

先に見たようにクラスター分析の出力はデンドログラムを使いますが、このように先にグループ数を決めておいて、平面上のポイントをまとめていく方法を「グループ分析」(Group analysis)と呼ぶことにします。頻度の和によって個体と属性のグループを作り、そうして出来上がった個体グループと属性グループの関係を分析することができます。グループ分析は個体数や属性数が多くて全体像がわかりにくいときに有効です。

7.4. 重回帰分析

重回帰分析(Multiple regression)とよばれる方法によって、次のような複数の説明変数(x1, x2, ...)と 1 個の目的変数(y: Y_n)をもつデータから、未知の目的変数を使って目的変数を予想する重回帰式を求めます。各説明変数に重み（ウェイト）W_p を掛けて重回帰式を作りますが、実際の結果 Y_n と重回帰式で求めた予測値ベクトル E_n の差が小さければ小さいほどその式が高く評価されます。そこで、実測値ベクトル Y_n と予測値ベクトル E_n の平方和が最小になるようにします。

たとえば、次のような成績表で、小テスト 3 回(x1, x2, x3)と、最終成績(POINT)の関係を見ます。

X_{np}	x1	x2	x3	POINT (y)
d1	6	8	5	12
d2	7	10	6	11
d3	8	4	8	13
d4	9	7	2	7
d5	10	9	4	14

ここで、POINT に該当する予測値 E_n を、切片 W(0)と各変数(X_{np})に重みとしての係数(W_p)を掛けたものを加算して作った式から求めます。[i = 1, 2, ..., n]

[1] $E(i) = W(0) + W(1)\,X(i, 1) + W(2)\,X(i, 2) + \ldots + W(p)\,X(i, p)$

[7] MacQueen, J. (1967) "Some methods for classification and analysis of multivariate observations", *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1* (Univ. of Calif. Press), 281-297.
http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992
[2017/01/07]: "(...) the k-means procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are, in fact, the means of the groups they represent (hence the term k-means)." (p.283)

この式の第 1 項 W(0)は回帰式の切片(intercept)を示します。この切片をすべての個体(1, 2, …, n)に共通に加えます。したがって、この列には単位ベクトル I_p を左積します[8]。

$$E(i) = I_p\, W(0) + X(i, 1)\, W(1) + X(i, 2)\, W(2) + \ldots + X(i, {}_p)\, W({}_p) \qquad [i = 1 \ldots n]$$

行列で示すと

$$E_n = X_{np}\, W_p \quad [X_{np} \text{の第 1 列は単位ベクトル}]$$

この式で求められた値と実測値 Y_n の間の残差のベクトルを R_n とします。

[2] $$R_n = Y_n - E_n = Y_n - X_{np}\, W_p$$

この残差 R_n の平方和 S を求めます。

$$\begin{aligned}
S &= R_n^T R_n = (Y_n - X_{np}\, W_p)^T (Y_n - X_{np}\, W_p) && \leftarrow [2] \\
&= [Y_n^T - (X_{np}\, W_p)^T] (Y_n - X_{np}\, W_p) && \leftarrow \text{転置行列の性質(T)} \\
&= Y_n^T\, Y_n - Y_n^T\, X_{np}\, W_p - (X_{np}\, W_p)^T\, Y_n + (X_{np}\, W_p)^T\, X_{np}\, W_p && \leftarrow \text{展開} \\
&= Y_n^T\, Y_n - Y_n^T\, X_{np}\, W_p - Y_n^T\, (X_{np}\, W_p) + W_p^T\, X_{np}^T\, X_{np}\, W_p && \leftarrow \text{T} \\
&= Y_n^T\, Y_n - 2\, Y_n^T\, X_{np}\, W_p + W_p^T\, X_{np}^T\, X_{np}\, W_p && \leftarrow \text{2 項と 3 項をまとめ}
\end{aligned}$$

この式の W_p は未知数です。重回帰分析の目的は、この S を最小化するために、S を変数のベクトル W_p で微分し（→後述）、その値がゼロベクトル(O_p^T)になるときの W_p を求めることです（多変数空間中の変数が形成する「曲面」の最小値の位置座標をイメージしてください）。

ここで、$S = Y_n^T\, Y_n - 2\, Y_n^T\, X_{np}\, W_p + W_p^T\, X_{np}^T\, X_{np}\, W_p$ の第 1 項 $Y_n^T\, Y_n$ には W_p がないので、W_p で微分するとゼロになります。第 2 項の $-2\, Y_n^T\, X_{np}\, W_p$ と第 3 項の $W_p^T\, X_{np}^T\, X_{np}\, W_p$ の微分については後述します。第 3 項の中の $X_{np}^T\, X_{np}$ は対称行列です。

$$\frac{\partial S}{\partial W_p} = \mathrm{Diff}(S, W_p) = -2\, X_{np}^T\, Y_n + 2\, X_{np}^T\, X_{np}\, W_p = O_p^T$$

$$\begin{aligned}
&2\, X_{np}^T\, X_{np}\, W_p = O_p^T + 2\, X_{np}^T\, Y_n && \leftarrow 2\, X_{np}^T\, Y_n \text{を移項} \\
&X_{np}^T\, X_{np}\, W_p = X_{np}^T\, Y_n && \leftarrow O_p^T \text{はゼロベクトル} \\
&(X_{np}^T\, X_{np})^{-1}\, (X_{np}^T\, X_{np})\, W_p = (X_{np}^T\, X_{np})^{-1}\, X_{np}^T\, Y_n && \leftarrow \text{（注[9]）} \\
&I_{pp}\, W_p = (X_{np}^T\, X_{np})^{-1}\, X_{np}^T\, Y_n && \leftarrow A\, A^{-1} = I_{pp} \\
&W_p = (X_{np}^T\, X_{np})^{-1}\, X_{np}^T\, Y_n && \leftarrow I_{pp}\, A = A
\end{aligned}$$

[8] 単位ベクトルを第 1 項に左積することで、第 1 項がそれに続く他の項と同じ構造になるので、全体の行列計算が可能になります。

[9] W_p を求めるためには W_p の係数を単位行列 I_{pp} にする必要があるので、両辺に$(X_{np}^T\, X_{np})^{-1}$ を左積します。

このようにして求めたベクトル W_p が下に示す「係数」(Value)の列です。

M.Coef.	Value
x1	.740
x2	.462
x3	1.157
Intercept	-3.819
Res.Ratio	.1237

「予測値」(E_n) は前述の式[1]で求めます。残差ベクトル(Resicual: Res)は、次の式で求めます。

$$Res = Y_n - E_n$$

なお、上表の Res.Ratio「残差比」は残差絶対値和と目的変数和の比です。残差がすべて 0 であれば、残差比は 0 となり、残差が目的変数に近づくにつれて 1 に近くなります。よって、この値が小さいほど回帰式のあてはまりがよいことになります。

$$Res.Ratio = Sm(AbsM(R_n)) / Sm(X_n)\ '$$

このようにして求めた回帰式を先の[1]

$$E_n = X_{np}\ W_p$$

によって次の導出変数(Derived)を計算します。データの目的変数と回帰式による導出変数を比較してください。

M.Regres.	POINT	Derived	Res.
d1	12	10.104	1.896
d2	11	12.926	-1.926
d3	13	13.207	-.207
d4	7	8.392	-1.392
d5	14	12.371	1.629

次に、先にもとめた係数ベクトル W_p を、目的変数が未知のデータ D_{np} に右積して、次の予測変数 Z_n を求めます。

$$Z_n = D_{np}\ W_p$$

次のデータ X.e の e1 は先の d1 と同じなので、同じ係数を掛けた予測変数(Expected)は当然同じになります。e2, e3 は変数の値が異なるので、それに応じて、予測変数が変化しています。

X.e	x1	x2	x3
e1	6	8	5
e2	6	8	2
e3	5	5	4
e4	7	5	4
e5	8	5	9

M.Predic.	Expected
e1	10.104
e2	6.633
e3	6.821
e4	8.301
e5	14.826

このような重回帰式のモデルは単回帰式(x1 のみ)でも同様に使うことができます。さまざまな要因(x1, x2, ...)から導出される予測変数は各種の予測の実務に役立ちますが、それだけでなく、それぞれの説明変数に掛ける係数の大小を見て、目的変数への影響度を評価します。先の例では v3 の係数(1.157)が他よりも大きいので、その重要度がわかります。

●多重共線性

次は、それぞれの説明変数と目的変数を合わせた相関行列です。これを見ると、x3 と POINT の相関が他と比べて高いことがわかります[10]。

C.Cor	x1	x2	x3	POINT
x1	1.0000	-.0687	-.4243	.0000
x2	-.0687	1.0000	-.3885	-.0080
x3	-.4243	-.3885	1.0000	.6207
POINT	.0000	-.0080	.6207	1.0000

重回帰分析をするとき、このような変数間の相関係数を見る必要があるのは、説明変数と目的変数の相関が変数のポジティブな評価に役立つだけでなく､説明変数どうしの相関がネガティブに問題を引き起こすためです。係数間に強い相関があるときは、そのことが影響して係数が容易でなくなります。このことは重回帰式が説明変数に重みを掛けた積の和になっていることから理解できます。たとえば説明変数 X(i, 1)と X(i, 2)の間に.98 などの強い相関があるとすると、回帰式の総和（積和）としての目的変数は一定なので、この 2 つの変数の値は競合し分け合うことになります。一方が強く働けば、他方を弱くしなければなりません。符号がプラスからマイナスに変わってしまうこともあります。もし、経験や直感から判断して、係数の符号（プラス･マイナス）が逆転していることなどが起きていれば、係数間の相関がある可能性が高いのです。これは**多重共線性**(muticollinearity)とよばれる問題です。極端な場合は変数間の相関係数が 1 のときです。これでは、2 つの変数に固有の情報がなく 1 つの情報だけで十分になります。重回帰式の中で使われる逆行列の計算（→後述）が不可

[10] しかし、データ数が少ないので、どの相関係数もあまり容易度は高くありません。

能になります。相関係数が高い場合も情報が少ないので、同様の問題が起きます。そのときは、重要な変数だけを残し、回帰式を単純化して、残りの重要な変数に注目すべきです。

回帰式に多くの係数を入れると､それだけあてはまりがよくなりますが、それは与えられたデータについてのあてはまりにすぎません。予測の一般性を高めるためには、実験を繰り返して適切な変数を選択し、なるべく少ない変数で予測式を求めるべきです。そうすれば、目的変数を説明する変数が何なのかを的確に、そして「きれいに」示すことができるからです。また、複数の相関が高い変数群の中から 1 つを選ぶことによって、変数のグルーピング（→「分類」）ができるので、変数間の関係の理解につながります。せっかく集めた変数の分布データを捨てるのが惜しい、ということでしたら、相関する変数の「どちらにも当てはまるケース」の頻度を計算し､これを新たな変数として使う､ということも考えられます｡または、相関する変数の「どちらかに当てはまるケース」を数えて、比べてみるとよいでしょう。このようにすれば、すべての変数の情報が生かされます。

●標準化回帰分析

説明変数と目的変数をそれぞれ標準化すると（→「標準得点」）、平均が 0 となり、その回帰直線は座標の原点(0, 0)を通るので、回帰式の切片がなくなります（→「相関係数」）。また、変数をその標準偏差で割ることによって､次表のように､説明変数のバラツキをなくした標準化された｢重み」が計算されます。変数の重みを比較するためには、この標準化回帰分析のほうが適しています。

S.Regres.	POINT	Derived	Res.
d1	12.000	10.104	1.896
d2	11.000	12.926	-1.926
d3	13.000	13.207	-.207
d4	7.000	8.392	-1.392
d5	14.000	12.371	1.629

S.Coef.	Value
x1	.433
x2	.394
x3	.957
Intercept	.000
Res.Ratio	.124

●逆行列

(1) 逆行列の定義

正方行列(X_{pp})について

$$X_{pp}\ Y_{pp} = I_{pp}\ (\text{単位行列}) \ \rightarrow\ Y_{pp} = X_{pp}^{-1}$$

となる正方行列(Y_{pp})は、X_{pp}の「逆行列」(Inverse matrix: X_{pp}^{-1})とよばれます。逆行列が関係する次の演算は統計の計算によく使われます。

(a) $X_{pp}\ X_{pp}^{-1} = I_{pp}$

X_{pp}	1	2
1	7	8
2	9	10

×

X_{pp}^{-1}	1	2
1	-5	4
2	4.5	-3.5

=

I	1	2
1	1	0
2	0	1

(b) $X_{pp}^{-1}\ X_{pp} = I_{pp}$

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

×

X_{pp}	1	2
1	7	8
2	9	10

=

I_{pp}	1	2
1	1	0
2	0	1

(2) 逆行列の性質

(a) $(X_{pp}^{-1})^{-1} = X_{pp}$

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

→

$(X_{pp}^{-1})^{-1}$	1	2
1	7	8
2	9	10

(b) $(X_{pp}\ Y_{pp})^{-1} = Y_{pp}^{-1}\ X_{pp}^{-1}$

X_{pp}	1	2
1	1	3
2	2	4

×

Y_{pp}	1	2
1	7	8
2	9	1

=

$X_{pp}Y_{pp}$	1	2
1	34	11
2	50	20

→

$(X_{pp}Y_{pp})^{-1}$	1	2
1	0.154	-0.085
2	-0.385	0.262

Y_{pp}^{-1}	1	2
1	-2.00	1.500
2	1.00	-0.500

×

X_{pp}^{-1}	1	2
1	-0.015	0.123
2	0.136	-0.108

=

$Y_{pp}^{-1}\ X_{pp}^{-1}$	1	2
1	0.154	-0.085
2	-0.385	0.262

(c) $(X_{pp}^{T})^{-1} = (X_{pp}^{-1})^{T}$

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{T}	1	2
1	7	9
2	8	10

→

$(X_{pp}^{T})^{-1}$	1	2
1	-5.0	4.5
2	4.0	-3.5

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

→

$(X_{pp}^{-1})^{T}$	1	2
1	-5.0	4.5
2	4.0	-3.5

(3) 逆行列演算の証明

次の演算はしばしば使われます。しっかりと理解しておくために証明を

しておきましょう。

[1] $I^{-1} = I$

$I\ I^{-1} = I$ [逆行列の定義: $X\ X^{-1} = I$, ここで $X = I$]
$I^{-1} = I$ [$I\ X = X$, $X=I$]

[2] $(A^{-1})^{-1} = A$

$A^{-1}\ (A^{-1})^{-1} = I$ [逆行列の定義: $A\ A^{-1} = I$]
$A\ A^{-1}\ (A^{-1})^{-1} = A\ I$ [両辺に A を左積]
$I\ (A^{-1})^{-1} = A\ I$ [逆行列の定義: $A\ A^{-1} = I$]
$(A^{-1})^{-1} = A$ [$X\ I = X$; $I\ X = X$]

[3] $(A\ B)^{-1} = B^{-1}\ A^{-1}$

$(A\ B)\ (A\ B)^{-1} = I$ [$X\ X^{-1} = I$, $X = A\ B$]
$(A\ B)\ (A\ B)^{-1} = A\ A^{-1}$ [$A\ A^{-1} = I$]
$(A\ B)\ (A\ B)^{-1} = A\ I\ A^{-1}$ [$A = A\ I$]
$(A\ B)\ (A\ B)^{-1} = A\ B\ B^{-1}\ A^{-1}$ [$I = B\ B^{-1}$]
$(A\ B)^{-1} = B^{-1}\ A^{-1}$ [両辺から A B を削除]

[4] $A\ A^{-1} = A^{-1}\ A$

$A\ A^{-1} = I$ [逆行列の定義: $A\ A^{-1} = I$]
$(A^{-1}\ A)\ (A\ A^{-1}) = (A^{-1}\ A)\ I$ [両辺に $A^{-1}\ A$ を左積]
$A^{-1}\ A\ A\ A^{-1} = A^{-1}\ A$ [$X\ I = X$, $X=A^{-1}\ A$]
$I\ A\ A^{-1} = A^{-1}\ A$ [$X\ I = X$, $X=A^{-1}\ A$]
$A\ A^{-1} = A^{-1}\ A$ [$I\ A = A$]

＊[2], [3]は足立(2005:110-111)を参照しました。

●逆行列の求め方

与えられた行列(X_{pp})と、初期値が単位行列である行列($Z_{pp}=I_{pp}$)を同時に変形していきます。X_{pp} が単位行列(I_{pp})になるように、X_{pp} と Z_{pp} に左から変形行列 T_{pp} を繰り返して掛けていきます。そのために

(a) 2 つの行を交換する T_{pp}
(b) 実数倍した 1 つの行全体に、実数倍した他の行を加算する T_{pp}

という 2 つの変換を使います。これらの変換を可能にする変形行列 T_{pp} を次々に左積すると、Z_{pp} が A_{pp} の逆行列になることを次の演算で確認しましょう(「Gauss の消去法」Gauss reduction)。

0. $X^{(0)}, Z^{(0)} = I$ ←X, Z の初期状態$^{(0)}$
1. $X^{(1)} = T^{(1)} X^{(0)}, Z^{(1)} = T^{(1)}I$ ←$X^{(0)}$と $Z^{(0)}$=I に $T^{(1)}$を左積
2. $X^{(2)} = T^{(2)} T^{(1)} X^{(0)}, Z^{(2)} = T^{(2)} T^{(1)}I$←さらに $T^{(2)}$を左積

(…) ←さらに $T^{(3)}, \ldots, T^{(k)}$を順次左積

3. $I = T^{(k)} \ldots T^{(2)} T^{(1)} X^{(0)}$ ←$X^{(0)}$に T を順次左積し I に至る
4. $Z^{(k)} = T^{(k)} \ldots T^{(2)} T^{(1)}I$ ←$Z^{(0)}$ =I に T を順次左積し $Z^{(k)}$を得る
5. $I\, X^{(0)-1} = T^{(k)} \ldots T^{(2)} T^{(1)} X^{(0)}\, X^{(0)-1}$ ←3 の両辺に $X^{(0)-1}$を右積
6. $X^{(0)-1} = T^{(k)} \ldots T^{(2)} T^{(1)}I$ ←5. I A = A; A A^{-1} =I
7. $Z^{(k)} = X^{(0)-1}$ ←4.右辺= 6.右辺、よって $Z^{(k)}$は $X^{(0)}$の逆行列になる

たとえば次の行列 $X^{(0)}$の逆行列を求めることを考えましょう。以下の演算のために、作業用の行列 $T^{(1)}$ と出力用の単位行列 $Z^{(1)}$= I を用意します。目的は $T^{(1)}, T^{(2)}, \ldots, T^{(k)}$の左積を繰り返して、$X^{(k)}$を単位行列にすることです。

X(0)	1	2	3
1	**0**	2	1
2	2	1	2
3	2	1	1

Z(0)	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

はじめに、X(1, 1)を 1 にするために次の演算をします。

R1 ← R1 / X(1, 1)

これは X の第一行 R1 を X(1, 1)で割って新たな R1 にする、ということです。ここでは、X(1, 1)が 0 なので割り算ができません。そのときは、第一列 C1 が 0 でない行と交換します。その結果 $X^{(1)}$となります。

R1 ← R2, R2 ← R1

$X^{(1)}$	1	2	3
1	**2**	**1**	**2**
2	**0**	**2**	**1**
3	2	1	1

$Z^{(1)}$	1	2	3
2	**0**	**1**	**0**
1	**1**	**0**	**0**
3	0	0	1

こうして新たな X(1, 1)←2 で先の除算をします。

R1 ← R1 / X(1, 1) ← R1 / 2

$X^{(2)}$	1	2	3
1	**2/2=1**	**1/2**	**2/2=1**
2	0	2	1
3	2	1	1

$Z^{(2)}$	1	2	3
1	**0/2=0**	**1/2**	**0/2=0**
2	1	0	0
3	0	0	1

次に R2 と R3 を R1 を使って、それぞれの C1 の値を 0 にします。ここでは R2 の X(2, 1)が 0 なので、R3 だけを次のようにして変えます。

R3 ← R3 – X(3, 1) R1 ← R3 – 2 R1

$X^{(3)}$	1	2	3
1	1	1/2	1
2	0	2	1
3	**2-2*1=0**	**1-2*(1/2)=0**	**1-2*1=-1**

$Z^{(3)}$	1	2	3
1	0	1/2	0
2	1	0	0
3	**0-2*0=0**	**0-2*1/2=-1**	**1-2*0=1**

これで C1 は完成です。次に同様なことを C2 で行います。

$X^{(4)}$	1	2	3
1	1	1/2	1
2	0	**2**	1
3	0	0	-1

$Z^{(4)}$	1	2	3
1	0	1/2	0
2	1	0	0
3	0	-1	1

今度は X(2,2)=2 は 0 でないので、そのまま R2 を 2 で割ります。

R2 ← R2 / X(2, 2) ← R2 / 2

$X^{(5)}$	1	2	3
1	1	1/2	1
2	**0/2=0**	**2/2=1**	**1/2**
3	0	0	-1

$Z^{(5)}$	1	2	3
1	0	1/2	0
2	**1/2**	**0/2**	**0/2**
3	0	-1	1

そして R1 と R2 の C2 を次の演算で、0 にします。

R1 ← R1 – X(1, 2) R2 ← R1 – 1/2 R2
R3 ← R3 – X(3, 2) R2 ← R3 – 0 R2

$X^{(6)}$	1	2	3
1	**1-(1/2)*0 =1**	**1/2-(1/2)*1 =0**	**1-(1/2)*(1/2)=3/4**
2	0	1	1/2
3	**0-0*0=0**	**0-0*1=0**	**-1-0*(1/2)=-1**

$Z^{(6)}$	1	2	3
1	**0-(1/2)*(1/2) =1/4**	**1/2-(1/2)*0 =1/2**	**0-(1/2)*0 =0**
2	1/2	0	0
3	**0-0*(1/2)=0**	**-1-0*0=-1**	**1-0*0=1**

これで C2 は完成です。次に同様なことを C3 で行います。

$X^{(7)}$	1	2	3
1	1	0	3/4
2	0	1	1/2
3	0	0	**-1**

$Z^{(7)}$	1	2	3
1	1/4	1/2	0
2	1/2	0	0
3	0	-1	1

R3 ← R3 / X(3, 3) ← R3 / -1

$X^{(8)}$	1	2	3
1	1	0	3/4
2	0	1	1/2
3	**0/-1=0**	**0/-1=0**	**-1/-1=1**

$Z^{(8)}$	1	2	3
1	1/4	1/2	0
2	1/2	0	0
3	**0/-1=0**	**-1/-1=1**	**1/-1=-1**

R1 ← R1 – X(1, 3) R3 ← R1 – 3/4 R3
R2 ← R1 – X(2, 3) R3 ← R1 – 1/2 R3

$X^{(9)}$	1	2	3
1	1-(3/4)x0 =1	0-(3/4)x0 =0	3/4-(3/4)x1 =0
2	0-(1/2)x0 =0	1-(1/2)x0 =1	1/2-(1/2)x1 =0
3	0	0	1

$Z^{(9)}$	1	2	3
1	1/4-(3/4)x0 =-1/4	1/2-(3/4)x1 =-1/4	0-(3/4)－1 =3/4
2	1/2-(1/2)x9 =1/2	0-(1/2)x1 =-1/2	0-(1/2)x-1 =1/2
3	0	1	-1

これらの演算の結果、次のようにXは単位行列になり、ZにXの逆行列が得られました。

$X^{(k)}$	1	2	3
1	1	0	0
2	0	1	0
3	0	0	**1**

$Z^{(k)}$	1	2	3
1	-1/4	-1/4	3/4
2	1/2	-1/2	1/2
3	0	1	-1

プログラムで実行すると確かにXの逆行列X^{-1}が得られ、XとX^{-1}の行列積を計算すると単位行列が得られます。

X	1	2	3
1	0	2	1
2	2	1	2
3	2	1	1

X^{-1}	1	2	3
1	-.250	-.250	.750
2	.500	-.500	.500
3	.000	1.000	-1.000

$X X^{-1}$	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

＊長谷川(2000:129-136)を参照しました。プログラムは縄田(1999:58-80)を参照しました。

プログラム[11]

```
Function Iv(ByVal Xpp) '逆行列(Gauss-Jordan 法. ver. 2013/06/28-2015/1/22)
  Dim TT$, P&, i&, j&, Tpp, Zpp, E: P = NC(Xpp): E = -15 'P=行数=列数
  TT$ = Xpp(0, 0): Zpp = Um(P) 'X 対象の行列 : Zpp 単位行列
  For i = 1 To P '1 列から P 列まで
    If Abs(Xpp(i, i)) < 10 ^ E Then '対角成分が 0 ならば行交換
      For j = i + 1 To P 'i+1 行から P 行まで
        If i < P And Abs(Xpp(j, i)) > 10 ^ E Then '非対角成分が 0
          Tpp = Um(P): Tpp(i, i) = 0: Tpp(j, j) = 0: Tpp(i, j) = 1: Tpp(j, i) = 1
                                                    '変形行列
          Xpp = X(Tpp, Xpp): Zpp = X(Tpp, Zpp) 'i 行と j 行を交換
          Exit For 'For j を脱出
        End If
      Next j
    End If
    If Xpp(i, i) = 0 Then '対角成分=0
      MsgBox Ln(29): Exit Function 'Msg「逆行列は存在しません。」
    End If
    For j = 1 To P '1 行から P 行まで、非対角成分=0, 対角成分=1
      If i <> j And Abs(Xpp(j, i)) > 10 ^ E Then
        Tpp = Um(P): Tpp(i, i) = 1 / Xpp(i, i) '変形行列 (Horikawa 2013)
        Xpp = X(Tpp, Xpp): Zpp = X(Tpp, Zpp) 'X(i, i) = 1
        Tpp = Um(P): Tpp(j, i) = -1 * Xpp(j, i) '変形行列
        Xpp = X(Tpp, Xpp): Zpp = X(Tpp, Zpp) 'Rj=Rj-X(j,i)*Ri→X(j,i) = 0
      End If
    Next j
  Next i
  Zpp(0, 0) = TT$ & "^":     Iv = Zpp '返し値
End Functio
```

●変形行列

単位行列の一部を変更した行変形用行列を作成し、これをある行列に左積すると、一定の行変形ができます。ここではそのような行列を「変形行列」(transformation matrix)とよぶことにします。これらを逆行列の計算に使います。

[11] Tpp(i, i) = 1 / Xpp(i, i)を、最終プロセスではなく各行のプロセスに置くことによって数値のオーバーフローを回避する方法は堀川遼太さんからいただいたアイデアです(2013)。ご本人の許可をいただき、お名前を載せて謝意を表します。

(a) R1 ← 0

T_{pp}	1	2	3
1	0	0	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	0	0	0
2	4	5	6
3	7	8	9

(b) R1 ← R2

T_{pp}	1	2	3
1	0	1	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	4	5	6
2	4	5	6
3	7	8	9

(c) R1 ~ R2 (交換)

T_{pp}	1	2	3
1	0	1	0
2	1	0	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	4	5	6
2	1	2	3
3	7	8	9

(d) R1 ← 3 R1 (倍数)

T_{pp}	1	2	3
1	3	0	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	3	6	9
2	4	5	6
3	7	8	9

(e) R2 ← R2 + R1

T_{pp}	1	2	3
1	1	0	0
2	1	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	1	2	3
2	5	7	9
3	7	8	9

(f) R2 ← R2 + 2 R1

T_{pp}	1	2	3
1	1	0	0
2	2	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	1	2	3
2	6	9	12
3	7	8	9

(g) R2 ← 3 R2 + 2 R1

T_{pp}	1	2	3
1	1	0	0
2	2	3	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

=

R_{pp}	1	2	3
1	1	2	3
2	14	19	24
3	7	8	9

最後の演算を見ると、変形行列の対角成分で自分の行を積算し、非対角成分でその列番にあたる行を積算していることがわかります。行のゼロ化[1]や行の移動[2][3]も同様です。

＊芝(1975: 197-199)を参照しました。

●行列の微分

多変量分析ではしばしば行列をベクトルで微分します。行列の積の成分を展開すればベクトルで微分した結果が行列とベクトルの積になることがわかります。

[1] はじめに、次のような行列 T_{pp} の W_p による微分について見ましょう。

$$T_{pp} = Y_n^T X_{np} W_p = [y_1, y_2, \ldots, y_n] \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

この行列 T_{pp} をベクトル $W_p = \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$ で微分する、ということの意味を理解するために T_{pp} を展開します。

$$\begin{aligned} T_{pp} = [&y_1 x_{11} + y_2 x_{21} + \ldots + y_n x_{n1}, \\ &y_1 x_{12} + y_2 x_{22} + \ldots + y_n x_{n2}, \\ &\ldots, \\ &y_1 x_{1n} + y_2 x_{2n} + \ldots + y_n x_{np}] \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix} \end{aligned}$$

$$\begin{aligned} = \quad &(y_1 x_{11} + y2\, x_{21} + \ldots + y_n x_{n1}) * w_1 \\ &+ (y_1 x_{12} + y_2 x_{22} + \ldots + y_n x_{n2}) * w_2 \\ &+ \ldots \\ &+ (y_1 x_{1n} + y_2 x_{2n} + \ldots + y_n x_{np}) * w_p \end{aligned}$$

偏微分の記号 $\frac{\partial S}{\partial a}$ を Diff(S, w)で示すと（「S を w で微分する」という意味です）

$$\begin{aligned} &Diff(T_{pp}, w_1) = y_1 x_{11} + y_2 x_{21} + \ldots + y_n x_{n1} && \leftarrow \text{上式の 1 行目} \\ &Diff(T_{pp}, w_2) = y_1 x_{12} + y_2 x_{22} + \ldots + y_n x_{n2} && \leftarrow \text{上式の 2 行目} \\ &\ldots \\ &Diff(T_{pp}, w_p) = y_1 x_{1p} + y_2 x_{2p} + \ldots + y_n x_{np} && \leftarrow \text{上式の p 行)} \end{aligned}$$

これらをまとめて示すと次のようになります。

$\mathrm{Diff}(T_{pp}, W_p) = \mathrm{Diff}(Y_n{}^T X_{np} W_p, W_p) = X_{np}{}^T Y_n$ [←縦ベクトル]

高等学校で既習の次の微分と比べてみてください。

$\mathrm{Diff}(yxw, w) = yx$

[2] 次は微分する項(Wp)が 2 乗されている場合です。たとえば

$$T_{pp} = W_p{}^T X_{pp} W_p = [w_1, w_2, \ldots, w_p] \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{12} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{1p} & x_{2p} & \ldots & x_{pp} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

を、ベクトル $W_p = \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$ で微分します: $\mathrm{Diff}(T_{pp}, W_p)$。ここでは X_{pp} を対称行列とします。

$$T_{pp} = W_p{}^T X_{pp} W_p$$

$$= [w_1 x_{11} + w_1 x_{12} + \ldots + w_1 x_{1p},$$
$$w_1 x_{21} + w_2 x_{22} + \ldots + w_2 x_{2p},$$
$$\ldots,$$
$$w_1 x_{n1} + w_2 x_{n2} w_2 + \ldots + w_p x_{np}\ ] \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

$$= w_1 x_{11} w_1 + w_1 x_{12} w_2 + \ldots + w_1 x_{1p} w_p$$
$$+ w_2 x_{12} w_1 + w_2 x_{22} w_2 + \ldots + w_2 x_{2p} w_p$$
$$+ \ldots$$
$$+ w_p x_{1p} w_1 + w_p x_{2p} w_2 + \ldots + w_p x_{pp} w_p$$

$$= x_{11} w_1{}^2 + w_1 x_{12} w_2 + \ldots + w_1 x_{1p} w_p$$
$$+ w_2 x_{12} w_1 + x_{22} w_2{}^2 + \ldots + w_2 x_{2p} w_p$$
$$+ \ldots$$
$$+ w_p x_{1p} w_1 + w_p x_{2p} w_2 + \ldots + x_{pp} w_p{}^2$$

この式で w_1 を含む成分は第 1 行と第 1 列の成分です。よって

$$\mathrm{Diff}(T_{pp}, w_1) = 2w_1 x_{11} + 2(w_2 x_{12} + \ldots + w_p x_{1p}) = 2(w_1 x_{11} + w_2 x_{12} + \ldots + w_p x_{1p})$$

同様に、w_2 を含む成分は第 2 行と第 2 列の成分です。よって

$$\mathrm{Diff}(T_{pp}, w_2) = 2w_2 x_{12} + 2(w_2 x_{22} + \ldots + 2w_p x_{2p}) = 2(w_2 x_{12} + w_2 x_{22} + \ldots + w_p x_{2p})$$

同様にして w_p を含む成分は最終の第 p 行と第 p 列の成分です。よって

$$\mathrm{Diff}(T_{pp}, W_p) = 2w_p x_{1p} + 2(w_p x_{2p} + \ldots + w_p x_{pp}) = 2(w_p x_{1p} + {}_2 x_{2p} + \ldots + w_p x_{pp})$$

以上をまとめて示すと次のようになります。

$$\mathrm{Diff}(T_{pp}, Wp) = \mathrm{Diff.}\ (Wp^T X_{pp} Wp, Wp) = 2\ X_{pp}\ Wp$$

高等学校で既習の次の微分と比べてみてください。

$$\mathrm{Diff}(wxw, w) = \mathrm{Diff}(xw^2, w) = 2xw$$

●数量化 1 類

次のように、説明変数が数量ではなく質的なデータ(v)を扱うとき、これを 0-1 に変換して、同様に重回帰分析をすることができます。この方法は「数量化 1 類」(Quantification method, first type)とよばれます。

X	v1	v2	v3	POINT
d1		v		12
d2	v	v	v	11
d3	v		v	13
d4	v	v		7
d5	v	v		14

X	POINT	Expected	Residual
d1	12.000	12.000	.000
d2	11.000	11.000	.000
d3	13.000	13.000	.000
d4	7.000	10.500	-3.500
d5	14.000	10.500	3.500

Weight	P: Intercept	v1	v2	v3	Std res.
Value	14.000	-1.500	-2.000	.500	2.214

この方法を使用するにあたって注意しなければならないのは、次のようなケースです。

X	v1	v2	v3	POINT
d1	v	v		12
d2	v	v	v	11
d3	v		v	13
d4	v	v		7
d5	v	v		14

X	v1	v2	v3	POINT
d1		v		12
d2	v	v		11
d3	v		v	13
d4	v	v		7
d5	v	v		14

上左表では v1 がすべて選択されていますので、この v1 には弁別する情報がありません。また、右表では v2 と v3.が相補分布(complementary distribution)をしています。この場合は、どちらかを選択すれば他方が決まっているので、どちらか 1 つにしか弁別する情報がないことになります。

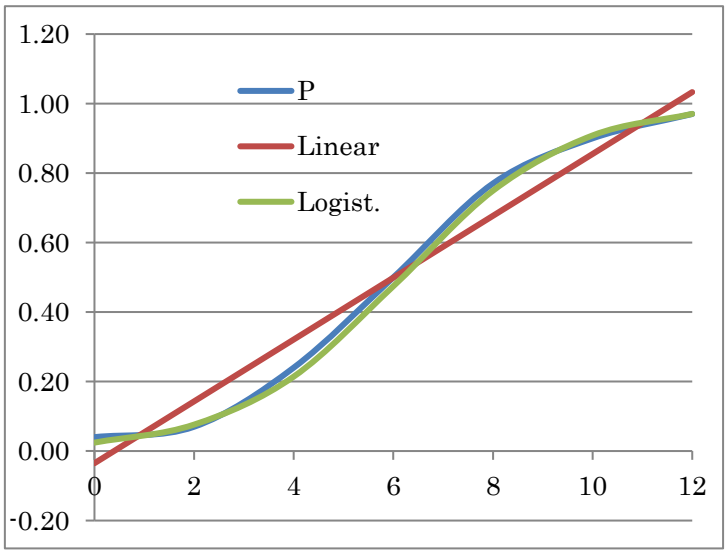
このような行列ではすべて逆行列が存在せず分析ができないので、データから該当する行を取捨選択しなければなりません。

●ロジット回帰分析

下表は変数 X と、それに対応する確率（または何らかの比率:P）を示します。たとえば、1 週間の学習時間(X)と英語のテストの正解率のようなものを考えます。確率や比率の範囲は[0, 1]です。

L	X	P	Linear	Logist.
1	0	0.04	-0.035	.024
2	2	0.07	0.143	.076
3	4	0.24	0.321	.214
4	6	0.50	0.499	.474
5	8	0.77	0.677	.749
6	10	0.90	0.855	.908
7	12	0.97	1.033	.970

上の Linear は単回帰分析による導出変数です。これをグラフにすると、次の直線のようになります。ここで、近似があまりよくないことと、X=0 で P がマイナスになり、X=12 で P が 1 を超えていることがわかりますが、これは率の範囲が[0, 1]であるので、現実的ではありません。一方、上表の Logist.はかなりよく P に近似しています。また、グラフを見ても、[0, 1]の範囲を超えることはなさそうです。



あらゆる変数に対して、上の Logist のように[0, 1]の範囲に納める方法としては、限定得点が考えられますが、直線による限定得点では上のような S 字型の P の分布では近似が良くありません。

次の表と図は確率 P、その関数であるロジット(Logit: L)、そしてロジッ

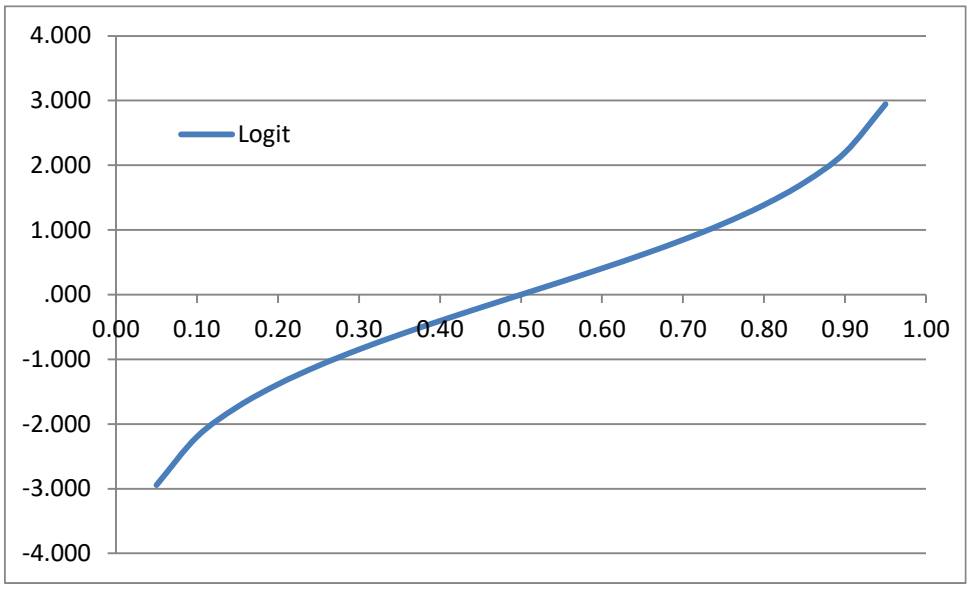
トから確率 P を導く逆関数 InvLogit を示します。ロジット(Logit)は

$$L = \mathrm{Ln}\ (P / 1 - P)$$

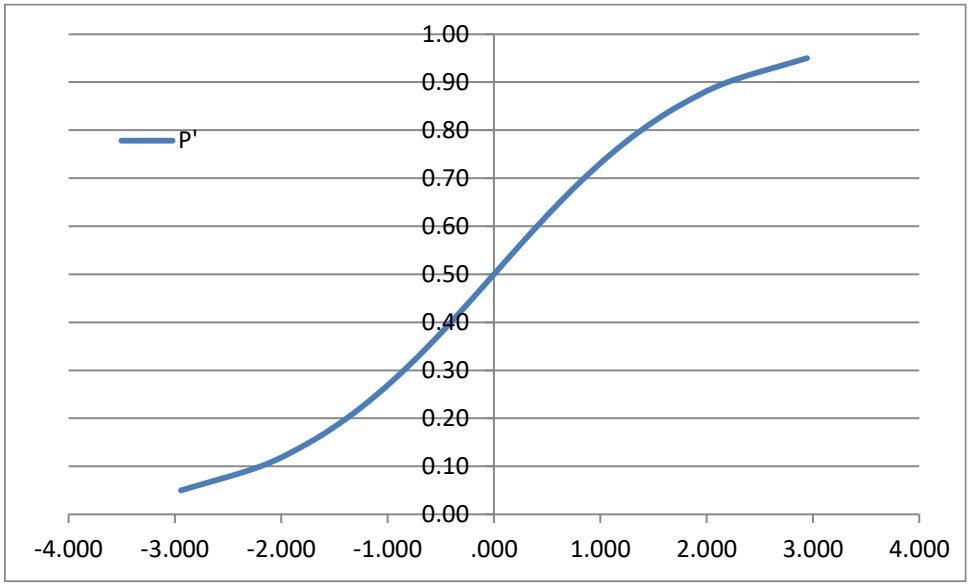
つまり、ロジットは、あることが起こる確率 P とそれが起こらない確率 1 − P の比率の自然対数を示します。

P	Logit	P'
0.05	-2.944	0.05
0.10	-2.197	0.10
0.15	-1.735	0.15
0.20	-1.386	0.20
0.25	-1.099	0.25
0.30	-.847	0.30
0.35	-.619	0.35
0.40	-.405	0.40
0.45	-.201	0.45
0.50	.000	0.50
0.55	.201	0.55
0.60	.405	0.60
0.65	.619	0.65
0.70	.847	0.70
0.75	1.099	0.75
0.80	1.386	0.80
0.85	1.735	0.85
0.90	2.197	0.90
0.95	2.944	0.95

下図は横軸が確率 P であり、それに応じてロジットがどのように変化するかを示しています。P の範囲は[0, 1]ですが、ロジットは範囲が自由で P=0 のときに－∞、P=1 のときに＋∞に漸近します。



次の図では横軸がロジット、縦軸が確率です。



上の確率 P は、ロジット(L)から次のようにして導出します(e: 自然対数の底)。

$$\mathrm{Ln}\,(P / (1 - P)) = L$$

$$P / (1 - P) = e^{L}$$

$$P = (1 - P)\, e^{L} = e^{L} - P\, e^{L}$$

$$P + P\, e^{L} = e^{L}$$

$$(1 + e^{L})\, P = e^{L}$$

$$P = e^{L} / (1 + e^{L}) = 1 / (1 / e^{L} + 1) = 1 / (e^{-L} + 1)$$

次の表は、確率 P をロジットに変換して重回帰分析をし、その導出変数を確率に戻して出力した結果です。この方法は**ロジスティック回帰分析** (Logistic regression: L.Reg.)とよばれます。

L	X	P
1	2	.017
2	4	.023
3	6	.037
4	8	.050
5	10	.070

L.Reg.	P	Derived	Res.
1	.017	.017	.000
2	.023	.024	-.001
3	.037	.035	.002
4	.050	.050	.000
5	.070	.071	-.001

L.Coef.	Value
X	.189
Intercept	.012
Res.Ratio	.020

説明変数が複数の場合は次のモデルで回帰分析をします。

$$L = Ln\ (P / (1 - P) = W(0) + W(1)\ X(i, 1) + W(2)\ X(i, 2) + \ldots + W(p)\ X(i, p)$$
$$= X_{np}\ W_p$$

同じデータをそのまま重回帰分析(M.Reg.)にかけると次の結果になりました。上のロジスティック回帰分析の結果の方が残差(Res, Res.Ratio)が少ないことがわかります。

L	X	P
1	2	.017
2	4	.023
3	6	.037
4	8	.050
5	10	.070

M.Reg.	P	Derived	Res.
1	.017	.013	.004
2	.023	.026	-.003
3	.037	.039	-.003
4	.050	.053	-.003
5	.070	.066	.004

M.Coef.	Value
X	.007
Intercept	-.001
Res.Ratio	.081

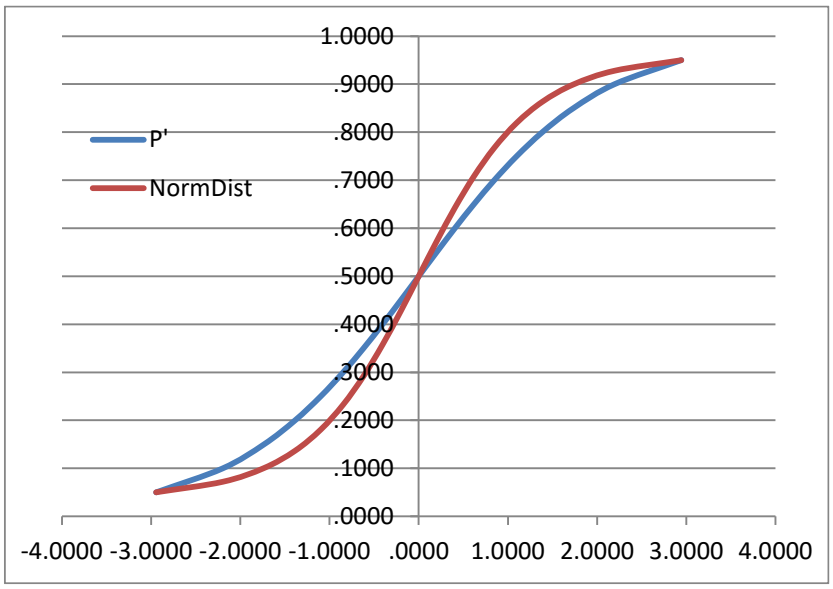
●正規回帰分析

先の図（ロジットを横軸に、確率を縦軸にしたグラフ）は累積正規確率分布とよく似ています（→「確率」）。次の表と図が示すように、その中心の座標は、ロジットに対応する確率(P)でも、正規分布(NormDist)でも同じですが(0.5)、傾きが少し違っています[12]。

Logit	P'	NormDist
-2.9444	.0500	0.0502
-2.1972	.1000	0.0721
-1.7346	.1500	0.1006
-1.3863	.2000	0.1367
-1.0986	.2500	0.1807
-.8473	.3000	0.2326
-.6190	.3500	0.2919
-.4055	.4000	0.3575

[12] NormDist 関数の引数である平均を 0.5 とし、標準偏差を Logit の範囲で求めました。

-.2007	.4500	0.4276
.0000	.5000	0.5000
.2007	.5500	0.5724
.4055	.6000	0.6425
.6190	.6500	0.7081
.8473	.7000	0.7674
1.0986	.7500	0.8193
1.3863	.8000	0.8633
1.7346	.8500	0.8994
2.1972	.9000	0.9279
2.9444	.9500	0.9498



ロジスティック回帰分析の回帰式の目的変数はロジットに対応する確率(P)を使いますが、その確率分布ではデータ（目的変数）の平均と分散（または標準偏差)が考慮されていません。どのようなデータの目的変数でも、すべて同じようにロジットに対応する確率分布をあてはめて一般化します。

ここで、重回帰式の目的変数（確率）がこの変数の平均と分散を考慮に入れた正規累積分布にしたがう、と見なし、正規累積分布の逆関数 NormInv で変換した数値を使って重さベクトルを算出する方法を**正規回帰分析**(Normal regression: N.Regres)と名づけて提案します。導出変数(Derived)の計算では、もとの目的変数の平均(m)と分散(v)を使った正規累積分布関数 NormDist(x, m, sqr(v), 1)を適用します。

次の表が先のロジスティック回帰分析と同じデータを使った、正規回帰分析(N.Regres)の結果です。残差(Res)と残差比(Res.Ratio)がさらに小さくなりました。

L	X	P

N.Regres.	P	Derived	Res.

N.Coef.	Value

1	2	.017
2	4	.023
3	6	.037
4	8	.050
5	10	.070

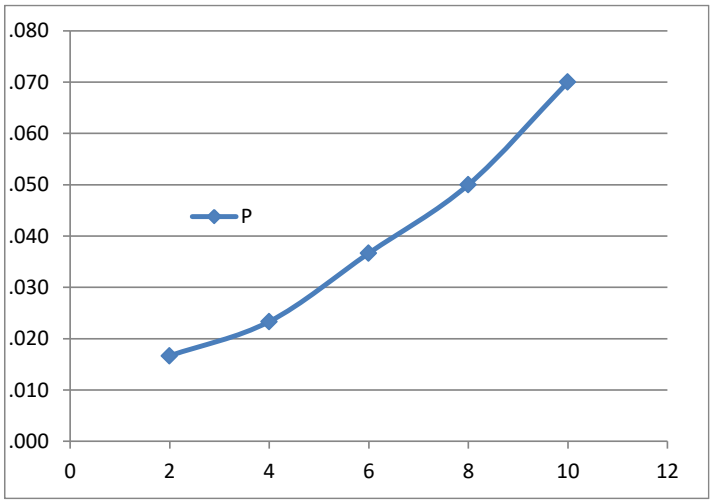
1	.017	.016	.000
2	.023	.024	-.001
3	.037	.035	.001
4	.050	.050	.000
5	.070	.070	.000

X	.002
Intercept	-.005
Res.Ratio	.015

正規回帰分析は目的変数が直線式でなく、むしろ正規分布（の一部分）のような分布をしているときに有効です。そこで、次のように複数の説明変数(X1, X2)があるときはその相関係数行列を計算して、P と相関が高い変数(X1)を使って P の散布図を描きます。

L	X1	X2	P
1	2	6	.017
2	4	5	.023
3	6	5	.037
4	8	3	.050
5	10	6	.070

C.Cor	X1	X2	P
X1	1.0000	-.2582	.9853
X2	-.2582	1.0000	-.1272
P	.9853	-.1272	1.0000



上の図を見ると分布は直線になっていないことがわかります。そこで、直線式による重回帰係数は適切でないことになります。次の 2 つの表によって重回帰分析と正規回帰分析の結果を比較すると（残差と残差比）、正規回帰分析のほうがこのデータに適していることが確認できます。

L	X1	X2	P
1	2	6	.017
2	4	5	.023
3	6	5	.037
4	8	3	.050
5	10	6	.070

M.Regres.	P	Derived	Res.
1	.0167	.0141	.0026
2	.0233	.0255	-.0022
3	.0367	.0393	-.0027
4	.0500	.0484	.0016
5	.0700	.0693	.0007

M.Coef.	Value
X1	.0069
X2	.0024
Intercept	-.0140
Res.Ratio	.0494

L	X1	X2	P
1	2	6	.017
2	4	5	.023
3	6	5	.037
4	8	3	.050
5	10	6	.070

N.Regres.	P	Derived	Res.
1	.0167	.0164	.0002
2	.0233	.0244	-.0010
3	.0367	.0355	.0012
4	.0500	.0499	.0001
5	.0700	.0705	-.0005

N.Coef.	Value
X1	.0016
X2	.0001
Intercept	-.0050
Res.Ratio	.0155

●指数回帰分析

下左表は乱数の標本を 2 セットを用意して、その標本数(N)を 10, 20, ..., 100 まで実験しながら求めた相関係数(Cor)の分散(V)を示します。標本数が多くなると分散がそれに応じて減少します（→「相関係数の容易度」）。

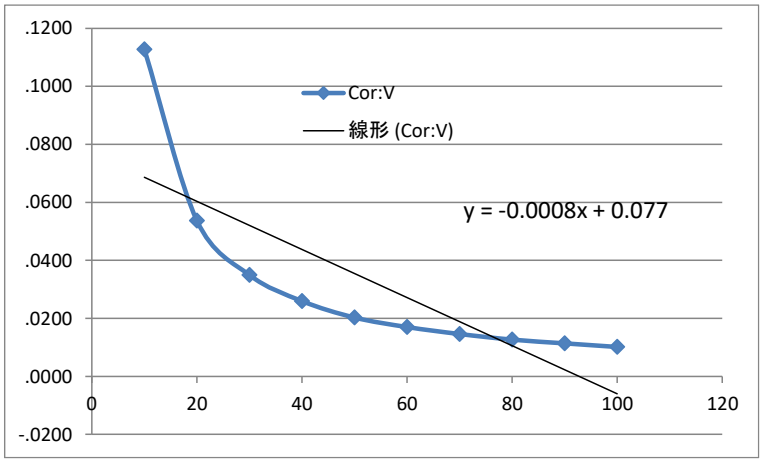
N:V	X	Y
1	10	.113
2	20	.054
3	30	.035
4	40	.026
5	50	.020
6	60	.017
7	70	.015
8	80	.013
9	90	.011
10	100	.010

重回帰分析は 1 変数でも可能なので（「単回帰分析」とよばれます）、上左表(N:V)をそのまま重回帰分析をすると、次の結果になりました。

M.Regres.	Y	Derived	Res.
1	.113	.069	.044
2	.054	.060	-.007
3	.035	.052	-.017
4	.026	.044	-.018
5	.020	.035	-.015
6	.017	.027	-.010
7	.015	.019	-.004
8	.013	.011	.002
9	.011	.002	.009
10	.010	-.006	.016

M.Coef.	Value
Cor:N	-.001
Intercept	.077
Res.Ratio	.455

エクセルで「近似曲線」「線形近似」でグラフを見ると、直線式には近似していないことがわかります。



このように線形近似できない分布の解決法として、次のような目的変数に指数を使った回帰式

$$\text{指数回帰式：} Y^E = W(0) + W(1)\ X(i, 1) \qquad [i = 1, 2, \ldots, n]$$

が考えられることがあります。上の例では説明変数 W が 1 つの単回帰式になりますが、それが複数の重回帰式は

$$Y^E = W(0) + W(1)\ X(i, 1) + W(2)\ X(i, 2) + \ldots + W(p)\ X(i, p)$$

ここで目的変数 Y の指数 E を-1 にすると、W と Y の関係が W = 1 / Y のように反比例になります。下表が Y の指数を-1 とした場合の重回帰分析の結果です。導出変数(Derived variable: DV)は次のように重回帰式の結果の逆数を使います。

$$DV = 1 / [.997\ X + (-1.168)]$$

X^-1.Regres.	Y	Derived	Res.
1	.113	.114	-.001
2	.054	.053	.000
3	.035	.035	.000
4	.026	.026	.000
5	.020	.021	.000
6	.017	.017	.000
7	.015	.015	.000
8	.013	.013	.000
9	.011	.011	.000
10	.010	.010	.000

X^-1.Coef.	Value
X	.997
Intercept	-1.168
Res.Ratio	.007

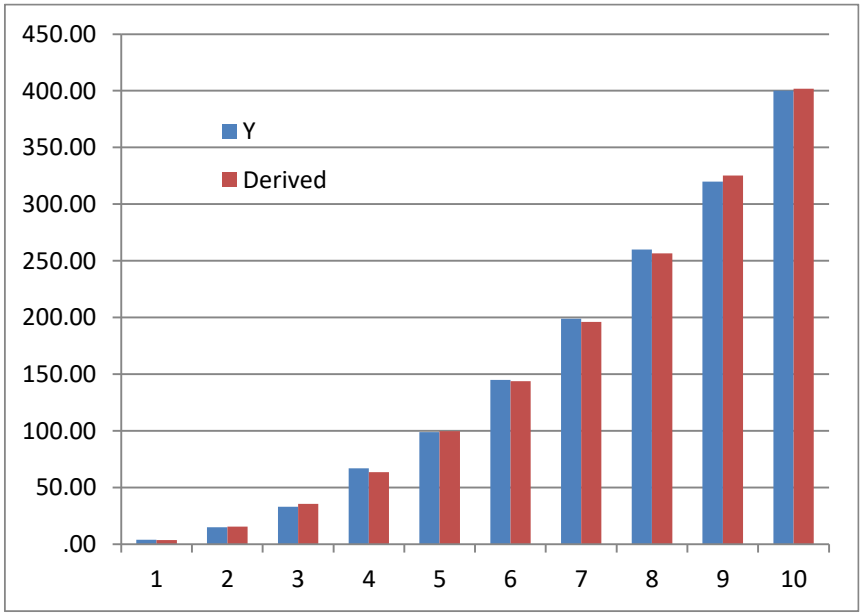
次は指数 E がプラスであると思われるデータ例です。

P	X	Y
1	2	4
2	4	15
3	6	33
4	8	67
5	10	99
6	12	145
7	14	199
8	16	260
9	18	320
10	20	400

ここで E = 2 として指数回帰分析をすると次の結果になりました。

X^2.Regres.	Y	Derived	Res.
1	4.00	3.77	.23
2	15.00	15.62	-.62
3	33.00	35.56	-2.56
4	67.00	63.59	3.41
5	99.00	99.72	-.72
6	145.00	143.93	1.07
7	199.00	196.23	2.77
8	260.00	256.63	3.37
9	320.00	325.12	-5.12
10	400.00	401.69	-1.69

X^2.Coef.	Value
X	1.01
Intercept	-.07
Res.Ratio	.01



実際のプログラムでは、指数を[-5, 5]の範囲で順次変えながら重回帰分

析を行い、それらの結果の中から、最少の残差を示す指数を選択します。最少残差の指数が 1 であれば、先に扱った重回帰分析と同じ結果になります。

●主成分回帰分析

後述の主成分得点の相関はゼロになるので（→主成分分析）、この特徴を利用すれば、先述の重回帰分析の多重共線性の問題を回避することができます。下左図は成績(E:英語, L: ラテン語, M:数学)と、年間に読んだ本(H:人文学；S:自然科学)の冊数を示す架空のデータ例です。H/(H+S)は人文書の相対頻度を示します[13]。はじめに、E, L, M の列だけを用いて主成分分析(PCA)をしました。下右表は主成分得点です。

D	E	L	M	H	S	H/(H+S)
d6	58	34	90	1	9	.10
d7	50	53	100	2	8	.20
d1	45	48	66	2	7	.22
d3	58	51	78	3	7	.30
d5	43	44	32	2	4	.33
d2	56	59	54	3	2	.60
d4	77	72	20	18	2	.90

PCA	#1	#2	#3
d6	-1.38	.79	-1.03
d7	-.92	.77	.83
d1	-.82	-.54	.32
d3	-.18	.59	.01
d5	-.51	-1.67	-.27
d2	.64	-.15	.37
d4	3.17	.22	-.24

下の主成分分析を見ると、第 1 主成分(#1)はラテン語(L), 英語(E)と、数学(M)を分けているので、総合的に「言語」を示す成分であると考えます。これが固有値の大部分を占める重要な成分です(2.03)。第 2 主成分(#2)は英語(E)と数学(M)がラテン語(L)に対立しているので、「現代性」を示す成分と見てよいでしょう。第 3 主成分(#3)はラテン語と数学に反応しているので「教養」を示していると考えられますが、その固有値が小さいのであまり大きな意味をもちません。

PCAv	#1	#2	#3
E	.57	.62	-.55
L	.63	.09	.77
M	-.52	.78	.34

C.Cor	#1	#2	#3
#1	1.00	.00	.00
#2	.00	1.00	.00
#3	.00	.00	1.00

PCAe	#1	#2	#3
E.value	2.03	.67	.30

下右表は左表の列最小値を引いて、最小値を 0 にした得点です。これで

[13] 傾向をつかむために、H/(H+S)の列で昇順にソートしました。

も主成分間の相関は 0 のままです[14]。

PCA	#1	#2	#3	H/(H+S)
d6	-1.38	.79	-1.03	.10
d7	-.92	.77	.83	.20
d1	-.82	-.54	.32	.22
d3	-.18	.59	.01	.30
d5	-.51	-1.67	-.27	.33
d2	.64	-.15	.37	.60
d4	3.17	.22	-.24	.90
Min	-1.38	-1.67	-1.03	

PCA	#1	#2	#3	H/(H+S)
d6	.00	2.46	.00	.10
d7	.47	2.43	1.86	.20
d1	.56	1.12	1.35	.22
d3	1.21	2.26	1.03	.30
d5	.87	.00	.76	.33
d2	2.02	1.52	1.39	.60
d4	4.55	1.89	.79	.90

次は、人文学の読書傾向指数 H/(H+S)を目的変数としたロジスティック回帰分析の結果です。これを見ても、「言語」の主成分(#1)と「現代性」(#2)の主成分が係数として強く働いていることがわかります。

L.Regres.	H/(H+S)	Derived	Res.
d6	.100	.097	.003
d7	.200	.205	-.005
d1	.222	.249	-.027
d3	.300	.304	-.004
d5	.333	.331	.003
d2	.600	.541	.059
d4	.900	.907	-.007

L.Coef.	Value
#1	.922
#2	.802
#3	.234
Intercept	-1.686
Res.Ratio	.040

■文字頻度の変遷と年代

下左表は 13~19 世紀の文字母数を揃えたスペイン語文献の特定の文字の頻度と　文献の成立年代(Y)を示します。下右表は重回帰分析の結果です。<*>は文字が略されている箇所の頻度を示します。

Obra	<*>	ñ	è	á	τ	Y
Cid	836				144	1207
Fazienda	902				157	1220
Alcalá	921				444	1230
GE	1,349				301	1270
Alexandre	877				78	1300
Lucanor	1,877				227	1330
Troyana	1,105				399	1350

Obra	Y	Expected	Residual
Cid	1207	1396	-189
Fazienda	1220	1382	-162
Alcalá	1230	1249	-19
GE	1270	1266	4
Alexandre	1300	1421	-121
Lucanor	1330	1241	89
Troyana	1350	1249	101

[14] 主成分得点の平均は 0 になり、そのようなデータの重回帰分析は結果が正しくありません。

LBA	1,366			146	1389	LBA	1389	1335	54
Alba	464	156		543	1433	Alba	1433	1485	-52
Especulo	1,024	52		215	1450	Especulo	1450	1419	31
Gramática	577	51	4	192	1492	Gramática	1492	1482	10
Celestina	573	41		131	1499	Celestina	1499	1491	8
Sumario	329	70		322	1514	Sumario	1514	1474	40
Diálogo	561				1535	Diálogo	1535	1492	43
Lazarillo	297	33		142	1554	Lazarillo	1554	1505	49
Casada	139	40			1583	Casada	1583	1598	-15
Quijote	165	57	3	2	1605	Quijote	1605	1621	-16
Buscón	93	47	7	1	1626	Buscón	1626	1617	9
Criticón	147	45	20		1651	Criticón	1651	1616	35
Instante	4	21	94	2	1677	Instante	1677	1641	36
Austria	7	60	39		1704	Austria	1704	1665	39
Autoridades		27	3	196	1726	Autoridades	1726	1780	-54
Picarillo	4	123	108		1747	Picarillo	1747	1798	-51
Delincuente		42		229	1787	Delincuente	1787	1831	-44
Ortografía		35		93	1815	Ortografía	1815	1694	121
Diablo		55		223	1841	Diablo	1841	1845	-4
Sombrero		89		222	1874	Sombrero	1874	1894	-20
Perfecta		63		184	1899	Perfecta	1899	1820	79

次は切片と変数の係数を示します。

Intercept	<*>	ñ	è	á	τ	Std res.
1554.853	- .112	1.475	.572	.936	- .457	70.948

略字<*>と接続詞の τ の係数がマイナスなので、年代の推移と逆相関していることがわかります。一方、スペイン語特有文字のエニェ ñ や、アクセント符号がついた母音文字は年代の推移と相関しています。しかし、標準残差が 70 なので、これらの文字の出現による予測はかなり困難であることがわかります。

■名詞・形容詞の頻度ランクに影響する要因

口語体のスペイン語の名詞と形容詞の頻度ランクに影響する要因として、文法的な品詞・性・数、語彙的な派生・合成の有無、語形の長さ、意味の多様性、そして実際的な使用範囲（→拡散度）などが考えられます。品詞(C.G.)の違いをダミー変数として名詞を 1, 形容詞を 2 とし、性(Gén.)については男性を 1, 共性を 2, 女性を 3 とし、数(Núm.)については単数を 1、単複同形を 2, 複数を 3 とします。派生(Deriv.)では、語根語を 1 とし、接頭辞・接尾辞のある派生語・合成語は 2 とします。意味の多様性を数量化

するために、ここでは便宜的に西和辞典の語義数を使います[15]。使用範囲（拡散度´Disp.）は多分野の頻度調査の結果を利用します[16]。最後に口語体のスペイン語のコーパスから頻度表を作成し、そのランクを降順にし(Rank)、さらに 100 位ごとのグループ(R.100)を計算しました[17]。

Forma	C.G.	Gén.	Núm.	Deriv.	Sílaba	Fon.	Sem.	Disp.	Rank	R.100
día	1	1	1	1	2	3	6	0.798	1	1
casa	1	3	1	1	2	4	8	0.679	2	1
años	1	1	3	1	2	4	4	0.917	3	1
(…)										

次にこれらの変数の間の相関行列を出力しました。

Forma	C.G.	Gén.	Núm.	Deriv.	Sílaba	Fon.	Sem.	Disp.	R.100
C.G.	1.000	-.047	-.026	.126	-.124	.091	.071	.074	.065
Gén.	-.047	1.000	-.050	.233	.159	.114	-.059	-.003	.037
Núm.	-.026	-.050	1.000	-.069	-.020	.197	.038	.042	.040
Deriv.	.126	.233	-.069	1.000	.521	.615	-.249	-.151	.200
Sílaba	-.124	.159	-.020	.521	1.000	.804	-.357	-.253	.145
Fon.	.091	.114	.197	.615	.804	1.000	-.305	-.193	.224
Sem.	.071	-.059	.038	-.249	-.357	-.305	1.000	.350	-.164
Disp.	.074	-.003	.042	-.151	-.253	-.193	.350	1.000	-.253
R.100	.065	.037	.040	.200	.145	.224	-.164	-.253	1.000

上表によれば品詞、性、数は頻度ランクとあまり相関がありません。派生・合成の有無、音節数、音素数は互いに相関しているので、それらの中から頻度ランクと一番相関が高い音素数で代表させることができます（正相関：音素数が多いほど頻度ランクは大きい）。意味の多様性と使用範囲も頻度ランクに関係する要因として加えます。どちらも弱い逆相関となっています（意味の多様性・使用範囲が大きいほどランク順位は小さい）。

そこで重回帰分析の入力データを次のように設定します。

Forma	Fon.	Sem.	Disp.	R.100
día	3	6	0.798	1
casa	4	8	0.679	1
años	4	4	0.917	1
(…)				

[15] 上田・ルビオ『プエルタ新スペイン語辞典』研究社
[16] Juilland, A and Chang-Rodríguez, E. (1964) *Frequency Dicitonary of Spanish Words*. London. Mouton
[17] Antonio Moreno Sandoval 氏より提供。

次が重回帰分析の結果です。

Forma	Fon.	Sem.	Disp.	R.100	R.100^	Res.
día	3.000	6.000	.798	1.000	4.289	3.289
casa	4.000	8.000	.679	1.000	4.737	3.737
años	4.000	4.000	.917	1.000	4.309	3.309

(…)

Mr.w.	Fon.	Sem.	Disp.	Interc.
Org.	.234	-.034	-2.371	5.685
Std.	.172	-.039	-.206	.000

上表を見ると残差(Res.)がかなりあることがわかります。このことは次の残差平均(Rn.m.)と残差比(Res.R.)を見るとさらに確認できます。

Eval.	Rn.m.	Yn.m.	Res.R.
Value	2.520	5.828	.432

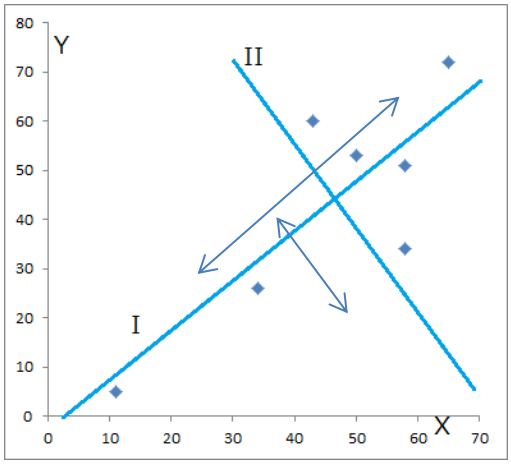
参考までに下に全体の相関係数を掲げます。

Correl.	Fon.	Sem.	Disp.	R.100
Fon.	1.000	-.305	-.193	.224
Sem.	-.305	1.000	.350	-.164
Disp.	-.193	.350	1.000	-.253
R.100	.224	-.164	-.253	1.000

7.5. 主成分分析

データの各変数に適当な重み（ウェイト）を共通に掛け、データの変数の分散を最大化し、同時に変数間の相関係数がゼロになるようにすると、そのような重みを掛けた新変数と個体は新たに総合的な意味をもつようになります。

簡単に 2 つの変数だけのデータで新変数を求めることを次のグラフで示すと、次の図の X-Y 軸を回転させた I-II 軸でデータの値を見ることになります。このとき I 軸で見たデータの値の分散（I 軸に平行な矢印）は最大になり、その次に II 軸で見たデータの値の分散（II 軸に平行な矢印）が続きます。



この新変数は**主成分**(Principal Component)とよばれます。たとえば生徒の外国語文解釈テストと外国語作文テストなどの結果を総合して、新しく外国語の総合能力を示すような数値が得られます。この方法は**主成分分析**(Principal Component Analysis: PCA)とよばれます。

データの分散が最大になるように重みを掛けて作られた主成分は、そのデータの特徴（情報）を最も的確に説明します。そして、この主成分と相関しない、ほかの主成分の中でその次に大きな分散を示すもの（第 2 主成分）は、それに続いて的確な説明をする主成分であると考えられます。両方の新変数間に相関をなくすことで、それぞれ独自の解釈が可能になります。そのような主成分は、もとの変数の数だけ存在しますが、そのデータを説明する力は分散が少なくなるにつれて次第に落ちてくるので、最初のいくつかの主成分までを考察するだけで十分です。

そのような重みを求めるために、はじめにデータ行列(D_{np})から縦平均M_pを引き、それを縦標準偏差S_pで割って標準化した行列X_{np}を使います(→「得点」「標準得点」）。

$$X_{np} = (D_{np} - M_p) / S_p$$

これに適当な未知の重みベクトル(W_p)を右積した変数ベクトルを Z_n とします。

[1] $$Z_n = X_{np} W_p$$

この標準化合成変数ベクトル Z_n の分散(V)を求めます。

[2]
$$V = (Z_n^T Z_n) / N \quad \leftarrow \text{分散の定義}$$
$$= (X_{np} W_p)^T (X_{np} W_p) / N \quad \leftarrow \text{それぞれの } Z_n \text{ に[1]を代入}$$
$$= W_p^T X_{np}^T X_{np} W_p / N \quad \leftarrow (A B)^T = B^T A^T \text{（→「行列」）}$$
$$= W_p^T (X_{np}^T X_{np} / N) W_p \quad \leftarrow N \text{ はスカラーなので移動可}$$
$$= W_p^T R_{pp} W_p \quad \leftarrow R_{pp} = X_{np}^T X_{np} / N$$
（→「関係」「相関行列」）

このような未知の重みベクトル W_p の条件として、その成分の 2 乗和を 1 とします[18]。

[3]　$W_p^T W_p = 1$

この条件[3]のもとで[2]分散(V)の最大値を求めるには

$$F = W_p^T R_{pp} W_p - L (W_p^T W_p - 1)$$　←(…)内が[3]の条件

という式 F を W_p で偏微分した値を 0 とします（ベクトルによる微分については→「重回帰分析」）。L は「ラグランジュ乗数」(Lagrange multiplier)とよばれます（→後述「ラグランジュ乗数法」）。

[4a]　$Df(F, W_p) = 2 R_{pp} W_p - 2 L W_p = 0$　←F を Wp で微分
[4b]　$R_{pp} W_p = L W_p$　←左辺 $L W_p$ を右辺に移項

上の[4b]式は「行列*ベクトル＝スカラー*ベクトル」という「固有方程式」(characteristic equation)の形になっています。ラグランジュ乗数(L)が固有方程式のなかのスカラーの位置にあり、これは「固有値」(eigen value)とよばれます。一方、固有方程式の左辺と右辺に共通するベクトル W_p は「固有ベクトル(eigen vector)）とよばれます。固有方程式の左辺にある 1 つの行列からプログラムは固有値と固有ベクトルを同時に求めます（→後述「固有値問題」）。

なお、固有値 L は次の演算によって分散(V)になることがわかります。

V	$= W_p^T R_{pp} W_p$	←[2]
	$= W_p^T L W_p$	←[4b] $R_{pp} W_p = L W_p$
	$= L W_p^T W_p$	←L はスカラーなので移動可
	$= L$	←[3] $W_p^T W_p = 1$

固有値も固有ベクトルも変数の数だけ存在します。それらを新しい合成変数（「成分」component）として、固有値（＝分散）の大きさによって順に成分番号(#1, #2, ..., #P: P は変数の数)をつけます。

下左表(D)は成績のデータ例です(E:英語, L: ラテン語, M:数学)。下右表(std.PCS)は、標準化したデータ行列に固有ベクトル(Wp)が掛けられた値で「主成分得点」(Principal Component Score: PCS)とよばれます(←[1] Z_n)。下右表(std.PC.s.)では、主成分得点の標準偏差が 1 になるように標準化しました（→「得点」「標準得点」）。

[18] このような条件をつけないと重みベクトルは無数に存在することになるからです。

D	E	L	M
h1	58	34	90
h2	50	53	100
h3	45	48	66
h4	58	51	78
h5	43	44	32
h6	56	59	54
h7	77	72	20

std.PC.s.	#1	#2	#3
h1	-.972	.963	-1.864
h2	-.643	.934	1.515
h3	-.578	-.663	.591
h4	-.124	.718	.013
h5	-.358	-2.035	-.491
h6	.446	-.182	.671
h7	2.228	.266	-.435

次が主成分得点（上右表:PC.s.）のそれぞれの変数(#1, #2, #3)の平均と分散を示します。

値	#1	#2	#3
平均値	.000	.000	.000
分散	2.026	.672	.303

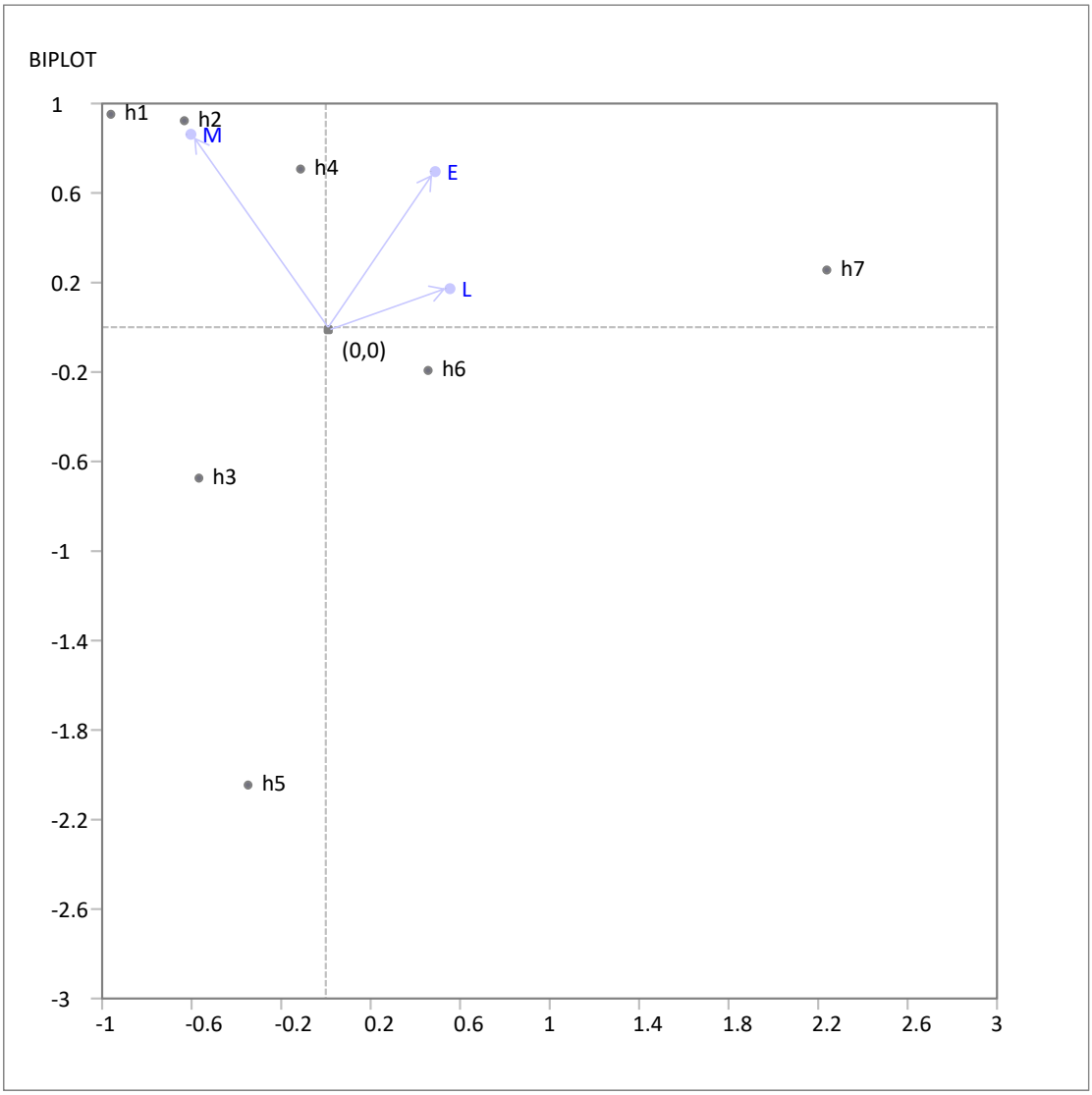
下左表(PcMr.e)は固有値(E.value)(とその寄与率(Ratio))、累積寄与率(Ac.ratio)を示します。累積寄与率から、第 1 成分(#1)と第 2 成分(#2)だけでほとんどの分散(情報)をカバーしていることがわかります(.899 = 89.9%)。そして、下右表(PcMr.v)は、それぞれの主成分の固有ベクトルを示します。これを見ると第 1 主成分(#1)は、英語・ラテン語 vs 数学、つまり「文系・理系」の軸を示し、第 2 主成分(#2)は英語・数学 vs ラテン語、つまり「現代・古典」の軸を示しているようです。

PcMr.e	#1	#2	#3
E.value	2.026	.672	.303
Ratio	.675	.224	.101
Ac.ratio	.675	.899	1.000

PcMr.v	#1	#2	#3
E	.569	.616	-.545
L	.635	.093	.767
M	-.523	.782	.338

●バイプロット

固有ベクトルの変数の成分の軸と個体の主成分得点の個体の軸は、同じ意味（たとえば「文系・理系」と「現代・古典」）を持つ、と考えられるので、第 1 成分(#1)と第 2 成分(#2)から成る平面上に変数と個体を同じ#1, #2 の平面でプロットした散布図を作ります。この図は「バイプロット」(Biplot)とよばれます。この図によって変数と個体の関係を明らかにすることができます。たとえば第 1 主成分(#1)については E, L, h6, h7 が近くにあり、第 2 主成分(#2)については E, M, h1, h2, h4 が近くなることがわかります。また、M の方向に h1, h2, h4 があることがわかります。このように変数と個体間の関係は、その「近さ」ではなく「方向」（向き）で見るべきです。一方、同じ個体間の関係は方向と近さを考慮します。そのとき、変数の方向が参考になります。



●固有値ベクトル・固有行列

下左表はデータ行列、下右表はその相関係数行列です。

D	E	L	M
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

R_{pp}	E	L	M
E	1.000	.643	-.335
L	.643	1.000	-.545
M	-.335	-.545	1.000

このような正方行列 R_{pp} において

$$R_{pp}\ W_p = L\ W_p$$

の等式（「固有値問題」(characteristic equation)とよばれます）が成り立つとき、この式の中の数値 L は「固有値」(eigen value) とよばれ、ベクトル Wp は「固有ベクトル」(eigen vector)とよばれます。ここで未知数の固有値(L)と固有ベクトル(W_p)は最大で R_{pp} の列（=行）の数だけあるので、ここではそれらの集合を「固有値ベクトル」(eigen value vector: L_p)と、「固有行列」(eigen matrix: E_{pp})とよぶことにします。よって、先の式は次のようになります[19]。

$$R_{pp}\ E_{pp} = L_p * E_{pp}$$

下左表が相関行列(R_{pp})、下中表がその固有行列(E_{pp})、下右表が両者の行列積($R_{pp}\ E_{pp}$)です。

R	M	S	L
M	1.000	.643	-.335
S	.643	1.000	-.545
L	-.335	-.545	1.000

X

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

=

R E	#1	#2	#3
M	1.152	.414	-.165
S	1.286	.062	.232
L	-1.060	.526	.102

次の左表が上の相関行列の固有値ベクトル(Lp)、中表がその固有行列(E_{pp})、右表が両者の積($Lp\ E_{pp}$)です。ここで上と下のそれぞれの右表が同じになることを確認してください($R_{pp}\ E_{pp} = Lp\ E_{pp}$)。

L	#1	#2	#3
Value	2.026	.672	.303

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

L E	#1	#2	#3
M	1.152	.414	-.165
S	1.286	.062	.232
L	-1.059	.526	.102

次のように固有行列の中のそれぞれの固有ベクトルは長さが 1 になり，内積がゼロになることを確認します。($E_{pp}^T\ E_{pp} = I_{pp}$ [単位行列])。

E^T	M	S	L
#1	.569	.635	-.523
#2	.616	.093	.782
#3	-.545	.767	.338

X

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

=

E^T E	1	2	3
1	1.000	.000	.000
2	.000	1.000	.000
3	.000	.000	1.000

●固有ベクトルの直交性

縦の固有ベクトルからなる行列 E_{pp} を構成する 2 つの固有ベクトル $E_{p(i)}$

[19] 一般に線形代数の本は、$L_p * E_{pp}$ のようにベクトルと行列の要素間の積を定義していないので、この演算を可能にするために、やや複雑なベクトル→行列の対角化、という操作をしますが、このテキストでは先にベクトルと行列の要素間の積を定義してこれを使います。→「行列」の章。

と $E_{p(j)}$の行列積は 0 になります。

$$E_{p(i)}{}^{T}\, E_{p(j)} = 0 \quad [i \neq j]$$

このことを次のようにして導きます。

1. $R_{pp}\, E_{p(j)} = L_{(j)}\, E_{p(j)}$ ←固有値問題の定義

1.の両辺に同じ操作をします。

2. $E_{p(i)}{}^{T}\, R_{pp}\, E_{p(j)} = E_{p(i)}{}^{T}\, L_{(j)} * E_{p(j)}$ ←1 の両辺に $E_{p(i)}{}^{T}$を左積
3. $= L_{(j)}\, E_{p(i)}{}^{T}\, E_{p(j)}$ ←$L_{(j)}$はスカラーなので移動可

2.の左辺を変形します。

4. $E_{p(i)}{}^{T}\, R_{pp}\, E_{p(j)} = [R_{pp}{}^{T}\, E_{p(i)}]^{T}\, E_{p(j)}$ ←$B^{T} A = (A B)^{T}$
5. $= [E_{p(j)}{}^{T}\, R_{pp}\, E_{p(i)}]^{T}$ ←$B^{T} A = (A^{T} B)^{T}$
6. $= [E_{p(j)}{}^{T}\, L_{(i)}\, E_{p(i)}]^{T}$ ←固有値問題：$R\, E_{p} = L\, E_{p}$
7. $= L_{(i)}\, [E_{p(j)}{}^{T}\, E_{p(i)}]^{T}$ ←L はスカラーなので移動可
8. $= L_{(i)}\, E_{p(i)}{}^{T}\, E_{p(j)}$ ←$(A^{T} B)^{T} = B^{T} A$

2.と 4.の左辺どうしは同じなので、右辺どうしも同じ、よって

9. $L_{(j)}\, E_{p(i)}{}^{T}\, E_{p(j)} = L_{(i)}\, E_{p(i)}{}^{T}\, E_{p(j)}$ ←3. = 8.
10. $[L_{(i)} - L_{(j)}]\, E_{p(i)}{}^{T}\, E_{p(j)} = 0$ ←左辺を右辺に移項して整理
11. $E_{p(i)}{}^{T}\, E_{p(j)} = 0$ ←$L_{(i)} \neq L_{(j)}$

ベクトル成分の積和が 0 であることは、それらのベクトルが直交していることを示します。また、前提として固有ベクトルの長さは 1 とします。

12. $E_{p(i)}{}^{T}\, E_{p(i)} = 1$

11.と 12.をすべての固有ベクトルについてみると、次の式になります。

13. $E_{pp}{}^{T}\, E_{pp} = I_{pp}$ ←E_{pp}は単位行列

＊固有行列の直交性については足立(2005)を参照しました。

●スペクトル分解

次の固有値問題の式

$$R_{pp}\, E_{pp} = L_{p}\, E_{pp}$$

の R_{pp} は次のように分解できます。これは「スペクトル分解」(spectoral decomposition)とよばれ、次に扱う「冪乗法」(べきじょうほう)による固有

値問題の解決で使います。

a. $R_{pp} = L_{(1)} E_{(1)} E_{(1)}^T + L_{(2)} E_{(2)} E_{(2)}^T + \ldots + L_{(p)} E_{(p)} E_{(p)}^T$

ここで(1), (2), ..., (p)は、それぞれ固有値 L とそれに対応する固有ベクトル E_p を示します。この式を導くために次を準備します。

b1. $E_{pp}^T E_{pp} = I_{pp}$ ←先述の固有行列 E の直交性を示す 13.の式
b2. $E_{pp}^{-1} E_{pp} = I_{pp}$ ←逆行列の定義: $X^{-1} X = I$
b3. $E_{pp}^T = E_{pp}^{-1}$ ←b1, b2
b4. $(E_{pp}^T)^{-1} E_{pp}^T = I_{pp}$ ←逆行列の定義: $X^{-1} X = I$
b5 $(E_{pp}^{-1})^T E_{pp}^T = I_{pp}$ ←逆行列の規則: $(X^T)^{-1} = (X^{-1})^T$
b6 $(E_{pp}^T)^T E_{pp}^T = I_{pp}$ ←b3
b7 $E_{pp} E_{pp}^T = I_{pp}$ ←転置行列の性質: $(X^T)^T = X$
b8 $E_{pp}^T E_{pp} = E_{pp} E_{pp}^T = I_{pp}$ ←b1, b7

これで準備ができたので固有値問題から始めます。

c1. $R_{pp} E_{pp} = L_p E_{pp}$ ←固有値問題
c2. $R_{pp} E_{pp} E_{pp}^T = L_p E_{pp} E_{pp}^T$ ←両辺に E_{pp}^T を右積
c3. $R_{pp} E_{pp}^T E_{pp} = L_p E_{pp} E_{pp}^T$ ←b8: $E_{pp}^T E_{pp} = E_{pp} E_{pp}^T$
c4. $R_{pp} E_{pp}^{-1} E_{pp} = L_p E_{pp} E_{pp}^T$ ←b3: $E_{pp}^T = E_{pp}^{-1}$
c5. $R_{pp} I_{pp} = L_p E_{pp} E_{pp}^T$ ←c4, b2: $E_{pp}^{-1} E_{pp} = I_{pp}$
c6. $R_{pp} = L_p E_{pp} E_{pp}^T$ ←R I = R

次に、c6.の $E_{pp} E_{pp}^T$ の行列積を展開します。

E_{pp}	(1)	(2)	(…)	(p)
1	e_{11}	e_{12}	…	e_{1p}
2	e_{21}	e_{22}	…	e_{2p}
…	…	…	…	…
p	e_{p1}	e_{p2}	…	e_{pp}

X

E_{pp}^T	1	2	…	p
(1)	e_{11}	e_{21}	…	e_{p1}
(2)	e_{12}	e_{22}	…	e_{p2}
(…)	…	…	…	…
(p)	e_{1p}	e_{2p}	…	e_{pp}

=

$e_{11}e_{11}+e_{12}e_{12}+\ldots+e_{1p}e_{1p}$	$e_{11}e_{21}+e_{12}e_{22}+\ldots+e_{1p}e_{2p}$	…	$e_{11}e_{p1}+e_{12}e_{p2}+\ldots+e_{1p}e_{pp}$
$e_{21}e_{11}+e_{22}e_{12}+\ldots+e_{2p}e_{1p}$	$e_{21}e_{21}+e_{22}e_{22}+\ldots+e_{2p}e_{2p}$	…	$e_{21}e_{p1}+e_{22}e_{p2}+\ldots+e_{2p}e_{pp}$
…	…	…	…
$e_{p1}e_{11}+e_{p2}e_{12}+\ldots+e_{pp}e_{1p}$	$e_{p1}e_{21}+e_{p2}e_{22}+\ldots+e_{pp}e_{2p}$	…	$e_{p1}e_{p1}+e_{p2}e_{p2}+\ldots+e_{pp}e_{pp}$

固有行列(E)を構成する縦ベクトル（固有ベクトル）$E_{(1)}$, $E_{(2)}$, … $E_{(p)}$について次の式を展開します。$E_{(1)}^T E_{(1)}$のような積でないのでスカラーにはなりません。はじめに、固有行列の第 1 行の固有ベクトルの積を見ます。

$E_{(1)} E_{(1)}^T =$

E	(1)
1	e_{11}
2	e_{21}
…	…
p	e_{p1}

X

E^T	1	2	…	p
(1)	e_{11}	e_{21}	…	e_{p1}

=

$e_{11}e_{11}$	$e_{11}e_{21}$	…	$e_{11}e_{p1}$
$e_{21}e_{11}$	$e_{21}e_{21}$	…	$e_{21}e_{p1}$
…	…	…	…
$e_{p1}e_{11}$	$e_{p1}e_{21}$	…	$e_{p1}e_{p1}$+

このように行列積 $E_{(1)}\ E_{(1)}{}^T$ の要素は、先の行列積 $E_{pp}\ E_{pp}{}^T$ の要素のそれぞれの第 1 項になります。2 番目のベクトルの次の積を見ます。

$E_{(2)}\ E_{(2)}{}^T$ =

E	(2)
1	e_{12}
2	e_{22}
…	…
p	e_{p2}

X

E^T	1	2	…	p
(2)	e_{12}	e_{22}	…	e_{p2}

=

$e_{12}e_{12}$	$e_{12}e_{22}$	…	$e_{12}e_{p2}$
$e_{22}e_{12}$	$e_{22}e_{22}$	…	$e_{22}e_{p2}$
…	…	…	…
$e_{p2}e_{12}$	$e_{p2}e_{22}$	…	$e_{p2}e_{p2}$

ここでも行列積 $E_{(2)}\ E_{(2)}{}^T$ の要素が行列積 $E_{pp}\ E_{pp}{}^T$ の要素のそれぞれの第 2 項になることを確認します。同様にして p 番目のベクトルの次の積は

$E_{(p)}\ E_{(p)}{}^T$ =

E	(p)
1	e_{1p}
2	e_{2p}
…	…
p	e_{pp}

X

E^T	1	2	…	p
(p)	e_{1p}	e_{2p}	…	e_{pp}

=

$e_{1p}e_{1p}$	$e_{1p}e_{2p}$	…	$e_{1p}e_{pp}$
$e_{2p}e_{1p}$	$e_{2p}e_{2p}$	…	$e_{2p}e_{pp}$
…	…	…	…
$e_{pp}e_{1p}$	$e_{pp}e_{2p}$	…	$e_{pp}e_{pp}$

よって、それぞれの固有ベクトルの積 $E_{(i)}\,E_{(i)}{}^T$ (i = 1, 2, …, p)を全部足すと、固有行列全体の積 $E\,E^T$ になります。

$$E_{pp}\,E_{pp}{}^T = E_{(1)}\,E_{(1)}{}^T + E_{(2)}\,E_{(2)}{}^T + \ldots + E_{(p)}\,E_{(p)}{}^T$$

よって

$$R_{pp} = L_p\,E_{pp}\,E_{pp}{}^T \quad \leftarrow c6.$$
$$= L_{(1)}\,E_{p(1)}\,E_{p(1)}{}^T + L_{(2)}\,E_{p(2)}\,E_{p(2)}{}^T + \ldots + L_{(p)}\,E_{p(p)}\,E_{p(p)}{}^T$$
$$\leftarrow L_{(1)}, L_{(2)}, \ldots, L_{(p)} \text{ はスカラー}$$

＊スペクトル分解については足立(2005)と岩崎・吉田(2006)を参照しました。

●冪乗法

R_{pp} の固有値ベクトルと固有行列を求めるために、**冪乗法**（べきじょうほう: Power method）を使います。この方法は最大固有値を求め、その残差行列を使って次のステップで残差行列の最大固有値を求める、というステップを次々に列の数だけ行います。

$$R_{pp}\,E_{pp} = Lp\,E_{pp}$$

はじめに、Lp を構成するそれぞれの固有値を L(1), L(2), …, L(p)とすると、これらの固有値の大きな方から順番に取り出す方法を次のように考えます。

E_{pp} を構成するそれぞれの縦ベクトルを $E_p(1), E_p(2), \ldots, E_p(p)$とすると、それらにそれらの和(Sp)の初期状態 $Sp^{(0)}$を次にようにします。

$$Sp^{(0)} = E_p(1) + E_p(2) + \ldots + E_p(p)$$

この両辺に R_{pp} を次々に左積していきます。

$Sp^{(1)} = R_{pp}\,Sp^{(0)} = R_{pp}\,E_p(1) + R_{pp}\,E_p(2) + \ldots + R_{pp}\,E_p(p)$
[両辺に R_{pp} を左積]
$= L(1)\,E_p(1) + L(2)\,E_p(2) + \ldots + L(p)\,E_p(p)$ [←$R_{pp}\,E_p = L\,E_p$]

$Sp^{(2)} = R_{pp}{}^2\,Sp^{(0)} = L(1)^2\,E_p(1) + L(2)^2\,E_p(2) + \ldots + L(p)^2\,E_p(p)$
[さらに両辺に R_{pp} を左積]
(…) [順次両辺に R_{pp} を左積]

$Sp^{(k)} = R_{pp}{}^{k}\ Sp^{(0)} = L(1)^k\ E_p(1) + L(2)^k\ E_p(2) + \ldots + L(p)^k\ E_p(p)$

ここで右辺の L(1), L(2), …, L(p)の中の最大のものを L(m)とします。

$L(m) > L(1), L(2), \ldots, L(p)$

先の式は

$Sp^{(k)} = L(1)^k\ E_p(1) + \ldots + L(m)^k\ E_p(m) + \ldots + L(p)^k\ E_p(p)$ [L(m)が最大 L]
$= L(m)^k\ [L(1)^k/L(m)^k\ A(p)1 + \ldots + E_p(m) + \ldots + L(p)^k/L(m)^k\ E_p(p)]$
[$L(m)^k$を外に出す]

k を十分に大きくすると[…]の中の $E_p(m)$以外は、その係数の分数がゼロに近づくので無視できるほど小さくなります。よって

$Sp^{(k)} \fallingdotseq L(m)^k\ E_p(m) \quad [k \to \infty, \quad L(p)^k/L(m)^k \to 0]$

最初の（最大の）固有値 L(1)と固有ベクトル $E_p(1)$を次の式で求めます。

$L(1) = [Sp^{(k)T}\ Sp^{(k)}]^{1/2}$ [L の長さは 1]
$E_p(1) = Sp^{(k)} / L(1)$ $[Sp^{(k)} \fallingdotseq L(m)^k\ E_p(m)]$

次に大きな固有値 L(2)と固有ベクトル $E_p(2)$を求めるための $R_{pp}(2)$は、最初の $R_{pp}(1)$から一定の行列を引いた残差行列になります。そのために R_{pp}を次のようにスペクトル分解(spectral decomposition）します。

$R_{pp} = Lp * E_{pp}\ E_{pp}{}^{T}$

この式を展開すると次のようなスペクトル分解の式になります。

$R_{pp} = v1\ E_{p1}\ E_{p1}{}^{T} + v2\ E_p2\ E_p2^{T} + \ldots + Lp\ E_{pp}\ E_{pp}{}^{T}$

そこで、上式から $v1\ E_{p1}\ E_{p1}{}^{T}$を除いた残差行列を次のステップの $R_{pp}(2)$とします。

$R_{pp}(2) = R_{pp}(1) - L(1)\ E_p(1)\ E_p(1)^{T}$

この新たな $R_{pp}(2)$を使って、先のプロセスを繰り返します。同じプロセスを、$R_{pp}(3), R_{pp}(4), \ldots, R_{pp}(p)$までのうち、望む固有値の数だけ繰り返して終了します。

次に最初の（最大の）固有値と固有ベクトルを算出します。固有値問題 $R_{pp}\ E_p = L_p\ E_p$の E_pは、それを定数倍しても成立するので無数に存在します。そこで縦ベクトル(E_p)の長さ（2 乗和）を 1 とする条件をつけます。E_{pp}のそれぞれの縦ベクトルを $E_{p(1)}, E_{p(2)}, \ldots, E_{p(p)}$とすると

[1]　　$E_{p(i)}{}^T E_{p(i)} = 1 \quad (i = 1, 2, \ldots, p)$

そして $E_{p(i)}$と $E_{p(j)}$ [i ≠ j] が直交する、という条件を加えます。

$$E_{p(i)}{}^T E_{p(j)} = 0 \quad (i, j = 1, 2, \ldots, p;\ i \neq j)$$

よって

$$E_{pp}{}^T E_{pp} = I_{pp} \text{ (単位行列)}$$

はじめに E_p の初期値を単位ベクトル(I_p)とし

$$E_p \leftarrow I_p$$

そして

[2]　　$E_p \leftarrow R_{pp} E_p$　　←Ep に Rpp を左積
[3]　　$L \leftarrow (E_p{}^T E_p)^{1/2}$
[4]　　$E_p \leftarrow E_p / L$　←Ep の長さを 1 にする[1]
[5]　　Ep の変化が大きければ[4]に戻る、小さければ終了し次へ

の[2]-[4]のプロセスを E_p に変化がなくなるまで繰り返すと、最初の（最大の）固有値 $L_{(1)}$と固有ベクトル $E_{p(1)}$が求められます。このプロセスを具体的に追ってみましょう。

R_{pp}	1	2
1	1	0.8
2	0.5	1

k	0	1	2	3	4	5	6
$E_p \leftarrow R_{pp} E_p$	1.000	1.273	1.280	1.281	1.281	1.281	1.281
	1.000	1.061	1.024	1.015	1.013	1.013	1.012
$L \leftarrow \sqrt{(E_p{}^T E_p)}$	1.414	1.657	1.640	1.634	1.633	1.633	1.632
$E_p \leftarrow E_p / L$	.707	.768	.781	.784	.784	.784	.784
	.707	.640	.625	.621	.620	.620	.620

上右表の k=0 の Ep は初期値です。その下の $L \leftarrow \sqrt{(E_p{}^T E_p)}$は Ep の長さを示します。その下の $E_p \leftarrow Ep / L$ で Ep の長さを 1 に揃えます。

次に k = 1 の列の Ep は Ep ←　Rpp Ep の行列積の結果です。その後は、先と同じことを繰り返します。そして Ep が変化しなくなるまで、K=3, 4, .. と繰り返して、最終的な Ep と L を求めます。

ここで、Rpp Ep = L Ep という関係になることは上の表から明らかです。すなわち、$R_{pp} E_p = (1.281, 1.012)^T = 1.632 * (.784, .620)^T = L * E_p$ となります。

次に先に見たように、R_{pp} をスペクトル分解した

$$R_{pp} = L_{(1)} E_{p(1)} E_{p(1)}{}^T + L_{(2)} E_{p(2)} E_{p(2)}{}^T + \ldots + L_{(p)} E_{p(p)} E_{p(p)}{}^T$$

から、今回算出した第 1 項 $L_{(1)} E_{p(1)} E_{p(1)}{}^T$ を除いた

$$R_{pp}^{(2)} = R_{pp}^{(1)} - L_{(1)} E_{p(1)} E_{p(1)}^{T}$$

を計算し、先のプロセスによって、$R_{pp}^{(2)}$の最大の固有値 $L_{(2)}$と固有ベクトル $E_{p(2)}$を求めます。以下同様に、$R_{pp}^{(p)}$までを求めて、最終的な固有値の集合である固有値ベクトル(L_p)と固有ベクトルの集合である固有行列(E_{pp})を完成します。

R	1	2
1	1	0.8
2	0.5	1

Ev(R)	#1	#2
E.value	1.632	.368

Em(R)	#1	#2
1	.784	.016
2	.620	1.000

＊冪乗法とプログラムについては白井(2009: 99-101)と Nakos and Joyner (1999:467-472)を参照しました。

●ダミー変数相関行列の固有値・固有ベクトル

下左表(Nnp)は名義尺度行列です。それをダミー変数で二値化した行列が下中表(Dnp)です。下右表(Cor.)はその相関行列です。

N_{np}	x1	x2
d1	A	C
d2	A	D
d3	A	D
d4	A	E
d5	B	C
d6	B	D
d7	B	E

D_{np}	A	B	C	D	E
d1	1	0	1	0	0
d2	1	0	0	1	0
d3	1	0	0	1	0
d4	1	0	0	0	1
d5	0	1	1	0	0
d6	0	1	0	1	0
d7	0	1	0	0	1

Cor.	A	B	C	D	E
A	1.000	-1.000	-.091	.167	-.091
B	-1.000	1.000	.091	-.167	.091
C	-.091	.091	1.000	-.548	-.400
D	.167	-.167	-.548	1.000	-.548
E	-.091	.091	-.400	-.548	1.000

ダミー変数行列(Dnp)の特徴は、上の A:B と C:D:E のように、互いに(1, 1)になったり、(0, 0)になったりせず、必ずどれかが 1 であって、そのほかはゼロになることです。たとえば、A が 1 であれば、かならず B が 0 になります。その逆も成り立ちます。C:D:E については、C が 1 であれば、D, E はかならず 0 です。そこで、A の値がわかれば B が決定され、また C, D の値がわかれば E の値が決定されているので、全体の自由度は 1+2 =3 ということになります。

このように情報が冗長な対称行列の固有値は自由度を超えるとすべてゼロになることが以下の出力でわかります。

Ev	#1	#2	#3
E.value	2.159	1.441	1.400

Em	#1	#2	#3
A	-.624	.333	.000
B	.624	-.333	.000
C	.204	.382	-.707
D	-.372	-.698	.000
E	.204	.382	.707

このようにダミー変数は 5 個ですが、その自由度は 3 なので、3 個の固有値・固有ベクトル(#1, #2, #3)になります。それでも、次のように、固有値・固有ベクトルの定義通り(Rpp Ep = L Ep)、相関行列と固有ベクトルの積と、固有値・固有ベクトルの成分積は同じになります。

Cor. Em	#1	#2	#3
A	-1.347	.480	.000
B	1.347	-.480	.000
C	.440	.551	-.990
D	-.803	-1.005	.000
E	.440	.551	.990

Ev*Em	#1	#2	#3
A	-1.347	.480	.000
B	1.347	-.480	.000
C	.440	.551	-.990
D	-.803	-1.005	.000
E	.440	.551	.990

このダミー変数相関行列の固有値・固有ベクトルの性質を、後述する主成分重回帰分析で確認します。

●ラグランジュ乗数法

条件付きの微分には**ラグランジュ乗数法**(Lagrange multiplier method)が使われます。次の関数

[1] $Y = f(x_1, x_2, \ldots, x_n)$

の極値を求めるために、Y の$(x_1, x_2, \ldots, x_n)$による偏微分

$$Df(Y, x_1)=0,\ Df(Y, x_2)=0,\ \ldots,\ Df(Y, x_n)=0$$

から $x_1, x_2, \ldots, x_n$ を求めます。このとき

[2] $G = g(x_1, x_2, \ldots, x_n) = 0$

というような別の条件がついていることがあります。このように条件付きの関数を微分するときには「ラグランジュの未定乗数法」L をつけて

[3] $W = Y - L\ G$
$= f(x_1, x_2, \ldots, x_n) - L\ g(x_1, x_2, \ldots, x_n)$

この W を次のように $x_1, x_2, \ldots, x_n$, L で微分し、W の極値を求めます。

[4]　　$Df(W, x_1) = 0, Df(W, x_2) = 0, \ldots, Df(W, x_p) = 0, \underline{Df(W, L) = 0}$

[3]の W= Y – L G を[4]の各式に代入すると、それぞれ次のようになります。

$$Df(W, x_1) = Df(Y, x_1) - L\ Df(G, x_1) = 0$$
$$Df(W, x_2) = Df(Y, x_2) - L\ Df(G, x_2) = 0$$
$$(\ldots)$$
$$Df(W, x_n) = Df(Y, x_n) - L\ Df(G, x_n) = 0$$

そして、最後の式（[4]下線）は次のようになります。

$$Df(W, L) = Df(Y - L\ G, L) = -G = 0 \quad [Df(Y)はゼロ]$$

よって

$$G = g(x_1, x_2, \ldots, x_n) = 0$$

このように W をそれぞれの未知数で微分すると、たしかに[2]の条件が満たされることがわかります。この理由から条件付き関数を微分するときはその条件に L という乗数(ラグランジュ乗数 Lagrange multiplier)をつけた式(3)を使って $x_1, x_2, \ldots, x_n, L$ を求める、という方法をとります。

＊小林(1967:89-90)を参照しました。

●集中分析

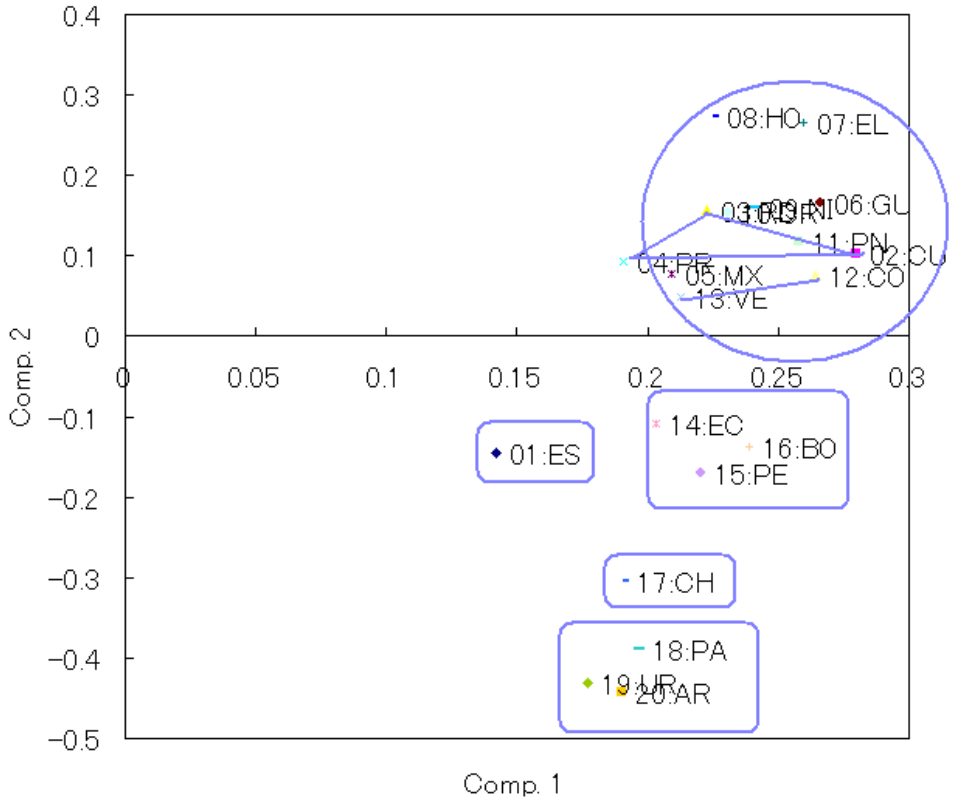
変数の重みと個体の得点を昇順でソートし、得点を並び替えると次のような集中化した得点になります。

PCA.Cct	Latin	English	Physics
B	88	28	20
C	64	43	32
A	59	56	54
F	48	45	66
E	51	58	78
G	22	32	90
D	16	50	100

■地域変異語彙の主成分分析

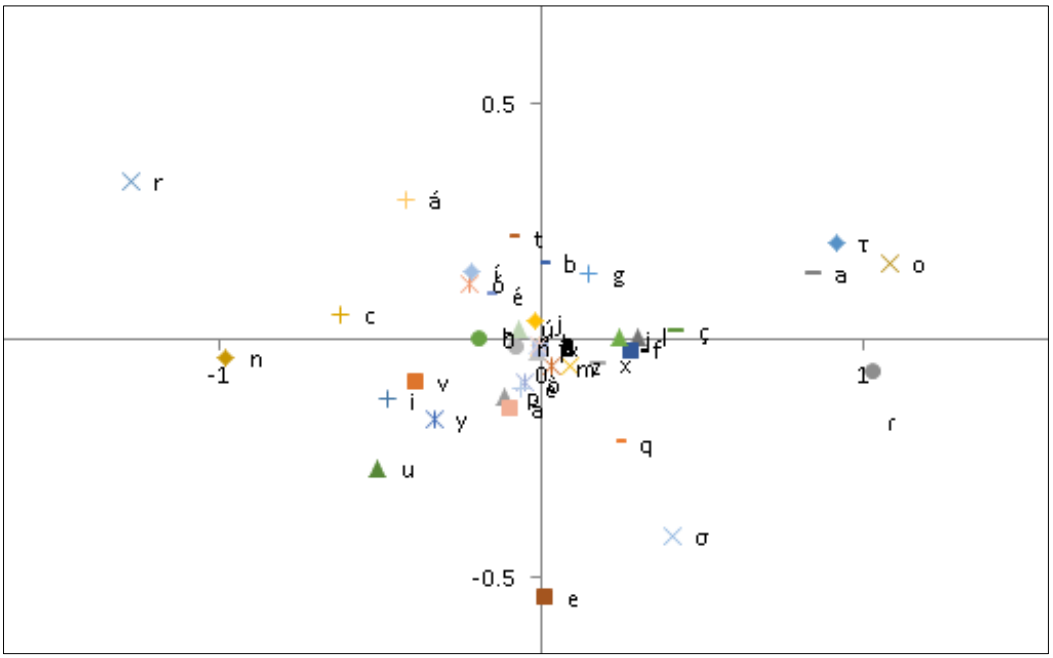
次の図は現代スペイン語の地域変異語彙841語を20か国で調査した結果を主成分分析し、第1主成分を行に、第2主成分を列にして各国をプロットしたものです。第1主成分（行）はスペイン(ES)とラテンアメリカ諸国を分けています。右上の〇で囲んだ国々はメキシコ(MX)・中米諸国(HO, EL,

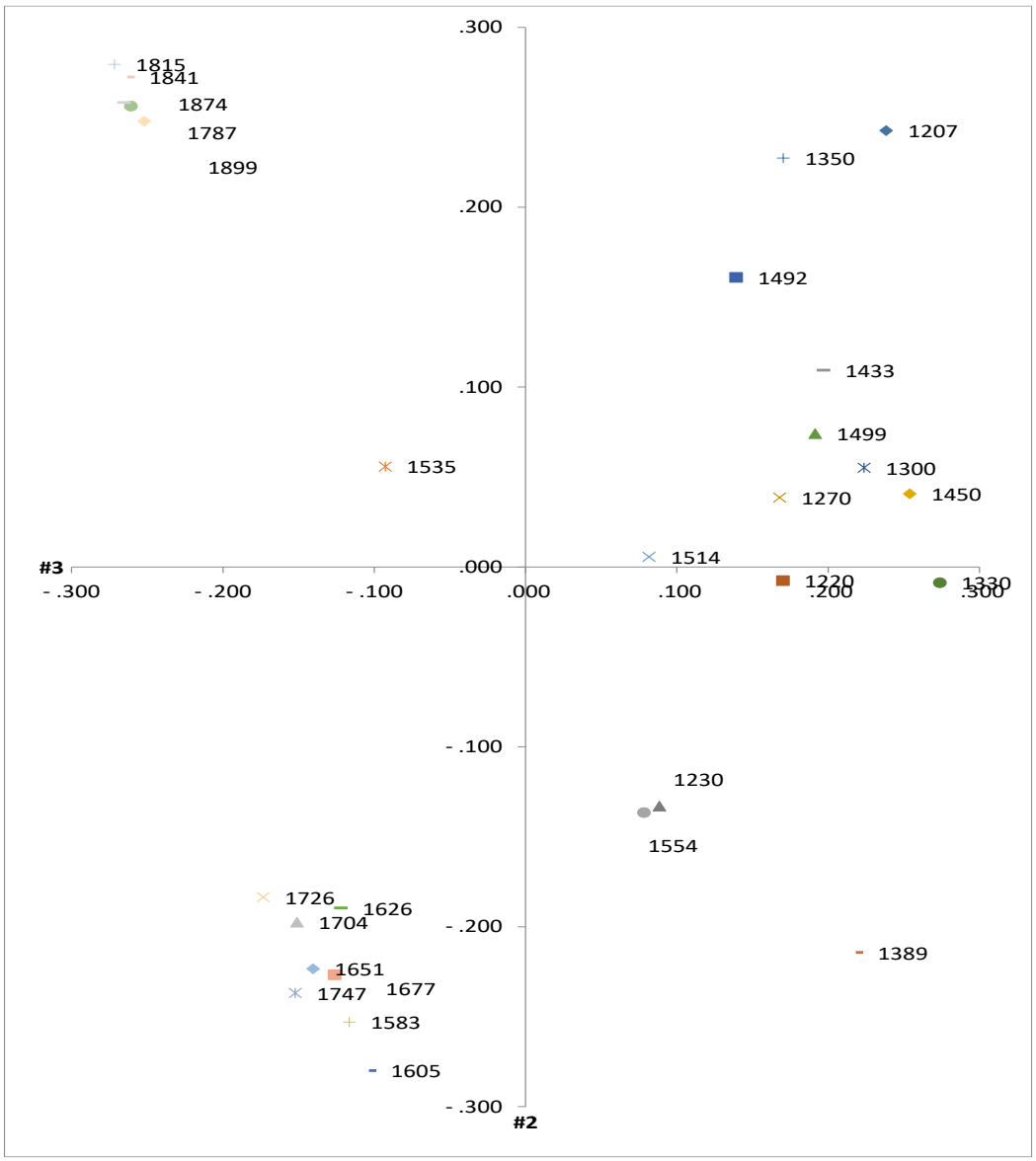
GI, PN)・カリブ海諸国(PR, CI, RD)・コロンビア(CO)・ベネズエラ(VE)です。その下にアンデス諸国(EC, BO, PE)、チリ(CH)、ラプラタ諸国(PA, IR, AR)が続きます。このようにラテンアメリカ諸国は第 2 主成分（列）によっておよそ南北に配置されます。このように地域変異語彙はバラバラに分布するのではなく、一定の地理的な連続性(continuum)を示しています。



■中世近代スペイン語文字使用頻度の主成分分析

13 世紀から 19 世紀までのスペイン語史の中に位置づけられる 28 作品をサンプルにし、使用されているすべての文字の頻度からなる行列を作成しました。それを主成分分析にかけると、第 1 主成分はとくにデータを特徴づけることがありませんが、第 2 主成分（中世と近代）と第 3 主成分（17-18 世紀と 19 世紀）の特徴が明らかに示されています。文字の変異に関しては、とくに s, d, r のバリアントが重要です。


0.5
-0.5
-1
1
r
á
t
b
g
τ
o
a
c
n
v
i
y
u
q
σ
e
r


.300
.200
.100
.000
- .100
- .200
- .300
#3
#2
- .300
- .200
- .100
.000
.100
.200
.300
1815
1841
1874
1787
1899
1207
1350
1492
1433
1499
1300
1270
1450
1535
1514
1220
1330
1230
1554
1726
1626
1704
1651
1747
1677
1583
1605
1389

■スペイン語の硬口蓋鼻子音の文字の発達

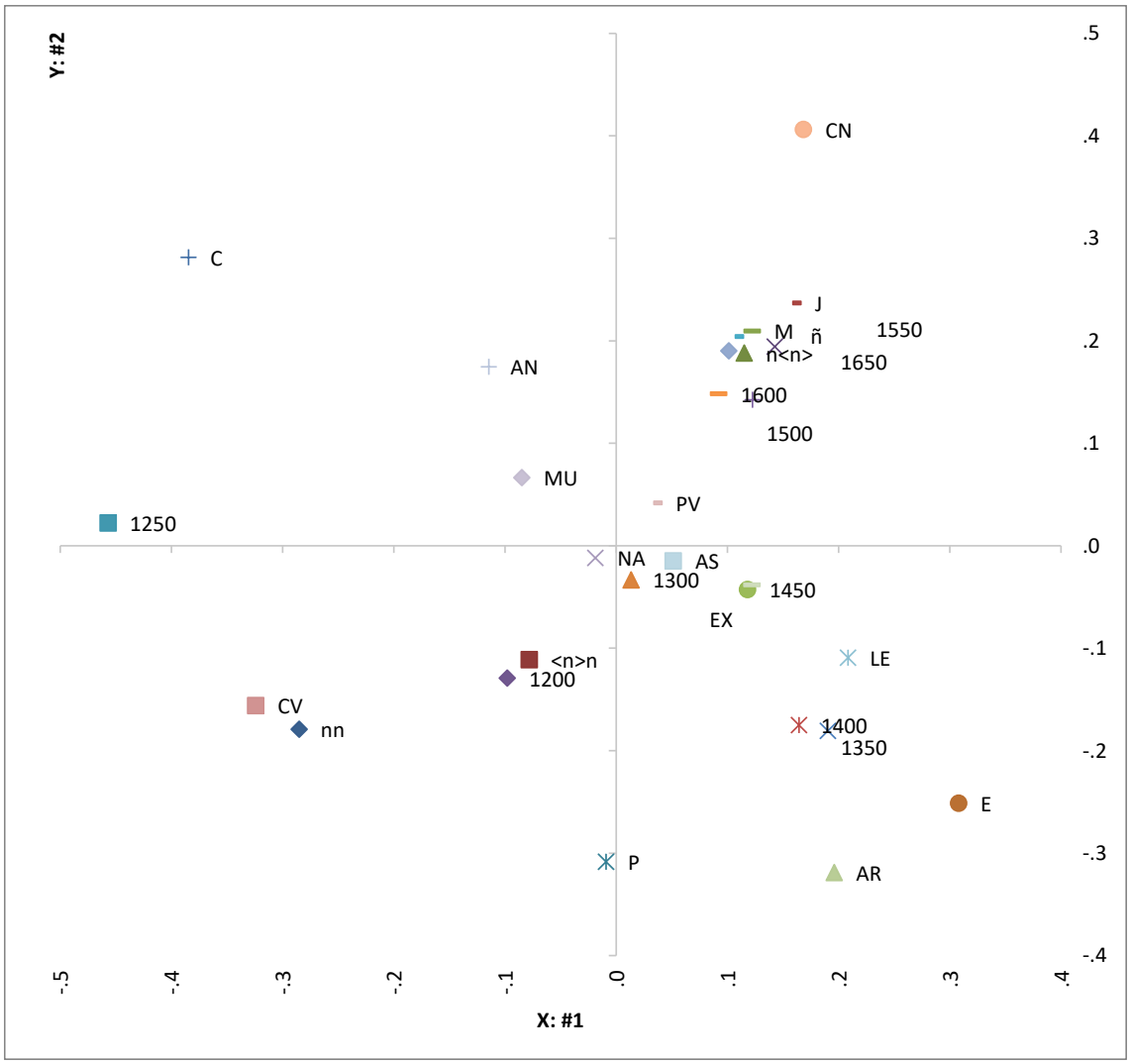
言語データ分析では、さまざまな検索式を使ってアイテム（検索された文字連続）を出力させます。そのときに検索したアイテムには地域・年代・文体・文字種などのパラメータが付加されています。その中から、単数・複数の説明変数と単数の目的変数の関係を論じるときの手段の 1 つとして名義主成分分析が役立ちます。

次は、中世スペイン語の公証文書に現れた文字 4 つの形態 nn, <n>n, n<n>, ñ を目的変数とし、それに関係する文書類(T: P 私文書；C 王室文書など)、年代(年代幅 50 年：Y50)、地域(R)を説明変数とする名義尺度データの一部です。

D	T	Y50	R	N
d1	P	1200	CV	nn
(…)				
d9	P	1200	CV	<n>n
(…)				
d677	C	1250	AN	n<n>
(…)				
d9570	C	1400	CV	ñ

(…)

次は第 1 主成分(#1)と第 2 主成分(#2)を x 軸と y 軸にして、それぞれの名義尺度をプロットした図です。この図から第 1 主成分が初期(13 世紀)と中期(14, 15 世紀)・後期(16, 17 世紀)を分け、第 2 主成分が中期(14, 15 世紀)と後期(16, 17 世紀)を分けていることがわかります。そして、初期の領域に nn, <n>n が配置され、後期の領域に n<n>, ñ が配置されています。旧カスティーリャ地方(CV)は初期の語形を特徴とし、新カスティーリャ地方(NV)は後期の語形を特徴としています。文書類については国事文書(C)が古い語形 nn, <n>n に、教会文書(E)と市会文書(M)と法令文書(J)が新語形 n<n>, ñ に近い位置を占めています。



●名義主成分分析

下左表(N)は名義変数行列で、下右表(Q)はそれをダミー変数で数値化した行列です（→名義尺度の数量化）。

N	x1	x2	y
d1	A	C	X
d2	A	D	X
d3	A	D	Y
d4	A	E	X
d5	B	C	X
d6	B	D	Y
d7	B	E	Z

Q	A	B	C	D	E	X	Y	Z
d1	1	0	1	0	0	1	0	0
d2	1	0	0	1	0	1	0	0
d3	1	0	0	1	0	0	1	0
d4	1	0	0	0	1	1	0	0
d5	0	1	1	0	0	1	0	0
d6	0	1	0	1	0	0	1	0
d7	0	1	0	0	1	0	0	1

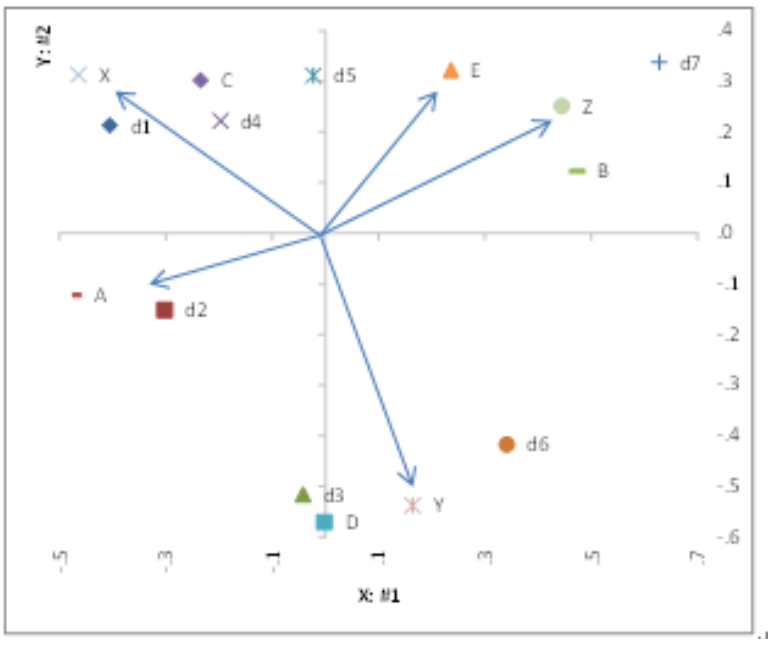
上右表(Q)を主成分分析すると、固有値・固有ベクトルは次のように 5 個まで算出されます。

Pca.d	#1	#2	#3	#4	#5
d1	-.405	.213	-.122	.178	.050
d2	-.302	-.151	.124	-.197	.172
d3	-.042	-.516	.108	.161	-.001
d4	-.197	.222	.358	-.062	-.197
d5	-.022	.312	-.451	-.056	-.035
d6	.341	-.417	-.221	-.072	-.087
d7	.628	.337	.205	.048	.099

Pca.v	#1	#2	#3	#4	#5
A	-.474	-.122	.407	.289	.105
B	.474	.122	-.407	-.289	-.105
C	-.235	.302	-.546	.487	.071
D	-.002	-.570	.009	-.395	.380
E	.236	.322	.537	-.055	-.487
X	-.464	.313	-.079	-.499	-.049
Y	.164	-.537	-.108	.354	-.436
Z	.445	.251	.252	.249	.632

Pca.e	#1	#2	#3	#4	#5
E.value	6.726	6.403	3.867	.926	.744
%	.841	.800	.483	.116	.093
Ac.%	.841	1.641	2.125	2.240	2.333

主成分分析では説明変数と目的変数をとくに区別することなく、両者を同一平面上で扱うことができます。次は、ダミー変数行列(Q)のデータ(d1...d7)と、変数(A...E, P, Q)と、目的変数(X, Y, Z)を同じ平面でプロットした散布図です。このような図はバイプロット(biplot)とよばれます。これを見ると、X の方向に d1, d4, d5, C があり、Y の方向に d3, d6, D があり、Z の方向に E, B, d7 があることがわかります。このように変数(A, B, C, X, Y, Z)と個体(d1, d2, ..., d7)間の関係は、その「近さ」ではなく「方向」（向き）で見るべきです。一方、同じ個体間の関係は方向と近さを考慮します。そのとき変数の方向が参考になります。



●名義化主成分分析

下左表(D)を縦平均の平均値を基準にして 2 名義化した行列が下右表です（→得点）。

D	E	L	M
h1	58	34	90
h2	50	53	100
h3	45	48	66
h4	58	51	78
h5	43	44	32
h6	56	59	54
h7	77	72	20

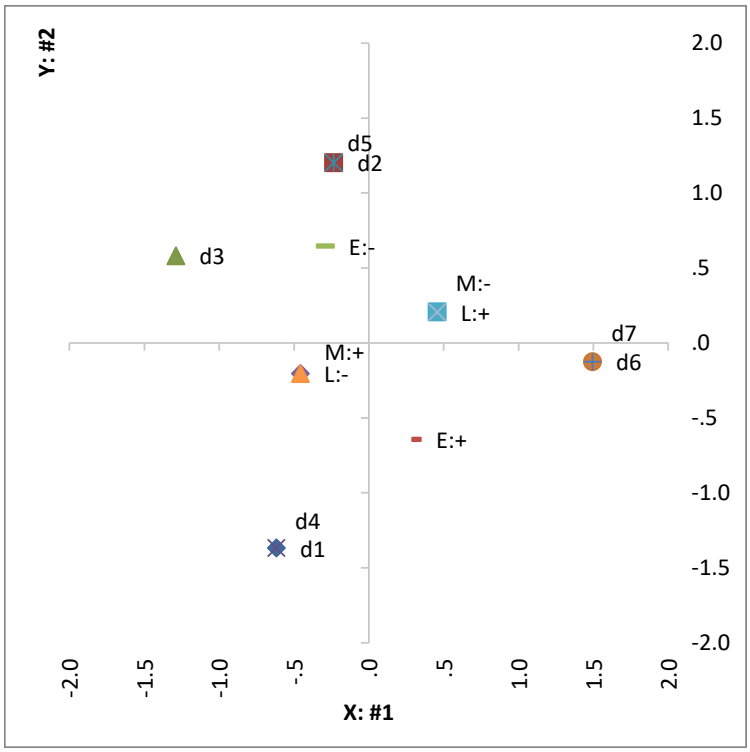
N2	E	L	M
h1	E:+	L:-	M:+
h2	E:-	L:+	M:+
h3	E:-	L:-	M:+
h4	E:+	L:-	M:+
h5	E:-	L:-	M:-
h6	E:+	L:+	M:-
h7	E:+	L:+	M:-

この行列を対象に名義主成分分析をすると次の結果になります。

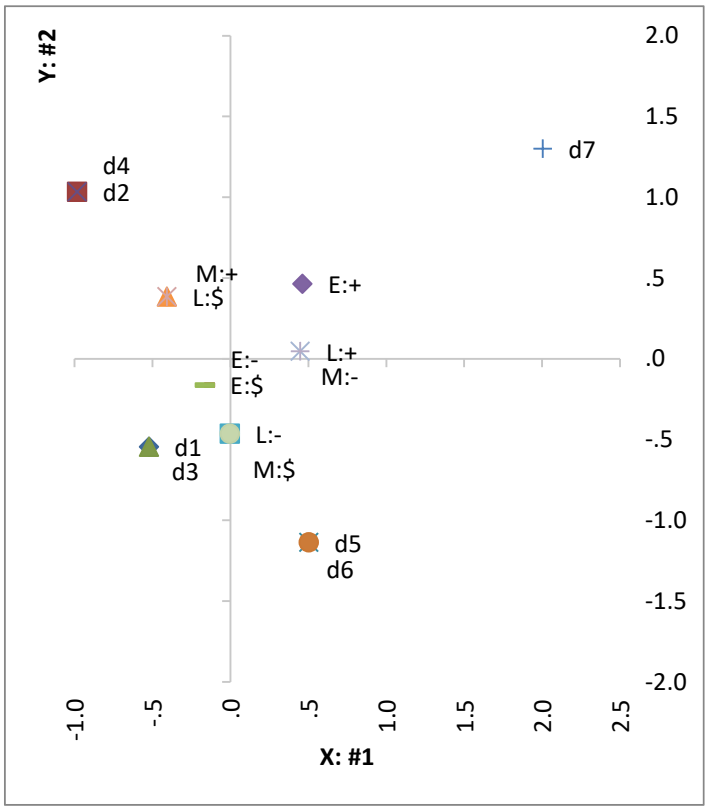
D	#1	#2	#3
h1	-.616	-1.367	.000
h2	-.234	1.202	-1.871
h3	-1.289	.581	.000
h4	-.616	-1.367	.000
h5	-.234	1.202	1.871
h6	1.495	-.125	.000
h7	1.495	-.125	.000

v	#1	#2	#3
E:+	.291	-.645	.000
E:-	-.291	.645	.000
L:-	-.456	-.205	.500
L:+	.456	.205	-.500
M:+	-.456	-.205	-.500
M:-	.456	.205	.500

PCA.e	#1	#2	#3
E.value	3.046	1.788	1.167
%	.508	.298	.194
Ac.%	.508	.806	1.000



この方法を名義化主成分分析(Nominalized Principal Component Analysis: NPCA)とよびます。NPCA は数値の弁別がおおまかになるので全体の傾向を探るために役立ちます。また、変数(E, L, M)のマイナス方向(E:-, L:-, M:-)も示されるので、その変数の正負の位置が参考になります。次のように 3 段階で数値名義化主成分分析をすると、さらに詳細な傾向がわかります。



7.6. 対応分析

対応分析(Correspondence analysis)とよばれる方法はフランスの Jean-Paul Benzécri が開発した多変数解析法です[20]。対応分析では、たとえば次のような得点表の個体（X_n: Ana, Juan, Mary, Ken)と変数(Y_p: English, Physics, Latin)の間の相関係数が最大になるように個体と変数に適当な負荷値を与え、その負荷値によってそれぞれの意味を探ります。

D_{np}	Y_1: English	Y_2: Physics	Y_3: Latin	S_n
X_1: Ana	9	14	18	41
X_2: Juan	17	7	11	35
X_3: Mary	15	13	14	42
X_4: Ken	5	18	8	31
T_p	46	52	51	N=149

[20] Benzécri の対応分析は日本の林知己夫が独自に開発した数量化Ⅲ類と同等のアルゴリズムです。

↓

C_{np}	Y_2: Physics	Y_3: Latin	Y_1: English	X_n
X_4: Ken	18	8	5	-.473
X_1: Ana	14	18	9	-.094
X_3: Mary	13	14	15	.108
X_2: Juan	7	11	17	.400
Y_p	-.361	.028	.377	

行和ベクトル S_n と個体ベクトル X_n の積、および列和ベクトル Tp と変数ベクトル Y_p との積の平均(Mx, My)をそれぞれ 0 とします。

$S_n = SumR(D_{np})$; $T_p = SumV(D_{np})$; $N=Sum(D_{np})$

[1a] Mx

$= [(9X_1 + 14X_1 + 18X_1)$
$+ (17X_2 + 7X_2 + 11X_2)$
$+ (15X_3 + 13X_3 + 14X_3)$
$+ (5X_4 + 18X_4 + 8X_4)] / 149$
$= (41X_1 + 35X_2 + 42X_3 + 31X_4) / 149$
$= S_n^T X_n / N = 0$

[1b] My

$= [(9Y_1 + 17Y_1 + 15Y_1 + 5Y_1)$
$+ (14Y_2 + 7Y_2 + 13Y_2 + 18Y_2)$
$+ (18Y_3 + 11Y_3 + 14Y_3 + 8Y_3)] / 149$
$= (46Y_1 + 52Y_2 + 51Y_3) / 149$
$= T_p^T Y_p / N = 0$

行和ベクトル S_n と個体ベクトル X_n の積、および列和ベクトル Tp と変数ベクトル Y_p との積の分散(Vx, Vy)をそれぞれ 1 とします。そのために、次のように行和 S_n と列和 T_p を対角線上に並べた「対角行列」をそれぞれ S_{nn}. T_{pp} として用意します。

[2] $S_{nn} = diag(S_n)$; $T_{pp} = diag(T_p)$ diag:対角行列

S_{nn}	X_1	X_2	X_3	X_4
X_1	41			
X_2		35		
X_3			42	
X_4				31

T_{pp}	Y_1	Y_2	Y_3
Y_1	46		
Y_2		52	
Y_3			51

[2a] $Vx = [41(X_1 - Mx)^2 + 35(X_2 - Mx)^2 + 42(X_3 - Mx)^2 + 31(X_4 - Mx)^2 / 149$
$= (41X_1^2 + 35X_2^2 + 42X_3^2 + 31X_4^2) / 149$ ←[1a] $Mx = 0$

$= X_n^T S_{nn} X_n / N = 1$

[2b] $Vy = [46(Y_1 - My)^2 + 52(Y_2 - My)^2 + 51(Y_3 - My)^2] / 149$
$= (46Y_1^2 + 52Y_2^2 + 51Y_3^2) / 149$ ←[1b] $My = 0$
$= Y_p^T T_{pp} Y_p / N = 1$

入力データ D_{np} を散布図と見なすと、その X 軸：X_n と Y 軸：Y_p の間の相関係数(R)は,

[3] $R = [9(X_1-Mx)(Y_1-My)$
$+ 14(X_1-Mx)(Y_2-My)$
$+ 18(X_1-Mx)(Y_3-My)$
$+ 17(X_2-Mx)(Y_1-My)$
$+\ldots$
$+ 8(X_4-Mx)(Y_3-My)] / 149$

$= (9X_1Y_1 + 14X_1Y_2 +\ldots + 8X_4Y_3) / 149$ ←$Mx = My = 0$
$= X_n^T D_{np} Y_p / N$

この相関係数 R が最大になるときの X_n, Y_p を求めることが対応分析の目的です。R を最大化するためには、それぞれの分散 $Vx = Vy = 1$ という条件を加えた次の式 Q を X_n と Y_p で微分し($Df(Q, X_n)$, $Df(Q,Y_p)$)、その結果をゼロベクトル(O_n, O_p)とします。Lx, Ly はラグランジュ乗数です[21]。

$Q = R - Lx (Vx - 1) - Ly (Vy - 1)$
$= (X_n^T D_{np} Y_p) / N - Lx [(X_n^T S_{nn} X_n) / N - 1] - Ly [(Y_p^T T_{pp} Y_p) / N - 1]$

[4a] $Df(Q, X_n) = D_{np} Y_p / N - 2 Lx S_{nn} X_n / N = O_n$ (ゼロ)

[4b] $Df(Q, Y_p) = D_{np}^T X_n / N - 2 Ly T_{pp} Y_p / N = O_p$ (ゼロ)

[5a] $D_{np} Y_p / N = 2 Lx S_{nn} X_n / N$ ←[4a]の第 2 項を右辺に移項
$X_n^T D_{np} Y_p / N = 2 Lx X_n^T S_{nn} X_n / N$ ←両辺に X_n^T を左積
$R = 2 Lx X_n^T S_{nn} X_n / N$ ←[3] $R = X_n^T D_{np} Y_p / N$
$R = 2 Lx$ ←[2a] $X_n^T S_{nn} X_n / N = 1$

[5b] $D_{np}^T X_n / N = 2 Ly T_{pp} Y_p / N$ ←[4b]の第 2 項を右辺に移項
$X_n^T D_{np} / N = 2 Ly Y_p^T T_{pp} / N$ ←行列移動；T_{pp} 対角行列
$X_n^T D_{np} Y_p / N = 2 Ly Y_p^T T_{pp} Y_p / N$ ←両辺に Y_p を右積
$R = 2 Ly$ ←[3] $R = X_n^T D_{np} Y_p / N$; [2b] $Y_p^T T_{pp} Y_p / N = 1$

[21] よって：
$Df(Q, Lx) = (X_n^T S_{nn} X_n) / N - 1 = 0$ ← [2a]
$Df(Q, Ly) = [(Y_p^T T_{pp} Y_p) / N - 1 = 0$ ← [2b]

[5a], [5b]から

[6] $R = 2\,Lx = 2\,Ly$

[7a] $D_{np}\,Y_p / N = 2\,Lx\,S_{nn}\,X_n / N$ ← [5a.1]

$D_{np}\,Y_p = 2\,Lx\,S_{nn}\,X_n$ ← 両辺に N を掛ける

$D_{np}\,Y_p = R\,S_{nn}\,X_n$ ← [5a.4] R = 2 Lx

$R\,S_{nn}\,X_n = D_{np}\,Y_p$ ←両辺交換

$S_{nn}\,X_n = D_{np}\,Y_p / R$ ←スカラーR移動

$S_{nn}^{-1}\,S_{nn}\,X_n = S_{nn}^{-1}\,D_{np}\,Y_p / R$ ←両辺に S_{nn}^{-1} を左積

$X_n = S_{nn}^{-1}\,D_{np}\,Y_p / R$ ←$S_{nn}^{-1}\,S_{nn} = I_{nn}$

[7b] $D_{np}^T\,X_n / N = 2\,Ly\,T_{pp}\,Y_p / N$ ← [5b]

$D_{np}^T\,X_n = R\,T_{pp}\,Y_p$ ← [6] R= 2 Ly

[8] $D_{np}^T\,X_n = R\,T_{pp}\,Y_p$ ← [7b]

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,Y_p / R = R\,T_{pp}\,Y_p$ ← [7a] $X_n = S_{nn}^{-1}\,D_{np}\,Y_p / R$

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,Y_p = R^2\,T_{pp}\,Y_p$ ← スカラーR 移動

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,T_{pp}^{1/2}\,Y_p = R^2\,(T_{pp})^{1/2}\,(T_{pp})^{1/2}\,Y_p$

←$(T_{pp}^{1/2})^{-1}\,T_{pp}^{1/2} = I_{pp}$; $(T_{pp})^{1/2}\,(T_{pp})^{1/2} = T_{pp}$ (後述)

ここで

[9] $(T_{pp})^{1/2}\,Y_p = A_p$

とすると

$(T_{pp}^{1/2})^{-1}\,(T_{pp})^{1/2}\,Y_p = (T_{pp}^{1/2})^{-1}\,A_p$ ←[9]の両辺に$(T_{pp}^{1/2})^{-1}$を左積

$I_{pp}\,Y_p = (T_{pp}^{1/2})^{-1}\,A_p$ ← $(T_{pp}^{1/2})^{-1}\,T_{pp}^{1/2} = I_{pp}$ (後述)

$Y_p = (T_{pp}^{1/2})^{-1}\,A_p$ ← I_{pp} は単位行列

[8]の第 3 行は

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,Y_p = R^2\,T_{pp}\,Y_p$ ← [8].3

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = R^2\,T_{pp}\,Y_p$ ← [9] $(T_{pp})^{1/2}\,Y_p = A_p$

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = R^2\,(T_{pp})^{1/2}\,(T_{pp})^{1/2}\,Y_p$

← $(T_{pp})^{1/2}\,(T_{pp})^{1/2} = T_{pp}$ (後述)

$D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = R^2\,T_{pp}^{1/2}\,A_p$ ← [9] $(T_{pp})^{1/2}\,Y_p = A_p$

$(T_{pp}^{1/2})^{-1}\,D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = (T_{pp}^{1/2})^{-1}\,T_{pp}^{1/2}\,R^2\,A_p$

←両辺に$(T_{pp}^{1/2})^{-1}$を左積

$(T_{pp}^{1/2})^{-1}\,D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = I_{pp}\,R^2\,A_p$ ←$(T_{pp}^{1/2})^{-1}\,T_{pp}^{1/2} = I_{pp}$

$(T_{pp}^{1/2})^{-1}\,D_{np}^T\,S_{nn}^{-1}\,D_{np}\,(T_{pp}^{1/2})^{-1}\,A_p = R^2\,A_p$ ←I_{pp} は単位行列

ここで

$$(T_{pp}^{1/2})^{-1}\, D_{np}^{T}\, S_{nn}^{-1}\, D_{np}\, (T_{pp}^{1/2})^{-1} = A_{pp}$$

とすれば

$$(T_{pp}^{1/2})^{-1}\, D_{np}^{T}\, S_{nn}^{-1}\, D_{np}\, (T_{pp}^{1/2})^{-1}\, A_p = R^2\, A_p \qquad \leftarrow \text{先の式}$$
$$A_{pp}\, A_p = R^2\, A_p \quad \leftarrow \quad (T_{pp}^{1/2})^{-1}\, D_{np}^{T}\, S_{nn}^{-1}\, D_{np}\, (T_{pp}^{1/2})^{-1} = A_{pp}$$

という固有値問題($A_{pp}\, A_p = S\, A_p$)になります。プログラムはこの固有値問題の既知の A_{pp} から未知の R^2 と A_p を同時に求めます。Y_p は[9]の最終式から求めます。

$$Y_p = (T_{pp}^{1/2})^{-1}\, A_p \qquad \leftarrow [9]$$

ここで、ベクトル Y_p は横和ベクトル S_n との積和の平均が 0、分散が 1 となるような小さな値です（←[1b], [2b])。そこでデータの規模に合わせるために、全体にデータの総和の根 $Sum(D_{np})^{1/2}$ を掛けます。また、さらに相関係数 R_p を掛けると相関係数の大きさを反映した座標になります[22]。

$$Yp' = Yp * Sum(D_{np})^{1/2} * R_p$$

X_n は[7a]の最終式から求めます。

$$X_n = S_{nn}^{-1}\, D_{np}\, Y_p / R$$

＊数理とプログラムは奥村(1986), 高橋(2005), 三野(2005)を参照しました。

●行列の 1/2 乗と -1/2 乗

非負正方行列 A_{pp} について、$X_{pp}\, X_{pp} = X_{pp}^2 = A_{pp}$ となる X_{pp} は A_{pp} の 1/2 乗 $A_{pp}^{1/2}$ と定義されます: $(A_{pp}^{1/2})^2 = A_{pp}$ 。

$$X_{pp}^2 = X_{pp}\, X_{pp} = A_{pp}; \quad X_{pp} = A_{pp}^{1/2}$$

また、非負正方行列 A_{pp} に逆行列 A_{pp}^{-1} が存在するとき、$Y_{pp}\, Y_{pp} = A_{pp}^{-1}$ となる Y_{pp} は A_{pp} の-1/2 乗 $A_{pp}^{-1/2}$ と定義されます: $(A_{pp}^{-1/2})^2 = A_{pp}^{-1}$ 。

$$Y_{pp}^2 = Y_{pp}\, Y_{pp} = A_{pp}^{-1}; \quad Y_{pp} = A_{pp}^{-1/2}$$

●対角行列の逆行列

T_{pp} が対角行列のとき、その逆行列 T_{pp}^{-1} は対角行列であり、その成分は T_{pp} の逆数になります。

$$T_{pp}^{1/2}\, T_{pp}^{1/2} = T_{pp}$$

[22] 高橋(2005: 127-129).

T_{pp}	1	2	3
1	A		
2		B	
3			C

T_{pp}^{-1}	1	2	3
1	1/A		
2		1/B	
3			1/C

T_{pp} が対角行列のとき、$T_{pp}^{1/2}$ は対角行列であり、その対角成分は T_{pp} の対角成分の平方根になります。

$$(T_{pp}^{1/2})^{-1}\ T_{pp}^{1/2} = T_{pp}$$

T_{pp}	1	2	3
1	A		
2		B	
3			C

$T_{pp}^{1/2}$	1	2	3
1	√A		
2		√B	
3			√C

●個体と変数の対応

下左表(D_{np})はデータ行列、下右表は個体の負荷値 X_n です。

Dnp	Y_1: English	Y_2: Physics	Y_3: Latin
X_1: Ana	9	14	18
X_2: Juan	17	7	11
X_3: Mary	15	13	14
X_4: Ken	5	18	8

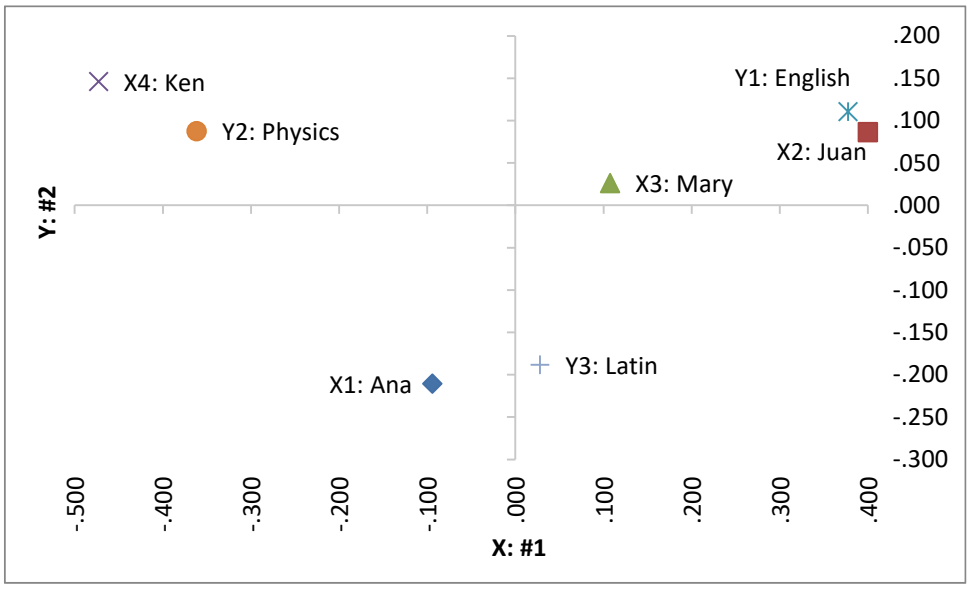
CA.d. (Xn)	#1	#2
x1: Ana	-.094	-.211
x2: Juan	.400	.086
x3: Mary	.108	.026
x4: Ken	-.473	.146

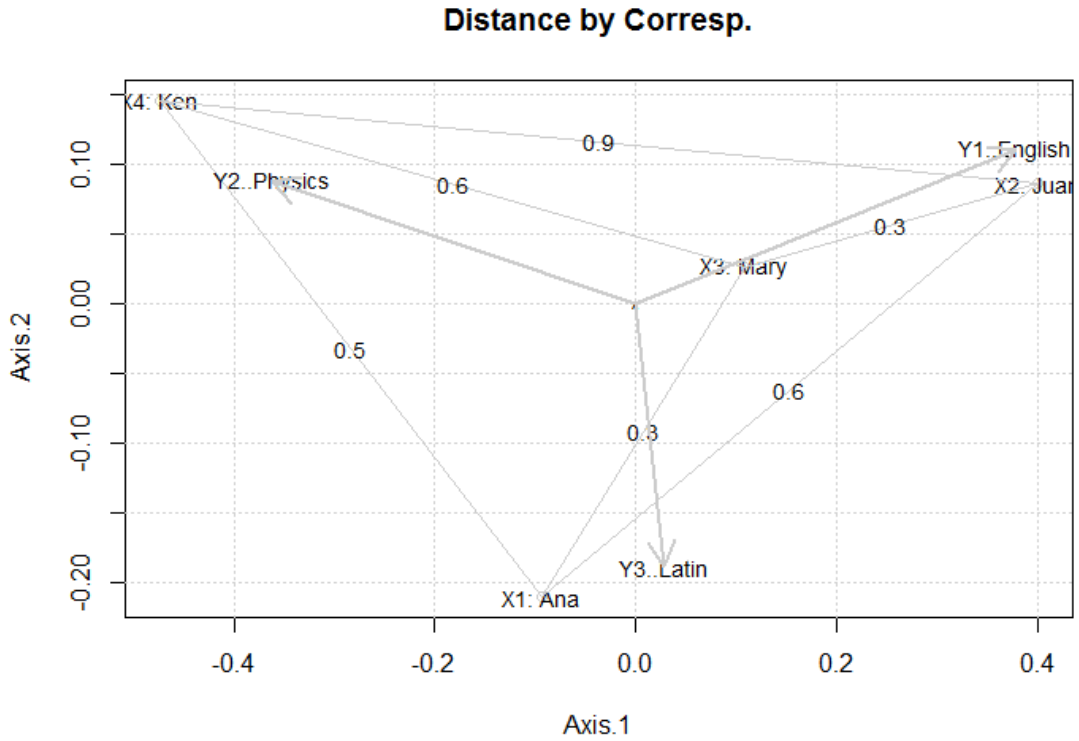
下左表は相関係数(Correl)を示し、下右表は変数の負荷値(Y_p)を示します。

Corresp.	1	2
Correl.	.300	.136

CA.v. (Yp)	#1	#2
y1: English	.377	.110
y2: Physics	-.361	.087
y3: Latin	.028	-.189

Xn と Yp を連続させた項目名付散布図(Item scatter)です。





これらの図を見ると、Ken と Physics, Juan と English、Ana と Latin がそれぞれ近い関係になることがわかります。Mary が全体の中で中立ですが、やや English に近づいています。なお、作図のプログラムが可能な R で出力した 2 番目の図では、科目(English, Physics, Latin)をバイプロットにしました。

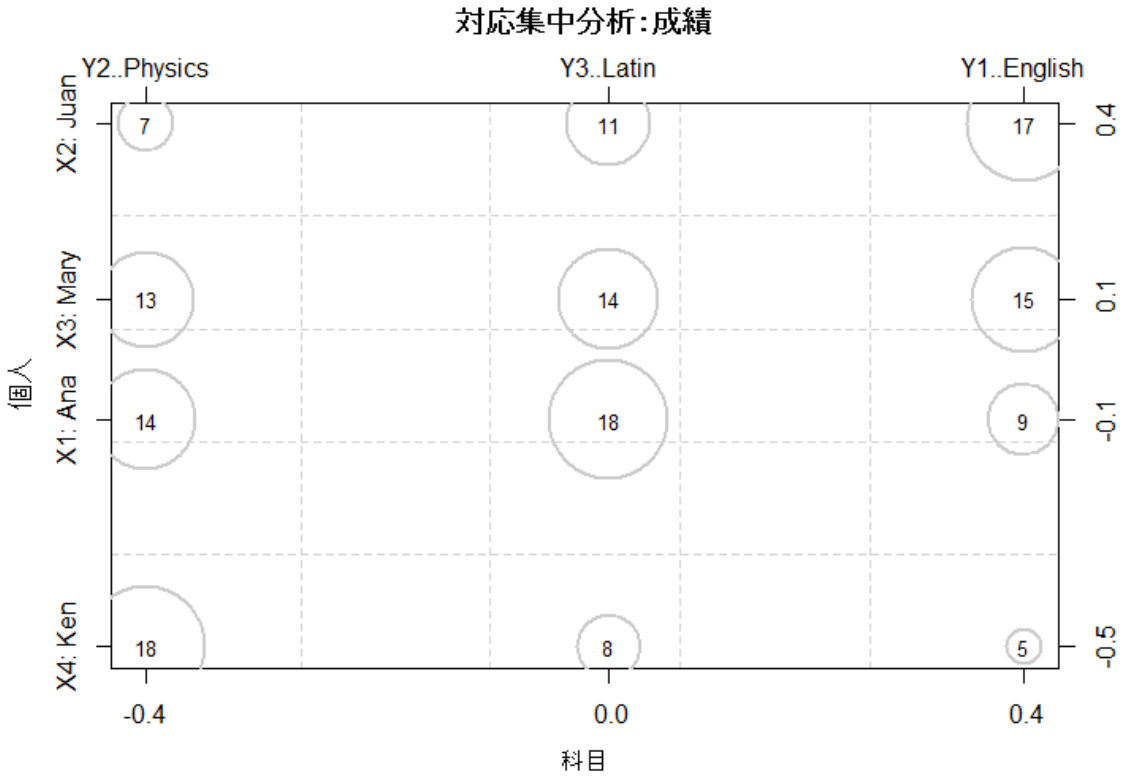
●対応集中分析

変数と個体の係数得点を昇順でソートし得点を並び替えると次のような集中化した得点になります。対応分析による集中化は $X_{1:4}$ と $Y_{1:3}$ の間の相

関係数が最も高い得点の分布を示します。

Cor.A	Y2: Physics	Y3: Latin	Y1: English
X4: Ken	18	8	5
X1: Ana	14	18	9
X3: Mary	13	14	15
X2: Juan	7	11	17

次は、両軸の負荷値と点数の大きさを考慮して作図した分布図です。



●片側対応分析

前述の対応分析では個体と変量に与える未知のベクトルを求めましたが、ここでは、どちらかを既知のベクトルとし、残るほうを未知のベクトルとします。この方法を「片側対応分析」(Unilateral correspondence analysis)とよびます。既知のベクトルとして、この成分の順番を固定するために、連続数 1, 2, …, N（または P)の標準得点を与えます。これを外的基準として固定し、未知の行、または未知の列のベクトルを求めます。そのとき、ベクトルの平均を 0 とし、分散を 1 として標準化します。片側対応分析の目的は（前述の両側）対応分析(Bilateral correspondence analysis)と同様に、データ行列の分布の相関係数を最大化することです。

はじめに、変数の横ベクトル Y_p を連続数の標準得点で固定し、個体のベ

クトル X_n を未知として、これを求めます。

（両側）対応分析の説明で使ったデータを下に再掲します。

D_{np}	Y_1: English	Y_2: Latin	Y_3: Science	S_n
X_1: Ana	9	14	18	41
X_2: Juan	17	7	11	35
X_3: Mary	15	13	14	42
X_4: Ken	5	18	8	31
T_p	46	52	51	N=149

この「片側（個体）対応分析」の目的は、変数ベクトル$(Y_1, Y_2, Y_3) = Y_p$ を既知として、相関係数 R が最大になるときの未知の個体$(X_1, X_2, X_3, X_4) = X_n$ のベクトルを求めることです。

個体ベクトル X_n と変数ベクトル Y_p の平均(Mx, My)をそれぞれ 0 とします。

$S_n = SumR(D_{np})$; $T_p = SumV(D_{np})$; $N = Sum(D_{np})$

$S_{nn} = dg(S_n)$; $T_{pp} = dg(T_{1p})$ [dg: 対角行列]

[1a] $Mx = (41X_1 + 35X_2 + 42X_3 + 31X_4) / 149 = S_n^T X_n / N = 0$

個体(X_n)の分散(Vx)を 1 とします。

[2] $Vx = [(41X_1 - Mx)^2 + (35X_2 - Mx)^2 + (42X_3 - Mx)^2 + (31X_4 - Mx)^2 / 149$

$= (41X_1^2 + 35X_2^2 + 42X_3^2 + 31X_4^2) / 149$ ←[1a] $Mx = 0$

$= X_n^T S_{nn} X_n / N = 1$

D_{np} を散布図と見なすと、その X 軸：X_n と Y 軸：Y_p の間の相関係数(R)は,

[3] $R = [9(X_1-Mx)(Y_1-My)$

$+ 14(X_1-Mx)(Y_2-My)$

$+ 18(X_1-Mx)(Y_3-My)$

$+ 17(X_2-Mx)(Y_1-My)$

$+\ldots$

$+ 8(X_4-Mx)(Y_3-My)] / 149$

$= (9X_1Y_1 + 14X_1Y_2 + \ldots + 8X_4Y_3) / 149$ ←$Mx = My = 0$

$= X_n^T D_{np} Y_p / N$

この R を最大化するためには、分散 Vx = 1 という条件を加えた次の式の S を X_n で微分し(Df(Q, X_n))、その結果をゼロベクトル(O_n)とします。L はラグランジュ乗数です。

$$
\begin{aligned}
Q &= R - L\,(Vx - 1)\\
&= (X_n^T\, D_{np}\, Y_p) / N - L\,(Vx - 1)\\
&= (X_n^T\, D_{np}\, Y_p) / N - L\,[(X_n^T\, S_{nn}\, X_n) / N - 1]
\end{aligned}
$$

[4] $Df(Q, X_n) = D_{np}\, Y_p / N - 2\, L\, S_{nn}\, X_n / N = O_n$ (ゼロ)

[5] $D_{np}\, Y_p / N = 2\, L\, S_{nn}\, X_n / N$ ←[4]の第 2 項を移項

$X_n^T\, D_{np}\, Y_p / N = 2\, L\, X_n^T\, S_{nn}\, X_n / N$ ←両辺に X_n^T を左積

$R = 2\, L$ ← [2] $X_n^T\, S_{nn}\, X_n / N = 1$; [3] $R = X_n^T\, D_{np}\, Y_p / N$

[6] $D_{np}\, Y_p = R\, S_{nn}\, X_n$

←[5] $D_{np}\, Y_p / N = 2\, L\, S_{nn}\, X_n / N$; [6] $R = 2\, L$

$R\, S_{nn}\, X_n = D_{np}\, Y_p$ ←両辺交換

$S_{nn}\, X_n = D_{np}\, Y_p / R$ ←スカラー R 移動

$S_{nn}^{-1}\, S_{nn}\, X_n = S_{nn}^{-1}\, D_{np}\, Y_p / R$ ←両辺に S_{nn}^{-1} を左積

$X_n = S_{nn}^{-1}\, D_{np}\, Y_p / R$ ←$S_{nn}^{-1}\, S_{nn} = I_{nn}$

先の両側対応分析と同様に、X_n に R（相関係数）を掛けます[23]。

$$X_n'' = X_n * R = S_{nn}^{-1}\, D_{np}\, Y_p / R * R = S_{nn}^{-1}\, D_{np}\, Y_p$$

[3]の R を最大化するためには、分散 $Vy = 1$ という条件を加えた次の式 Q を Y_p で微分し($Df(Q,Y_p)$)、その結果をゼロベクトル(O_p)とします。L はラグランジュ乗数です。

$$
\begin{aligned}
Q &= (X_n^T\, D_{np}\, Y_p) / N - L\,[Vy - 1]\\
&= (X_n^T\, D_{np}\, Y_p) / N - L\,[(Y_p^T\, T_{pp}\, Y_p) / N - 1]
\end{aligned}
$$

[4b] $Df(Q, Y_p) = D_{np}^T\, X_n / N - 2\, L\, T_{pp}\, Y_p / N = O_p$ (ゼロ)

[5b] $D_{np}^T\, X_n / N = 2\, L\, T_{pp}\, Y_p / N$ ←[4b]の第 2 項を移項

$X_n^T\, D_{np} / N = 2\, L\, Y_p^T\, T_{pp} / N$ ←$A^T\, B = B^T\, A$; T_{pp} 対角行列

$X_n^T\, D_{np}\, Y_p / N = 2\, L\, Y_p^T\, T_{pp}\, Y_p / N$ ←両辺に Y_p を右積

$R = 2\, L$ ←[3] $R = X_n^T\, D_{np}\, Y_p / N$; [2b] $Y_p^T\, T_{pp}\, Y_p / N = 1$

---

[23] 一方、個体のベクトルを固定して、変数のベクトルを求めるときは、[2]以下を次のようにします。

[1b] $My = (46Y_1 + 52Y_2 + 51Y_3) / 149 = T_p^T\, Y_p / N = 0$

[2b] $Vy = [(46Y_1 - My)^2 + (52Y_2 - My)^2 + (51Y_3 - My)^2] / 149$

$= (46Y_1^2 + 52Y_2^2 + 51Y_3^2) / 149$ ←[1b] $My = 0$

$= Y_p^T\, T_{pp}\, Y_p / N = 1$

[3] $R = X_n^T\, D_{np}\, Y_p / N$

[6b] $D_{np}^T X_n = R T_{pp} Y_p$ ←[5b] $D_{np}^T X_n / N = 2 L T_{pp} Y_p / N$; [6] $R = 2 L$

$R T_{pp} Y_p = D_{np}^T X_n$ ←両辺交換

$T_{pp} Y_p = D_{np}^T X_n / R$ ←スカラーR移動

$T_{pp}^{-1} T_{pp} Y_p = T_{pp}^{-1} D_{np}^T X_n / R$ ←両辺に T_{pp}^{-1} を左積

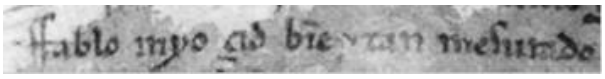
$Y_p = T_{pp}^{-1} D_{np}^T X_n / R$ ←$T_{pp}^{-1} T_{pp} = I_{pp}$

ここでも先と同様にR（相関係数）を掛けて除去します。

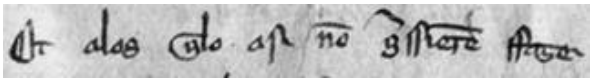
$Y_p" = Y_p * R = T_{pp}^{-1} D_{np}^T X_n / R * R = T_{pp}^{-1} D_{np}^T X_n$

■中世・近代スペイン語公証文書書体の年代推移

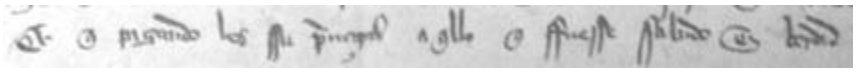
次は古スペイン語の手稿本や公証文書で用いられた書体の例です。



(a) Cid, 1207, Letra gótica libraria {7} *ffablo myo çid bien e tan meſurado*



(b) CODEA:0287, Madrid, 1340, Letra de albalaes
{14} *Et a los que lo aſi non quiſſieren ffazer*



(c) CODEA:3931, Madrid 1386, Letra gótica cortesama
{31} *E que pagando los ſſu prinçipal aquello que ffueſſe ſabido en verdad*

上の(a)は「手稿本ゴチック体」、(b)は「勅書体」、(c)は「宮廷ゴチック体」の写真です。他にも十数種の書体がありました。次の表はCODEAコーパスに収められた文書数の年代推移を示します[24]。

[24] 資料はCODEA (Corpus de documentos Españoles Anteriores a 1800)，分析はLYNEALを使いました。
http://corpuscodea.es/ [2019/5/24]
https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/codea.htm [2019/5/24]

Letra	Carolina	Cort	De al.	De priv	G. c.	G. c. al	G. c. p.	G. lib	G. r.	Gótica	H. c	H. r.	Prec	Proc.
1075	1													
1100	2													
1125	1													
1150	7							1		1				
1175	2							5	1					
1200	3							14	5	10				
1225	3			1				48	5	12				
1250	1		2	4	14			45	4	8				
1275			14	5	32	22		13	3	1				
1300			4		46	1		8						
1325			1		29			3		1				
1350			1		25			3					5	
1375		6			7		3						6	
1400		4			2			1		1		1	20	
1425		12							1	1			3	
1450		30							2					1
1475		9							3	1		1		
1500		20							4			1	3	
1525		8							1		1	1		2
1550		1									3	11		
1575		1									2	3		2
1600											4	3		
1625											8			1
1650											4			
1675											7			

次に、縦軸の年代を固定し、横軸だけを対象とした片側対応分析をすると、次のような結果になりました。次の表で、左上から右下にかけて推移する段階的なグラデーションが観察されます。ここで、(1)から(14)の書体が年代に沿って変化していくことがわかります。

Letra	(1) Carolina	(2) G. lib	(3) De priv	(4) Gótica	(5) G. c. al	(6) De al.	(7) G. c.	(8) G. r.	(9) G. c. p.	(10) Prec	(11) Cort	(12) H. r.	(13) Proc.	(14) H. c
1075	1													
1100	2													
1125	1													
1150	7	1		1										
1175	2	5						1						
1200	3	14		10				5						
1225	3	48	1	12				5						
1250	1	45	4	8		2	14	4						
1275		13	5	1	22	14	32	3						
1300		8			1	4	46							
1325		3		1		1	29							
1350		3				1	25			5				
1375							7		3	6	6			
1400		1		1			2			20	4	1		
1425				1				1		3	12			
1450								2			30		1	
1475				1				3			9	1		
1500								4		3	20	1		
1525								1			8	1	2	1
1550											1	11		3
1575											1	3	2	2
1600												3		4
1625													1	8
1650														4
1675														7

Letra: (1) Carolina, (2) Gótica libraria, (3) De privilegios, (4) Gótica, (5) Gótica cursiva [albalaes], (6) De albalaes, (7) Gótica cursiva, (8) Gótica redonda, (9) Gótica cursiva [precortesana], (10) Precortesana, (11) Cortesana, (12) Humanística redonda, (13) Procesal, (14) Humanística cursiva

＊資料：CODEA +2015: «Corpus de Documentos Españoles Anteriores a 1700» (Pedro Sánchez Prieto Borja, GITHE: (Grupo de Investigación de Textos para la Historia del Español, Universidad de Alcalá). Contiene 1502 textos provenientes de distintas regiones de España de l siglo XI al XVII.

■アンダルシア方言の音声特徴の地域差

次表はスペイン、アンダルシア地方の 8 県で得られた資料中の、語末子音の脱落による先行母音の開母音化の相対頻度を示します。

R / N * 100	H	SE	CA	MA	CO	J	GR	AL
1533B:miel:el>e+	17	10	9	15	20	29	46	30
1533C:miel:el>e:	11	6	4	16	12	11	16	3
1615A:caracol:-ól>ó+(:)	2	3	3	5	15	14	19	11
1615B:caracol:-ól>ó(:)	18	27	15	16	3	1	6	2
1616A:árbol:-ol>o+	2	1			6	8	6	6
1616B:árbol:-ol>o	23	30	17	26	18	11	23	11
1618A:sol:-ól>ó+(:)	7	9	3	13	13	12	19	11
1618B:sol:-ól>ó(:)	15	21	15	13	1	1	6	1
1623A:beber:-ér>é+l	2			1	10	11	19	20
1623B:beber:-ér>é+	4	7	3	6	13	17	15	8
1623C:beber:-ér>é	19	24	15	19	2		4	
1626A:tos:-ós>ó+h	6	2	2	4	7	13	17	9
1626C:tos:o++				2	7	10	18	12
1626C:tos:-ós>ó+	7	7	5	13	18	17	27	19
1626D:tos:-ós>ó	10	22	11	9	2	1	2	
1627A:nuez:-éθ>é+h	5	2	1		2	3	8	3
1627B:nuez:-éθ>é+	7	13	5	17	20	25	39	26
1627C:nuez:e++				5	14	18	26	18
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1629A:voz:-óθ>óh	5	3		1	1	1	5	3
1629B:voz:-óθ>ó+	3	5	3	12	22	30	44	30
1629C:voz:-óθ>ó	18	23	14	13	2	1	2	1
1689A:niños:-os>-o+	1	2		4	22	31	44	30
1689B:niños:-os>oh[os)	4	1			2	3	8	8
1690A:pared:-éd>é+	6	8		10	17	19	24	11
1693A:redes:redes>rede	3	2	1	1	4	12	8	18
1693B:redes:redes>re+	14	6	4	12	3	6	16	6
1693C:redes:redes>reh	1		2		1	4	7	
1694A:clavel:-él>-él	3	2	1	3	6	5	11	15
1694B:clavel:-él>é+,		6	3	15	20	24	40	29
1694C:clavel:-él>ér						1	5	1
1695A:claveles:e-es>-e-e+		2		4	2	2	4	3
1695B:claveles:e-es>-e+-e+		1		7	18	24	33	21
1695C:claveles:-e-es>-e-e:		1		3	1	2	1	1
1695D:claveles:e-es>-e-eh	3	1			5	4	9	5

この分布表を対応分析(両側)にかけると次のような結果になりました。

Cor.A	AL	J	GR	CO	MA	H	SE	CA
1694C:clavel:-él>ér	1	1	5					
1626C:tos:o++	12	10	18	7	2			
1623A:beber:-ér>é+l	20	11	19	10	1	2		
1627C:nuez:e++	18	18	26	14	5			
1689A:niños:-os>-o+	30	31	44	22	4	1	2	
1695B:claveles:e-es>-e+-e+	21	24	33	18	7		1	
1616A:árbol:-ol>o+	6	8	6	6		2	1	
1693A:redes:redes>rede	18	12	8	4	1	3	2	1
1629B:voz:-óθ>ó+	30	30	44	22	12	3	5	3
1694B:clavel:-él>é+,	29	24	40	20	15		6	3
1695D:claveles:e-es>-e-eh	5	4	9	5		3	1	
1694A:clavel:-él>-él	15	5	11	6	3	3	2	1
1615A:caracol:-ól>ó+(:)	11	14	19	15	5	2	3	3
1689B:niños:-os>oh[os)	8	3	8	2		4	1	
1626A:tos:-ós>ó+h	9	13	17	7	4	6	2	2
1690A:pared:-éd>é+	11	19	24	17	10	6	8	
1693C:redes:redes>reh		4	7	1		1		2
1627B:nuez:-éθ>é+	26	25	39	20	17	7	13	5
1626C:tos:-ós>ó+	19	17	27	18	13	7	7	5
1623B:beber:-ér>é+	8	17	15	13	6	4	7	3
1533B:miel:el>e+	30	29	46	20	15	17	10	9
1695A:claveles:e-es>-e-e+	3	2	4	2	4		2	
1618A:sol:-ól>ó+(:)	11	12	19	13	13	7	9	3
1627A:nuez:-éθ>é+h	3	3	8	2		5	2	1
1695C:claveles:-e-es>-e-e:	1	2	1	1	3		1	
1533C:miel:el>e:	3	11	16	12	16	11	6	4
1629A:voz:-óθ>óh	3	1	5	1	1	5	3	
1693B:redes:redes>re+	6	6	16	3	12	14	6	4
1616B:árbol:-ol>o	11	11	23	18	26	23	30	17
1615B:caracol:-ól>ó(:)	2	1	6	3	16	18	27	15
1618B:sol:-ól>ó(:)	1	1	6	1	13	15	21	15
1623C:beber:-ér>é			4	2	19	19	24	15
1627C:nuez:-éθ>é		1	1	3	9	12	16	12
1629C:voz:-óθ>ó	1	1	2	2	13	18	23	14
1626D:tos:-ós>ó		1	2	2	9	10	22	11

左上の区画の頻度は西地方(MA, H, SE, CA)で開母音化が少ないことを示し、右下の区画には、逆に東地方(AL, J, GR, CO)で開母音化が多くなっ

ていることを示しています。これは一般的な傾向であって、上右や下左の区画にも多くの数値があるので例外が多いことがわかります。

＊資料：『アンダルシア言語民俗地図』(Manuel Alvar y Antonio Llorente: *Atlas lingüístico y etnográfico de Andalucía*, 1973)

■集中領域：ラテンアメリカの「農夫」を示す語

次の表は、個体と属性の関係 Cahuzac (1980)のラテンアメリカの「農夫」を示す語の国別分布を対応分析した結果です[25]。

F.A.	PA	UR	AR	BO	CH	PE	EC	CU	MX	RD	PN	PR	CO	C5	VE
03 camilucho	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
04 campero	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
12 comparsa	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
16 changador	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
18 chuncano	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
20 estanciero	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
41 piona	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
21 gaucho	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
39 partidario	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
28 invernador	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0
46 viñatero	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0
43 rondín	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
30 lampero	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
47 yanacón	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
09 campusano	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
26 huertero	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0
24 guaso	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0

上の図を見ると、語 3, 4, 12, 16, 18, 20, 41, 21 が PA, UR, AR に集中していることがわかります。このような部分を「集中領域」と呼びます。集中化した分布図では、左上から右下にかけての対角線に近接する領域で、多くの集中領域が見つかります。そこで、プログラムによって個体（語形）と属性（国名）のそれぞれの組み合わせを始点として、それより右下にあるポイントを次々にサーチし、反応点の平均値と有意度の積が最大になるポイントを定め、平均値と有意度の積を降順にソートして出力させます。

[25] Cahuzac, Philippe. (1980) "La División del español de América en zonas dialectales: Solución etnolingüística o semántico-dialectal." *Lingüística Española Actual*, 10, pp. 385-461.

C1	A1	C2	A2	X1	Y1	X2	Y2	Suma	Total	Media	ProEx
03 camilucho	PA	20 estanciero	AR	1	1	3	6	18	18	1.000	.847
04 campero	PA	41 piona	AR	1	2	3	7	18	18	1.000	.847
12 comparsa	PA	21 gaucho	AR	1	3	3	8	18	18	1.000	.847
16 changador	PA	21 gaucho	AR	1	4	3	8	15	15	1.000	.819
03 camilucho	UR	20 estanciero	AR	2	1	3	6	12	12	1.000	.779
04 campero	UR	41 piona	AR	2	2	3	7	12	12	1.000	.779
12 comparsa	UR	21 gaucho	AR	2	3	3	8	12	12	1.000	.779
16 changador	UR	39 partidario	AR	2	4	3	9	12	12	1.000	.779
18 chuncano	PA	21 gaucho	AR	1	5	3	8	12	12	1.000	.779
18 chuncano	UR	28 invernador	AR	2	5	3	10	12	12	1.000	.779
20 estanciero	UR	46 viñatero	AR	2	6	3	11	12	12	1.000	.779
20 estanciero	PA	46 viñatero	AR	1	6	3	11	17	18	.944	.762
41 piona	UR	46 viñatero	AR	2	7	3	11	10	10	1.000	.741
24 guaso	PE	36 montubio	EC	6	17	7	21	10	10	1.000	.741
41 piona	PA	46 viñatero	AR	1	7	3	11	14	15	.933	.721
21 gaucho	UR	46 viñatero	AR	2	8	3	11	8	8	1.000	.688

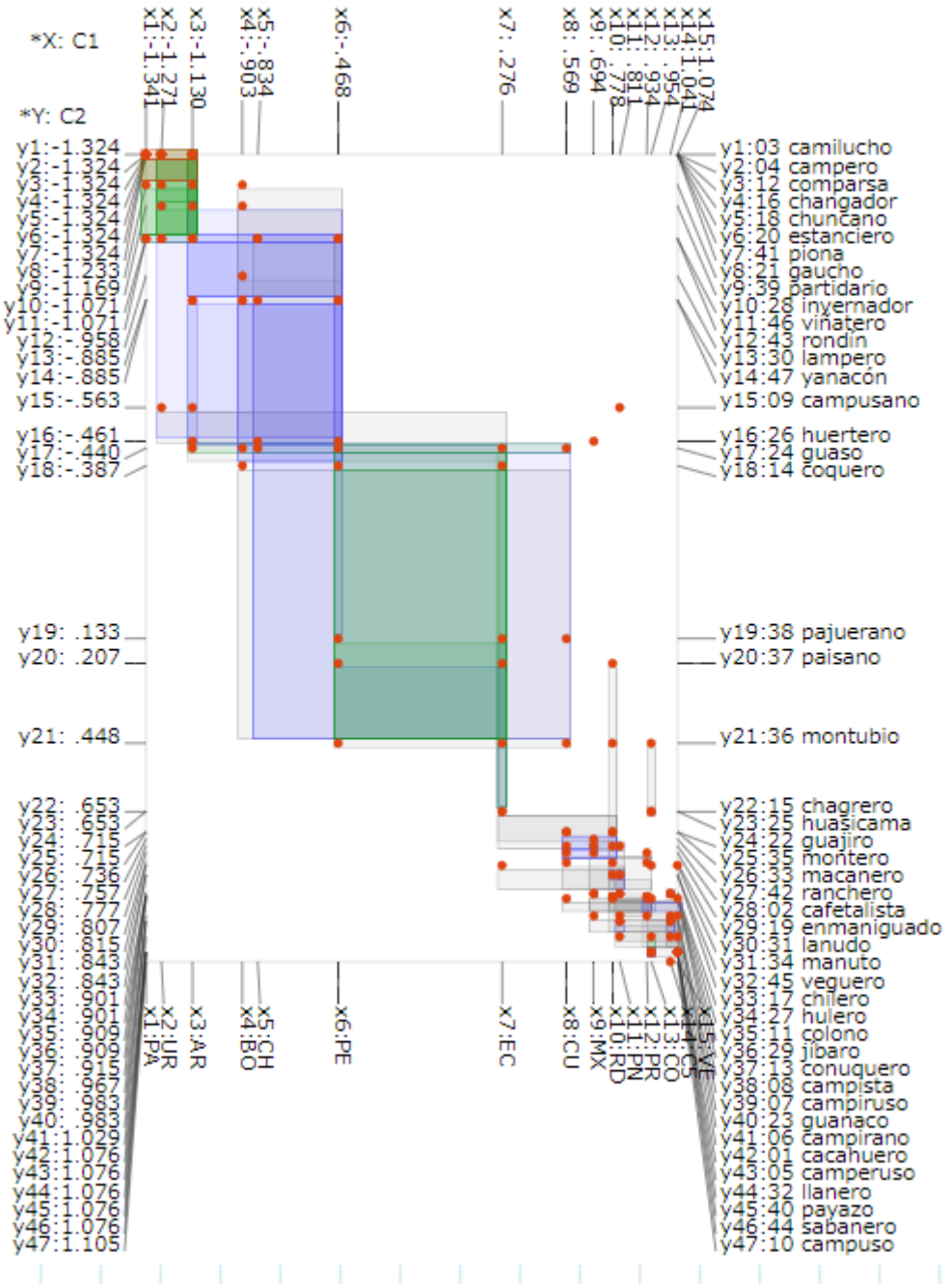
たとえば、最初の 3 camilucho:PA は座標(1, 1)に位置するので、ここを始点とします[26]。それが座標(3, 6)までサーチすると[27]、(1,1)と(3,6)に囲まれた部分の総頻度数 Sum=18、個数の総和 Total = 18、そして次の期待確率(ProEx)を求めます（→「確率」）。

ProEx = BinE(Sum, Total, 0.95)

ソートするときの第 1 キーは ProEx, 第 2 キーは Mean, 第 3 キーは個数 Count としました。3 camilucho を始点とする集中領域の期待確率(ProEx)は.847 です。

[26] X 軸は左から右に進みますが、Y 軸は上から下に降りて進みます。

[27] 集中化したデータを見ると(3, 8)まで領域を拡大できることがわかります。しかし、ここでは最大領域を探すことが目的ではなく、5 x 5 を最大のユニットとして、その範囲内の期待確率が最大になる領域をサーチします。(3, 8)までの領域は(1, 1)ではなく、(1, 2), (1, 3)からの領域でも最大領域になるので、集中領域の結果は同じです。



■スペイン語文字エニェ ñ の歴史

次は同じ資料から得られた古文書中のエニェ ñ の分布を示すものです。

ñ	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
<n>n	6	14	1		2							
gn						10	3					
n	6	36	33	19	33	31	52	19	8	10	4	4
n<n>	8	416	329	236	367	520	522	40	2			
nn	12	80	4	1		1						
pn<n>					15	9	9					

ñ					1	9	58	90	29	67	45

片側対応分析の方法を適用して、これを年代を固定して、文字形態の順番を変えると次の表が得られます。

ñ	1200	1250	1300	1350	1400	1450	1500	**1550**	1600	1650	1700	1750
nn	12	80	4	1		1						
<n>n	6	14	1		2							
n<n>	8	416	329	236	367	520	522	**40**	2			
n	6	36	33	19	33	31	52	**19**	8	10	4	4
pn<n>					15	9	9					
gn						10	3					
ñ						1	9	**58**	90	29	67	45

この表から全体の頻度の推移がわかり、その推移の中でとくに 1550 でエニェ ñ が他を凌いで優勢になったことがわかります。

■スペイン語文字アチェ h の歴史

スペイン語の文字 h は発音されません。発音されないのになぜ書かれるようになったのでしょうか？その歴史を探るために同じ資料を分析します。次の 2 つの片側対応分析による頻度分布を見ると、動詞 haber の活用形 he や ha が中世でも h を伴って使われていたことがわかります。

<he>	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
e		2	6	4								
h<e>	1	32	10	7	12	34	20		1			4
he		7	3			2	1	1	4	2		

<ha>	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
a	2	12	18	12	3	3	20	9	32	6	6	1
ha	1	1	11	12	14	40	55	4	12	3		8

ところが次の表を見ると他の活用形では h は中世では使われていません。

<haber>	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
au<er>			2	14	1		2					
auer	9	45	24	15	24	7	4	10	13	7		
aue<r>			1		4	1						
av<er>						33	29					
aver					4	11	22	4	1			

aber			1				1	3	1		1	1
hauer							1	1	8	2	1	3
haber									2			2
haver												2

<había>	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
auie	4	12										
auie<n>		1										
auje		1		1	1							
auya				1								
auja		4	3	7	2	13	3	1				
avja						5	2					
avia						4	8	2	1			
auia	4	4	3					30	2	9		
abia								3	16			
hauia								7	4	10		8
habia								1				5
havia												3

次のような名詞でも同じです。

«hombre»	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
omne	3	79	58	26	44	49	9					
ome		5	6	2	7	4	11	1				36
ombre		2	3	1		3	4				1	14
hombre							5	4	15	1	1	1

これは、he, ha, han などの単音節語が接続詞 e, 前置詞 a と区別される必要があったので、ラテン語形 HABERE にあった h を使用したものと考えられます。その h が後に動詞全体の活用形に適用され、さらに他のラテン語 H に由来する名詞などにも広がったようです。従来は単にラテン語起源だけで説明されることが多かったのですが、それでは、なぜラテン語の文字を復活しなければならなかったのか、その理由がわかりませんでした。

7.7. 因子分析

「因子分析」(Factor analysis)は主成分分析と逆の考え方をする分析法です。主成分分析ではデータ行列の全変数を説明するような軸を探しますが、因子分析では、各成分が互いにできるだけ異なるようなベクトル（因子 factor）を探します。成績データを例にすると、たとえば英語と国語を説明

するような変数（文系因子）が、数学と理科を説明するような変数（理系因子）と明確に異なるようにします。ここでは因子分析の多くの手法の中から芝(1975:90-103)にしたがって Horst の「バリマックス法」(Varimax method)を説明します。

この分析法の目的は未知の因子(A_1, A_2, …, A_p) ができるだけ互いに異なるようにするために、因子ベクトル(A_p)の分散(V)を最大化することです。以下は簡略化して分散の分母(N)を外し、次のように変動を使います(V*)。M は A_p の平均を示し、P は A_p の成分の個数です。

$$
\begin{aligned}
V^* &= \Sigma (a_i - m)^2 && \leftarrow \text{変動の定義 } (i = 1, p)\\
&= \Sigma (a_i^2 - 2\, m\, a_i + m^2) && \leftarrow \text{展開}\\
&= \Sigma a_i^2 - 2\, m\ \Sigma a_i + \Sigma m^2 && \leftarrow \Sigma \text{ を分配}\\
&= \Sigma a_i^2 - 2\, m\ \Sigma a_i + p\, m^2 && \leftarrow \Sigma m^2 = p\, m^2 \ (p \text{ は } a_p \text{ の成分の個数})\\
&= \Sigma a_i^2 - 2\ \Sigma a_i\ \Sigma a_i / p + p\, m^2 && \leftarrow\ m = (\Sigma a_i) / p\\
&= \Sigma a_i^2 - 2\, (\Sigma a_i)^2 / p + p\, m^2 && \leftarrow\ (\Sigma a_i) \text{ をまとめる}\\
&= \Sigma a_i^2 - 2\, (\Sigma a_i)^2 / p + p\, (\Sigma a_i)^2 / p^2 && \leftarrow\ m^2 = [(\Sigma a_i) / p]^2\\
&= \Sigma a_i^2 - 2\, (\Sigma a_i)^2 / p + (\Sigma a_i)^2 / p && \leftarrow\ 3 \text{ 項を約分}\\
&= \Sigma a_i^2 - (\Sigma a_i)^2 / p && \leftarrow 2 \text{ 項と } 3 \text{ 項の加減}
\end{aligned}
$$

これを p^*p の行列で示すと次のようになります(→後述：●単位行列・単位ベクトルの利用)。

$$V^* = A_p^T\, (I_{pp} - I_p\, I_p^T / p)\, A_p$$

ここでΣa_iの計算で負値が相殺されるのを防ぐため、a ではなく a^2 とした「分散」(V**)を求めます。$A_p^{(2)}$はベクトル A_p のすべての成分を 2 乗した成分をもつベクトルを示します。

[1] $$V^{**} = A_p^{(2)\,T}\, (I_{pp} - I_p\, I_p^T / p)\, A_p^{(2)}$$

ここで次の対角行列

$$A_{pp} = \begin{bmatrix} a_1 & \square & \square & \square \\ \square & a_2 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & a_p \end{bmatrix}$$

を導入すると、先の式[1]は

[1b] $$V^{**} = A_p^T\, A_{pp}\, (I_{pp} - I_p\, I_p^T / p)\, A_{pp}\, A_p$$

となります（→後述：●単位行列・単位ベクトルの利用）。

これから求める因子ベクトル A_p はデータ行列の相関行列 R_{pp} に未知のベクトル T_p を右積したものとします。

[2]　　$A_p = R_{pp} T_p$

T_pの長さを 1 と規定します。

[2b]　　$T_p^T T_p = 1$

$T_p^T T_p = 1$ [2b]という条件付きで V**の最大値を求めるために Lagrange 乗数 l をつけた次の式 W を設定します。

$$W = V^{**} - l (T_p^T T_p - 1)$$
$$= A_p^T A_{pp} (I_{pp} - I_p I_p^T / p) A_{pp} A_p - l (T_p^T T_p - 1) \leftarrow [1b)]$$
$$= T_p^T R_{pp}^T A_{pp} (I_{pp} - IpIp^T / p) A_{pp} R_{pp} T_p - l (T_p^T T_p - 1) \quad \leftarrow [2]$$

この W を未知の T_p で微分した式 Df(W, T_p)を 0 とします。

$$Df(W, T_p) = 2 [R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / p) A_{pp} R_{pp} T_p - l T_p] = 0$$

[3]　　$R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / p) A_{pp} R_{pp} T_p = l T_p$　　←$l T_p$を右辺に

左辺　$= R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / p) A_{pp} A_p$　　← [2] $A_p = R_{pp} T_p$
$= R_{pp}^T (A_{pp} I_{pp} A_{pp} A_p - A_{pp} I_p I_p^T A_{pp} A_p / p)$　　←R_{pp}^Tだけを外に
$= R_{pp} (A_{pp} I_{pp} A_{pp} A_p - A_{pp} I_p I_p^T A_{pp} A_p / p)$　　←R_{pp}は対称行列
$= R_{pp} (A_{pp} A_{pp} A_p - A_{pp} I_p I_p^T A_{pp} A_p / p)$　　←$A_{pp} I_{pp} = A_{pp}$
$= R_{pp} (A_{pp} A_p^{(2)} - A_{pp} I_p I_p^T A_{pp} A_p / p)$　　←$A_{pp} A_p = A_p^{(2)}$
$= R_{pp} (A_{pp} A_p^{(2)} - A_p A_p^T A_p / p)$　　←$I_p^T A_{pp} = A_p^T$
$= R_{pp} (A_p^{(3)} - A_p A_p^T A_p / p)$　　←$A_{pp} A_p^{(2)} = A_p^{(3)}$

よって[3]は次の[3b]になり、この[3b]が成立するときに V は最大化します。

[3b]　　$R_{pp} (A_p^{(3)} - A_p A_p^T A_p / p) = l T_p$

ここで

[4]　　$W_p = A_p^{(3)} - A_p A_p^T A_p / p$

とすると，[3b]は次の[3c]になります。W_p, l, T_p は未知数です。

[3c]　　$R_{pp} W_p = l T_p$

Ap は次の式で示され、「構造ベクトル」とよばれます（→後述）。

[5]　　$A_p = R_{pp} W_p / (W_p^T R_{pp} W_p)^{1/2}$

このままでは[4]と[5]を同時に満足する A_p, W_p を計算できません。[5]の式の右辺の Wp は[4]の式の左辺であり、[4]の式の右辺の Ap は[5]の式の左

辺であるので、互いに依存しているからです。そこでプログラムではじめに W_p に適当な値を入れ、[5]によって A_p を求め、それを使って[4]で W_p を求めます。そして、さらに[5]に戻って、[4]で得られた W_p から A_p を求め、[4]で W_p を求めます。このような繰り返しを A_p に変化がなくなるまで行うと、[4]と[5]を満足させる A_p と W_p の値が決まります。この A_p が最初の構造（因子）ベクトルです。

1 つの因子が見つかった後は残差の相関行列(R_{pp})から、順次同じプロセスで因子を探します。毎回因子ベクトルが得られたら、Wp と標準化データ行列(Z_{np})を使って因子得点行列(S_{np})を計算します。

●単位行列・単位ベクトルの利用

行列の演算は、その成分を展開すると理解できます。

[1] $V^* = \Sigma a_i^2 - (\Sigma a_i)^2 / p = A_p^T (I_{pp} - I_p I_p^T / p) A_p$

この右辺の成分を確かめます。

$$A_p^T (I_{pp} - I_p I_p^T / p) A_p$$

$$= A_p^T \left(\begin{bmatrix} 1 & \square & \square & \square \\ \square & 1 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} [1, 1, \ldots, 1] / p \right) A_p$$

$$= A_p^T \left(\begin{bmatrix} 1 & \square & \square & \square \\ \square & 1 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \ldots & \ldots & \ldots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} / p \right) A_p$$

$m = 1 / p$ とすると

$$V^* = A_p^T \left(\begin{bmatrix} 1 & \square & \square & \square \\ \square & 1 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & 1 \end{bmatrix} - \begin{bmatrix} m & m & \ldots & m \\ m & m & \ldots & m \\ \ldots & \ldots & \ldots & m \\ m & m & m & m \end{bmatrix} \right) A_p \qquad \leftarrow \text{行列を展開}$$

$$= [a_1, a_2, \ldots, a_p] \begin{bmatrix} 1-m & -m & \ldots & -m \\ -m & 1-m & \ldots & -m \\ \ldots & \ldots & \ldots & \ldots \\ -m & -m & \ldots & 1-m \end{bmatrix} A_p \qquad \leftarrow \text{行列引き算}$$

$$= [a_1(1 - m) + a_2(-m) + \ldots + a_p(-m),$$
$$a_1(-m) + a_2(1 - m) + \ldots + a_p(-m),$$
$$\ldots$$
$$a_1(-m) + a_2(-m) + \ldots + a_p(1 - m)] a_p \qquad \leftarrow \text{行列積}$$

$$= [a_1 - m a_1 - m a_2 - \ldots - m a_p] a_1$$

$+ [-m\ a_1 + a_2 - m\ a_2 - \ldots + m\ a_p]\ a_2$

$+ \ldots$

$+ [-m\ a_1 - m\ a_2 - \ldots + a_p - m\ a_p]\ a_p$　　←[…]の中を展開

$= [a_1 - m\ (a_1 + a_2 + a_p)]\ a_1$

$+ [a_2 - m\ (a_1 + a_2 + a_p)]\ a_2$

$+ \ldots$

$+ [a_p - m\ (a_1 + a_2 + a_p)]\ a_p$　　←m でくくる

$= [a_1^2 - m\ (a_1 + a_2 + a_p)\ a_1]$

$+ [a_2^2 - m\ (a_1 + a_2 + a_p)\ a_2]$

$+ \ldots$

$+ [a_p^2 - m\ (a_1 + a_2 + a_p)\ a_p]$　　←各項を展開

$= a_1^2 + a_2^2 + \ldots + a_p^2 - m\ (a_1 + a_2 + \ldots + a_p)(a_1 + a_2 + \ldots + a_p)$　←整理

$= a_1^2 + a_2^2 + \ldots + a_p^2 - m\ (a_1 + a_2 + \ldots + a_p)^2$　←2 乗

$= a_1^2 + a_2^2 + \ldots + a_p^2 - (a_1 + a_2 + \ldots + a_p)^2 / p$　←$m = 1 / p$

$= \Sigma\ a_i^2 - (\Sigma\ a_i)^2 / p \quad = \quad V^*$　←[1]

[2]　$V^{**} = A_p^{(2)\,T}\ (I_{pp} - I_p\ I_p^T / p)\ A_p^{(2)}$

$= A_p^T\ A_{pp}\ (I_{pp} - I_p\ I_p^T / p)\ A_{pp}\ A_p$　… [1b]

上の等式が成立することを $A_p^T\ A_{pp}$ と $A_{pp}\ A_p$ の成分で確認します。

$$A_p^T\ A_{pp} = [a_1, a_2, \ldots, a_p] \begin{bmatrix} a_1 & \square & \square & \square \\ \square & a_2 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & a_p \end{bmatrix} = [a_1^2, a_2^2, \ldots, a_p^2] = A_p^{(2)\,T}$$

$$A_{pp}\ A_p = \begin{bmatrix} a_1 & \square & \square & \square \\ \square & a_2 & \square & \square \\ \square & \square & \ldots & \square \\ \square & \square & \square & a_p \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{bmatrix} = \begin{bmatrix} a_1^2 \\ a_2^2 \\ \ldots \\ a_p^2 \end{bmatrix} = A_p^{(2)}$$

よって

$$A_p^T\ A_{pp}\ (I_{pp} - I_p\ I_p^T / p)\ A_{pp}\ A_p = A_p^{(2)\,T}\ (I_{pp} - I_p\ I_p^T / p)\ A_p^{(2)} = V^{**}$$

●構造ベクトル

標準化されたデータ行列の Z_{np} の相関行列は(n:データの個数)

[1]　$R_{pp} = Z_{np}^T\ Z_{np} / n$

Z_{np}に重みベクトルW_pを右積して合成した変数ベクトルF_nとします。

[2]　$F_n = Z_{np}\ W_p$

合成変数ベクトルF_nの分散$V(F_n)$は、F_nの平均は0なので

[3]　$V(F_n) = F_n^T\ F_n / n$
$= (Z_{np}\ W_p)^T\ Z_{np}\ W_p / n$
$= W_p^T\ Z_{np}^T\ Z_{np}\ W_p / n$
$= W_p^T\ R_{pp}\ W_p$　← [1]

合成変数ベクトルF_nを標準化したベクトルG_nは

[4]　$G_n = F_n / V(F_n)^{1/2}$　←F_nを標準化
$= Z_{np}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2}$　←[2], [3]

この標準化されたデータ行列Z_{np}と合成変数ベクトルG_nとの相関係数ベクトルをA_pとすると

$A_p = Z_{np}^T\ G_n / n$
$= Z_{np}^T\ Z_{np}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2} / n$　←[4]
$= R_{pp}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2}$　←[1]

この最終式が先の本文の[5]になります。

$$A_p = R_{pp}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2}$$

＊芝(1975)を参照しました。同書はA_pを「構造ベクトル」とよび、その重要性を強調しています。

●因子集中分析

変数の重みと個体の得点を昇順でソートし、得点を並び替えると次のような集中化した得点になります。

Fct.cct	c. Clear	b. Sharp	d. Hard	a. Big	e. Heavy
7.sa	2	2	1	-2	-2
1.pa	2	2	2	-1	-3
3.ta	1	2	2	-1	-2
11.ra	2	-2	-3	1	0
5.ka	1	3	3	0	-1
10.na	0	-1	-2	0	0
9.ma	-1	-1	-2	0	0
2.ba	-3	-3	1	2	2
4.da	-3	-1	1	2	2
8.za	-2	-1	0	2	3
6.ga	-3	-2	2	3	3

■音象徴の実験

下左図は 1 音節の音の個人的な感覚を 5 つの-3 ~ 3 の尺度で記入したものです。たとえば pa と聞いてとても clear という感じがすれば 3, 逆にとても dark という感じがあれば-3 とします。どちらでもなければ 0 でその間に 2, 1, 0, -1, -2 という段階をつけてみました。いわゆる「音象徴」(sound symbolism)に関する個人的な実験です。

Ss	Big	Sharp	Clear	Hard	Heavy
1.pa	-1	2	2	2	-3
2.ba	2	-3	-3	1	2
3.ta	-1	2	1	2	-2
4.da	2	-1	-3	1	2
5.ka	0	3	1	3	-1
6.ga	3	-2	-3	2	3
7.sa	-2	2	2	1	-2
8.za	2	-1	-2	0	3
9.ma	0	-1	-1	-2	0
10.na	0	-1	0	-2	0
11.ra	1	-2	2	-3	0

FA.d.	#1	#2	#3	#4
1.pa	-1.329	.823	-.733	.196
2.ba	.989	.246	-2.171	-.667
3.ta	-.946	.823	-.057	-.463
4.da	1.134	.274	.664	-.761
5.ka	-.400	1.390	1.430	.912
6.ga	1.535	.813	-.684	.708
7.sa	-1.355	.305	-.138	-.397
8.za	1.203	-.221	1.416	.522
9.ma	-.127	-1.328	.601	-1.663
10.na	-.286	-1.311	.257	-.589
11.ra	-.419	-1.815	-.585	2.202

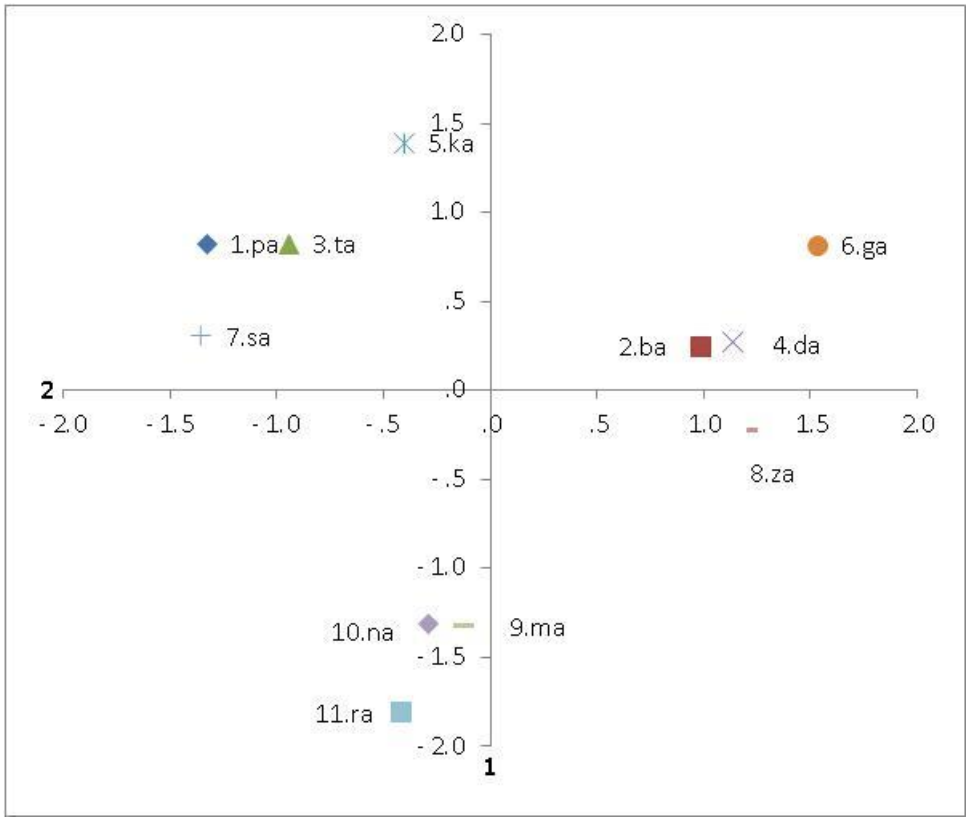
FA.v.	#1	#2	#3	#4
a. Big	.960	-.063	-.106	.217
b. Sharp	-.728	.557	.399	.004
c. Clear	-.940	-.063	.072	.327
d. Hard	.008	1.000	-.013	-.015
e. Heavy	.979	-.150	.016	.048

上右図が因子得点行列、下図が因子行列です。

第 1 因子は Big と Heavy に強く反応しているので「重厚さ」を示しているようです。第 2 因子は Hard と Sharp に反応しているので「切れ味」のようなものを示していると思います。それぞれの因子は次の相関係数行列が示すように無相関になります。このことは軸が直交していることを意味します。

Correlation	1	2	3	4
1	1.000	.000	.000	.000
2	.000	1.000	.000	.000
3	.000	.000	1.000	.000
4	.000	.000	.000	1.000

それぞれの音節の得点とそのグラフ（x 軸=第 1 因子；y 軸=第 2 因子；）を示すと、第 1 因子では有声音と無声音が対立し、第 2 因子では、破裂音（閉鎖音）と摩擦音・鼻音・流音の対立していることがわかります。



7.8. 多重条件分析

言語現象に限らず自然や社会の現象を記述するデータには単一の条件ではなく複数の条件が組み合わさって一定の結果に関係していることがあります。ここでは、たとえば、ある言語形式の使用が、歴史、地理、社会、文体、…の変数によって変化する場合の分析法を考えます。

7.8.1. 多重条件リスト

多重条件分析(Multiple Conditional Analysis)と呼ぶ方法によって、たとえば上の表の条件 c1, c2, c3, c4 と結果の E の間の関係について分析し、単一の条件や多重に結合する条件の影響度を計測します。

M	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z

はじめに条件(c)と結果(E)の該当するセルにタイトル列のデータに出力します。

MA.c1	X	Y	Z
A	d1	d3	
A	d2		
B			d4
B			d5

c2	X	Y	Z
C	d1		
D	d2	d3	d4
E			d5

c3	X	Y	Z
F	d1		
F	d2		
G		d3	
H			d4
H			d5

c4	X	Y	Z
I	d1		
J	d2		
K		d3	
L			d4
M			d5

ここで A が条件で X が結果であるとすると、A→X という関係を示しているデータが d1, d2 であることがわかります。このことは上のような単一条件でも、また下のような二重条件でも同様です。さらに三重条件、四重条件まで条件の組み合わせを増やすことができます。

MA.c1+c2	X	Y	Z
A + C	d1		
A + D	d2	d3	
B + D			d4
B + E			d5

MA.c1+c3	X	Y	Z
A + F	d1		
A + F	d2		
A + G		d3	
B + H			d4
B + H			d5

MA.c1+c4	X	Y	Z
A + I	d1		
A + J	d2		
A + K		d3	
B + L			d4
B + M			d5

(…)

MA.c1+c2+c3+c4	X	Y	Z
A + C + F + I	d1		
A + D + F + J	d2		
A + D + G + K		d3	
B + D + H + L			d4
B + E + H + M			d5

7.8.2. 多重条件頻度

次にそれぞれの条件と結果の組み合わせに該当するデータの頻度を計算します。

(1) 単一条件頻度

MA.f.c1	X	Y	Z
A	2	1	
B			2

c2	X	Y	Z
C	1		
D	1	1	1
E			1

c3	X	Y	Z
F	2		
G		1	
H			2

c4	X	Y	Z
I	1		
J	1		
K		1	
L			1
M			1

出力の X, Y, Z 列は結果列(E)の各成分の絶対頻度です。

(2) 二重条件頻度

すべての条件の中から 2 つの組合せについて頻度を計算します。

MA.f.c1+c2	X	Y	Z
A + C	1		
A + D	1	1	
B + D			1
B + E			1

MA.f.c1+c3	X	Y	Z
A + F	2		
A + G		1	
B + H			2

MA.f.c1+c4	X	Y	Z
A + I	1		
A + J	1		
A + K		1	
B + L			1
B + M			1

さらに、c2+c3, c2+c4, C3+c4 も同様にして計算します。

(3) 三重条件頻度

すべての条件の中から 3 つの組合せについて頻度を計算します。

MA.f.c1+c2+c3	X	Y	Z
A + C + F	1		
A + D + F	1		
A + D + G		1	
B + D + H			1
B + E + H			1

MA.f.c1+c2+c4	X	Y	Z
A + C + I	1		
A + D + J	1		
A + D + K		1	
B + D + L			1
B + E + M			1

MA.f.c1+c3+c4	X	Y	Z
A + F + I	1		
A + F + J	1		
A + G + K		1	
B + H + L			1
B + H + M			1

MA.f.c2+c3+c4	X	Y	Z
C + F + I	1		
D + F + J	1		
D + G + K		1	
D + H + L			1
E + H + M			1

7.8.3. 多重条件係数

多重の条件と結果との関連度を調べるために次のような**多重条件係数** (Coefficient of multiple condition: CMC) を考えます。

(1) 単一条件係数

次のような単一条件での条件係数 CMC は先述の卓立化 Jaccard 係数を使います。CMC(1)の(1)は条件が 1 つであることを示します。

条件 (c)	結果 (e)	ウェイト(w)	頻度(f)	区分
+1 (有)	+1 (有)	w1:(+1)(+1) = +1	f 1	a
+1 (有)	−1 (無)	w2:(+1)(−1) = −1	f 2	b
−1 (無)	+1 (有)	w3:(−1)(+1) = −1	f 3	c
−1 (無)	−1 (無)	w4:(−1)(−1) = +1	f 4	d

たとえば、次の A:X の場合の多重条件係数を計算します。

MA.f.c1	X	Y	Z
A	2	1	
B			2

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

区分:a は条件(+):結果(+)の場合で、A:X=2 になります。区分 b は条件(+):結果(−)の場合で、A:Y+A:Z=1 です。区分 c は条件(−):結果(+)の場合ですがデータにはありません。区分 d は条件(−):結果(−)の場合で、上の表の B:Z=2 がそれにあたります。しかし、Jaccard 係数(J)は a / (a + b + c)という式を使うので d は考慮しません。

J = 2 / (2 + 1 + 0) = 2 / 3 ≒ .666

条件係数(CMC)は、 (N:行数 ; P:列数; abs:絶対値)

CMC = w1*f1*(N+P–2) / [w1*f1*(N+P–2)+abs(w2)*f2 + abs(w3)*f3]

次の A:X の単一条件の CMC(1)を次の分布表に適用すると、

MA.f.c1	X	Y	Z
A	f1: 2 (a)	f2: 1 (b)	
B			f4: 2 (d)

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

CMC(1) = 1*2*(2+3－2) / [1*2*(2+3－2)+1*1+0]
= 6 / 7 ≒ .857

A:Y の単一条件係数 CMC(1)は、

MA.f.c1	X	Y	Z
A	f2: 2 (b)	f1: 1 (a)	
B			f4: 2 (d)

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

CMC(1) = 1*1*(2+3–2) / [1*(2+3–2)*2+1*2+0] = 3 / 5 = .600

(2) 二重条件係数

二重条件ではウェイトを拡大して加算します。ウェイト(w)は条件の和と結果を積算します。区分は条件の和の正負と結果の正負で決まります。条件がゼロの場合は区分がありません。

条件(c1)	条件(c2)	結果(e)	ウェイト(w)	頻度(f)	区分
+1	+1	+1	w1:(+1+1)(+1) = +2	f 1	a
+1	+1	−1	w2:(+1+1)(−1) = −2	f 2	b
+1	−1	+1	w3:(+1−1)(+1) = 0	f 3	−
+1	−1	−1	w4:(+1−1)(−1) = 0	f 4	−
−1	+1	+1	w5:(−1+1)(+1) = 0	f 5	−
−1	+1	−1	w6:(−1+1)(−1) = 0	f 6	−
−1	−1	+1	w7:(−1−1)(+1) = −2	f 7	c
−1	−1	−1	w8:(−1−1)(−1) = +2	f 8	d

上のウェイトはそれぞれ+2, −2 になっていて、これは+1, −1 としても CMC の計算に変わりありません。しかし、三重条件、さらに多重条件では異なる数値になるので、一般化するために、二重条件でもこのままにしておきます。

MA.f.c1+c2	X	Y	Z
A + C	1		
A + D	1	1	
B + D			1
B + E			1

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

たとえば上の[A+C]:X の二重条件係数 $CCM^{(2)}$は（ウェイト w は絶対値とします）、

MA.f.c1+c2	X	Y	Z
A + C	f1:1 (a)		
A + D	f3: 1 (-)	f4: 1 (-)	
B + D			f8: 1 (d)
B + E			f8: 1 (d)

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

$CCM^{(2)}$ = w1*f1*(N+P–2) / [w1*f1*(N+P–2)+w2*f2+w7*f7]
= 2*1*(4+3－2) / [2*(4+3－2) + 2*0 + 2*0]
= 10 / 10 = 1.000

ここで条件[A+D]の行は、条件[A+C]との比較において、c1=+1, c2=−1 になるので加算すると 0 になり、a, b, c, d のどの区分にも入りません。

次に上の[A+D]:X の二重条件係数.833 は次のように計算します。

MA.f.c1+c2	X	Y	Z
A + C	f3:1 (-)		
A + D	f1: 1 (a)	f2: 1 (b)	
B + D			f8: 1 (d)
B + E			f8: 1 (d)

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

$CCM^{(2)}$ = w1*f1*(N+P–2) / [w1*f1*(N+P–2)+abs(w2)*f2+abs(w7)*f7]
= 2*1*(4+3－2) / [2*1*(4+3－2) + 2*1 + 0]
= 10 / 12 = .833

ここでも条件[A+C]は c1=+1, c2=−1 になるので加算すると 0 になり、a, b, c, d のどの区分にも入りません。

(3) 三重条件係数

次は三重条件ですが、条件の数が増えるだけで、計算方法は同じです。

c1	c2	c3	結果 (e)	ウェイト (w)	頻度 (f)	区分
+1	+1	+1	+1	w1:(+1+1+1)(+1) = +3	f 1	a
+1	+1	+1	−1	w2:(+1+1+1)(−1) = −3	f 2	b
+1	+1	−1	+1	w3:(+1+1−1)(+1) = +1	f 3	a
+1	+1	−1	−1	w4:(+1+1−1)(−1) = −1	f 4	b
+1	−1	+1	+1	w5:(+1−1+1)(+1) = +1	f 5	a
+1	−1	+1	−1	w6:(+1−1+1)(−1) = −1	f 6	b
+1	−1	−1	+1	w7:(+1−1−1)(+1) = −1	f 7	c
+1	−1	−1	−1	w8:(+1−1−1)(−1) = +1	f 8	d
−1	+1	+1	+1	w9:(−1+1+1)(+1) = +1	f 9	a
−1	+1	+1	−1	w10:(−1+1+1)(−1) = +1	f 10	b
−1	+1	−1	+1	w11:(−1+1−1)(+1) = −1	f 11	c
−1	+1	−1	−1	w12:(−1+1−1)(−1) = +1	f 12	d
−1	−1	+1	+1	w13:(−1−1+1)(+1) = −1	f 13	c
−1	−1	+1	−1	w14:(−1−1+1)(−1) = +1	f 14	d
−1	−1	−1	+1	w15:(−1−1−1)(+1) = −3	f 15	c
−1	−1	−1	−1	w16:(−1−1−1)(−1) = +3	f 16	d

三重条件ではそれぞれウェイトが異なりますが、これまでと同じように条件(c1, c2, c3)の和の正負と結果(e)の正負の組み合わせから区分 a, b, c, d を決めます。

MA.f.c1+c2+c3	X	Y	Z
A + C + F	1		
A + D + F	1		
A + D + G		1	
B + D + H			1
B + E + H			1

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

たとえば上の[A+C+F]:X の多重条件係数 CMC は（ウェイト w は絶対値とします）、

MA.f.c1+c2+c3	X	Y	Z
A + C + F	f1: 1 (a)		
A + D + F	f5: 1 (a)		
A + D + G		f8: 1 (d)	
B + D + H			f16: 1 (d)
B + E + H			f16: 1 (d)

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

CMC = (w1*f1+w3*f3+w5*f5+w9*f9)*(N+P–2)
/ (w1*f1+w3*f3+w5*f5+w9*f9)*(N+P–2)

+ (w2*f2+w4*f4+w6f6+w10*)
+ (w7*f7+w11*f11+w13*f13+w15+f15)]
= (3*1+1*1)*(5+3－2)
/ (3*1+1*1)*(5+3－2)
+ 0
+ 0
= 24 / 24 = 1

[A+D+F]:X の多重条件係数 CMC は、

MA.f.c1+c2+c3	X	Y	Z
A + C + F	f1: 1 (a)		
A + D + F	f5: 1 (a)		
A + D + G		f4: 1 (b)	
B + D + H			f16: 1 (d)
B + E + H			f16: 1 (d)

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

CMC = (w1*f1+w3*f3+w5*f5+w9*f9)*(N+P–2)
/ (w1*f1+w3*f3+w5*f5+w9*f9)*(N+P–2)
+ (w2*f2+w4*f4+w6f6+w10*)
+ (w7*f7+w11*f11+w13*f13+w15+f15)]
= (3*1+1*1)*(5+3－2)
/ (3*1+1*1)*(5+3－2)
+ 1*1
+ 0
= 24 / (24 + 1) = .960

ここでは[A+D+G]の条件が、[A+D+F]と[+, +, −]のように 2 回一致するので、f4、区分[b]になり、分母をわずかに増やしています。

プログラムでは四重条件まで計算して出力します。入力行列の列数が 5 以上の場合もすべての可能な四重条件までを計算すれば、実際上不都合がほとんでないからです。このことは多変量分析で固有値・固有ベクトルを算出するときと同様です。逆に五重条件などを出力してもほとんど分析が不可能になります。

●多重条件分析による相対化の実験

言語資料は次のような分布を示すことがおおいのですが、しばしば、データを数量化したとき、それが正しく相対化されているかどうか、問題になることがあります。

ID	1.Rasgo	Entorno	2.A25	3.Tipo	Letra
2	/y/	#V_V	a0925	1.V	i
3	/y/	#V_V	a0925	1.V	i
4	/y/	#V_V	a0950	1.V	i
4	/i/	#V_V	a0950	1.V	i
5	/y/	#V_V	a0975	1.V	i
5	/y/	#V_V	a0975	1.V	i
5	/y/	#V_V	a0975	1.V	i
9	/y/	#V_V	a1225	4.Gc	y
9	/y/	#V_V	a1225	4.Gc	y

次のような簡略化したデータを使って相対化の実験をします。M1 には 5 データあります。

M1	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z

次の M2 には M1 の d5 をさらに 5 回繰り返して追加してあります。

M2	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z
d6	B	E	H	M	Z
d7	B	E	H	M	Z
d8	B	E	H	M	Z
d9	B	E	H	M	Z
d10	B	E	H	M	Z

M2 のように母数が多くなると、当然頻度が多くなり、このことを考慮しない絶対頻度による分析の不備が指摘されます。このことは次の M1.f と M2.f.のそれぞれの絶対頻度(f)を比べると明らかです。M2.f の B:Z=7 はデータ行列に該当する多くのデータ(d6, ..., d10)が含まれているためです。

M1.f	X	Y	Z
A	2	1	
B			2

M1.c	X	Y	Z
A	.857	.600	
B			1.000

M2.f	X	Y	Z
A	2	1	
B			7

M2.c	X	Y	Z
A	.857	.600	
B			1.000

ところが、それぞれの右の表の条件係数(c)については、M1.c と M2.c では変化がありません。B:Z の値はどちらも 1.000 を示しますが、これは条件 B と競合する行が他にないためです。このような分布を「排他分布」(Exclusive distribution)とよびます。排他分布は、分布行列全体で示されることもあれば、一定の個別の分布だけで示されることもあります。上の例では、A:X, A:Y, B:Z はどれも排他分布を示しています。

上の例のように分布が排他的であれば、入力データの偏りは、条件係数に影響しません。このことは先の Jaccard 係数の式 a / (a + b + c)を見れば明らかです。この式の b, c の数値が排他分布を示す B:Z で 0 になるからです。

また、A:X, A:Y にも影響しないのは、M1 と M2 の差分 d6-10 に A:X, A:Y が含まれないためです。

次の条件 c2 の場合は様子が少し異なります。

MA1.f.	X	Y	Z
C	1		
D	1	1	1
E			1

MA1.c.	X	Y	Z
C	.800		
D	.571	.667	.571
E			.800

MA2..f.	X	Y	Z
C	1		
D	1	1	1
E			6

MA2.c.	X	Y	Z
C	.800		
D	.571	.667	.333
E			.960

上の E:Z を見ると、絶対頻度(f)の増加によって、条件係数(c)の値も上昇しています。これは、E:Z が排他分布ではなく、D:Z と競合しているので(「競合分布」(Competitive distribution)とよびます）、その影響を受けるからです。このように、条件係数は絶対頻度のように自己だけの値で評価するのではなく、自己と他者との関係を評価します。しかし、自己の数値（絶対頻度）が非常に大きくなっても(1→6)、極端な上昇を示しません(.800→.960)。

このように条件係数は、データが排他分布を示すときは、絶対頻度の変化に影響されず、また、データが競合分布を示すときでも絶対頻度のような極端な変化を示さないので、データの分布の解釈がより正確になります。

一方、絶対頻度は収集されたデータの状態をそのまま示しますので、相対化すると見失われてしまう実態を見るために役立ちます。分析には両者を考慮を入れるべきです。

■スペイン語の硬口蓋有声摩擦音と文字<j>

次は 10-13 世紀のスペイン北部で記された文献中の<i>, <j>, <y>の文字について多重条件分析（二重条件）をした結果の抜粋です。下左表が絶対頻度、下右表が多重条件係数です。どちらも<j>の列で降順に並べ替えてあります。

MA.f.Fonema+a25	<i>	<j>	<y>
/i/ + 1200	1465	451	1
/i/ + 1175	805	122	2
/i/ + 1225	333	111	18
/ly/ + 1200	201	56	
/ʒ/ + 1250	21	52	
/i/ + 1150	386	42	1
/ʒ/ + 1225	4	37	
/ly/ + 1225	15	26	
/i/ + 1250	233	21	70
/ʒ/ + 1275	1	21	
/ly/ + 1175	116	20	
/ʒ/ + 1200	13	20	
/i/ + 1275	70	18	33

MA.t.Fonema+a25	<i>	<j>	<y>
/i/ + 1200	.988	.922	.020
/i/ + 1225	.955	.896	.547
/i/ + 1175	.984	.834	.070
/ly/ + 1200	.695	.784	
/i/ + 1150	.967	.741	.061
/ʒ/ + 1250	.142	.710	
/i/ + 1275	.831	.684	.853
/ʒ/ + 1225	.031	.667	
/ʒ/ + 1200	.122	.659	
/i/ + 1250	.941	.640	.890
/ly/ + 1225	.117	.585	
/ʒ/ + 1275	.008	.498	
/ly/ + 1175	.530	.483	

この表を見ると、たとえば *fijo, ojo, concejo* などのように、<i>ではなく<j>が音素/ʒ/を表示するのは 13 世紀から多くなったことがわかります。多重連関係数を使うと絶対頻度だけでわかりにくい場合に相対的な判断ができるようになります。

7.9. 関係分析

行列の変数間の関係を調べる方法として「連関規則」(Association rules)が使われますが[28]、ここで私たちは変数間だけでなく個体間の関係と変数・個体間の関係を見るために「連関規則」の考え方を応用した「関係分

[28] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. http://www.almaden.ibm.com/cs/quest/papers/sigmod93.pdf [2016/4/25]
尾崎隆『ビジネスに活かすデータマイニング』技術評論社(2014) pp.163-184.
豊田秀樹『データマイニング入門』東京書籍(2008) pp.149-183.

析」(Relation Analysis)を提案します。

単数の個体や属性の関係だけでなく、それぞれをセットにして組み合わせる方法も提示します。

因子間の連繋の細部を見るための数表（関係表）と、それぞれの因子を繋ぐネットワークを描くプログラムを開発して使います[29]。

7.9.1. 一次元分析

はじめに行列の個体間の関係を見ます。実際には大きな行列を使いますが、ここでは簡単化して 5 行 4 列の行列 D1 を例として説明します。

下表の{i1, i2, i3, i4, i5}を「個体」と呼び、{A, B, C, D}を「属性」と呼び、個体の関係をさまざまな指標によって示します。例として個体 i1 と i3 使います。

D1	A	B	C	D
i1	1	1	0	0
i2	0	0	1	0
i3	0	1	0	0
i4	0	0	1	1
i5	1	1	1	0

はじめに i1 と B の間のそれぞれの和 S(i1)と S(i3)を求めます。

S(i1) = 1 + 1 = 2
S(i3) = 1

行列全体の総和 S は

S = 9

よって、i1 と i3 の間のそれぞれの確率 P(i1)と P(i3)は

P(i1) = S(i1) / S = 2 / 9 = .222
P(i3) = S(i3) / S = 1 / 9 = .111

ここで i1 と i3 が同時に出現する回数 C(i1, i3)は 2 列目の属性 B だけで一致しているので

C(i1, i3) = 1

次に i1 が出現した回数は 2 回で(S(i1)=2)、その 2 回のうち、i3 と一致しているのは 1 回なので(C(i1, i3) = 1)、「i1 が出現したときに i3 が出現する」

[29] http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/index.html [2016/12/26]

という条件付き確率 P(i3|i1)は[30]

P(i3|i1) = C(i1, i3) / S(i1) = 1 / 2 = .500

ここで、たとえば「単語 i1 が単語 i3 と共起する割合」である P(i3|i1)と、「データ内の全単語の中で単語 i3 が占める割合」（＝確率 P(i3）の差(Difference: Dif)に注意します。たとえ P(i3|i1)が高くても、そもそも i3 の確率 P(i3）がはじめから十分に高ければ、i1 の出現が i3 の出現に関係している、とは必ずしも言えないからです。

Dif = .500 - .111 = .389

さらに、とくにこのように小さな数値のデータでは有意度(Significance: Sig)が示されるべきなので、次の式で右側累積二項確率の 1 の補数を計算します。

Sig = binS(C, S(i1), P(i3)) = binS(1, 2, .111) = .790

この式は、一般に期待される確率 P(i3) = .111 をもつ事象が S(i1)という試行数の中で C 回起こる右側累積二項確率を求め、その値を 1 から引いた数値を示します。この有意度(.790)はかなり低いので、Dif をそのまま認めることができません。そこで、最終的な評価として A と有意度 Sig の積(A*S)を使います。

D*S = D * Sign = .389 * .790 = .307

プログラムは次の表を出力します。この表を「関係表」(Relation table: [R])と呼びます。

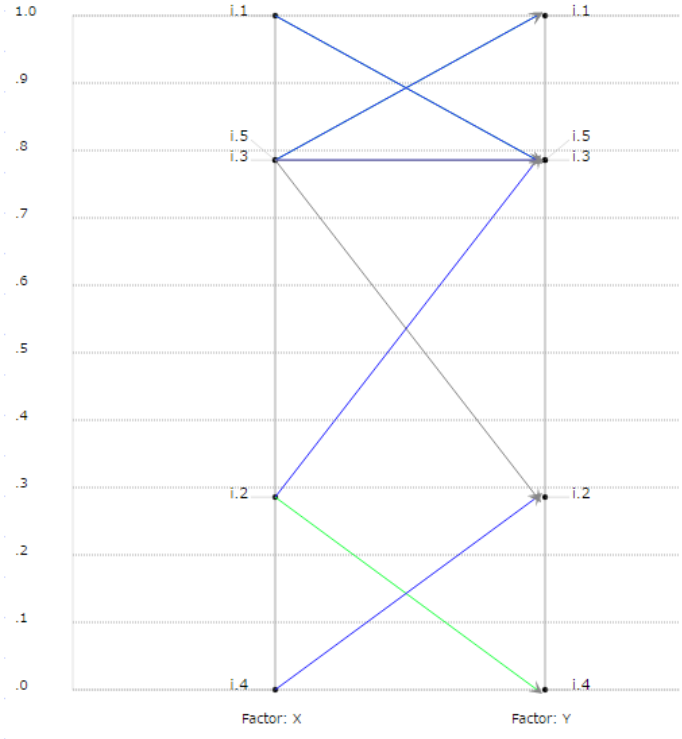
→	[R]	X	Y	N(X)	N(Y)	C	S(X)	S(Y)	S	P(X)	P(Y)	P(Y|X)	Dif.	Sign.	D*S
1	1	2	4	i.2	i.4	1	1	2	9	.111	.222	1.000	.778	.778	.605
2	2	3	1	i.3	i.1	1	1	2	9	.111	.222	1.000	.778	.778	.605
3	3	1	5	i.1	i.5	2	2	3	9	.222	.333	1.000	.667	.889	.593
4	4	2	5	i.2	i.5	1	1	3	9	.111	.333	1.000	.667	.667	.444
5	5	3	5	i.3	i.5	1	1	3	9	.111	.333	1.000	.667	.667	.444
6	6	5	1	i.5	i.1	2	3	2	9	.333	.222	.667	.444	.874	.388
7	7	1	3	i.1	i.3	1	2	1	9	.222	.111	.500	.389	.790	.307
8	8	4	2	i.4	i.2	1	2	1	9	.222	.111	.500	.389	.790	.307
9	9	5	2	i.5	i.2	1	3	1	9	.333	.111	.333	.222	.702	.156
10	10	5	3	i.5	i.3	1	3	1	9	.333	.111	.333	.222	.702	.156

上表の X, Y は対象の 2 変数の行番号と列番号、N(X), N(Y)は変数名、C が X, Y の共起回数、S(X), S(Y)はそれぞれの変数の和、P(X), P(Y)はそれぞ

[30] 条件付き確率は「連関規則」(Association rules)のよる分析では「信頼度」(Confidence)とよばれます(尾崎(2014: 165166, 豊田(2008: 148))。

れの変数の確率、P(Y|X)は条件確率、Dif は条件確率から Y の確率 P(Y)を引いた差、Sign は有意度、D*S は差 Dif と有意度の積を示します。私たちは最終列の積を重視するので、この列を降順でソートして出力するようにプログラムを作成しました。

次は、データを主成分分析し、第 1 主成分の主成分得点を[0, 1]の範囲に限定化して上下に配置し、関係の強さを線の色で表した図です。これを「関係図」(Relation Chart)と呼びます。グレー→青(.25 以上)→緑(.5 以上)→赤(.75 以上)の順で D*S の数値が高くなります。



＊入力行列を転置すれば、次のように個体ではなく属性の関係分析ができます。

→	[R]	X	Y	N(X)	N(Y)	C	S(X)	S(Y)	S	P(X)	P(Y)	P(Y|X)	Dif.	Sign.	D*S	*Reset*
1	1	1	2	A	B	2	2	3	9	.222	.333	1.000	.667	.889	.593	1
2	2	4	3	D	C	1	1	3	9	.111	.333	1.000	.667	.667	.444	2
3	3	2	1	B	A	2	3	2	9	.333	.222	.667	.444	.874	.388	3
4	4	3	4	C	D	1	3	1	9	.333	.111	.333	.222	.702	.156	4
5	5	1	3	A	C	1	2	3	9	.222	.333	.500	.167	.444	.074	5
6	6	3	1	C	A	1	3	2	9	.333	.222	.333	.111	.471	.052	6
7	7	2	3	B	C	1	3	3	9	.333	.333	.333	.000	.296	.000	7

●非負行列

上の例では 0 と 1 からなるデータを対象にしましたが、ここで共起回数(C)の概念を拡張し、次のような頻度行列も扱えるようにします。さらに 0.56 や 10.3 などの小数であっても可能にします。つまり負値を含まない行

列（非負行列 non negative matrix）を分析対象に含めます。

D2	A	B	C	D	E
i1	10	19	14	7	12
i2	11	7	10	0	1
i3	0	0	1	12	1
i4	0	1	2	3	3

たとえば、属性 A→B の連繋を見るために、はじめに、その共起数(Cooccurence: C)を次のように定義します。

C(A, B) = Σ (i) min[A(i), B(i)]
= min(10, 19) + min(11, 7) + min(0, 0) + min (0, 1)
= 10 + 7 + 0 + 0 = 17

上の式は、A 列と B 列の各ペアのうち、小さい方の数(minimum: min)を足し上げていくことを意味します。この A→B の連繋では、min(10, 19) = 10, min(11, 7) = 7, min(0, 0) = 0, min(0, 1) = 0、よって

c(A, B) = 10 + 7 + 0 + 0 = 17

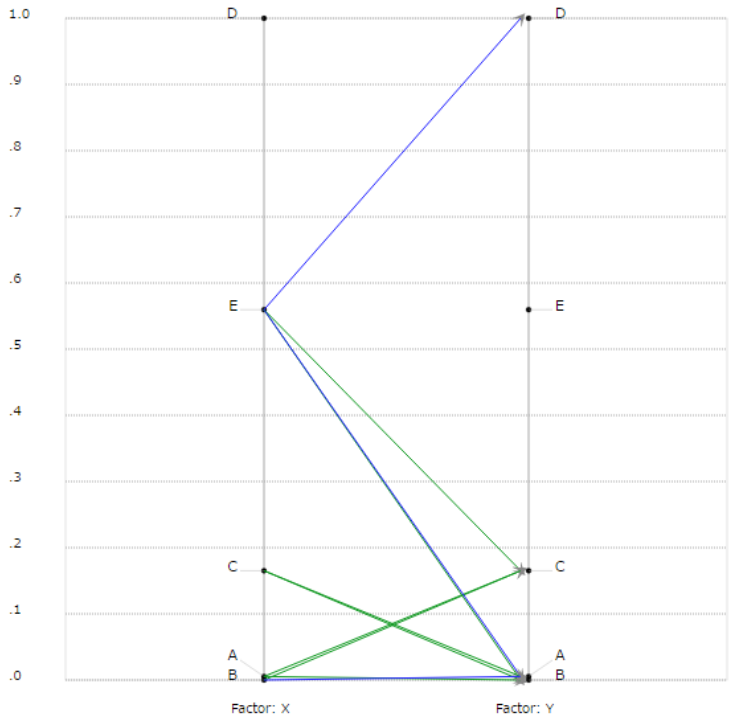
になります[31]。たとえば min(10, 19) = 10 が、10 と 19 の「共通部分」なので、それを「共起数」と見なします。よって

A の和：S(A) = 21
B の和：S(B) = 27
A, B の共起回数：C = C(A, B) = 17
総和：S = 114
A の確率：P(A) = S(A) / S = 21 / 114 = .184
B の確率：P(B) = S(B) / S = 27 / 114 = .237
条件確率：P(B|A) = C / S(A) = 17 / 21 = .810
差：Dif = P(B|A) - P(B) = .810 - .237 = .573
有意度：Sig = binS(C, S(A), P(B)) = binS(17, 21 .237) = 1.000
積：D*S = .573 * 1.000 = .573

次の関係表は上の非負行列の属性{A, B, C, D}のすべての組み合わせで、それぞれの値を計算し、積 D*S によって逆順にソートした結果です。A→B の連繋は[R]の 6 に示されています。

[31] 実際には、このケースでは 2 つの引数を比較するので、「最小値」というよりも、「より小さな値」を探すことになります。しかし、その関数 min はセット X 因子（左側）と Y 因子の間の共起数を求めるときにも使いますから、そのときは引数は全部で 3 または 4 になるので、その場合と統一させた機能（最小値）を min に与えておきます。

[R]	X	Y	N(X)	N(Y)	C	S(X)	S(Y)	S	P(X)	P(Y)	P(Y|X)	Dif.	Sign.	D*S
1	1	3	A	C	20	21	27	114	.184	.237	.952	.716	1.000	.716
2	5	3	E	C	16	17	27	114	.149	.237	.941	.704	1.000	.704
3	5	2	E	B	14	17	27	114	.149	.237	.824	.587	1.000	.587
4	2	3	B	C	22	27	27	114	.237	.237	.815	.578	1.000	.578
5	3	2	C	B	22	27	27	114	.237	.237	.815	.578	1.000	.578
6	1	2	A	B	17	21	27	114	.184	.237	.810	.573	1.000	.573



なお、この計算法は先の(0, 1)のデータにあてはめても同じ結果になるので、(0, 1)行列と頻度行列のどちらの行列でも同じプログラムで処理できます。そして同様に小数点のある一般的な非負行列（全成分が[0, 1]の範囲にある小数点ある数値の行列）も分析できます。

7.9.2. 二次元分析

これまで個体間や属性間のそれぞれの関係を分析しましたが、次に個体と属性の間の関係を分析する方法を探ります。

関係分析を適用するときは、次のデータ行列の属性{A, B, C, D, E}を X 因子とし、個体{i1, i2, i3, i4}を Y 因子とします。これは頻度データですが、(0, 1)データでも小数データでも同じ計算方法を使うので、一般の非負行列に適用できます。たとえば、個体{i.1, i.2, i3, i.4}が変異語で、属性{A, B, C, D, E}が言語テキストであるとすれば、変異語とテキストとの関係を考えることになります。

D2	A	B	C	D	E	sH
i1	10	19	14	7	12	62
i2	11	7	10	0	1	29
i3	0	0	1	12	1	14
i4	0	1	2	3	3	9
sV	21	27	27	22	17	S:114

ここで、個体{i1, i2, i3, i4}と属性{A, B, C, D, E}の組み合わせからなる、それぞれのセルの値を共起回数(C)とします。たとえば[i1:A]のセルは 10 ですから、ここで「i1 が A と 10 回共起している」ということになります。同様にして[i2:A]の共起回数は 11 です[32]。

はじめに、個体の和 S(X)が表の横和(Sh=62)であり、属性の和 S(Y)が縦和(Sv=21)であることを確認します。よって、個体(X)の確率 P(X)は横和(62)を総和(S = 114)で割った値(.544)になります。Y 因子の確率(Yp)は横和(62)を総和(114)で割った値(.184)です。

個体の和 S(X): S(i1) = 62
属性の和 S(Y): S(A) = 21
i1, A の共起回数：C = C(i1, A) = 10
総和：S = 114
個体の確率 P(X) = S(i1) / S = 62 / 114 = .544
属性の確率 P(Y) = S(A) / S = 21 / 114 = .184

条件確率 P(Y|X)は共起回数(C)を個体の和 S(X)で割った値です。

P(Y|X) = C(i1, A) / S(X) = 10 / 62 = .161

二次元の両側関係規則では個体が X 因子になるので、ここで計算した条件確率は、先の一次元の片側関係規則の条件確率（C / S(X): 共起回数／X 因子の和）に対応します。つまり、個体 i1 の全体(62)の中で、A に一致する確率は 10 / 62 = .161 ということになります。

次に条件確率 P(Y|X)と Y の確率 P(Y)の差(Dif)は

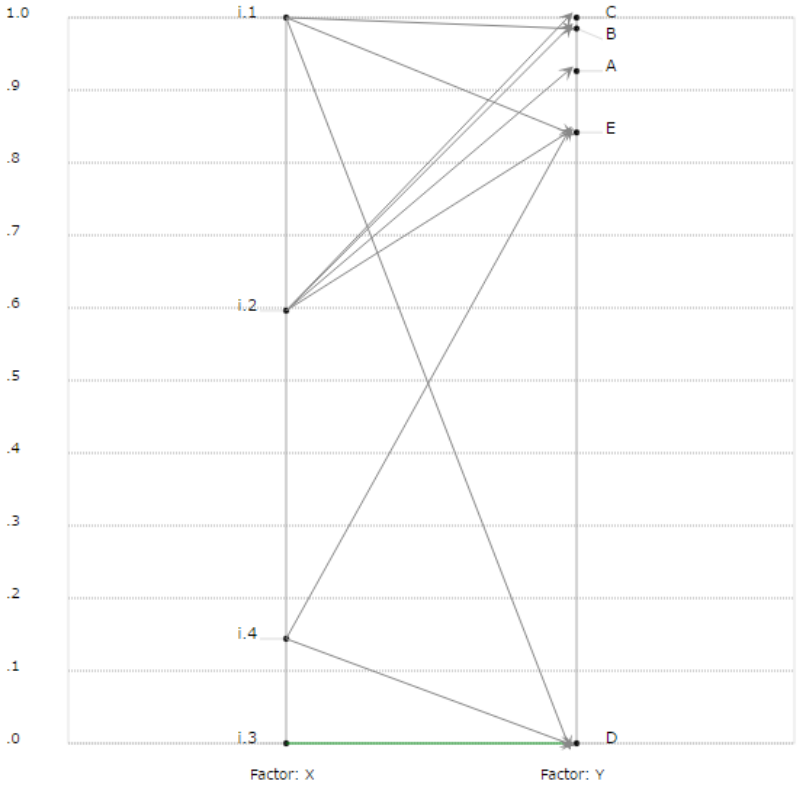
Dif = P(Y|X) - P(Y) = .161 - .184 = -.023
Sig = binS(C, S(i1), P(A)) = binS(10, 62, 21/114) = .272
D*S = -.023 * .272 = -.006

上のように、確率差(Dif)がマイナスの値になる、ということは、関係性がないことを示しています。確かに、Y の確率が.184 であるのに、X→Y の条件付き確率が.161 であるのですから、X との関係はほとんど考えられません。

[32] よって、(0, 1)データでは共起回数は必然的に 0 または 1 に限られます。

以下に関係度数(R)を降順でソートしてTがプラス値である上位8位までを出力しました。

→	[R]	X	Y	N(X)	N(Y)	C	S(X)	S(Y)	S	P(X)	P(Y)	P(Y|X)	Dif.	Sig.	D*S
1	1	3	4	i.3	D	12	14	22	114	.123	.193	.857	.664	1.000	.664
2	2	2	1	i.2	A	11	29	21	114	.254	.184	.379	.195	.989	.193
3	3	4	5	i.4	E	3	9	17	114	.079	.149	.333	.184	.861	.159
4	4	4	4	i.4	D	3	9	22	114	.079	.193	.333	.140	.757	.106
5	5	2	3	i.2	C	10	29	27	114	.254	.237	.345	.108	.873	.094
6	6	1	2	i.1	B	19	62	27	114	.544	.237	.306	.070	.872	.061
7	7	1	5	i.1	E	12	62	17	114	.544	.149	.194	.044	.794	.035
8	8	2	2	i.2	B	7	29	27	114	.254	.237	.241	.005	.451	.002
9	9	2	5	i.2	E	1	29	17	114	.254	.149	.034	-.115	.009	-.001
10	10	1	4	i.1	D	7	62	22	114	.544	.193	.113	-.080	.032	-.003



7.10. 選択軸分析

これまでに見てきたように、一般に属性や個体のプロットを二次元の平面に描くには、その座標として主成分分析・対応分析・因子分析などの多変量解析で得られた値が使われます（→「分析」の「主成分分析」・「対応分析」・「因子分析」）。これらの方法を理解するには高度な数学的準備が必要ですが、このセクションで扱う「選択軸分析」(Analysis by selective axes)と呼ぶ方法は相関行列、類似行列または距離行列さえ理解すれば簡単に実行できます（→「関係」）。以下では属性間の関係を見ますが、データ行列を転置すれば個体間の関係を見ることができます。最後に個体と属

性の関係を示すプロット図の描き方を説明します。

7.10.1. 相関行列による選択軸分析

下左表はデータ行列の例であり(M)、下右表はその相関行列 Co(M)です。

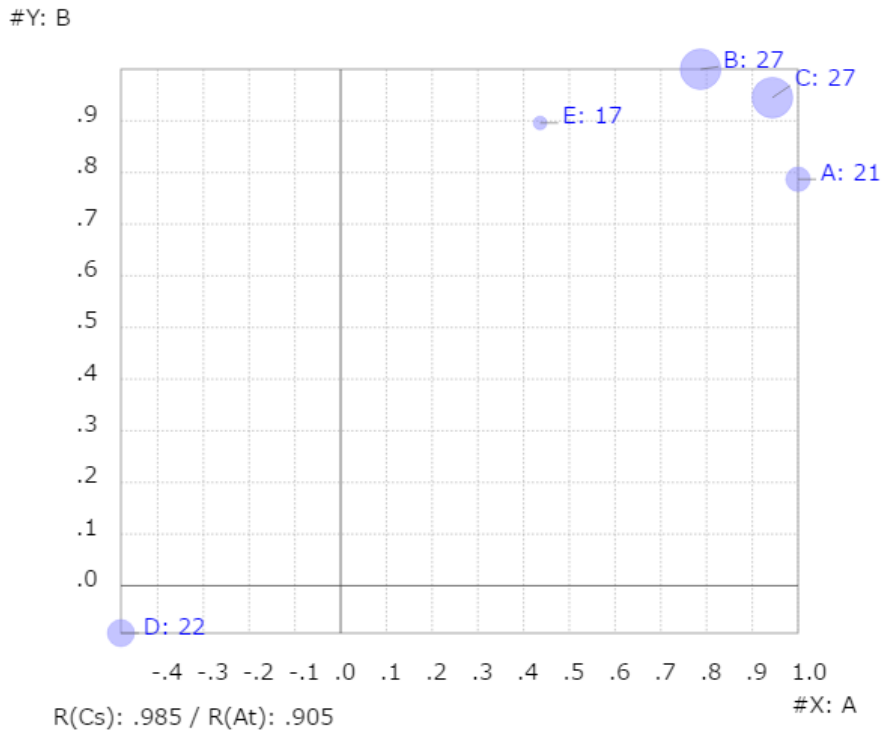
M	A	B	C	D	E
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

Co(M)	X:A	Y:B	C	D	E
A	1.000	.787	.944	-.480	.436
B	.787	1.000	.945	-.092	.896
C	.944	.945	1.000	-.331	.709
D	-.480	-.092	-.331	1.000	.140
E	.436	.896	.709	.140	1.000

上右の相関行列 Co(M)では、それぞれの列(A, B, C, D, E)と他の列との組み合わせの全部について、それぞれの相関係数を算出しています[33]。ここでたとえば A と B の列だけに注目すると、A 列が A と A, B, ..., E の関係の強さを相関係数で表現し、B 列が B と A, B, ..., E の関係の強さを相関係数で表現していることがわかります[34]。そこで相関行列 Co(M)の A 列を横軸(X 軸)とし、B 列を縦軸(Y 軸)として、A, B, ..., E の座標 A(1.000, .787), B(.787, 1.000), C(.944, .945), D(-.480, -.092), E(.436, .896)をプロットすると次の図になります。

[33] 行列(M)の個体数がわずかに 4 個なので、これらの相関係数には有意性がほとんどありません。しかし、ここでは説明の便宜のためにこのように簡単な例を使います。

[34] ここでは A:B 選択軸は縦方向に選択しましたが、相関行列は対称なので横方向に選択しても同じです。しかし後で見るように行列計算（行列積）の右積項に使うので、初めから縦方向の選択にしておくとよいでしょう。



【図-1】

この図はデータ行列(X)の全列(A, B, C, D, E)について、それらと A 列と B 列との関係（相関）を示しています。A のプロット(1.000, .787)は自己(A)との相関が完全な 1.000 になりますが、A 軸（横軸）に重ならないのは、B との関係が.787 であることを示しているからです。よって A 軸（横軸）はそれぞれの列と A の「関係（相関）」を示し、B 軸（縦軸）はそれぞれの列と B の「関係（相関）」を示しています。この A, B のように選択された列に関心や典型性・代表性があれば、その関心・典型的・代表的な列と各列の関係を見るときに上のような図が役立ちます。たとえば、A がスペインの首都 Madrid, B がアルゼンチンの首都 Buenos Aires, C がスペインの地方都市 Salamanca, D がスペインの地方都市 Sevilla, E がアルゼンチンの地方都市 Mendoza で集められたテキストであり、個体 h1, h2, h3, h4 が言語特徴の出現回数を示すようなデータを想定します。そうすると上の図によって A, B, ..., E が A, B とどのような関係の強さを示しているのかがわかります[35]。

さて、上のような A:B のような軸の選択の可能性は A:B 以外にも A:C, A:D, A:E, B:C, B:D, ...のように多くのペアが考えられます[36]。これらの多くのペアの中から全体のプロットが図の中で最大に分散するようなペアを見

[35] ここでは分析者の関心、データの典型性や代表性などを考慮して 2 つの軸を選択しましたが、分析のテーマ（対象、目的）を示す属性を 1 つ選択して X 軸とし、Y 軸としてはこの X 軸と最小絶対相関係数を示す属性を選択することもあります。最小絶対相関係数についてはこの後すぐに説明します。

[36] 属性の列数が p だとすると p*(p-1)/2 = 5*4 = 10 個の属性のペアがあります。

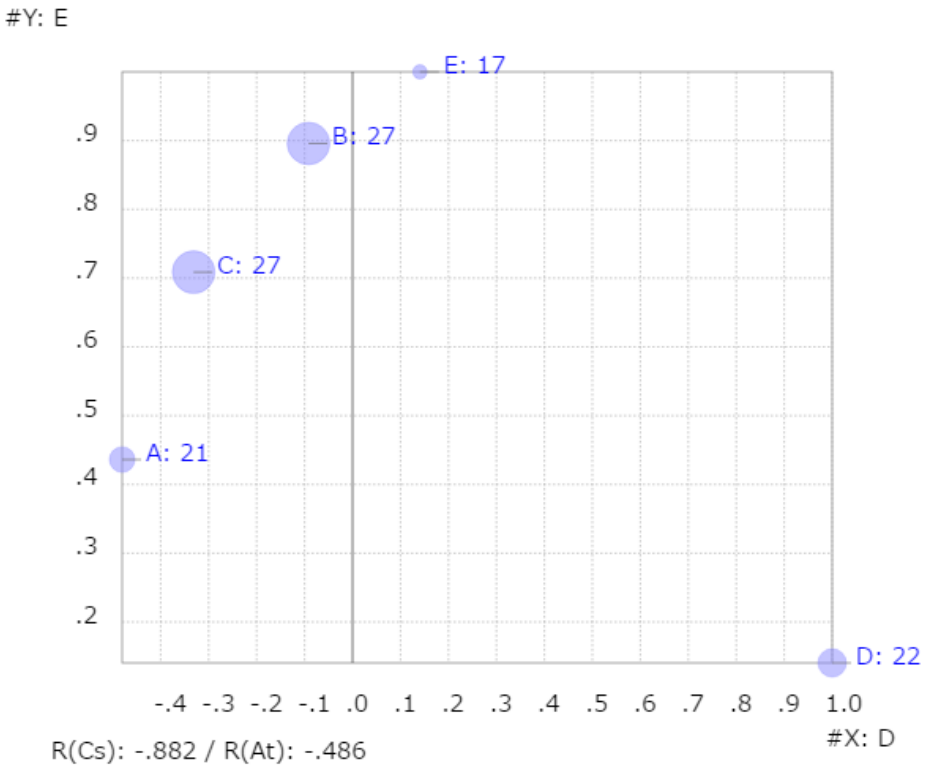
つけてペアのそれぞれを横軸と縦軸にして全体をプロットすると、それぞれの特徴を一番よく示す図ができるはずです。一般に相関係数が 0 に近い 2 つの列のプロットは分散して分布するので、先の相関行列 Co(M)の列間の相関係数の絶対値の最小値を探します[37]。そこで、プログラムで相関行列 Co(M)の相関行列 Co(Co(M))を求めると次が出力されました。

Co(Co(M))	A	B	C	D	E
A	1.000	.905	.981	-.993	.549
B	.905	1.000	.970	-.867	.853
C	.981	.970	1.000	-.961	.700
D	-.993	-.867	-.961	1.000	-.486
E	.549	.853	.700	-.486	1.000

上の行列 Co(Co(M))の絶対値の最小値は-.486 なので、その値を示す D 列と E 列を軸とすると、次のようにプロットの分散が最大になります[38]。

[37] 「相関行列 Co(X)の中の相関係数の最小値」(Co(X)の-.092)ではなく、「相関行列 Co(X)の列間の相関係数の絶対値の最小値」(Co(Co(X))の-.486)です。「相関係数の絶対値の最小値」でなくて単に「相関係数の最小値」とすると、逆相関を示すマイナス値が求められます。たとえば Co(A, D) = -.993 を使うと、プロットが左上から右下に向かう線上に固まるので、全体の分散が小さくなってしまいます。ここでの目的は分散をできるだけ大きくすることです。

[38] 相関行列 Co(M)の相関行列 Co(Co(M))を見ると A:B の相関係数が.905 になることがわかります。【図-1】で正の相関があるようですが、外れ値 D が強く働いています。一方、D:E の相関係数が負の-.486 なので全体の分布は逆方向になっています。ここでも右下の外れ値 D が強く働いています。



【図-2】

先の A:B【図-1】よりも今回の D:E【図-2】のどちらを見ても、それぞれの列が A, B, C, E というグループと D というグループに分かれていることがわかります。そこで、A, B, C, E をまとめて 1 つの軸とし、それと D 軸を比較しましょう。(A, B, C, E)を 1 つの軸にするために、データ行列の A, B, C, E を平均にした A.B.C.E という列を用意します。

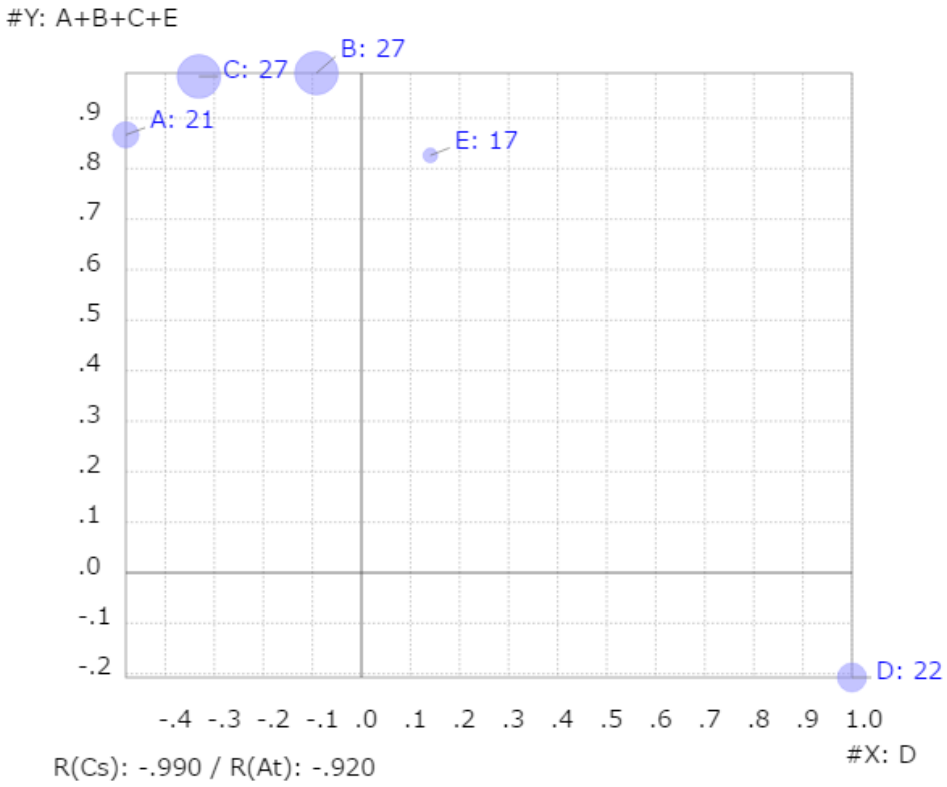
M'	A	B	C	D	E	A.B.C.E.
h1	10	19	14	7	12	13.75
h2	11	7	10	0	1	7.25
h3	0	0	1	12	1	.50
h4	0	1	2	3	3	1.50

この相関行列は

Co(M')	A	B	C	X:D	E	Y:A.B.C.E.
A	1.000	.787	.944	-.480	.436	.867
B	.787	1.000	.945	-.092	.896	.989
C	.944	.945	1.000	-.331	.709	.983
D	-.480	-.092	-.331	1.000	.140	-.207
E	.436	.896	.709	.140	1.000	.826
A.B.C.E.	.867	.989	.983	-.207	.826	1.000

上表 Co(X')の D 列と A.B.C.E 列をそれぞれ横軸と縦軸としてプロットし

ます。



【図-3】

先の【図-2】と比べると両者はかなりよく似ていますが、【図-3】では E が軸 A.B.C.E を代表しているのではなく、むしろ B, C がそれを代表していることがわかります[39]。

7.10.2. 近接行列による選択軸分析

相関行列による選択軸分析のためには少なくとも 3 個の個体が必要です[40]。一方、属性の近接行列を使えば、たとえ個体が 1 つであっても計算できます。

次は先のデータ行列(X)とその規定近接係数行列(Regular Proximity: RP)です（→「規定近接係数」）。

[39] ここでは単純にプロット図を見てグループを決めましたが、より複雑なプロット図を扱うときはクラスター分析法を使って 2 つの群に分割します。

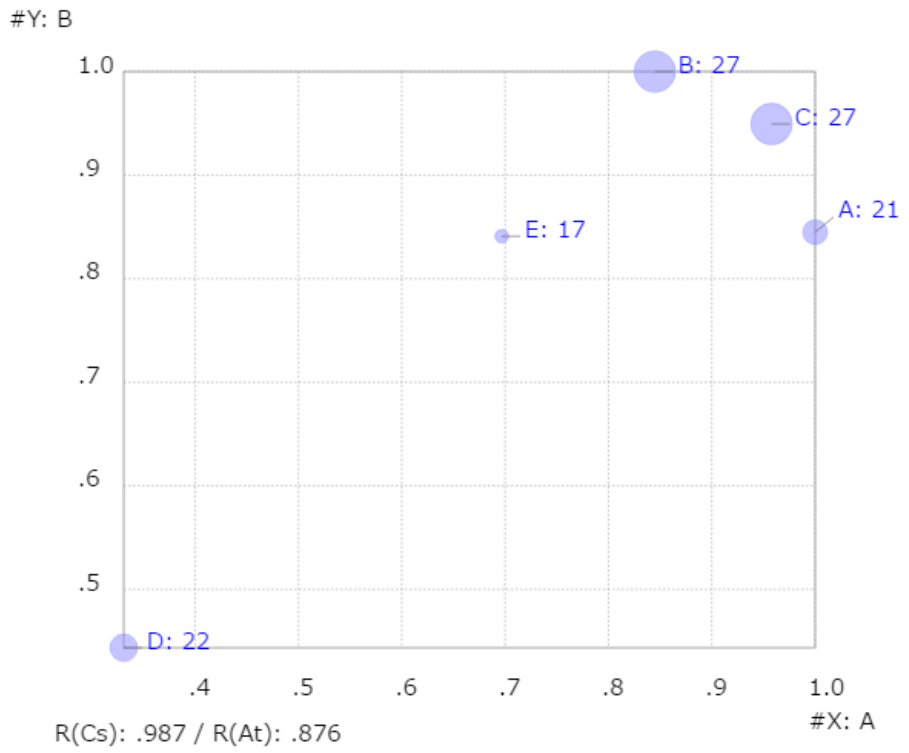
[40] 個体が 1 個の属性間の相関係数はどれも 0.000 になり、2 個では 1.000 または-1.000 になるので無意味な情報になります。実際には個体が 3 個のデータでもその属性の相関係数にはほとんど意味がありません（→「相関係数」）。

M	A	B	C	D	E
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

RP	A	B	C	D	E
A	1.000	.845	.958	.331	.697
B	.845	1.000	.949	.444	.841
C	.958	.949	1.000	.461	.811
D	.331	.444	.461	1.000	.588
E	.697	.841	.811	.588	1.000

次は横軸(X)を A，縦軸(Y)を B としたときのプロットです。この図によってそれぞれのデータが A と B とどのような関係を示すのかがわかります。A 軸（横軸）はそれぞれの列と A の「近さ（距離の限定逆数）」を示し、B 軸（縦軸）はそれぞれの列と B の「近さ（距離の限定逆数）」を示しています。

RP(A, B)	X: A	Y: B
A	1.000	.845
B	.845	1.000
C	.958	.949
D	.331	.444
E	.697	.841



【図-4】

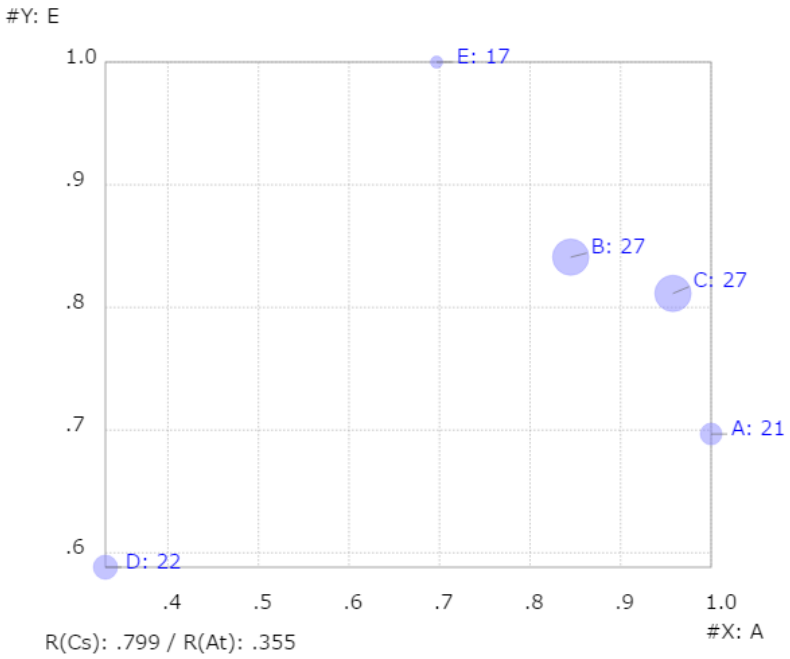
軸の選択の可能性の 1 つとして相関係数の絶対値が最小になる 2 つの軸を考えましょう。そこで、先の正規化近接行列(ND)の列を相関行列 Co(RP)を計算すると次のようになるので、全体の相関係数の中から最もゼロ(0)

に近い値を探すと Co(A, E) = .355 が見つかります[41]。

Co(RP)	A	B	C	D	E
A	1.000	.876	.978	-.980	.355
B	.876	1.000	.956	-.898	.661
C	.978	.956	1.000	-.970	.498
D	-.980	-.898	-.970	1.000	-.430
E	.355	.661	.498	-.430	1.000

よって、次のように RP(A, E)のプロットがすべての組み合わせ RP(X, Y)の中で最もデータの分布を差異化します。

RP(A, E)	X: A	Y: E
A	1.000	.697
B	.845	.841
C	.958	.811
D	.331	.588
E	.697	1.000



【図-5】

この 2 軸 A:E による【図-5】は先の 2 軸 A:B による【図-4】と大きく異なります。距離行列や近接行列はそれぞれの列の位置の遠近を見るので、この視点の違いによって距離の関係が大きく変わるためです。

また、【図-4】と【図-5】は先の相関行列による選択軸分析の結果とも異なります。相関行列は列の「動きの類似度」を示し、距離行列は列の「位

[41] ここで先の CS(B, D)の逆相関の強さ(-.891)を確認しておきましょう。

置の遠近」を示すので[42]、両者の性質が異なりますから分析の目的によって適した方法を選ばなければなりません。

7.10.3. 個体の選択軸分析

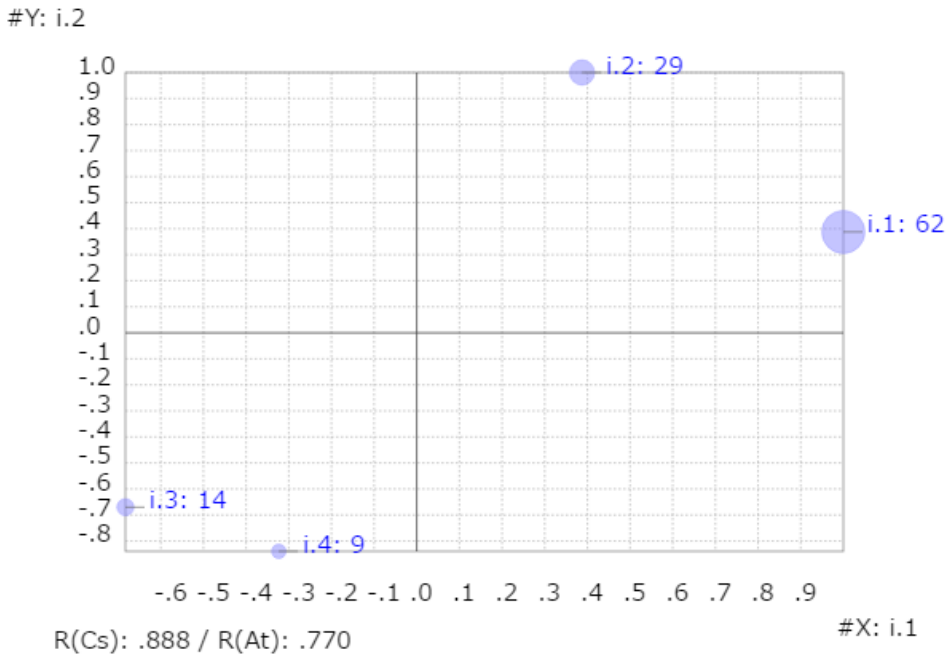
データの属性間の選択軸分析ではなく個体間の選択軸分析を実行するには、分析対象とするデータを転置すれば可能です。下左表はデータ行列(M), 下中表はその転置行列 T(M), そして下右表は転置行列の相関行列 Co(T(M))です。

M	A	B	C	D	E
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

T(M)	h1	h2	h3	h4
A	10	11	0	0
B	19	7	0	1
C	14	10	1	2
D	7	0	12	3
E	12	1	1	3

Co(T(M))	X:h1	Y:h2	h3	h4
h1	1.000	.387	-.683	-.323
h2	.387	1.000	-.670	-.840
h3	-.683	-.670	1.000	.586
h4	-.323	-.840	.586	1.000

次が個体 h1, h2 を選択した選択軸分析の結果です。この結果から個体は大きく h1-h2 と h3-h4 に分けることができ、とくに h3 が h1, h2 から大きく離れていることが確認できます。



【図-6】

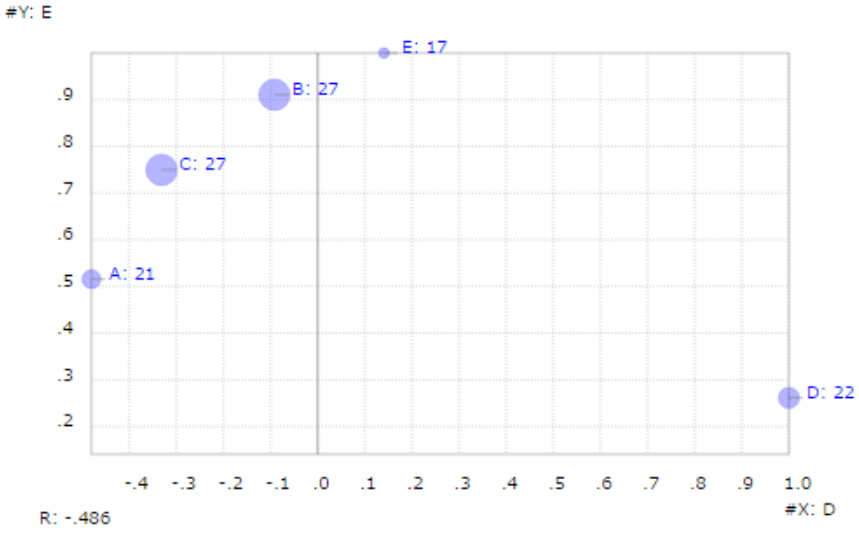
[42] 一方、類似行列は列の「一致度」を示します。

7.10.4. 属性・個体の選択軸分析

次に個体と属性を同じ平面にプロットする方法を考えます。そのために、選択軸分析によって属性の座標を決め、その同じ軸を使って個体の位置(座標)を求めます。次は属性(A, B, ..., E)の最小絶対値相関係数を示す D:E 選択軸分析の結果です。Sn1 は行和を示し、n 行 1 列の行列(縦ベクトル)です。

Mnp	A	B	C	D	E	Sn1
h1	10	19	14	7	12	62
h2	11	7	10	0	1	29
h3	0	0	1	12	1	14
h4	0	1	2	3	3	9
S1p	21	27	27	22	17	114

Co(M)	A	:B	C	D	E
A	1.000	.787	.944	-.480	.436
B	.787	1.000	.945	-.092	.896
C	.944	.945	1.000	-.331	.709
D	-.480	-.092	-.331	1.000	.140
E	.436	.896	.709	.140	1.000



【図-7】

この平面の中に個体 h1, h2, h3, h4 を合理的にプロットする方法を考えます。たとえば h1 は属性(A, B, ..., E)について(10, 19, ..., 12)という値を持っているので、これらの値の重心(平均)の X 座標と Y 座標を求めます。

次表(Ap2)は相関行列 Co(X)の中で選択した X 座標と Y 座標を並置した p 行 2 列の行列です。

Ap2	X: D	Y: E
A	-.480	.436
B	-.092	.896
C	-.331	.709
D	1.000	.140
E	.140	1.000

この行列は属性 A, B, ..., E の位置を示すので、これと h1(10, 19, ..., 12)

の値を考慮して h1 の重心の座標を加重平均を使って決めます[43]。h1 の重心の X 座標 X(h1)と Y 座標 Y(h1)は

X(h1) = (10*-.480 + 19*-.092 + 14*-.331 + 7*1.000 + 12*.140) / 62 = -.040
Y(h1) = (10*.436 + 19*.896 + 14*.709 + 7*.140 + 12*1.000) / 62 = .714

こうして h1 の重心の座標は(-.040, .714)になります。同様にして h2, h3, h4 の X 座標と Y 座標を求めると次の表 Cn2 になります。

Cn2	X:D	Y:E
h.1	-.040	.714
h.2	-.314	.661
h.3	.844	.242
h.4	.296	.637

この n 行 2 列の行列 Cn2 は行列関数 D, X を使えば次の式で求められます。プログラムでは煩雑な行列の積和(X)と割り算(D)のアルゴリズムを避けるためにこの式を用います。

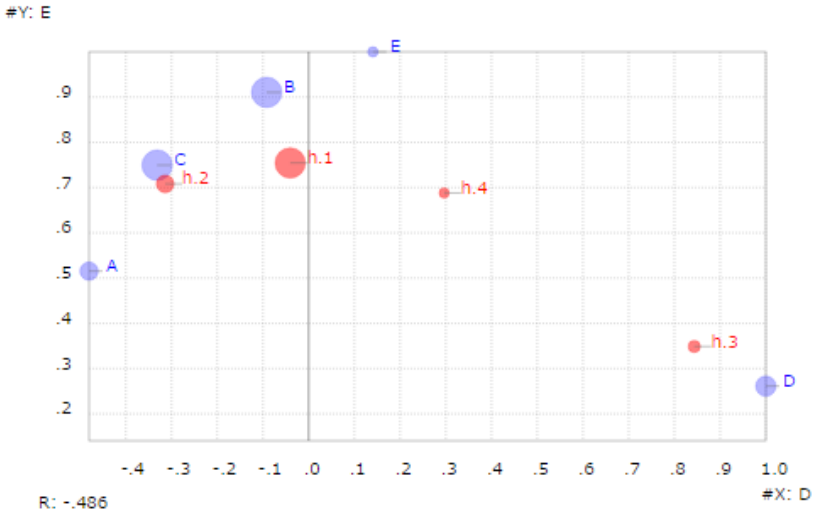
Cn2 = D(X(Mnp, Cp2), X(Mnp, Up1)) = D(Pn2, Sn1)

ここで Mnp は n 行 p 列の入力行列を示し、Cp2 は p 行 2 列の選択軸 Co(M)の属性座標行列を示します。Mnp に Cp2 を右積すると各要素が Mnp の行と Cp2 の列の積和になる Pn2 という n 行 2 列の行列になります。Up1 は p 行 1 列の成分がすべて 1 の行列です。Mnp に Up1 を右積すると行和を示す縦ベクトル Sn1 になります。たとえば D(Xnp, Ynp)は Xnp の要素を Ynp の要素でそれぞれ割ることを示します。D(Pn2, Sn1)の式は 2 列の行列 Pn2 を 1 列の行列 Sn1 で割ることになるので、このような場合は小さな方の行列 Sn1 を大きな方の行列 Pn2 の規模に拡張して Sn2 とし、Pn2 の各要素を Sn2 の各要素で割ります。Sn1→Sn2 という行列の拡張は具体的には次のようになります。

[43] 重心の X 座標を求めるには属性(1.000, .787, …, .436)の単純な平均ではなく、個体のもつ属性値に従って、それぞれの重みを勘案した加重平均値を使います。重心の Y 座標についても同様です。このように重心を使って個体と属性を関連させる方法は東京大学大学院情報学環の倉田博史先生からご教示をいただきました(2017/2/11)。

Sn1	1	→	Sn2	1	2
h1	62		h1	62	62
h2	29		h2	29	29
h3	14		h3	14	14
h4	9		h4	9	9

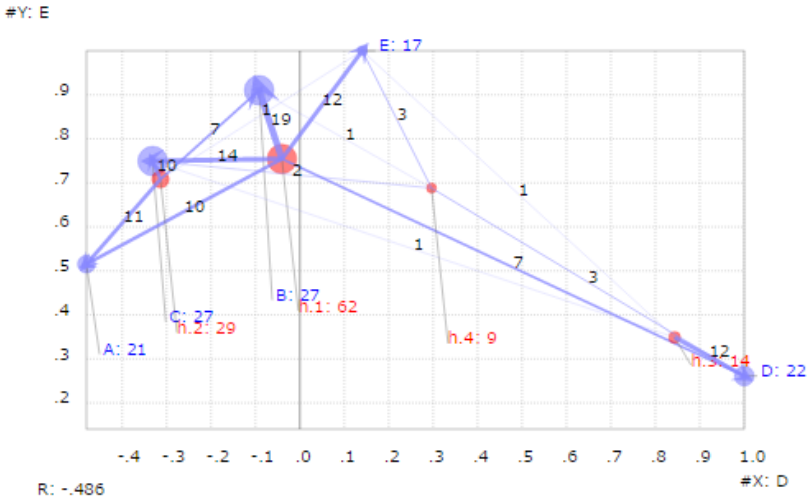
このようにして計算された Cn2 を x, y の座標を使ってプロットすると



【図-8】

たとえば個体 h1 はとくに属性 B に近く、個体 h2 は属性 A, B, C の重心（平均）となっていることがわかります。データ行列(M)を見ると個体 h3 は属性 D に近いのですが(=12)、上図では属性 C, E からもわずかに引かれています。個体 h4 はおおよそ全体の重心に近い位置になります。かりに個体 h5 が(0, 0, 0, 0, 3)という値をもつならば属性 E と一致して同じ位置になります。

個体 h2 の成分は(11, 7, 10, 0, 1)なので属性 C(=10)よりも属性 A(=11)に近くなるはずだと思われるかもしれません。しかし、ここで属性 A, C だけでなく属性全体との関係を考えると、その重心は(.910, .897)になり、上のようなプロット図になります。次の図のように個体と属性を線で結ぶことで、全体の位置関係が明確になります。この図を見ると h.2 が A (11), B (7), C (10)のほぼ重心に位置し、E による影響(1)はほとんどないことがわかります。当然 D による影響(=0)はありません。



【図-9】

このように個体(h1, h2, h3, h4)は、その属性の値の重心（加重平均）とするので、どれも属性(A, B, ..., E)で囲った線（多角形、ここでは 5 角形）の中に入ります。

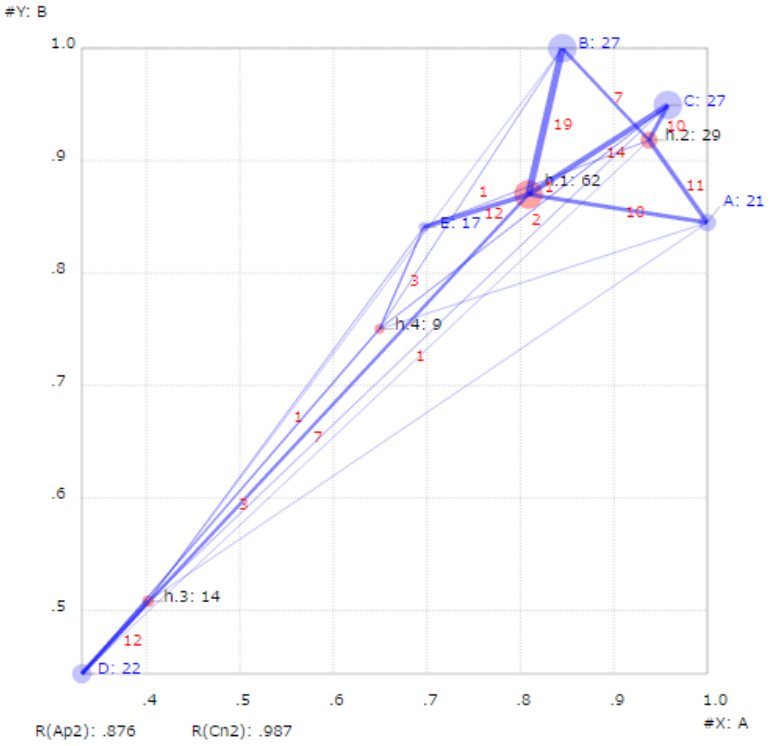
7.10.5. 軸の選択法

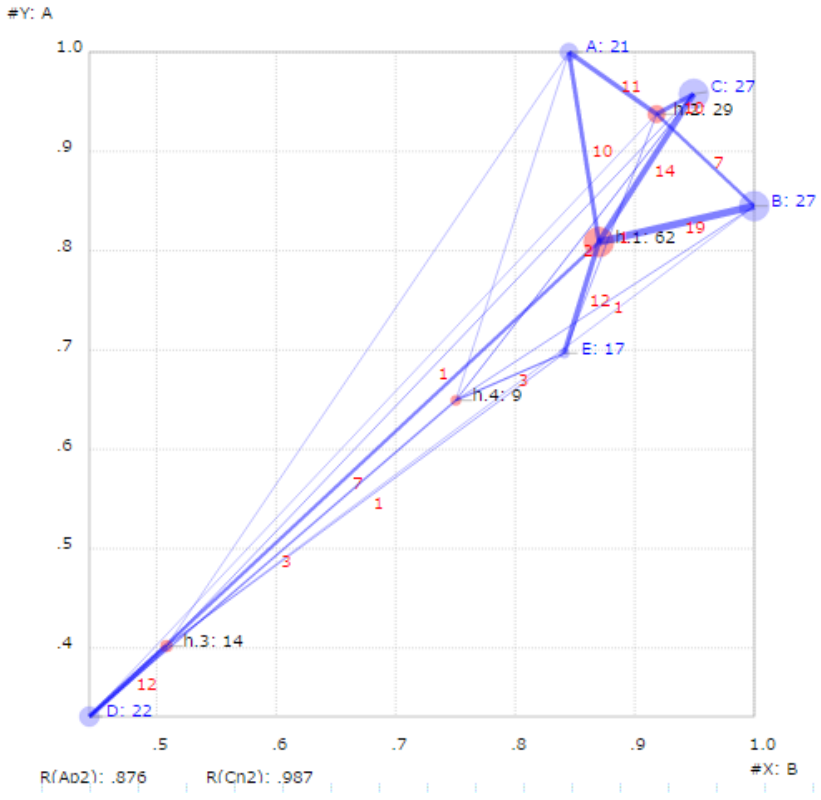
選択軸分析において、X 軸と Y 軸の選択については様々な方法が考えられます。次は入力行列 M とその規定近接係数行列(RP)です。これらを使って考えられる 2 軸の選択法を説明します。

M	A	B	C	D	E
h.1	10	19	14	7	12
h.2	11	7	10	0	1
h.3	0	0	1	12	1
h.	0	1	2	3	3

RP	A	B	C	D	E
A	1.000	.845	.958	.331	.697
B	.845	1.000	.949	.444	.841
C	.958	.949	1.000	.461	.811
D	.331	.444	.461	1.000	.588
E	.697	.841	.811	.588	1.000

(1) 単一自由選択法：たとえば上の RP の A, B をそれぞれ X 軸、Y 軸に選択します。A と B との関係から全体を観察する必要があるときにこの選択法を使います。A:B 以外にも A:C, A:D, A:E, B:C, ...のように様々な 2 軸の選択が可能です。また、A:B は B:A と本質的には同じですが、X 軸と Y 軸が交代するので、次のようにプロット図は対称的に異なります。





(2) 単一最小関係選択法：先に見たように X:A, Y:E の軸で RP の相関係数の絶対値が最小になります。この最小関係 2 軸選択法によって、データのプロットが最大に分散します。（→「近接行列による選択軸分析」）

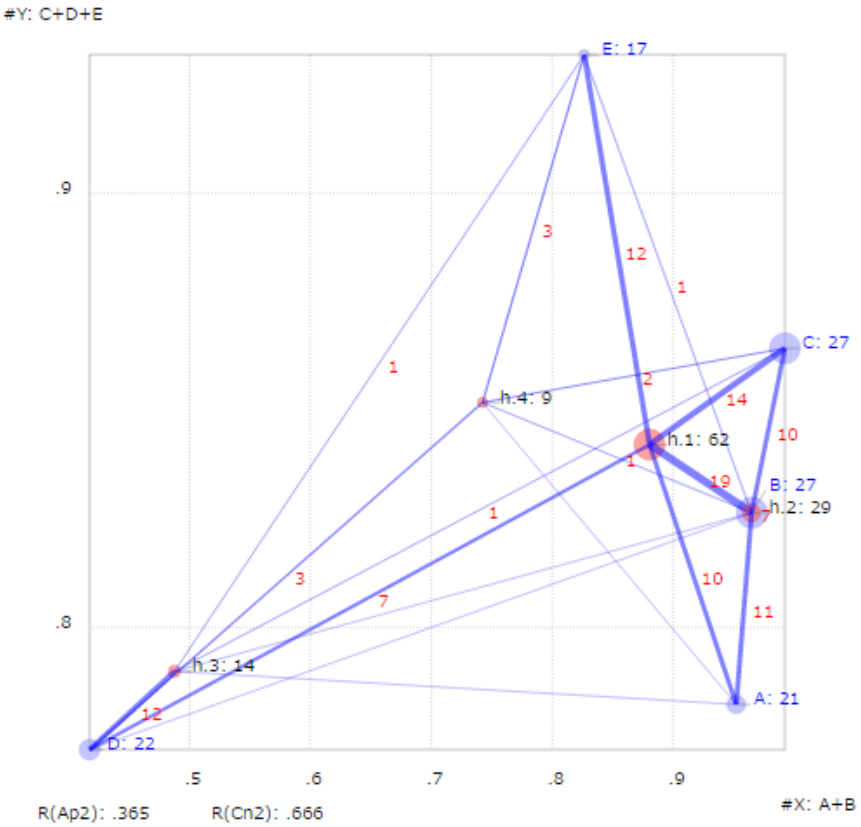
(3) 複合自由選択法：軸の選択において X:A+B, Y:C のように X 軸を A+B

の複合とすることも可能です。そのとき、X は入力行列の平均とします。たとえば h.1.: [A+B]は h1:A=10 と h1:B=19 の平均(10+19)/2 = 14.5 とします。

Cn2	X: A+B	Y: C
h.1	14.500	14.000
h.2	9.000	10.000
h.3	.000	1.000
h.4	.500	2.000

次は X:A+B, Y:C+D+E の選択軸分析の結果です。

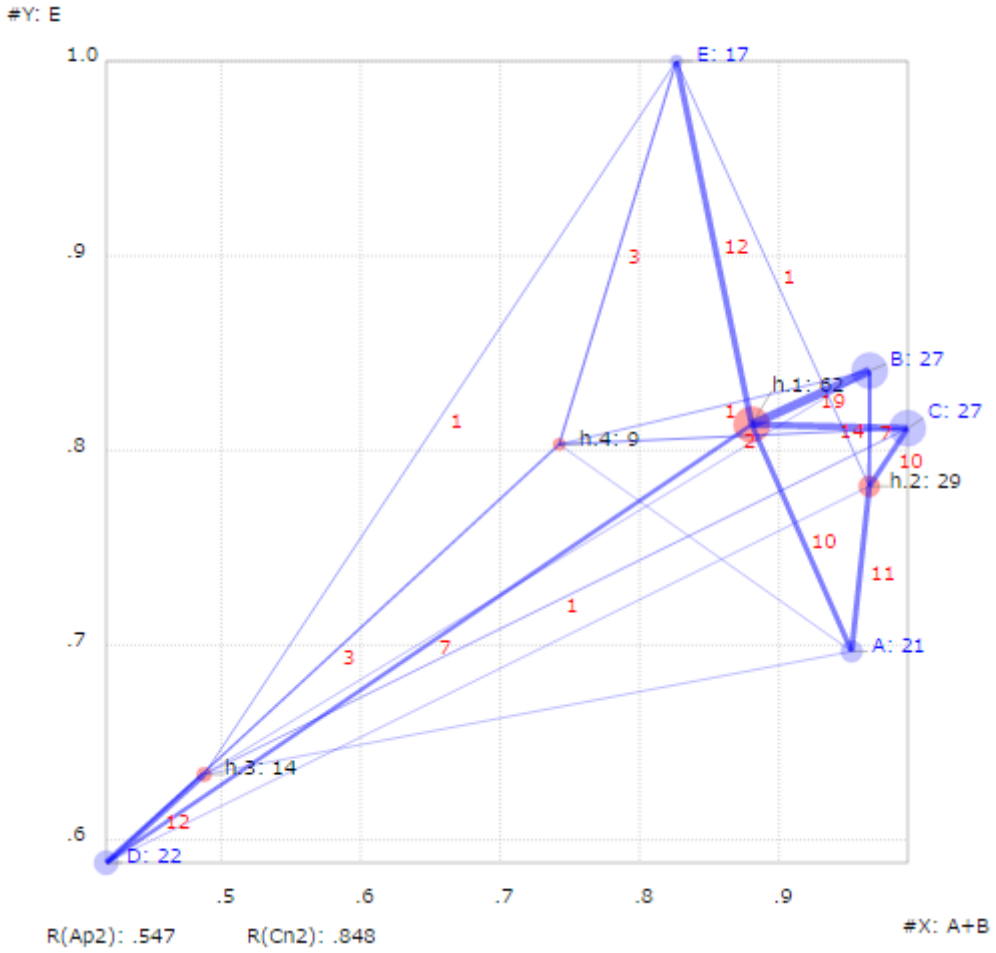
Cn2	X: A+B	Y: C+D+E
h.1	14.500	11.000
h.2	9.000	3.667
h.3	.000	4.667
h.4	.500	2.667



(4) 最小関係対選択法：1 つの軸を（単一・複合）選択し、他方の軸としてその軸と相関係数絶対値が最小を示す対の軸を選択します。次は X:A＋B を選択したときに最小関係対となる Y:E を選択した結果です。

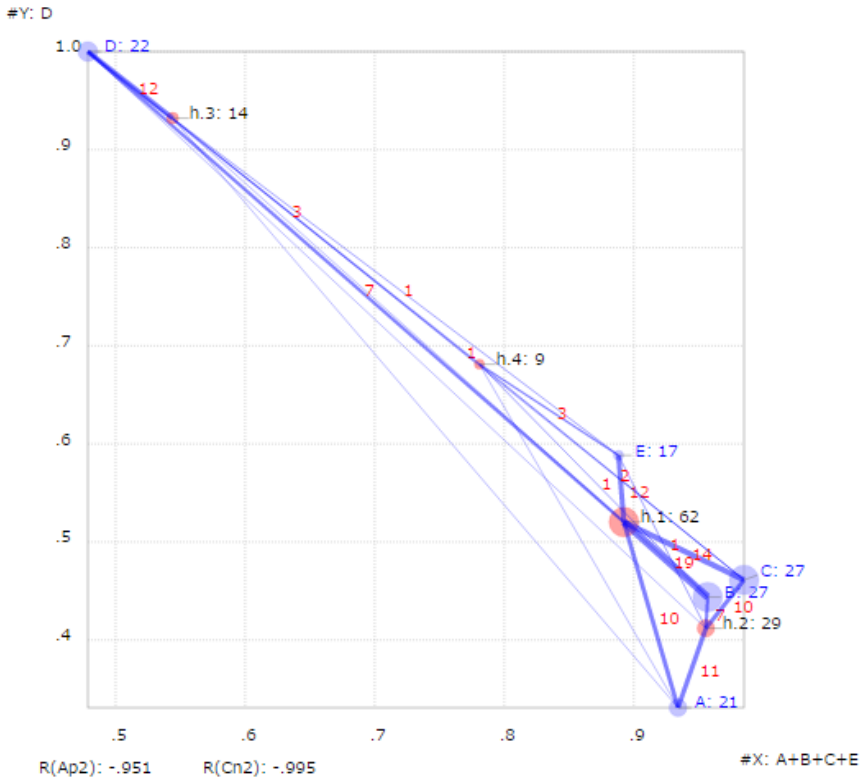
Cn2	A+B	E
h.1	14.500	12.000

h.2	9.000	1.000
h.3	.000	1.000
h.4	.500	3.000



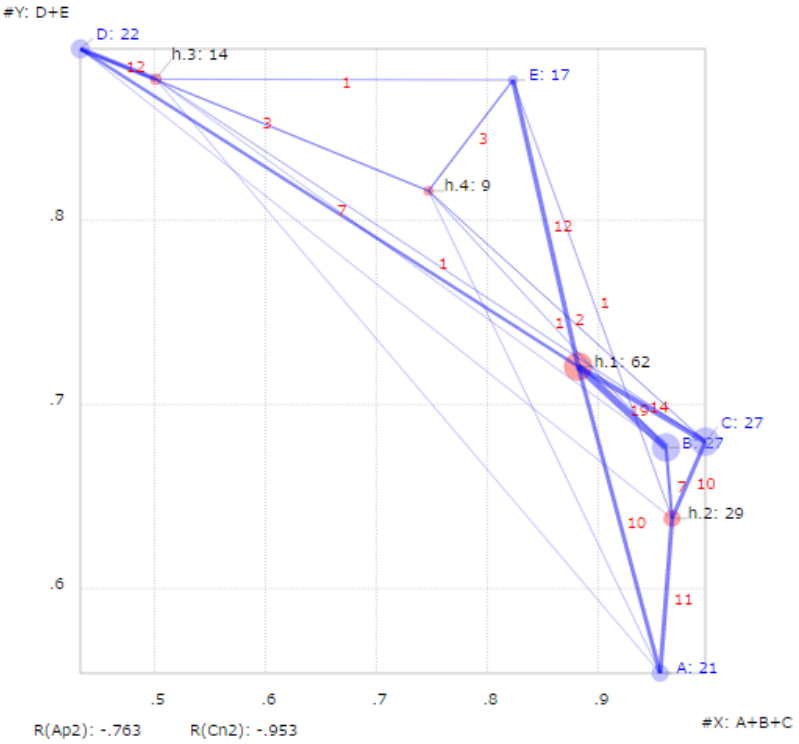
(5) 複合クラスターグループ軸選択法：複合軸として非階層クラスター分析の 2 グループを作る複合列を選択します。

Cn2	A+B+C+E	D
h.1	13.750	7.000
h.2	7.250	.000
h.3	.500	12.000
h.4	1.500	3.000



(6) 複合残余列選択法：X 軸で選択した単一・複合列を除いたすべての列の複合を Y 軸とします。次は X:A+B+C; Y:D+E の例です。

Cn2	A+B+C	D+E
h.1	14.333	9.500
h.2	9.333	.500
h.3	.333	6.500
h.4	1.000	3.000



●多変量解析と選択軸分析

先に見たように多変量解析（主成分分析・対応分析・因子分析）を理解するためには高度な数学的準備（行列の微分、逆行列、固有値問題など）と熟練したプログラミング技術、そして多くの分析の実際的な適用の経験が必要ですが、ここで紹介した選択軸分析の論理はとても単純です[44]。また、主成分分析・対応分析・因子分析の軸の解釈が同じデータであっても分析者の観察や主観によって異なることがあります。また、そもそも軸の解釈が困難になることさえあります。そして、プロット図で選択された 2 軸は全体の固有値・固有ベクトルの中から最大の 2 個だけを使うため情報の消失があって、果たしてそれらが全体を完全に代表しているかどうかについては保証されていません[45]。また、たとえば主成分分析で得られる属性の固有ベクトルに対応する固有値（分散を示す）は分散が最大の第 1 主成分と分散がそれより小さな第 2 主成分を同じように扱って X 軸と Y 軸にすることに根拠があるかどうか疑問を感じます。対応分析・因子分析についても同様です。

一方、この選択軸分析によるプロットの軸には該当する列の名称がそのまま使われるので、そもそも「解釈」する必要がなく、むしろそれぞれの軸が他の軸とどのような関係にあるのかを観察するだけでよいのです。さらに実際の分析において自由に軸を選択・合成できるので、分析者は一定

[44] 最大距離軸・自由選択軸・最小相関軸の方法を実行するためには距離行列と相関行列を作成するプログラムだけを作成するだけで十分です。
[45] 累積寄与率を見ると、どの程度までの情報を利用しているかがわかります。

の方法で出力された結果に縛られることがなく、納得がいくまで実験を重ねることも可能です。複合選択列・クラスター選択列・複合残余列によって合成された軸によってプロットされた図は対象の相関行列・類似行列・近接行列の全情報を使っているので、情報の損失はありません[46]。

しかし、選択軸分析では分析者が自由に軸を選択・合成するので、分析者の恣意（自由の裏返し）が避けられません。さらに選択した行列の種類（相関行列・類似行列・近接行列）や選択した軸の違いによってプロット図は当然大きく異なります。よって分析者はそれらの違いを意識して使用し、発表するときにも明示化して説明できなければなりません[47]。この点で選択軸分析は、入力行列を分析にかければ選択の余地がなく一義的な結果が出力される主成分分析・対応分析・因子分析と異なります[48]。

主成分分析・対応分析・因子分析のセクションで説明したように、それぞれの分析の目的が異なります。したがって、それぞれの分析法の優劣を排他的に論じるよりも、特徴を理解した上で相補的に使用し、結果を比較するとよいでしょう。

7.11. 分布分析

言語データ分析では、言語形式の有無だけでなく、その勢力を見るために各種頻度を調べることが重要です。さらに、頻度が片寄っているか、全体に拡散しているかも考慮されます。その拡散度を見るためには、たとえば同時に調査した各地で得られた資料を比較し、標準偏差や変動係数などが用いられますが、規定化した拡散度も有用です。

しかし、これらの分散に基づく係数は多くの比較材料がなければ計算できません。1 つのテキストまたは同時代・同地域で収集した同質のテキストの集合の中での線状的な分布を調べるときには、たとえば「線状拡散度」(Linear Dispersion: Lin.D.→4.6.1)が役立ちます。これは検索された言語形式の間隔が一定であるか、または片寄りがあるかを調べるために使います。線状拡散度の最初の位置と最後の位置の計算法を単純化した係数を「一様性」(Uniformity: Unif.)と呼びます[49]。

さらに、テキストの全語数をたとえば 10 のブロックに等分割して、それ

---

[46] 一方、個別の選択軸を使うと分析対象の視点を限定され焦点化されます。
[47] 自由には責任が伴います。
[48] 因子分析には様々な方法があるので、選択された方法の中では結果の一義性は確保されますが、方法の選択自体は自由です。
[49] 線状拡散度における最初の要素の位置と最後の要素の位置の計算法は、全体が左右に移動しても逸脱度に影響しないために行った処理ですが、全体が中央にあるほうが分布が平均化している、と考えると、最初の要素は左のスペースの間隔をもち、最後の要素は最後のスペースをもつ、と考えることも可能です。これら 2 つの方法は、とくに要素数が少ないときに、結果が大きく異なります。

ぞれのブロックにおける当該言語形式の頻度間で標準偏差・規定標準偏差・変動係数・拡散度などを計算することによって分布の一様性を調べることができます。ここでは 10 のブロックに等分割して調べた拡散度を「ブロック拡散度」(Block dispersion)と呼びます。線状拡散度はすべての「横の間隔」のバラツキを示しますが、ブロック拡散度はブロックの「縦の量」のバラツキを示します。「10 ブロック拡散度」をここでは簡単に「拡散性」(Dispersion: Disp)と呼びます。

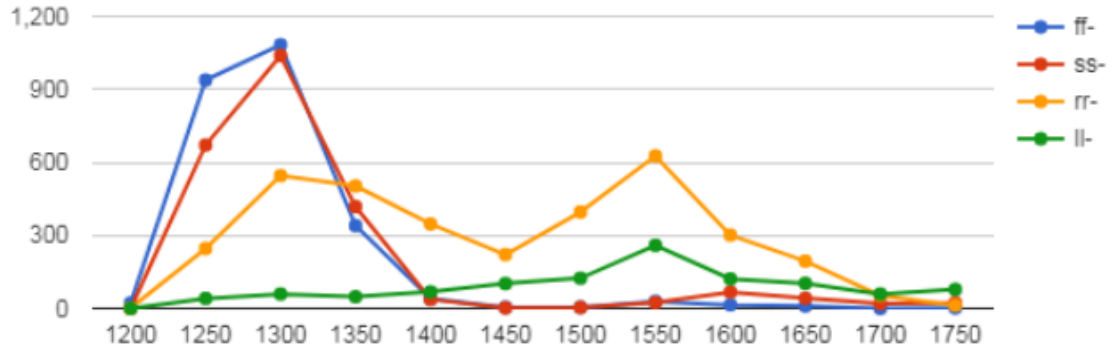
頻度(F)・一様性・拡散性を総合した値を「安定使用度」(Stable Usage: SU)と呼び、次の式で定義します。

$$SU = F * \sqrt{(Unif. * Disp)}$$

安定使用度によって、偏った頻度ではなく、安定した頻度を見るのに役立ちます。

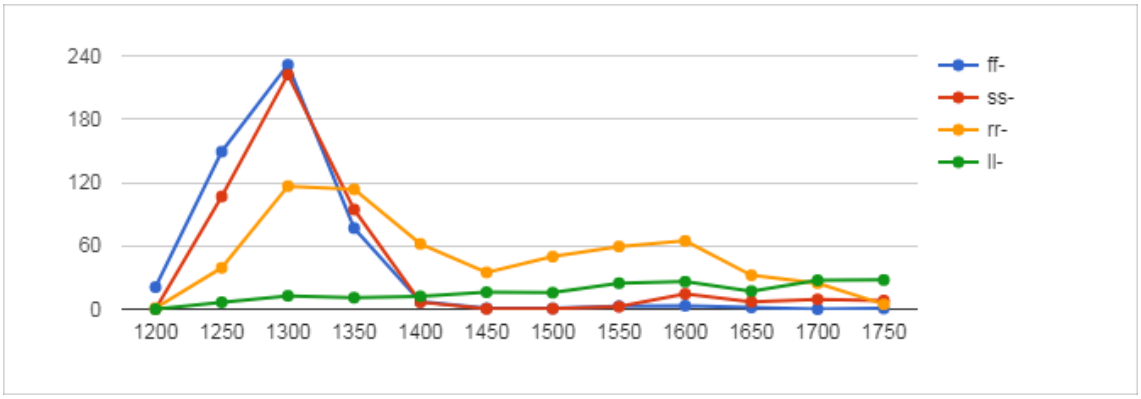
■中世スペイン語の語頭の重子音

→	FA	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	ff-	22	938	1083	340	40	5	5	29	15	10	0	2
2	ss-	0	671	1039	418	37	2	4	24	67	42	20	23
3	rr-	1	246	545	503	347	221	396	626	301	194	53	13
4	ll-	0	41	59	48	68	102	124	259	121	103	59	78

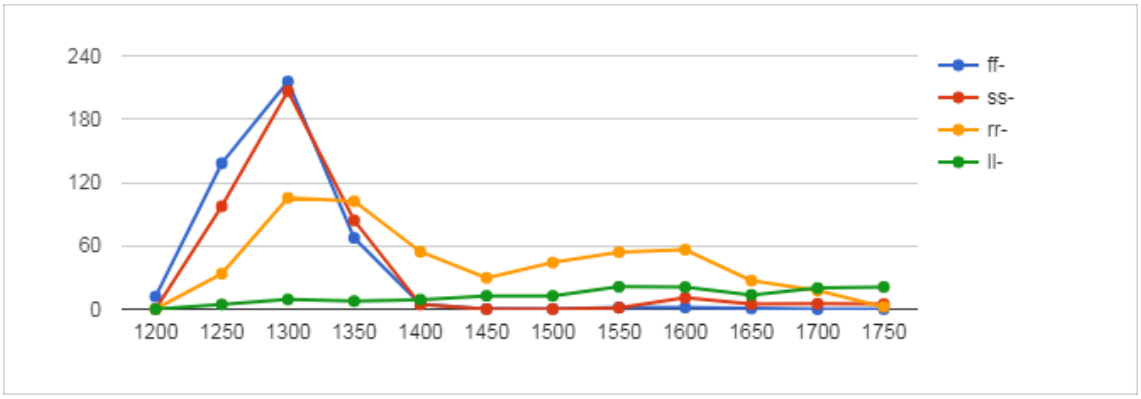


→	Atributo	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	Letra	53063	319211	243054	232732	281720	332140	416799	543404	235297	298462	107174	144919
2	Palabra	10546	62865	46774	44296	56167	63547	79569	105496	46555	60390	21546	28221
3	Documento	24	131	67	44	50	66	109	293	154	76	43	96

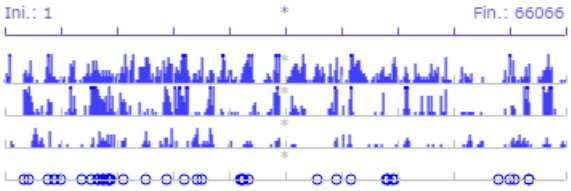
→	FN.PI.:10000	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	ff-	20.9	149.2	231.5	76.8	7.1	0.8	0.6	2.7	3.2	1.7	0.0	0.7
2	ss-	0.0	106.7	222.1	94.4	6.6	0.3	0.5	2.3	14.4	7.0	9.3	8.1
3	rr-	0.9	39.1	116.5	113.6	61.8	34.8	49.8	59.3	64.7	32.1	24.6	4.6
4	ll-	0.0	6.5	12.6	10.8	12.1	16.1	15.6	24.6	26.0	17.1	27.4	27.6



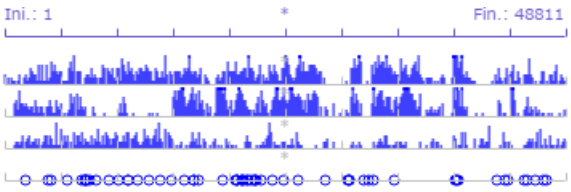
→	FP.PI.:10000	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	ff-	11.9	138.2	215.7	67.5	4.8	0.2	0.2	1.7	1.6	0.7	0.0	0.1
2	ss-	0.0	97.5	206.6	84.0	4.3	0.0	0.1	1.3	10.7	4.7	5.1	4.7
3	rr-	0.0	33.6	105.3	102.2	54.4	29.6	44.2	54.0	56.4	27.1	17.6	2.2
4	ll-	0.0	4.4	9.2	7.6	9.0	12.6	12.6	21.2	20.9	13.4	20.0	21.0



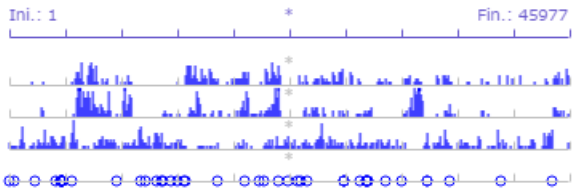
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1250	ff-	27.2	.885	.490	17.9
1250	ss-	19.2	.833	.332	10.1
1250	rr-	6.6	.850	.396	3.8
1250	ll-	.9	.722	.518	.6



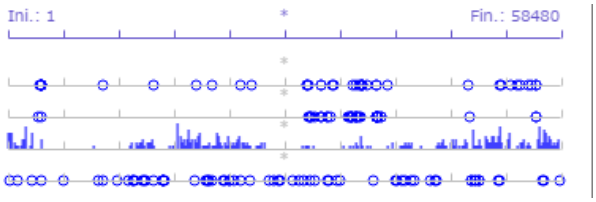
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1300	ff-	41.5	.935	.550	29.8
1300	ss-	39.7	.828	.457	24.4
1300	rr-	20.3	.832	.519	13.3
1300	ll-	1.8	.778	.601	1.2



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1350	ff-	12.8	.777	.390	7.0
1350	ss-	16.0	.727	.336	7.9
1350	rr-	19.4	.898	.554	13.7
1350	ll-	1.4	.783	.529	.9



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1400	ff-	1.0	.719	.452	.6
1400	ss-	.9	.392	.245	.3
1400	rr-	10.8	.764	.447	6.3
1400	ll-	1.8	.841	.565	1.2



7.12. 一般性・特殊性分析

次のサンプルデータのそれぞれの変数(v1:v5)が他の全部の変数と共通する度合いを、その「一般性」と名づけ、その補数を「特殊性」とします。はじめに、一般性の度合いを計算する方法を説明します。

G	v1	v2	v3	v4	v5
d1	1	1	1	0	0
d2	1	1	0	1	1
d3	0	1	0	1	0
d4	1	0	1	1	1

H	sumH
d1	3
d2	4
d3	2
d4	4

それぞれの変数がもつ、他の変数全体との共通性を計算するために、次のような縦和(sumV)は役立ちません。個別の変数の性質をとらえているだけで、他の変数については考慮していないからです。

V	v1	v2	v3	v4	v5
SumV	3	3	2	3	2

そこで、上表(G)の横和(sumH)がそれぞれの行の情報をもっているので、これを利用します。入力行列に行和(3, 4, 2, 4)ベクトル(H)から、入力行列(G)を引くと、次の行列(H-G)になります。

sumH(Gnp) - Gnp

H-G	v1	v2	v3	v4	v5
d1	2	2	2	3	3
d2	3	3	4	3	3
d3	2	1	2	1	2
d4	3	4	3	3	3

この行列(H-G)のそれぞれのセルは、その行にある他の列の和を示します。たとえば、[d1:v1]の 2 は入力行列(G)の[d1:v2]から[d1:v5]の数値の和を示します。

次に、入力行列(G)と(H-G)を掛けると、次の行列(Xnp)になります（行列

積ではなく要素積です→「行列」）。

Xnp = Inp * (sumH(Gnp) - Gnp)

Xnp	v1	v2	v3	v4	v5
d1	2	2	2	0	0
d2	3	3	0	3	3
d3	0	1	0	1	0
d4	3	0	3	3	3

このようにすると、入力行列(Inp)で 1 があるセルの数値が、新行列(Xnp)で Gmp の横和 - Gnp の値に変わりました。それぞれのセルで、この数値が多ければ多いほど、その数値の一般性が高い、ということになります。次が Xnp の縦和を「絶対値による一般性」(Generality by absolute value: G.a.val.)とよびます。その総和が H2 です。

G.a.val.	v1	v2	v3	v4	v5
Sum	8	6	5	7	6

H2	Suma
Suma	32

このベクトル(8, 6, 5, 7, 6)をその総和(32)で割ると、新行列(Xnp)の縦和の相対値になります。

G.r.val.	v1	v2	v3	v4	v5
M*H/H2	.250	.188	.156	.219	.188

これを「相対値による一般性」(Generality by relative value: G. r.val.)とします。その補数が「相対値による特殊性」(Peculiarity by relative value: P.r.val.)です。

P.r.val. = 1 - G.r.val.

P.r.val..	v1	v2	v3	v4	v5
1-G.r.val.	.750	.813	.844	.781	.813

先に見たように(→「得点」)、相対値は変数の数が多くなると一般に非常に小さな数値になるので、評価がしにくくなります。逆に「相対値による特殊性」は非常に大きな値になって、やはり扱いが困難です。そこで、次のように相対値を卓立化（→「相対得点」）して、「卓立化相対値による一般性」(Generality by prominent relative value: G.p.r.val.)と「卓立化相対値による特殊性」(Peculiarity by prominent relative value: P.p.r.val.)を計算します。

p.r.val.G.	v1	v2	v3	v4	v5
Suma	.571	.480	.426	.528	.480

P.p.r.val. = 1 - G.p.r.val.

p.r.val.P.	v1	v2	v3	v4	v5
Suma	.429	.520	.574	.472	.520

以上をまとめて、「一般性・特殊性分析」の結果の全体(Generality)を示します。

Generality	v1	v2	v3	v4	v5
Sum	3	3	2	3	2
G.a.val.	8	6	5	7	6
G.r.val.	.250	.188	.156	.219	.188
P.r.val.	.750	.813	.844	.781	.813
G.p.r.val.	.571	.480	.426	.528	.480
P.p.r.val.	.429	.520	.574	.472	.520

なお、変数ではなく個体についての「一般性・特殊性分析」をするためには、はじめに入力行列を転置してから同じプログラムを実行します。

Tr(M)	d1	d2	d3	d4
v1	1	1	0	1
v2	1	1	1	0
v3	1	0	0	1
v4	0	1	1	1
v5	0	1	0	1

Generality	d1	d2	d3	d4
Sum	3	4	2	4
G.a.val.	5	7	4	6
G.r.val.	.227	.318	.182	.273
P.r.val.	.773	.682	.818	.727
G.p.r.val.	.469	.583	.400	.529
P.p.r.val.	.531	.417	.600	.471

●値データと小数値データの一般性・特殊性

ここで提案した一般性・特殊性の計算は、次のような値データ行列(Fnp)についても、そのまま適用できます。先の(0, 1)データでは、他の数値の和に 0 または 1 を掛け合わせただけでしたが、ここでは、Fnp のそれぞれの数値が 0 以上の値をもつので、それが一般性・特殊性の大きさに影響します。

F	v1	v2	v3	v4	v5
d1	2	3	4	0	0
d2	1	2	0	5	2
d3	0	4	0	4	0
d4	1	0	3	1	1

G.a.val. = Fnp * [sumH(Fnp) - Inp]

Generality	v1	v2	v3	v4	v5
Sum	4	9	7	10	3
G.a.val.	28	50	29	46	21
G.r.val.	.161	.287	.167	.264	.121
P.r.val.	.839	.713	.833	.736	.879
G.p.r.val.	.434	.617	.444	.590	.354
P.p.r.val.	.566	.383	.556	.410	.646

さらに、次の相対得点などのような小数のデータであっても、非負値であれば、同様にして、その一般性・特殊性を計算することができます。

RSv	v1	v2	v3	v4	v5
d1	.500	.333	.571	.000	.000
d2	.250	.222	.000	.500	.667
d3	.000	.444	.000	.400	.000
d4	.250	.000	.429	.100	.333

Generality	v1	v2	v3	v4	v5
Sum	1.000	1.000	1.000	1.000	1.000
G.a.val.	1.015	.850	.769	.848	.908
G.r.val.	.231	.194	.175	.193	.207
P.r.val.	.769	.806	.825	.807	.793
G.p.r.val.	.546	.490	.459	.489	.510
P.p.r.val.	.454	.510	.541	.511	.490

■スペイン語語彙バリエーションの一般性・特殊性分析

世界の 21 か国で使用されるスペイン語の語彙のバリエーションを調査し、それぞれの国で使われる語形変異体を集計した結果、9886 行のデータ行列が得られました。

CPT	EXPLICA	INGLÉS	Forma	ES	GE	CU	RD	PR	MX	GU	HO	EL	NI	CR	PN	CO	VE	EC	PE	BO	CH	PA	UR	AR	Suma	Año
A001	Prenda de	Jacket	gabán					1																	1	2015
A001	Prenda de	Jacket	vestón																		1				1	2015
A002	Prenda de	Cardigan	camperita																					1	1	2016
A002	Prenda de	Cardigan	chaleco																		1				1	2015
A002	Prenda de	Cardigan	jompa						1																1	2016
A002	Prenda de	Cardigan	rebeca	1																					1	2015
A003	Prenda qu	T-shirt	camisa					1																	1	2015
A003	Prenda qu	T-shirt	pulóver			1																			1	2015
A004	Prenda de	Sweater	chomba																		1				1	2015
A004	Prenda de	Sweater	enguatada			1																			1	2015
A004	Prenda de	Sweater	saco													1									1	2015
A004	Prenda de	Sweater	saco de lana														1								1	2015
A004	Prenda de	Sweater	sudadera					1																	1	2015
A004	Prenda de	Sweater	tricota																			1			1	2015
A005	Chaqueta	Windbreal	chompa															1							1	2015
A005	Chaqueta	Windbreal	chupa	1																					1	2015
A005	Chaqueta	Windbreal	jompa						1																1	2016
A005	Chaqueta	Windbreal	parca																		1				1	2015
A006	Traje usad	Overalls	braga														1								1	2015
A006	Traje usad	Overalls	enterizo			1																			1	2015
A006	Traje usad	Overalls	jardinero																					1	1	2015
A006	Traje usad	Overalls	mecánico			1																			1	2015
A007	Prenda de	Poncho	jorongo						1																1	2015
A007	Prenda de	Poncho	manta																		1				1	2015

http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/ (2016/2/25)

次がその分析結果です。

Generality	ES	GE	CU	RD	PR	MX	GU	HO	EL	NI	CR	PN
Sum	4829	1645	4210	2203	3514	4253	1918	2575	1381	3715	986	2179
a.freq.G.	29547	18870	36195	24438	31971	36053	21711	26488	16703	34355	12603	23170
r.freq.G.	.057	.037	.070	.048	.062	.070	.042	.052	.033	.067	.025	.045
r.freq.P.	.943	.963	.930	.952	.938	.930	.958	.948	.967	.933	.975	.955
p.r.freq.G.	.550	.433	.602	.500	.570	.601	.469	.521	.402	.589	.335	.486
p.r.freq.P.	.450	.567	.398	.500	.430	.399	.531	.479	.598	.411	.665	.514

CO	VE	EC	PE	BO	CH	PA	UR	AR
835	2635	2253	1827	1793	2336	3025	1803	3542
10143	27918	24348	20550	20987	16476	29472	19893	31978
.020	.054	.047	.040	.041	.032	.057	.039	.062
.980	.946	.953	.960	.959	.968	.943	.961	.938
.287	.535	.499	.454	.460	.398	.549	.446	.570
.713	.465	.501	.546	.540	.602	.451	.554	.430

先に述べたように、入力行列列和(Sum)はスペイン語圏のそれぞれの国の一般性・特殊性を示すためには役立ちません。行和加重得点の列和（絶対頻度）によって、一般性・特殊性の大小関係を察知することができますが、同じスケールで変数を比較するためには相対頻度を使わなければなりません(r.freq.G.)。しかし、この相対頻度による一般性の数値は変数が多いため(21 か国)、どれも非常に小さくて比較しにくくなっています。（逆に特殊性(r.freq.P.)は全体が大きくなります。）そこで、行和加重得点の卓立化相対頻度を使った一般性・特殊性を下の 2 行(p.r.freq.G, p.r.freq.P.)に示しました。これら 2 つの数値が変数間の比較に適しています。

次に、諸国間の大小関係を見るために順位得点（降順）を出力します。「一般性」を示す 3 つの指数の順番はすべて同じになり、その「特殊性」の指数はその逆順になります。しかし、列和(Sum)は「一般性・特殊性」に関

係しないので順位が異なっています。

DRSh	ES	GE	CU	RD	PR	MX	GU	HO	EL	NI	CR	PN
Sum	1	18	3	12	6	2	14	9	19	4	20	13
a.freq.G.	6	17	1	10	5	2	13	9	18	3	20	12
r.freq.G.	6	17	1	10	5	2	13	9	18	3	20	12
r.freq.P.	16	5	21	12	17	20	9	13	4	19	2	10
p.r.freq.G.	6	17	1	10	5	2	13	9	18	3	20	12
p.r.freq.P.	16	5	21	12	17	20	9	13	4	19	2	10

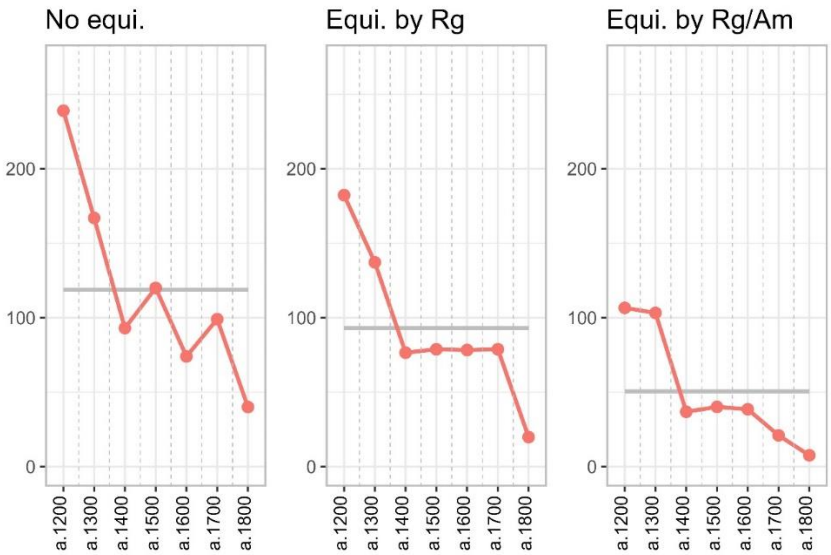
CO	VE	EC	PE	BO	CH	PA	UR	AR
21	8	11	15	17	10	7	16	5
21	8	11	15	14	19	7	16	4
21	8	11	15	14	19	7	16	4
1	14	11	7	8	3	15	6	18
21	8	11	15	14	19	7	16	4
1	14	11	7	8	3	15	6	18

7.13. 均等分析

次の表は「動詞活用形＋代名詞」(enclisis, 例：detúvose)の歴史的分布を示しています(sf:安全頻度, M.mean:安全頻度の平均値)。

a. No.Eq..	sf	b. Eq..Rg.	M.mean	c. Eq..Rg.Am.	M.mean
a.1200	239	a.1200	182.3	a.1200	106.6
a.1300	167	a.1300	137.2	a.1300	103.2
a.1400	93	a.1400	76.5	a.1400	36.8
a.1500	120	a.1500	78.8	a.1500	40.1
a.1600	74	a.1600	78.3	a.1600	38.5
a.1700	99	a.1700	78.8	a.1700	21
a.1800	40	a.1800	19.8	a.1800	7.6

【表-1(a,b,c)】「動詞活用形＋代名詞」の歴史的分布



【図-1】(a,b,c)「動詞活用形＋代名詞」の歴史的分布

ここで，a.1200 の高い頻度(=239)が示しているのは，歴史的な a.1200 の特徴ではなく，むしろ一定の地域的・社会文体的データの特徴である可能性があります。そのことを以下で確認します。

次の表は「動詞活用形＋代名詞」の地理・歴史的分布を示しています(安全頻度)。列が歴史的変化(a.1200-a.1800)を表し，行が地理的変異に表しています(AN: Andalucía, AR: Aragón, CN: Castilla la Nueva, CV: Castilla la Vieja, EX: Extremadura, LE: León) (→***)。

*	a.1200	a.1300	a.1400	a.1500	a.1600	a.1700	a.1800
AN	203	61	54	72	104	83	50
AR	14	40	42	20	133	73	7
CN	250	187	86	136	90	95	43
CV	265	167	117	75	72	36	16
EX	99	120	92	137	33	66	3
LE	173	267	71	53	37	134	19
Equi	-0.287	-0.047	-0.07	0.256	-0.092	0.139	0.384
Median	188	143.5	78.5	73.5	81	78	17.5
Mean	167.3	140.3	77	82.2	78.2	81.2	23
M.mean	182.3	137.2	76.5	78.8	78.3	78.8	19.8

【表-2】「動詞活用形＋代名詞」の地理・歴史的分布(安全頻度)
(AN: Andalucía, AR: Aragón, CN: Castilla la Nueva, CV: Castilla la Vieja, EX: Extremadura, LE: León); a.1200 (1200-1299), ..., a.1800 (1800-1899).

この表を見ると，たとえば a.1200 年代の大数平均値(M.mean=182.3)は，

先に見た a.1200 の安全頻度(=239)よりも低い値ですが，むしろこの数値(大数平均値)が地域間の頻度差を均等に考慮して得られた数値なので，a.1200 年代の状況を正しく表示していると考えます。なぜならば，上の表の数値はどれもそれぞれの年代・地域で同じ方法で計算された安全頻度なのでそのまま比較が可能だからです。その大数平均値ならば全地域を均等に評価して得られた数値になるはずです(平均値よりも大数平均値のほうが正確です→***)。一方，こうした地域差をまったく考慮しないで，年代だけの範囲で安全頻度を計算すると，その数値は年代の特徴というよりも，高頻度の地域の特徴を示していることになります(表-1)。この問題の解決法として，ここで提案した方法を「均等分析」(英:equilibrated analysis，西:análisis equilibrado)と呼びます。実際に，均等分析をしないで地域差を無視した各年代の安全頻度と，地域差を同等に考慮した均等分析による安全頻度を表-1(a,b)と図-1(a,b)で比べると，その結果はかなり大きく異なっています。

さらに，地域差だけでなく社会文体差も勘案するとよいでしょう。次の表は，縦軸に地域：社会文体にして集計した結果を示します[50]。

*	a.1200	a.1300	a.1400	a.1500	a.1600	a.1700	a.1800
AN:C	219	24	28	65	NA	NA	NA
AN:E	8	28	14	20	62	NA	11
AN:J	NA	336	NA	19	112	60	0
AN:M	0	0	NA	33	16	0	23
AN:P	NA	11	68	63	167	89	47
AR:C	NA	0	NA	NA	5	NA	NA
AR:E	8	19	40	13	0	NA	0
AR:J	NA	82	0	0	289	108	NA
AR:M	6	0	0	NA	NA	0	0
AR:P	4	43	34	16	48	23	8
CN:C	322	230	97	70	21	0	0
CN:E	140	31	14	12	91	0	NA
CN:J	0	149	61	161	94	115	15
CN:M	NA	145	76	71	0	0	35
CN:P	245	214	57	173	90	80	36
CV:C	386	220	159	53	NA	NA	NA
CV:E	175	129	62	87	80	0	0
CV:J	120	235	NA	34	57	20	8
CV:M	316	0	0	66	0	0	0
CV:P	167	110	33	51	22	34	13
EX:C	85	118	NA	NA	NA	NA	NA

[50] NA (not available)は資料に存在しない(採録されなかった)ケースです。

EX:E	NA	62	27	48	3	0	0
EX:J	NA	37	21	25	0	14	NA
EX:M	0	447	0	0	0	0	4
EX:P	60	54	118	154	34	73	0
LE:C	79	214	9	7	NA	NA	NA
LE:E	191	174	33	24	20	0	16
LE:J	346	NA	41	10	NA	0	7
LE:M	0	NA	11	0	15	0	5
LE:P	55	337	117	72	20	138	8
Equi	0.406	0.285	0.382	0.449	0.698	1.000	0.348
Mean	127.5	123.2	44.8	49.9	51.9	32.8	10.7
Median	85.0	96.0	33.0	34.0	21.5	0.0	7.5
M.mean	106.5	103.1	36.7	40	38.4	20.4	7.3

【表-3】「動詞活用形＋代名詞」の地域：社会文体・歴史的分布(安全頻度)
(C: documento cancillerezco, E: documento esclesiástico, J: documento judicial, M: documento municipal, P: documento particular)

この表の M.mean の行が地域差：社会文体差を考慮した均等分析の結果を示しています。以上をまとめて，再度，表-1, 図-1 を見直すと，３者の大きな傾向として下降線が見られますが，やはり，それぞれがかなり異なる曲線を辿っていることがわかります。

***

ここで扱ったことは数量的変異を示すクロス集計表に一般に見られる問題です。クロス集計表の縦軸 r(1),r(2),...,r(n)と横軸 c(1),c(2),...,c(p)が交差する位置 r(i):c(j)に比較的大きな数値があるとき,それが r(i)に因るものか, c(j)によるものか，わからないことが多いのです。実際は r(i)と c(j)が「共起」したときの数値だからです。

たとえば，次の【表-4】(「動詞活用形＋代名詞」の頻度)の LE:a.1700 に比較的大きな数値があります(=134)。この中で CV:1.1200 の数値(=314)が特出しています。この列の平均値(83.5)は CV の外れ値が大きく影響しています。さらに，その平均値が a.1200 の特徴であるのか，または CV の特徴であるのかがわかりません。

*	a.1200	a.1300	a.1400	a.1500	a.1600	a.1700	a.1800
AN	60	12	15	33	24	73	20
AR	4	20	30	6	29	23	2
CN	58	82	51	226	146	82	21
CV	314	121	111	58	18	11	6
EX	4	36	20	146	56	11	1
LE	61	99	42	15	6	40	5

【表-4】「動詞活用形＋代名詞」の頻度

一般のアンケート調査では，はじめから分類をしないでランダム化(randomization)した多数のサンプルを使用します(伊藤 2017, 中室・津川 2017)。そうすればサンプルの偏り(たとえば地方差)の影響がなくなることが期待されます。しかし,そのときはデータが多数でなければなりません。少数だとランダム化したサンプルに偏りができてしまうからです。また，多数のサンプルであっても，人口が大きな大都市や地方のサンプルに偏ったデータになります。

統計数理研究所が 1953 年から 5 年ごとに継続している「日本人の国民性調査」では，面接の対象者を抽出するのに「層化多段抽出法」というサンプリング方法が用いられています。その手順を次に引用します。「まず全国の市区町村を人口規模によって 6 つの層に分け、それぞれから 300～400 程度の町丁字を人口に比例する確率によって選ぶ。さらにそれぞれに割り当てた人数を住民基本台帳から等間隔に抜き出す。こうして選んだ人たちの自宅を調査員が訪ね、面接調査を実行する。」(統計数理研究所 2020: 6-7)。このサンプリング法の具体的な手順は次のように説明されています：「全国の市区町村を、区部・人口 20 万人以上の市部・人口 10 万人以上の市部・人口 10 万人未満の市部・郡部・沖縄県の 6 つに層化しました。次に、各層から合計 400 町丁字等を確率比例抽出で選びました。最後に、抽出した町丁字等の住民基本台帳から、その地点に割り当てた人数（平均 16）の標本を等間隔抽出法で選びました。」[51]ここで最初に設定した 6 つの層の人口規模が異なることに注目します。規模が異なる 6 つの層からさらに小さな単位の町丁字等が選ばれています。このような方法を採れば，一般のランダム化サンプリングで起こりやすい大都市の偏りが防げます。

しかし，私たちの歴史資料では現存するデータしか入手できず，アンケート調査のようにサンプルをコントロールすることは困難です。たしかに現存するデータを地方ごとにランダム化サンプリングをすることも考えられますが，上の【表-4】が示すように，サンプルの数はそれほど大きくはありません(これは多くの言語資料の特徴です)。少ないデータからサンプ

[51] https://www.ism.ac.jp/kokuminsei/page9/page10/index.html

リングすると，偏りが生じることは避けられません。そこで，人口規模を勘案してその偏りを排除した層化多段抽出法のように，規模の異なる地域や社会文体ごとに計算した安全頻度を平均化することで地域と社会文体の偏りを回避します。私たちの方法ではサンプリングするのではなく，それぞれのブロックの全数を調査します。

このようにして，すべての頻度と母数から計算した安全頻度の大数平均値(【表-2】【表-3】)は，特定の地方が影響した数値ではなく，どの地方も均等に同じ基準と方法で処理されているので，それぞれの年代の推移を正しく表示していると考えられます。そのとき，外れ値が大きく影響する平均値ではなく，外れ値の影響を少なくしながら外れ値の情報を含める大数平均値(M.mean; 英:majority mean, 西:media de mayoría→***)を使用します。

参考：

伊藤公一郎. 2017.『データ分析の力 因果関係に迫る思考法』光文社.
中室牧子・津川友介. 2017.『「原因と結果」の経済学』ダイヤモンド社.
統計数理研究所.2020『響き合う人とデータ.統数研プロジェクト紹介』統計数理研究所.
https://www.ism.ac.jp/kouhou/project/index.html