対応分析のグラフを適切に解釈する条件

Standard Coordinate, Principal Coordinate を理解する

藤本　一男

概要

本稿は、対応分析 (Correspondence Analysis) の出力を適切に評価するために、Standard Coordinate（標準座標）、Principal Coordinate（主座標）という尺度についての理解が必要であることを述べる。対応分析は、クロス表の行変数と列変数の関係を同時布置（対称マップ）として表現できることを特徴としている。しかし、行変数内、または列変数内のカテゴリーポイント間の距離は、数理的に定義されているが、行変数と列変数の間の距離は定義されていない。つまり、この同時布置は不正確なのである。この関係の理解しにくさが､対称マップの見た目から解釈することが蔓延することを後押ししている。

だが、この一見、わかりにくい関係を理解することが、同時布置（対称マップ）の有効性を含めて対応分析のパワーを引きだすためには必要なのである。

1　問題の所在

パソコンの高性能化、R による高機能な統計処理環境の普及によって、対応分析のような高度な多変量解析も身近なものになっている。しかし、身近（簡単な操作で結果が手に入る）になったからといって、誰でもがその手法を使いこなすことができるわけでもなく、結果に対して誤った解釈をしてしまう危険性もある[注1)]。

1　これは、Excel の普及によって、文字としての数値（つまりカテゴリー）が、見た目数値を数値として演算が可能になり、西里たちが、警鐘をならしつづけているリッカート・スケールのアプリオリ

対応分析においては、この危険性は繰り返し指摘されてきながら（後に列挙）、そもそもこの手法のメリットが、わかりやすいグラフ表示にあるとされているため、この危険性がかならずしも理解されているわけではない（西里 2007）。

この危険性を理解するのに必要なことは、standard coordinate と principal coordinate の関係を理解すること、そして、これらを図示するための biplot をめぐる議論を理解することである。

これらの概念は、たとえば、Greenacre の”Corespondence Analysis in Practice 2nd Edition”（CAiP2）にあるようなステップを踏むことで、獲得される。しかし、それは必ずしも簡単な道ではない。

そこで本稿では、先の功罪の原因にもなっている R で提供されている機能を活用し、図示の方法を比較しながら考えていく。対応分析の結果（result）は、plot で図示されるが、そこには、上述の状況が反映された座標の選択オプションが提供されている。そのオプションの機能を理解することで、今問題にしている「危険性」に対応することができるようになる。

2　princlipal coordinate と standard coordinate

対応分析の数理は、付録 A に書いてある通りである。今、問題にしている図示に関係するものは、princlipal coordinate と standard coordinate の二つである。

standard coordinate が空間を形成し、そこに principal coordnate が投影される。列を standard coordinate とし、行を principal coordinate で投影する、もしくはその逆である。先述の対応分析のメリットである、行変数と列変数の同時布置は、行、列ともに principal coordinate を用いて実現される。しかし、ここで問題がある。

行も列も、principal cooredinate というのは可能であるが（対称マップ）、行

な適用が治らないのと似ている。統計解析ソフトであれば、同じ 1 であっても、カテゴリーとしての 1 と数値としての 1 は、適用できる演算がまったく異なる。それは、尺度という視点から見て、当然のことである。しかし、統計処理ソフトではない Excel は、（おそらくは）あえてそこを踏み越えて、カテゴリーの 1 も数値として扱えるような処理を可能にしている。尺度の定義が明確でないと、統計処理の入り口にたてないはずである。西里は、西里 2007 で、カテゴリカル・データに対する、リッカート・スケールのアプリオリな適用を批判して、双対尺度法による分析を対置した。これが、データ構造を破壊せずに、分析する出発点だからである。

も列も standard coordinate という組み合わせは不可能である。それは、行変数によって生成される平面と、列変数によって生成される平面が同一ではないからである。

3　事例でみるグラフ出力

対応分析は、パッケージを使うことでクロス表で表現されている「地域」と「犯罪」の関係を図示することができる（図 3.3）。実行スクリプトは付録 B を参照。

入力として用いたデータは、Clausen 1998 の第 2 章で使われているものである。

表 3.1　ノルエウェイの犯罪統計

	強盗	詐欺	破壊
オスロ	395	2456	1758
中部地域	147	153	916
北部地域	694	327	1347

Clausen1989 第 2 章から引用

この図について、Clausen は、以下のように説明をしている。「グラフを見れば、3 つの顕著なクラスターが示されていること、また、国内の各地域が、特定の犯罪タイプとグループにまとめられるということが見てとれる。オスロにおいては、詐欺が相対的に多く発生しており、破壊と強盗は、それぞれ、中部地域と北部地域で発生している。」（Clausen 1998:14, 藤本 2015:16）

この一文を、下記グラフをみながら読むと、わかったような気になる。確かに、オスロと詐欺が、北部地域と強盗が、中部地域と破壊が、近くに表示されている。

しかし、この図からは、オスロでは詐欺が、北部では強盗が、中部では破壊が、発生している、と理解できなくもない。

分析の対象にして表に対して、mosaicplot を描いてみると、この行ポイント（地域）と列ポイント（犯罪の種類）の近接性は「傾向」であることがわかる。強盗や破壊はオスロでも発生している（図 3.1, 3.2）。

このように、行と列の変数のカテゴリーを同一の平面に表示できることが、対応分析のメリットではあるのだが（Clausen：藤本 2015:4）、それ故の危険も理解しておかなくてはならない。

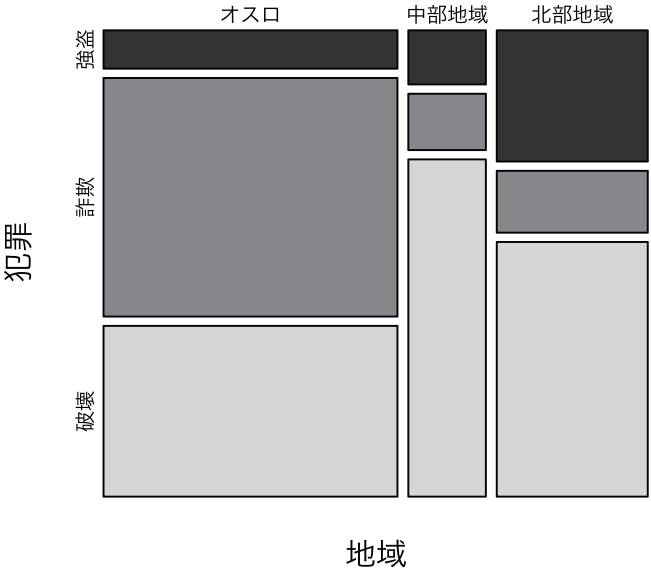

図 3.1　地域ごとの犯罪比率（行分析）

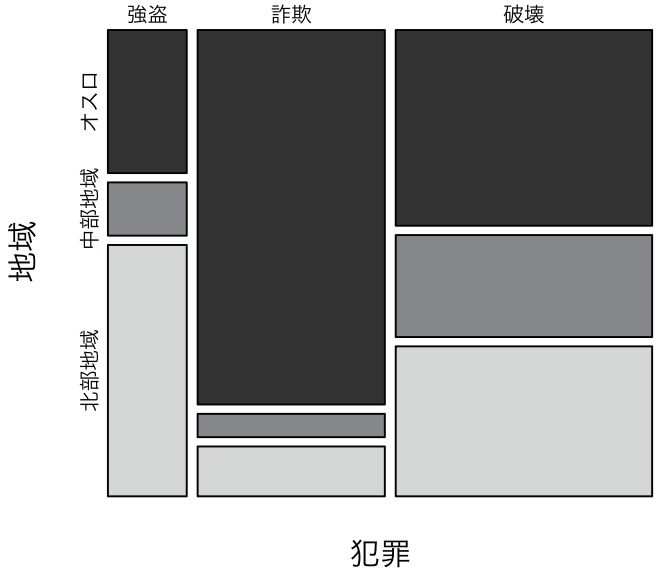

図 3.2　犯罪ごとの地域比率（列分析）

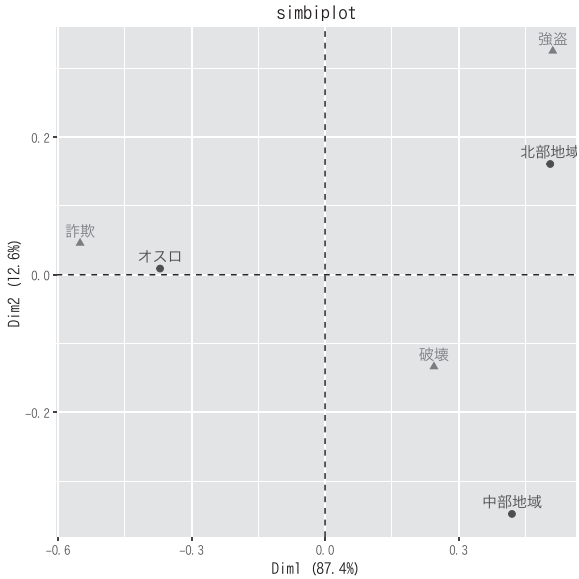

図 3.3　表 3.1 に対応分析を行った結果
：map オプションは “symmetric” である。

4　plot のオプションを変更したグラフを見る

ここで、対応分析の結果（リザルト）を図示するにあたり、default ではないオプションによる表示を確認することで、今確認した図示（symmetric Map、対称マップ）が唯一ではないことをみておく。

オプションに、map=” rowprincipal” を指定して表示する。このオプションの意味は後述するが、このオプションは、row=行座標を principal coordinate（主成分座標）に設定する指示で、列座標を standard coordinate に指定しその空間に行ポイント投影する、ということになる。ちなみに、default の対応分析ならではメリットと言われる同時布置（対称マップ）では、行ポイントも列ポイントも principal coordinate である。

行ポイントの相互の関係に変化はないものの、列ポイントは大きくかわる。この関係を理解することが、default として設定されている同時布置（symmetric map：対称マップ）を理解するための条件である。

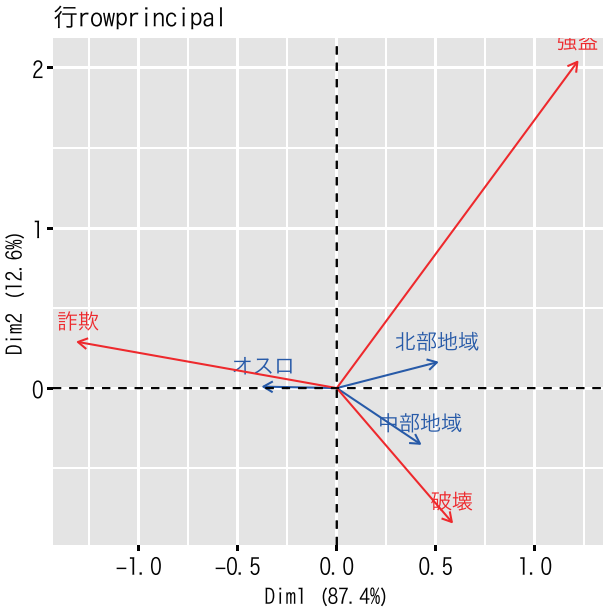

図 4.1　表 3.1 に対応分析を行った結果
：map オプションは “rowprincipal” である。

5　対応分析グラフ解釈の危険性への「警告」

対応分析は、その分析の過程で、**行変数**に対する処理と**列変数**に対する処理は同じプロセスをたどり、主成分分析で行うのと同じような次元縮減を行うが、そこで、手に入る、縮減された空間の軸（系の分散を体現）が同じになる、つまり、固有値、特異値が同じになるということ、また遷移公式によって、行のスコアは、対する列のすべての情報が反映されている、ということが説明される（Clausen 1998:20, 藤本 2015:21-22）。

こうした行と列の「対応性」「双対性」の説明（これが成立するのは、行、列それぞれのプロファイル、反応パターンを χ^2 距離で表現することを出発点としたことによる。）のあとに、行と列の同時布置の説明がでてくるとすると、この同時布置が占める行と列の変数カテゴリーの関係も「一目瞭然」のように思ってしまうことになるだろう。

しかし、変数内カテゴリー間の関係については、距離が明確に定義されて

いるが、変数間のカテゴリーについては、定義されていないのである。

この異なる変数間のカテゴリーの距離については、Clausen もおりに触れて強調している。

「この距離は、同じ変数内のカテゴリー・ポイント間でのみ計算可能であり、別の変数内のカテゴリーとの距離ではない、ということを銘記されたい。」(Clausen：藤本 2015:14)

「なお、各変数内のポイント間の距離だけが定義されるのであって、異なる変数ポイント間の距離が定義されるのではない、ということは再度強調しておく。」(ibid)

このように二回も強調されている。その上、最後のまとめのところでは以下のように更に強調するのである。

「この手法の不利な点は、異なるデータ・セットのポイント間の距離は定義されないことである。」(Clausen：藤本 2015:52)

しかしながら、同時布置の「わかりやすさ」の前に、これらの警告は顧みられない。

こうして同時布置の問題は今でも、「対応分析」をめぐる根本問題の一つである。

6　双対尺度法からの指摘

「対応分析」と数理的には等価といわれる「双対尺度法」を提唱するトロント大学名誉教授の西里静彦は、次のように述べている。

「数量化された行の重みと列の重みは同一空間にはない。特異値が 1 の場合のみ、両者は同一空間に存在するが、特異値が 1 より小さい場合には、行空間と列空間は合致しないので、あえて行変量と列変量を同じ空間（グラフ）に表現しようというのであれば、行の変量を列の空間に投影するか、列の変量を行の空間に射影しなくてはならない。その時の射影子の役割を果たすものが特異値である」(西里 2010:128) という部分をあげ次のようにコメントしている。

「これは理解できそうな文章であるが、実際にはあまり理解されていない」(ibid) と。つづけて、次のようにも述べる。「そして理論的には不正確であるのもかかわらず常套手段として用いられている対称尺度化による ρkyik ρkxjk のグラフが図 .... である。」(ibid:129)

このように多少なりとも数理的な説明がされている文献ではほぼ触れられ

ている「危険」性であるが、これが、ソフトの解説になると、省かれてしまい、「わかりやすいグラフ」が一人歩きすることになる[注2)]。

以上のように、危険性が強調されているにもかかわらず、同時布置が普及してしまったのは、便利だからであろうという。

西里は以下のようにコメントしている。

「そこでだされたのが対称化グラフ（symmetric scaling）あるいはフレンチ・プロットといわれるもので、射影された行の重みと射影された列の重みを同じグラフにプロットするもので図 .... がその例である。本書でもこの方法のグラフをインシデンスデータの場合に用いたがこれは別の空間にある二つのセットの変量を同じ空間に無理やり押し込めてしまう方法なので、グラフは大雑把な関係を示すだけで正確さに欠く。フランスのルバールら[注3)]はこのグラフに見られる行変数と列変数の距離は正確さを欠くのでグラフの解釈には注意を要すると警告している。」

「しかし、このグラフが常套手段となってしまった今日、どれだけの研究者がルバールなどの警告に注意を払うであろうか。」「このような空間の問題が念頭にあり、西里（1980a）はグラフをほとんど用いずに双対尺度法を解説した。しかし、『数量化というのはグラフによるデータ解析であるのにグラフがでてこない』ということで西里の書は特にフランスの研究者から批判を受けた。しかし正直な話、理論的な問題を含むグラフ法を広めることが正しくない。」（西里 2010:131）

西里が触れた大隅・ルバールの警告は以下である。

「主成分分析の場合と同様の方法で、一組の点の要素間の解釈をすることは理にかなっている。したがって「職種」として”教育･研究”と”健康管理サービス”とを指摘した「仕事の利点」の側の回答は類似の分布を示す。もう一方の組ついては、”社会的な地位”と”仕事への関心度”が「職種」間で似た分布を示している。このように一方のデータ集合の一つの点がもう一方の集合のすべての点に対する相対的なづけを解釈するのも理にかなったことである。しかし、特別な場合を除いて、相異なる集合に属する二つの点の親近性を解釈することは非常に危険である。」（大隅・ルバール 1994:75）[注4)]。

2　西里は、「全情報解析」を提案（西里 2010:127）

3　Lebart, Morineau & Tabard 1977：訳：大隅、馬場 1994

4　なお、大隅・ルバールでは、このあと、p.86-87 の解説で、「比率」という言葉で関係を述べている。

7　対応するルールはあるのか

以上みてきたように、対応分析の、このグラフ表示をめぐる問題は、手法の理解の根幹にかかわる。同時に、この問題は、多くの論争をひきおこしてきた（Greenacre 2007:267）。

この問題をめぐっては、西里は、グラフ表示こそが対応分析であるというスタンスに異をとなえたが、先に紹介したように、グラフ表示をしない解説には、フランスの研究者から批判されたと書いている。対応分析とは、行と列の関係をわかりやく図示する手法であるとする、彼らとは対極に位置している発想であろう。

facotoextra のドキュメントでは、symmetric map（対称マップ）では、行と列の関係を解釈するのは「不可能」という。「不可能」と言ってしまうと、ではなんのために、対称マップのオプションを残しておくのだ、ということになる。

Clausen、大隅・ルバールも、先述のように警告を発する。大隅は、このルバールたちの翻訳において、事例として日本のデータを分析しコメントを付している。こうした事例に丁寧に学ぶしかないのだが残念ながらそうした説明をしているテキストは見当たらない。

Greenacre は警告を発したあとで、それではこの問題に対処する黄金律はあるのかと自問する（Greenacre 2007:267）。彼は、そこで、一方を Standard coordinate に、他方を Principal Coordnate にマッピングしたものは、位置関係は正確であるが、データ構造を把握することができない、として、Symmetric Map の意義を（再）確認し、その上で、一方のベクトルと他方のベクトルが同じ方向を向いている場合には、Symmetric Map での表示での問題にはならず、そうでない場合は、問題になるという。そこでは、Gabriel 2002 が紹介されているが、ca で実装されている plot.ca のマップオプションにある、map= ”rowgab”, ”colgab” の gab は、Gabriel の gab である[注5)]。rowgab では、行を principal coordinate で、列を standard coordinate に対応するポイントの質量 mass をかけたものを用いてる。このオプションにある、”rowgreen”, ”colgreen” は、この standard coordinate に対する重みが、質量の平方根であるとことが違っている。この green は、Greenacre の Green であり、この

5　Gabriel は多変量データを二次元に表現する biplot を提案した。Gabriel 1971。一方を Standard Coordinate で他方を Principal Coordinate で表現する非対称マップは、biplot である。

rowgreen, colgreen で実装されている手法は、Greenacre 2007:103 に biplot の補正「Calibration of biplot」として提案されている[注6)]。

西里は、「数量化のあらたなステップ」として西里－クラベルの「全情報解析」（西里 2010：177）が、この問題を解決するものとして展望している。数量化理論の歴史は日本、フランスともに、60 年、50 年の歴史をもち、数理的な基礎は 1950 年代には完成していたといわれている。しかし、行ポイントと列ポイントの間の距離の問題は、未だに、最前線である。

Greenacre 2007 は Epilogue で、次の版（第 3 版）では、解決していることを望むと書いていたが、第 3 版（CAiP3 2017）でも未解決である。

8　どのように使えばいいのか

対応分析は、クロス表に集計されたカテゴリカル・データの行と列の関係を解析する手法である。であるななら、当然のことであるが、対応分析によって得られるグラフのみで分析を終わらせるわけはないのであって、独立性の検定をはじめ、カテゴリカル・データに対して、開発されてきた手法を用いて、丁寧に解析していくことが必要なのはいうまでもない。そうしたアプローチは、VCD（Visualizing Categorical Data）の名の下で、探索的データ解析を実現する総合的なアプローチとして発展を続けている。対応分析は、その重要なピースの一つである（Frendly:2016）。

9　対応分析パッケージと関連パッケージ

最後に本稿での言及した座標概念を理解するために使用する、R で提供されているパッケージに即して解説をしておく。

先述の standard coordinate と principal coordinate の扱いは、対応分析をめぐって論争が続いていた領域である（Greenacre CAiP2、Epilogue p267）。そこで、こうした事情を考慮した plot の実装が、パッケージ ca[注7)] ではなされていた。

6　この map オプションにある最後の一つ、Symbiplot は、ca のドキュメントおよび fviz_ca のドキュメントでは、「行と列の座標が特異値と等しくなるような尺度」と説明されている。これは Greenacre 2007:268 にある、SPSS の「Symmetrical normalization」のことである。なお、Greenacre は、この方法は評価していないが、SPSS の描画との関係で参照目的で加えているとしてされている。

7　http://www.carme-n.org/?sec=ca

それならば対応分析パッケージとして ca を使えばいいのだが、ca では summary を出力する際に、変数名がアルファベット表記を前提にした略語処理がおこなわれ、日本語カテゴリー名などが空白になってしまう。[注8)]

そのため、筆者は、Clausen の翻訳（藤本 2015）における検算、および解説において、FactoMineR[注9)] の CA を使用してきた経緯がある。

ところが、Alboukadel Kassambra と Fabian Mundt が、FactoMineR を意識した tools パッケージとして factoextra を公開し（現在 version1.0.3 2016-03-31）、そのなかで、ca の plot が有している「map オプション」を実装したのである。加えて、入力する対応分析の result オブジェクトは、FactoMineR の CA に限らず、ca［ca］、また、coa［ade4］、corresp［MASS］も入力オブジェクトとして使えるようにデザインされている。

こうしてパッケージ factoextra が提供する描画 function である、fvi_ca_biplot() を用いれば、FactoMineR で CA 処理したものを、ca の plot と同様のオプションで図示できるようになったのである。

fviz_ca は map オプションとして以下のものを用意している。

- ”symmetric”
- ”rowprincipal”, ”colprincipal”:
- ”symbiplot”
- ”rowgab”, ”colgab”
- ”rowgreen”, ”colgreen”

8　内部的にデータは保持されているので、summary ではなく result オブジェクトに直接アクセスすれば、変数名などは保存されてはいることが確認できる。

9　http://factominer.free.fr/

参考文献

［1］林知己夫 , 2011（1984）,『調査の科学』筑摩書房（講談社）

［2］林知己夫 , 2001,『データの科学』朝倉書店

［3］西里静彦 , 2007,『データ解析への洞察数量化の存在理由』関西学院大学出版会

［4］Clausen, Erik, 1998, “Applied Correspondence Analysis” SGAE,（訳＋解説：藤本一男 , 2015,『対応分析入門』オーム社）

［5］大隅・ルバール他 , 1994（1977）,『記述的多変量解析法』日科技連

［6］Greenacre, 1984, “Theory and Application of Correspondence Analysis”, Academic Press. http://www.carme-n.org/?sec=books5 PDF でダウンロート可能。

［7］Greenacre. M, 2007, “Correspondence Analysis in Practice Second Edition”, Chapman & Hall/CRC

［8］Frendly. M, 2016, “Discrete Data Analysis with R”, CRC Press

［9］Husson. F, 2011, “Exploratory Multivariate Analysis by Example Using R”, CRC Press

付録 A　対応分析の数理的要約

分析対象表（プロファイル行列 P）の標準化残差 S を求める。ここで、r は行和ベクトル。c は列和ベクトルである。いわゆる周辺度数。

1. ステップ 1

$S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$

この S の特異値分解（SVD）を行う。D の要素が特異値。

$D_r^{-1/2}$ は $_r^{-1/2}$ を要素とした対角行列。T は転置行列。

$S = \mathrm{U}D_\alpha V^T$ ここで、$U^TU = V^TV = I$

2. 座標の計算

ここで得られた結果から標準座標（standard coordinate）（Φ、Γ）、主軸座標（principal coordinate）（*F*、*G*）が計算される。

$\Phi = D_r^{-1/2}\mathrm{U}$

$\Gamma = D_c^{-1/2}\mathrm{V}$

$F = D_r^{-1/2}UD_\alpha = \Phi D_\alpha$

$G = D_c^{-1/2}VD_\alpha = \Gamma D_\alpha$

3. 固有値 / 特異値

特異値分解の結果得られた *Dα* は固有値を成分とした対角行列である。この操作の結果、n 次元空間であった行列 P は、固有値（として表現される P の分散＝慣性）ごとの軸に分解されていく。

付録 B　使用した R のスクリプト

対称マップ、非対称マップの違いを確認するためには、実際に対応分析を行い、それを図示する。あわせて、関連統計量を確認するというアプローチが可能である。以下に示すのは、そうした際に参考にしていただきたい。

```
library(FactoMineR)
library(factoextra)

.tbl2.1 <- matrix(c(395, 2456,1758,
                    147, 153, 916,
                    694, 327, 1347),byrow=T,3,3)
dimnames(.tbl2.1) <- list(地域=c("オスロ","中部地域","北部地域"),
                          犯罪=c("強盗", "詐欺","破壊"))
.tbl2.1

res.CA <- CA(.tbl2.1) #
plot.CA(res.CA,title="表2.1")
fviz_ca_biplot(res.CA,map="symmetric",title="Symmetric")
fviz_ca_biplot(res.CA,map="rowprincipal",title="rowprincipal")
```