

CA/MCAの数理（その 2）

津田塾大学 数学・計算機科学研究所

藤本一男

kazuo.fujimoto2007@gmail.com

履歴

- 2025/09/04 計量分析セミナー2025夏用に改訂
- 2023/09/06 計量分析セミナー2023夏用に作成

CA/MCAのリザルトを解釈するために

- 生成される二つの空間
- 分解される分散と座標軸
- SVDの結果から得られる行ポイント、列ポイントの座標軸
 - 主座標と標準座標
- 各ポイントの座標軸への寄与率（絶対的寄与率、contribution）
- 各ポイントと座標軸の相関。 $\cos^2$

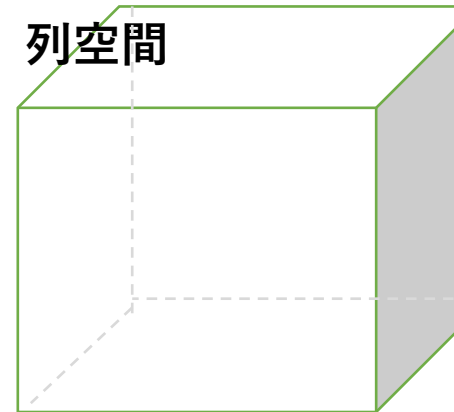
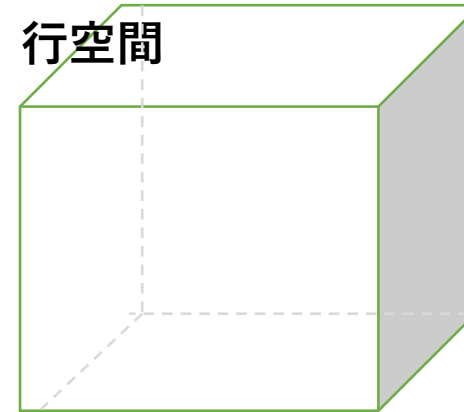
分析手順

- 分析対象のデータを前処理
 - 度数分布表など、通常の前処理を行う。
 - NAを"NA"など文字列に変換
 - CAやMCAのfunctionが求めている属性に変換
 - table、data.frame
 - factor
- CA/MCAを行う
- 慣性率を確認する。固有値の棒グラフ
- 行マップ、列マップを描き、座標軸を解釈する
 - ここが最初の関門

CA は 行空間と列空間を生成する

	col1	...	coln	rowSum
row1				
:				
rown				
colSum				

m x n 行列



空間生成とは、座標軸の生成を意味する
dim1...dimn.

> これらの座標軸は、データ表の慣性（分散）が空間生成に用いられた変数が分解されたもの。

> これらの慣性（分散）は、行空間と列空間で同じ値となる。（ここに、行分析と列分析の相互浸透があらわれている。cf カイ2乗距離をもちいているため。主成分分析PCAでは、それはできない。）

CAの三つのresult

	強盗	詐欺	破壊	行和
オスロ				
中部				
北部				
列和				



固有値 λ	Dim1	Dim 2
λ		

行座標 F	Dim1	Dim2
オスロ		
中部		
北部		

列座標 G	Dim1	Dim2
強盗		
詐欺		
破壊		

集計データ

回答者	変数 1	変数 2	...	変数n
1				
2				
3				
:				
n				



多重対応分析
MCAによる
空間生成
Active変数

生成された座標軸

回答者	分散率	累積分散率	修正分散率	累積修正分散率
Dim.1				
Dim.2		割合		
Dim.3				
:				
Dim.n				

個体空間座標値

回答者	Dim.1	Dim.2	...	Dim.n
1				
2		座標値		
3				
:				
n				

変数空間座標値

変数	Dim.1	Dim.2	...	Dim.n
変数1				
変数2				
変数3		座標値		
:				
変数n				

射影される変数座標値

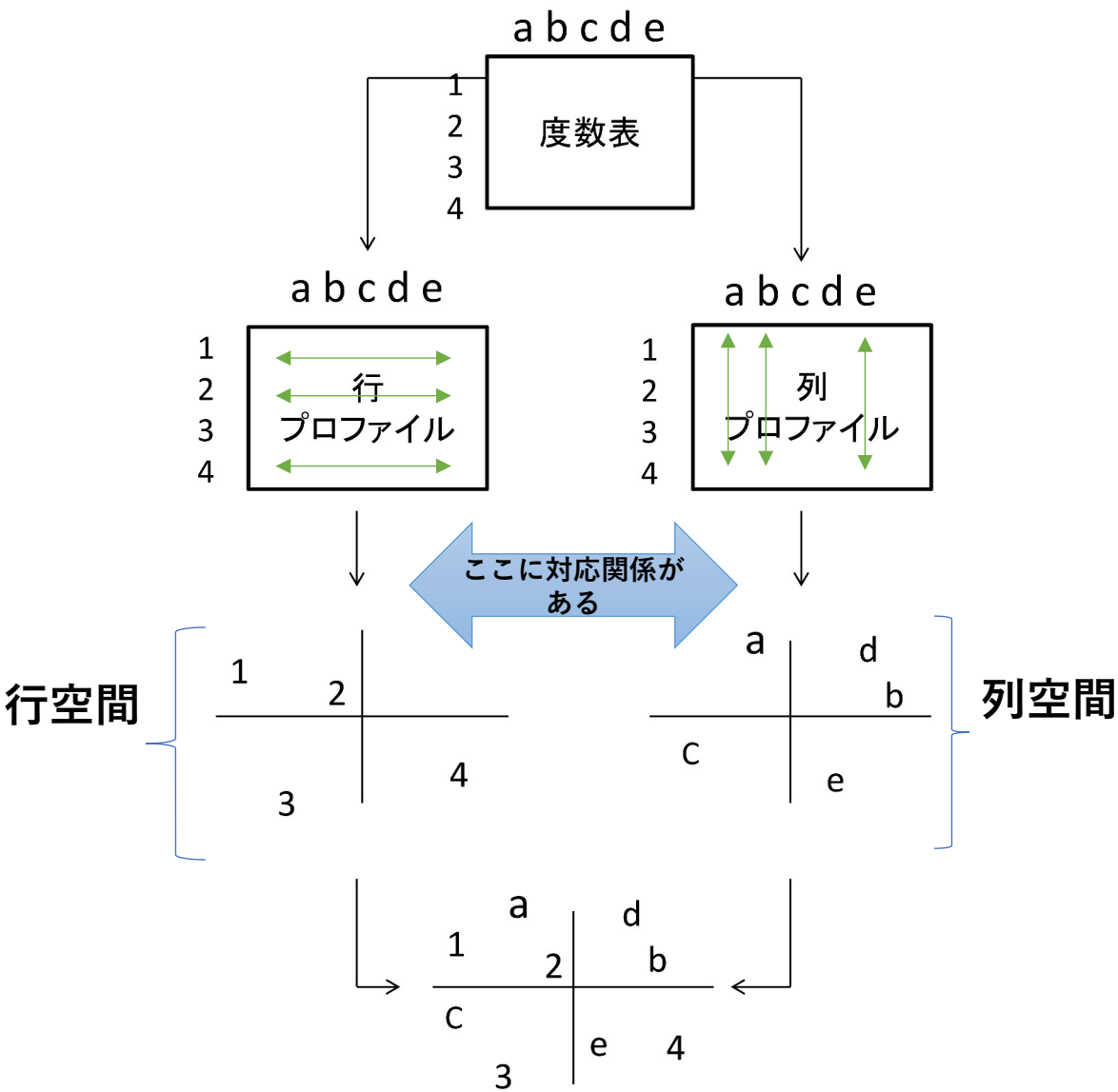
Passive変数	Dim.1	Dim.2	...	Dim.n
1				
2		座標値		

主座標と標準座標

二つの空間と座標系

- ・ 二つの空間の重ね合わせ
- ・ これは、それぞれのグラフを見てもらうのがよい。
- ・ 数理的に追いたい人は、『対応分析の理論と実践』を
読んでください。

マップ	行ポイント の座標	列ポイント の座標	option
対称マップ	主座標	主座標	Symmetric
	主座標	標準座標	rowprinci pal
	標準座標	主座標	colprintci pal
不可能！	標準座標	標準座標	✖



『対応分析入門』 p6、図1.2に加
筆

追加変数による空間解釈

行と列の関係

- 各行ポイントは、列ポイント全てと結びついている。
- 各列ポイントは、行ポイント全てと結びついている。
- その関係は、transition formula（遷移公式、推移公式）として以下のようにかける。『対応分析の理論と実践』 p246
 - F （行主座標）= $D_r^{-1}P\Gamma$ （ P は元表、 Γ は列標準座標）
 - G （列主座標）= $D_c^{-1}P^T\Phi$ （ P^T は元表の転置。 Φ は行標準座標）
- 標準座標は、平均0分散1にスケールリングされています（標準化）。

追加変数（サブリメンタリ・ポイント）

- 元表の行和、列和は、各行、各列の質量（weight）と呼ばれる。
 - この質量がある行や列が、行空間、列空間を生成する。
- ところで、行ポイント、列ポイントは、遷移公式によって結びついている。
 - 例：
 - MCAで変数空間で選択されたカテゴリの組み合わせによって、個体空間での個体の「位置」が決まります。
- そこで、**質量をもたないプロファイル**を考えると、そのプロファイルは、**反対側の空間に座標をもつことができる**。
- こうして、**空間生成には寄与せずに、内部構造を分析するための変数を考えることができる**。
 - 例えば：
 - 空間生成には、性別、年齢変数は用いずに、追加変数として生成された空間にplotする。
 - MCAでの構造化データ解析はこの仕組みを活用します。

CAの入力表と二つのプロフィール

0.56 0.15 0.29

行分析
行プロフィール

	強盗	詐欺	破壊	行和
オスロ	395	245 6	175 8	4609
中部	147	153	916	1216
北部	694	327	134 7	2368
列和	123 6	293 6	402 1	8493

列分析
列プロフィール

	強盗	詐欺	破壊	行和
オスロ	0.09	0.53	0.38	1
中部	0.12	0.13	0.75	1
北部	0.29	0.14	0.57	1
平均行比率	0.15	0.36	0.49	1

	オスロ	中部	北部	列和
強盗	0.32	0.12	0.56	1
詐欺	0.84	0.05	0.11	1
破壊	0.44	0.23	0.33	1
平均列比率	0.56	0.15	0.29	1

*元表を転置して配置している

このクロス表の変数、ノルウェイの都市名と
犯罪名は、Clausen1987=2015で使われて
いる事例のもの。

CAの入力表と二つのプロフィール

	強盗	詐欺	破壊	行和
オスロ	0.09	0.53	0.38	1
中部	0.12	0.13	0.75	1
北部	0.29	0.14	0.57	1
平均行プロ ファイル	0.15	0.36	0.49	1

行比率として要素計算したものが、
行プロフィール。
列和の行は、平均行プロフィールとなる。

	オスロ	中部	北部	列和
強盗	0.32	0.12	0.56	1
詐欺	0.84	0.05	0.11	1
破壊	0.44	0.23	0.33	1
平均列比率	0.56	0.15	0.29	1

列比率として要素計算したものが、
列プロフィール。
行和の列は、平均列プロフィールとなる。

*元表を転置して配置してある

行と列関係：対応関係

- 行の各ポイントは、列のすべてのポイントに関連している。
- 列の各ポイントは、行のすべてのポイントに関連している。
- 行と列の「対称性」。これを実現しているのが、点間距離をカイ2乗距離で定義していること。

遷移公式：Transition Formula

$$* F = RGD_{\lambda}^{-1/2}$$

行座標 F	Dim1	Dim2
オスロ		
中部		
北部		



$$** f_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{r_i} \right) g_{ik}$$

	強盗	詐欺	破壊
オスロ	行プロファイル R		
中部			
北部			

列座標 G	Dim1	Dim2
強盗	列座標G	
詐欺		
破壊		

固有値 λ	Dim1	Dim 2
λ	固有値 λ	

CA処理
前
CA処理
後

*Greenacre1984:64 **
Greenacre2017=2020:108

2025/9/4

CA/MCAの数理 (2)

遷移公式：Transition Formula

$$G = C F D_{\lambda}^{-1/2}$$

列座標 G	Dim1	Dim2
強盗		
詐欺		
破壊		



$$g_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{c_j} \right) f_{ik}$$

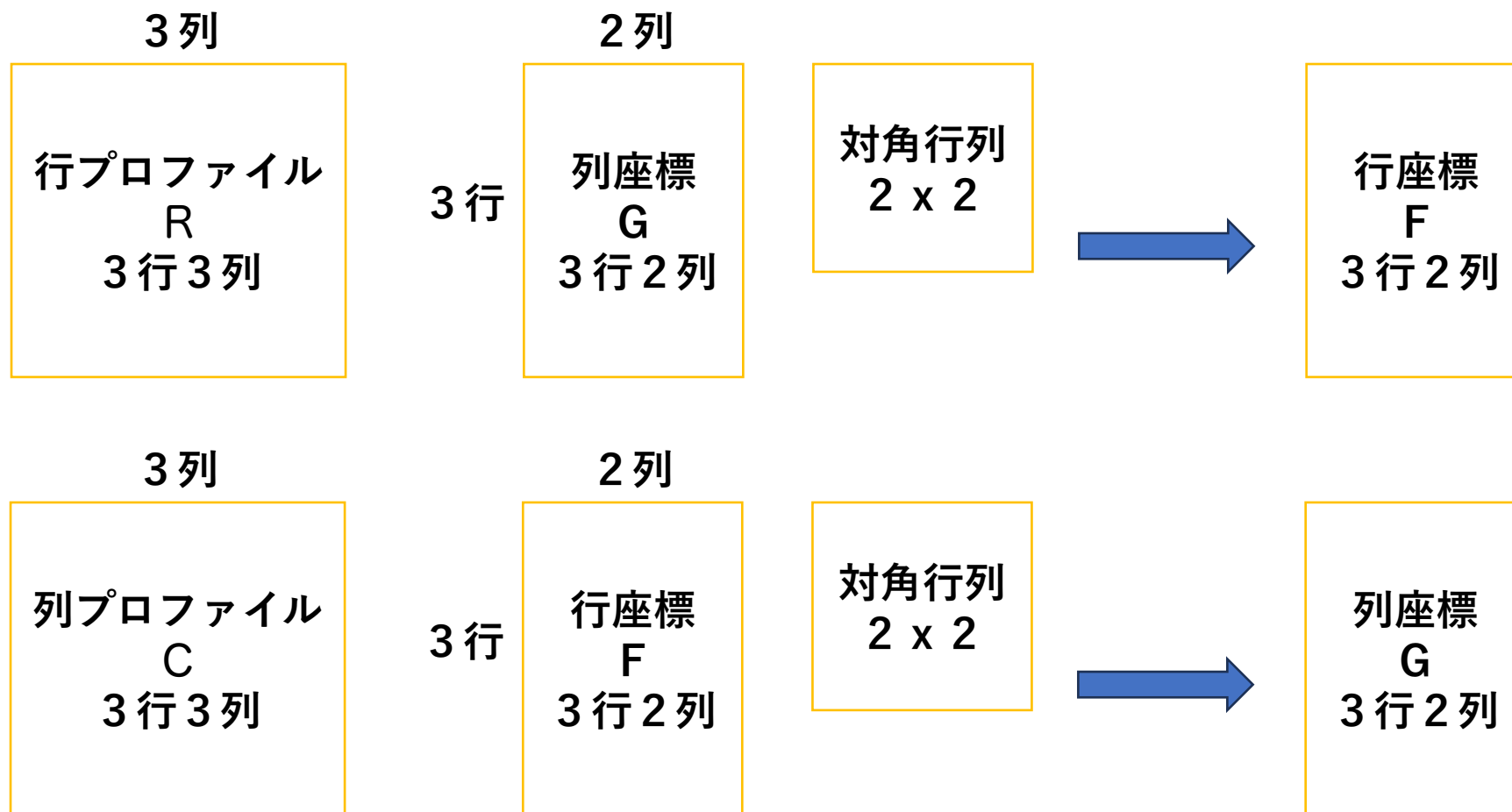
	オスロ	中部	北部
強盗	列プロファイル C		
詐欺			
破壊			

行座標 F	Dim1	Dim2
オスロ	行座標 F	
中部		
北部		

固有値 λ	Dim1	Dim 2
λ	固有値 λ	

CA処理
前
CA処理
後

行列の演算形式：遷移公式



行も列も
形式は同じ

追加変数を計算できる仕組み

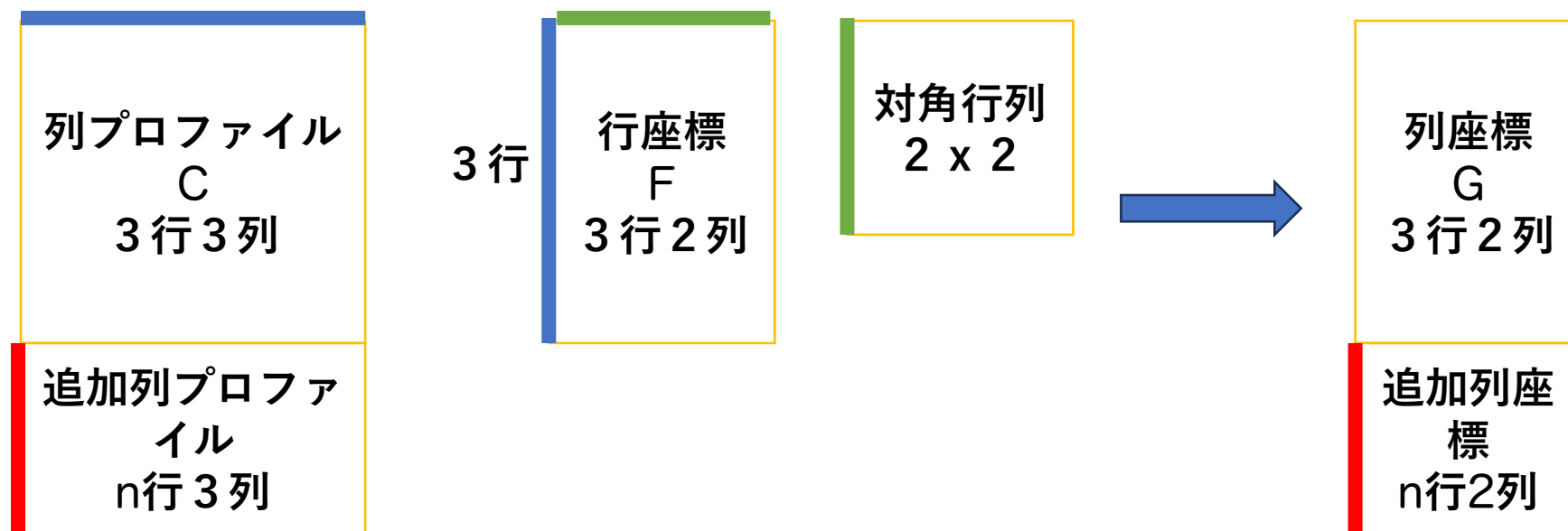
追加変数/カテゴリ/個体

- 空間生成に関係するポイント
 - Active変数
- CAの時はあまり問題にならないが、MCAでは次のことが問題になる。
 - 関連のない（まざると解釈不能になるカテゴリ）も、SVDを行えば数学的には何らかの座標を獲得できる。
 - しかし、解釈不能….
- そこで、空間生成には、同質（まざっても解釈可能という意味）なカテゴリを用いる。Active変数
- しかし、その連関はなくても、関連を見たい変数はある。
 - 年齢、性別、年収…..

追加変数の射影というアイデア

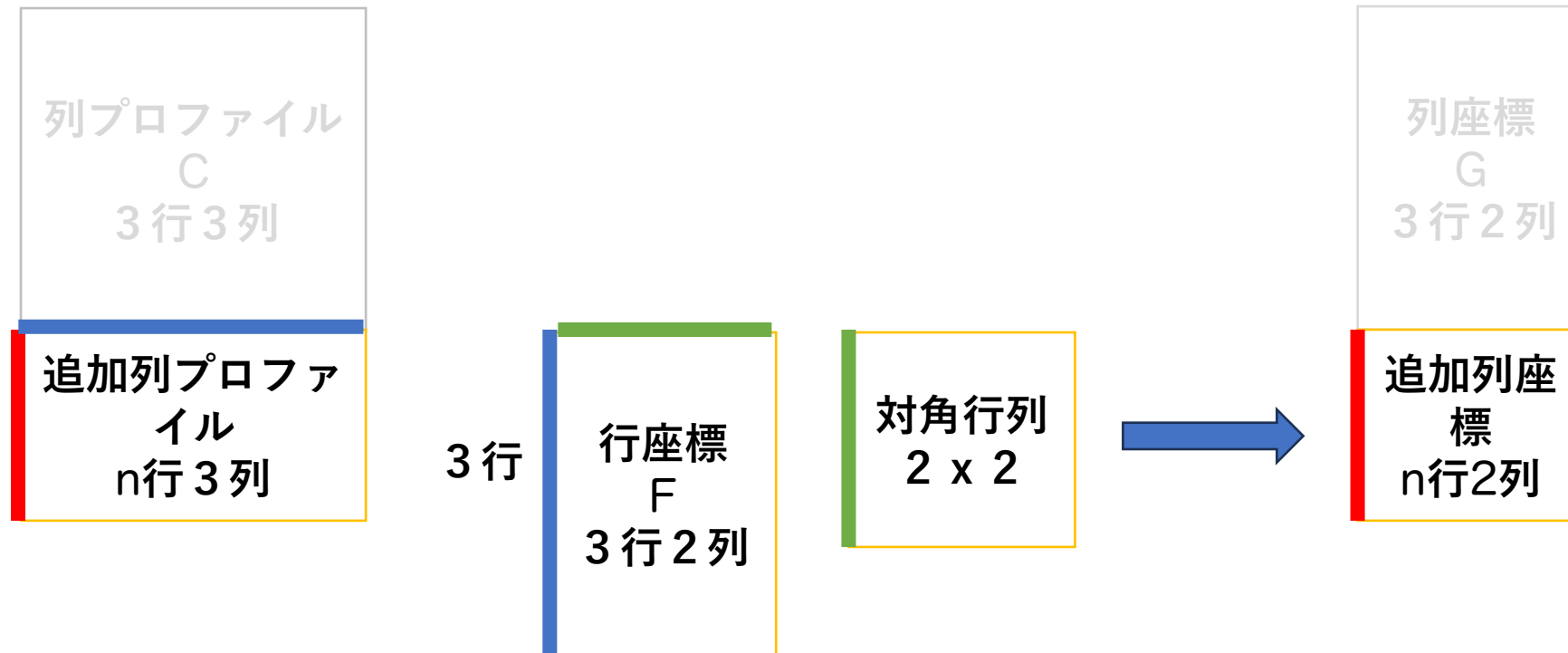
- 空間生成を行う変数の選択
 - Active変数
- これで生成される二つの空間の構造を、例えば性別、年代、もしくはその合成変数で分析したいときには、それら（性別など）を追加変数（サプリメンタリ変数）として、射影する。
- 射影するには、座標が取得される必要がある。
- この空間生成には寄与しないが、座標を取得できる方法を、「追加変数」として得る。

追加変数の座標が計算できるしくみ

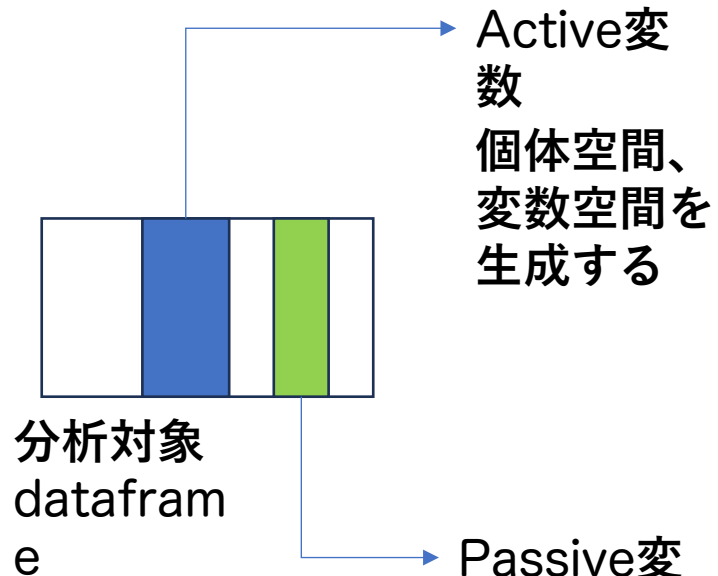


行列の掛け算では、右行列の列数と左行列の行数が同じであれば、計算可能。
つまり、列プロファイルに追加列プロファイルを加えると、列座標と固有値の
対角行列をかけて、追加行座標が得られることになる。
ここで、この追加列は、空間にはなんら影響を与えていない。

実際の計算では、追加変数分だけでよい



Active変数と追加変数の関係



指示行列化(index
matrix)
行和 r 、列和
 c
標準化残差行列

junkカテゴリ指定

特異値分解
SVD

$UD_{\alpha}V^t$

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

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

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

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

指示行列化(index
matrix)
転置して行和を求め、行比率行列
(プロフィール)化 (C_{sup})

$$\text{Passive変数の座標 } G_{sup} = C_{sup} F D_{\lambda}^{-1/2}$$

Active変数の列座標
とPassive変数 (追
加
変数) の列座標 (とも
に主座標) で散布図を
描く。

CAとMCA

CAとMCA

- CAは2変数、クロス集計表を入力とし、MCAは行が個体、列が変数の表（集計表）を入力する。
- しかし、functionの内部では、変数は回答カテゴリに分解されて、0/1でコーディングされたindicator（指示）行列に展開されて、その表に対してCAが行われている。
 - もう一つのMCAはburt行列という2変数ごと総当たりクロス表に対するCA。今回は、指示行列版MCAしか使いません。『対応分析の理論と実践』第18章
- この指示行列には厳しい制約がある。
 - 0/1で表示されるけれども、MA回答のような0/1展開ではない。

指示行列の重要性

- 変数内は、複数のカテゴリに分割されている。
 - その中に1が立つものが必ず1つあること。
 - つまり、行和は、変数総数になる。
 - これがMA回答の0/1とは異なる部分。
- ではMA回答はどうコーディングするのか。
 - その変数内の回答を合計すると1になるように配分する。
 - 選択肢が10個あって、3つ選ばれていたら、一つには1/3を配分。
 - 「いくつまで」「いくつでも」「いくつ」によって、コーディング方法がかわるので、やっかい。
 - 「参考資料」の「MAコーディングの問題」参照。

この指示行列ルールがGDAでは重要

- 平方和の分解で、
 - 全分散 = 群間分散 + 群内分散
が成り立つ前提。
- MCAをやっている実践的に直面する、ジャンクカテゴリの処理に関する。
 - speMCA カテゴリ選択MCA
 - CSA 個体選択MCAともに、全体の行和、列和は維持しており、全体のMCAとの比較を可能にする。（GDAのところで説明します。）