§9－4　主成分分析（pp. 202-203）

- 主成分分析は、データを要約して解釈する点で因子分析に類似している統計手法である。
- 主成分分析では、限りなく少ない成分（合成変数：component）へ相関関係にある観測変数を分解し、データを要約することで、新しい合成変数を作ることを目的とする。
- 活用例の一つとして「L. S. R, W, 語彙」の点数を圧縮して「総合英語力」のような成分を作ることが可能（竹内・水本, 2012, 183）

因子分析との違い

1. 主成分分析では、すべての分散がグルーピングされる

   cf. 因子分析では、観測変数の分散が共通分散と独自分散とに分けられ、共通分散のみを因子の推定に使用する。

(式 9.3)

第 1 主成分＝(主成分負荷量 1×観測変数 1）＋(主成分負荷量 2×観測変数 2)＋…

…（主成分負荷量 n×観測変数 n）

cf. 式 9.2

観測変数 1＝（因子負荷 a1×共通因子 1）＋（因子負荷 b1×共通因子 2）・・・＋誤差 1

観測変数 2＝（因子負荷 an×共通因子 1）＋（因子負荷 bn×共通因子 2）・・・＋誤差 n

2. 主成分分析は観測変数から合成変数への影響関係を推定する

   cf. 因子分析は、観測変数間の相関関係の原因となる潜在因子を推定し、観測変数への影響を想定する。

- 図 9.31 のモデルの矢印は主成分負荷量(component loading)を表し、因子分析における因子負荷量と同等の意味を持つ。

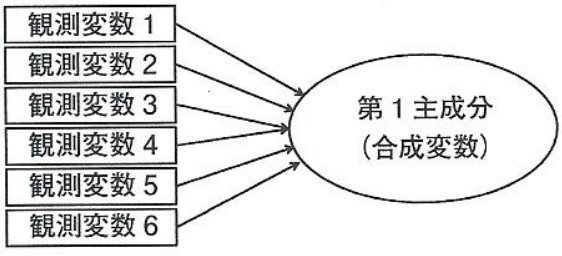


図 9.31　主成分分析モデル

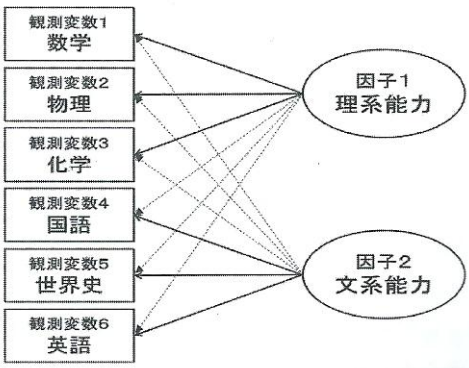


図 9.1　因子と観測変数

*矢印の向きが逆なのに注目

3. 主成分分析では、第一主成分が分散を最大限に説明するように計算が行われる

cf. 因子分析では、複数の因子を仮定することによって、観測変数全体の分散の説明を試みる。

・ただし、主成分分析も主成分の回転を行うことが可能であるため、因子分析と同様の分析結果が得られる場合がある。

4. 多重共線性や単一性が問題にならない

主成分分析ではデータの要約に主眼を置くため、相関が高いほど情報の集約性が高まり、多重共線性や単一性が問題にならない。

cf. 因子分析では、変数間の相関が高すぎる($r \geqq .90$)とエラーが生じることがある

→重回帰分析において、多重共線性や単一性が問題になる場合に、主成分分析で得られた主成分得点(component score)を独立変数として、重回帰分析を行う(主成分回帰：principle component regression)こともある。

■ 主成分分析は複数の観測変数を集約し、主成分得点によって変数化や数値化をすることが目的となる。

→多数の観測変数をまとめて数値の変数（主成分）にし、重回帰分析などの独立変数や観測変数とすることが分析の目的の場合には適した分析手法と言える。

(どういった時に主成分分析を用いるのが適しているか)

例 1. コーパスの構築・分析

石川(2010)は主成分分析の活用例として、5 つの助動詞の頻度データをまとめ、総合的な助動詞頻度指標を取り出す場合を例示している。このケースでは 5 語の合計頻度や平均頻度を出す場合に、粗頻度や調整頻度を単純に足しあわせただけでは、総合頻度指標としては不十分だとする一方、主成分分析を使えば、5 語それぞれの頻度について、変数の分散や単語間の相関関係を加味しつつ、全体を最も効率良く代表する値を取り出すことができる。としている。

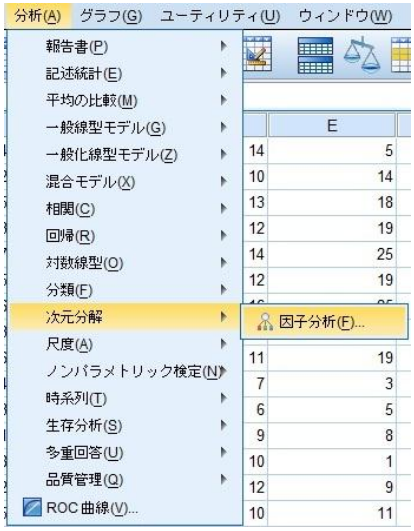
例 2. テキストマイニング利用時

石田(2008)では、鴎外と漱石のテキストについて、8 種類の助詞と読点の頻度情報についてまとめ（この時点でコーパスに近い)、それぞれの作品における助詞と読点の組み合わせの関係についての分析を例として、主成分分析をテキストマイニングに応用する際の例を紹介している。

■ 探索的因子分析を用いても、因子を抽出し、それに基づく観測変数のグルーピングを行い、尺度値などを算出することで、因子の数値化やスコア化が可能だが、観測変数の背後に共通して存在する因子の推定である点で異なる。

分析の手順

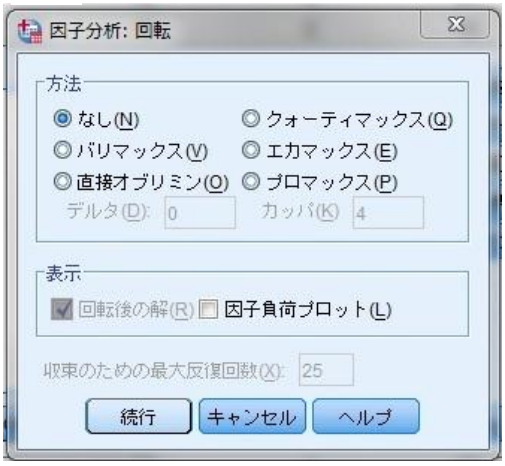
1.


分析(A)
グラフ(G)
ユーティリティ(U)
ウィンドウ(W)
報告書(P)
記述統計(E)
平均の比較(M)
一般線型モデル(G)
一般化線型モデル(Z)
混合モデル(X)
相関(C)
回帰(R)
対数線型(O)
分類(F)
次元分解
尺度(A)
ノンパラメトリック検定(N)
時系列(T)
生存分析(S)
多重回答(U)
品質管理(Q)
ROC 曲線(V)...
因子分析(F)...
E

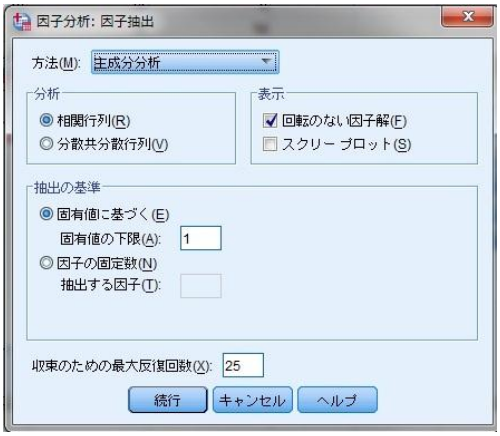
2.


因子分析
EF2
EF4
EF5
EX4
EF5R
REGR factor score ...
REGR factor score ...
REGR factor score ...
REGR factor score ...
REGR factor score ...
変数(V):
VL1
VL3
VL4
VL2
EF3
EF2R
EF1
ケース選択変数(C)
値(L)
記述統計(D)...
因子抽出(E)...
回転(T)...
得点(S)...
オプション(O)...
OK
貼り付け(P)
戻す(R)
キャンセル
ヘルプ

3.


因子分析: 回転
方法
なし(N)
クォーティマックス(Q)
バリマックス(V)
エカマックス(E)
直接オブリミン(O)
プロマックス(P)
デルタ(D): 0
カッパ(K): 4
表示
回転後の解(R)
因子負荷プロット(L)
収束のための最大反復回数(X): 25
続行
キャンセル
ヘルプ

4.


因子分析: 因子抽出
方法(M): 主成分分析
分析
相関行列(R)
分散共分散行列(V)
表示
回転のない因子解(F)
スクリープロット(S)
抽出の基準
固有値に基づく(E)
固有値の下限(A): 1
因子の固定数(N)
抽出する因子(T):
収束のための最大反復回数(X): 25
続行
キャンセル
ヘルプ

5.


因子分析: 因子得点
変数として保存(S)
方法
回帰(R)
Bartlett 法(B)
Anderson-Rubin 法(A)
因子得点係数行列を表示(D)
続行
キャンセル
ヘルプ

結果

説明された分散の合計

成分	初期の固有値			抽出後の負荷量平方和		
	合計	分散の %	累積 %	合計	分散の %	累積 %
1	4.594	41.760	41.760	4.594	41.760	41.760
2	2.198	19.983	61.743	2.198	19.983	61.743
3	.992	9.015	70.758			
4	.681	6.194	76.952			
5	.603	5.483	82.436			
6	.496	4.512	86.948			
7	.455	4.135	91.083			
8	.323	2.940	94.023			
9	.242	2.197	96.220			
10	.233	2.122	98.342			
11	.182	1.658	100.000			

因子抽出法: 主成分分析

第 2 主成分までに 50%が目安（ただし明確な基準はない）

KMO および Bartlett の検定

Kaiser-Meyer-Olkin の標本妥当性の測度		.843
Bartlett の球面性検定	近似カイ2乗	679.051
	自由度	55
	有意確率	.000

成分行列[a]

	成分	
	1	2
VL1	.829	-.078
VL3	.833	-.162
VL4	.787	-.062
VL2	.859	.027
EF3	.809	-.159
EF2R	.730	-.239
EF1	.694	-.119
EX3	.276	.723
EX2	.199	.740
EX1	.172	.741
EX4R	.205	.668

因子抽出法: 主成分分析

a. 2 個の成分が抽出されました

共通性

	初期	因子抽出後
VL1	1.000	.693
VL3	1.000	.721
VL4	1.000	.622
VL2	1.000	.738
EF3	1.000	.679
EF2R	1.000	.590
EF1	1.000	.496
EX3	1.000	.600
EX2	1.000	.586
EX1	1.000	.578
EX4R	1.000	.488

因子抽出法: 主成分分析

因子分析の結果と比較

- このデータでは２つの主成分が抽出された。それらの累積寄与率は 61.74%であったため、これらの１１項目のうち、２つの主成分で説明される割合が 61.73%であることがわかる。
- 成分行列では、第一主成分がすべて正の重み、第二主成分では VL1, VL3, VL4, EF3, EF2R, EF1 が負の重みを示し、VL2, EX3, EX2, EX1, EX4R が正の重みをつけている。そのため、第二主成分は、これらの 2 群のどちらが顕著にあらわれているのかを示す指標になる。
- 主成分分析では２つの主成分が抽出された一方で、因子分析では 3 因子を仮定している。
- 第二主成分の傾向から、負の成分は、VL（価値観）、EF（努力項目）に集中しており、正の成分は EX（充実感）に観測されている。