

SPSSXによるクラスター分析とQモード因子分析

浅野, 弘明
京都府立医科大学衛生学教室

<https://doi.org/10.15017/1470125>

出版情報：九州大学大型計算機センター広報. 22 (6), pp.710-714, 1989-11-25. 九州大学大型計算機センター
バージョン：
権利関係：



SPSSXによるクラスター分析とQモード因子分析

浅野 弘明*

1. はじめに

この広報の別の解説記事の中で、SPSS から SPSSX へ移行するための基本的な事項を解説していますが、その記事を書いたあとで SPSSX を利用して気が付いた点がありましたので、ここにあらためて紹介します。

新しい分析方法も取り入れられて、X版は分析方法がかなり充実した様に思われます。しかし、『変数に対するクラスター分析』と『Qモード因子分析』が見かけ上なくなりました。この2つの分析は、「CLUSTER」あるいは「FACTOR」単独では計算処理が出来ません（「QFACTOR」はX版には含まれません）。変数間あるいはケース間の種々の距離行列を求める「PROXIMITIES」や、相関行列を求める「PEARSON CORR」を前処理として利用することにより計算可能となります。英文マニュアル「SPSSX USER'S GUIDE 2nd EDITION」38章「PROXIMITIES」の例2，例3，40章「CLUSTER」11節「READ」の項で、この点に関して解説されていますが、今回、JCLを含めた形で実行例を紹介いたします。

2. 統計解析における変数とケース

通常、データ解析では身長のような変数と個人にあたるケースは明確に区別されています。しかし、変数とケースの区別は、表形式データにおける『列』と『行』にあたり、数学的には転置行列（行と列を入れ換えた行列）を取ることでその役目を入れ換えることが出来ます。定義しだいで変数とケースは入れ替わる訳で、変数とケースの区別は決して先天的に決定している訳ではありません。実際のデータ解析の場面でも、しばしばこの問題が生じます。一例として、ある学校での試験結果を考えて見ます。容易に、『生徒を理系・文系に分ける』と『試験科目の類似性を検討する』の2方向の分析が浮かびます。数量化分析Ⅲ類の様に、ケースと変数の役目がそれほど本質的でない分析もありますが、多くの多変量解析法においては、変数とケースは厳密に区別されています。しかし、SPSSXには変数とケースを入れ替える機能はありません。データ行列の行と列を入れ替えるためには、データ構造を変換するか、読み込みを工夫する必要があります。これは、生データに対して処理する場合でも結構大変ですが、実行形式で保存されたファイルでは更に複雑な処理が必要になります。しかし、多変量解析法の多くは、相関行列などの変数間あるいはケース間の距離行列が計算の基本となっています。すなわち、必要な距離行列さえ求めれば本質的な結果が得られることになります。

X版では、この距離行列を用いて計算処理する方法を多くの分析方法で取り入れています。特に、多次元尺度構成法「ALSCAL」では、様々なタイプの距離行列を扱う必要が生じます。また、分析の性質上、変数とケースの両者が分析の対象となります。これらの距離行列を求めるプログラムが

平成元年9月29日受理

* 京都府立医科大学衛生学教室

「PROXIMITIES」です。求める距離が変数間かケース間かの指定以外に、どのような尺度で距離を求めるかもきめ細かく指定できます（機能の詳細は、英文マニュアル38章「PROXIMITIES」を参照下さい）。ただし、計算された距離行列を次の処理で利用するためには、少々複雑な手続きが必要になります。以下では、クラスター分析と因子分析を例に取り、この距離行列の処理方法について解説いたします。

3. 変数に対するクラスター分析

かつてのクラスター分析「CLUSTER」では「TYPE」サブコマンドにより、ケースに対する分析（Qモード）か変数に対する分析（Rモード）かの指定が出来ました（指定がなければケースに対する分析（Qモード））。しかし、X版では「TYPE」サブコマンドがなくなり、変数に対する分析（Rモード）の指定出来なくなりました。「PEARSON CORR」（例1）か「PROXIMITIES」（例2）により相関行列を外部に出力し、その行列を新たに読み込んで分析に用いることによりRモード（変数に対する分析）が可能になります。また、例3の様に、相関行列を他の距離行列（例3ではユークリッド距離）とすることも可能です。

[例1] 「PEARSON CORR」で求めた相関行列によるクラスター分析

```

行番号
1:  //A799999AN  JOB    CLASS=A
2:  // EXEC SPSSX
3:  DATA LIST FILE=FT08F001 LIST / V1 TO V20
4:  PROCEDURE OUTPUT OUTFILE=CORRMAT
5:  PEARSON CORR  V1 TO V20
6:  OPTION 7
7:  INPUT MATRIX FILE=CORRMAT / FREE
8:  CLUSTER      V1 TO V20
9:  /READ=SIMILAR
10: FINISH
11: //FT08F001 DD DSN=??????. KYOTO. DATA, DISP=SHR
12: //CORRMAT DD UNIT=WORK, SPACE=(TRK,(10,10))
13: //
```

[概略説明]

行番号

- 1: ジョブ文
- 2: EXEC文（SPSSXの呼び出し）
- 3: FT08F001のDD名に対応するデータの読み込み
- 4: 相関行列をDD名『CORRMAT』で出力するためのコマンド
（出力先はワークファイルを利用するのが最も簡単）
（DD名CORRMATは変更可能、ただし、12行目のDD文と一致させる）
- 5: 相関行列の計算
- 6: 相関行列の外部出力を指定（省略不可）
- 7: 出力された相関行列の読み込み

(FILE = の指定は、4 行目の名前と同じにする)

8 : クラスタ分析の指定

9 : 対称行列からの読み込みを指定 (省略不可)

10 : SPSSX 終了

11 : データに対する DD 文

12 : 相関行列をワークファイルに作成するための DD 文

[例 2] 「PROXIMITIES」で求めた相関行列によるクラスタ分析例

行番号

```

1 : //A79999AN JOB CLASS = A
2 : // EXEC SPSSX
3 : DATA LIST FILE = FT08F001 LIST / V1 TO V20
4 : PROCEDURE OUTPUT OUTFILE = CORRMAT
5 : PROXIMITIES V1 TO V20
6 : /VIEW = VARIABLES
7 : /MEASURE = CORR
8 : /WRITE
9 : INPUT MATRIX FILE = CORRMAT / FREE
10 : CLUSTER V1 TO V20
11 : /READ = SIMILAR
12 : /PLOT = DENDROGRAM
13 : FINISH
14 : //FT08F001 DD DSN = ??????. KYOTO. DATA, DISP = SHR
15 : //CORRMAT DD UNIT = WORK, SPACE = (TRK, (10, 10))
16 : //
```

[概略説明]

行番号

1 ~ 4 : 例 1 と同じ

5 ~ 8 : 「PROXIMITIES」による相関行列の計算

「VIEW」 : 計算する距離行列が、変数間かケース間かの指定
(ケースの場合は省略可、変数の場合は省略不可)

「MEASURE」 : 距離の種別の指定
(CORR で相関行列の指定、省略不可)

「WRITE」 : 外部出力指定 (例 1 の 7 行目に対応、省略不可)

9 ~ 12 : クラスタ分析の指定 (例 1 と同様)

但し、樹状図を「PLOT = DENDROGRAM」により指定している

12 ~ 15 : 例 1, 10 ~ 12 行と同じ

[例 3] 「PROXIMITIES」で求めた距離行列によるクラスタ分析例

例 2 の 7 行目を下の行で置き換える。

```

7 : /MEASURE = EUCLID
    (「EUCLID」がユークリッド距離の指定)
```

4. Qモード因子分析

通常、因子分析は変数に対する分析ですが、例えば、体格変数により個人を大柄、小柄、長身やせ型、肥満型などに群分けをしたい場合などに、Qモードの因子分析が利用できます。以前は「QFACTOR」によりQモードの因子分析が可能でしたが、X版には「QFACTOR」は含まれません。クラスター分析での例2、例3の様に、「PROXIMITIES」により、個人間の距離行列を求め因子分析で利用する必要があります。

『変数に対するクラスター分析』においては、クラスター分析で用いる変数がデータ読み込み時に定義されていますので、新たに宣言する必要はありませんでした。しかし、Qモード因子分析では、ケースを変数として扱うための一連の処理が必要になります(例4、9～14行)。特に、「NUMERIC」コマンドを利用して、ケースを変数として宣言(登録)する必要があります(例4、10行)。

例を示しますが、距離行列(相関行列)を外部ワークファイルに出力する方法などは、クラスター分析例と同様です。なお、例4でのケース数は全部で30としています(ケース数に制限はありません)。

[例4] Qモード因子分析例

行番号

```

1:  //A79999AN  JOB  CLASS=A
2:  // EXEC SPSSX
3:  DATA LIST FILE=FT08F001 LIST / V1 TO V20
4:  PROCEDURE OUTPUT OUTFILE=CORRMAT
5:  PROXIMITIES V1 TO V20
6:      /VIEW=CASE
7:      /MEASURE=CORR
8:      /WRITE
9:  INPUT PROGRAM
10: NUMERIC CASE1 TO CASE30
11: INPUT MATRIX FILE=CORRMAT / FREE
12: END INPUT PROGRAM
13: FACTOR VARIABLES=CASE1 TO CASE30
14:     /READ
15: FINISH
16: //FT08F001 DD DSN=??????.KYOTO. DATA, DISP=SHR
17: //CORRMAT DD UNIT=WORK, SPACE=(TRK,(10,10))
18: //
```

(解説)

行番号

- 1: ジョブ文
- 2: EXEC文(SPSSXの呼び出し)
- 3: FT08F001のDD名に対応するデータの読み込み
- 4: 相関行列をDD名『CORRMAT』で出力するためのコマンド
(DD名CORRMATは変更可能、ただし、17行目のDD文と一致させる)
- 5~8: 「PROXIMITIES」による相関行列の計算
「VIEW」 : 計算する距離行列が、変数間かケース間かの指定
(ケースの場合は省略可、変数の場合は省略不可)

「MEASURE」 : 距離の種別の指定

(CORR で相関行列の指定, 省略不可)

「WRITE」 : 外部出力指定 (省略不可)

9~12 : 出力された相関行列の読み込み

(10 行目の「NUMERIC」宣言はケースの数だけ必要)

(FILE = の指定は, 4 行目の名前と同じにする)

13~14 : 因子分析の指定

(14 行目は省略不可)

15 : SPSSX 終了

16 : データに対する DD 文

17 : 相関行列をワークファイルに作成するための DD 文

Q モード因子分析を行う際に, 例えば身長(m)と体重(g)の様に値の桁数が異なる変数が含まれると, ケース間での相関係数が 0.9 以上の大きな値ばかりになることがあり, 分析結果が歪められる可能性があります. この様な恐れのある場合に有効な方法として, 各変数を Z スコア (標準化得点) や偏差値に変換しておくことが考えられます. 「PROXIMITIES」コマンド中で 「STANDARDIZE = VARIABLES Z」を指定することにより, 変数の標準化を行うことができます. 具体的には, 例 4 の 6 行目に 「 /STANDARDIZE = VARIABLES Z」(サブコマンドですので第 1 カラムは空白) を挿入すれば, 各変数値が標準化されます.