主成分分析とクラスター分析

SHIMURA Masato
jcd02773@nifty.ne.jp

2016 年 7 月 1 日

目次

多変量解析

多変量解析は多数の変数をデータの多元的特性をなるべく損なわないで、より低次元の空間に置き換えることである。

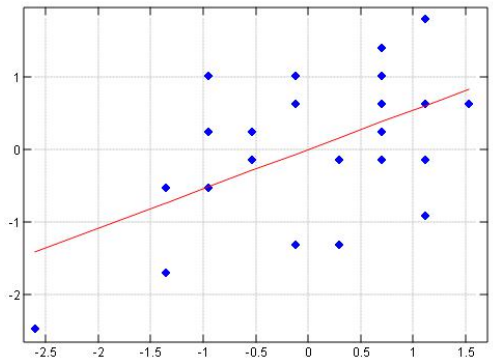
JAPLA で配布された鈴木義一郎著「統計分析へのいざない」（マーケティングサイエンス研究所）に解説のある多変量解析の内、主成分分析とクラスター分析のアルゴリズムとスクリプトを紹介する。

1　データの取扱い

鈴木義一郎名誉教授常用のサンプルデータ
ミスユニバース代表のプロポーションである (身長, 体重,3 サイズ）

```
   STYLE
165 53 86 56 92
160 47 84 52 92
166 55 86 64 89
164 56 90 60 95
168 55 87 56 87
164 54 87 57 92
168 54 94 58 97
169 55 88 57 92
169 53 86 58 93
166 56 84 57 90
165 53 85 55 90
163 49 84 59 90
164 52 87 58 90
167 53 86 59 88
169 58 89 60 90
169 51 84 60 90
166 50 86 59 87
168 53 88 60 88
165 54 88 62 90
167 50 88 58 89
170 55 88 60 90
168 57 84 62 92
168 56 85 62 94
163 52 83 60 88
```

身長と体重を表す次の散布図は標準化されているが、縦、横各 0 ポイントの線で区切っても大まかな傾向はつかめ、これだけで非階層クラスターに分類できそうだ。



1.1　ダイレクト入力

J の *ijs* 画面でのダイレクト入力の例

```
STYLE=: 165 53 86 56 92 160 47 84 52 92 166 55 86 64 89 164 56 90 60 95
STYLE=: STYLE,168 55 87 56 87 164 54 87 57 92 168 54 94 58 97 169 55 88 57 92
STYLE=:STYLE,169 53 86 58 93 166 56 84 57 90 165 53 85 55 90 163 49 84 59 90
STYLE=: STYLE , 164 52 87 58 90 167 53 86 59 88 169 58 89 60 90 169 51 84 60 90
STYLE=:STYLE, 166 50 86 59 87 168 53 88 60 88 165 54 88 62 90 167 50 88 58 89
STYLE=: STYLE, 170 55 88 60 90 168 57 84 62 92 168 56 85 62 94 163 52 83 60 88
STYLE=: 24 5 $ STYLE
```

2　主成分分析

2.1　主成分分析のアルゴリズム

主成分分析のアルゴリズムを大まかに述べれば次のとおりである

1. データを標準化する
2. 相関行列を求める
3. 相関行列の固有値と固有ベクトルを求める
4. 固有値の大きいものから幾つかの主成分を抽出する
5. 固有ベクトルを標準化したデータに反映する

2.2　標準化

データから単位を取り去る標準化を行う

1. 標準化したデータ

```
      stand STYLE
   _0.53298 _0.145407  _0.22758  _1.06586  0.566029
   _2.59613  _2.47193  _1.06788  _2.64005  0.566029
   _0.12035  0.630099  _0.22758   2.08253 _0.668943
  _0.945609   1.01785   1.45301  0.508333     1.801
   0.704909  0.630099  0.192568  _1.06586  _1.49226
  _0.945609  0.242346  0.192568 _0.672312  0.566029
   0.704909  0.242346   3.13361 _0.278763   2.62431
    1.11754  0.630099  0.612716 _0.672312  0.566029
    1.11754 _0.145407  _0.22758 _0.278763  0.977686
   _0.12035   1.01785  _1.06788 _0.672312 _0.257286
   _0.53298 _0.145407 _0.647729  _1.45941 _0.257286
   _1.35824  _1.69642  _1.06788  0.114785 _0.257286
  _0.945609  _0.53316  0.192568 _0.278763 _0.257286
   0.292279 _0.145407  _0.22758  0.114785   _1.0806
    1.11754   1.79336   1.03286  0.508333 _0.257286
    1.11754 _0.920913  _1.06788  0.508333 _0.257286
   _0.12035  _1.30867  _0.22758  0.114785  _1.49226
   0.704909 _0.145407  0.612716  0.508333   _1.0806
   _0.53298  0.242346  0.612716   1.29543 _0.257286
   0.292279  _1.30867  0.612716 _0.278763 _0.668943
    1.53017  0.630099  0.612716  0.508333 _0.257286
   0.704909    1.4056  _1.06788   1.29543  0.566029
   0.704909   1.01785 _0.647729   1.29543   1.38934
   _1.35824  _0.53316  _1.48803  0.508333   _1.0806
```

各行の合計は 0 になっている。

```
   require 'numeric'
   clean  +/  stand STYLE
0 0 0 0 0
```

2. *Script*

ここまでのスクリプトを整理してみよう

平均	`mean=:+/ % #`	$\bar{x} = \frac{\sum x_1}{n}$
残差	`dev=:-"1 mean`	$x - \bar{x}$
残差平方和	`ss=:+/@:*:@dev`	$ss = \sum(x - \bar{x})^2$
標準偏差	`sd=:%:@(ss % #)`	$sd = \sqrt{\frac{ss}{n}}$
標準化	`stand=:dev % "1 sd`	$\frac{x - \bar{x}}{sd}$

各残差を求め標準偏差で割ったもの

2.3 相関行列

1. 相関行列を求める

標準化したデータ *(X_0)* の分散共分散行列は相関行列 *(R)* になり、データ相互間の影響が一瞥できる。

```
   cor STYLE
       1 0.542494  0.326565   0.372425  0.0115011
0.542494        1    0.3131   0.449056   0.241926
0.326565   0.3131         1  0.0605705   0.424285
0.372425 0.449056 0.0605705          1 _0.0649716
0.0115011 0.241926  0.424285 _0.0649716          1
```

2. *Script*

内積	`mp=: +/ . *`	$X \cdot X$
共分散	`cov=: (|: mp ]) % #`	
相関行列	`cor=: cov@stand`	

共分散は データの内積を n で割ったもの。左側を先に転置 *(`|:`)* しておく。

相関行列は先に標準化したデータの共分散である。

2.4 固有値と固有ベクトル

1. 相関行列の固有値を求める。

ルベリエ・ファディーエフ法を用いた。

*1

$$\begin{array}{ccccc} 1-\lambda & 0.542494 & 0.326565 & 0.372425 & 0.0115011 \\ 0.542494 & 1-\lambda & 0.3131 & 0.449056 & 0.241926 \\ 0.326565 & 0.3131 & 1-\lambda & 0.0605705 & 0.424285 \\ 0.372425 & 0.449056 & 0.0605705 & 1-\lambda & -0.0649716 \\ 0.0115011 & 0.241926 & 0.424285 & -0.0649716 & 1-\lambda \end{array}$$

ルベリエ・ファディーエフ法はマトリクスを練り上げて多項式を作成し、p. で解く。
この特性方程式は次のようになる

$$f(x) = -0.3468 + 2.58768\lambda - 7.11894\lambda^2 + 8.9141\lambda^3 - 5\lambda^4 + \lambda^5$$

```
   }. ,.   char_lf cor STYLE
+----------------------------------------+
|2.14898 1.30735 0.675427 0.509665 0.358586|  NB. 固有値
+----------------------------------------+
|_0.3468 2.58768 _7.11894 8.9141 _5 1      |  NB. 特性方程式
+----------------------------------------+
```

2. 固有ベクトルを求める（ 各縦ベクトル）

```
   pick_evec cor STYLE
0.521033 _0.236935  0.534739   0.27597  0.557029
0.570928 _0.104455 _0.198549  0.481168 _0.626247
0.419977  0.483214  0.401878 _0.591852 _0.279874
0.399578 _0.488062 _0.513792 _0.551293  0.185005
 0.25792  0.679155 _0.499144  0.195203   0.43009
```

この固有ベクトルの行列は直行行列になる。

2.5 主成分行列

1. 主成分を抽出する固有値の大きいものから第一主成分、第二主成分.. とする。
第 *1* 主成分の固有ベクトルのを見る。すべてに反応している → 全体のボリューム
第 *2* 主成分は *BH* が反応
固有値の合計に占める比率からこの *2* 主成分で７０％程度説明できる。

*1 LAPACK は固有ベクトルが反転するので今回は回避した（Appendix　参照）

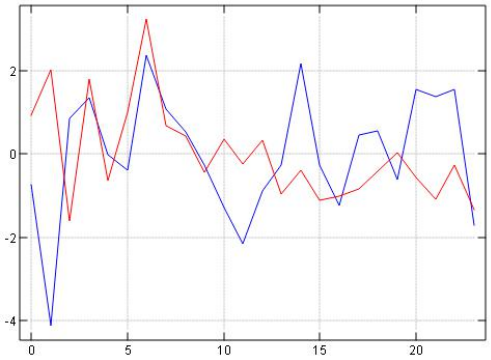
```
   tmp0,  (tmp1=.+/\ tmp0=.\:~  >1{char_lf  cor STYLE),: tmp % 5
 2.14898  1.30735 0.675427 0.509665 0.358586   NB. 固有値 ソート済
 2.14898  3.45632  4.13175  4.64141        5   NB. 累計
0.429795 0.691264  0.82635 0.928283        1   NB. 累積比率
```

2. 主成分行列

標準化したデータに相関行列の固有ベクトルを左からかける

```
 a=. (stand STYLE) mp pick_evec cor STYLE
    _0.7362  0.936127 _0.0824939  0.615734 _0.0958763
   _4.12136   2.03024  _0.252701  0.292094   0.155811
   0.861054  _1.61799   _1.01701 _0.873998  _0.300371
    1.36629    1.7949   _1.28395 _0.559852  _0.702183
_0.00288101  _0.63305    1.62171  0.680054  _0.894831
  _0.396108   1.00434  _0.413484  0.222809  _0.613332
    2.38716   3.24024   0.421466 _0.877535    0.44099
    1.07669  0.678023   0.781621  0.730085   0.175482
   0.544456  0.440489   0.190222  0.717664    1.14618
   _0.26507 _0.440424  _0.221754   1.40899  _0.640632
   _1.28225  0.366024   0.361812  0.920647  _0.405195
    _2.1452 _0.247758   _0.74919 _0.672574   0.515251
  _0.893963  0.334108 _0.0507581 _0.528014  _0.408964
  _0.259151 _0.953949   0.574104 _0.128828  _0.125957
    2.17669  _0.37585   0.523853  0.229547  _0.806268
  _0.255225  _1.10744   0.218527  0.166855    1.48148
   _1.24446  _1.01425   0.789895 _0.882781   0.195636
   0.466001 _0.837749   0.930248 _0.729246 _0.0584787
   0.569255 _0.409947  _0.624044   _1.1575  _0.491132
  _0.621462 0.0452568    1.13949 _0.888565   0.471596
     1.5511 _0.555129   0.806617 0.0323629   0.269655
    1.38491  _1.07768   _1.27941   0.89922   0.294372
    1.55233    _0.275   _1.44452  0.624694   0.773712
   _1.71261  _1.32352  _0.940251 _0.241859  _0.376939
```

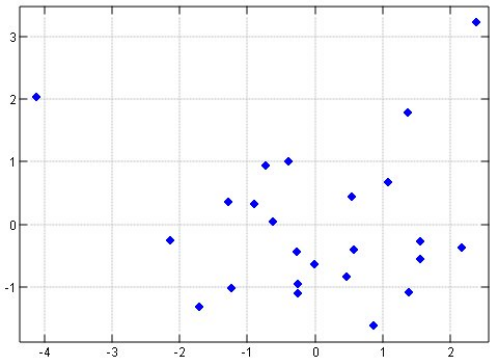
3. 第一主成分（青）と第二主成分（赤）のグラフ



２つの主成分スコアがともに高い女性は *2* 人

4. 第 *1* 主成分（*X* 軸）と第 *2* 主成分（*Y* 軸）の散布図

```
require 'plot jpeg'
'marker' plot{  |:2{."1 a
 pd 'save jpg c:/temp/principal0.jpg'
```



右上に突出しているのが児島明子である

5. *Script*

```
NB.principal=: (stand STYLE) mp pick_evec cor STYLE

   principal=: stand mp pick_evec@cor   NB. fork
```

2.6　この簡単なアルゴリズムが何故主成分分析になるのか

主成分分析のアルゴリズムとスクリプトは拍子抜けするほど簡明だが、その基礎となる理論は深淵である。

1. 主成分分析は *1901* 年にピアソンが考案し、*1933* 頃ホテリングが独立に提唱して、主成分分析と名付けた。*(WikiPedia* による。*)*
   *2

   工学、統計、数学など分野によりいろいろな呼び名があるようだ。
2. 相関行列から固有値と固有ベクトルを求めることはラグランジュの未定定数法を解いて最適化することである
3. 対称行列は２次形式の形をしている。ルベリエ・ファディーエフ法で固有値を求め、固有ベクトルを得る過程は２次形式の解法と同じである。

*2 WikiPedia では　特異値分解とレイーリーで極値を説明している。

例　2次曲線 $2x^2 - 2xy + 2y^2 = 9$ を標準化する

2 次形式　.

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

ルベリエ・ファディーエフ法　ルベリエ・ファディーエフ法は2次形式と馴染みが良い。

- マトリクス

```
   A=:2 _1,: _1 2

   A
 2 _1
_1  2
```

- 固有値

```
   char_lf A
+-+---+------+
|1|3 1|3 _4 1|
+-+---+------+
```

- 特性方程式

$$fx = 3 - 4x + x^2 = (x-3)(x-1) = 0$$

- 固有ベクトル

```
a=.   pick_evec A
 0.707107 _0.707107
_0.707107 _0.707107
```

- 固有ベクトルを用いて対角化する

```
   a mp A mp a    NB. mp +/ . *  (内積)
3 0
0 1
```

楕円　楕円方程式となる

$$u^2 + 3v^2 = 9$$

$$\frac{u^2}{3^2} + \frac{v^2}{\sqrt{3}^2} = 1$$

4. ラグランジュの未定乗数法

最適化はラグランジュの未定乗数法を用いるのが早道である。

- 条件式

$$u^2 + v^2 = 1$$

$$Q(u, v) = au^2 + 2buv + cv^2 = x^{'}Ax$$

- ここで

$$x = \begin{pmatrix} u \\ v \end{pmatrix}, \qquad A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

-

$$F(u, v) = Q(u, v) - \lambda(u^2 + v^2 - 1) = au^2 + 2buv + cv^2 - \lambda(u^2 + v^2 - 1)$$

- $F(u, v)$ を u と v で夫々偏微分して、$\frac{1}{2}$ し, 2 次形式にする

$$au + bv - \lambda u$$

$$bu + cv - \lambda v$$

- .

$$A\lambda\mathbf{x} = 0$$

$$A\lambda = A - \lambda I = \begin{pmatrix} a - \lambda & b \\ b & c - \lambda \end{pmatrix}$$

5. 主成分分析のメインの式

$u^2 + v^2 = 1$ の条件式の下で

$U(a, b, c)$ の最小値を与える a, b を求める

回帰分析は (2 次の場合）x を説明変数として y の挙動を回帰直線で求める。

主成分分析は x, y を総合した主軸を求める。

3　クラスター分析

クラスター分析の *J* のアドオンが *Dentrite* として *J6* のアドオンのパッケージで提供されいたが、*J8* ではまだ提供されていない。

しかし、*jsoftware.com* の *ShowCase/Essay* に *Dendrite* として簡単なスクリプトと解説が入っている。このクラスター分析をスクリプトを *J8* で通るようにして実践してみる

クラスター分析にはいろいろな手法が提案されている。色々な手法で試してみて、安定した結果が得られるケースが好ましいようだ。

J のアドオンのクラスター分析は *Roger Hui* と *Oreg　Kovchenko* によって作成され、簡潔で高い機能が装備されている。

距離	ユークリッド距離
クラスターの組み合わせ	*Kruskal* 法
樹形図	*viewmat* を用いる

距離	ユークリッド距離
	$P(x_1, y_1) \Longleftrightarrow Q(x_2, y_2)$
	$d = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$
		`dist=:+/&.:*:@:-"1/~`
		Script を途中で 2 分し、`&.` で *2* 乗したものを合計した後、 2 乗の逆関数の平方根作用させる名人芸。
		易しく書き直すと
		`dist2=: %:@:+/@:*:@:-"1/~`

Kruscal 法	*graph* 理論で各頂点が *1* 辺を持てる場合に全体を統合する最短、最適な組み合わせを求める方法。（*Joseph B. Kruskal 1956*）をクラスターの組み合わせに応用したもの。	

3.1 Worked Example

サンプルデータ　*data3*（鈴木）*7* 種類の犬の下顎骨に関する *6* 種類の計測値

```
        dat3
        x1  x2  x3   x4   x5   x6
 9.7   21 19.4  7.7   32 36.5  NB. 0
 8.1 16.7 18.3    7 30.3 32.9  NB. 1
13.5 27.3 26.8 10.6 41.9 48.1  NB. 2
11.5 24.3 24.5  9.3   40 44.6  NB. 3
10.7 23.5 21.4  8.5 28.8 37.6  NB. 4
 9.6 22.6 21.1  8.3 34.4 43.1  NB. 5
10.3 22.1 22.1  8.1 32.3   35  NB. 6
```

標準化　.

```
        stand dat3
_0.499959 _0.498656 _0.943027   _0.741868  _0.49148 _0.612915
 _1.51806  _1.92814  _1.35097      _1.391 _0.864003  _1.30554
  1.91802    1.5957   1.80129      1.9474   1.67792   1.61887
 0.645401  0.598387  0.948325    0.741868   1.26157  0.945483
 0.136352  0.332437  _0.20132 1.64728e_15   _1.1927 _0.401281
 _0.56359 0.0332437 _0.312576   _0.185467 0.0344349  0.656891
_0.118172 _0.132975  0.058277   _0.370934 _0.425741 _0.901507
```

ユークリッド距離　.

```
        dist2 stand dat3
           0 2.07032 5.88927 3.71427 1.65306 1.69794 1.22772
2.07032       0 7.76655 5.52131 3.46837 3.45657 2.92645
5.88927 7.76655       0 2.32849 4.98943 4.60909 5.13077
3.71427 5.52131 2.32849       0 3.16846 2.41247 3.06706
1.65306 3.46837 4.98943 3.16846       0 1.80327 1.15101
1.69794 3.45657 4.60909 2.41247 1.80327       0 1.74308
1.22772 2.92645 5.13077 3.06706 1.15101 1.74308       0
```

階層化　ユークリッド距離をもとに階層化を行う。上から下へ順に階層化がなされる。

クラスカルアルゴリズムはグラフ理論を用いて、描いた距離ノード上の最短距離から順にグラフの線上に太線を塗り重ねて行って、全部塗り終わったら木が完成するというものである。

*3

*3 詳細は Wikipedia にクラスカル法としてあがっている

クラスカル法によるクラスターの生成過程の一覧

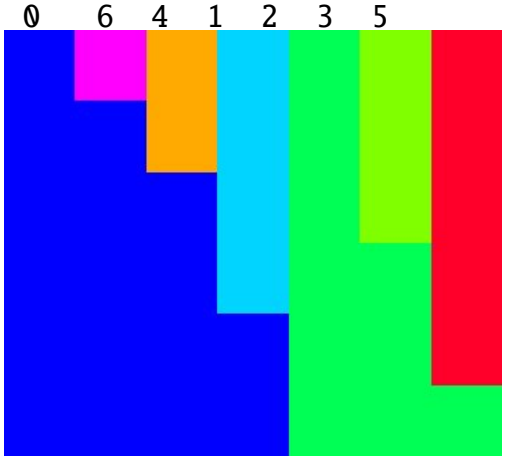
```
   boxclust 1{:: (# mstc edgesort) dist2 dat3
+-------------+-----+---+-+-+-+
|0 6          |1    |2  |3|4|5| NB. cluster 0-6
+-------------+-----+---+-+-+-+
|0 4 6        |1    |2  |3|5| | NB. cluster 0-4-6
+-------------+-----+---+-+-+-+
|0 4 6        |1    |2 3|5| | | NB. cluster 2-3
+-------------+-----+---+-+-+-+
|0 1 4 6      |2 3  |5  | | | | NB. cluster 0-1-4-6
+-------------+-----+---+-+-+-+
|0 1 4 6      |2 3 5|   | | | | NB. cluster 2-3-5
+-------------+-----+---+-+-+-+
|0 1 2 3 4 5 6|     |   | | | |  NB. Whole
+-------------+-----+---+-+-+-+
```

デンドログラム　デンドログラムで表す距離と構造を求める

```
   (icicle (# mstc edgesort)) dist2  dat3
+---+-------------+-------+
|0 6|0 6|4|1|2|3|5|3.37046|
|4 6|0 6 4|1|2|3|5| 4.6669|
|2 3|0 6 4|1|2 3|5|5.98665|
|0 1|0 6 4 1|2 3|5|6.21289|
|3 5|0 6 4 1|2 3 5|7.25741|
|0 5|0 6 4 1 2 3 5|7.42563|
+---+-------------+-------+
```

デンドログラム　　*viewmat* で表示

```
silview dat3
```




```
0 6 4 1 2 3 5
| | | | | | |
⊥ ⊥ | | | | | 3.37
|   | | | | |
⊥ – ⊥ | | | | 4.67
|     | | | |
|     | ⊥ ⊥ | 5.99
|     | |   |
⊥ – – ⊥ |   | 6.21
|       |   |
|       ⊥ – ⊥ 7.26
|       |
⊥ – – – ⊥     7.43
    |
```

viewmat の図は慣れが必要。樹形図は簡単に手書きできる

3.2 Script

1. *dist2* ユークリッド距離

```
NB. Cluster analysis
NB. Script is written in Jsoftware.com Essay/Dendrite
 dist=:+/&.:*:@:-"1/~
 dist2=: %:@(+/)@:*:@:-"1/~   NB. same dist
```

2. *boxclust*

```
NB. Usage:  7 mstc dat1
mstc=: 4 : 0
 z=. 0 2$f=. i.k=. x[w=. 0$~0,x
 for_e. y do.
  if. ~:/j=. |.^:(>/) e{f do.
   z=. z,e
   if. 1=k=.<:k do. z;w,x#0 return. end.
   w=.w,f=. ({.j) (f I.@:= {:j)} f
  end.
 end.
 assert. 0 [ 'graph is not connected'
)
```

```
  boxclust=: (</. i.@#)"1
NB. USAGE:   boxclust 1{::7 mstc dat1
NB. USAGE:  |:edgesort dist2 dat2
NB. Usage:D=: boxclust 1{:: (# mstc edgesort) dist2 dat2

   od2=: 2&# #: i.@*:   NB. 2 coord odometer
   edgesort=: ((] /: ({˜ <"1)) (#˜ </"1)) od2@#
```

3. *icicle*

```
NB. Usage:   (icicle (# mstc edgesort)) dist2 4|. dat2
nok=: 3 : 0
  z=. (i.{:$y),}:y
  for_i. i.&.<:#y do. z=. z /:"1 i{z end.
)

trimbox=: {.@(];.0˜ _2 _2&({.!.1)@(-&2)@{:@$)@":
iceview=: (# {. }.) ([: trimbox </.)"1 {.
linkage=: {˜ <"1
icicle=: {.@] , <@:iceview@nok@(1&{::)@] , (<@,.@linkage 0&{::)
```

4. *silview*

```
silnok=: 3 : 'nok 1{:: (# mstc edgesort) dist y'
silview=: viewmat@silnok
NB. Usage: silview dat2
```

付録 A　CSV ファイルの取扱い

A.1　CSV ファイル

CSV ファイルからの入力

1. *CSV* ファイル *(*カンマファイル）
   CSV ファイルの例

```
165, 53, 86, 56, 92
160 ,47, 84, 52, 92
166 ,55, 86, 64, 89
164, 56, 90, 60, 95
168, 55, 87, 56, 87
164, 54, 87, 57, 92
```

2. *CSV* ファイルの読み込み
   CSV ファイルにはデータのみで、コメントは書き込めない

```
require 'files csv'
readcsv 'c:/temp/style2.csv'
```

```
+----+---+----+---+---+
|165 | 53| 86 | 56| 92|
+----+---+----+---+---+
|160 |47 | 84 | 52| 92|
+----+---+----+---+---+
|166 |55 | 86 | 64| 89|
+----+---+----+---+---+
|164 | 56| 90 | 60| 95|
+----+---+----+---+---+
```

- `DAT=:  ".@> readcsv 'c:/temp/style2.csv'`
- 読み込んだ状態ではボックスに入った文字列
-
- ボックスを開いて *(* > *(*数値化 *(* `".` *)* する
- 空欄があると左詰めされるので *0* やダミーデータを入れておく
- *CSV* ファイルの作成は *EXCEL* で打ち込んで *CSV* で出力すると便利

付録 B　EXCEL ファイルの読み書き

EXCEL や *LibreCalc* 本体は用いないで、ファイルのみを用いる。

B.1　xls 形式-bif8

tara は *EXCEL* や *LibreCalc* の *bif8* 型式のファイルの読み書きを行う。

bif8 は *xls* で保存されるカンマデータの型式で、*EXCEL2003* までに用いられた。新しい *EXCEL* では *xls* 形式で保存すればよい。

最新の *EXCEL* は *xml* を用いてデータが組上げられている。

xml に対応した *taraxml* も提供されている。

1. *tara* は addon/tables に置かれている
   J の *File/Open* でロードするのが簡潔
2. *tara* での *EXCEL* ファイルの 読み込み

```
   DAT=:   readexcel 'c:/temp/style2.xls'
   DAT
+---+--+--+--+--+
|165|53|86|56|92|
+---+--+--+--+--+
|160|47|84|52|92|
+---+--+--+--+--+
|166|55|86|64|89|
+---+--+--+--+--+
|164|56|90|60|95|
+---+--+--+--+--+
|168|55|87|56|87|
+---+--+--+--+--+
```

3. ボックスを > で開く
4. *EXCEL* への書き込み
   DAT2 writexlsheets jpath ' temp/tarasmp1.xls'

B.2　xml 形式の xlsx

taraxml も提供されている。*(*読み込めなかった*)*

付録 C　LAPACK

線形計算で定評のある *LAPACK* は *J8* でも用いることが出来る。

addons/math/lapack に置かれている。ここにチュートリアル *(lapack.ijt)* があるので *JAPACK* の全貌はこれを印刷して読むか、*Help/Studio/Lab* で動かしてほしい。

実数の固有値と固有ベクトルを求めるのは次の *2* 本のファイルを読み込む

- *lapack.ijs*
- *dgeev.ijs* または *geev.ijs*

J8 ではロケールが要求されているようだ。 _jlapack_

```
   ,. }. dgeev_jlapack_ cor stand STYLE
+-------------------------------------------------+
|2.14898 1.30735 0.358586 0.675427 0.509665       |
+-------------------------------------------------+
|_0.521033 _0.236935 _0.557029  0.534739   0.27597|
|_0.570928 _0.104455  0.626247 _0.198549  0.481168|
|_0.419977  0.483214  0.279874  0.401878 _0.591852|
|_0.399578 _0.488062 _0.185005 _0.513792 _0.551293|
| _0.25792  0.679155  _0.43009 _0.499144  0.195203|
+-------------------------------------------------+
```

付録 D　主成分分析　手計算で手順を追う

有馬・石村に今では貴重な手計算の例題があった。ところどころ計算機の力を借りながら手順を追ってみよう。

例題　サンプルデータ　関東 7 都県の人口 *10* 万人当たりのスポーツ施設 x_1 と教育施設 x_2

```
   dat4
22.9 13.7
24.9 16.2
19.3 11.3
  22 10.4
28.6 24.9
42.6 26.5
41.3 20.3
```

社会生活統計指標 *(*総理府統計局*)1982*

情報損失量を求める

22.9	13.7	$\lvert 22.9a_2 - 13.7a_1 + a_0\rvert$
24.9	16.2	$\lvert 24.9a_2 - 16.2a_1 + a_0\rvert$
⋮		

多項式　.

$$6339.32a_2^2 + 2420.33a_1^2 + 7a_-^2 - 7686.86a_2a_1 + 403.21_2a_0 \quad - 246.6a_1a_0$$

Script　手計算は大変なので *Script* を書いた

```
arima0=: 3 : 0
a2=.  1 _1 1   *(L:0)  { y,.1   NB. dat4
+/  ,/ > ,.op L:0 */˜ L:0 a2
)

arima1=: 3 : '+/ >*/˜ L:0 {1 _1 1 * "1  y ,.1 '
  arima0 dat4
6339.32 2420.33 7 _7686.86 403.2 _246.6
```

行列形式　.

```
   arima1 dat4
 6339.32 _3843.43  201.6
_3843.43  2420.33 _123.3
   201.6   _123.3      7
```

ラグランジュの未定剰余法　　与式

$$F(a_2, a_1, a_0, \lambda) = U(a_2, a_1, a_0) - \lambda(a_2^2 + a_1^2 - 1)$$

$$= 6339.32a_2^2 + 2420.33a_1^2 + 7a_0^2 - 7686.86a_2a_1 + 403.2a_2a_0 - 246.6a_1a_0 - \lambda(a_2^2 + a_1^2 - 1)$$

偏微分　.

$$\begin{cases} \dfrac{\partial F}{\partial a_2} &= 12678.6a_2 - 7696.86a_1 + 403.2a_0 - 12\lambda a_2 = 0 \\ \dfrac{\partial F}{\partial a_1} &= 4840.66a_1 - 7686a_2 - 246.6a_0 - 2\lambda a_1 = 0 \\ \dfrac{\partial F}{\partial a_0} &= 403.2a_2 - 246.6a_1 + 14a_0 = 0 \end{cases}$$

[*4]

整理する

$a_0 = 28.8a_2 - 17.6143$ として上 *2* 式に代入して a_0 を消す

当初データの平均値が現れる

```
  403.2 246.6 % 14
28.8 17.6143

   mean dat4
28.8 17.6143
```

$$\begin{cases} 533.2a_2 - 292.5a_1 - \lambda a_2 = 0 \\ -292.5a_2 + 248.2a_1 - \lambda a_1 = 0 \end{cases}$$

行列表示　.

$$\begin{pmatrix} 533.2 - \lambda & -292.5 \\ -292.5 - \lambda & 248.2 \end{pmatrix} \begin{pmatrix} a_2 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

[*4] $\dfrac{\partial F}{\partial a_2} = 2a_2 + a_1 + a_0 - 2\lambda a_2 = 0$

```
固有値と固有ベクトルを求める          a=: 533.2 _292.5,. _292.5 248.2
         char_lf a
      +-+---------------+--------------+
      |1|716.065 65.3348|46784 _781.4 1|
      +-+---------------+--------------+
         pick_evec a
        0.84793 _0.530109
      _0.530109  _0.84793
本文の方法     固有ベクトルは等しくなる
          cov dat4
      905.617 549.061
      549.061 345.761

         char_lf cov dat4
      +-+---------------+----------------+
      |1|1241.99 9.38736|11659 _1251.38 1|
      +-+---------------+----------------+

         pick_evec cov dat4
      0.852703  0.522396
      0.522396 _0.852703
```

References

有馬　哲　石村貞夫「多変量解析のはなし」東京図書　*1987*

鈴木義一郎「統計分析へのいざない」マーケティングサイエンス研究所　*1998*

藤原偉作「楽しく学べる多変量解析法」　現代数学社　*1985*