

## 主成分分析

### 1. 目的

例題を提示して説明します。

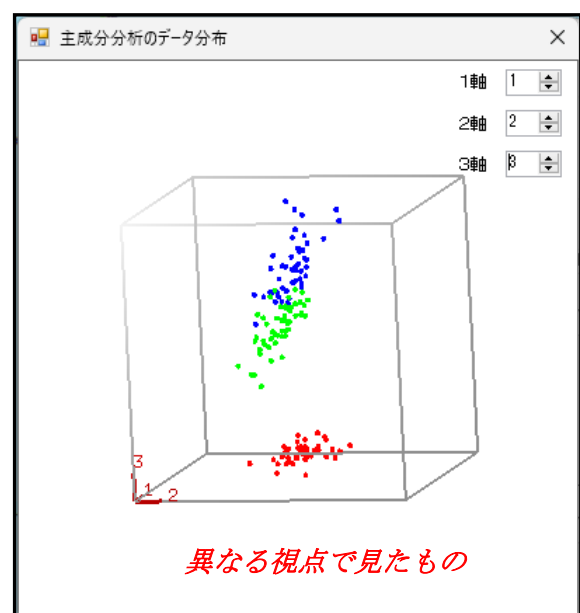
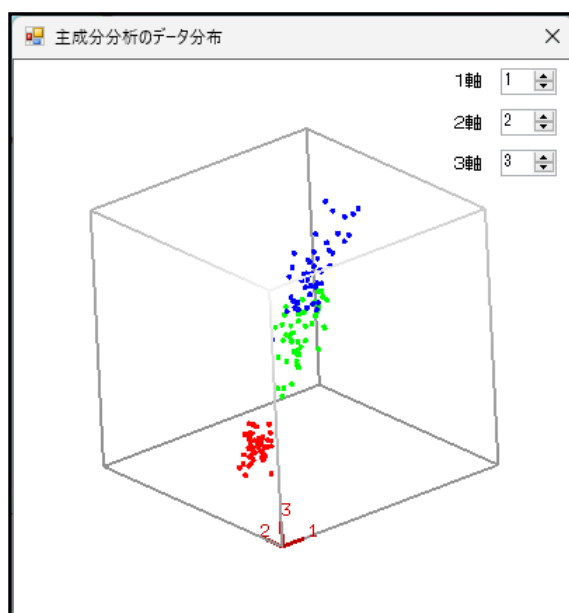
植物のアヤメの3種類（セトーサ、ヴェルシコロール、ヴィルジニカ）について、がくの長さ、がくの幅、花弁の長さ、花弁の幅を計測した150件のデータがあります。（FisherのIrisデータセットと呼ばれます）

ID	種類	がくの長さ	がくの幅	花弁の長さ	花弁の幅
1	セトーサ	5.1	3.5	1.4	0.2
...	.....	.....	.....	.....	.....
50	セトーサ	5.0	3.3	1.4	0.2
51	ヴェルシコロール	7.0	3.2	4.7	1.4
...	.....	.....	.....	.....	.....
100	ヴェルシコロール	5.7	2.8	4.1	1.3
101	ヴィルジニカ	6.3	3.3	6.0	2.5
...	.....	.....	.....	.....	.....
150	ヴィルジニカ	5.9	3.0	5.1	1.8

上記の計測情報から、種類を区別するための特徴を数量で表現できないかを考えます。

具体的には、がくの長さ、幅、花弁の長さ、幅の4種類の変量より特徴的な指標に集約して、解釈しやすくすることです。

上記の4変量のデータを、まず3次元で表現すると以下のようなようです。



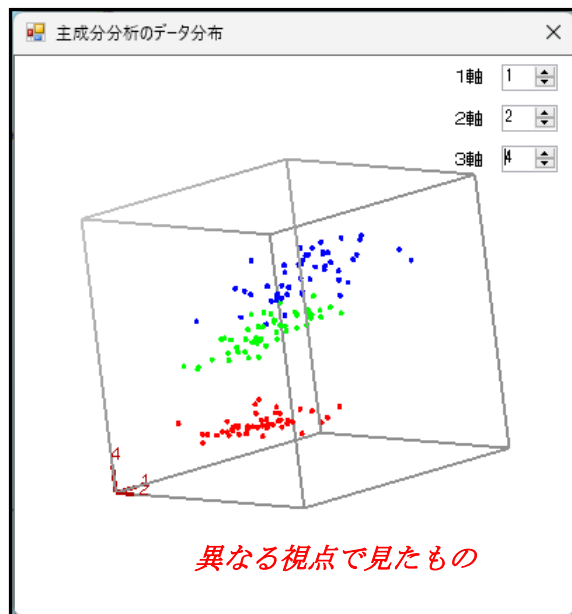
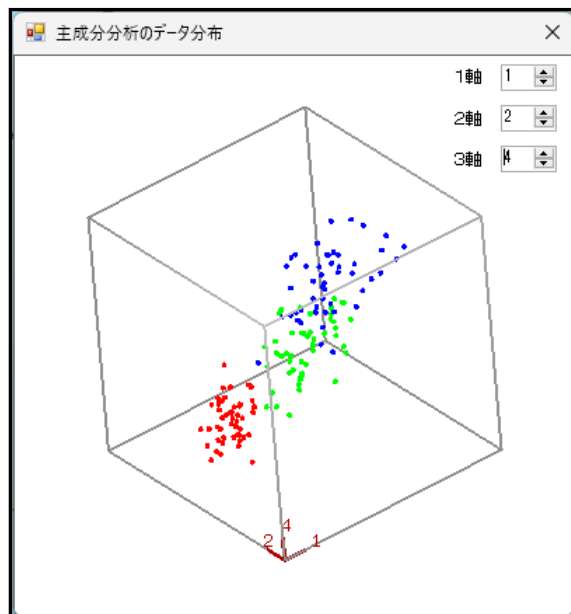
4変量のデータですので、4次元で表現できれば最善ですが、残念ながら、私たちは4次元でものを見る能力がありません。（ある高名な数学者は、4次元でものを見て思考できるという話を聞いたことがあります....）

したがって、3変量に絞って3次元で表現したのが 上記の絵です。  
この場合は 第1軸にがくの長さ、第2軸にがくの幅、第3軸に花卉の長さを選んで表示したものです。

赤い点がセトーサ、緑の点がヴェルシコロール、青の点がヴィルジニカ となります。

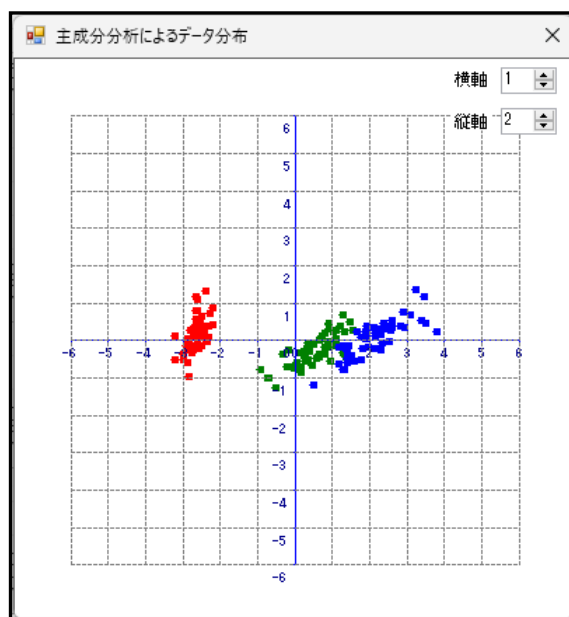
3次元空間上で なんとなく 3種類が区別できそうです。

ついでに、第3軸に花卉の幅に変更して表示したものが以下です。



先と同様に、赤い点がセトーサ、緑の点がヴェルシコロール、青の点がヴィルジニカです。  
3次元空間上でなんとなく3種類が区別できますが、変量の組み合わせで、位置が異なってくると 判断に困ります。

今度は、4つの変量を集約して表現したものが以下です。



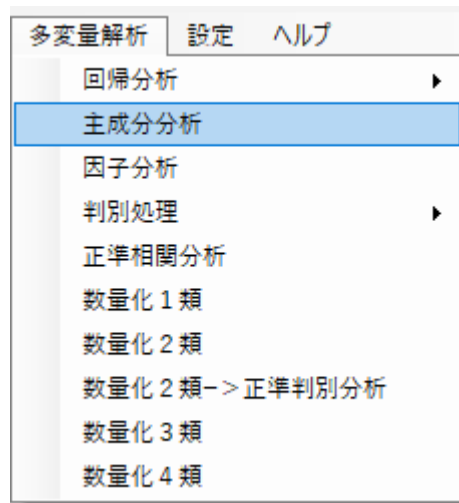
2つの指標(2次元)にまとめてますので、理解はしやすいです。

このように 複数の変量から 少ない指標に集約して表示する技術を 主成分分析と言います。

## 2. 使用法

### (1) メニューの選択

メニューの「多変量解析→主成分分析」を選択します。



### (2) パネルが表示されます。

主成分分析

計算実行 ☒ 標準化する ☐ 固有値のスクリーンプロット 因子数 : 0 主成分スコア

固有値、寄与率、固有ベクトル:

	N0	主成分	1	2	3	4	5

主成分負荷量:

	N0	主成分	1	2	3	4	5	寄与率

主成分スコア:

	N0	ID	1	2	3	4	5

変量ごとの平均、分散:

	項目	数	合計	平均	(標本)分散

☐ 先頭行をラベルとして使用

	N0	ID	Group	X1	X2	X3	X4	X5
▶*								

(3) データの入力

パネルのグリッド（下の部分）にデータを入力します。

表データを貼り付け ☐ 先頭行をラベルとして使用 クリア データ分布 相関/分散行列

	N0	ID	Group	X1	X2	X3	X4	X5
▶								

以下のように前もって表計算ソフトで定義しておきます。

ID	あやめの種類	がくの長さ	がくの幅	花卉の長さ	花卉の幅
1	セトーサ	5.1	3.5	1.4	0.2
2	セトーサ	4.9	3.0	1.4	0.2
...	.....	.....	.....	.....	.....
150	ヴィルジニカ	5.9	3.0	5.1	1.8

データの項目は、

(1) ID

(2) Group （所属を示す文言）

これは何でもいいのですが、種類であったり、所属するクラスだったり、データの種別を示すものです。

全部 同じ扱いでよければ 同じ文言でも構いません。

(3) X1 : データ 1

(4) X2 : . . . . .

というものです。

データの赤線部分をコピーします。表題部分も取り込みますので、

「先頭行をラベルとして使用」にチェックを入れます。

表データを貼り付け

をクリックすると、グリッド部分にコピーされます。

変量ごとの平均、分散:

	項目	数	合計	平均	(標本)分散
▶	がくの長さ	150	876.5	5.843333	0.6811222
	がくの幅	150	458.6	3.057333	0.1887129
	花卉の長さ	150	563.7	3.758	3.095503

表データを貼り付け ☒ 先頭行をラベルとして使用 クリア データ分布 相関/分散行列

	N0	ID	あやめの種類	がくの長さ	がくの幅	花卉の長さ	花卉の幅
▶	1	1	セトーサ	5.1	3.5	1.4	0.2
	2	2	セトーサ	4.9	3.0	1.4	0.2
	3	3	セトーサ	4.7	3.2	1.3	0.2

下のグリッドには入力データがそのまま反映されます。

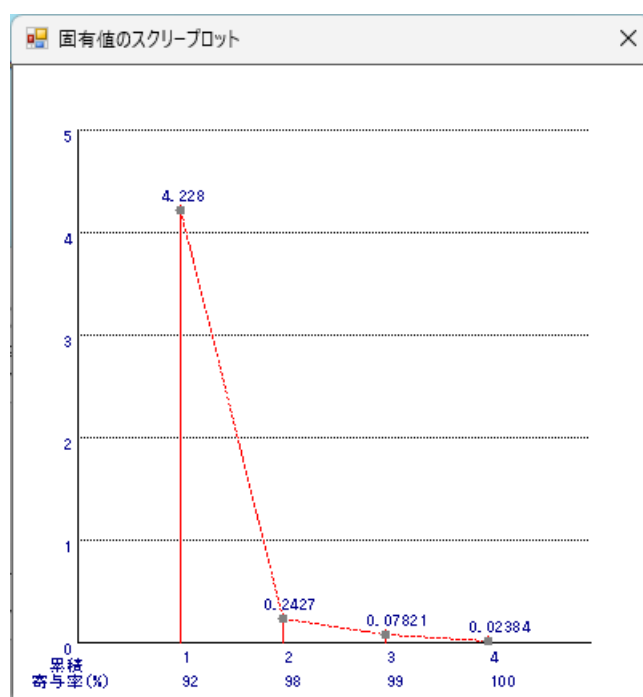
上のグリッドには、変量毎の集計が表示されます。

#### (4) 固有値のスクリープロット

例題のデータは、4つの変量で表現されています。  
したがって、最大4つの指標が定義されます。

しかし4つの変量を4つの指標で説明されても、うれしくありません。  
せいぜい2変数程度の少ない指標で、データの特徴が捉えられる方が  
良いからです。

固有値のスクリープロットをクリックすると、その目安を判断できます。



データは4つの変量ですので、4つの固有値が計算されます。  
大きい順に、 4.228、0.2427、0.07821、0.0238 です。  
この場合の固有値は、データの分散を表現していると言えます。  
下に 累積寄与率 (%) が表示されていて、左から 92、98、99、100  
となっています。  
これは、最初の 4.228 で、全体の傾向の 92%が表現できている、  
ということを意味します。  
続く 0.2427 で、最初と合わせて、全体の傾向の 98%が表現できて  
いることを意味します。

通常は、全体の 95%までの傾向が表現できるところで、指標の数を絞ります。  
ですので、ここでは 2つの指標に 絞ることにします。

(5) 因子数の決定

デフォルトでは、元のデータの 변수数 4 が設定されています。

因子数 : 4

ここで、指標の数を 2 にしますので、2 を設定します。

因子数 : 2

(6) 計算実行

因子数を決定した段階で、**計算実行** を押して、計算実行します。

固有値、寄与率、固有ベクトル:						使用法	考え方
	No.	主成分	1	2			
▶	-	固有値	4.228242	0.24267...			
	-	寄与率(%)	92.46187	5.306648			
	-	累積寄与率(%)	92.46187	97.76852			
	1	がくの長さ	0.36138...	0.65658...			
	2	がくの幅	0.0045...	0.72015...			
主成分負荷量:							
	No.	主成分	1	2	寄与率		
▶	1	がくの長さ	0.89740...	0.39060...	0.95790...		
	2	がくの幅	-0.3987...	0.82522...	0.84000...		
	3	花弁の長さ	0.997874	-0.0483	0.99809		
主成分スコア:							
	No.	ID	1	2			
▶	1	1	-2.6841...	0.31939...			
	2	2	-2.7141...	-0.1770...			
	3	3	-0.8889	-0.1449			

上記 3 つの表で 計算結果が示されました。

いづれも グリッドが小さいため、データの全部を見ることはできません。  
スクロールして、見ることはできますが、グリッドの左上隅のセルをクリック  
して、Ctrl-C を押すと、表形式のデータがコピーされます。  
その際、グリッドが 以下のように青くなるのを確認してください。

ここをクリックして Ctrl-C を押す

固有値、寄与率、固有ベクトル:						使用法	考え方
▶	No.	主成分	1	2			
	-	固有値	4.228242	0.24267...			
	-	寄与率(%)	92.46187	5.306648			
	-	累積寄与率(%)	92.46187	97.76852			
	1	がくの長さ	0.36138...	0.65658...			
	2	がくの幅	0.0045...	0.72015...			

表計算ソフト上で、Ctrl-V を押すと、データがそのままコピーされます。

No.	主成分	1	2
–	固有値	4.2282E+00	2.4267E-01
–	寄与率(%)	9.2462E+01	5.3066E+00
–	累積寄与率(%)	9.2462E+01	9.7769E+01
1	がくの長さ	3.6139E-01	6.5659E-01
2	がくの幅	-8.4523E-02	7.3016E-01
3	花弁の長さ	8.5667E-01	-1.7337E-01
4	花弁の幅	3.5829E-01	-7.5481E-02

上記表の 最初の 3 行は 主成分の 固有値、寄与率、累積寄与率を表示しています。

表現に必要な指標の数を 累積寄与率の 95% を超えるところまでとしましたので、ここでは 指標数を 2 としたわけです。

続く 4 行は、主成分ベクトルを表示しています。

各変量に対する 係数ベクトル  $\{a\} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$  になります。

第 1 主成分ベクトルは、 $\{a_1\} = \begin{pmatrix} 3.6139E-01 \\ -8.4523E-02 \\ 8.5667E-01 \\ 3.5829E-01 \end{pmatrix}$

第 2 主成分ベクトルは、 $\{a_2\} = \begin{pmatrix} 6.5659E-01 \\ 7.3016E-01 \\ -1.7337E-01 \\ -7.5481E-02 \end{pmatrix}$

であることを示しています。

2 番目の表を同様に 表計算ソフトのシート状にコピーすると、以下のようになります。

この表は 主成分と元データとの相関係数である 主成分負荷量 および寄与率を示すものです。

No.	主成分	1	2	寄与率
1	がくの長さ	8.9740E-01	3.9060E-01	9.5790E-01
2	がくの幅	-3.9875E-01	8.2523E-01	8.4000E-01
3	花弁の長さ	9.9787E-01	-4.8381E-02	9.9809E-01
4	花弁の幅	9.6655E-01	-4.8782E-02	9.3659E-01



3 番目の表は主成分スコアを示すものです。

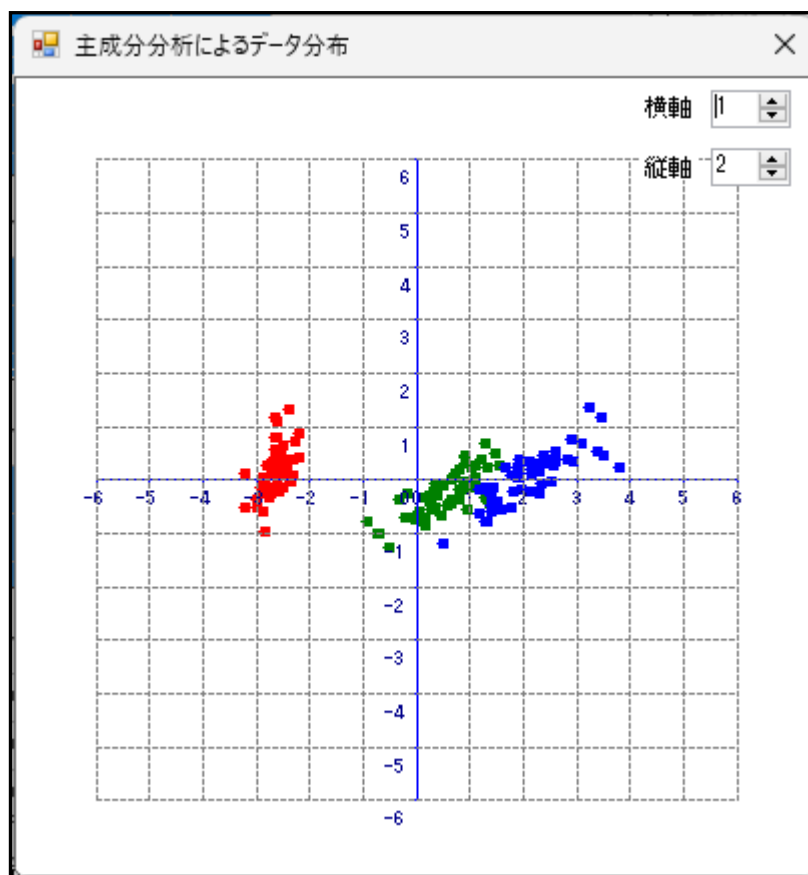
2 つの指標に集約された座標軸上での、各サンプルの位置を表示します。

No .	ID	1	2
1	1	-2.6841E+00	3.1940E-01
2	2	-2.7141E+00	-1.7700E-01
3	3	-2.8890E+00	-1.4495E-01
148	148	1.7643E+00	7.8859E-02
...	.....	.....	...
149	149	1.9009E+00	1.1663E-01
150	150	1.3902E+00	-2.8266E-01

各点の位置の確認は、

主成分スコア

を押してできます。



3 次元上で見るより、解釈しやすいと 思われます。