VI-2. データの構造.

VI-2-1. 主成分分析(PCA)

VI-2-1-1. 主成分分析（PCA）とは何か

主成分分析（Principle component analysis）を最も簡単に説明すると、観察された変数の座標空間にプロットされたデータ分布を、固有ベクトルを座標軸とする座標空間に移し替える技術です。それは、実際に見えるものの関係の絵を、見えない潜在的な成分の関係に組み立てなおすことです。ここでは一旦、潜在的な成分どうしは独立していることにして、観察された変数は、潜在的成分の組み合わせ（一次結合）によって説明できることにします。正直に言えば、より上位のレベルでの潜在的な因子同士に相関があっても、確かめようがありませんから、その独立性を仮定することの現実性など分かりません。それでも、主成分分析をすることによって、より抽象化された視点から現象の構造が見えてくるかもしれません。

主成分分析の前提条件は、データが正規分布していることですが、この条件はあまり厳しくありません。データの頻度分布が単峰形ならば主成分分析ができるとゆるく考えておいてください。

VI-2-1-2. 主成分分析の操作

データは、もともとの実態を何かのベクトルへ映し出した写像として描かれています。数学的に、主成分分析はその写像を固有ベクトルへの写像に作り変えます。それは、もとのデータの座標を含み固有ベクトルと直交する平面と原点への距離を計算することにほかなりません。

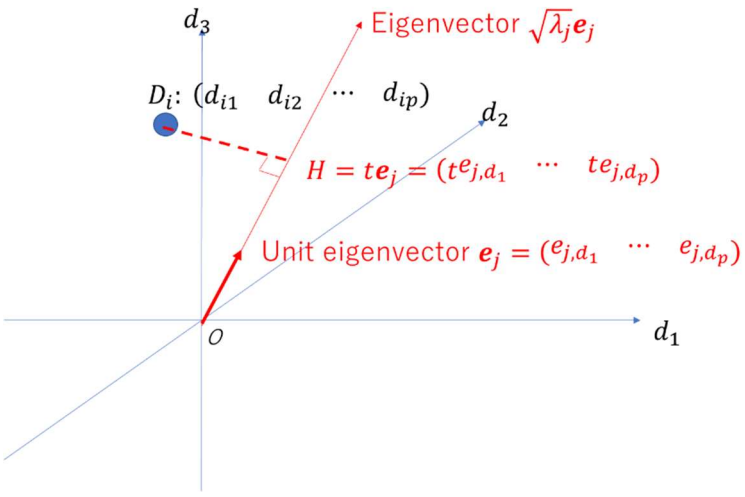

図 73. 観測された変数の座標から固有ベクトルの座標への変換

図 73 にその座標変換のデータと固有ベクトルの関係を示しました。標本サイズがn、観測変数がp個として、データを$D_i: (d_{i,1} \quad \cdots \quad d_{i,p})$ $((i = 1, \cdots, n)$として、単位行列化した固有ベクトルを $\boldsymbol{e}_j: (e_{j,d_1} \quad \cdots \quad e_{j,d_p})$, とします。 Hは原点から平面に下した垂線の脚で、ベクト

ル$\overrightarrow{OH}$長さがtの固有ベクトルです。ここではj番目の固有ベクトル（j番目の固有ベクトル）について示しました。図には長さが標準偏差の固有ベクトルも示しました($\sqrt{\lambda_j}\boldsymbol{e}_j$).

$$\overrightarrow{OH} = t\boldsymbol{e}_j = (te_{j,x_1} \quad \cdots \quad te_{j,x_p})$$

$$\overrightarrow{HD_i} = (d_{i,1} - te_{j,x_1} \quad \cdots \quad d_{i,p} - te_{j,x_p})$$

$$\overrightarrow{HD_i} \perp \overrightarrow{OH}$$

内積 $\overrightarrow{HD_i} \cdot \overrightarrow{OH} = 0$

$$(d_{i,1} - te_{j,x_1} \quad \cdots \quad d_{i,p} - te_{j,x_p}) \begin{pmatrix} te_{j,d_1} \\ \vdots \\ te_{j,d_p} \end{pmatrix} = td_{i,1}e_{j,d_1} - t^2{e_{j,d_1}}^2 + \cdots + td_{i,p}e_{j,d_p} - t^2{e_{j,d_p}}^2$$

$$= t\left(d_{i,1}e_{j,d_1} - t{e_{j,d_1}}^2 + \cdots + d_{i,p}e_{j,d_p} - t{e_{j,d_p}}^2\right)$$

$$= t\left(d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p} - t\left({e_{j,d_1}}^2 + \cdots + {e_{j,d_p}}^2\right)\right)$$

$$= t\left(d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p} - t\right)$$

$\because$ $\boldsymbol{e}_j$ は単位ベクトル

${e_{j,d_1}}^2 + \cdots + {e_{j,d_p}}^2 = 1$

$$t\left(t - d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}\right) = 0$$

$$t = 0 \; or \; t = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

前提条件から

$$t \neq 0$$

したがって

$$t = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

これがデータi のj番目の主成分の主成分得点です。

$$\mathrm{PCS}_{i,j} = t_{ij} = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

このような計算で以下の主成分得点表が作れます.

主成分得点

	PC1	PC2	…,	PCp
標本番号				
1	$\mathrm{PCS}_{1,1}$	$\mathrm{PCS}_{1,2}$	…	$\mathrm{PCS}_{1,p}$
2	$\mathrm{PCS}_{2,1}$	$\mathrm{PCS}_{2,2}$	…	$\mathrm{PCS}_{2,p}$
⋮	⋮	⋮	⋱	⋮
n	$\mathrm{PCS}_{n,1}$	$\mathrm{PCS}_{n,2}$	…	$\mathrm{PCS}_{n,p}$

これらをもとに２つあるいは３つの主成分を直交軸として分布図を書くことが出来ます。

VI-2-1-3. 分散共分散行列の対角化と主成分分析

二つの主成分分析の方法があります。データが同じでも、この二つの主成分分析の結果とその解釈は異なります。分散共分散行列を対角化するのが一つの方法で、もう一つは相関行列を対角化します。分散共分散行列も相関行列も対称行列で二次形式です。そのような行列の対角化によって得られた固有値が、上記の説明で使った固有値であることを示します。その過程で、二次形式の行列の空間幾何学と対称行列の性質を使います。行列が正定置であれば、二次形式の行列は、多次元空間において傾いた超楕円を表し、固有ベクトルは超楕円の軸を表しています。これは V-2-4 で説明しました。もう一つ V-2-2 で紹介した、対称行列($\boldsymbol{P}$)の性質、その対角化行列の転置行列($\boldsymbol{P^T}$)が逆行列($\boldsymbol{P^{-1}}$)だということも使います。具体的には、次の式を使います。

一般的な対角化の式

$$\boldsymbol{P^{-1}VP = \Lambda}$$

$\boldsymbol{V}$が二次形式ならば

$$\boldsymbol{P^T = P^{-1}}$$

したがって$\boldsymbol{V}$の対角化は次の式になります。

$$\boldsymbol{P^TVP = \Lambda}$$

そもそも、分散共分散行列は次のように作ります。

$$\boldsymbol{V = DD^T}$$

$$\boldsymbol{P^TVP = P^TDD^TP = (P^TD)(P^TD)^T}$$

$$\boldsymbol{P^TD} = \begin{pmatrix} \boldsymbol{e_1}^T \\ \boldsymbol{e_2}^T \\ \vdots \\ \boldsymbol{e_p}^T \end{pmatrix} \begin{pmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1p} & d_{2p} & \cdots & d_{np} \end{pmatrix}_{p\times n}$$

$$= \begin{pmatrix} e_{11} & \cdots & e_{j1} & \cdots & e_{p1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{1k} & \cdots & e_{jk} & \cdots & e_{pk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{1p} & \cdots & e_{jp} & \cdots & e_{pp} \end{pmatrix} \begin{pmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1p} & d_{2p} & \cdots & d_{np} \end{pmatrix}_{p\times n}$$

$$= \begin{pmatrix} \sum_{j=1}^{p} e_{j1}\, d_{1j} & \sum_{j=1}^{p} e_{j1} d_{2j} & \cdots & \sum_{j=1}^{p} e_{j1} d_{nj} \\ \sum_{j=1}^{p} e_{j2}\, d_{1j} & \sum_{j=1}^{p} e_{j2} d_{2j} & \cdots & \sum_{j=1}^{p} e_{j2} d_{nj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{p} e_{jp}\, d_{1j} & \sum_{j=1}^{p} e_{jp} d_{2j} & \cdots & \sum_{j=1}^{p} e_{jp} d_{nj} \end{pmatrix}_{p\times n}$$

$e_{jk}d_{ij}$ を $d_{i,j}e_{j,k}$と書き換えます。

$d_{i,j}e_{k,d_1} + \cdots + d_{i,p}e_{k,d_p} = \sum_{j=1}^{p} e_{jk}d_{ij} = t_{ik} = \mathrm{PCS}_{i,k}$

（これは、固有ベクトルと直交する、点$\boldsymbol{d}$を含む超平面と原点の距離）

$$\boldsymbol{P^T D} = \begin{pmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ t_{12} & t_{22} & \cdots & t_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1p} & t_{2p} & \cdots & t_{np} \end{pmatrix}_{p\times n} = (\boldsymbol{t_1} \quad \boldsymbol{t_2} \quad \cdots \quad \boldsymbol{t_n})$$

$$\boldsymbol{t_i} = \begin{pmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{ip} \end{pmatrix} = \begin{pmatrix} \mathrm{PCS}_{i,1} \\ \mathrm{PCS}_{i,2} \\ \vdots \\ \mathrm{PCS}_{i,p} \end{pmatrix}$$

$$\boldsymbol{D^T P} = (\boldsymbol{P^T D})^{\boldsymbol{T}} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{np} \end{pmatrix}_{n\times p} = \begin{pmatrix} \boldsymbol{t_1}^T \\ \boldsymbol{t_2}^T \\ \vdots \\ \boldsymbol{t_2}^T \end{pmatrix}$$

$$\boldsymbol{P^T V P} = \boldsymbol{P^T D D^T P} = \begin{pmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ t_{12} & t_{22} & \cdots & t_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1p} & t_{2p} & \cdots & t_{np} \end{pmatrix}_{p\times n} \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{np} \end{pmatrix}_{n\times p}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} {t_{i1}}^2 & \sum_{i=1}^{n} t_{i1}t_{i2} & \cdots & \sum_{i=1}^{n} t_{i1}t_{1p} \\ \sum_{i=1}^{n} t_{i2}t_{i1} & \sum_{i=1}^{n} {t_{i2}}^2 & \cdots & \sum_{i=1}^{n} t_{i2}t_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} t_{1p}t_{i1} & \sum_{i=1}^{n} t_{ip}t_{i2} & \cdots & \sum_{i=1}^{n} {t_{i2}}^2 \end{pmatrix}_{p\times p}$$

二次形式の対称行列の固有ベクトルは互いに直交しています。したがって、その直交ベクトルに投影したベクトル同士も直交しています。

$$\sum_{i=1}^{n} t_{ij}t_{ik} = \delta_{jk} \sum_{i=1}^{n} t_{ij}t_{ik}$$

$$\delta_{jk} = \begin{cases} 1 \ (j = k) \\ 0 \ (j \neq k) \end{cases}$$

δ_{jk} は クロネッカーのデルタ

$$\boldsymbol{P^T V P} = \boldsymbol{P^T D D^T P} == \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}_{p\times p}$$

以上のように、λ_i は主成分iの平方和　(固有ベクトル i 上のデータの平方和)です。

VI-2-1-4. 主成分分析の結果の記述.

主成分分析の意義の一つはデータの集約です。主成分分析の計算では、元データの変数の数と同じ数の主成分が出来ます。しかし、いくつかの主成分は、元のデータの変数の分散

に比べて大きく、またある主成分は、元のデータの分散に比べて分散が小さいはずです。現象に対する影響の小さい主成分について考える必要はあまりないでしょう。そこで、まず、重要な主成分を選び出します。具体的には、全分散に占める割合の大きい分散を持つ主成分を選ぶのです。主成分の分散は固有値λです。全分散は、対角成分のトレース、つまり、対角因子の和です。

$$V_{total} = \sum_{j=1}^{p} \lambda_j$$

	寄与率	累積寄与率
PC1	$\frac{\lambda_1}{V_{total}}$	$\frac{\lambda_1}{V_{total}}$
PC2	$\frac{\lambda_2}{V_{total}}$	$\frac{\lambda_1+\lambda_2}{V_{total}}$
⋮	⋮	⋮
PCp	$\frac{\lambda_p}{V_{total}}$	$\frac{\lambda_1+\lambda_2+\cdots+\lambda_p}{V_{total}}$

例えば、現象の 70%まで説明したいときには、累積寄与率 0.7 までに含まれる主成分だけを取り上げて、残りは意味のない変動として取り扱います。このやり方は単純ですが少し機械的すぎて、実際上困ります。たとえば、0.7 の周辺に、小さな分散の主成分がたくさん存在した場合、どうしたらよいか判断できません。よくある別の方法は、スクリー・プロットというやり方です。スクリー・プロトとは図 74 に示したように、固有値の大きいものから順番に折れ線グラフを書くという方法です。

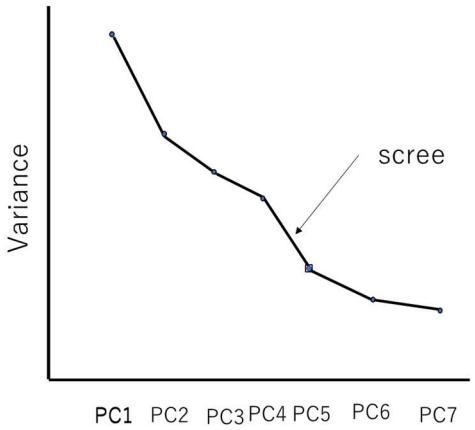

図 74. 主成分の分散のスクリー・プロット

第 4 主成分と第 5 主成分の間に大きな差があります。ここがスクリー（崖）になっています。この前後で考慮すべき主成分とそうでない主成分を分けて、第 5 主成分以下を切り捨

てます。
この選択法は曖昧さを含んでいます。どこにスクリーが出来るかは場合によるからです。
主成分分析を相関行列から始めた場合には、もう少し数学的な方法が考えられます。

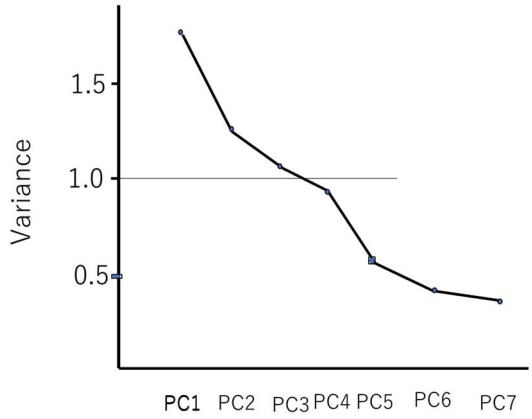

図 7 5. 相関行列の主成分分析のスクリー・プロット

相関行列の主成分分析では、すべての変数の分散が 1 です。相関行列の主成分分析は、すべての変数の分散を主成分に振り分けますが、その平均は 1 のままです。このことは、分散が 1 より大きい主成分は、分散を吸収する影響力の強い主成分です。反対に分散が 1 より小さい主成分は、盈虚力の小さな主成分です。ですから、分散が 1 より大きな主成分を選ぶというのも一つの選択です（図 75 参照）。
実際に公開されているソフトウエア―は結果の解釈のためにいくつかの機能が付け加えられています。最も一般的な主成分の解釈のための指標は主成分負荷量です。この指標は、主成分と実際に観測された変数の値の関係の強さを表しています

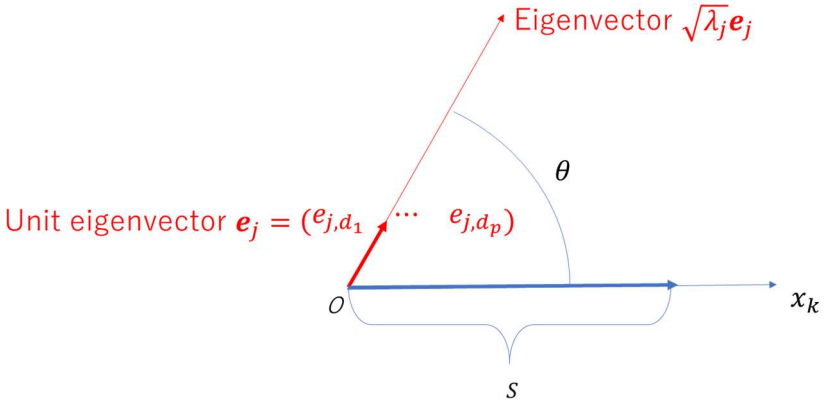

図 76. 固有ベクトルと観測変数の内積と相関係数$(k = 1, \cdots p)$

図 76 に固有ベクトルと元のデータ変数のベクトルを示しました。この図では、固有ベクトルの長さを固有値の平方根と S にしました。実際には、長さを$\sqrt{\lambda_j}$に固定する必要はありません。任意の実数で良いのです。ベクトルの関係性とは相関係数で相関係数は二つのベクトルがなす角度だからです。ここでことさら$\sqrt{\lambda_j}$ にしたのは、主成分負荷量と関係づける

ためです。
内積には次の二つの定義がありました。

$$\boldsymbol{V_1} \cdot \boldsymbol{V_2} = [\boldsymbol{V_1}][\boldsymbol{V_2}] \cos\theta = (v_{1,1} \quad \cdots \quad v_{1,n})(v_{2,1} \quad \cdots \quad v_{2,n})^T$$

$$\boldsymbol{V_1} = (v_{1,1} \quad \cdots \quad v_{1,n}), \qquad \boldsymbol{V_2} = (v_{2,1} \quad \cdots \quad v_{2,n})$$

$\boldsymbol{V_1}$と$\boldsymbol{V_2}$ の相関係数は$\cos\theta$です。
図 76 に示した固有ベクトルと変数のベクトルの場合には内積は以下の通りです。

$$\sqrt{\lambda_j} S \cos\theta_{j,k} = \left(\sqrt{\lambda_j} e_{j,d_1} \quad \cdots \quad \sqrt{\lambda_j} e_{j,d_p}\right) \begin{pmatrix} 0 \\ \vdots \\ S \\ \vdots \\ 0 \end{pmatrix} = \mathrm{S}\sqrt{\lambda_j} e_{j,d_k}$$

$$\cos\theta = e_{j,d_k}$$

相関係数は e_{j,d_j}だということになります。
主成分負荷量とは(PCL)、一つの変数の固有ベクトルに対する相関の大きさです。その大きさを、主成分の偏差の長さのベクトルに変数のベクトルを投影したときの、投影された部分の大きさだと考えると、次のような式になります。

$$\mathrm{PCL}_{j,k} = \sqrt{\lambda_j} r_{j,k} = \sqrt{\lambda_j} e_{j,d_k}$$

これを使って次のような表ができます。

主成分負荷量

	PC1	PC2	…,	PCp
偏差	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	…	$\sqrt{\lambda_p}$
変数				
変数 1	$\sqrt{\lambda_1} e_{1,d_1}$	$\sqrt{\lambda_2} e_{2,d_1}$	…	$\sqrt{\lambda_p} e_{p,d_1}$
変数 2	$\sqrt{\lambda_1} e_{1,d_2}$	$\sqrt{\lambda_2} e_{2,d_2}$	…	$\sqrt{\lambda_p} e_{p,d_2}$
⋮	⋮	⋮	⋱	⋮
変数 p	$\sqrt{\lambda_1} e_{1,d_p}$	$\sqrt{\lambda_2} e_{2,d_p}$	…	$\sqrt{\lambda_p} e_{p,d_p}$

これで、主成分と各変数の関係はわかりますが、変数間で$\sqrt{\lambda_i}$を比較することの意味については考える必要があります。変数に異なる最小単位で測られたデータが含まれていた場合、$\sqrt{\lambda_i}$ は小さな単位で測られたデータで大きくなります。もう少し理論的に表現すると、変数の分散は変数間で異なります。おそらく、分散が大きく異なった主成分の間で、分散の大きさを比較しても意味がないでしょう。主成分分析には、分散共分散行列から固有値を求めるやり方と、相関行列から固有値を求めるやり方があります。二つのやり方で、異なる結果が出ます。分散共分散行列の主成分分析は標準化されていないデータの主成分分析です。相関行列の主成分分析は標準化されたデータの主成分分析です。分析の目的が違うのです。分散共分散行列の主成分分析でも、分散の違いが無視できないのであれば、次の

表を作った方が良いかもしれません。

主成分と変数の相関行列

	PC1	PC2	…,	PCp
Variable				
Variable 1	e_{1,d_1}	e_{2,d_1}	…	e_{p,d_1}
Variable 2	e_{1,d_2}	e_{2,d_2}	…	e_{p,d_2}
⋮	⋮	⋮	⋱	⋮
Variable n	e_{1,d_p}	e_{2,d_p}	…	e_{p,d_p}

計算過程で示したように、e_{j,x_k}は相関係数です。

$$e_{j,d_k} = r_{jk}$$

この表を視覚化して表す一つの方法は、図 77 のような図を作ることです。図中の円は、PCa − PCb平面で切った超球の切断面です。半径は 1 です。

Vc は標準偏差の長さの変数 c ベクトルの、PCa − PCb平面への投影で、PCa − PCbの座標はPCa と PCbとの相関係数です。

$$\mathrm{V}c = (r_{ac} \quad r_{bc})$$

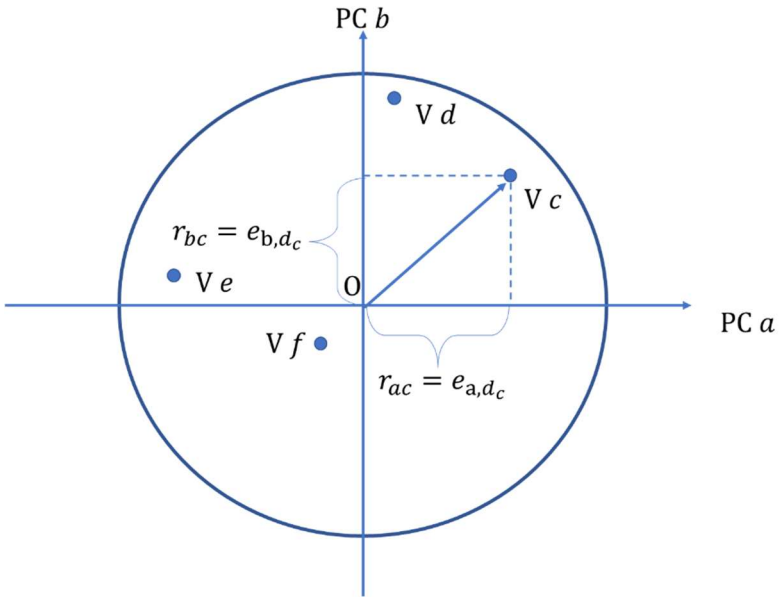

図 77. 二つの主成分の変数への貢献度の視覚的表現

$$\left|\overrightarrow{\mathrm{OV}c}\right|^2 = {r_{ac}}^2 + {r_{bc}}^2$$

$$\left|\overrightarrow{\mathrm{OV}c}\right| = \sqrt{{r_{ac}}^2 + {r_{bc}}^2}$$

$\left|\overrightarrow{\mathrm{OV}c}\right|^2$はPC$a$ と PCbの貢献度です。 Vcが、PCa と PCbで完全に説明できるのならば、

$$\left|\overrightarrow{\mathrm{OV}c}\right|^2 = {r_{ac}}^2 + {r_{bc}}^2 = 1$$

となり、ベクトル$\overrightarrow{\mathrm{OV}c}$の長さは、

$$\left|\overrightarrow{\mathrm{OV}c}\right| = \sqrt{{r_{ac}}^2+{r_{bc}}^2} = 1$$

この図では、$\left|\overrightarrow{\mathrm{OV}c}\right|$の長さは、0.8 ぐらいでしょう。ですから、この変数の分散への主成分a主成分bの寄与率は $0.8^2 = 0.64$ぐらいです。これは、$\mathrm{V}c$の分散の半分以上が主成分a と主成分bで説明できるということですから、無視できません。そこで、私たちは主成分aと主成分bは、$\mathrm{V}c$を説明する主要な主成分だと判定します。また、主成分a と主成分bの寄与率はほぼ同じぐらいです。$\mathrm{V}d$は円周に近いのですが、r_{ad}が小さいので、$\mathrm{V}d$は主として主成分bによって説明できると推定します。$\mathrm{V}e$も円周に近いのですが、r_{ae}の値が負です。また、r_{be}は小さな値です。ですから、この変数は主成分aに逆相関しています。$\mathrm{V}f$は原点に近づいています。このことは、$\mathrm{V}f$が主成分a と 主成分bにあまり関係がないことを示しています。$\mathrm{V}f$は$\mathrm{PC}a - \mathrm{PC}b$ある角度をもって交差しているのです。したがって、このベクトルについては、別の平面に投影して関係性を考える必要があります。 最も簡単な、結果の解釈の方法は、二つの主成分を座標軸とする分布図に、変数のベクトルを投影してみることです。普通は、比較のために、変数のベクトルの長さを標準偏差にします。

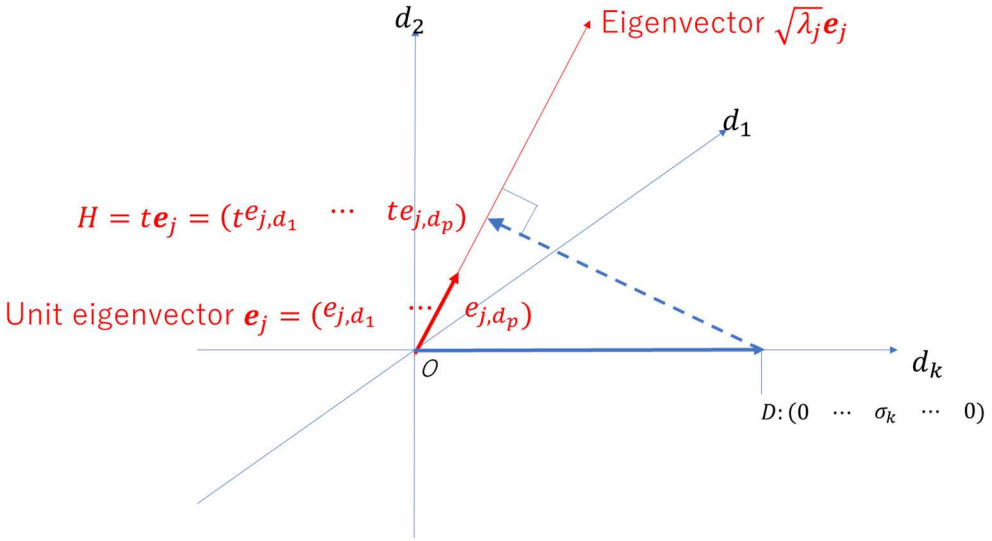

図 78. 変数ベクトルx_kの主成分ベクトルへの投影

変数ベクトルの投影の計算は主成分のベクトルへの座標変換と同じです。変数ベクトルの長さは標準偏差σ_kです。その座標は D=(0 $\cdots$ σ_k $\cdots$ 0).

$$\overrightarrow{\mathrm{OH}} \perp \overrightarrow{\mathrm{DH}}$$

$$\overrightarrow{\mathrm{DH}}=(te_{j,d_1} \quad \cdots \quad te_{j,d_p}) - (0 \quad \cdots \quad \sigma_k \quad \cdots \quad 0) = \left(te_{j,d_1} \quad \cdots \quad te_{j,d_k} - \sigma_k \quad \cdots \quad te_{j,d_p}\right)$$

$$\overrightarrow{\mathrm{OH}} \cdot \overrightarrow{\mathrm{DH}}=0$$

$$\boldsymbol{e}_j \cdot \overrightarrow{\mathrm{DH}}=0$$

$$(e_{j,d_1} \quad \cdots \quad e_{j,d_p}) \begin{pmatrix} te_{j,d_1} \\ \vdots \\ te_{j,d_k} - \sigma_k \\ \vdots \\ te_{j,d_p} \end{pmatrix} = 0$$

$$t\left({e_{j,d_1}}^2 + \cdots + {e_{j,d_p}}^2\right) - \sigma_k e_{j,d_k} = 0$$

$$t = \sigma_k e_{j,d_k}$$

$$\because {e_{j,x_1}}^2 + \cdots + {e_{j,x_p}}^2 = 1$$

これらを使えば次の表が作れます。

	主成分					
変数	PC1	PC2	⋯	PCk	⋯	PCp
変数 1	$\sigma_1 e_{1,d_1}$	$\sigma_1 e_{2,d_1}$	⋯	$\sigma_1 e_{k,d_1}$	⋯	$\sigma_1 e_{p,d_1}$
変数 2	$\sigma_2 e_{1,d_2}$	$\sigma_2 e_{2,d_2}$	⋯	$\sigma_2 e_{k,d_2}$	⋯	$\sigma_2 e_{p,d_2}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
変数 k	$\sigma_k e_{1,d_k}$	$\sigma_k e_{2,d_k}$	⋯	$\sigma_k e_{k,d_k}$	⋯	$\sigma_k e_{p,d_k}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
変数 p	$\sigma_p e_{1,d_p}$	$\sigma_p e_{2,d_p}$	⋯	$\sigma_p e_{k,d_p}$	⋯	$\sigma_p e_{p,d_p}$

この表から、適当な 2 つの主成分を選び出して、二つの主成分のベクトルが作る平面上に、標準偏差の長さの変数ベクトルを投影することが出来ます。たとえば、PC1-PC2 平面に投影すると、それぞれの変数の PC1-PC2 の座標は以下のようになります。

Variable 1: $(\sigma_1 e_{1,d_1} \quad \sigma_1 e_{2,d_1})$

Variable 2: $(\sigma_2 e_{1,d_2} \quad \sigma_2 e_{2,d_2})$

⋮ ⋮

Variable k: $(\sigma_k e_{1,d_k} \quad \sigma_k e_{2,d_k})$

⋮ ⋮

Variable p: $(\sigma_p e_{1,d_p} \quad \sigma_p e_{2,d_p})$

これを、PC1-PC2 平面上のデータの分布に重ね合わせたのが図 79 です。

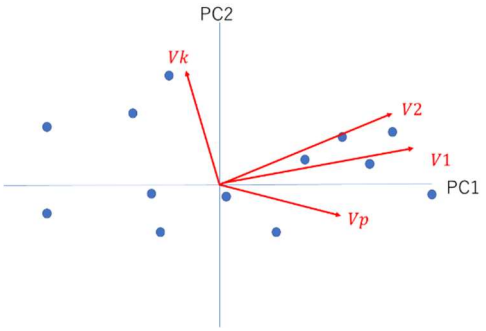

図 79. データの散布図と元の変数ベクトルの写像の重ね合わせの例

この操作は、ほぼ、対応分析（Correspondence analysis）と同じです。それはそれとして、変数 1 と変数 2 のベクトルの向きはほぼ主成分 1 と一致していて、十分な長さがあります。いつでもそうなるのではありませんが、主成分 1 は何かその現象や出来事の大きさにかかわる成分であることが多く、質的な特性、たとえば、差異、比率、不安定性、曖昧さのような質的な性質は第二主成分以降に来ることが多いようです。もし、主成分 1 が量的な特性を表しているのであれば、変数 1、変数 2 はともに大きさにかかわる変数のはずです。このような同一性を内的整合性あるいは内部的な一貫性といいます。内的一貫性は主成分分析ではあまり重要な意味を持たないのですが、因子分析、特に質問票を使った社会心理の調査の分析などでは重要な概念です。複数の質問に対する回答が高い相関性を持っているときには、それらの質問は、同じ内容のものを別の表現で訊いたことになるからです。これは、同じ内容を表す心理的傾向を違った角度から調べる必要がある心理学の分野では重要です。しかし、内的整合性は、他の分野では否定的にとらえられることがあります。同じ質問の繰り返しは無駄が多く、回答者に負担をかけて、質問票調査の質を低下させるからです。内的整合性は、クロンバックのα値（Cronbach's α value）で評価します。高いクロンバックのα値が得られた時、質問票をできるだけコンパクトにするために、次の質問票調査では、代表的な少数の質問に絞り込んだり、反対に、内的整合性を生かして、分析するときに変数を結合して、より感度の良い指標を作って分析します。

VI-2-1-5. 主成分分析の結果の解釈

主成分分析の数学的な意味は分かりやすいのですが、その結果の解釈はしばしば難しいことになります。主成分分析で何をしているかという説明は様々あるのですが、それは分析の目的によります。一番一般的な説明は、主成分分析は、データセットを構成している構成要素をいくつかの主要な構成要素にまとめているのだという説明です。専門知識があれば、主成分分析を使わなくても知識と経験で、すべて因子の中から代表的な因子を選び出すことが出来ます。現象の背景にある潜在的な因子を見つけることが主成分分析の目的だと説明することもできます。確かに、主成分分析によって、現象の背景にある仕組みを見つけ出せることがありますが、そういう運のよい例はまれです。主成分分析は、構成要素の軸（主成分）で結果を表現します。それぞれの軸は直交していて独立しています。その軸が示すものは時には何かの大きさだったり相違だったりしますが、その軸の意味は、しばしば、我々の日常言語では意味が解りません。結果の解釈は因子分析（Factor analysis：FA）の方が容易です。特に、斜交回転を使った因子分析は結果の解釈が自然に無理なくできます。斜交回転を使った因子分析のように、科学的な事前情報や日常生活の経験に合う分かりやすいベクトルに一致するように、独立性を無視して軸を回転したくなるのは自然なことです。似たような分析に見えますが、主成分分析と因子分析の目的は全く異なります。主成分分析は、現象の構造を理解する方法として重要です。主成分分析は、多次元空

間中のデータを変数間の相関を取り除いて別の空間に移し替えます。その本質はスペクトル分解です。

生物の分布は環境によって決まります。一方、環境の物理的要素は水温と塩分のように本来は独立しています。しかし、沿岸帯で塩分と水温を測定すると、実際には相関があります。たとえば、夏場に河口から沖に向けて測点を作って、塩分と水温を測定すると、水温は河口から離れるにしたがって減少し、塩分は増加します。その結果、塩分と水温には明瞭な逆相関の関係が生まれます。同様に、リン酸や硝酸濃度のような化学環境の測定項目についても相関が生まれます。生物は、そのような環境因子の組み合わせに適応しているのです。生物の分布を一つの水質項目という要素に還元して説明してもあまり意味はありません。もちろん、要素還元主義は生物の生理的な反応の理解には有効です。しかし、単純な要素還元主義では、行動や生物の進化メカニズムなど複雑なシステムの発見にはあまり油工ではないでしょう。複雑系を理解するためには、現象と関係がないかもしれない変数も含めて、網羅的なデータの蓄積が必要です。たとえば、何の事前情報もなく、沿岸における魚種の分布がどのように決まっているのかを知ろうとすれば、様々な場所で魚を捕獲し、魚種名、個体数、体重、体長など生物データと水深、塩分、水温、透明度。底質、サンプリング時間など、様々な物理・化学的データを集めるしかないでしょう。その中のいくつかの変数は相関し、別のものは逆相関し、残りは関係がないでしょう。確かに、このやり方は、よくデザインされているとは言えませんが、事前情報がなければ、こういう見通しのないやり方をとるしかありません。これは、仮説検証ではありません。科学とは仮説検証だと考える人もいます。これは誤りです。科学は、まず、無原則に積み重なったデータセットの中から、仮説を作らなくてはなりません。その目的のために、分析者は変数ごとに、データの分布を確かめるために頻度分布を作ります。ある分析者は、X-Y プロットを作って変数間の関係を見つけようとします。別の分析者は相関分析をします。最も一般的には、分散共分散行列、あるいは、相関行列を作って、潜在的な相関関係を見つけます。まだ、コンピュータが今のように普及していなかった頃のことです。筆者は、日本沿岸の養殖ノリの生産量を決める因子について分析しました。50 年以上にわたり、日本各地の海苔養殖場の単位面積当たりの生産量と価格のデータをあつめて、生産量と価格について、地域間の総あたりの行列をつくりました。そして、実際の地図上で、相関する地域を色で塗り分けたのです。これには２か月以上かかりました。その結果、ノリの生産にかかわる重要の要素を見つけることが出来ました。著者は、遠く離れた地域間で、同調的に変動する地域を見つけたのです。そして、それがなぜ同調するのか考えた結果、背景にある要因を見つけました。そのやり方は洗礼されていない稚拙なやり方です。もはやそんなやり方をする人はいないでしょう。その時点では、著者は主成分分析を知りませんでした。仮に知っていたとしても、コンピュータを持っていなかったので計算できなかったでしょう。しかし、やっていることは主成分分と同じです。今ならば、若い時の自分自身に、環境データを含めて、すべてのデータを入れて主成分分析をすることを薦めるでしょう。主

成分分析の最も重要な機能は、潜在的な背景の仕組みを含めて、データ間の関係を視覚化して見せることです。無知で若かった筆者は、背景の仕組みが見つかることを漠然と期待して、総当たりの相関行列を作りました。相関はいつでも因果関係を示しているとは限りませんし、遠く離れたノリ養殖漁場間に因果関係が存在する可能性はほとんどないでしょう。若かった筆者は、日本の沿岸環境とノリの生理学を知っていたので、遠く離れた養殖漁場に変動を作り出す潜在的な因子に気が付きました。漠然とした見通しで始めてたまたまうまくいったのです。つまり、その分野の情報が必要で、主成分分析だけでは結論に到達しないのです。しかし、このことは、主成分分析をする動機付けとしては十分でしょう。コンピュータを使えば、主成分分析は簡単だからです。描かれた絵の意味が一でもわかるわけではないにしても、主成分分析で全体像が描けます。まず、主成分分析をやってみるべきです。