【第 3 章補足資料】

＜主成分分析の計算方法の解説：第 3 章表 3.2 についての補足＞

数学と国語のテストの例題を用いて，主成分分析の計算方法の詳細を示す．

1.　1 次元指標の形成―分散共分散行列の固有値問題

表○○の数学と国語のテストの例題では，これから要約の記述統計として，x，y の標準偏差，分散，相関係数，共分散は

$$s_x=20,\ s_y=15,\ s_x^2=400,\ s_y^2=225,\ r_{xy}=0.5,\ s_{xy}=20\cdot 15\cdot 0.5=150$$

となっている．いま，数学，国語の成績 x，y をもとに総合点数(指標とよぶ)

$$f=l_1x+l_2y$$

を作り，学力が f でうまく総合判断しやすいよう f が大きく広がることを目的とする．(l_1，l_2 の符号組み合わせで f の意味が異なってくるから，2 通りの f が予想される)そのためには，分散の公式から

$$s_f^2=l_1^2s_x^2+2l_1l_2s_{xy}+l_2^2s_y^2$$

であるが，l_1，l_2 の値を制約しておかないと最大化には意味がないから

$$\text{条件}\qquad l_1^2+l_2^2=1\qquad(\text{ベクトルの長さ}=1)$$

のもとで，

$$s_f^2=l_1^2s_x^2+2l_1l_2s_{xy}+l_2^2s_y^2$$

の最大値を求める．条件付き最適化の手続きに従い λ をラグランジュの未定乗数として，関数

$$\mathcal{L}(l_1,\ l_2,\ \lambda)=l_1^2s_x^2+2l_1l_2s_{xy}+l_2^2s_y^2-\lambda(l_1^2+l_2^2-1)$$

の(無条件)最大化の条件を求めよう．すなわち l_1，l_2 の連立方程式が得られる

$$\partial\mathcal{L}/\partial l_1=2l_1s_x^2+2l_2s_{xy}-2\lambda l_1=0,$$
$$\partial\mathcal{L}/\partial l_2=2l_1s_y^2+2l_1s_{xy}-2\lambda l_2=0$$

行列形式の方が見やすく，次の分散共分散行列の「固有値問題」(Eigenvalue problem)となる：

$$\begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}\begin{pmatrix} l_1 \\ l_2 \end{pmatrix}=\lambda\begin{pmatrix} l_1 \\ l_2 \end{pmatrix}$$

【第 3 章補足資料】

ここでλ を「固有値」，ベクトル $(l_1,\ l_2)$ を「固有ベクトル」という．この解き方の定跡は，まず

$$\begin{pmatrix} {s_x}^2-\lambda & s_{xy} \\ s_{xy} & {s_y}^2-\lambda \end{pmatrix}\begin{pmatrix} l_1 \\ l_2 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

となるが，右辺が 0 であるため l_1，l_2 がともに 0 はもちろん解である．これは無意味であり，有意味な(「自明でない」non-trivial)解が(一つでも，ただし実際は無限に)あるための必要十分条件は，係数の行列式

$$\begin{vmatrix} {s_x}^2-\lambda & s_{xy} \\ s_{xy} & {s_y}^2-\lambda \end{vmatrix}=0$$

である(よく知られているので証明略)．これを展開して結局

$$\lambda^2-({s_x}^2+{s_y}^2)\lambda-({s_x}^2{s_y}^2-{s_{xy}}^2)=0 \qquad \text{(固有方程式)}$$

を得る．ここでできたこれの 2 根が固有値 λ_1，λ_2，それに対応する l_1，l_2 がそれぞれの固有ベクトルである．実際，λ_1，λ_2 をそれぞれ代入して l_1，l_2 を解く際，当然不定(解は一つに定まらない)となるので，ここでは ${l_1}^2+{l_2}^2=1$ のように長さ 1 に規格化(normalize)する．

例　ここの具体例では，固有値問題

$$※ \qquad \begin{pmatrix} 400 & 150 \\ 150 & 225 \end{pmatrix}\begin{pmatrix} l_1 \\ l_2 \end{pmatrix}=\lambda\begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \qquad \text{(連立方程式)}$$

を解く固有方程式は

$$\lambda^2-625\lambda+67500=0$$

となる．ここで 625，67500 が重要だが後で触れる．解の公式より

$$\lambda=\frac{625\pm\sqrt{625^2-4\cdot 67500}}{2}=\frac{625\pm\sqrt{120625}}{2}$$

より，2 つの固有値は大きさの順に，

$$\lambda_1=486.15,\quad \lambda_2=138.85$$

次に 固有ベクトル$(l_1,\ l_2)$ の一例は，$\lambda_1=486.15$，$\lambda_2=138.85$ それぞれに対し，

【第 3 章補足資料】

元の連立方程式※に戻って (1，0.5743)，(0.5743，−1) などが得られる．このベクトルは長さ 1 にではない．長さ 1 に規格化した固有ベクトルは

$$\left(\frac{1}{\sqrt{1^2+0.5743^2}},\ \frac{0.5743}{\sqrt{1^2+0.5743^2}}\right)=(0.8672,\ 0.4980)$$

同様に，

$$\left(\frac{0.5743}{\sqrt{0.5743^2+(-1)^2}},\ \frac{-1}{\sqrt{0.5743^2+(-1)^2}}\right)=(0.498,\ -0.8672)$$

のように求められる．予想通り f は 2 通り得られて

$$f_1=0.8672x+0.4980y,\qquad f_2=0.4980x-0.8672y$$

ここで，今後結果をみるために次の数学的事項を指摘しておこう．固有値は 2 次方程式の 2 解だから，高校来の知識を使い，

ii） 固有値は実数，しかも正（>0）
iii） 固有値の和 ＝ 対角項の和すなわちトレース：$s_x^2+s_y^2=625$
iv） 固有値の積 ＝ 行列式の値：$s_x^2s_y^2-s_{xy}^2=67500$
v） 異なった固有値に対する固有ベクトルは直交

2. 主成分を抜き出す

以上，x，y データから得られた 2 通りの指標

$$f_1=0.8672x+0.4980y,\qquad f_2=0.4980x-0.8672y$$

を「第 1 主成分」1st principal component, PC1，「第 2 主成分」2nd principal component, PC2 という．各係数の符号に注意しておこう．主成分の性質は以下に述べる通り．うまくできている．

1） PC1 の分散は

$$\begin{aligned}s_{f_1}^2&=0.8672^2\cdot400+2\cdot0.8672\cdot0.4980\cdot150+0.4980^2\cdot225\\&=300.8+129.6+55.8\\&=496.2\\&=\lambda_1\end{aligned}$$

【第 3 章補足資料】

PC2 の分散は同様に $s_{f_2}{}^2=138.9=\lambda_2$ で，第 1 主成分のそれより小さい．

すなわち，各主成分 PC1，2 の分散は対応する固有値に等しい．実際 i)から必ず＞0 となる

$$s_{f_1}{}^2+s_{f_2}{}^2=\lambda_1+\lambda_2=625$$

したがって，PC1，PC2 の重要度(寄与率 contribution)の割合は，それぞれ

$$486.15/625=0.778(77.8\%),\quad 138.85/625=0.222(22.2\%)$$

たしかに PC1 は，文字通り重要であり，PC2 はこれを補っている．

2)　第 1 主成分 f_1 と第 2 主成分 f_2 の共分散の計算は($\bar{x}=\bar{y}=0$に注意)

$$\Sigma(0.8672x+0.4980y)(0.4980x-0.8672y)/10\quad(\Sigma：サンプル上での和)$$

$$=0.8672\cdot0.4980\cdot400+(0.4980^2-0.8672^2)\cdot150-0.4970\cdot0.8672\cdot225$$

$$=172.8-75.6-92.2$$

$$=0$$

で相関関係はない(無相関)．これは固有ベクトルの直交iv)からも導かれる．

(x，y)サンプル中の分散は互いに無相関のこれらの 2 つの主成分 f_1，f_2 の要因に完全に分解される．

以上の結果，次のように要約される(実際の計算はコンピュータが行う)．

主成分結果(要約)

	第 1 主成分 (PC1)	第 2 主成分 (PC2)	計
数学 x	0.8672	0.4980	
国語 y	0.4980	−0.8672	
分散(固有値)	486.15	138.85	625.00
寄与率(%)	77.8	22.2	100.0

③　主成分の解釈とネーミング

x，y には実体的な意味があるが，

$$f_1=0.8672x+0.4980y,\qquad f_2=0.4980x-0.8672y,$$

【第 3 章補足資料】

は数理上の操作の結果で固有の意味は与えられていないため，各主成分は何をあらわすかを解釈し，命名するのがふつうである．これはコンピュータにはできず人間(のみ)が行う．機械学習への応用でも変わらない．

表○○にある係数(主成分への「負荷量」という)をみると，第 1 主成分には数学 x，国語 y の学力がともに + で入っており，第 1 主成分は「総合的能力」(ある意味での「実力」)の要素を表すことができる．第 2 主成分には，数学 x のみが + で，国語 y のような文系的傾向は − となっているから，文系-理系を対照的に表す軸の上で，「理系的能力」の要素をあらわすと解釈できる．

もとより，このような「解釈」や「ネーミング」は一通りではなく，なにがしかの主観やその分野のさまざまな知識にもとづくが，これは複雑多様な現象の反映である．これを容認してこそ主成分分析による現象のモデル化が可能であり，力を発揮する．