主成分分析：基礎理論

1　はじめに

多変量解析法[1)2)]は，1960年代にパーソナルコンピュータの普及とともに，心理学や経済学等の人間行動を分析する手法として発展し，1970年代に化学の分野に応用[3)~5)]され始め現在に至っている。主成分分析の分析化学への応用は，複数の試料の分布図から類似性を判断するとき，分布図が二次元か三次元でしか図示できないので，4個以上の測定値や定量値｛(混合物の含有率)(カテゴリー)｝を持つ複数（p個）の試料の分布状態を，なんらかの方法で3個以下の数値に集約する必要から始められた。従って，主成分分析はp個の試料中のn個のカテゴリーの内容を変形することなく，累積寄与率が90％以上となる2～3個の数値（第一～第三主成分得点）に集約する方法として利用されている。

2　計算方法

2･1　各主成分の計算

表1に示した，n個のカテゴリーを持つp個の試料は，(1)に示した式で各主成分｛Z_m｝が計算される。

$$\begin{aligned} Z_1 &= l_{11}x_1 + l_{12}x_2 + \cdots + l_{1n}X_n \\ Z_2 &= l_{21}x_1 + l_{22}x_2 + \cdots + l_{2n}X_n \\ &\cdots\cdots\cdots\cdots \\ Z_m &= l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mn}X_n \end{aligned} \quad \cdots\cdots(1)$$

ここで，式(1)の係数｛l_{km}｝は，

$$l_{k1}^2+l_{k2}^2+\cdots+l_{kn}^2=\sum_{l=1}^{n}(l_{ki})^2=1(k=1, 2, \cdots, m) \quad \cdots\cdots(2)$$

となる。

このとき，第一主成分｛Z_1｝の係数｛l_{1i}｝(i=1, 2, …, n)は，式(2)の条件下で｛Z_1｝の分散が最大になるように定められる。ついで，第二主成分｛Z_2｝の係数｛l_{2i}｝(i=1, 2, …, n)は，式(2)の条件を満たし，第一主成分｛Z_1｝と第二主成分｛Z_2｝間が完全に無相関になる条件下で｛Z_2｝の分散が最大になるように定められる。以下同様の方法を繰り返し，第m主成分｛Z_m｝までの係数｛l_{mi}｝(i=1, 2, …, n)が各主成分｛$Z_1, Z_2, \cdots, Z_{m-1}$｝と無相関の条件下で，各主成分｛$Z_1, Z_2, \cdots, Z_{m-1}$｝の分散が最大になるように定められる。

各試料中のn個のカテゴリー｛$x_1, x_2, \cdots, x_n$｝を直交変換して得られた第一～第m主成分｛$Z_1, Z_2, \cdots, Z_m$｝の，分散の大きなものから順に第一主成分｛Z_1｝，第二主成分｛Z_2｝, …，第m主成分｛Z_m｝とする。

2･2　主成分寄与率の計算

第一～第m主成分｛$Z_1, Z_2, \cdots, Z_m$｝を，p個の試料に与える新しいm個の指標｛X_i｝(i=1, 2, …, n)とすると，もとのn個のカテゴリー｛$x_1, x_2, \cdots, x_n$｝に対する重相関係数｛R_i｝(i=1, 2, …, n)の二乗の和が最大となるように，第一～第m成分｛$Z_1, Z_2, \cdots, Z_m$｝を決める。このことは，第一～第m主成分｛$Z_1, Z_2, \cdots, Z_m$｝を一次関数として示すと，式(3)のように｛$X_1, X_2, \cdots, X_n$｝を求めることになる。

$$\begin{aligned} X_1 &= b_{11}z_1 + b_{12}z_2 + \cdots + b_{1m}z_m \\ X_2 &= b_{21}z_1 + b_{22}z_2 + \cdots + b_{2m}z_m \\ &\cdots\cdots\cdots\cdots \\ X_n &= b_{n1}z_1 + b_{n2}z_2 + \cdots + b_{nm}z_m \end{aligned} \quad \cdots\cdots(3)$$

式(3)で計算した｛$X_1, X_2, \cdots, X_n$｝と，n個のカテゴリー｛$x_1, x_2, \cdots, x_n$｝を式(4)に代入し，

$$\sum_{l=1}^{n}\sum_{b=1}^{p}(x_{bi}-X_{bi})^2/s_i^2 \quad (S_i：標準偏差) \quad \cdots\cdots(4)$$

計算値が最小値となるように，第一～第m主成分｛$Z_1, Z_2, \cdots, Z_m$｝を決めることになる。この方法で計算した第一～第m主成分｛$Z_1, Z_2, \cdots, Z_m$｝を表2に示す。

第j主成分｛Z_j｝の分散｛$V(Z_j)$｝は，

$$V(Z_j) = \frac{\Sigma(Z_{kj} - \overline{Z_j})^2}{p-1} \quad (k = 1, 2, \cdots, m) \quad \cdots\cdots(5)$$

で計算される。

第j主成分｛Z_j｝の寄与率｛f_1｝は，式(6)から第j主成分｛Z_j｝の分散｛$V(Z_j)$｝と，各主成分の分散｛$V(Z_1), V(Z_2), \cdots, V(Z_m)$｝の合計値を用いて計算する。

$$f_j = \frac{V(Z_j)}{V(Z_1) + V(Z_2) + \cdots + V(Z_j) + \cdots V(Z_m)} \quad \cdots\cdots(6)$$

第j主成分までの累積寄与率｛T_j｝は，式(7)から第j主成分｛Z_j｝までの分散｛$V(Z_1), V(Z_2), \cdots, V(Z_j)$｝の合計値と，各主成分の分散｛$V(Z_1), V(Z_2), \cdots, V(Z_m)$｝の合計値を用いて計算する。

$$T_j = \frac{V(Z_1) + V(Z_2) + \cdots + V(Z_j)}{V(Z_1) + V(Z_2) + \cdots + V(Z_j) + \cdots + V(Z_m)} \quad \cdots\cdots(7)$$

主成分分析では，累積寄与率｛T_j｝が90％以上となるまでの主成分｛$Z_1, Z_2, \cdots, Z_j$｝を使用するが，この方法でn個のカテゴリーは2個｛Z_1, Z_2｝か3個｛Z_1, Z_2, Z_3｝の主成分に集約される。

2･3　主成分得点の計算

主成分分析の結果，n個のカテゴリーはm個の主成分に変換されるが，図示は各主成分から計算されたp個の試料の主成分得点を使用する。第j主成分｛Z_j｝の主成分得点の二乗和｛S_j｝は式(8)から，

$$S_j = \sum_{l=1}^{m} Z_{ji}^2 = \sum_{l=1}^{m} (l_{j1}x_{i1} + l_{j2}x_{i2} + \cdots + l_{jn}x_{in})^2 \quad \cdots\cdots(8)$$

となり，式(2)の条件下で主成分得点の二乗和｛S_j｝が最大となるように式(1)の係数を決定する。このことは，p個の試料から第j主成分｛Z_j｝の軸に垂線を下したときの｛Z_j｝軸との交点の数値となる。参考として，図1に8試料から第一主成分｛Z_1｝に下した状態を示す。

図1に示したカテゴリーが2個の場合の第j主成分｛Z_j｝の主成分得点の二乗和｛S_j｝は式(9)から計算する。

表1　p個の試料が持つn個の測定値や定量値

試料	測定値や定量値（カテゴリー）			
	x_1	x_2	……	x_n
1	x_{11}	x_{12}	……	x_{1n}
2	x_{21}	x_{22}	……	x_{2n}
j	x_{j1}	x_{j2}	……	x_{jn}
p	x_{p1}	x_{p2}	……	x_{pn}

表2　p個の試料が持つn個の測定値や定量値から計算されるm個の主成分

試料	主　成　分			
	z_1	z_2	……	z_m
1	z_{11}	z_{12}	……	z_{1m}
2	z_{21}	z_{22}	……	z_{2m}
j	z_{j1}	z_{j2}	……	z_{jm}
p	z_{p1}	z_{p2}	……	z_{pm}

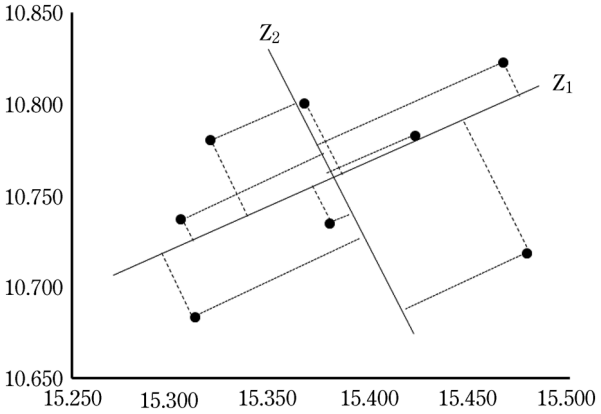

図 1　主成分に対応する座標の変換

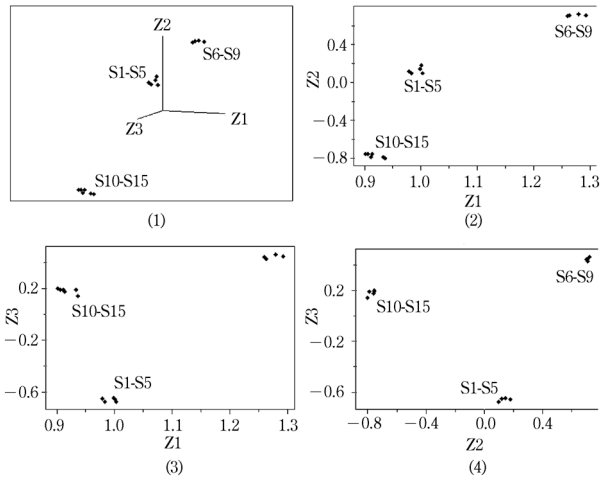

図 2　主成分得点の図示方法

表 3　計算例のための入力数値と計算で得られた主成分得点

試料	測定値								主成分得点		
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Z_1	Z_2	Z_3
S1	0.1505	0.0018	0.5227	0.4150	0.0505	0.8677	0.3142	0.3377	0.9836	0.0995	−0.6767
S2	0.1544	0.0015	0.5860	0.4375	0.0509	0.8164	0.3856	0.3022	1.0011	0.1798	−0.6565
S3	0.1598	0.0012	0.5033	0.4551	0.0523	0.8476	0.3571	0.3262	0.9990	0.1441	−0.6478
S4	0.1589	0.0014	0.5708	0.4041	0.0531	0.8085	0.3321	0.3406	0.9794	0.1191	−0.6495
S5	0.1562	0.0019	0.5546	0.4248	0.0513	0.8558	0.3126	0.3576	1.0025	0.1004	−0.6762
S6	0.3574	0.4112	0.3137	0.9972	0.6223	0.0629	0.7519	0.1838	1.2798	0.7234	0.4654
S7	0.3875	0.4053	0.3345	0.9584	0.6083	0.0611	0.7263	0.1886	1.2603	0.7016	0.4469
S8	0.3253	0.3911	0.3739	0.9814	0.6221	0.0638	0.7053	0.1859	1.2635	0.7090	0.4280
S9	0.3667	0.4234	0.3930	0.9394	0.6513	0.0631	0.7540	0.1841	1.2929	0.7109	0.4483
S10	0.0079	0.6428	0.0595	0.0228	0.2408	0.3340	0.0697	0.9135	0.9016	−0.7567	0.2012
S11	0.0076	0.6336	0.0606	0.0253	0.2038	0.3926	0.0660	0.9673	0.9362	−0.8021	0.1451
S12	0.0070	0.6541	0.0606	0.0241	0.2444	0.3615	0.0657	0.9498	0.9337	−0.7880	0.1903
S13	0.0073	0.6130	0.0607	0.0249	0.2253	0.3270	0.0625	0.9731	0.9111	−0.7874	0.1910
S14	0.0072	0.6090	0.0608	0.0265	0.2582	0.3597	0.0668	0.9299	0.9138	−0.7576	0.1754
S15	0.0073	0.6497	0.0620	0.0243	0.2425	0.3473	0.0651	0.9064	0.9060	−0.7567	0.1937

$$S_j = \sum_{l=1}^{m} Z_{ji}^{\,2} = \sum_{l=1}^{m} (l_{j1}x_{i1} + l_{j2}x_{i2})^2$$
$$= (l_{j1})^2 \sum_{l=1}^{m} x_{i1}^{\,2} + 2\, l_{j1}l_{j2} \sum_{l=1}^{m} x_{i1}x_{i2} + (l_{j2})^2 \sum_{l=1}^{m} x_{i2}^{\,2} \quad \cdots\cdots (9)$$

3　計算例

表 3 に示した測定値を用いた主成分分析で，表 3 右側に示した主成分得点を計算する。式(1)，(2)による計算方法は複雑なので，計算経過を確認したい場合は文献[8]を参照されたい。主成分分析のソフトウエアを用いた場合は，表 3 の測定値を代入すれば各主成分得点が算出され，3 通りの二次元図か 1 通りの三次元の図として図示されてくる。主成分分析を分析化学に応用するときには，数学的な計算過程よりも，計算から得られた内容が何を意味しているのかを把握することのほうがはるかに重要である。従って，まず結果の明らかな試料（例えば表 3 の測定値）を用いて主成分分析を行い，得られた計算結果を分析化学の観点から解析して，計算内容を理解した後に使用する必要がある。

表 3 の測定値を用いて主成分分析を行った結果では，累積主成分寄与率が 90 % 以上となるまでに第三主成分 $\{Z_1, Z_2, Z_3\}$ までを必要（第一主成分寄与率 64.8518，第二主成分寄与率 22.4962，第三主成分寄与率 12.5612）としたので，主成分分析結果は図 2(1)の三次元か，使用するソフトウエアによっては，図 2(2)，(3)，(4)のように 3 図からなる二次元の図で表示してくる。ほとんどの場合，図 2(1)～(4)から明らかなように，累積主成分寄与率が 90 % 未満でも，分析目的試料（未知試料）の帰属グループや試料の分離状態の判断には，図 2(2)のように第一主成分 $\{Z_1\}$ と第二主成分 $\{Z_2\}$ の結果だけで，定性分析結果を誤りなく判断することができる。最終主成分得点を定量分析に用いる方法もあるが，詳細は文献[6]を参照されたい。

4　スケーリング処理

主成分分析では，入力した数値(測定値や定量値)を，カテゴリーごとにスケーリング処理を行う場合があり，スケーリング処理はオートスケーリングを使用する場合が多いので，オートスケーリングの計算方法を示す。それ以外のスケーリング処理については，文献[6,7]を参照されたい。オートスケーリング（AS_j）[6,7]は，複数の試料が持っている測定値や定量値（カテゴリー）の分散を標準偏差で除す方法である。

$$AS_j = \frac{X_{ij} - \overline{X_{ij}}}{S_{ij}} \qquad S_{ij} = \frac{\{\Sigma(X_{ij} - \overline{X_{ij}})^2\}^{1/2}}{n-1}$$

i：測定値（定量値）；j：試料
n：1 個の試料中のカテゴリー数
S_{ij}：標準偏差

スケーリング処理は，主成分分析に用いた m 個の試料間のカテゴリーの桁数を揃えるので，用いる未知試料が異なるとスケーリング処理した各カテゴリーの数値も異なってくる。従って，分析化学に使用する場合は，安易に測定値や定量値をスケーリング処理すべきではない。

文　献

1) 奥村忠一，久米　均，芳賀敏郎，吉澤　正：“多変量解析法”，pp. 159−167，(1971)，(日科技連)．
2) 田中　豊，脇本和昌：“多変量統計解析法”，pp. 53−66，(1983)，(現代数学社)．
3) M. A. Sharaft, D. L.Ilian, and B. R. Kowalski : “*Chemometrics*”, (1986), (J. Wiley & Sons, New York)．
4) 桐嶋鐵郎：“ケモメトリックス”，pp. 72−77，(1992)．
5) 宮下芳勝，佐々木慎一：“ケモメトリックス―科学パターン認識と多変量解析―”，pp. 19−37，(1995)，(共立出版)．
6) 三井利幸：“改定分析化学のための多変量解析法”，pp. 123−161，(2014)，(一粒書房)．
7) 三井利幸：“ケモメトリックスの基礎と応用”，pp. 114−120，(2003)，(アイピーシー)．

〔数値解析研究所　三井利幸〕