Technical Sheet
No. 21-25

データ解析入門 3 <主成分分析によるデータマイニング>

キーワード：主成分分析、ローディングプロット、因子負荷量、寄与率、多変量解析

はじめに

化学分析やアンケート分析などでは、しばしば多変量データを扱います。主成分分析（PCA）は多変量データの可視化などに用いられる解析手法です。本稿では、前稿（No. 21-24）[1]に引き続き、ワインの化学分析に関するデータセット[2]を用います。ワインの分析結果に PCA を実行し、ワインを特徴づける変数を推察します。ワインのデータセットやデータの可視化については、前稿をご参照ください。

PCA によるデータの可視化

本稿で用いるワインのデータセットは、サンプル数が 178 個、全 13 変数の多変量データです。前回示したスコアプロットを図 1 に再掲します [1]。各サンプルは 3 つのクラスに分類されており、図 1 ではクラスごとにデータを色分しました。なお、データの前処理として標準化 [1]を行っています。

図中の矢印で強調したように、各クラスはそれぞれ異なった主成分スコアをとる傾向にあります。例えば、緑クラスでは、第 1 主成分スコアはゼロを中心にばらつき、一方で第 2 主成分スコアは正になる傾向にあります。この傾向は他の 2 クラスとは異なります。つまり、主成分スコアにより各クラスが特徴づけられています。それでは、第 2 主成分軸とはどのような意味を持つ座標軸なのでしょうか。

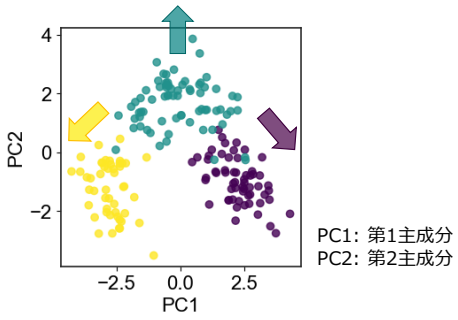

図 1　スコアプロット

主成分は元の変数の線形結合

PCA により得られる主成分は元の変数の線形結合で表されます。つまり、i 番目のサンプルの第 2 主成分スコア $t_{i,2}$ は、

$$t_{i,2} = p_{1,2}x_{i,1} + p_{2,2}x_{i,2} + \cdots + p_{13,2}x_{i,13} \quad (1)$$

という式により計算されます。ここで、$x_{i,j}$ (j = 1, 2,⋯, 13) は i 番目のサンプルの変数 x_j の数値です。そのため、$p_{j,2}$ は第 2 主成分スコアに対する変数 x_j の寄与の度合いと考えることができます。ここで、式(1)の 13 個の重み $p_{j,2}$ をまとめたベクトル $\boldsymbol{p}_2$ を、

$$\boldsymbol{p}_2 = (p_{1,2}, p_{2,2}, \cdots, p_{13,2})^{\mathrm{T}} \quad (2)$$

と表します。このベクトル $\boldsymbol{p}_2$ は第 2 主成分軸の方向を示します。なお、紙面を節約するため、行ベクトルを転置しています。次に、178 サンプルの第 2 主成分スコアをまとめて、

$$\boldsymbol{t}_2 = (t_{1,2}, t_{2,2}, \cdots, t_{178,2})^{\mathrm{T}} \quad (3)$$

とします。$t_{i,2}$ は i 番目のサンプルの第 2 主成分スコアです。また、データ行列 [1]を X とすると、行列 X は 178 行×13 列になります。行列 X の 13 列をそれぞれ列ベクトル $\boldsymbol{x}_1$〜$\boldsymbol{x}_{13}$ とすると、

$$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{13}) \quad (4)$$

と表現できます。

以上のベクトルおよび行列を用いると、第 2 主成分スコア $\boldsymbol{t}_2$ は、

$$\boldsymbol{t}_2 = X\boldsymbol{p}_2 = p_{1,2}\boldsymbol{x}_1 + p_{2,2}\boldsymbol{x}_2 + \cdots + p_{13,2}\boldsymbol{x}_{13} \quad (5)$$

という関係にあります。同様の関係は第 1 主成分においても成り立ち、第 1 主成分スコア $\boldsymbol{t}_1$ は、

$$\boldsymbol{t}_1 = X\boldsymbol{p}_1 \quad (6)$$

となります。なお、$\boldsymbol{p}_1 \neq \boldsymbol{p}_2$ であり、各主成分軸は直交しているため両ベクトルの内積は $\boldsymbol{p}_1 \cdot \boldsymbol{p}_2 = 0$ です。

ローディングプロット

ここからは、主成分軸に寄与する元の変数を推察します。今回、標準化を行っているため、各変数 x_j のスケールは統一されています。そのため、$\boldsymbol{p}_1$ と $\boldsymbol{p}_2$ の各要素を横軸および縦軸にプロットすることで、両主成分軸への寄与を確認できます。

しかしながら、標準化が行われていない場合、元の変数のスケールの差異が $\boldsymbol{p}_1$ と $\boldsymbol{p}_2$ の各要素の大きさに影響を与えてしまいます。そのような影響を取り除くため、主成分スコア $\boldsymbol{t}_1$ と $\boldsymbol{x}_j$ (j = 1, 2,⋯, 13)　の

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

相関係数を第 1 主成分軸の解釈に用いることができます。同様に、第 2 主成分軸の解釈には $\boldsymbol{t}_2$ と $\boldsymbol{x}_j$ の相関係数を用います。主成分スコアと元の変数との相関係数は因子負荷量、因子負荷量の散布図はローディングプロットと呼ばれます。

図 2 に第 1 および第 2 主成分のローディングプロットを示します。第 2 主成分に関する因子負荷量に着目すると、変数 x_1 や変数 x_{10} は負の相関係数をもつことがわかります。したがって、スコアプロットで第 2 主成分スコアが大きくなる傾向にあった緑クラスのサンプルでは、x_1 や x_{10} などの値は小さくなる傾向にあると思われます。簡潔な記述のため、これまで明示してきませんでしたが、x_1 は「alcohol」、x_{10} は「color_intensity」に対応します [2]。したがって、アルコール濃度や色強度などがワインを特徴づける因子であると推察できます。同様にデータマイニングを行うことで、第 1 主成分に寄与する元の変数も推察することができます。

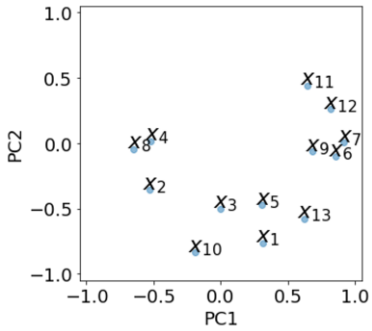

図 2　ローディングプロット

寄与率とは

PCA ではデータのばらつきが最大になる方向に主成分軸を作成しました。分散という統計量はデータのばらつきの指標であり、PCA では主成分スコアの分散の合計を全情報量とし、全情報量に対する各主成分スコアの分散の割合を寄与率とします。以下に寄与率の計算式を示します（式(7)）。

$$\text{寄与率} = \frac{\text{第}j\text{主成分スコアの分散}}{\text{主成分スコアの分散の合計}} \qquad (7)$$

図 3 に寄与率とその累積（累積寄与率）を示します。第 1～第 13 主成分に向かうにつれ、寄与率が小さくなっていることがわかります。また、第 2 主成分までの累積寄与率は約 55 %であり、比較的多くの情報が第 2 主成分までに縮約されています。

図 4 に第 1～第 5 主成分スコアの散布図行列を示します。第 1 主成分スコアや第 2 主成分スコアを用いると、クラスごとの分布に違いが確認できます。一方、寄与率の小さな主成分スコアのみでは各クラスの重なりが大きくなっています。一般に、寄与率が小さい主成分ではノイズの影響が大きくなります。そのため、次元削減により元々の情報量がどの程度保持されているかを確認することが重要になります。言い換えますと、多変量データの次元削減過程における情報損失に留意する必要があります。

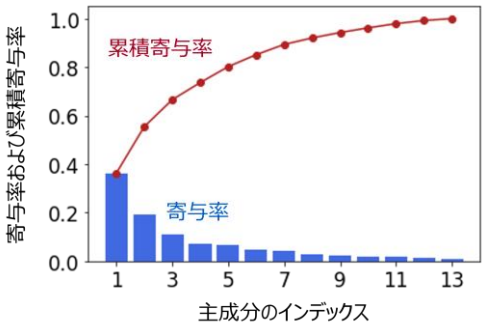

図 3　寄与率および累積寄与率

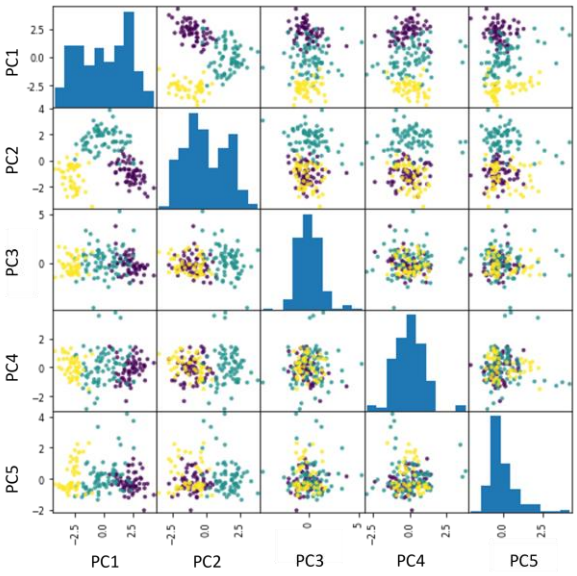

図 4　第 1～第 5 主成分スコアの散布図行列

おわりに

スコアプロットで興味深い特徴が見つかった場合、ローディングプロットと見比べることでデータマイニングを行うことができます。例えば、ガスクロマトグラフなどの分析結果を解析することで、酒類などの製品開発の指針が得られることがあります。

参考文献

1) 永廣卓哉 「データ解析入門 2　<主成分分析によるデータの可視化>」, ORIST テクニカルシート, No. 21-24 (2021)

2) UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data (accessed on April 28th, 2021)

発行日　2021 年 12 月 20 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@tri-osaka.jp