Technical Sheet
No. 21-24

データ解析入門 2 <主成分分析によるデータの可視化>

キーワード：主成分分析、スコアプロット、多変量解析、可視化、次元削減

はじめに

主成分分析（PCA）は、マーケティングや化学分析などの様々な分野で用いられる多変量解析手法です。一般に、高次元データの特徴をそのまま理解することは困難ですが、PCA では高次元データの特徴的な変動の抽出を図ります。本稿では、PCA の基本的なコンセプトを解説します。

多変量データの取り扱い

ここでは、ワインの化学分析に関するデータセット[1]を PCA により解析します。図 1 にデータセットの一部を示しますが、全 178 サンプルのワインについて、アルコール濃度などの 13 項目が評価されています。このような多変量データは行列で表現でき、178 行×13 列のデータ行列となります。このとき、各変数（x_1～x_{13}）は横方向に並び、13 次元空間における各座標軸に対応します。13 次元空間におけるデータの様子を想像することは難しいため、まずは便宜上選んだ 2 つの変数 x_1、x_3 の散布図を確認してみます。図 2(a)に示すように、2 つの座標軸からなる 2 次元平面に 178 個のデータ点がプロットされています。座標軸が 13 個になると、もはや図示できませんが、2 変数の場合と同様に考え、178 個のデータ点が 13 次元空間にプロットされている様子をイメージします。このように考えると、データ行列の各列には x_1～x_{13} の各座標軸上の値（スコア）が格納されていると理解できます。

図 1　データ行列の概念図

データの前処理

解析手法の性質上、PCA ではデータの前処理として中心化を行います。中心化とは、データ行列の各要素から、対応する変数の平均値を引く操作のことであり（式(1)）、中心化によりデータ点が原点周りに移動します（図 2(b)）。また、今回のデータでは、変数 x_1～x_{13} の単位およびスケールはそれぞれ異なります。そこで、中心化を実行後、データを各変数の標準偏差で割ることで、変数のスケールを統一します（図 2(c)）。この操作を標準化といいます（式(2)）。標準化によりスケールの違いに由来する影響を解析結果から取り除くことができます。そのため、PCA に限らず、変数の単位が異なる場合などでは、データの標準化がしばしば重要になります。

$$x_j \leftarrow x_j - \bar{x_j} \tag{1}$$

$$x_j \leftarrow \frac{x_j - \bar{x_j}}{\sigma_j} \tag{2}$$

なお、添え字 j は変数 x のインデックス、$\bar{x_j}$ は変数 x_j の平均値、σ_j は変数 x_j の標準偏差を表します。

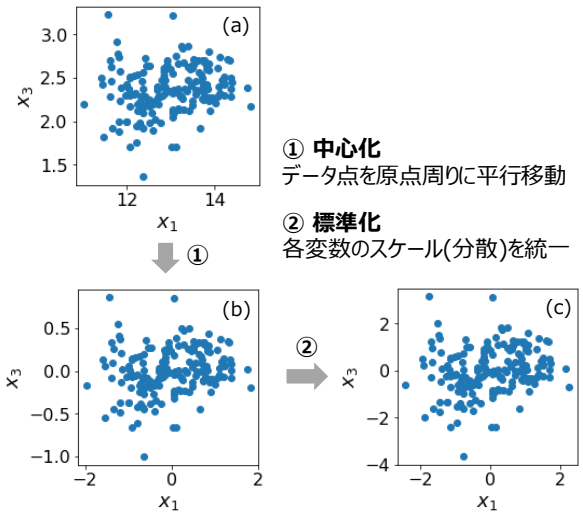

図 2　変数 x_1 および変数 x_3 の散布図：
(a) 前処理なし、(b) 中心化後、(c) 標準化後

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）

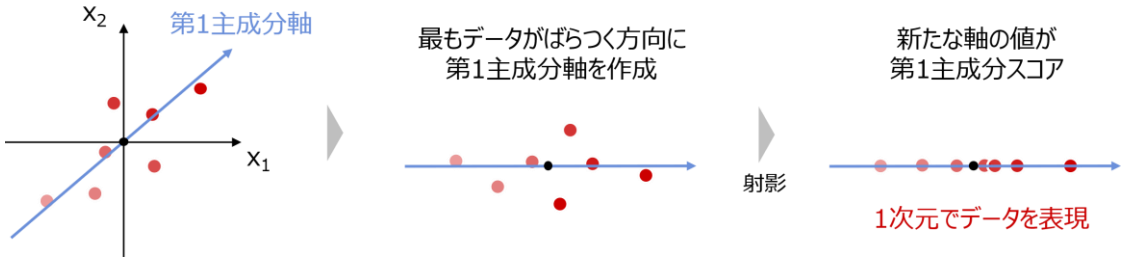

図 3　2 次元平面における PCA による次元削減

PCA のコンセプト

ここまでで、13 次元という多変量データを図示できないことを確認しました。一方で、2 変数であれば散布図を作成できました。ただし、図 2 では全 13 変数のうち、2 変数をプロットしているにすぎず、データ全体の特徴をとらえきれていません。

PCA では、変数 x_1〜x_{13} の線形結合で表される新たな軸（第 1 主成分軸）を作成し、多変量データの特徴を少ない変数でとらえられるようにします。つまり、i 番目のワインの第 1 主成分スコア $t_{i,1}$ は、

$$t_{i,1} = p_{1,1}x_{i,1} + p_{2,1}x_{i,2} + \cdots + p_{13,1}x_{i,13} \qquad (3)$$

と表されます。ここで、$x_{i,j}$ (j = 1, 2,⋯, 13) は i 番目のサンプルの変数 x_j の数値であり、$p_{j,1}$ は後述の処理から求まる重みです。なお、ここでいう主成分（Principal component）とは、PCA で抽出する新たな変数のことであり、化学組成などにおける「主な成分」という意味ではありません。

次に、PCA による主成分軸の求め方を説明します。まず、第 1 主成分のスコアが最も大きく変動するような第 1 主成分を見つけだし、その軸への情報の縮約を図ります。簡単のため、2 変数から 1 変数への次元削減を考えます（図 3）。今、求めたい第 1 主成分軸の方向は、第 1 主成分スコアが最もばらつく方向です。そこで、第 1 主成分軸方向のデータのばらつきを最大化します。これにより第 1 主成分軸に元データの主要な変動をできるだけ写しとる操作が実行されます。

2 次元平面における次元削減はこれで終了ですが、多変数を扱う場合は、第 1 主成分軸に直交する中で、次にスコアが最もばらつく第 2 主成分軸を決定します。続いて、第 1 および第 2 主成分軸に直交する中で、スコアが最もばらつく軸を第 3 主成分軸とする、という操作を繰り返すことで、全ての主成分軸が得られます。

スコアプロットによる可視化

上述の操作により、第 1〜第 13 主成分軸が求まります。PCA により得られる第 1 および第 2 主成分スコアの散布図（スコアプロット）を示します（図 4(a)）。冒頭では述べませんでしたが、本データセットのワインの栽培者はそれぞれ異なり、栽培者ごとに 3 つのクラスラベルが割り振られています。そこで、図 4 では 3 つのクラスごとにプロットを色分けしました。PCA により得られたスコアプロットでは、クラスごとにデータが集まっていますが（図 4(a)）、ただ単に変数 x_1、x_3 をプロットするだけでは、データの特徴を十分にとらえられていません（図 4(b)）。色分けがされていなければ、3 種類のワインの存在を見出すことは困難です（図 2(c)）。このように、PCA による次元削減を実行することで、データの特徴を可視化できる場合があります。また、可視化に用いる主成分の組み合わせを変更することで、より示唆に富む特徴が見つかることもあります。

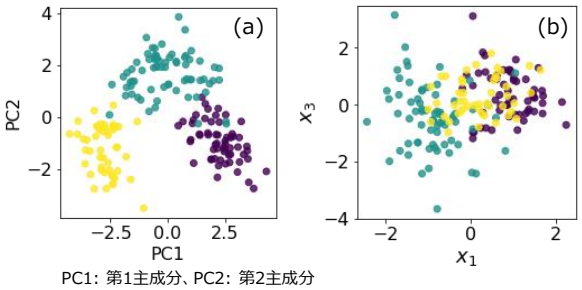

図 4　(a) スコアプロット、(b) 図 2(c)を色分けした散布図

おわりに

本稿では、データ解析のための前処理や PCA の概念について紹介しました。PCA は有用な線形次元削減手法であり、データの特徴を確認するためにしばしば用いられます。次稿では、本稿で用いたワインのデータセットを題材に、PCA によるデータマイニングについて紹介します。

参考文献

1) 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