

## データ解析入門 16 <部分的最小二乗回帰>

**キーワード：**部分的最小二乗法、PLS 回帰、ハイパーパラメータ、交差検証

### はじめに

前稿<sup>1)</sup>では、重回帰分析(MLR)によりアルカンの沸点<sup>2)</sup>を予測しました。本稿では、多重共線性に対してより頑健であり、様々な分野においてよく用いられる部分的最小二乗(PLS: Partial least squares)回帰を紹介します。PLS には成分数というハイパーパラメータが存在します。そこで、交差検証(CV: Cross validation)と呼ばれるハイパーパラメータ調整のための手法についても説明します。

### 重回帰分析(MLR)の計算原理

まず、MLR の計算原理について確認します。MLR では複数の説明変数から目的変数を予測する回帰モデルを最小二乗法により構築します。MLR で求める関数  $f(x)$  は式(1)で示す線形結合で表されます。

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M \quad (1)$$

ここで、 $\beta_i (i = 1, \dots, M)$  は変数  $x_i$  の回帰係数です。前稿<sup>1)</sup>にて紹介したように、MLR では連立方程式  $y = X\beta$  について近似解を計算します。 $y$ 、 $X$ 、 $\beta$  はそれぞれ目的変数ベクトル、データ行列、回帰係数ベクトルです。一般的に、MLR では最小二乗法に基づき、式(2)から回帰係数を算出します。

$$\beta = (X^T X)^{-1} X^T y \quad (2)$$

全データ数  $N$  が変数の数  $M$  を上回る状況では、行列  $X^T X$  が正則となり、逆行列  $(X^T X)^{-1}$  から回帰係数を求めることができます[図 1(a)]。しかしながら、 $M > N$  の場合、独立な式が不足するため、式(2)から回帰係数を一意に求めることができなくなります

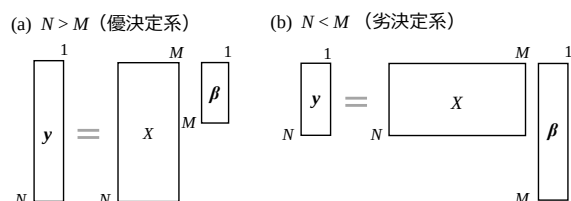


図 1 連立方程式  $y = X\beta$  の概念図

[図 1(b)]。

変数の数がデータ数を上回るというケースは特段珍しいわけではなく、たとえばスペクトルデータのような高次元データを取り扱う場合、図 1(b)に示す劣決定系に直面するかもしれません。

劣決定問題の解法として、正則化と呼ばれる手法が知られています。正則化法では、回帰係数が大きくなりすぎないように回帰係数にペナルティを課すことで、逆行列を計算できるようにします。一方、高次元データを次元削減し、回帰分析を実行可能にするという指針も考えられます。後者の代表的な手法の 1 つが PLS 回帰です。

### PLS 回帰のアルゴリズム

PLS 回帰では次元削減された変数を説明変数として回帰分析します。次元削減後の変数は潜在変数  $t$  などと呼ばれます。主成分分析と異なり、目的変数を加味しながら次元削減を行います。

計算手順の一例を図 2 に示します。ここでは、1 つの目的変数を予測する場合のアルゴリズムを示しています。まず、 $y$  との共分散を最大にする重みベクトル  $w_1$  を計算します。次に、重みベクトル  $w_1$  とデータ行列  $X$  から潜在変数  $t_1$  を計算します。 $t_1$  が得られると、最小二乗法によりローディングベクトル  $p_1$  を計算し、行列  $X$  から  $t_1 p_1^T$  の影響を除去します。また、最小二乗法により  $t_1$  の回帰係数  $q_1$  を計算し、 $y$  を更新します。以上の操作を繰り返します。

上述のアルゴリズムにより、PLS 回帰の各潜在変数は直交するように抽出されます。したがって、MLR で問題となった多重共線性の問題を回避することができます。また、潜在変数の数を  $N$  よりも小さくすることで、劣決定系であっても PLS 回帰が可能となります。ここで、潜在変数の数は成分数とも呼ばれ、PLS 回帰モデルにおけるハイパーパラメータです。ハイパーパラメータとは、推論や予測の枠組みの中で決定されないパラメータのことであり<sup>3)</sup>、機械学習の予測精度を左右するため適切な値に調整する必要があります。

### 交差検証(クロスバリデーション)

交差検証(クロスバリデーション、CV)は、訓練

| 目的変数 $y$ と相関する潜在変数 $t$ を抽出し、PLS回帰モデルの説明変数とする    |                                   |                                         |
|-------------------------------------------------|-----------------------------------|-----------------------------------------|
| 1. 重みベクトル $w_1$ (長さ 1)を計算                       | $w_1 = \frac{X^T y}{\ X^T y\ }$   | $w_2 = \frac{X_2^T y_2}{\ X_2^T y_2\ }$ |
| 2. データ行列 $X$ と $w_1$ から潜在変数 $t_1$ を計算           | $t_1 = X w_1$                     | $t_2 = X_2 w_2$                         |
| 3. 最小二乗法によりローディングベクトル $p_1$ を計算                 | $p_1 = \frac{X^T t_1}{t_1^T t_1}$ | $p_2 = \frac{X_2^T t_2}{t_2^T t_2}$     |
| 4. $X$ から $t_1 p_1^T$ の影響を取り除き、 $X$ を $X_2$ に更新 | $X_2 = X - t_1 p_1^T$             | $X_3 = X_2 - t_2 p_2^T$                 |
| 5. 最小二乗法により潜在変数 $t_1$ の回帰係数 $q_1$ を計算           | $q_1 = \frac{y^T t_1}{t_1^T t_1}$ | $q_2 = \frac{y_2^T t_2}{t_2^T t_2}$     |
| 6. $y$ から $t_1 q_1$ の影響を取り除き、 $y$ を $y_2$ に更新   | $y_2 = y - t_1 q_1$               | $y_3 = y_2 - t_2 q_2$                   |
| 7. 潜在変数の数だけ上記操作を繰り返す                            |                                   |                                         |

図 2 PLS 回帰のアルゴリズム

データを、たとえば 5 分割し、80 % のデータでモデルを訓練し、残りの 20 % のデータで予測精度を検証します。この場合、全 5 通りの訓練および検証データの分割が考えられ、決定係数  $R^2$  などの予測精度を 5 回評価します。その後、 $R^2$  などの平均値を計算します。この操作を各ハイパーパラメータにおいて実行し、最も良好なスコアが得られた値を採用します。

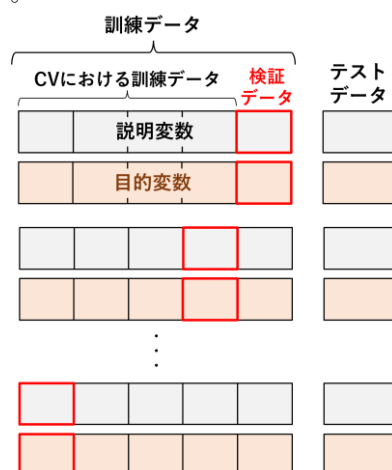


図 3 交差検証の概念図

次に、アルカンの沸点データセット<sup>2)</sup>を用い、10 分割した交差検証 (10-fold CV) の結果を示します (図 4)。横軸は PLS 回帰の成分数 (ハイパーパラメータ) であり、縦軸は  $R^2$  を表します。成分数を増やすとモデルの表現力が高まるため、訓練データへの過剰適合に注意しなければなりません。図 4 の青点は全ての訓練データで学習し、全ての訓練データを予測した結果です。この結果と 10-fold CV の結果 (橙点) との比較から、4 以上の成分数における過剰適合の傾向が示唆されました。

## PLS 回帰の予測結果

図 5 に、RDKit により得た分子屈折率 (MolMR) を説明変数とした単回帰分析、複数の RDKit 記述

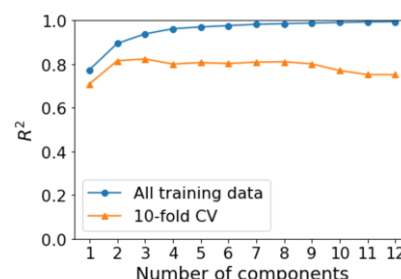


図 4 PLS 回帰の学習結果 (訓練データ 50 件)

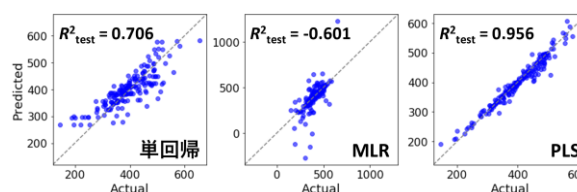


図 5 予測結果 (訓練データ 100 件)

子を説明変数とした MLR および PLS 回帰 (成分数 2) の予測結果を示します。訓練データ (96 次元) は 100 件としました。PLS 回帰では、MLR より高い  $R^2$  が確認されました。今回、説明変数の次元が訓練データ数とほとんど同じであり、過剰適合に注意すべき状況でした。そのため、シンプルな単回帰の方が MLR よりも高い  $R^2$  を示しました。

## おわりに

PLS 回帰は多重共線性に頑健な線形回帰モデルとして広く用いられてきました。また、PLS 回帰を非線形回帰モデルに拡張する手法や PLS 回帰をベースとする変数選択法なども提案されています。

## 参考文献

- 1) 永廣卓哉: データ解析入門 15 <回帰分析の概説>、ORIST テクニカルシート、No. 22-28 (2022)
- 2) E. S. Goll and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **39**, 974-983 (1999)
- 3) <https://ja.wikipedia.org/wiki/ハイパーパラメータ> (accessed on October 7th, 2022)