

『AI・データサイエンスのための数学入門』学習補助資料

PCA：座標を変えて、少数の軸を残す

中心化・射影・復元をつなぐ補助ノート

1 「軸を変える」と「次元を減らす」を分ける

PCA は、まずデータの平均を原点に移し、ばらつきの大きい方向へ座標軸を取り直します。**すべての主成分を残す段階では、情報は失われません。**一部の軸の座標を捨てたときに、次元削減と近似が起こります。

行列の大きさを確認する

N 個の観測、 p 個の変数をもつデータを考えます。本ノートでは、共分散の分母を本文と同じ N に統一します。

記号	大きさ	意味
$\tilde{X}$	$N \times p$	各列から、その列の平均を引いたデータ
$\Sigma = \tilde{X}^\top \tilde{X} / N$	$p \times p$	分散共分散行列
$V_k = (v_1, \dots, v_k)$	$p \times k$	単位長の主成分方向を列に並べた行列
$Z = \tilde{X} V_k$	$N \times k$	各観測の新しい座標、主成分スコア

中心化が必要な理由

例えば身長と体重なら、中心化後の値は「平均身長からの差」「平均体重からの差」です。PCA は、平均の周りで観測がどう広がるかを調べます。中心化せず $X^\top X / N$ を用いると、ばらつきに加えて平均位置の影響も入ります。

$$\frac{1}{N} X^\top X = \Sigma + \bar{x} \bar{x}^\top.$$

ここで X の各行は観測値の転置、 $\bar{x}$ は平均ベクトルです。

方向・スコア・固有値の役割

方向 v_j は変数をどう組み合わせるか、スコア z_{ij} は各観測の新しい座標、固有値 λ_j はその座標全体の分散です。3 つを区別すると、PCA の式を読みやすくなります。

分母 N と $N-1$ の違い

共分散の分母を $N-1$ に変えると、行列全体と固有値が $N/(N-1)$ 倍になります。同じ中心化データなら、主成分方向・スコア・寄与率は変わりません。標準化も行う場合は、標準偏差の定義もそろえます。

2 固有値が「射影後の分散」になる理由

単位方向 $\mathbf{v}$ に射影したスコアベクトルは $\mathbf{z} = \tilde{\mathbf{X}}\mathbf{v}$ です。中心化によりスコアの平均は 0 なので、

$$\frac{1}{N} \sum_{i=1}^N z_i^2 = \frac{1}{N} \mathbf{z}^\top \mathbf{z} = \mathbf{v}^\top \frac{\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}}{N} \mathbf{v} = \boxed{\mathbf{v}^\top \Sigma \mathbf{v}}.$$

特に $\Sigma \mathbf{v}_j = \lambda_j \mathbf{v}_j$ 、 $\|\mathbf{v}_j\| = 1$ なら分散は λ_j です。単位長にしないと、方向を変えずにベクトルを大きくするだけで分散を増やせてしまいます。

第 2 主成分以降は、直交する方向から探す

第 1 主成分と直交する単位方向に制限して分散を最大化すると、第 2 主成分が得られます。同様に、既に選んだ方向に直交する範囲で順に最大化します。

正規直交固有ベクトルを並べた V を使うと、全スコアの共分散は

$$\frac{1}{N} \mathbf{Z}^\top \mathbf{Z} = \mathbf{V}^\top \Sigma \mathbf{V} = \text{diag}(\lambda_1, \dots, \lambda_p).$$

したがって異なる主成分スコアの共分散は 0 です。ただし、無相関は一般に独立を意味しません。

数値例：共分散行列 $\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$

固有値は 6, 2、方向は $\mathbf{v}_1 = (1, 1)^\top / \sqrt{2}$ 、 $\mathbf{v}_2 = (1, -1)^\top / \sqrt{2}$ です。中心化点 $(3, 1)^\top$ のスコアは $(2\sqrt{2}, \sqrt{2})$ です。

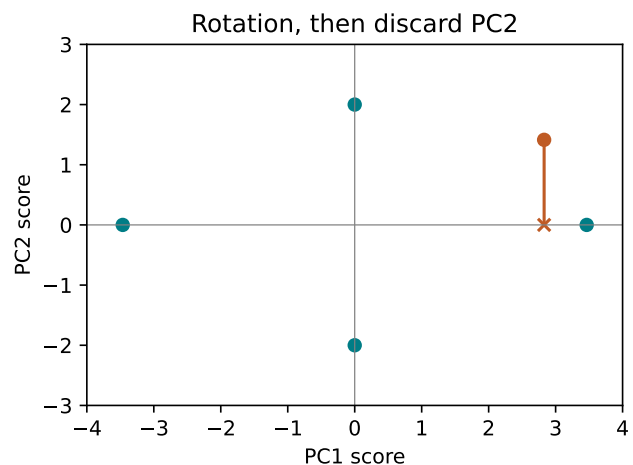
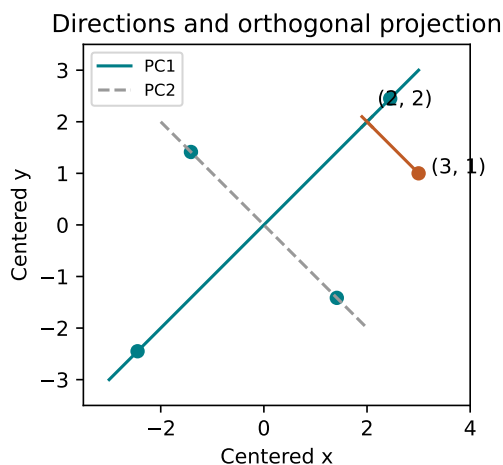


図 1：青の 4 点は共分散が上記の行列となる例。橙は別の点 (3, 1) の変換と射影です。PC1 だけ残すとスコアは $(2\sqrt{2}, 0)$ となり、元の座標では (2, 2) に復元されます。

3 PCA と低ランク近似：何を最小にしているか

k 個の主成分から中心化データを復元すると、

$$\hat{\tilde{X}} = ZV_k^\top = \tilde{X}V_kV_k^\top.$$

元の単位のデータへ戻すには、各行に平均ベクトルの転置を加えます。1 点なら

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \sum_{j=1}^k z_j \mathbf{v}_j.$$

射影で捨てる部分は、残す部分と直交する

各点の中心化ベクトルを、残す方向の成分と残りに分けます。ピタゴラスの定理より、全点で足し合わせると

$$\|\tilde{X}\|_F^2 = \|\tilde{X}V_kV_k^\top\|_F^2 + \|\tilde{X} - \tilde{X}V_kV_k^\top\|_F^2.$$

左辺は固定されています。したがって、残す分散を最大化することは、元の空間での二乗復元誤差を最小化することと同じです。上位 k 方向を選ぶと、

$$\|\tilde{X} - \hat{\tilde{X}}\|_F^2 = N \sum_{j=k+1}^p \lambda_j.$$

1 観測当たりの平均二乗距離は $\sum_{j>k} \lambda_j$ です。1 成分当たりの平均二乗誤差なら、さらに p で割ります。

SVD との対応

$\tilde{X} = UDV^\top$ を特異値分解とすると、

$$\Sigma = \frac{1}{N}VD^\top DV^\top, \quad \lambda_j = \frac{d_j^2}{N}.$$

主成分方向は右特異ベクトル、スコアは $\tilde{X}V_k = U_kD_k$ です。打ち切り SVD が PCA による復元を与えます。特異値は d_j 、分散は d_j^2/N なので、二乗を忘れないようにします。

0 の固有値と、方向の一意性

共分散行列は半正定値なので固有値は非負です。中心化データの階数は高々 $\min(N-1, p)$ です。固有値 0 の方向には観測のばらつきがありません。固有ベクトルの符号は自由で、同じ固有値が重なる空間内では軸の選び方も一意ではありません。

4 寄与率・標準化・解釈の注意

本文の固有値 (50, 25, 15, 7, 3) では、全分散 100、累積寄与率は 50%、75%、90%、97%、100% です。

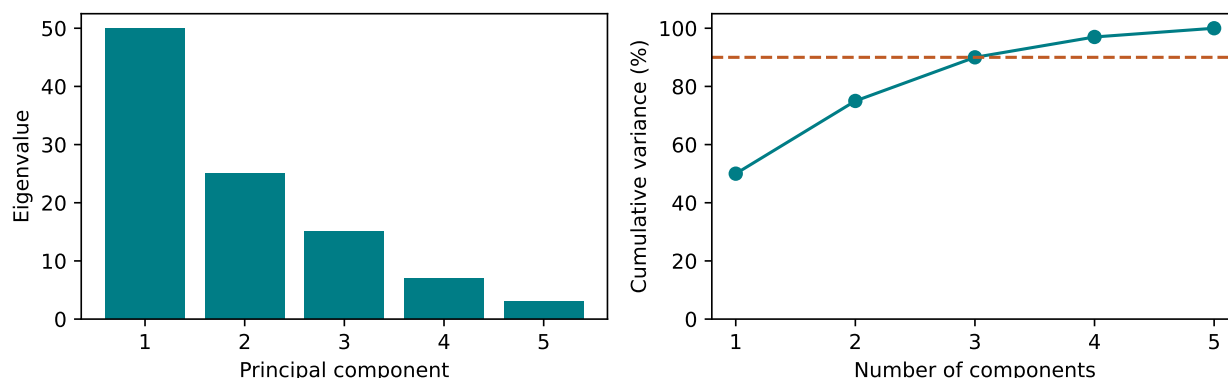


図 2：左は固有値を並べたスクリープロット、右は累積寄与率。2 次元表示では 75%、3 次元なら 90% の分散を保持します。90% は普遍的な採用基準ではなく、目的に応じた目安です。

中心化と標準化は別の操作

中心化は $x_{ij} - \bar{x}_j$ 、標準化は $(x_{ij} - \bar{x}_j)/s_j$ です。標準化して PCA を行うことは、同じ分母で計算した相関行列に基づく PCA に対応します。分散 0 の変数は標準化できません。

例えば無相関な 2 変数の共分散が $\text{diag}(10000, 1)$ なら、PC1 は第 1 変数方向です。標準化後は単位行列となり、両方向の分散は等しく、PC1 は一意に決まりません。単位差を除く利点がある一方、標準化はどの誤差を重視するかも変える操作です。

「情報」と「分散」は同じ意味ではない

寄与率は、選んだスケールにおける二乗偏差の保持率です。分類に重要な差が小分散方向にあれば、その差を失うことがあります。また、大分散方向が測定誤差や外れ値を表す場合もあります。PCA は目的変数を使わないため、予測に最適な方向とは限りません。

本文の 5 教科の例を、復元から見る

スコアが $z_1 = 0, z_2 = 6$ なら、2 主成分から復元する中心化得点は

$$0\mathbf{v}_1 + 6\mathbf{v}_2 = (-3, 3, -3, 3, 0)^\top.$$

元の $(4, 8, -2, 6, -16)^\top$ との差は $(7, 5, 1, 3, -16)^\top$ です。PC1 スコアが 0 でも、すべての教科が平均的とは限りません。スコアが正であることも、各教科の偏差の符号を一律には決めません。

5 本文の練習問題：解答と補足

問 1：分散共分散行列

$\bar{X} = \bar{Y} = 0$ です。分母を 3 として、

$$s_{XX} = \frac{1+0+1}{3} = \frac{2}{3}, \quad s_{YY} = \frac{4+0+4}{3} = \frac{8}{3}, \quad s_{XY} = \frac{2+0+2}{3} = \frac{4}{3}.$$

したがって

$$\Sigma = \frac{1}{3} \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix}.$$

共分散は正で、実際には $Y = 2X$ という完全な線形関係があります。

補足：固有値は $10/3, 0$ 、PC1 方向は $(1, 2)^\top / \sqrt{5}$ です。PC1 スコアは $(-\sqrt{5}, 0, \sqrt{5})$ で、1 次元だけから元の全点を正確に復元できます。

問 2：寄与率

全分散は $6 + 2 = 8$ です。PC1 の寄与率は $6/8 = 75\%$ 、PC2 は $2/8 = 25\%$ 。 $(1, 1)^\top$ 方向の方がばらつきが大きくなります。スコア計算では、方向を単位ベクトルに直します。

問 3：スコアと限界

$$z_1 = \frac{3+1}{\sqrt{2}} = 2\sqrt{2}, \quad z_2 = \frac{3-1}{\sqrt{2}} = \sqrt{2}.$$

2 軸とも残すと $z_1 \mathbf{v}_1 + z_2 \mathbf{v}_2 = (3, 1)^\top$ です。PC1 だけなら $(2, 2)^\top$ となり、誤差は $(1, -1)^\top$ 、二乗距離は 2 です。これは捨てたスコアの二乗 z_2^2 と一致します。

小分散方向にある分類・予測上の重要な特徴や、少数の観測の特徴が失われる可能性があります。また、線形な射影なので、曲面に沿った非線形構造を少数の軸で十分に表せるとは限りません。

新しいデータにも同じ座標系を使う

学習データで求めた平均・標準偏差・主成分方向を保存し、新しいデータにもそのまま適用します。評価用データを含めて前処理や PCA を学習すると、予測性能の評価に情報が漏れることがあります。

6 追加問題と解答

問 1：次元の選択 固有値が $(9, 4, 1)$ のとき、90% 以上の分散を残す最小の主成分数と、1 観測当たりの二乗復元誤差を求めよ。

問 2：平均を戻す 平均が $(10, 20)^T$ 、PC1 方向が $(1, 1)^T/\sqrt{2}$ 、スコアが $3\sqrt{2}$ だった。PC1 だけを用いた元の座標での復元値を求めよ。

問 3：固有ベクトルの符号 v_1 を $-v_1$ に変えると、スコア、固有値、復元値はどう変わるか。

問 4：SVD と誤差 $N = 5$ 、中心化データの特異値が $(6, 2)$ だった。共分散の固有値、PC1 寄与率、ランク 1 復元の残差平方和を求めよ。

問 5：標準化の判断 単位が大きく違う変数を扱うとき、標準化の利点と、無条件に行うべきでない理由を述べよ。

問 6：高い寄与率の意味 「累積寄与率 99% なら、分類に必要な情報も 99% 残る」は正しいか。

解答・解説

- 全分散 14。PC1 だけでは $9/14 \approx 64.3\%$ 、2 軸なら $13/14 \approx 92.9\%$ なので **2 主成分**。捨てる固有値は 1 なので平均二乗距離は 1 です。
- $(10, 20)^T + 3\sqrt{2}(1, 1)^T/\sqrt{2} = (13, 23)^T$ 。
- スコアの符号が反転します。固有値は変わらず、 $(-z_1)(-v_1) = z_1 v_1$ なので復元値も変わりません。
- 固有値は $36/5, 4/5$ 、PC1 寄与率は $36/40 = 90\%$ 。残差平方和は捨てた特異値の二乗 **4** です。
- 単位や数値スケールによる支配を抑えられます。ただし、元の分散の差が意味をもつ場合や、微小な変動がノイズ中心の場合には、標準化が目的に合わないこともあります。
- 正しくありません。分散の保持率であって、分類性能の保持率ではありません。捨てた 1% の方向が分類を決めることもありえます。

講義で押さえること

中心化し、単位方向へ射影する。固有値は分散、スコアは座標。少数の軸からの復元は近似であり、その誤差は捨てた固有値で測れる。寄与率と課題に必要な情報は区別する。