主成分解析 (PCA: Principal Component Analysis)

山内 仁喬

2021 年 10 月 31 日

主成分解析 [1] は, 多次元空間の中で生体分子がどのような挙動を示しているのか解析する際に用いられる. 蛋白質の揺らぎは異方性が高いため, 物理量とも関連付けられることができる [2, 3, 4]. また, シミュレーションで得られた構造のクラスタリングをする際にも主成分解析は有用である.

1 主成分解析の基礎

主成分解析では, 原子座標 $\mathbf{Q}$ を原子座標の線形結合として表される独立な集団座標 (=主成分座標) $\mathbf{\Xi} = \{\xi_{ak}\}$ への線形変換を考える. 原子座標から主成分座標への変換行列を $\mathbf{A}^{\mathrm{t}}$ とすると

$$\mathbf{\Xi} = \mathbf{A}^{\mathrm{t}}\mathbf{Q} \tag{1}$$

とかける. 以降の便利のため, k 番目の構造における座標ベクトルを $\mathbf{Q}_k$, $\mathbf{\Xi}_k$ と書くことにする. 基底が正規直交であるとすると,

$$\mathbf{A}^{\mathrm{t}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{t}} = \mathbf{I} \tag{2}$$

が成立する. ここで $\mathbf{I}$ は単位行列である. 主成分分析では, 新たに得られた軸に置いてデータの分散が最大になるように線形変換を施す. 変換後の分散は

$$\begin{aligned}
\tilde{\sigma}^2(\mathbf{A}) &= \sum_{k=1}^{K} w_k(\mathbf{\Xi}_k - \langle\mathbf{\Xi}\rangle)^{\mathrm{t}}(\mathbf{\Xi}_k - \langle\mathbf{\Xi}\rangle) \\
&= \sum_{k=1}^{K} w_k \left\{\mathbf{A}^{\mathrm{t}}(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)\right\}^{\mathrm{t}} \left\{\mathbf{A}^{\mathrm{t}}(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)\right\} \\
&= \sum_{k=1}^{K} w_k \mathrm{Tr}\left[\mathbf{A}^{\mathrm{t}}(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)^{\mathrm{t}}\mathbf{A}\right] \\
&= \mathrm{Tr}\left[\mathbf{A}^{\mathrm{t}} \sum_{k=1}^{K} w_k(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)(\mathbf{Q}_k - \langle\mathbf{Q}\rangle)^{\mathrm{t}}\mathbf{A}\right] \\
&= \mathrm{Tr}\left[\mathbf{A}^{\mathrm{t}}\mathbf{\Sigma}\mathbf{A}\right]
\end{aligned} \tag{3}$$

と計算される. ここで, データの平均を計算するためにデータの重み w_k を導入した. また, $\mathbf{\Sigma}$ は分散共分散行列である. また, 途中の式変形において $\mathbf{x}^{\mathrm{t}}\mathbf{x} = \mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{t}})$ が成り立つことを使用た. $\mathbf{A}^{\mathrm{t}}\mathbf{A}$ の制約のもとで, 式 (3) を最大にする $\mathbf{A}$ を求める問題に帰着する:

$$J(\mathbf{A}) \equiv \mathrm{Tr}[\mathbf{A}^{\mathrm{t}}\mathbf{\Sigma}\mathbf{A}] - [(\mathbf{A}^{\mathrm{t}}\mathbf{A} - \mathrm{I})\mathbf{\Sigma}] \tag{4}$$

つまり, Σ は対角行列である. $\mathbf{A}$ で偏微分すると

$$\frac{\partial J(\mathbf{A})}{\partial \mathbf{A}} = 2\mathbf{\Sigma A} - 2\mathbf{A\Sigma} = 0 \tag{5}$$

ここで, (n, m) 行列 $\mathbf{X}$, n 次元正方行列 $\mathbf{Y}$ について

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}\{\mathbf{X}^{\mathrm{t}}\mathbf{Y}\mathrm{X}\} = (\mathbf{Y} + \mathbf{Y}^{\mathrm{t}})\mathbf{X} \tag{6}$$

を用いた. ゆえに,

$$\begin{aligned} \mathbf{\Sigma A} &= \mathbf{A\Sigma} \\ \mathbf{A}^{\mathrm{t}}\mathbf{\Sigma A} &= \mathbf{\Sigma} \end{aligned} \tag{7}$$

となり, $\mathbf{A}^{\mathrm{t}}$ は $\mathbf{\Sigma}$ を対角化する行列となる.

2 主成分解析のタンパク質への応用

N 個の原子からなる系を考える. ここでは各原子の座標 x, y, z を, より一般的な形である q で表すことにする. 実際に主成分解析を行う場合, 全原子を扱わずに主鎖の原子座標や C_α 原子の座標のみに適用することが多い. その他の内部座標 (例えば二面角や各アミノ酸残基の重心座標) を用いてもよい. 分子シミュレーションによって, K 個の構造を得たとする. 各瞬間座標を, それぞれの平均座標からの変位に変換して, 質量の平方根で重み付けをする. 瞬間構造の座標を行方向, 時系列を列方向にまとめると, $3N \times K$ 行列

$$\mathbf{Q} = \{Q_{ik}\} = \{\sqrt{m_i}\,(q_{ik} - \langle q_i \rangle)\} \tag{8}$$

を得る. ここで, $\langle \mathbf{q}_i \rangle$ は座標 $\mathbf{q}_{ik}$ の平均構造である. $\langle \mathbf{q}_i \rangle$ は以下の手順で求められる: まず仮の $\mathbf{q}'_{ik}$ [*1] を定める. 続いて, $\mathbf{q}'_{ik}$ に対して K 個の瞬間構造を全て best fit させて平均構造を求める. さらに求めた平均構造に対して瞬間構造全てを best fit し平均構造を計算する. この操作を繰り返すことで平均構造を収束させる. 通常は数回のイテレーションで平均構造に収束するはずである.

主成分解析では, 原子座標 $\mathbf{Q}$ を原子座標の線形結合として表される独立な集団座標 (=主成分座標) $\mathbf{\Sigma} = \{\sigma_{ak}\}$ への線形変換を考える. 原子座標から主成分座標への変換行列を $\mathbf{P}^{\mathrm{t}}$ とすると

$$\mathbf{\Sigma} = \mathbf{P}^{\mathrm{t}}\mathbf{Q} \tag{9}$$

とかける. ここで, 分散共分散行列 $\mathbf{R}$ を与える:

$$\begin{aligned} \mathbf{R} &= \langle \mathbf{Q}^{\mathrm{t}}\mathbf{Q} \rangle \\ &= \left\{ \sqrt{m_i m_j}\, \langle (q_{ik} - \langle q_i \rangle)\,(q_{jk} - \langle q_j \rangle) \rangle \right\} \\ &= \left\{ \sqrt{m_i m_j}\, (\langle q_{ik} q_{jk} \rangle - \langle q_i \rangle \langle q_j \rangle) \right\}. \end{aligned}$$

この行列は明らかに対称行列である. 各構造に対して重み w_k が課せられているとすると, 平均は

$$\begin{aligned} \langle q_i \rangle &= \sum_{k=1}^{K} w_k q_{ik} \\ \langle q_i q_j \rangle &= \sum_{k=1}^{K} w_k (q_{ik} q_{jk}) \end{aligned}$$

[*1] 自由度の数が同じであれば, どんな座標値でもいい. 例えば, 初期構造 $\mathbf{q}_{i1}$ などを用いる.

と計算される. ただし重み w_k は規格化されているものとする.

原子座標から主成分座標への変換行列 $\mathbf{P}^{\mathrm{t}}$ は, 原子座標の分散共分散行列 $\mathbf{R}$ の標準固有値問題 (対角化) から得られる.

$$\mathbf{RP} = \mathbf{P\Lambda} \tag{10}$$

ここで $\mathbf{\Lambda}$ は対角行列である. その成分 $\lambda_a (a = 1, 2, \cdots, 3N)$ は固有値であり, 主成分座標方向の分散を表す: $\lambda_a = \langle \sigma_{ak}^2 \rangle$. 事前に原子座標に対して best fit をして並進と回転の自由度を取り除いている場合, 6 つの固有値がゼロとなる. 行列 $\mathbf{P}$ の各列は, 分散共分散行列 $\mathbf{R}$ の固有ベクトル $\mathbf{u}_a$ である:

$$\mathbf{P} = \left[\begin{array}{cccc} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{3N} \end{array} \right] . \tag{11}$$

固有ベクトル $\mathbf{u}_a$ は a 番目の主成分軸に対応する. また, 固有ベクトルは規格化されているものとする. つまり, 変換行列は $\mathbf{PP}^{\mathrm{t}} = \mathbf{P}^{\mathrm{t}}\mathbf{P} = \mathbf{I}$ を満たす. 原子座標 $\mathbf{Q}$ は平均構造からの変位であることから, 座標軸の原点は分布の平均値になっている.

主成分座標 $\mathbf{\Sigma} = \{\sigma_{ak}\}$ を得るには, $\mathbf{Q}$ を主成分座標に射影すればよい:

$$\mathbf{\Sigma} = \{\sigma_{ak}\} = \mathbf{P}^{\mathrm{t}}\mathbf{Q}. \tag{12}$$

$\{\sigma_{ak}\}$ の a 行目の列ベクトル $(k = 1, 2, \cdots, K)$ は第 a 主成分座標の時系列 $\sigma_a(t_k)$ であり, 具体的に次のように計算される:

$$\sigma_a(t_k) = \left[\begin{array}{cccc} u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(3N)} \\ u_2^{(1)} & u_2^{(2)} & \cdots & u_2^{(3N)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{3N}^{(1)} & u_{3N}^{(2)} & \cdots & u_{3N}^{(3N)} \end{array} \right] \left[\begin{array}{c} Q_1(t_k) \\ Q_2(t_k) \\ \vdots \\ Q_{3N}(t_k) \end{array} \right] \tag{13}$$

$$= \left[\begin{array}{c} u_1^{(1)}Q_1(t_k) + u_1^{(2)}Q_2(t_k) + \cdots u_1^{(3N)}Q_{3N}(t_k) \\ u_2^{(1)}Q_1(t_k) + u_2^{(2)}Q_2(t_k) + \cdots u_2^{(3N)}Q_{3N}(t_k) \\ \vdots \\ u_{3N}^{(1)}Q_1(t_k) + u_{3N}^{(2)}Q_2(t_k) + \cdots u_{3N}^{(3N)}Q_{3N}(t_k) \end{array} \right] . \tag{14}$$

λ_a は σ_{ak} の k に関する分散であることから次の式が成立する:

$$\lambda_a = \langle \sigma_{ak}^2 \rangle = \sum_{i=1}^{3N} v_{ik}^2 \langle q_{ak}^2 \rangle. \tag{15}$$

また, 分散の総和への主成分 a の寄与率は

$$\frac{\lambda_a}{\sum_{a'=1}^{3N-6} \lambda_{a'}} \tag{16}$$

から計算することができる.

$\mathbf{\Sigma}$ を標準偏差である $\lambda^{-\frac{1}{2}}$ でスケールして

$$\mathbf{U}^{\mathrm{t}} = \lambda^{-\frac{1}{2}}\mathbf{\Sigma} = \lambda^{-\frac{1}{2}}\mathbf{P}^{\mathrm{t}}\mathbf{Q} \tag{17}$$

とすると, $\mathbf{U}^{\mathrm{t}}\mathbf{U} = \mathbf{I}$ である. $K \times 3N$ 行列 $\mathbf{Q}^{\mathrm{t}}$ の特異値分解

$$\mathbf{Q}^{\mathrm{t}} = \mathbf{U}\lambda^{\frac{1}{2}}\mathbf{P}^{\mathrm{t}} \tag{18}$$

からも変換行列 $\mathbf{P}^{\mathrm{t}}$ を得ることができる.

エネルギー面が完全に多次元のパラボラとなっている時は, 得られる変換行列は基準振動解析の値と一致する. この時, 固有値 λ_a は固有振動数 ω_a と, 次の関係を持つ [2, 3, 4]:

$$\lambda_a = \frac{k_B T}{\omega_a^2}. \tag{19}$$

3 PCA の実行例

簡単な例題を通して, PCA がどのように実行されるかを確認する. このような例は, プログラムを実装したときのデバッグに役に立つ.

3.1 PCA を解析的実行した例 1

次のデータの集合 $\{\mathbf{x}_i\}$ を考える:

$$\mathbf{x}_1 = (1,2), \quad \mathbf{x}_2 = (2,4), \quad \mathbf{x}_3 = (3,6) \tag{20}$$

あるいは, データの平均がゼロになるようにシフトしたデータの集合 $\{\mathbf{x}'_i\}$

$$\mathbf{x}'_1 = (-1,-2), \quad \mathbf{x}'_2 = (0,0), \quad \mathbf{x}'_3 = (1,2) \tag{21}$$

を考えても得られる主成分軸は変わらないが, 今回は元の集合 $\{\mathbf{x}_i\}$ を考えていく. [*2] 分散共分散行列は,

$$\begin{aligned}
\begin{bmatrix} \langle x^2\rangle - \langle x\rangle^2 & \langle xy\rangle - \langle x\rangle\langle y\rangle \\ \langle xy\rangle - \langle x\rangle\langle y\rangle & \langle y^2\rangle - \langle y\rangle^2 \end{bmatrix} &= \begin{bmatrix} \frac{1+4+9}{3} - \left(\frac{1+2+3}{3}\right)^2 & \frac{2+9+18}{3} - \left(\frac{1+2+3}{3}\right)\left(\frac{2+4+6}{3}\right) \\ \frac{2+9+18}{3} - \left(\frac{1+2+3}{3}\right)\left(\frac{2+4+6}{3}\right) & \frac{4+16+36}{3} - \left(\frac{2+4+6}{3}\right)^2 \end{bmatrix} \\
&= \begin{bmatrix} 0.666\ldots & 1.333\ldots \\ 1.333\ldots & 2.666\ldots \end{bmatrix}
\end{aligned}$$

と計算される. 分散共分散行列の固有値と固有ベクトルを求めると

$$\lambda_1 = 3.333\ldots, \quad \mathbf{u}_1 = (0.45, 0.89)^t \tag{22}$$

$$\lambda_2 = 0.000\ldots, \quad \mathbf{u}_2 = (-0.89, 0.45)^t \tag{23}$$

を得る. 最後に, 元のデータを主成分軸上に射影する. 通常平均値が主成分軸の原点になるようにするので, シフトしたデータの集合 $\{\mathbf{x}'_i\}$ を主成分軸上に射影して, 新たな座標の集合 $\mathbf{y}$ を得る.

$$\mathbf{y} = \begin{bmatrix} 0.45 & 0.89 \\ -0.89 & 0.45 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \end{bmatrix} \tag{24}$$

$$\simeq \begin{bmatrix} -2.23 & 0 & 2.23 \\ 0 & 0 & 0 \end{bmatrix} \tag{25}$$

3.2 PCA を解析的実行した例 2

次のデータの集合 $\{\mathbf{x}_i\}$ を考える:

$$\begin{aligned}
&\mathbf{x}_1 = (1,1), \quad \mathbf{x}_2 = (3,3), \quad \mathbf{x}_3 = (5,5), \\
&\mathbf{x}_4 = (2,4), \quad \mathbf{x}_5 = (4,2)
\end{aligned}$$

[*2] 実際は, シフトしたデータの集合 $\{\mathbf{x}'_i\}$ を用いた方が, 分散共分散行列の $\langle x\rangle$ や $\langle y\rangle$ 項が 0 となるため, 計算が楽である.

なお, データ点の平均値は $\mathbf{x}_{\rm ave} = (3,\ 3)$ である. データの平均がゼロになるようにシフトしたデータの集合 $\{\mathbf{x}'_i\}$ は

$$\begin{aligned}
&\mathbf{x}'_1 = (-2,-2), \quad \mathbf{x}'_2 = (0,0), \quad \mathbf{x}'_3 = (2,2),\\
&\mathbf{x}'_4 = (-1,1), \qquad \mathbf{x}'_5 = (1,-1),
\end{aligned}$$

である. 分散共分散行列は,

$$\begin{aligned}
\begin{bmatrix} \langle x^2\rangle - \langle x\rangle^2 & \langle xy\rangle - \langle x\rangle\langle y\rangle \\ \langle xy\rangle - \langle x\rangle\langle y\rangle & \langle y^2\rangle - \langle y\rangle^2 \end{bmatrix} &= \begin{bmatrix} \frac{4+0+4+1+1}{5} & \frac{4+0+4-1-1}{5} \\ \frac{4+0+4-1-1}{5} & \frac{4+0+4+1+1}{5} \end{bmatrix}\\
&= \begin{bmatrix} 2 & 1.2 \\ 1.2 & 2 \end{bmatrix}
\end{aligned}$$

分散共分散行列の固有値と固有ベクトルを求めると

$$\lambda_1 = 3.2 \quad \mathbf{u}_1 = (0.71, 0.71)^t \tag{26}$$

$$\lambda_2 = 0.8 \quad \mathbf{u}_2 = (-0.71, 0.71)^t \tag{27}$$

を得る. 最後に, 元のデータを主成分軸上に射影する. 通常平均値が主成分軸の原点になるようにするので, シフトしたデータの集合 $\{\mathbf{x}'_i\}$ を主成分軸上に射影して, 新たな座標の集合 $\mathbf{y}$ を得る.

$$\mathbf{y} = \begin{bmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{bmatrix} \begin{bmatrix} -2 & 0 & 2 & -1 & 1 \\ -2 & 0 & 2 & 1 & -1 \end{bmatrix} \tag{28}$$

$$\simeq \begin{bmatrix} -2.84 & 0 & 2.84 & 0 & 0 \\ 0 & 0 & 0 & 1.42 & -1.42 \end{bmatrix} \tag{29}$$

参考文献

[1] Martin Karplus and Joseph N. Kushick. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules*, Vol. 14, No. 2, pp. 325–332, 1981.

[2] Akio Kitao, Fumio Hirata, NobuhiroGō, Nobuhiro Go. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical Physics*, Vol. 158, No. 2-3, pp. 447–472, dec 1991.

[3] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, Vol. 9, No. 2, pp. 164–169, apr 1999.

[4] 北尾彰朗. 主成分分析を使って眺めたタンパク質のエネルギー地形. 統計数理, Vol. 49, No. 1, pp. 43–56, 2001.