バイオマンでも多様体が知りたい－00
～PCAをノリで理解しないことから開始～

岡野雄士

2023/02/17

1　今回の目標

実は世の中のさまざまなアルゴリズムには数学的構造が背景に潜んでいるが、我々はそれを意識しないでも十分に生活できている。例えば、パソコンで文字を打つという操作の背景には、代数構造が潜んでおり、パソコンは入力情報を代数構造に従って文字を表示しているのである。しかし、このことを全く意識しなくとも文字を打つことはできる。それはなぜだろうか？あくまでこれは私見であるが、それは数学的な妥当性が”誰か”によって担保されているからではないだろうか。我々は誰かが証明した定理や法則を既知のものとして、地道な思索に耽ることを回避し、文明の進化を加速させてきた。

しかし、今日の生物学はどうであろうか。single cell tech. によって、以前に我々が知り得たものよりも圧倒的にリッチな情報を適切に処理することが求められている中で様々なツールが開発され、そのアルゴリズムの背景を知らなくとも簡単にこのようなデータを処理できてしまう便利な時代となっている。しかし、これらのアルゴリズムの妥当性は誰が証明してくれるのだろうか？そして何よりも、生物学的な妥当性は誰が確認するのだろうか？生物学的妥当性は決して一意的ではないであろう。線虫の腸管で示された事項がヒトの脳でも自明に事実となるであろうか？生物学は博物学的に各論を集約してきた歴史があるため、極論としては自分の興味のある対象については自分で吟味するしかないであろう。

ここまでダラダラとポエムめいたことを書いてしまったが、今回の目標は以下の通りである。

1. 生物学研究において数学の必要性を感じる
2. 線形代数・微分積分・確率統計といった分野を身近に感じる
3. それらの足がかりとして主成分分析（PCA）を数式で理解する

1.1　注意

個人的な備忘録と生物学者への説明の練習が主眼なので、無駄な部分は適宜スキップしてください。誤字脱字は適宜修正します。書き殴りなので、常体と敬体がぐちゃぐちゃで読みにくいと思いますが、そこはご愛嬌ということで...

1.2　ノーテーションについて

ここでは、今後用いるノーテーションの一部を先に紹介しておく。なるべく一般的なものを用いるようにするが、一応確認しておいてもらいたい。

$\mathbb{R}$: 実数　　$\forall$: 任意の　　$\exists$: ある（〇〇が存在する）

$^{\forall}x \in \mathbb{R}$: 実数値をとる任意の変数 x

$p(X = x)$: 確率変数 X の値が x となる確率

$^{\exists}\mathbf{x} \in \mathbb{R}^n$: ある実数 n 次元ベクトル $\mathbf{x}$

2　序論：これ、わかりますか？

2.1　あれもこれもベクトル

線形代数というと、ベクトルや行列を扱うことは有名であると思う。私も学部一年で線形代数を学んだ時に単位はもらえたため、それくらいのことは覚えていたが、正直なところ、当時は線形代数が生物学にどう生きているか全く理解していなかった。しかし、世の中は実にベクトルに溢れていると言える。そして、細胞も「見方によっては」ベクトルであると言える。ここでは、この認識のギャップを埋めつつ、「見方によっては」のニュアンスを掴んでもらいたい。

実は、ベクトルを扱う上で、それと同時にベクトルの帰属する全体集合であるベクトル空間を考えていることになるのであるが、ベクトル空間とは、以下の定義を満たすものであれば何でもベクトル空間なのである（厳密な定義はまた別の機会におこなう）。

$Def)$　ベクトル空間 V :

1.要素同士の足し算をすることができる

2.要素に対してスカラー積（例：実数倍）を行うことができる。

3.要素同士の足し算とスカラー積の順番を入れ替えられる

例 : $^{\forall}a \in \mathbb{R}, ^{\forall}x, y \in V,\ a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$

4.スカラー同士の足し算とスカラー積の順番を入れ替えられる

例 : $^{\forall}a, b \in \mathbb{R}, ^{\forall}x \in V,\ (a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$

中でも注目してもらいたいのが、「足し算ができる」という性質と、「スカラー積ができる」という性質であり、これらをまとめて「線形性」と呼ぶ。幾分か当たり前のように見えただろうか。だとすれば、それは自然科学の様々な現象がベクトルを使って表記されているから（すなわち、線形的な数学的構造に依存しているから）であろう。また、ベクトルと聞いた際に、「高校数学で扱った三角形の問題で使ったな」のような、幾何ベクトルのイメージから脱却してもらえただろうか。もちろん、高校の幾何ベクトルもベクトルの一側面であることには間違いないのであるが、幾何ベクトルだけがベクトルなのではなく、線形性があるもののことをベクトルというのである。

さて、先ほど細胞も「見方によっては」ベクトルと言ったが、これはどういうことだろうか。例として、single cell RNA sequencing（scRNA-seq）のデータを考える。例えば、ある細胞についての D 個の遺伝子群の転写量が $\log_2(TPM+1)$ で与えられているとしよう（$^\forall\mathbf{x} \in \mathbb{R}^D$ とする）。また、別の細胞に対応するデータ $^\forall\mathbf{y} \in \mathbb{R}^D$ を考えると、$\mathbf{x}+\mathbf{y}$ を考えることは何ら不自然ではないであろう。また、何らかの実数 $^\forall\alpha \in \mathbb{R}$ によってスカラー倍して、$\alpha\mathbf{x}$ を考えることも全く自然なことであろう。実際、全遺伝子について発現量の平均値を計算することに疑問を持たないのであれば、足し算やスカラー積の操作についても前提として許していることになる。そして、このように線形性をもつため、scRNA-seq データはベクトルであり、scRNA-seq データはベクトル空間に存在することになる。

しかし、これは数学上の話であって、生物学的な観点からの妥当性は如何であろうか。すなわち、細胞をベクトルとして表現するモデルは生物学的に妥当なのか？ということである。そのためには、生物学的な意味と数学的な意味を突き合わせて考える必要がある。しかし、生物学と数学には大きな違いがあり、「生物学では起こりそうなことの性質を議論する」一方で、数学では「どんなに極端な状況でも成立する性質を議論する」ことである。具体的に言うのであれば、細胞の議論をベクトル空間での議論に置き換えることで、対応する細胞が存在し得ないようなベクトルについても許してしまうことになる。生物学的に言えば、細胞と細胞の遺伝子発現量を足した $\mathbf{x}+\mathbf{y}$ のような状況は細胞の遺伝子発現量といえるのだろうか。また、$1234.456\mathbf{x}$ のようにスカラー倍した状況は細胞の遺伝子発現量と言えるのであろうか。この生物学的な妥当性は「見方によっては」正しいし、逆も然りであろう。

生物学で使われている数学を数式や定義といった、数学で用いられている表記で書き表すことは、その理論の前提や帰結がもつ性質を吟味することにつながるため、我々生物学者はそこに生物学的解釈が伴うか注視する余地が生まれる。一方で、数学的理論の背景事項に対する厳密性が失われると、生物学的観点との照会が困難となる。だからこそ、数学は数学のまま語るべきであり、生物学者は数学を用いる前のモデル化のプロセスにも目を光らせるべきではないだろうか。

2.2 PCA って何ですか？

あなたは、「PCA って何ですか？」と聞かれた時に一行程度の文章で説明できるであろうか。実際の数式を見る前の自分は、「線形的な次元削減」と認識していたが、これは本質的ではない。なぜならば、次元削減と対応する部分は、累積寄与率を減らす人為的な操作に付随するものであって、原理上、累積寄与率は 100%まで計算できるし、実際にお使いのライブラリでもそのように計算しているはずである。

また、「分散が最大になるように第一主成分を取ってきて、それと直交するなかで分散が最大になる第二主成分をとって...」のような説明法もよくされていると思うが、その説明から数式を書き起こせる人がどのくらいいるだろうか。私も最初に PCA の概念に触れた時にはそのように教わったが、何となくイメージできても数式を想起できなかった。前述のように、数式に書き起こすことは、前提や帰結を注視することに役立つので、なるだけ数式がイメージしやすいような説明ができるのに越したことはない。

私が用意した回答は、「分散共分散行列の固有値分解と固有ベクトルによる基底変換」である。数式がイメージしやすく、短く纏まっているであろう。実はこのドキュメントのオチは、PCA が分散共分散行列の固有値分解であることを説明することなので、既知の場合は以下は不要となる。この説明と既出の説明との関連性において少しでも疑問点があれば、このドキュメントを通じて、次元の話や、軸の直交性の話、固有値分解の話などはカバーしていきたいと思うので、お付き合いいただきたい。

2.3 一次独立と軸の直交性は全くの別物

一次独立も直交な軸も、お互いに影響がなさそうなイメージがして、同じような議論に思えてしまうであろうか。しかし、これらは全く違う話をしている。RNA-seq データそのものや、PCA を用いたデータ解析結果の議論において、一次独立性や直交性などのタームは非常に重要であるため、正確に理解していることが非常に重要である。結論から言えば、基底同士は定義より一次独立であるが、それが必ずしも直交している必要はない。斜交座標系が良い例である。

これらの違いがわかりやすい他の例としては、3 次元内積空間も挙げられる。3 次元内積空間とは、その名の通り、内積が定義された 3 次元ベクトル空間のことである。直交性は 3 次元に限らず、内積空間の任意の 2 つのベクトル $^{\forall}\mathbf{x}, \mathbf{y}$ について、$\mathbf{x} \cdot \mathbf{y} = 0$（但し、$\mathbf{x} \cdot \mathbf{y}$ は $\mathbf{x}$ と $\mathbf{y}$ の内積）となることで定義される。一方で、一次独立性に関して 3 次元ベクトル空間では、$\mathbf{x} \times \mathbf{y} \neq 0$（但し、$\mathbf{x} \times \mathbf{y}$ は $\mathbf{x}$ と $\mathbf{y}$ の外積）と同値条件であることが知られている。このように、一次独立性と軸の直交性は全く別の話をしているのである。

3　PCA に必要な線形代数

3.1　一次独立性・基底・次元

3.2　内積と直交性

3.3　基底変換行列

3.3.1　導入：「基底による表記」と「座標系による表記」

例えば、$\mathbb{R}^2$ に関して、基底をなす（お互いに一次独立な）$^\forall\mathbf{x},\mathbf{y}$ を考える。このとき、$^\forall\alpha,\beta\in\mathbb{R}$ に関して、$\alpha\mathbf{x}+\beta\mathbf{y}=:\mathbf{z}$ となるようなベクトル $\mathbf{z}$ に関して、今後以下の 2 通りの表記を行うこととする。特に、2 つ目の「座標系による表記については、基底を対応する係数が書かれる括弧内での位置によって表しているため、基底の表記を拝借して xy 座標系と呼ぶことにする。

1. 基底による表記：$\alpha\mathbf{x}+\beta\mathbf{y}$
2. 座標系による表記：$\begin{pmatrix}\alpha & \beta\end{pmatrix}$ または (α,β)

座標系（ないしは成分）での表示の際には、横ベクトルで書くことで、縦ベクトルで表記された基底のベクトルとの行列積で、$\begin{pmatrix}\alpha & \beta\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}=\alpha\mathbf{x}+\beta\mathbf{y}$ のように基底による表記との整合性を取れるようにする。また、データセットのように複数の点を行列として表す際には、縦方向に伸ばすことでデータ数を表すことにする。そのため、データ数が n で d 次元のデータは、$n\times d$ 行列として表現することにする。

3.3.2　正方行列と基底変換

正方行列とは、$n\times n$ 行列のように、行数と列数が同じ行列のことを指す。正方行列はベクトルとの行列積によって、基底の変換を意味するという性質がある。

例えば、$\mathbb{R}^2$ に関して、基底をなす（お互いに一次独立な）$^\forall\mathbf{x},\mathbf{y}$ をそれぞれ以下のように変換して、新たに基底をなす $\mathbf{u},\mathbf{v}$ に対して座標系を変換することを考える。

$$\mathbf{u}:=2\mathbf{x}$$
$$\mathbf{v}:=\frac{\mathbf{y}}{3}$$

このとき、行列 $A:=\left(\begin{smallmatrix}2 & 0\\ 0 & \frac{1}{3}\end{smallmatrix}\right)$ と定義すると、これは以下のように基底を成分にもつベクトルとの行列積によって、基底を変換できていることがわかる。

$$\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}=A\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$$

次に、行列の成分表示によって、xy 座標系が uv 座標系に移される過程を考える。例として、xy 座標が (α, β) のベクトルの uv 座標系での座標系 (α^*, β^*) として、これらの値を考えると、$\mathbf{u} = \frac{1}{2}\mathbf{x}$ かつ $\mathbf{v} = 3\mathbf{y}$ なので、$\alpha\mathbf{x}+\beta\mathbf{y} = \frac{\alpha}{2}\mathbf{u}+3\beta\mathbf{v}$ となるため、uv 座標は $(\alpha^*, \beta^*) = (\frac{\alpha}{2}, 3\beta)$ となる。そのため、A の逆行列である、$A^{-1} = \left(\begin{smallmatrix} \frac{1}{2} & 0 \\ 0 & 3 \end{smallmatrix}\right)$ を用いると、以下のように書き表すことができる。

$$\begin{pmatrix} \alpha^* & \beta^* \end{pmatrix} = \begin{pmatrix} \alpha & \beta \end{pmatrix} A^{-1}$$

基底そのものと座標の値は反変的であるため、多少分かりにくい点もあるが、A^{-1} を xy 座標系から uv 座標系への変換行列と捉えると、成分表示した場合でも、正方行列との行列積によって基底が変換されるということが幾分かわかりやすくなるであろう。

3.4 行列式

行列式とは、正方行列に対して定義される指標である。一般には、行数及び列数を置換（並び替え）したときの符号と対応する成分の相乗の総和であるが、ここではより実用的な計算方法のみ扱う。

3.4.1 2 次の正方行列の場合

2 次の正方行列（$^\forall A := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$）に対して行列式 $detA$ は以下のように定義される

$$detA = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

3.4.2 余因子展開

3.4.3 サラスの公式

特に、4×4 正方行列以上の行列式を求める際に、3×3 正方行列の余因子展開の結果を公式として用いると計算が楽になる。

$$\begin{aligned} detA = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\quad - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} \end{aligned}$$

図示して覚える方法が有名であるが、個人的には、行数を固定して考えて、$a_{11}a_{22}a_{33}$ から $a_{1i}a_{2j}a_{3k}$ を作るためには、何回列数をスワップすればいいかを考え、一回のスワップで $a_{1i}a_{2j}a_{3k}$ となる場合はマイナス、二回スワップが必要な場合はプラスという覚え方が間違えにくいと思う。実際に、奇数回

のスワップを奇置換といい負の符号を割り当て、偶数回のスワップを偶置換といい正の符号を割り当てるのが行列式の定義で行なっていることである。

3.4.4 例題

末尾に解答解説を用意したので適宜確認されたい。

1. $\begin{vmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{vmatrix}$

2. $\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{vmatrix}$

3. $\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 3 & 5 & 4 \end{vmatrix}$

4. $\begin{vmatrix} 1 & 1 & 3 \\ 1 & 2 & 5 \\ 2 & 0 & 2 \end{vmatrix}$

3.4.5 基底変換行列の行列式

基底変換行列の行列式は、変換による基底の向きの変化や、変換による基底の伸縮など、幾何学的に重要な様々な情報の尺度である。また、行列式が 0 となる変換行列は、逆変換に対応する逆行列がないため、変換行列による座標系の変換は点と点の 1 対 1 対応関係がないことを意味している。

3.5 固有値分解

3.5.1 固有値・固有ベクトルの幾何的な意味

固有値・固有ベクトルとは基底変換に紐づいた概念である。任意の変換行列 A があり、その変換行列で零ベクトルでない、ある特殊なベクトル $\mathbf{v}$ を線形変換すると、変換前のベクトル $\mathbf{v}$ と変換後のベクトル $A\mathbf{v}$ に関してちょうどある実数 λ によってスカラー倍したとき等しくなるような、特殊な関係（$A\mathbf{v} = \lambda\mathbf{v}$）が成立しているとする。このとき、$A$ による変換は $\mathbf{v}$ の方向に関して、λ 倍拡大させる変換と捉えることができる。

この $\mathbf{v}$ のような性質をもつベクトルを固有ベクトルといい、$\mathbf{v}$ に対する λ のようなスカラーのことを固有値という。そのため、固有値は固有ベクトル方向への”空間の拡大率”のように捉えることができる。

3.5.2 固有値分解の計算方法

3.5.3 例題

末尾に解答解説を用意したので適宜確認されたい。

4 PCA に必要な確率統計

分散共分散行列の定義までさらっと触れられれば十分なので、確率論などの込み入った話は置いておいて、ごくごく簡単に紹介する。

4.1 一変量の分散

一変数の確率変数の分散 $var(X)$ は X の期待値 $E[X]$ を用いて以下のように定義される。

$$var(X) := E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

また一般に、確率変数 X について、標本平均を $\bar{X}$ とすると、$E[\bar{X}] = E[X]$ となるため、標本平均は確率変数の期待値の推定量として用いられる。一方で、標本分散を S^2 とすると、$S^2 := \overline{(X - \bar{X})^2} = \overline{X^2} - (\bar{X})^2$ であるが、標本数が $^\forall k \in \mathbb{N}$ のとき、$E[S^2] = \frac{k-1}{k} var(X)$ となる。そのため、不偏分散 $U^2 := \frac{k}{k-1} S^2$ が確率変数の分散の推定量として用いられる。

4.2 多変量の分散共分散行列

多変量の分散共分散行列の例として、$^\forall n \in \mathbb{N}$ について、$X_1, X_2, \cdots, X_n$ の n 変量に関する分散共分散行列 Σ は以下のように定義される

$$\Sigma := \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

このとき、$^\forall i, j \in \{1, 2, \cdots, n\}$ に関する σ_{ij} は以下のように定義される。

$$\sigma_{ij} := E[(X_i - E[X_i])(X_j - E[X_j])]$$

この定義は、各成分が分散または共分散に一致するため、Σ は分散共分散行列とよばれる。また、$\mathbf{X} := (X_1, X_2, \cdots, X_n)^\top$ のようなベクトルを用いれば、Σ は以下のように書き表すことができ、一変量の分散を自然に多変量に拡張したことがわかる。

$$\Sigma = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^\top]$$

5　PCA について

$^{\forall}n, d \in \mathbb{N}, n \geq d$ のとき、$n \times d$ 行列であるデータセットに対応する、確率変数 $X_1, X_2, \cdots X_d$ の分散共分散行列で $d \times d$ 行列の Σ を考える。このとき、データから真の $E[X_i]$ ($^{\forall}i \in \{1, 2, \cdots, d\}$) を求めることができないため、代わりに推定量である $\overline{X_i}$ を用いることになる。このとき、標本分散的に Σ を計算しても、不偏分散のように計算しても、結局 $\frac{n-1}{n}$ 倍のスカラー倍がかかるかかからないかの問題なので、PCA をする上では理論上影響しない。

5.1　主成分スコア・寄与率

5.2　軸の直交性について

固有値分解の重要な性質として、任意の実対称行列 $^{\forall}A \in \mathbb{R}^d$ の固有ベクトルは正規直交基底となることが挙げられる。さらに言えば、固有ベクトルを並べた行列を P とすると、P は必ず直交行列となる。

今回、固有値分解の対象となった Σ は定義より、実対称行列であるため、固有ベクトルは

5.3　固有ベクトルによる基底変換

5.4　補足 1：なぜ最大の分散をとる操作が固有値分解になるのか？

ラグランジュの未定定数法

5.5　補足 2：応用

5.5.1　$n \times d$ vs $d \times n$

5.5.2　データリスケーリング

5.5.3　$n < d$ の場合

6　例題の解答・解説

6.1　行列式

1. $\begin{vmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{vmatrix} = cos^2\theta + sin^2\theta = 1$

2. $\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{vmatrix} = 1\cdot2\cdot1+1\cdot1\cdot2+2\cdot1\cdot1-1\cdot1\cdot1-2\cdot2\cdot2-1\cdot1\cdot1=-4$

3. $\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 3 & 5 & 4 \end{vmatrix} = 1\cdot2\cdot4+1\cdot1\cdot3+2\cdot1\cdot5-1\cdot1\cdot5-2\cdot2\cdot3-1\cdot1\cdot4=0$

4. $\begin{vmatrix} 1 & 1 & 3 \\ 1 & 2 & 5 \\ 2 & 0 & 2 \end{vmatrix} = 1\cdot2\cdot2+1\cdot5\cdot2+3\cdot1\cdot0-1\cdot5\cdot0-3\cdot2\cdot2-1\cdot1\cdot2=0$

6.1.1 補足：一次従属性と行列式

3. と 4. の行列式は 0 となったが、3. は行基本変形、4. は列基本変形をするとそれぞれ 3 行目と 3 列目の成分を全て 0 にすることが可能なことにお気づきだろうか。このような行列については、それぞれ行または列が一次従属な関係になっているのである。3. を例にとって、以下のように各行を何らかのベクトルとして表現するように書き換えてみると一次従属であることがわかりやすい。

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 3 & 5 & 4 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{x}+\mathbf{y}+2\mathbf{z} \\ \mathbf{x}+2\mathbf{y}+\mathbf{z} \\ 3\mathbf{x}+5\mathbf{y}+4\mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{a}+2\mathbf{b} \end{pmatrix}$$

この場合、仮に $\mathbf{x}, \mathbf{y}, \mathbf{z}$ が基底であっても、行列式が 0 となる変換行列をかけると、基底の数が減ってしまうのである。実際に $\mathbf{a}, \mathbf{b}$ は基底であっても $\mathbf{c}$ はこれらの一次結合で表せるため、基底ではない。

6.2 固有値分解