統計学の基礎理論③
多次元データと次元縮約

東京科学大学　ILA国府台
中林　潤

JH人材育成課　バイオインフォマティシャン育成講座　④

内容

- データの次元
- 主成分分析（Principle Component Analysis PCA）
- 非線形圧縮

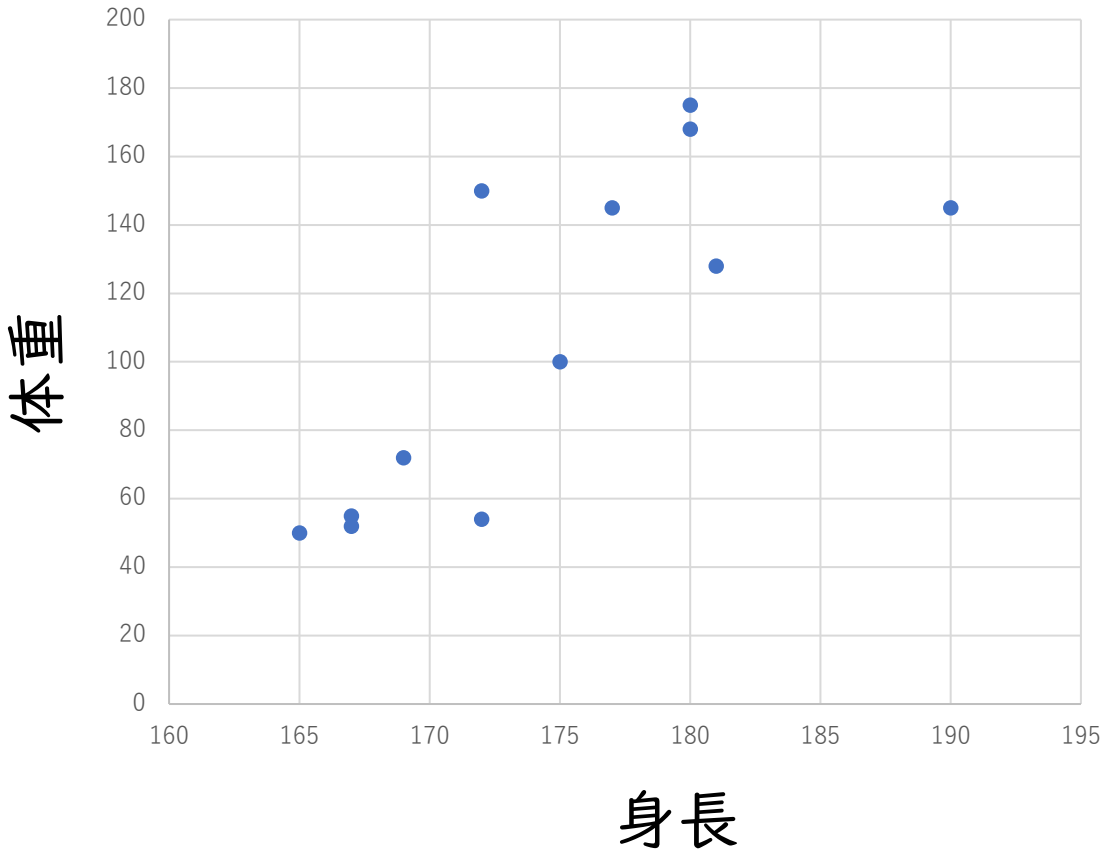
データの次元

- データに含まれる特徴量の数

サンプル	身長	体重	胸囲
①	172	54	87
②	167	55	90
③	165	50	91
④	167	52	92
⑤	169	72	88

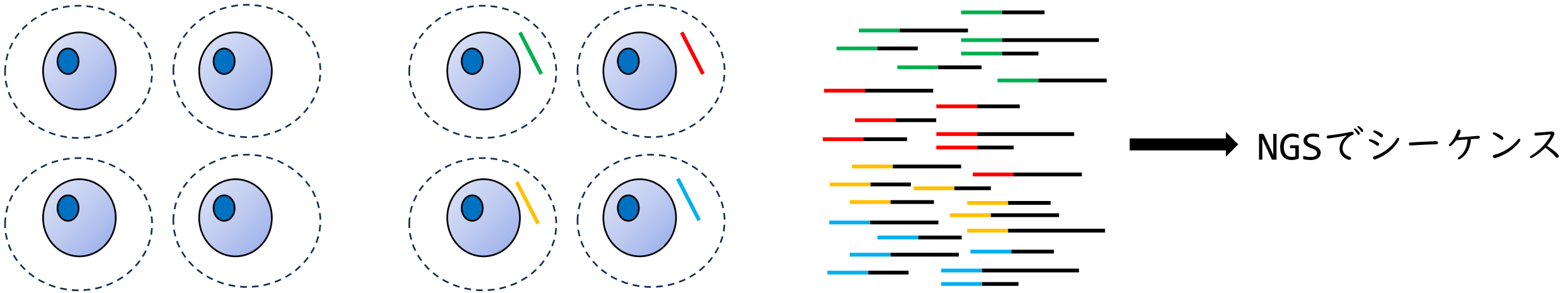
身長、体重→2次元
身長、体重、胸囲→3次元

散布図



データの多次元化

- scRNA-seq
  個々の細胞に発現している全遺伝子の発現量を測定する。



細胞を油滴で分離　バーコード配列を添加　RNAを抽出しライブラリを作成

scRNA-seqの例

- Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.
  Zeisel A. et al. ***Science*** 2015 Mar 6;347(6226):1138-1142

- マウス脳細胞でscRNA-seqを行った論文

- 3005細胞，19946遺伝子の発現量を測定→19946次元のデータ

次元縮約

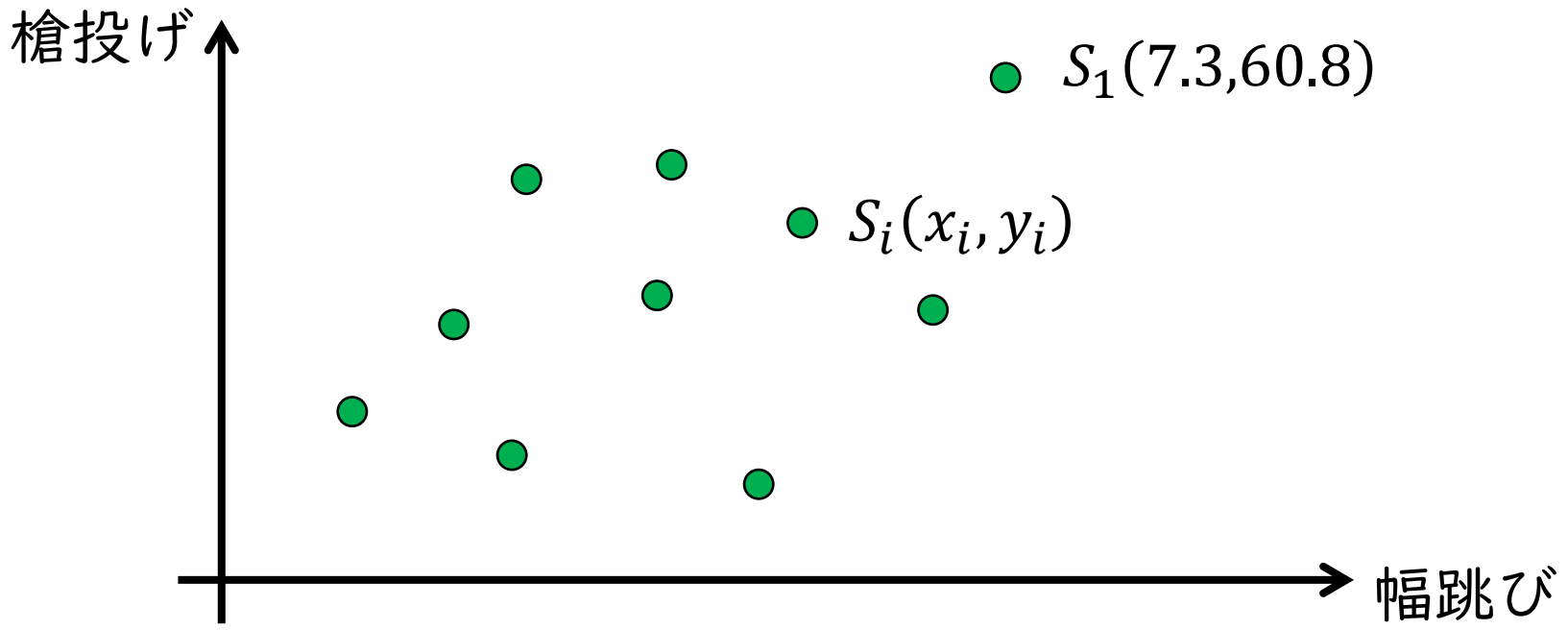
- 高次元データを低次元に変換する。
- 元のデータの持つ情報をなるだけ減らさないように次元を減らす。
- データの圧縮や可視化のための行う。
- データを要約し解釈するのに役立つ。

Principle Component Analysis PCA

- 多変量解析法の一つ
- 線形圧縮
- 多数の変数から、データ全体の値のばらつきを最もよく表す少数の変数を合成する。

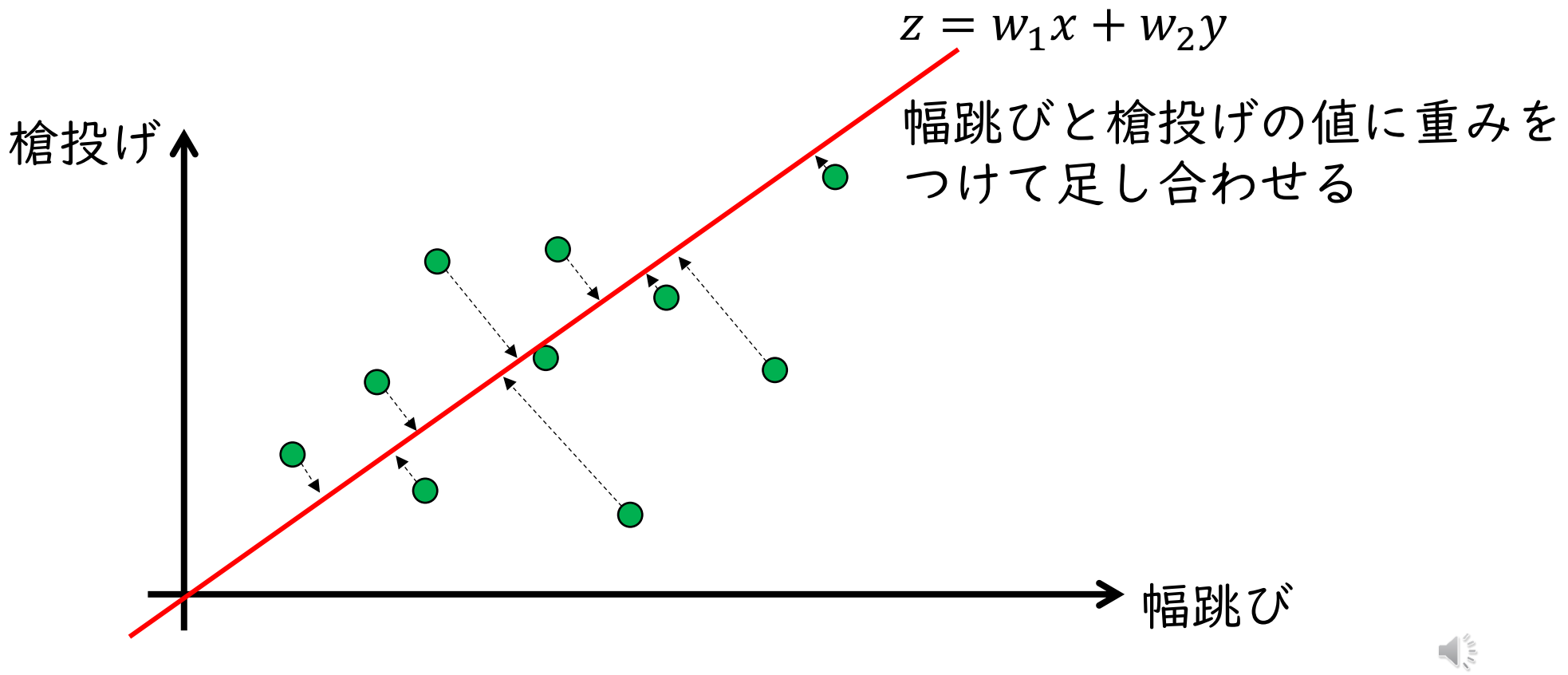
2次元→1次元の例

体力測定のデータ
走り幅跳び、槍投げの2次元データ



2変数からデータ全体の値のばらつきを良く表す1変数を合成する

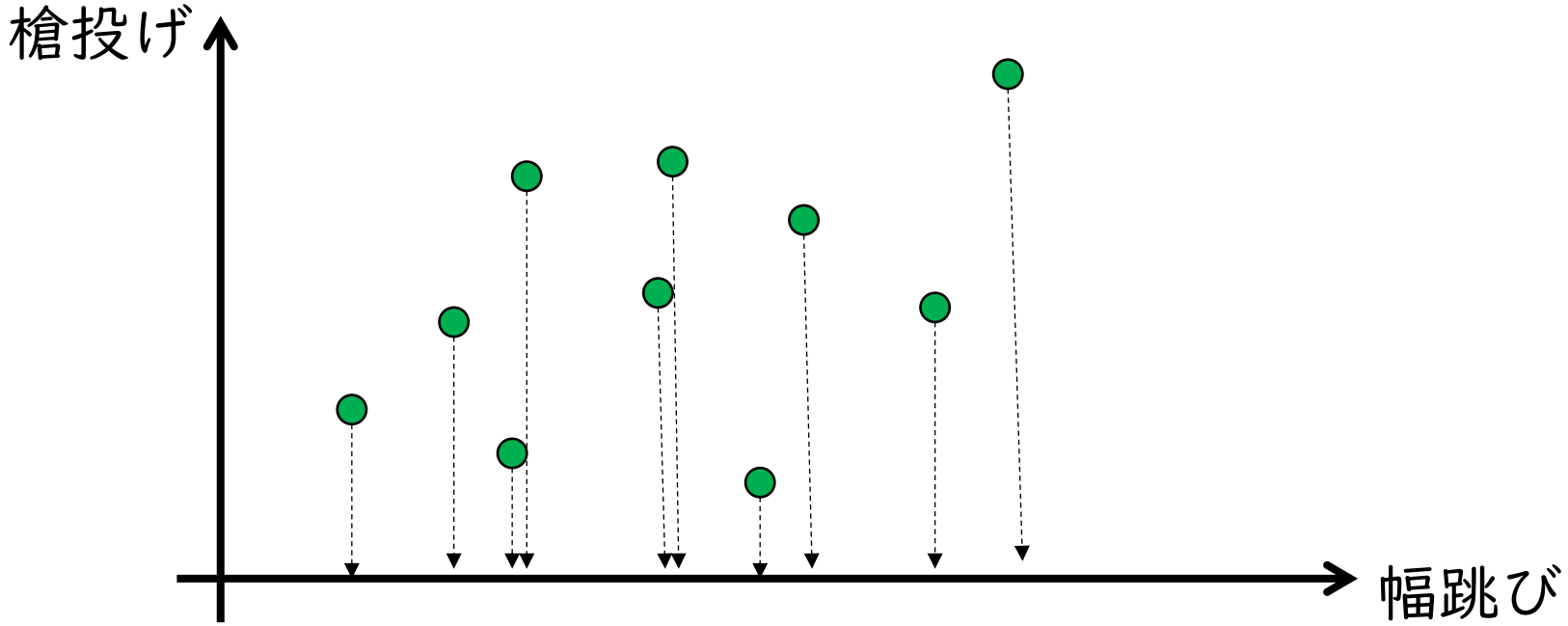
2次元→1次元の例



重みづけの値をどう決めるか？

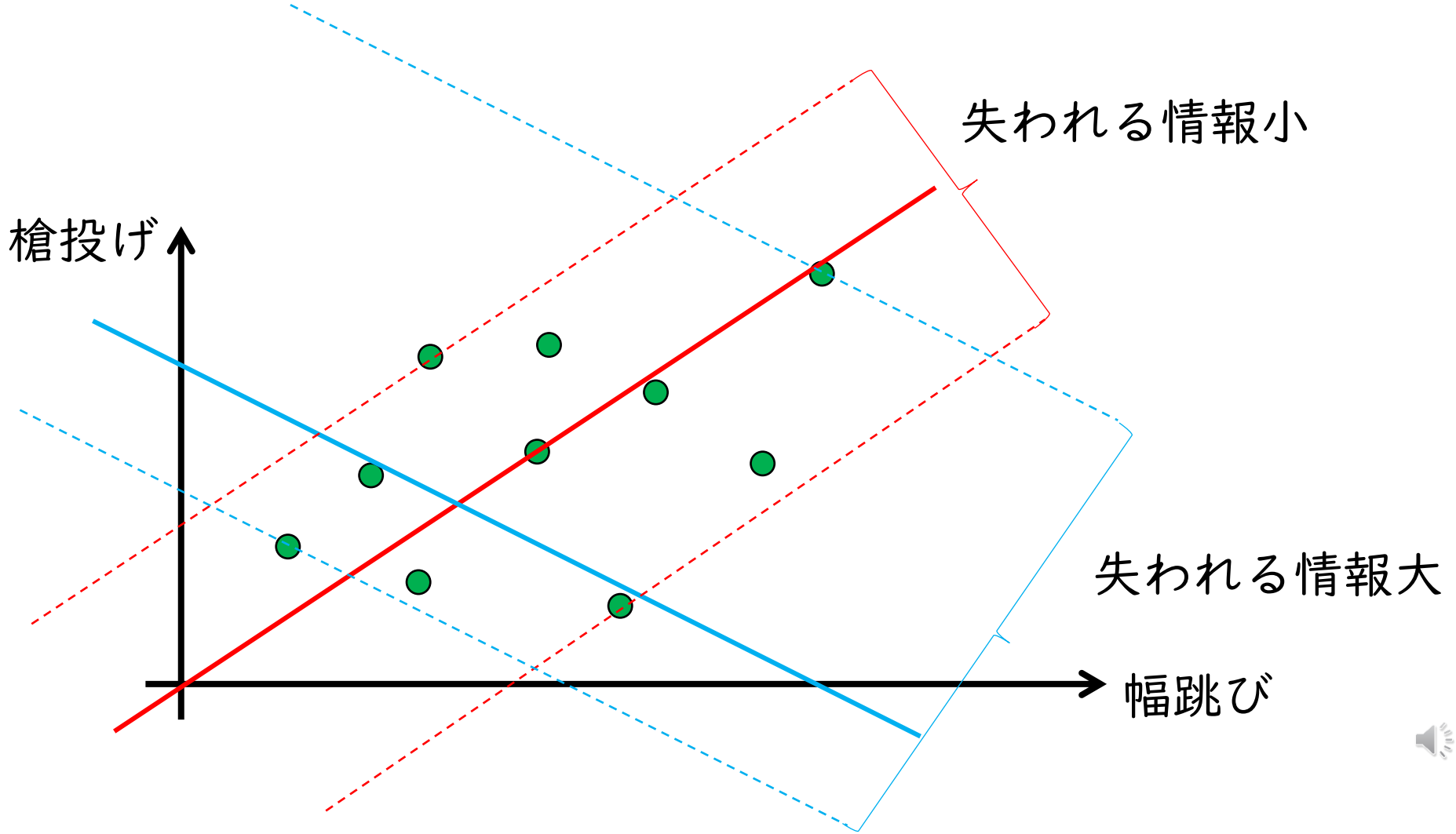
2次元→1次元の例

$z = w_1 x + w_2 y$
$w_2 = 0$の場合



槍投げの情報は失われてしまう。

2次元→1次元の例



データのばらつきに沿う軸を求める

分散共分散行列

- 分散と共分散を行列の形式で表したもの
- 分散

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

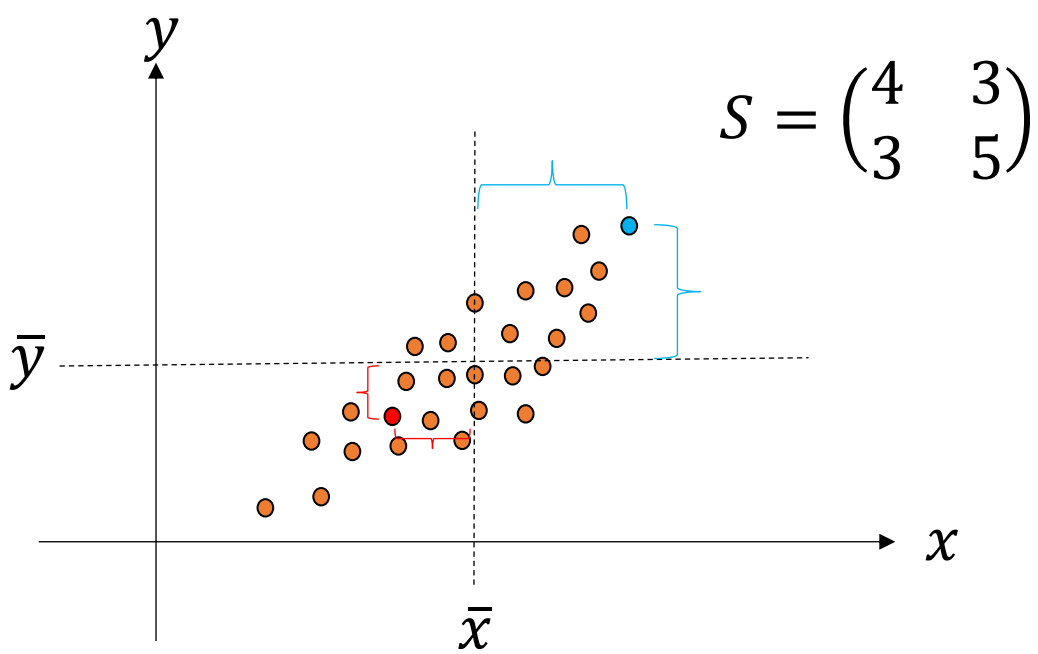
- 共分散

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

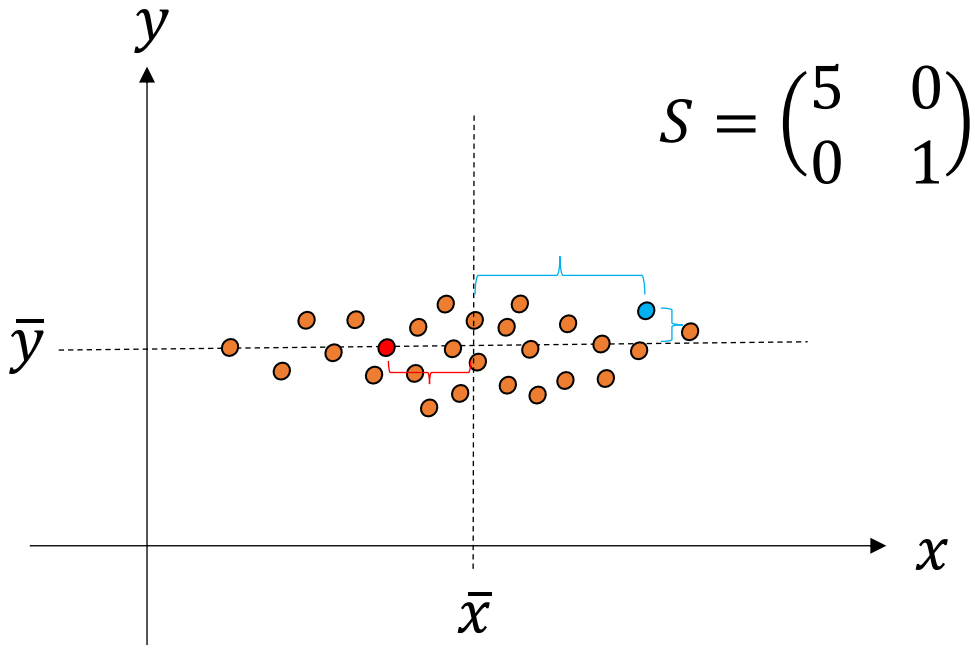
- 分散共分散行列

$$S = \begin{pmatrix} \sigma_{11}\sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$$

分散共分散行列



$$S = \begin{pmatrix} 4 & 3 \\ 3 & 5 \end{pmatrix}$$

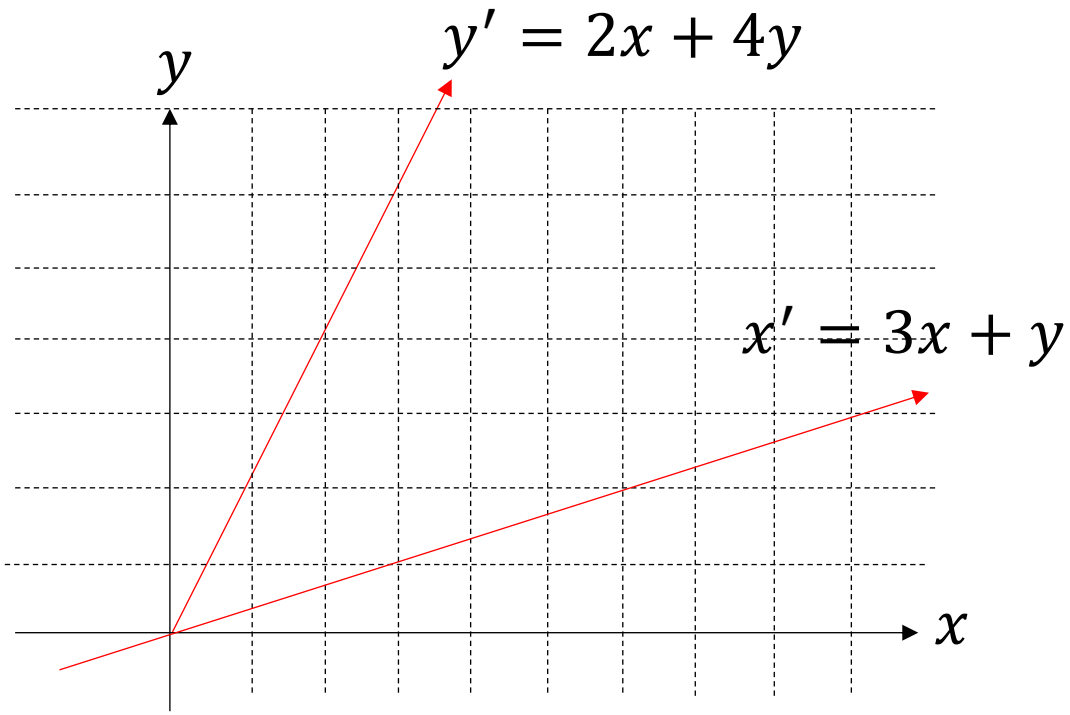


$$S = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

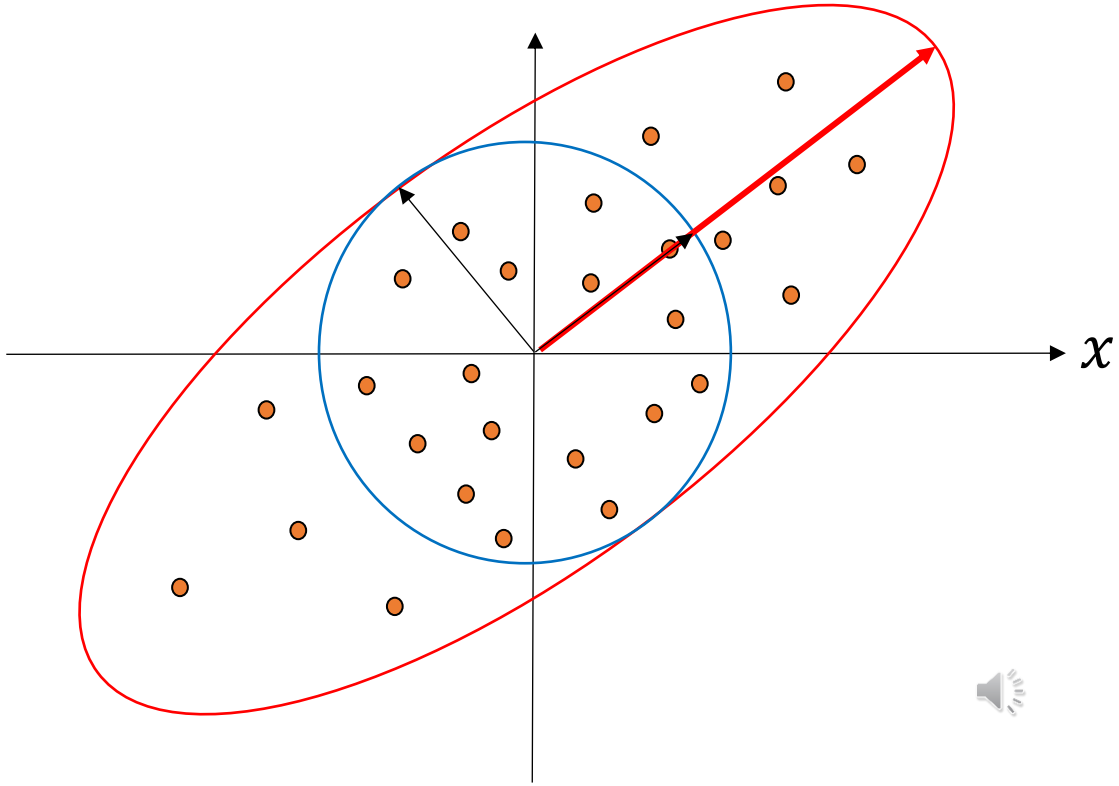
分散共分散行列と写像

行列は写像を表す

$$\begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{matrix} 3x + y \\ 2x + 4y \end{matrix}$$



$$S = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}$$

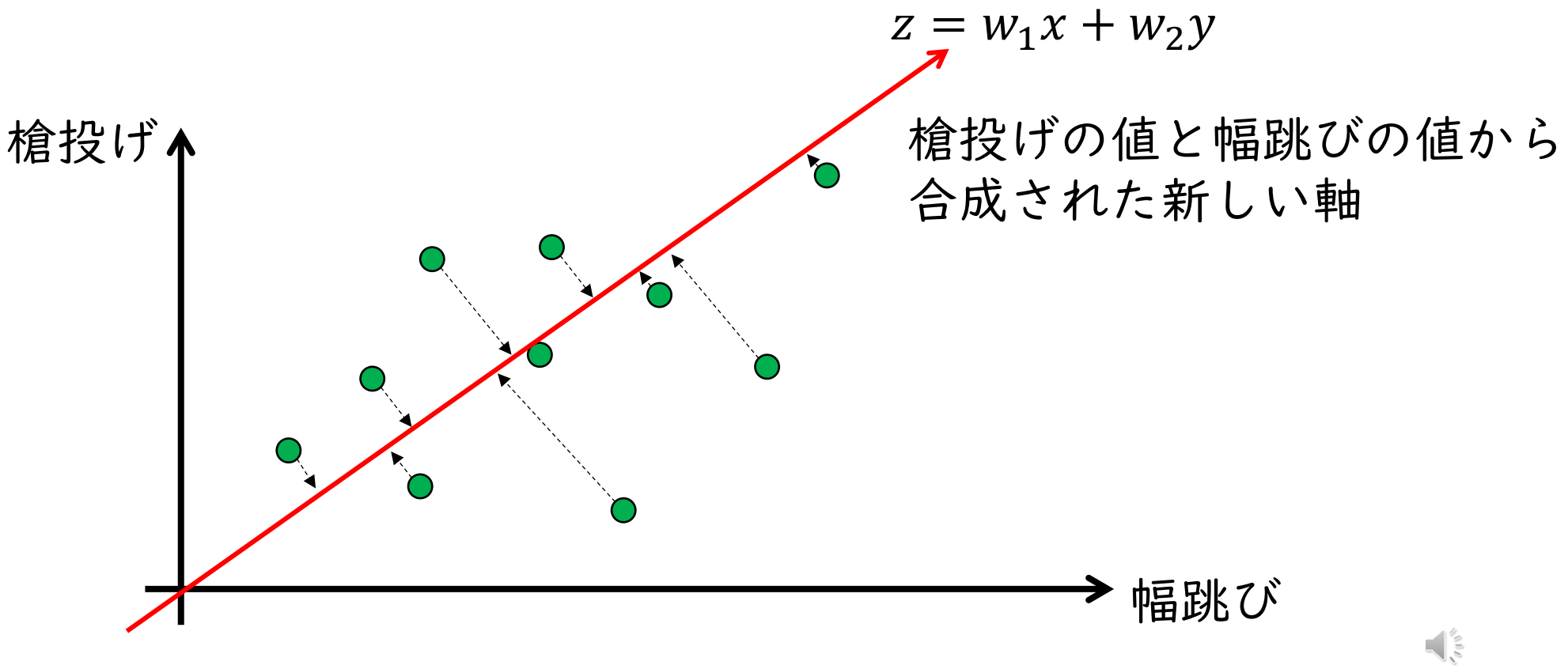


固有ベクトル

- n次正方行列Aで線形変換を行ったとき、向きが変わらないベクトル
  $A\vec{x} = \lambda\vec{x}$
  λ：固有値

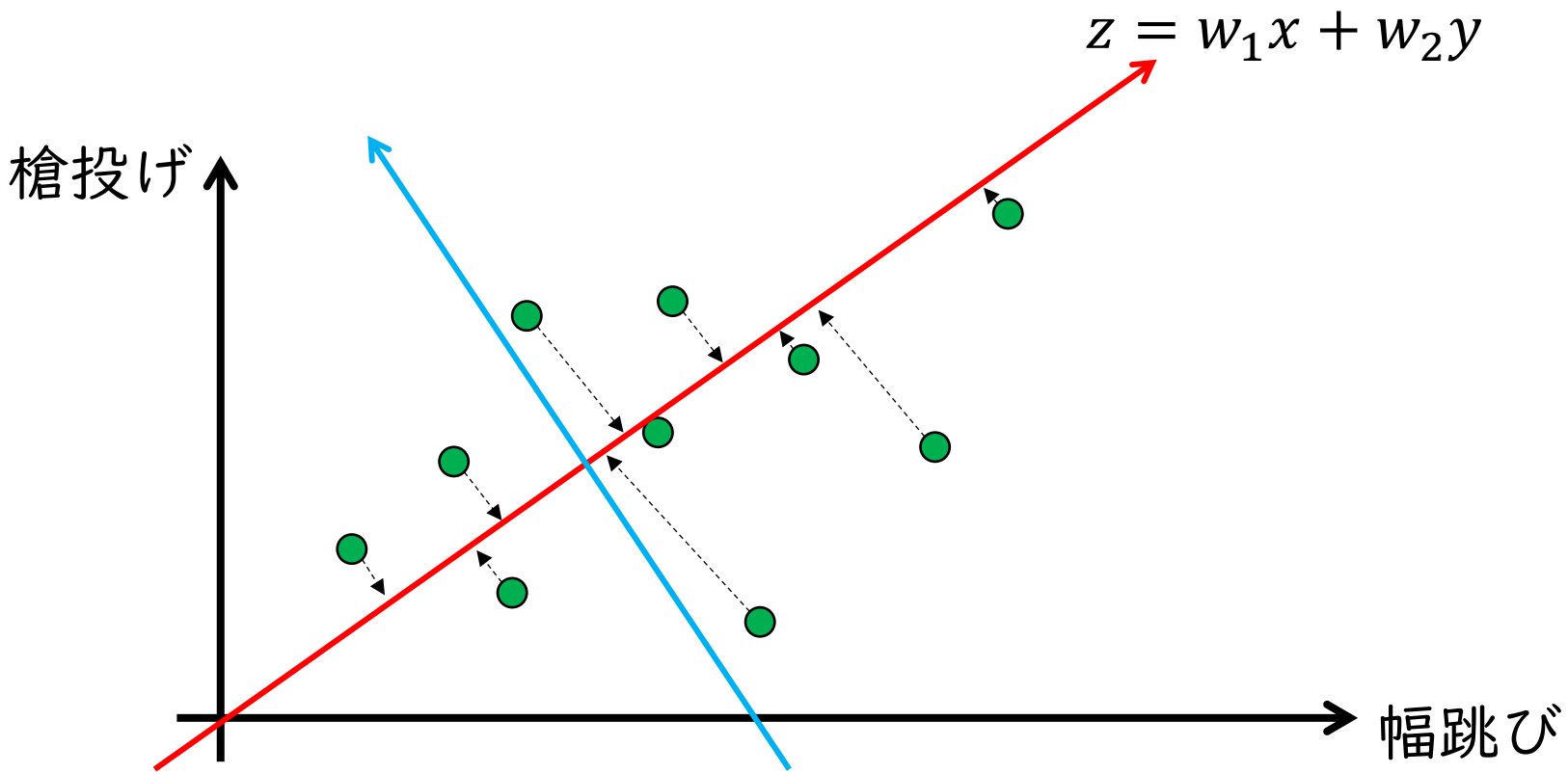
- 分散共分散行列の最大固有値を求め、その固有ベクトルを第1主成分とする。

第1主成分



分散共分散行列の固有ベクトルを第1主成分とする。

第2主成分



第1主成分に直交する成分を第2主成分とする。

主成分分析

- 分散共分散行列の固有ベクトルを第1主成分とする。
- 第1主成分と直交する成分を第2主成分とし、第n主成分まで軸を取る。
- 任意のn次元までデータの次元を縮約する。
- 主成分は元の軸から合成された新しい軸。
- データのばらつきをなるだけ損なわない縮約。

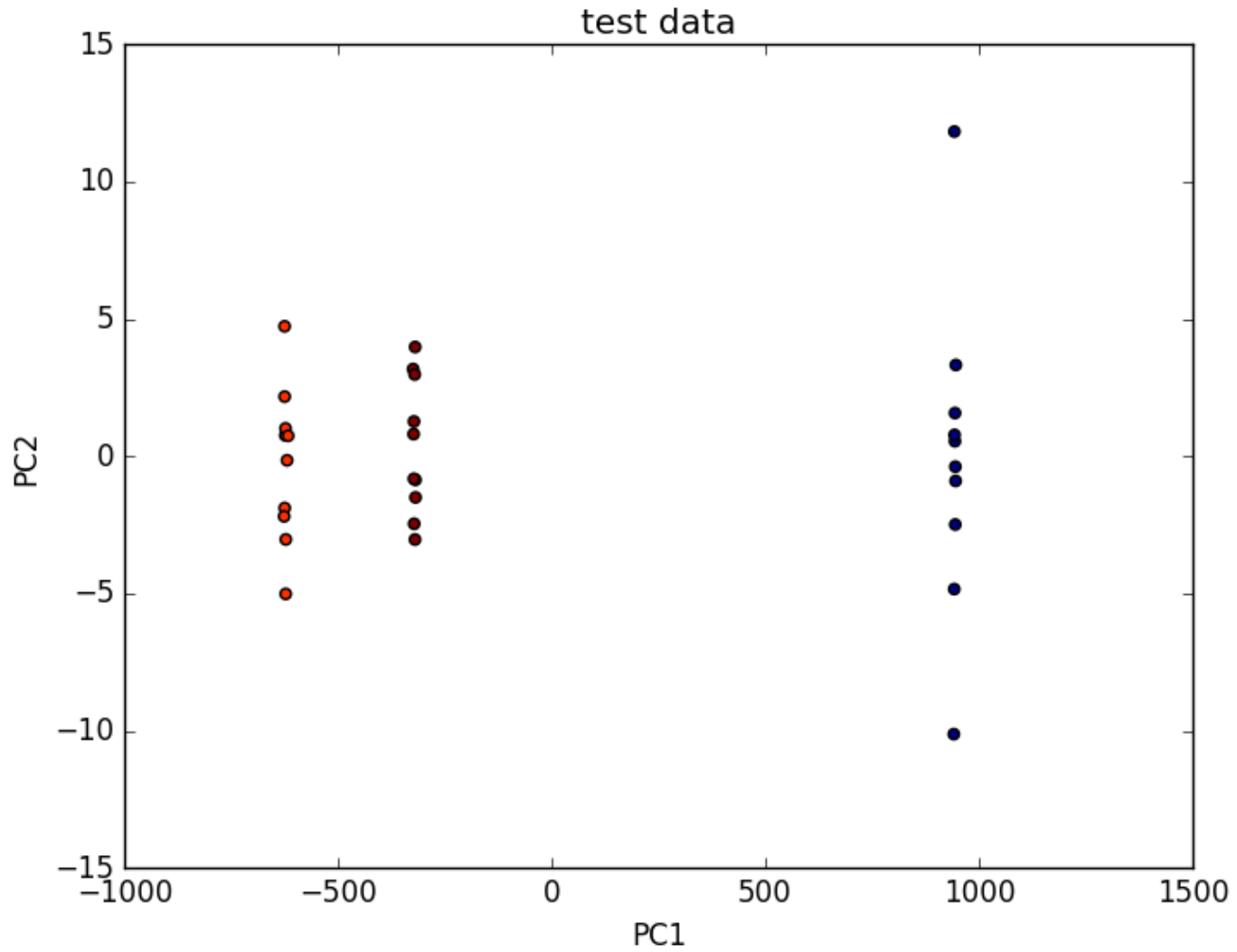
主成分分析の実例

- 模擬データ
  10次元　30サンプル
  サンプルA～J：平均10の乱数
  サンプルK～Y：平均100の乱数
  サンプルZ～AI：平均500の乱数

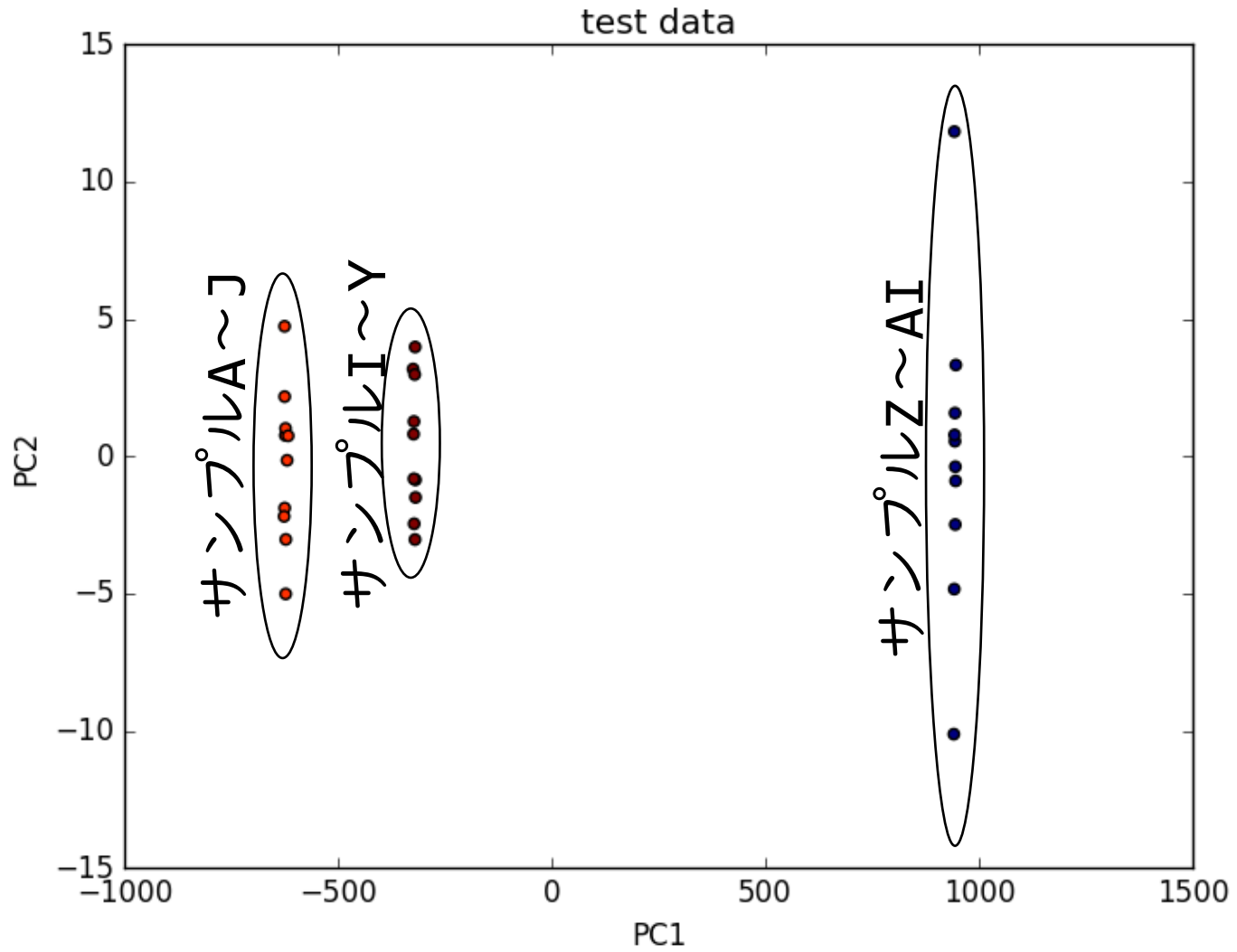
	サンプルA	サンプルB	サンプルC	サンプルD	サンプルE	サンプルF	サンプルG	サンプルH	サンプルI	サンプルJ
特徴量1	4.595383	4.883367	3.010195	4.000774	1.138695	6.153429	7.408813	3.78352	4.359661	12.90997
特徴量2	4.871761	3.651973	1.393781	4.181326	2.039158	7.926498	7.023114	9.606039	5.608627	3.758899
特徴量3	3.533661	1.904725	3.663627	2.980091	0.969703	7.50953	5.001108	5.659776	2.579873	7.84611
特徴量4	2.290189	3.179252	6.227462	3.680637	2.702148	1.131448	5.147689	7.889077	6.745012	5.826612
特徴量5	8.550237	2.949927	6.781291	7.529187	3.130017	5.635492	4.799112	5.253043	5.674672	5.275087
特徴量6	4.205498	5.879531	3.080952	5.610137	12.08058	2.549865	7.461188	8.151719	2.433683	5.844444
特徴量7	5.960999	4.105216	2.320368	5.981021	4.574415	7.100958	3.003179	6.920395	4.644592	8.720155
特徴量8	8.258824	6.022997	6.120873	3.718331	5.738754	4.95501	5.034265	3.195478	2.38449	6.755224
特徴量9	2.979031	4.590665	6.401722	0.367454	5.204553	3.431815	3.613974	3.57669	1.644468	5.703054
特徴量10	6.72884	6.179765	9.021657	5.438645	5.328348	4.620061	1.287124	7.304506	3.194148	6.512263

- PCAで2次元に縮約

主成分分析の実例



主成分分析の実例

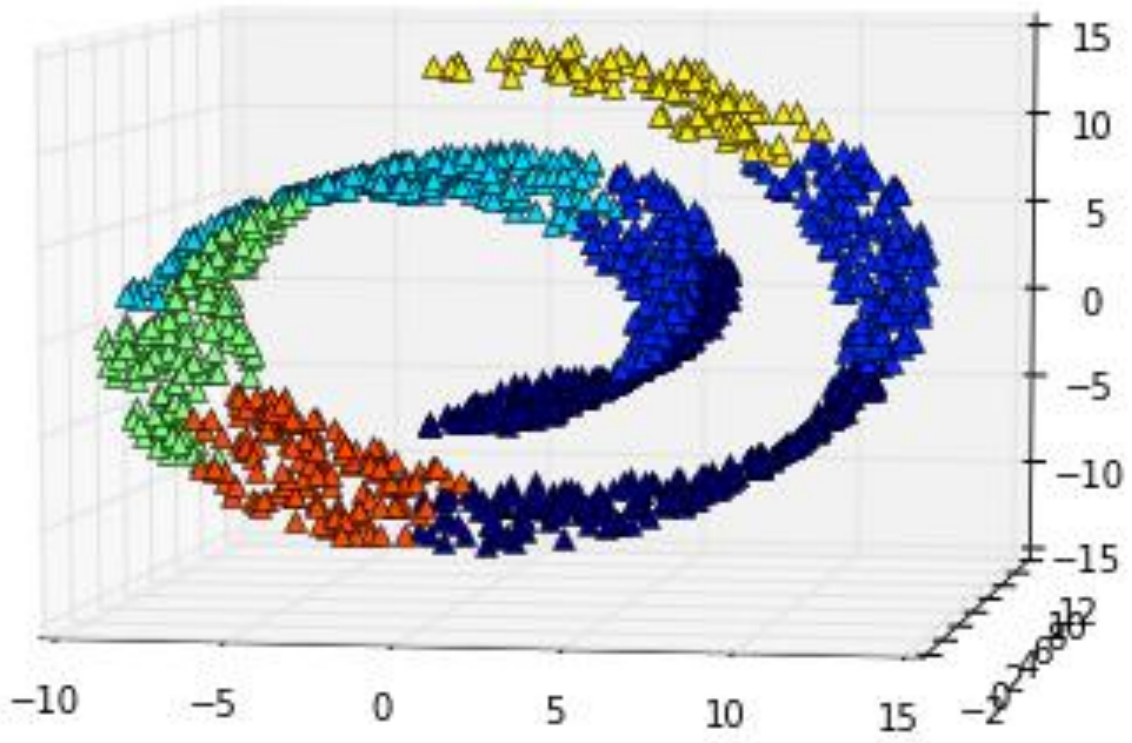


元のデータの情報が保たれている。

主成分分析の問題点

スイスロール

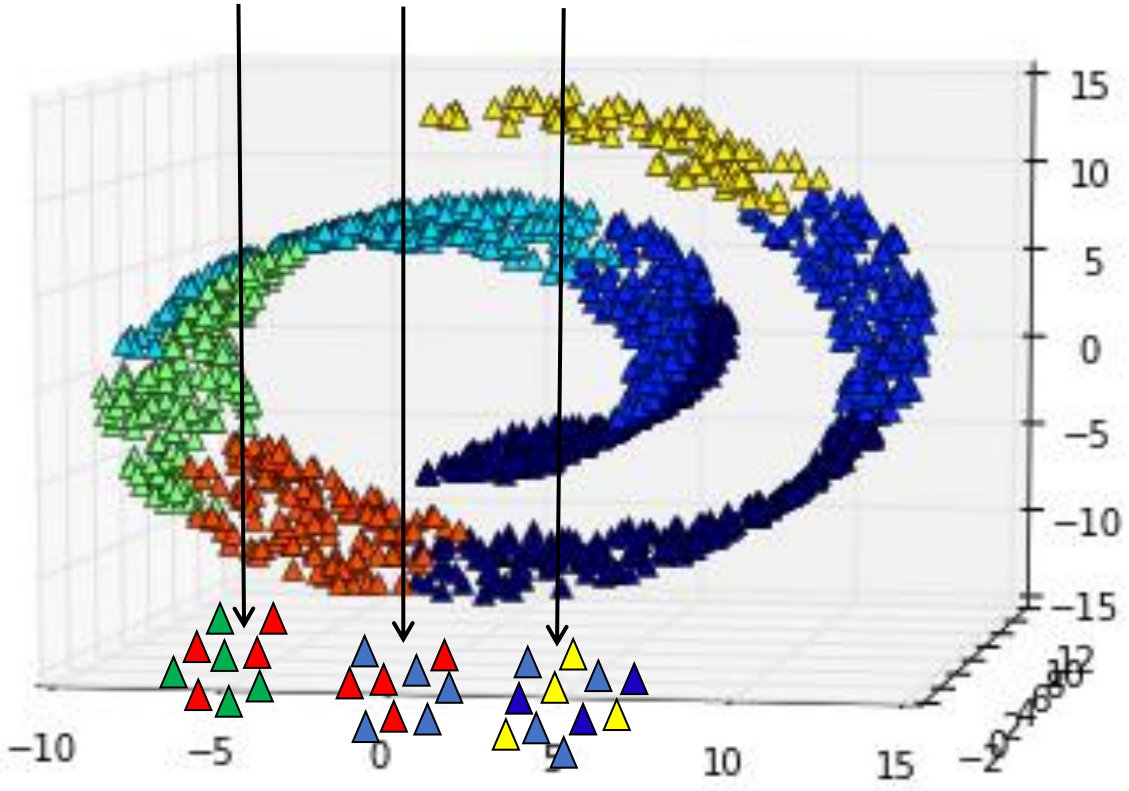
3次元空間内に巻いた平面（2次元）の構造を持つ。



主成分分析の問題点

主成分分析：線形圧縮

高次元空間内に非線形の構造があると、情報の喪失が起こる。



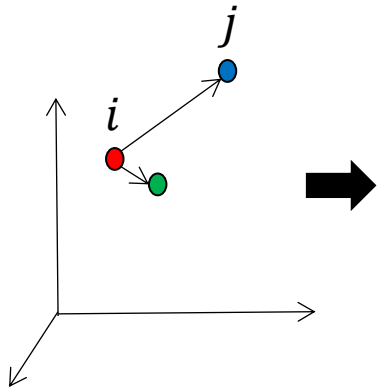
非線形圧縮

- tSNE（t-Distributed Stochastic Neighbor Embedding）
  非線形圧縮法の一つ
  高次元空間内の局所構造を保ったまま低次元に変換する。

tSNE

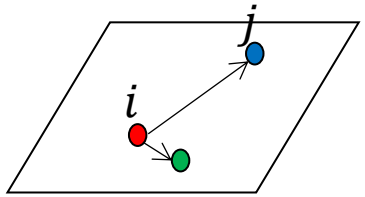
高次元空間上の距離を正規分布で表現

$$p_{j|i} = \frac{e^{\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i{}^2\right)}}{\sum_{k \neq i} e^{\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i{}^2\right)}}$$



低次元空間上の距離をt分布で表現

$$q_{j|i} = \frac{e^{\left(-\left\|y_i - y_j\right\|^2\right)}}{\sum_{k \neq i} e^{\left(-\left\|y_i - y_k\right\|^2\right)}}$$



カルバック・ライブラー情報量：$D(p\|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$

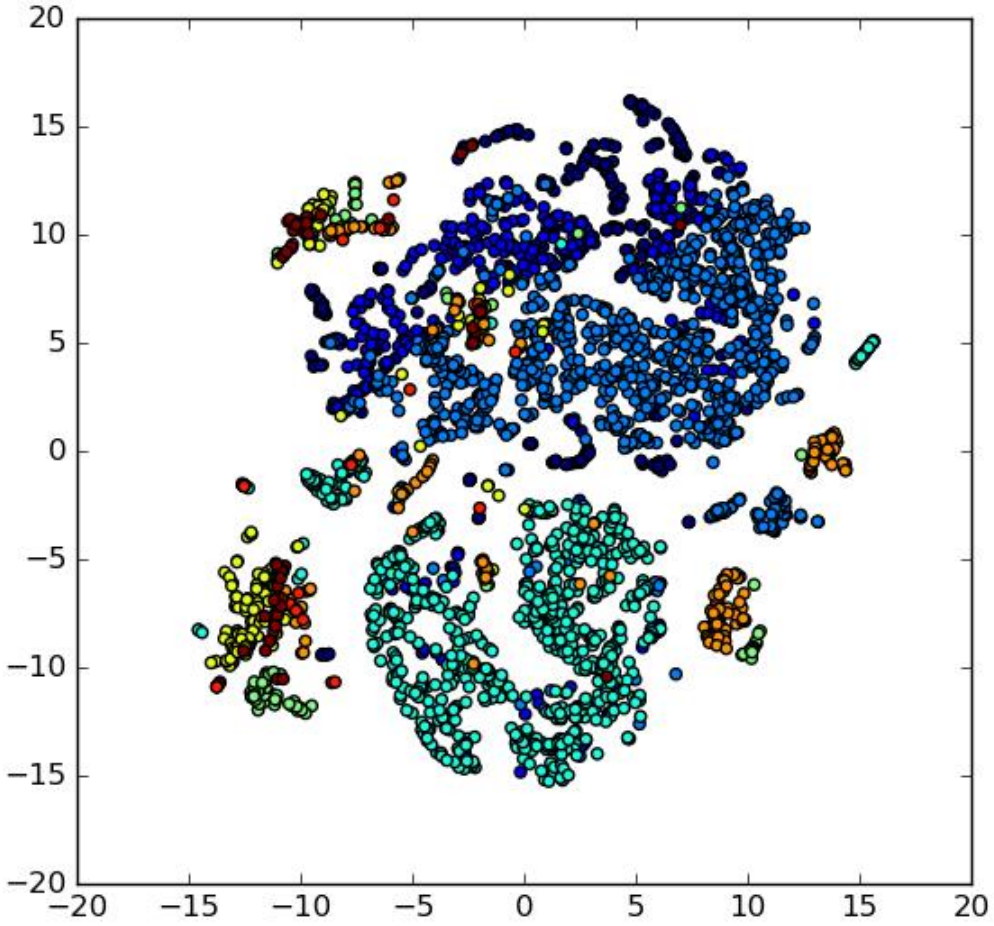
カルバック・ライブラー情報量が最大となるよう、低次元空間の座標を決める。
高次元空間内で距離の近い点同士は、低次元に変換後も近傍にいる確率が高いとする。

tSNEの実例

19946次元のデータ

↓

2次元に縮約



- interneurons
- pyramidal SS
- pyramidal CA1
- oligodendrocytes
- microglia
- endothelial-mural
- astrocytes ependymal

Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.
Zeisel A. et al. ***Science*** 2015 Mar 6;347(6226):1138-1142

まとめ

- 次元縮約
  多次元のデータを、元の情報をなるだけ保持して低次元に変換する。

- Principle Component Analysis（PCA）
  分散共分散行列の固有ベクトルを第1主成分とする。
  データ値のばらつきを保つ。

- t-Distributed Stochastic Neighbor Embedding（tSNE）
  非線形圧縮法
  高次元空間内の局所構造を保つ。