環境情報論第11回

応用編：主成分分析1

神山 翼, @t_kohyama,
tsubasa@is.ocha.ac.jp,
理3-703

今日は，データの重要な部分を
客観的に抜き出す方法を勉強します

応用編：主成分分析1

分散が最大になる方向に座標を回転する分析手法

共分散行列の固有値問題を解き
その固有ベクトルでデータ空間の基底を張り直せば良い

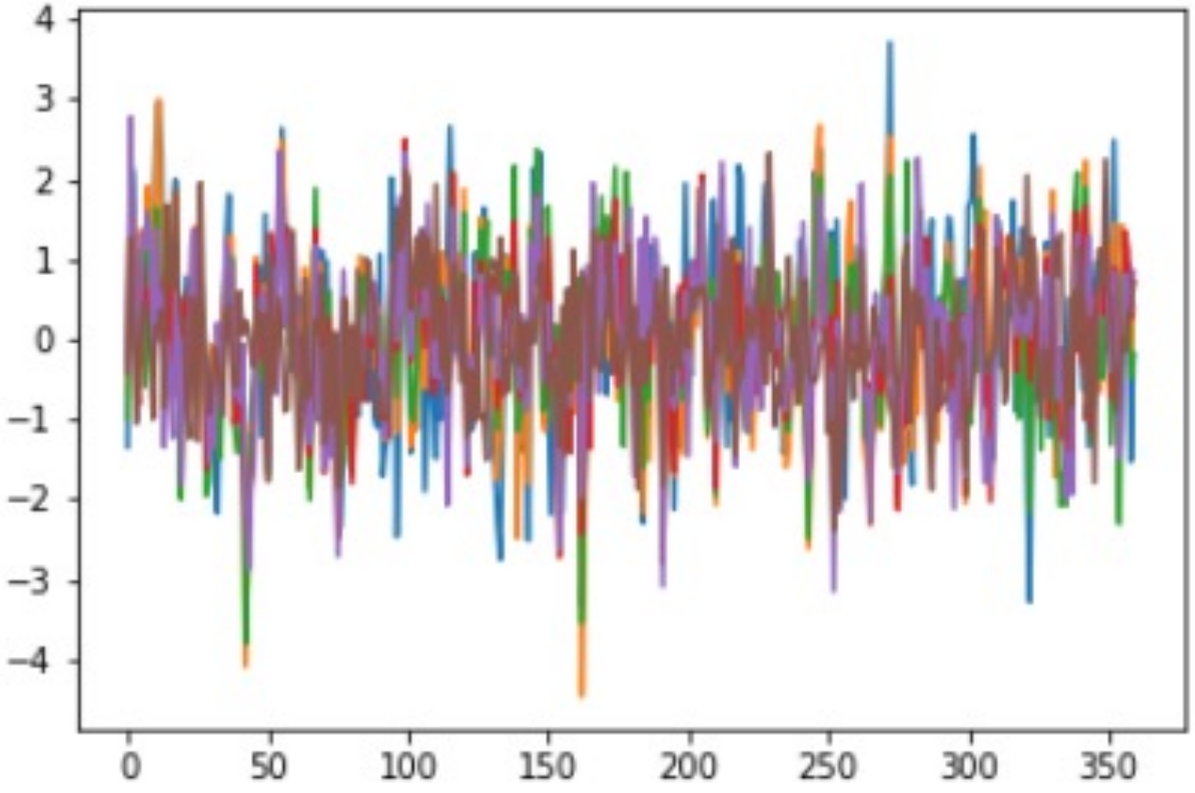
データの中で卓越する変動成分を
少ないデータ数で表現できるようになるのが目標

主成分分析

データの「主成分」を見つける

例1:

札幌・仙台・東京・大阪・福岡・那覇の気温偏差から，「日本で最も目立つ変動」を一つ抜き出したい



似ている時系列が6つもある

→これらをブレンドして
「日本代表」を1つ作れば
十分なケースが結構ある
（**次元圧縮**）

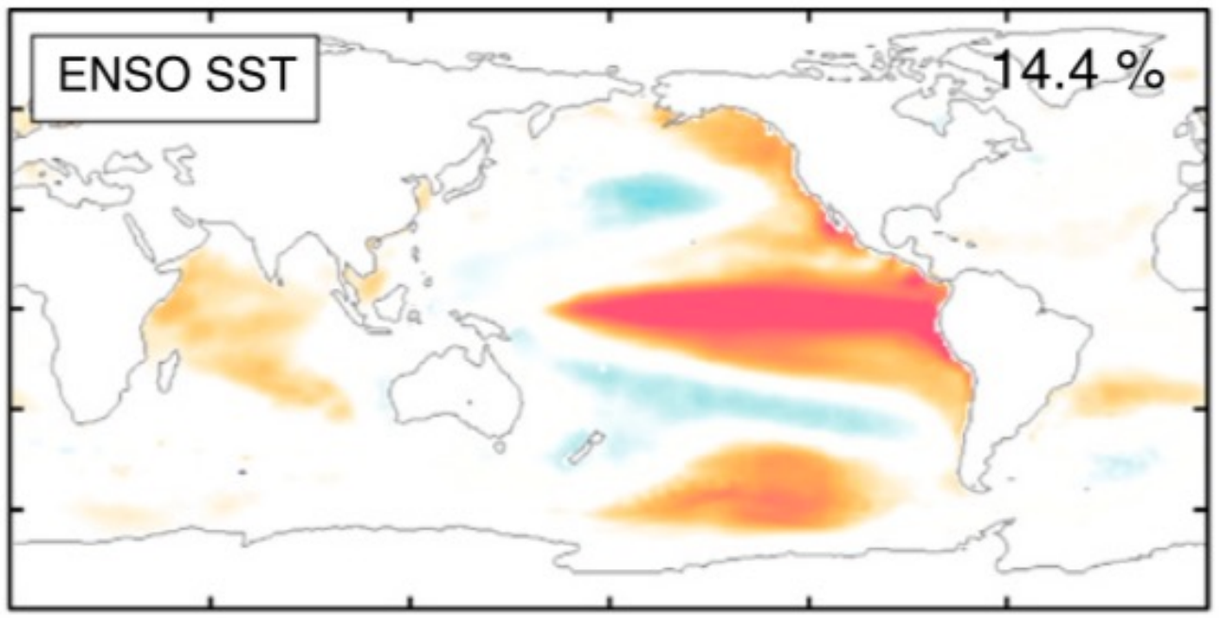
主成分分析

データの「主成分」を見つける

例2:

海面水温データの中から

「最も目立つ変動」を取り出したい



Niño3.4等のインデックスは

主観的に定義されている

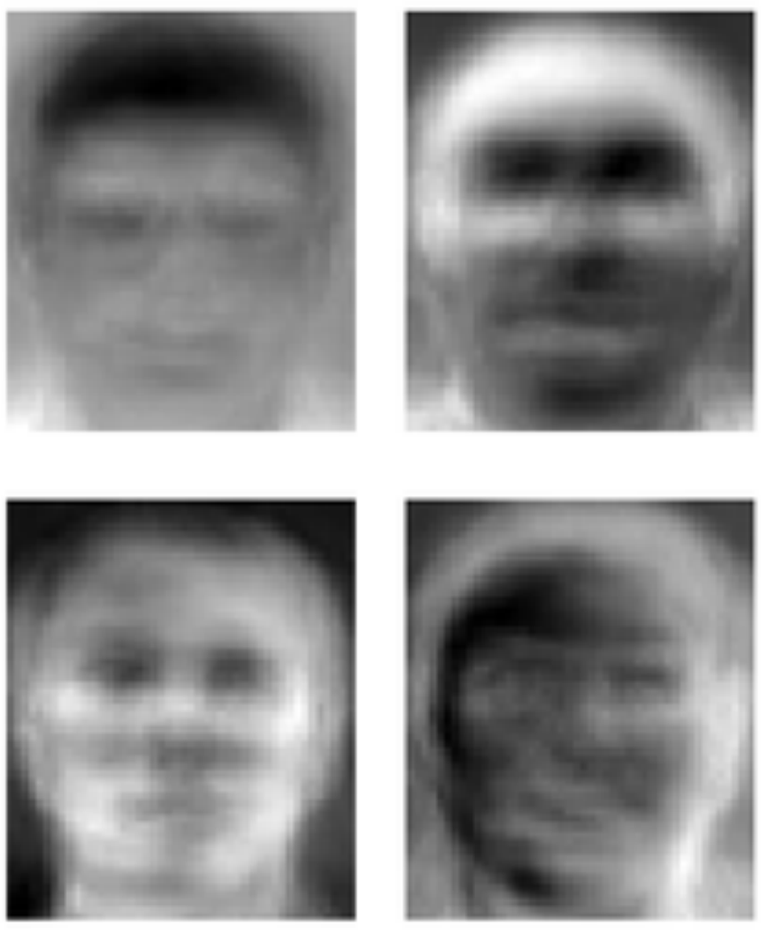
→客観的にインデックス

を定義したい（**特徴抽出**）

Kohyama and Hartmann (2016)

主成分分析

データの「主成分」を見つける



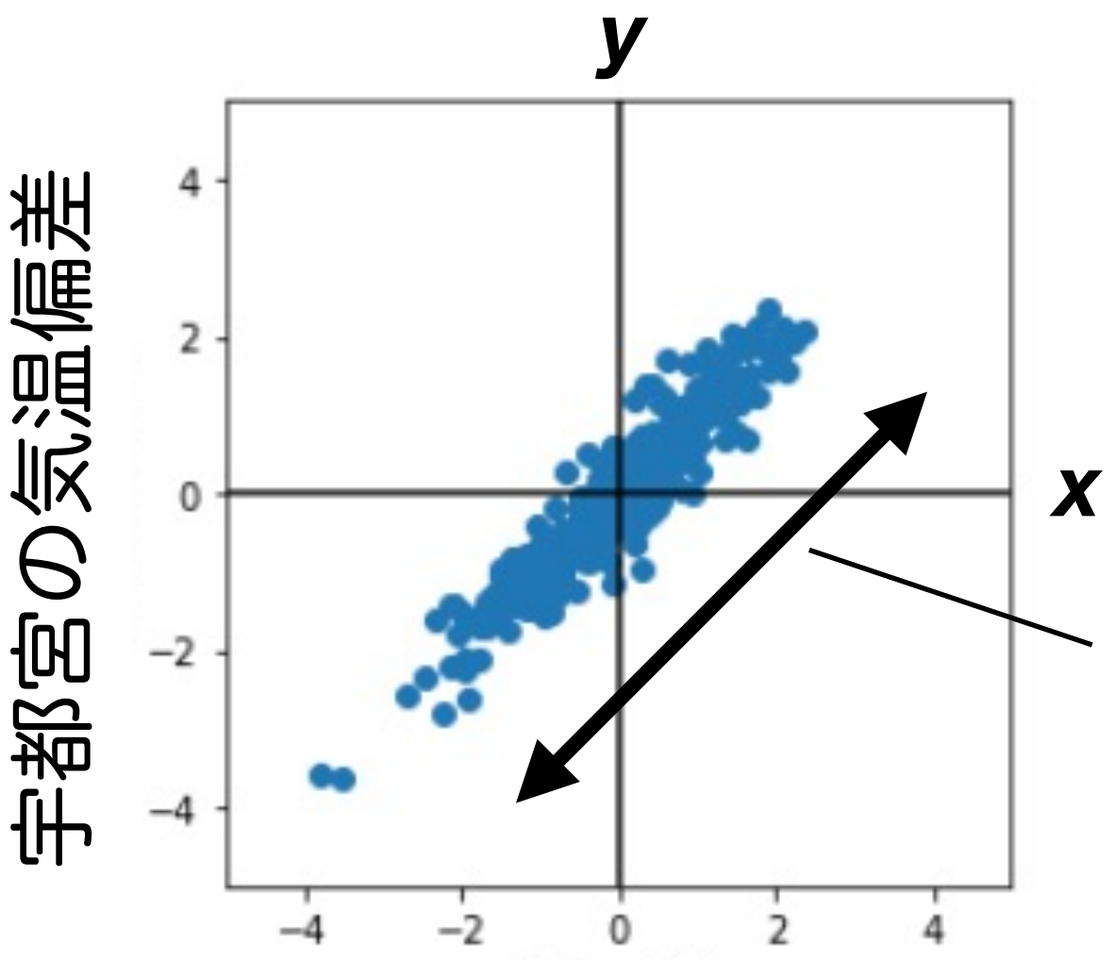
例3: 人間の顔はどこが一番
「人によって違う」部分なのか
を知りたい

少ないデータで（**次元圧縮**）
人を見分ける（**特徴抽出**）
顔認証システムに応用されている

https://ja.wikipedia.org/wiki/固有顔

データの「主成分」とは何か？

まずはデータが2つの場合（2次元の場合）を考えてみよう



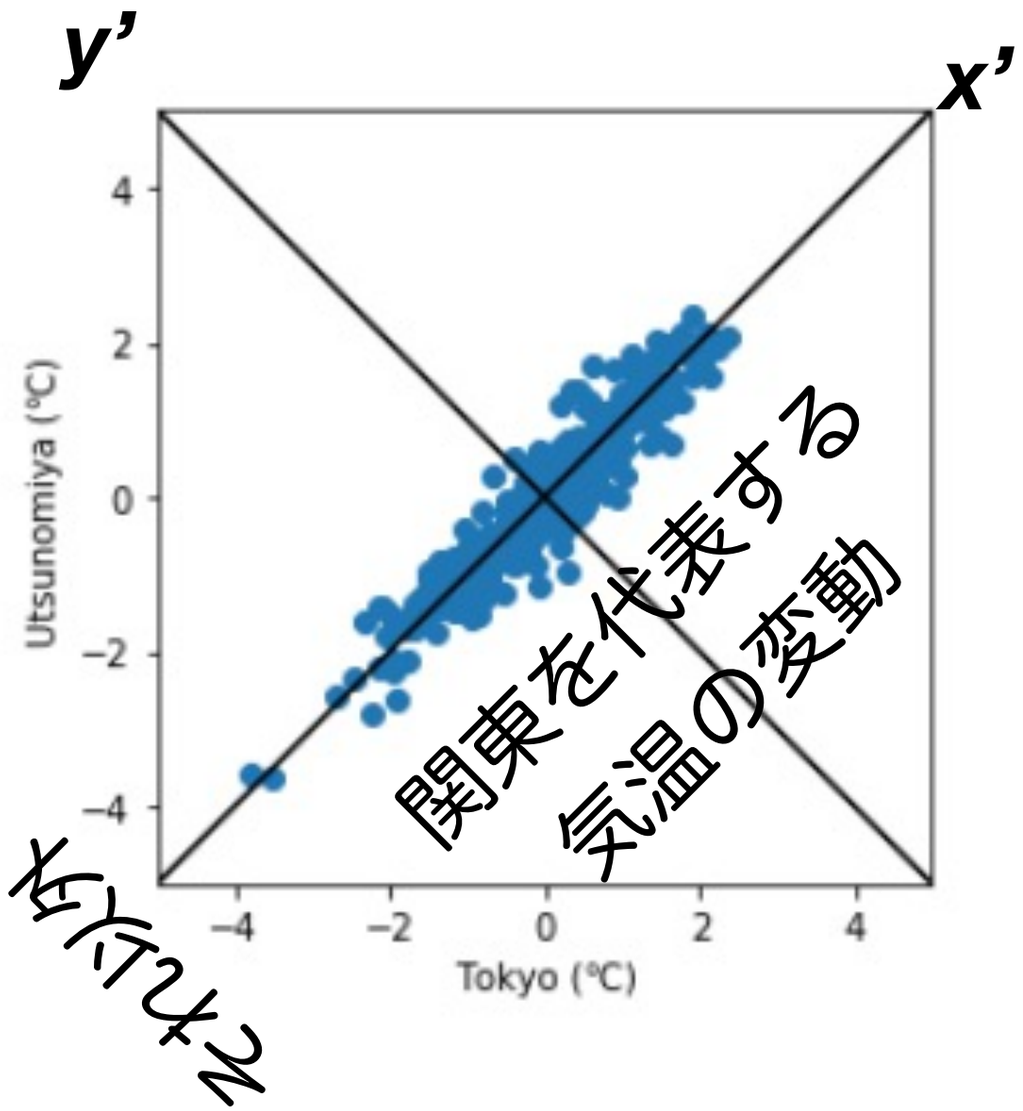
2次元の「データ空間」に
点が散らばっていると解釈する

データの散らばり（分散）が
最大の方向（＝変動の主成分）
を見つけたい

分散が最大となる方向に座標を回転

軸を取り直すと，もっともそのデータを説明できる

情報を最大限に残せる



座標軸を回転させることで，

分散が最大になる方向を探して

新しく軸を取り直す

＝主成分分析

x’のデータだけを見れば

データ空間内の最も本質的な

情報が「圧縮」されている

主成分の見つけ方

「まさかこんなところでも役に立つとは線型代数」シリーズ

データ行列Xを定義

東京の気温偏差

$$X := \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,N} \end{pmatrix}$$

宇都宮の気温偏差

Ｎヵ月分を横に並べる

共分散行列を計算

共分散は高校の復習(?)

共分散 (covariance)

$$\begin{aligned}\mathrm{cov}(\vec{x_1}, \vec{x_2}) &:= (x_{1,1}x_{2,1} + x_{1,2}x_{2,2} + \cdots + x_{1,N}x_{2,N})/N \\ &= (\vec{x_1} \cdot \vec{x_2})/N\end{aligned}$$

共分散行列

numpyなら
計算は瞬殺！
↓

$$C := \begin{pmatrix} \mathrm{cov}(\vec{x_1}, \vec{x_1}) & \mathrm{cov}(\vec{x_1}, \vec{x_2}) \\ \mathrm{cov}(\vec{x_2}, \vec{x_1}) & \mathrm{cov}(\vec{x_2}, \vec{x_2}) \end{pmatrix} = XX^T/N$$

共分散行列の固有ベクトルを計算

「え，対角化ってデータ解析で使うんですか…?」

$$C\vec{e_1} = \lambda_1\vec{e_1},\ C\vec{e_2} = \lambda_2\vec{e_2} \iff CE = E\Lambda$$

$$\iff E^{-1}CE = \Lambda$$

$$E = \begin{pmatrix} \vec{e_1} & \vec{e_2} \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

ただし，$\lambda_1 \geq \lambda_2 > 0$

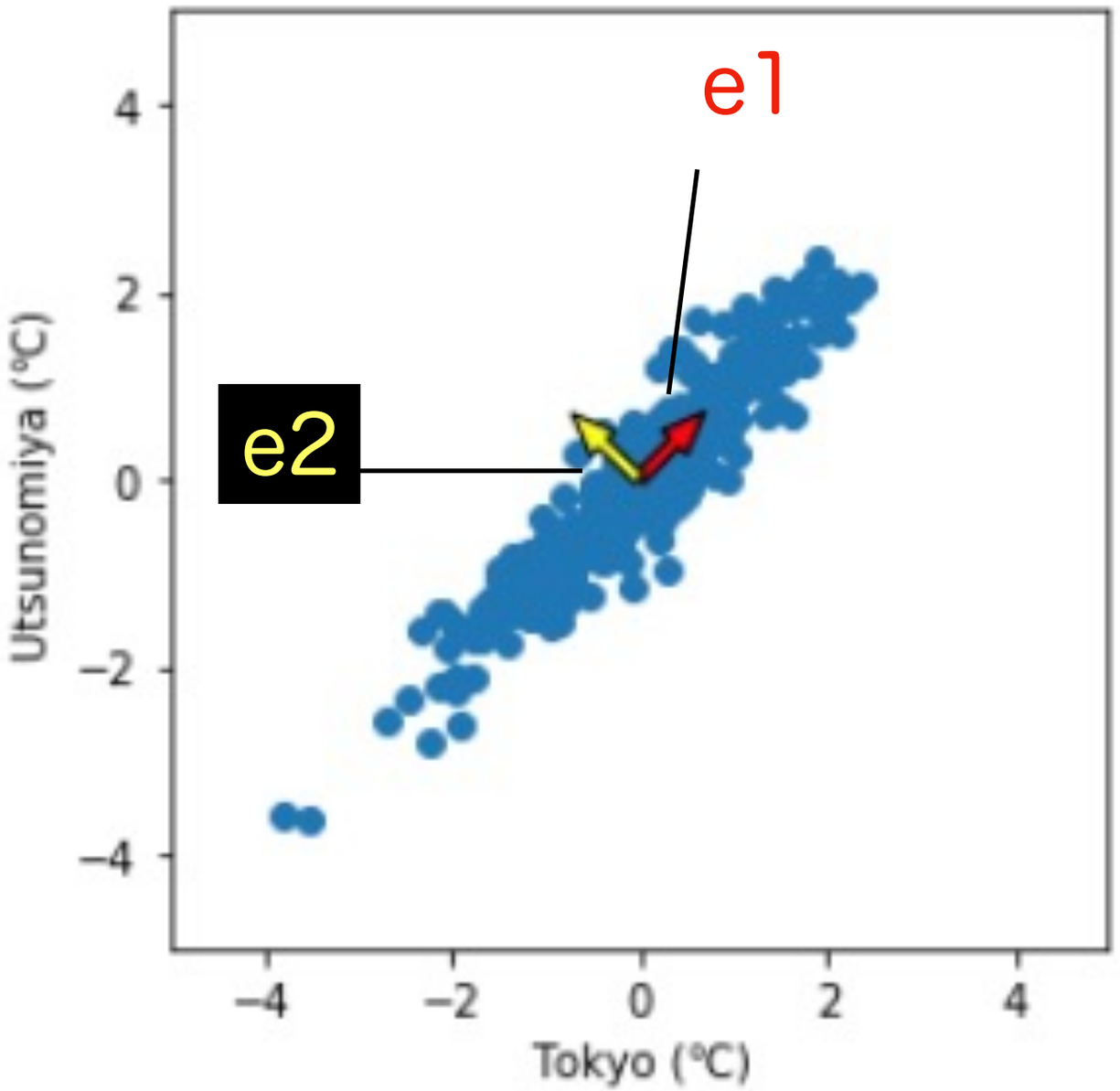
C は実対称行列なので

固有値は実数，固有ベクトルは直交

（「線型代数学4」の復習）

固有ベクトルの方向が分散最大の方向

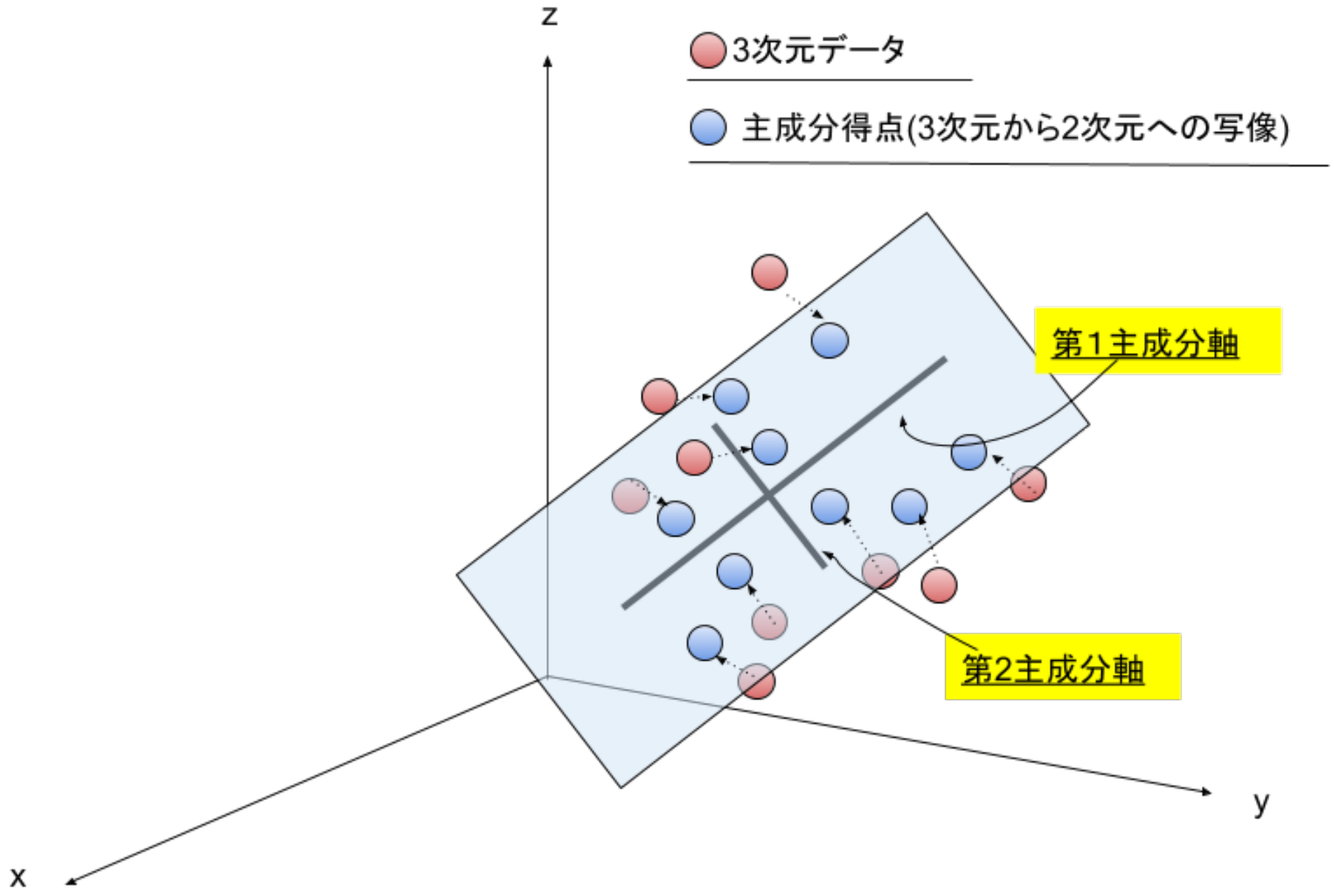
最大の固有値に対応する固有ベクトルが
主成分の軸の向き



軸を見つけてからの手続きは
次週のお楽しみ。

この方法ならn次元に拡張してもいけそう

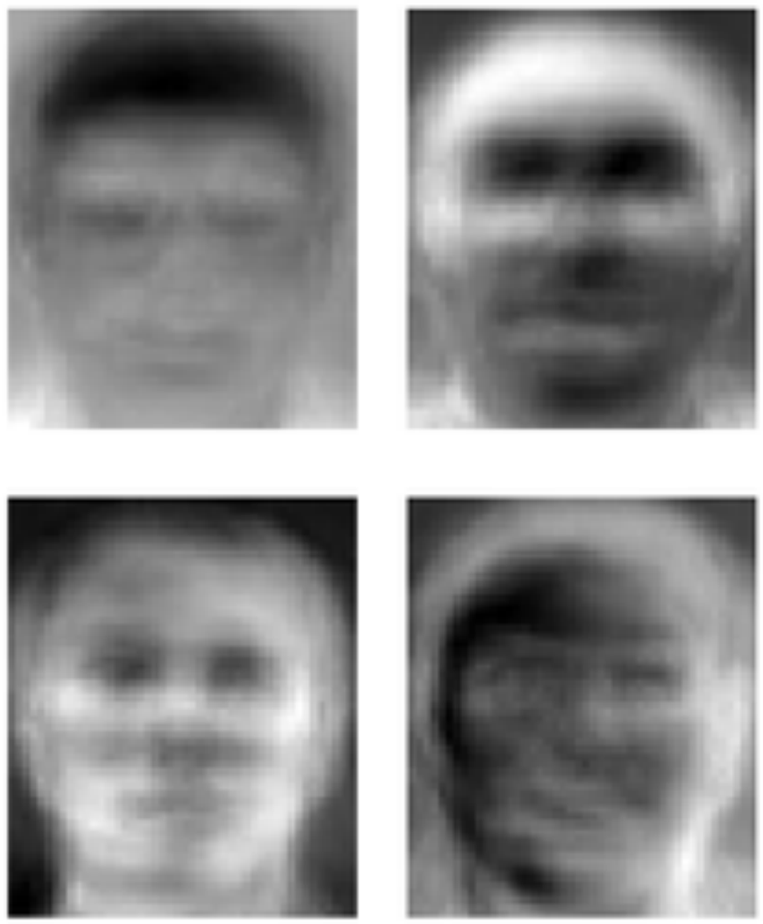
6次元のデータ空間は描けないけど
固有ベクトルなら同じ手続きで求められる



https://qiita.com/NoriakiOshita/items/460247bb57c22973a5f0

顔認証の話をもう一度

少ないデータで顔認証したいとき
どこの類似度を見れば良いか？



「固有顔」の方向の
類似度のみを判定すれば良い

↓

データ空間内の特定の次元にのみ
注目すれば，**少ないデータで**
特徴を抽出できる

https://ja.wikipedia.org/wiki/固有顔

たのしい課題たち

A: 主成分分析の説明

B:6次元データについて同じことをやってみる

C: 実対称行列の固有値固有ベクトルの性質の証明

D: 固有ベクトルの方向が分散最大方向と
一致することの証明（概略のみでOK）

今日は，データの重要な部分を
客観的に抜き出す方法を勉強します

応用編：主成分分析1

分散が最大になる方向に座標を回転する分析手法

共分散行列の固有値問題を解き
その固有ベクトルでデータ空間の基底を張り直せば良い

データの中で卓越する変動成分を
少ないデータ数で表現できるようになるのが目標

本日の導入パートは以上です。
何でも良いのでZoomの方に
授業に関係のあるコメントを
してください（出席代わり）。

コメント拾いが終わったら,
早速今日のプログラミングに進みましょう。