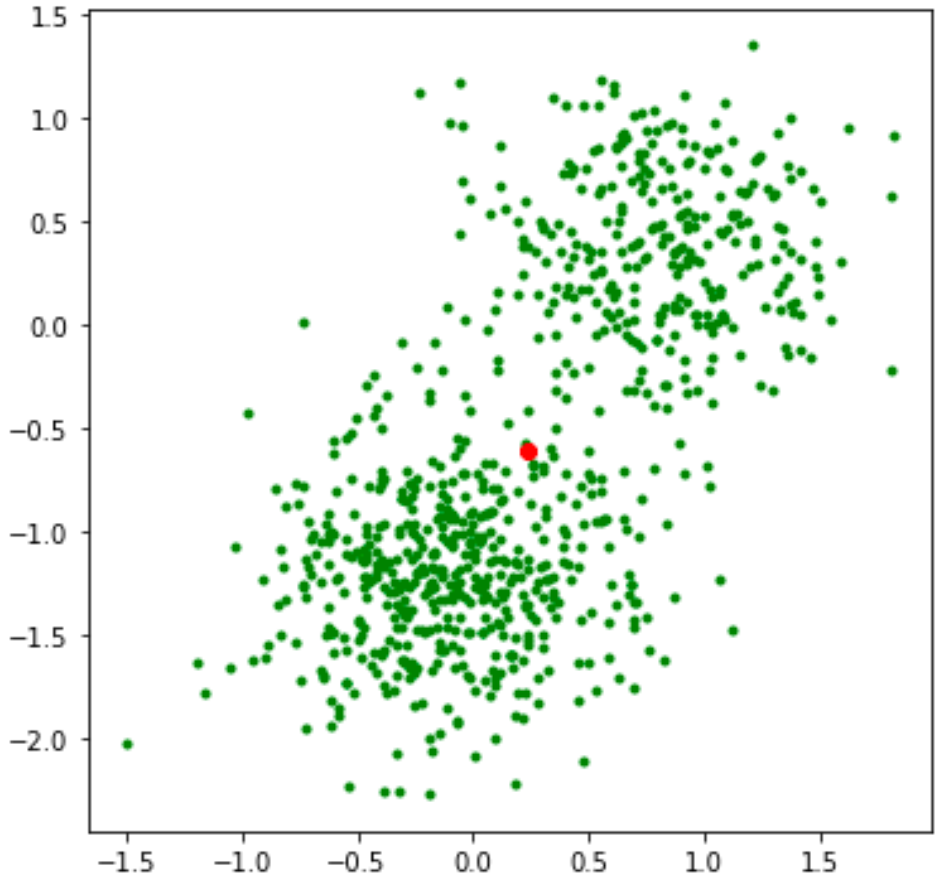
教師なし学習

正田 備也

masada@rikkyo.ac.jp

データ集合を要約する

- 平均ベクトルで要約する
  - すべてのベクトルの和を求めて個数で割る



例）右図のような2次元ベクトルの集まり

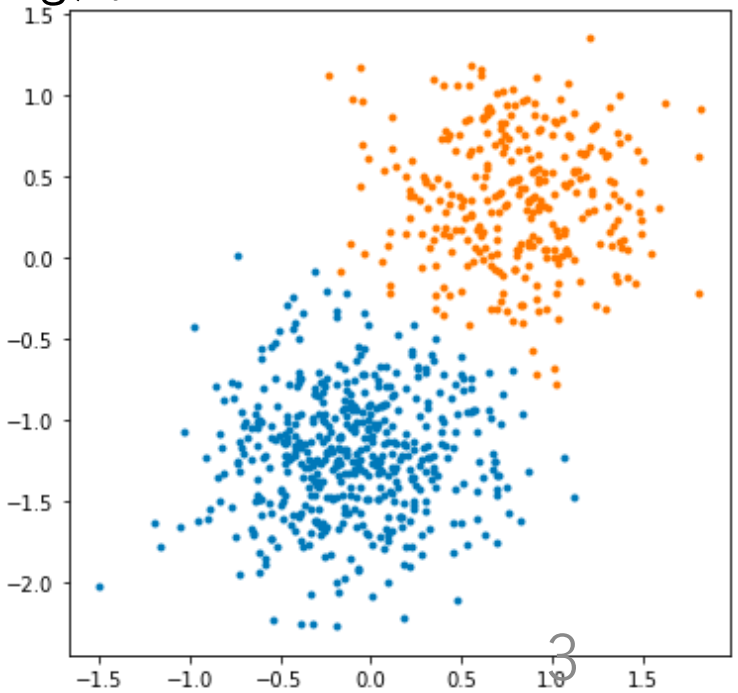
- 平均は、全体をうまく要約していると言えるか？
- データがひとつにまとまって分布していないように見える
- どうする？

教師なし学習 (unsupervised learning)

- 与えられているのは、入力ベクトルの集まりだけ

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$

- 正解となるクラスラベルy_iは対応づけられていない
- そうであっても・・・
- データの自然なまとまりを見つけたい
  - 右図はK-meansクラスタリングの実際の結果。
- そこで、教師なし学習の出番！
  - 次元圧縮も、もうひとつの重要な用途。



階層的クラスタリング

- 入力ベクトルがバラバラの状態から始める
  - つまり、どのクラスタも1つのデータ点だけを含む状態から始める。
- 一番距離が近いクラスタのペアをマージする
  - どのような基準でクラスタのペアを選ぶかで、クラスタリング結果が違ってくる。

- 「一番距離が近い」というところの近さの基準をどうするか？
  - これをlinkage criterionと呼ぶ

linkage criterion
どの2つのクラスタをマージするか？

1. Ward criterion （分散を利用した方法）
   - クラスタの重心からの距離の2乗の和を、すべてのクラスタにわたって合算したものが、最小になるように、クラスタをマージする。
2. Complete linkage criterion （最大値を最小化する方法）
   - 異なるクラスタに属するデータ点の距離の最大値が最小のクラスタをマージする
3. Average linkage criterion （平均値を最小化する方法）
   - 異なるクラスタに属するデータ点の距離の平均値が最小のクラスタをマージする
4. Single linkage criterion （最小値を最小化する方法）
   - 異なるクラスタに属するデータ点の距離の最小値が最小のクラスタをマージする

階層的クラスタリングの応用例

- 単語ベクトルをクラスタリング
  - Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft), Section 6.9.
    - 元々階層構造や派生構造をもつ対象に使うと、うまくいくことがある。
    - 意味や範疇が近い単語は階層構造の上でも近くに位置している（？？？）

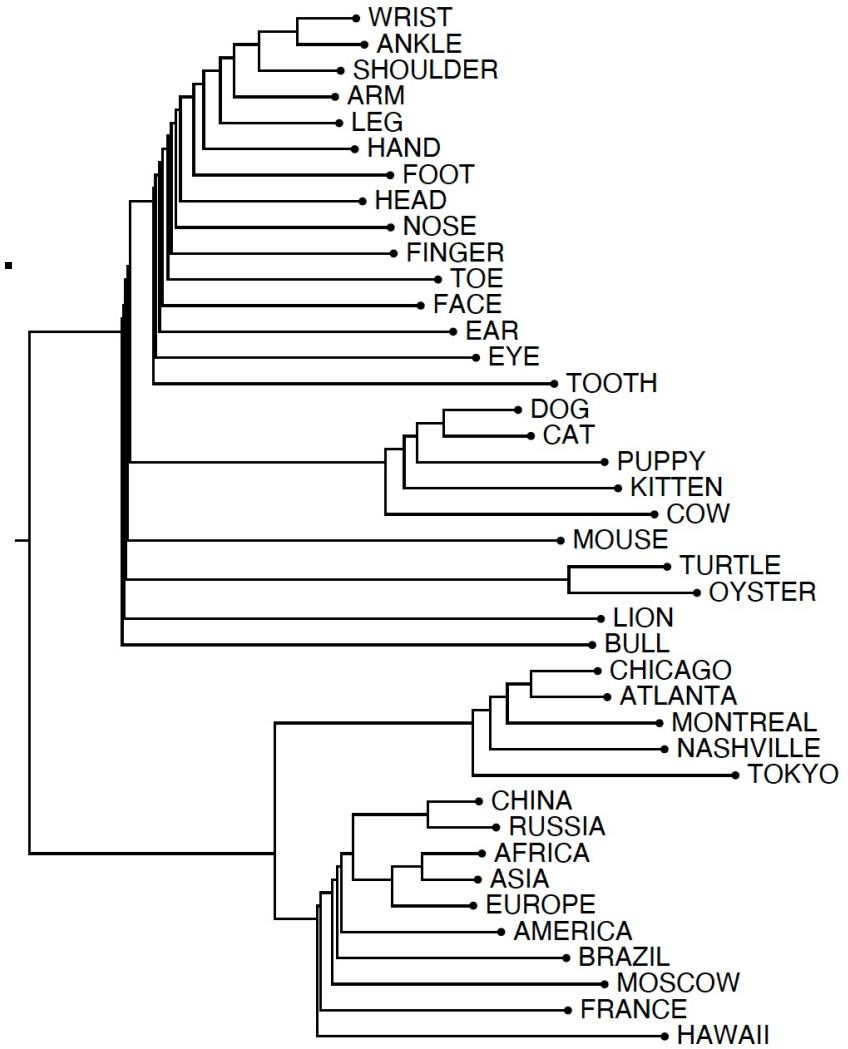
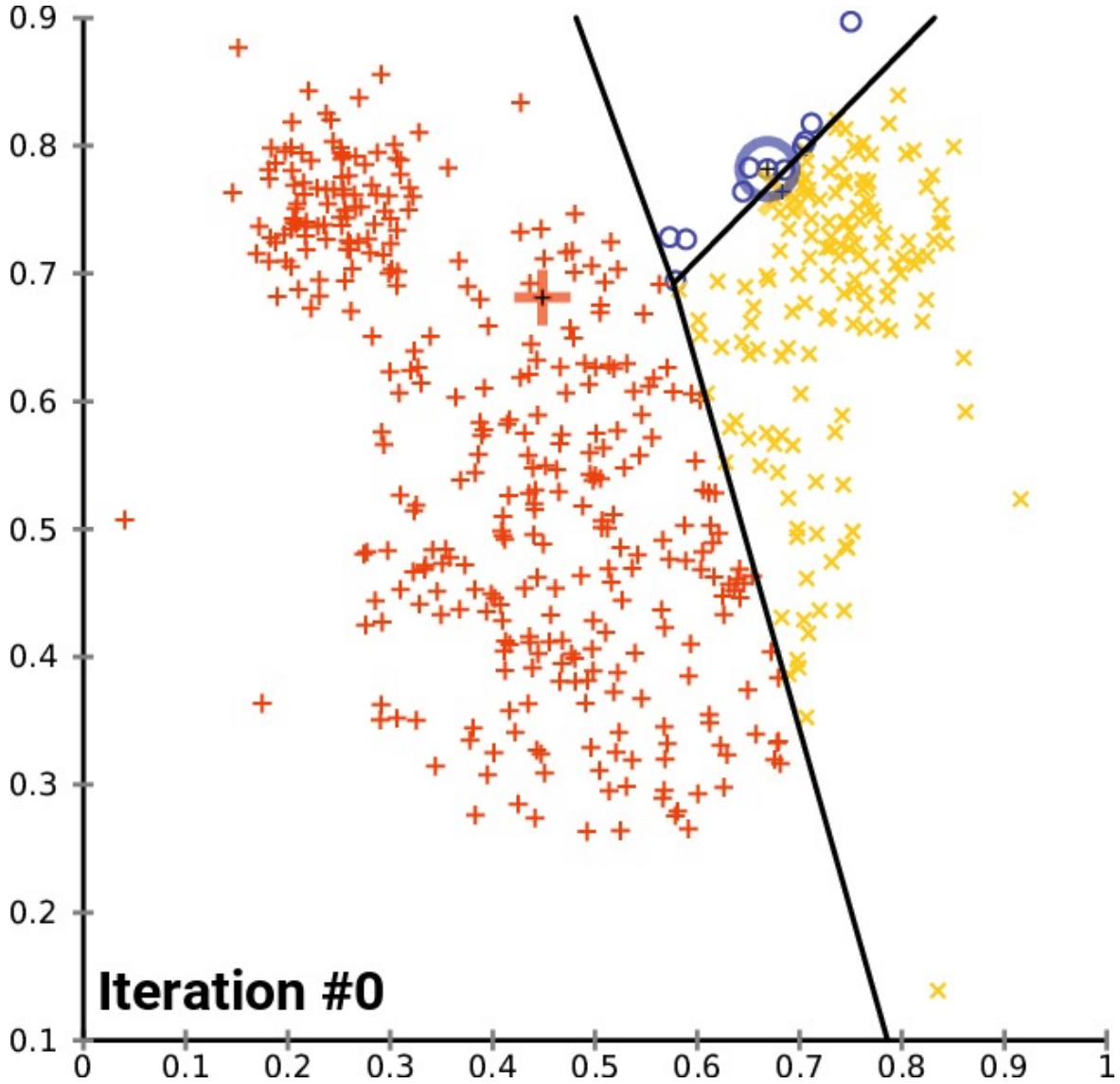


Figure 9: Hierarchical clustering for three noun classes using distances based on vector correlations.

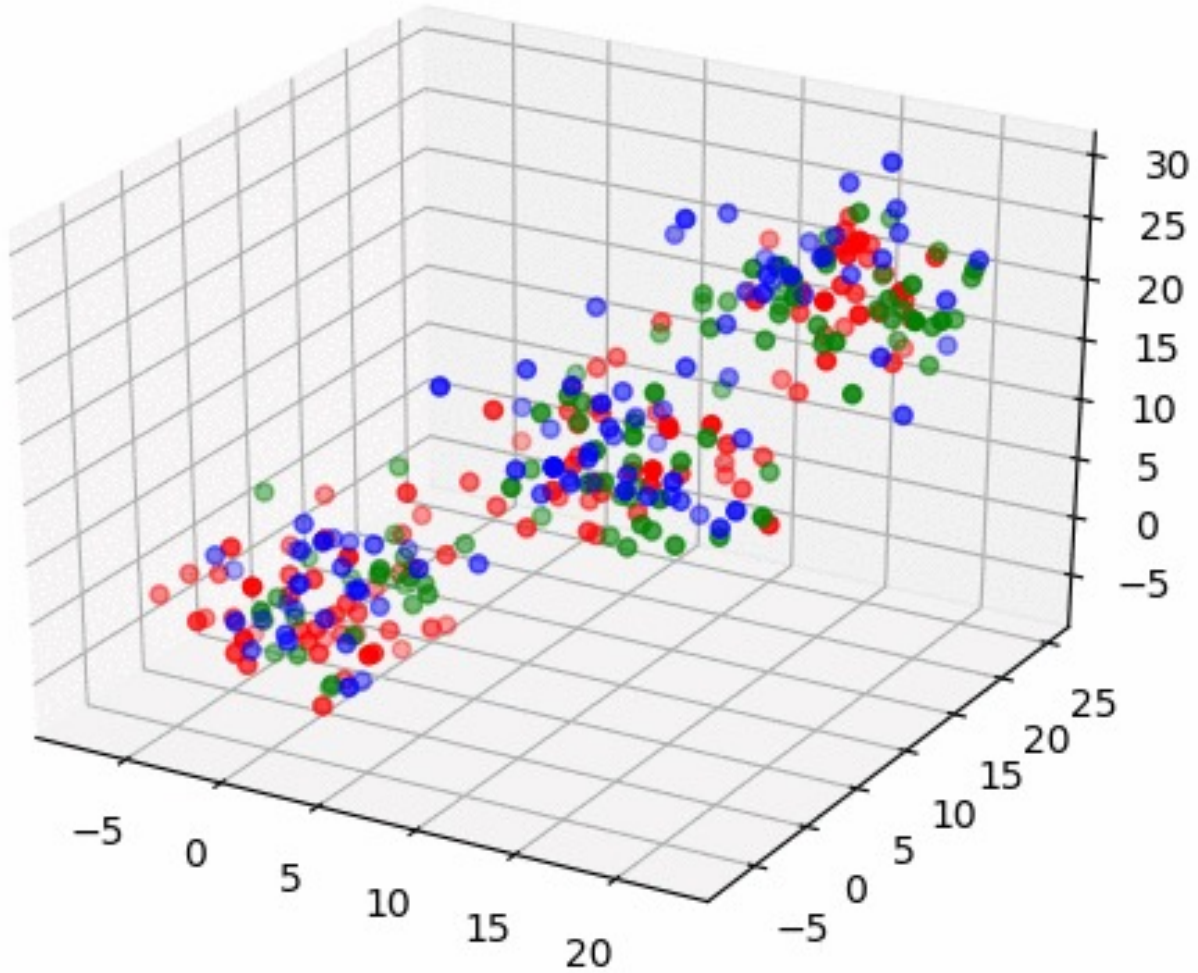
K-平均法 (K-Means)

- とても良く知られているクラスタリング手法
- 与えられたベクトル集合をK個の排他的なクラスタに分ける

1. K個のクラスタの重心(centroid)を適切に初期化
2. N個のベクトルの各々を、最も近い重心に対応するクラスタへ属させる（K個の重心までの距離を求め、最小値を探す）
3. 構成されたクラスタの重心を求め直す
4. クラスタの重心がまだ変動しているなら、2. へ戻る



https://commons.wikimedia.org/wiki/File:K-means_convergence.gif



http://marubon-ds.blogspot.com/2018/04/introduction-to-k-means-algorithm-and.html

クラスタリング手法の出力

- model-basedなクラスタリングは2種類の情報を出力
  1. 各入力ベクトルのクラスタへの所属の情報
  2. クラスタの数理的な表現（例えば重心）

- K-平均法の場合
  1. 各入力ベクトルは、ちょうどひとつのクラスタに属する
  2. 各クラスタは、その重心で表現される

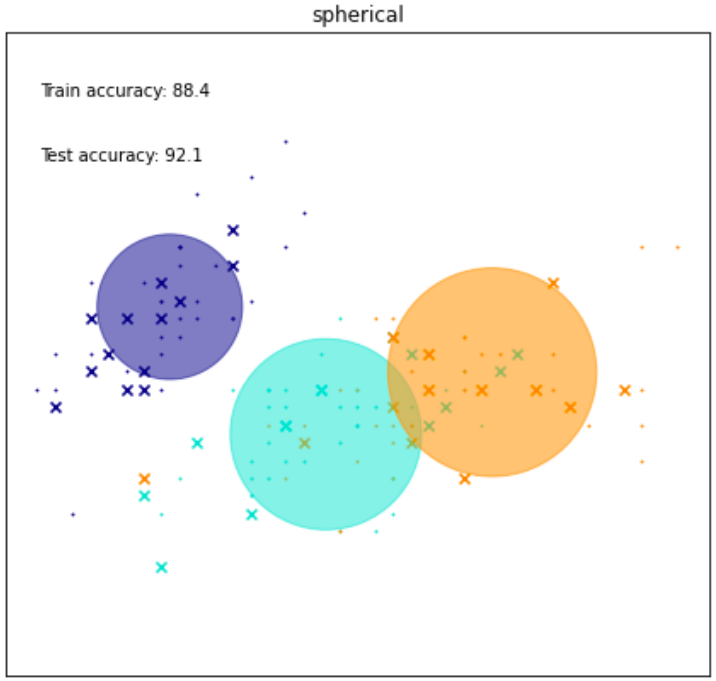
混合ガウス分布によるクラスタリング

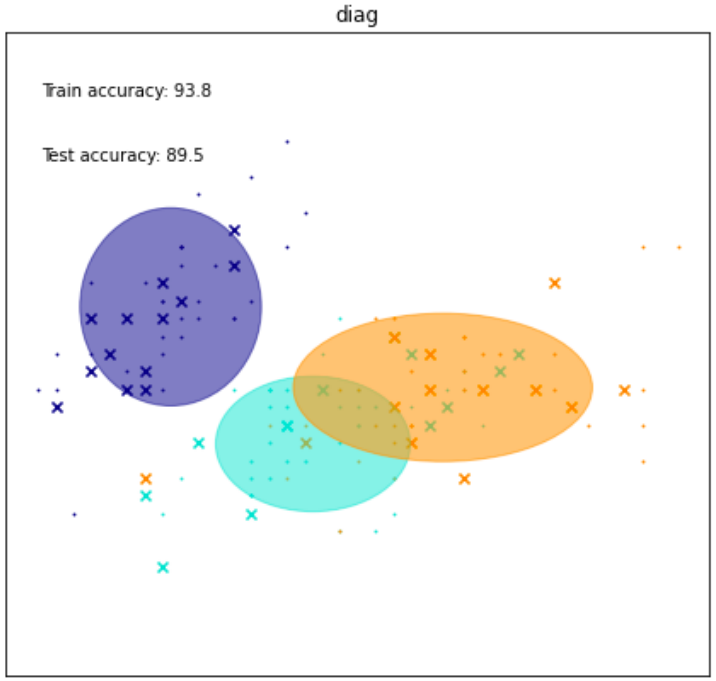
- K-平均法の仮定を緩めた手法
  - 詳細は秋学期の「統計モデリング１」で。

1. 入力ベクトルはちょうど一つのクラスタに属さなくてよい
   - 各クラスタへの所属確率（所属の度合い）が得られる
2. 各クラスタの分布は同心円状の広がりを持っていなくてよい
   - 横に広がった円でいいし、傾いていてもいい
     - クラスタ形状の情報は多変量正規分布の共分散行列によって与えられる

https://github.com/tomonari-masada/courses/blob/master/SML2020/Gaussian_Mixture_example01.ipynb

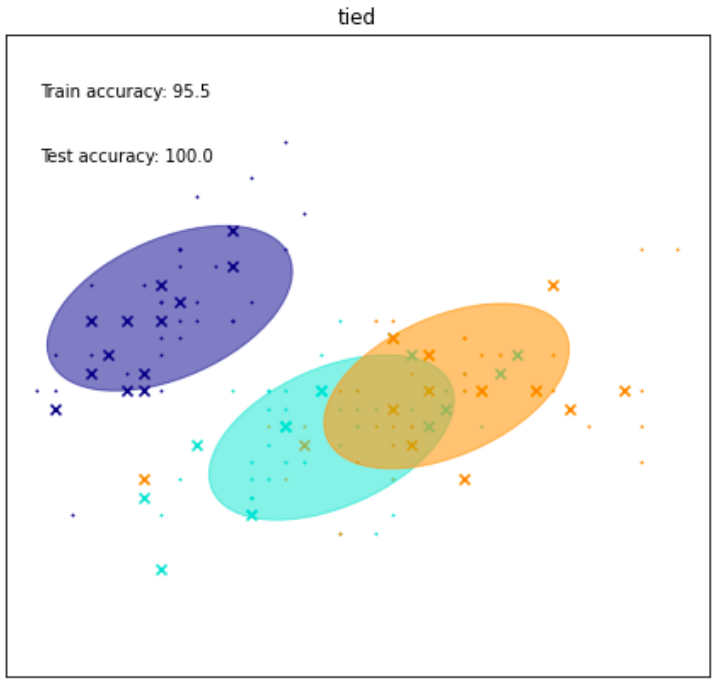
共分散行列が
すべての対角成分
が等しい対角行列

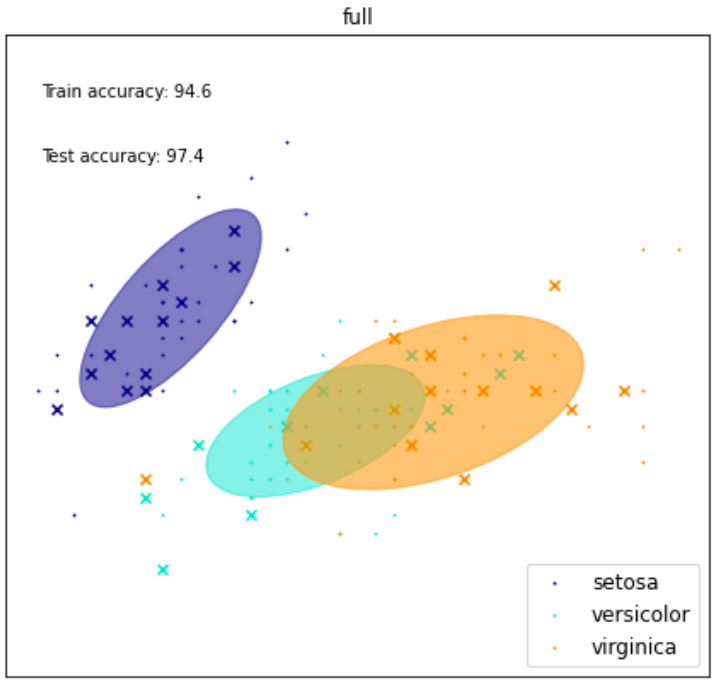




共分散行列が
対角成分が様々な
対角行列

共分散行列が
対角行列でないが
成分は共有



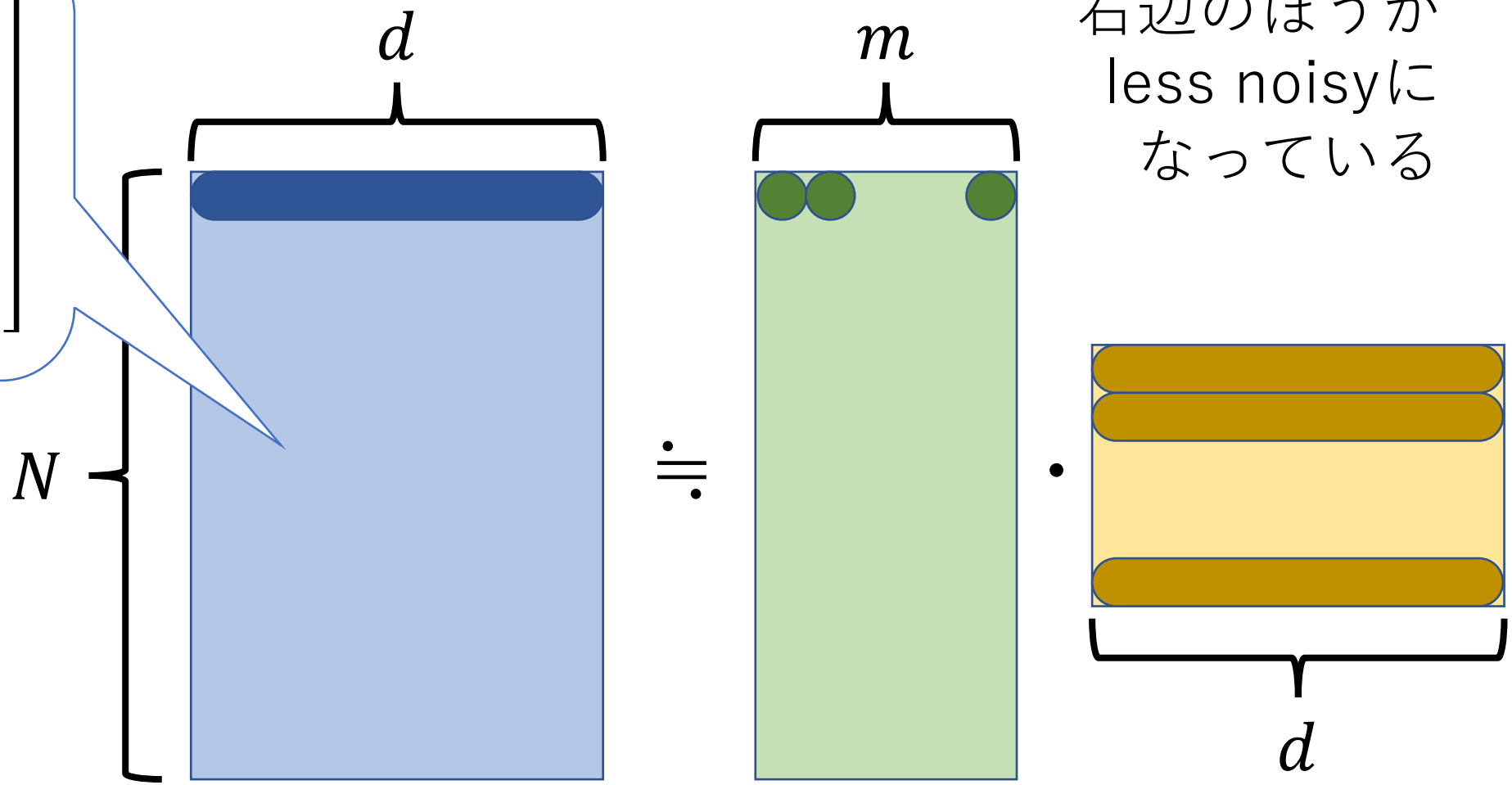


共分散行列が
対角行列でない
成分も共有しない

次元圧縮
dimensionality reduction

（復習）PCAで計画行列をless noisyにする

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,d} \end{bmatrix}$$



右辺のほうが
less noisyに
なっている

左辺は
元々の
計画行列

演習11-1

- 次の行列の積を求めなさい

$$\begin{bmatrix} -7 & 6 \\ 8 & -4 \\ 7 & 2 \end{bmatrix} \begin{bmatrix} -2 & 4 & -6 \\ -6 & -1 & -9 \end{bmatrix}$$

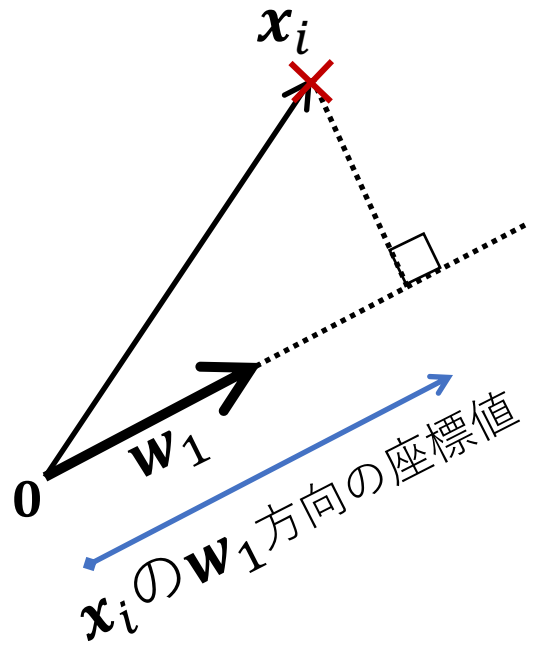
PCAの考え方

- 高次元空間に散らばったデータ点の集合が与えられている
    - データ点はベクトルです（原点から発してその点まで達する矢印のイメージ）
  - データの重心が原点になるように、全体の位置をシフトさせておく。
- データ点が最も広く散らばっている方向を見つける
  - この方向を表すベクトルが、主成分
- その方向と直交する超平面上に、全てのデータ点を押しつぶす
  - データ点の散らばっている空間の次元が、ひとつだけ下がる
- 次元を下げた空間で、また同じことをする
  - つまり、データ点が最も広く散らばっている方向を見つける

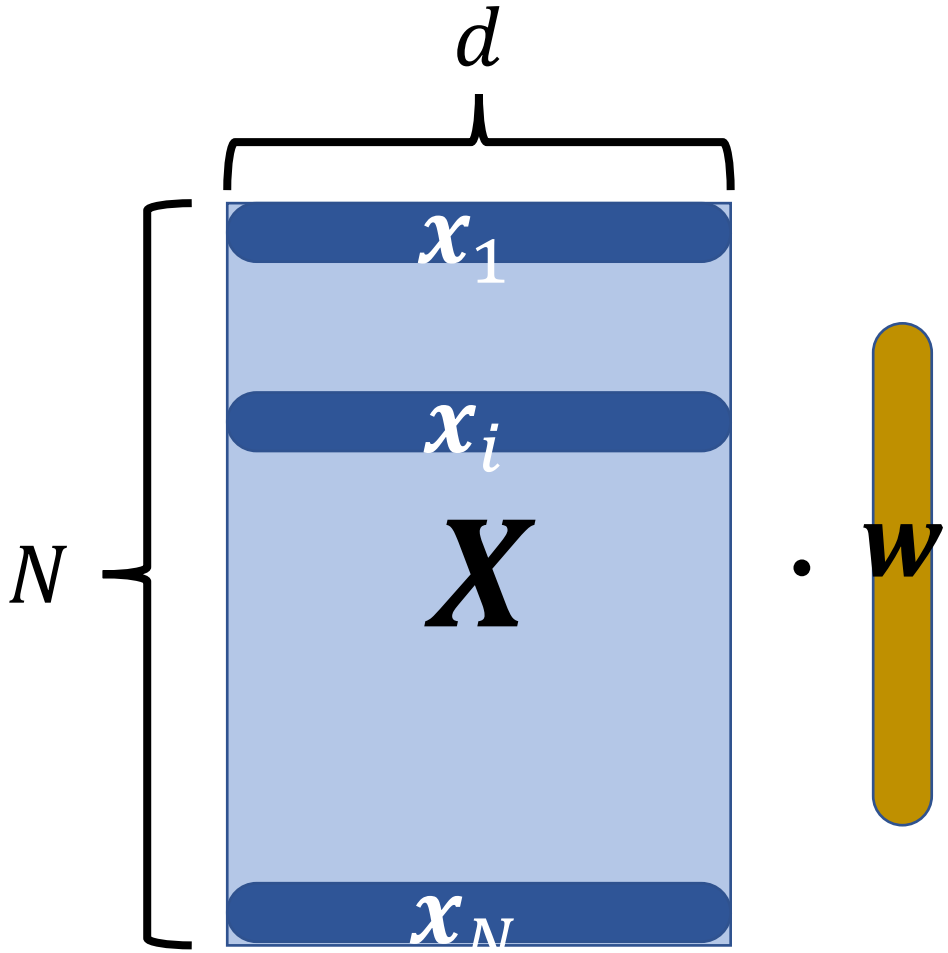
「最も広く散らばっている方向」

- 式で書くと以下の通り

$$\boldsymbol{w}_1 = \underset{\|\boldsymbol{w}\|=1}{\arg\max} \|\boldsymbol{X}\boldsymbol{w}\|^2$$



  - $\boldsymbol{X}$は計画行列（ただし重心を原点へ移動した後のもの）
  - $\boldsymbol{X}\boldsymbol{w}$は列ベクトル
    - $\boldsymbol{X}$の各行ベクトル$\boldsymbol{x_i}$と$\boldsymbol{w}$との内積の値が要素として並ぶ、列ベクトル。
    - この内積の値は、 $\boldsymbol{X}$の各行ベクトル$\boldsymbol{x_i}$の、$\boldsymbol{w}$の方向への座標値。
      - つまり、 $\boldsymbol{w}$の方向への＋かーの符号がついた長さ



$\boldsymbol{Xw}$のイメージ

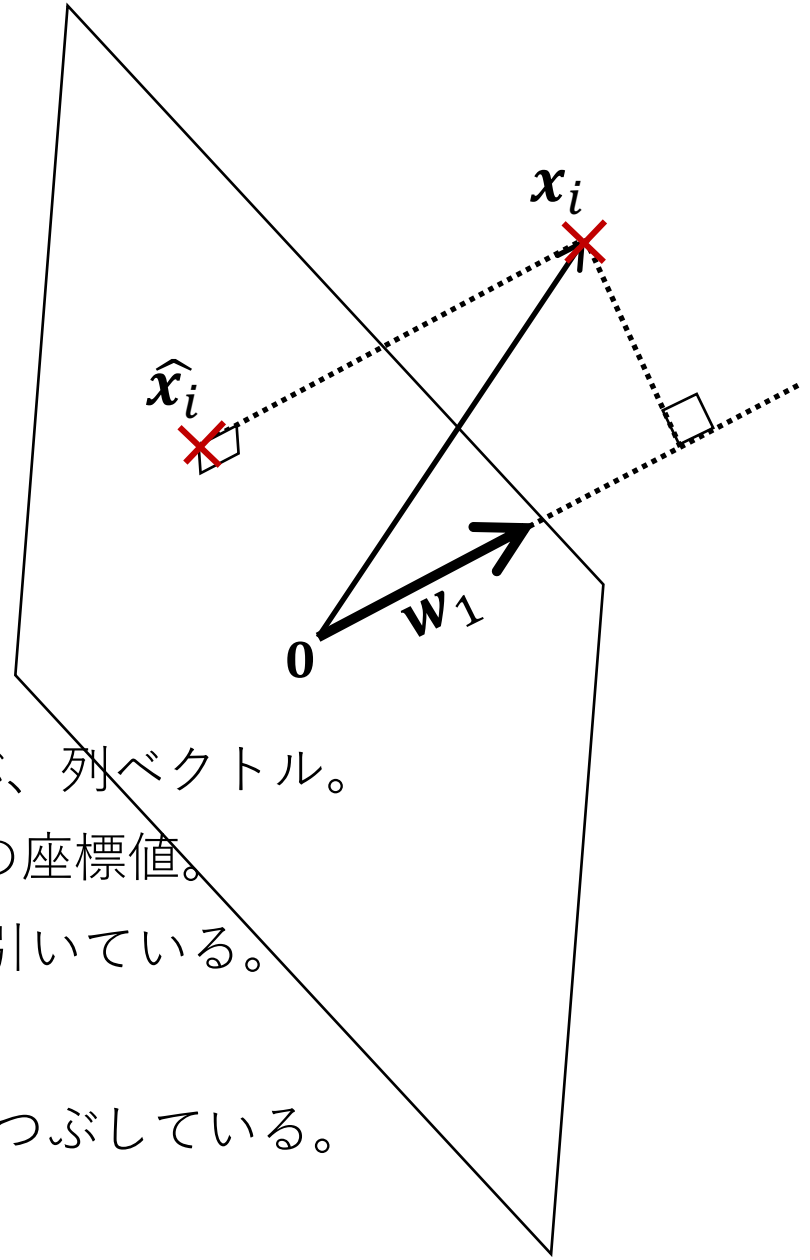
各データ点$\boldsymbol{x_i}$と$\boldsymbol{w}$との内積$\boldsymbol{x_i}^T\boldsymbol{w}$を、まとめて書いているだけ。

「データ点を押しつぶす」

- 式で書くと以下の通り

$$\hat{\boldsymbol{X}} = \boldsymbol{X} - (\boldsymbol{X}\boldsymbol{w}_1)\boldsymbol{w}_1{}^T$$

  - $\boldsymbol{X}\boldsymbol{w}$は列ベクトル
    - $\boldsymbol{X}$の各行ベクトル$\boldsymbol{x}_i$と$\boldsymbol{w}$との内積の値が要素として並ぶ、列ベクトル。
    - この内積の値は、 $\boldsymbol{X}$の各行ベクトル$\boldsymbol{x}_i$の、$\boldsymbol{w}$の方向への座標値。
    - その座標値に$\boldsymbol{w}_1$をかけて、元の$\boldsymbol{X}$の行ベクトル$\boldsymbol{x}_i$から引いている。
    - こうすると、 $\boldsymbol{x}_i$から$\boldsymbol{w}_1$方向の成分を消せる。
    - つまり、$\boldsymbol{x}_i$を、$\boldsymbol{w}_1$と直交し原点を通る超平面上へ押しつぶしている。



PCAの（実際の）アルゴリズム

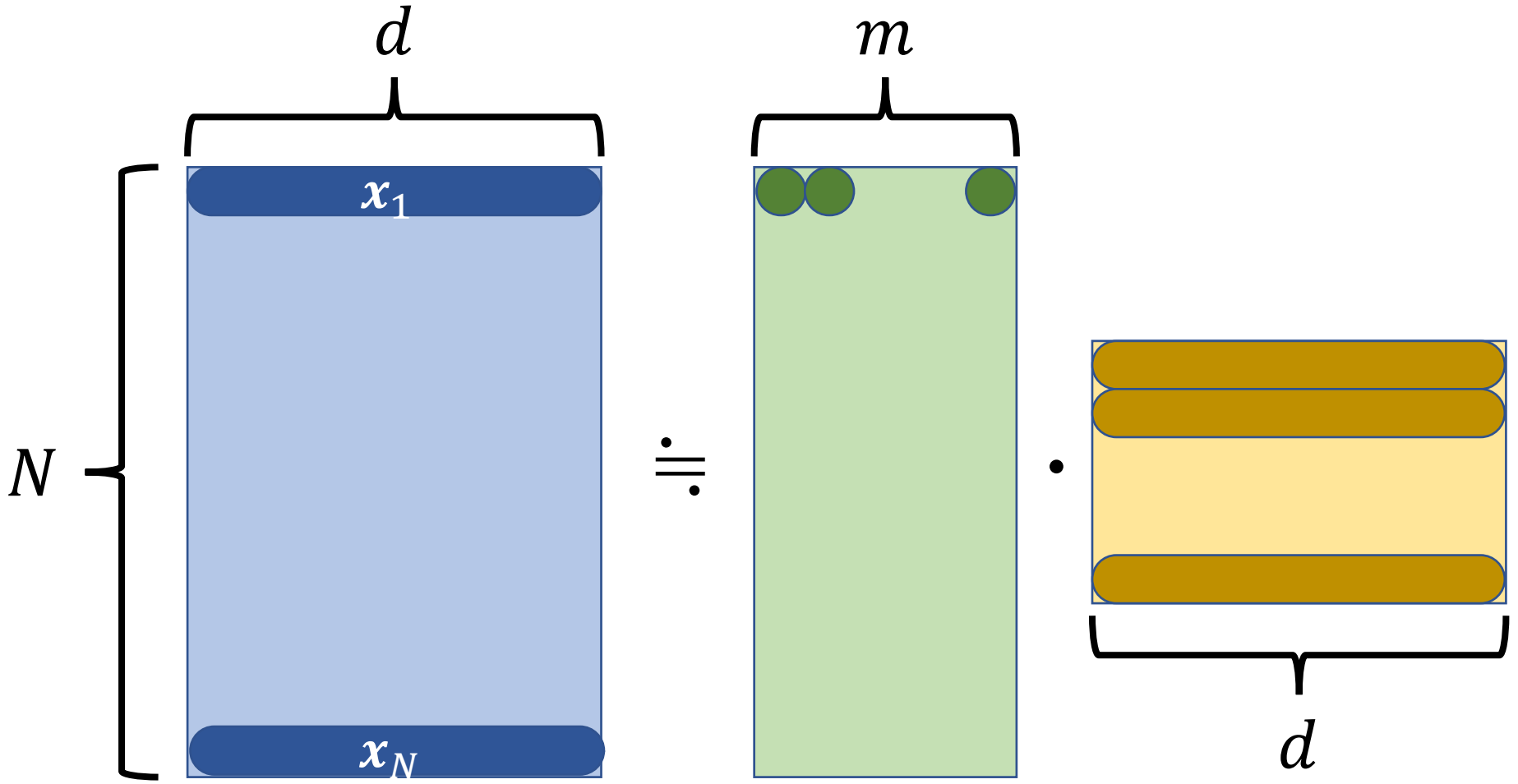
1. データを中心化する（平均を引く）
2. 共分散行列$\boldsymbol{C} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$を計算する
3. 共分散行列$\boldsymbol{C}$の固有値と固有ベクトルを求める
4. 最も大きい$\boldsymbol{m}$個の固有値に対応する固有ベクトルを選ぶ

- このアルゴリズムで求まる固有ベクトルと、「最も広く散らばった方向」が同じになることは、ちゃんと証明できる
  - 等式制約の場合のラグランジュ未定乗数法を使う。

http://people.cs.pitt.edu/~iyad/PCA.pdf

PCA, NMF, PLSIなど

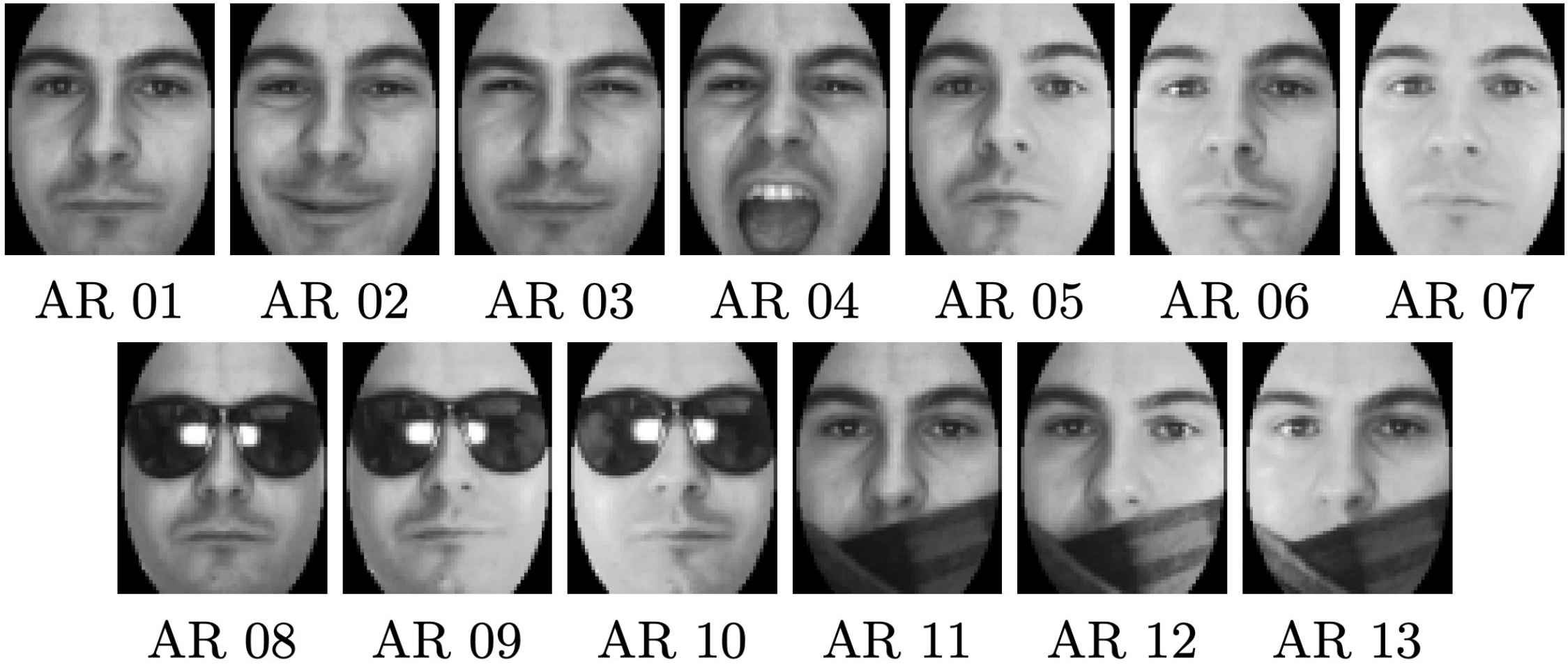
- 計画行列を、小さい2つの行列の積で表現し直す



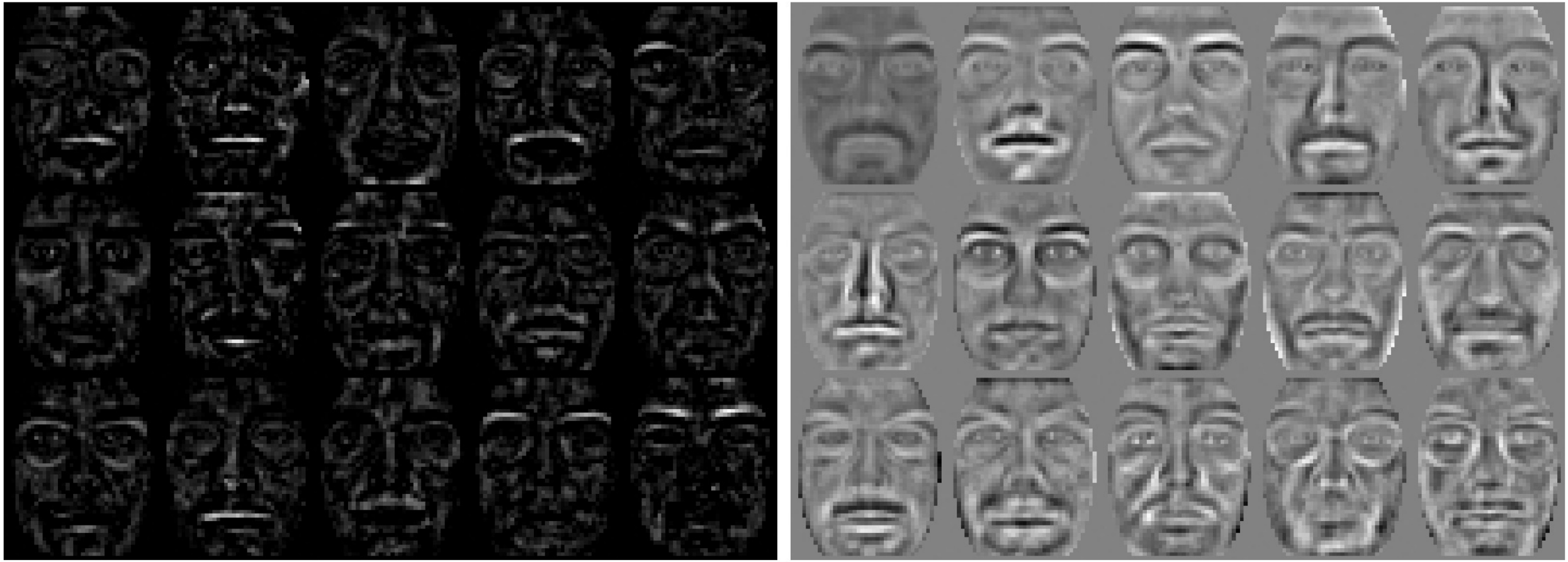
NMF (Nonnegative Matrix Factorization)

- PCAの問題点
  - 元々どの属性値も負の値を取り得ないデータには適用しにくい
  - 主成分は負の値を含みうるが、このような負の値は解釈が難しい

- NMFの特徴
  - PCAでいう主成分に相当するベクトルが非負ベクトルとして得られる
  - 得られる非負ベクトルは、お互いに直交しないこともありうる
    - ただし、疎なベクトルになるので、ほぼ直交しているともみなせる。



http://xrm.phys.northwestern.edu/research/pdf_papers/2002/guillamet_lectnotcompsci_2002.pdf



(a) NMF bases. (b) PCA bases.

Fig. 2. Bases obtained by both techniques, PCA and NMF

http://xrm.phys.northwestern.edu/research/pdf_papers/2002/guillamet_lectnotcompsci_2002.pdf

PLSI (probabilistic latent semantic indexing)

- NMFの確率モデル版
- 詳細は来年夏学期の「統計モデリング２」で
- EMアルゴリズムでパラメータを推定する

LDA (latent Dirichlet allocation)

- PLSIをベイズ化した確率モデル
- 詳細は来年夏学期の「統計モデリング２」で
- 変分ベイズ推定で事後分布のパラメータを推定する

課題11

- word2vecという手法を使うと、単語をベクトル化できる
  - word2vecの詳細は、秋学期の「自然言語処理特論」で。
- 日本語の単語について、word2vecでベクトルを求めたデータセットがあるので、このデータセット上でクラスタリングをおこなってみる
- 各クラスタに属する単語ベクトルを、クラスタの重心に近い順に並べたとき、上位に来る単語がどんな単語かを見てみる
  - クラスタ重心に近い単語を、簡単に「重要語」と呼ぶことにする。
- クラスタリングの手法を変えたり、クラスタ数を変えたりしたとき、各クラスタの重要語がどう変化するか、調べてみる