

データサイエンス 第4回

～多変量解析・次元削減～

情報理工学系研究科
創造情報学専攻
中山 英樹

本日の内容

- 多変量解析
 - 次元圧縮
 - 目的変数なしの場合
 - 主成分分析、LLE、MDSなど
 - 目的変数ありの場合
 - 判別分析、LFDAなど
-

本題に入る前に

- データの種類にはいろいろあり、**尺度**を意識することが重要

データの種類	尺度の種類	尺度の意味	可能な計算	例
量的データ	比尺度	原点(0という値)と比率に意味がある	＋、－、×、÷	身長、体重、金額
	間隔尺度	値の間隔に意味がある	＋、－	知能指数
質的データ	順序尺度	順序に意味がある	度数, 最頻値, 中央値	マラソンの順位
	名義尺度	区別するだけ	度数, 最頻値	性別、血液型

例えば…

- 5段階評価のアンケート

(1)悪い (2)やや悪い (3)ふつう (4)良い (5)とても良い

- 順序尺度。平均に意味はあるか？
 - 正しくデータを表す代表値となるかは不明
- カテゴリをつけず“5点満点”なら比尺度？

多変量解析とは

- 大規模、高次元なデータから本質的な情報（できれば低次元）を抽出するための統計的手法群の総称
 - 目的変数がない場合

説明変数	手法
量的データ(比尺度)	主成分分析、因子分析
量的データ(間隔尺度)	クラスター分析、多次元尺度構成法、数量化Ⅳ類
質的データ	数量化Ⅲ類、対応分析

- 目的変数がある場合

目的変数	説明変数	手法
量的データ	量的データ	回帰分析、正準相関分析
	質的データ	数量化Ⅰ類
質的データ	量的データ	判別分析
	質的データ	数量化Ⅱ類

ダミー変数

ダミー変数

多変量解析による次元圧縮

- 生データは一般に極めて高次元
 - 例) 文書、画像 数十万~数百万次元
 - **次元の呪い**：適切な学習が難しくなる
 - 人間にとっても意味が掴みにくい（可視化できない）
- 実際のデータは冗長であり、本質的に重要な構造は低次元で表現できる（場合が多い）

主成分分析：Principal Component Analysis (PCA)

- p次元の特徴ベクトル $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ を、元のデータの構造をできるだけ保ったまま低次元へ圧縮したい

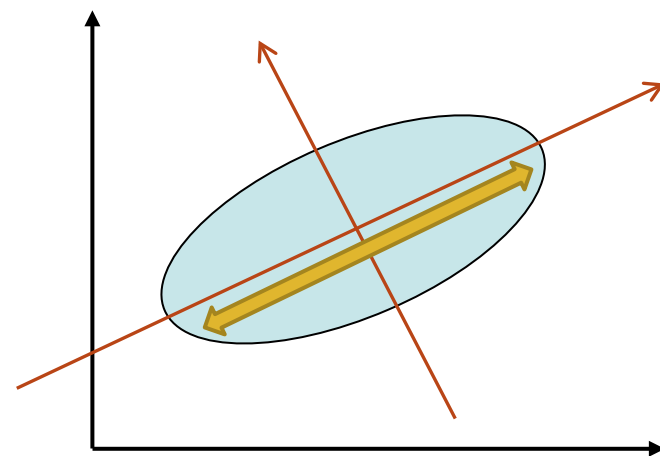
線形射影： $z = a_1x_1 + a_2x_2 + \dots + a_px_p = \mathbf{a}^T \mathbf{x}$ (ただし $\mathbf{a}^T \mathbf{a} = 1$)

- データの分布を最もよく記述する軸は？

⇒ 分散最大基準

$$\text{var}(z) = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2$$

を最大化する \mathbf{a} を求めたい



PCA : 分散最大基準による導出

$$\begin{aligned}\text{var}(z) &= \frac{1}{n} \sum_{i=1}^n (z_i - z)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \bar{\mathbf{x}})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \bar{\mathbf{x}})(\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \bar{\mathbf{x}})^T \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{a}^T (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{a} \\ &= \mathbf{a}^T \left(\frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \right) \mathbf{a} \\ &= \mathbf{a}^T \underline{C_X} \mathbf{a}\end{aligned}$$

\mathbf{x} の共分散行列

$$J_{PCA} = \mathbf{a}^T C_X \mathbf{a} \text{ を}$$

$$\mathbf{a}^T \mathbf{a} = 1 \text{ のもとで最大化}$$



$$J'_{PCA} = \mathbf{a}^T C_X \mathbf{a} - \lambda (\mathbf{a}^T \mathbf{a} - 1) \text{ を最大化}$$

(ラグランジュの未定乗数法)

$$\frac{\partial J'_{PCA}}{\partial \mathbf{a}} = 2C_X \mathbf{a} - 2\lambda \mathbf{a} = 0 \text{ (停留点)}$$



$$C_X \mathbf{a} = \lambda \mathbf{a}$$

※行列の微分についてはmatrix cookbook等を参照

<http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>

PCA : 平均二乗誤差最小基準による導出

主成分空間に射影した点の元の空間における座標は

$$\hat{\mathbf{x}}_i = z_1 \mathbf{a}_1 + z_2 \mathbf{a}_2 + \cdots + z_m \mathbf{a}_m = \sum_{j=1}^m \mathbf{a}_j^T \mathbf{x}_i \mathbf{a}_j$$

$$\varepsilon^2(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2$$

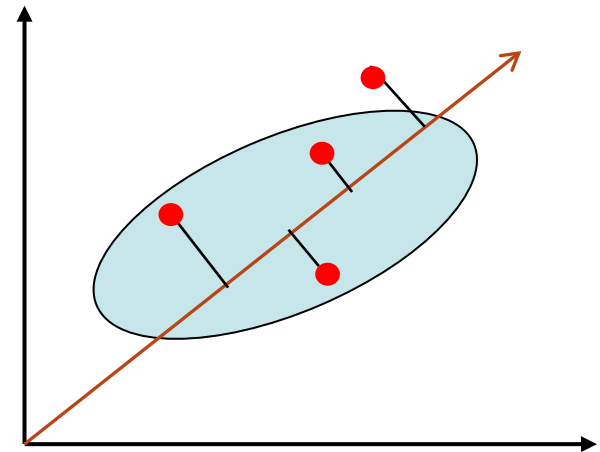
$$= \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i - \sum_{j=1}^m \mathbf{a}_j^T \mathbf{x}_i \mathbf{a}_j \right\|^2$$

$$= \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \sum_{j=1}^m \mathbf{w}_j^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w}_j$$

$$= \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \sum_{j=1}^m \mathbf{w}_j^T R_X \mathbf{w}_j$$

定数

結局こちらを最大化



自己相関行列 $R_X = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$
の固有値問題に帰着

$$R_X \mathbf{a} = \lambda \mathbf{a}$$

PCA : つづき

- 複数の無相関な軸が、固有値に対応する固有ベクトルとして得られる
 - 固有値の大きさがその軸（固有ベクトル）におけるデータの分散の大きさに対応
- 累積寄与率を参考に主成分（固有ベクトル）の数を決める
 - i番目の主成分の寄与率：
$$\lambda_i / \sum_{j=1}^p \lambda_j$$
 - m番目の主成分までの累積寄与率：
$$\sum_{j=1}^m \lambda_j / \sum_{j=1}^p \lambda_j$$

(固有値は降順にならんでいるものとする)

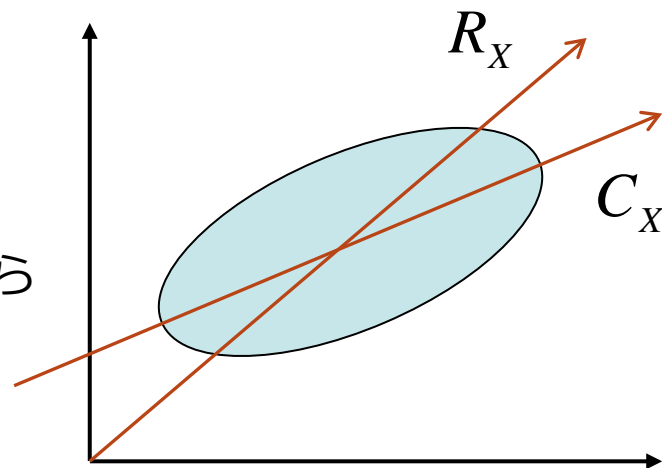
注意

- 共分散行列、自己相関行列、相関係数行列と、それぞれの固有値問題で張られる部分空間の違いに注意

- 自己相関行列（相関係数行列ではない！）

$$R_X = C_X + \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

- 二乗誤差基準で導出した場合は、一般にはこちら
 - 座標原点を中心に分散を見た場合に相当
 - 特徴に非負制約がある場合に有効
- 最初にデータから平均を引いておけば分かりやすい
 - ただし、いつもそれが適切とは限らない



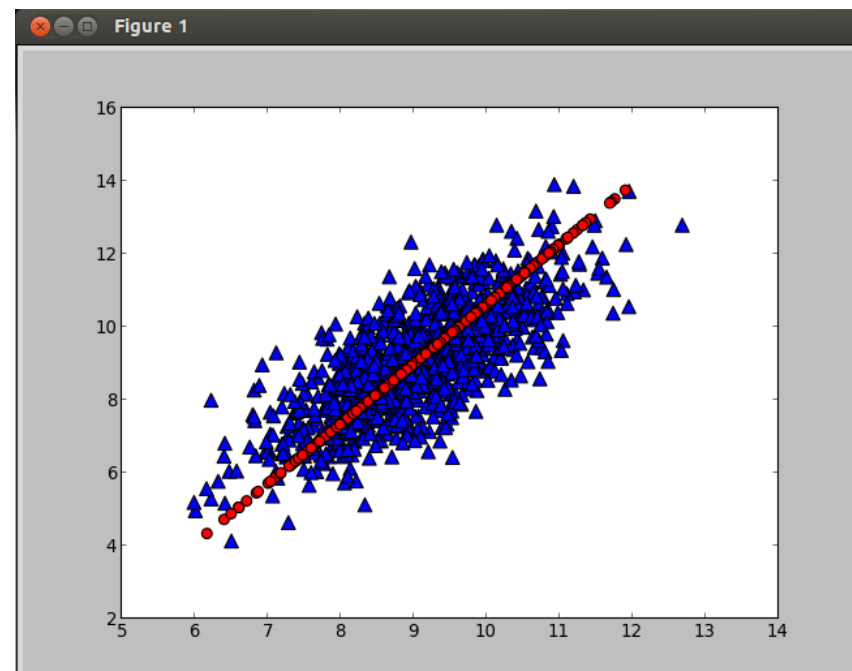
（参考） 相関係数行列

- 元データの各特徴を平均0、分散1に正規化したあとの共分散行列に等しい

サンプル

```
>>> python pca_test1.py
```

```
dataMat = pca.loadDataSet('testSet.txt')  
lowDMat, reconMat = pca.pca(dataMat, 1)
```



pca.pca(dataMat, 1)の第二引数を2に変えたらどうなるか？

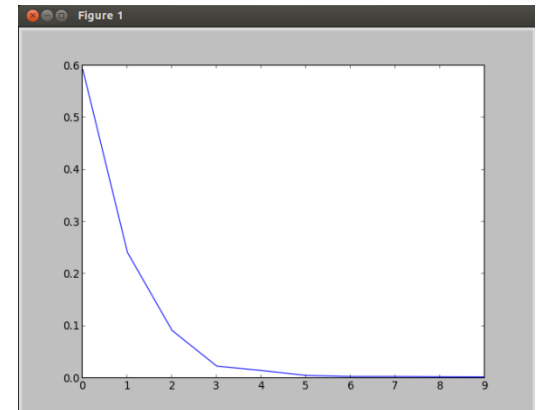
SECOM dataset

- 半導体製造ラインのモニタリングデータ
 - 故障の早期発見などが目的
 - UCI Machine Learning Repository
<http://archives.ics.uci.dcu/ml/datasets/SECOM>
- 590次元、欠損値多数

```
>>> import numpy as np
>>> import pca
>>> dataMat = pca.replaceNaNWithMean() #pca.py参照 (pandasで書いた方がよい?)
>>> meanVals = np.mean(dataMat, axis=0)
>>> meanRemoved = dataMat - meanVals
>>> covMat = np.cov(meanRemoved, rowvar=0)
>>> eigVals,eigVecs = np.linalg.eig(np.mat(covMat))

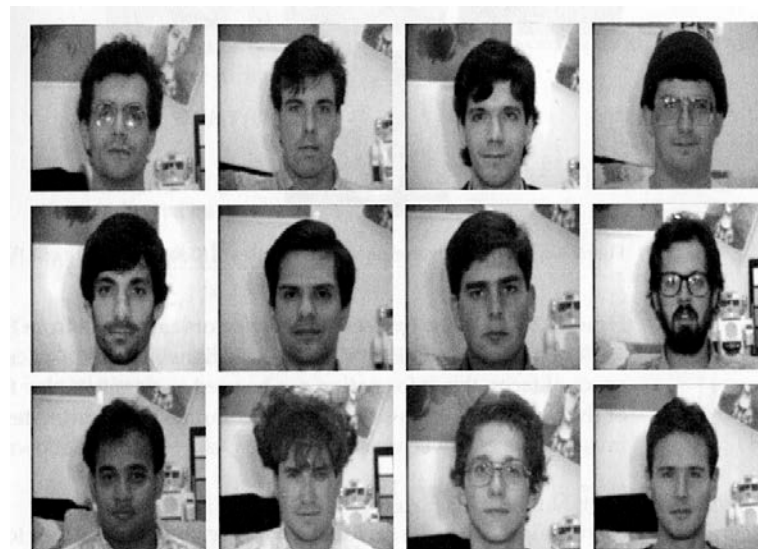
>>> eigVals

>>> import matplotlib
>>> import matplotlib.pyplot as plt
>>> plt.plot(eigVals[:10]/np.sum(eigVals)) #上位10主成分までの寄与率を表示
>>> plt.show()
```



画像認識への応用

- 固有空間 = 固有ベクトルが張る空間
- 学習サンプルから固有空間を求める
- 画像はベクトル1個（画素値）



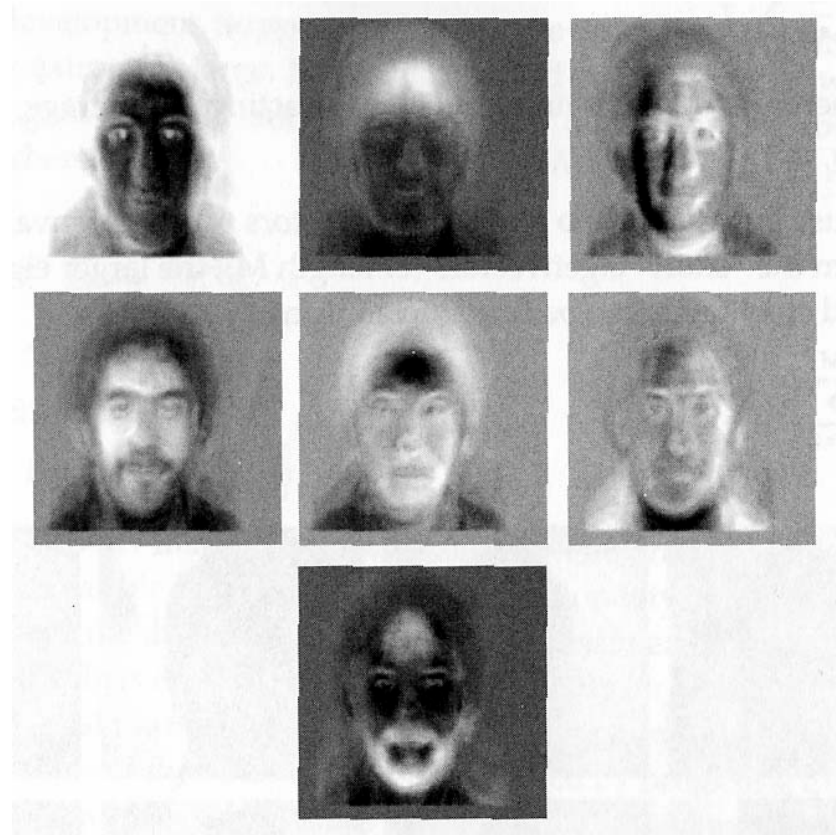
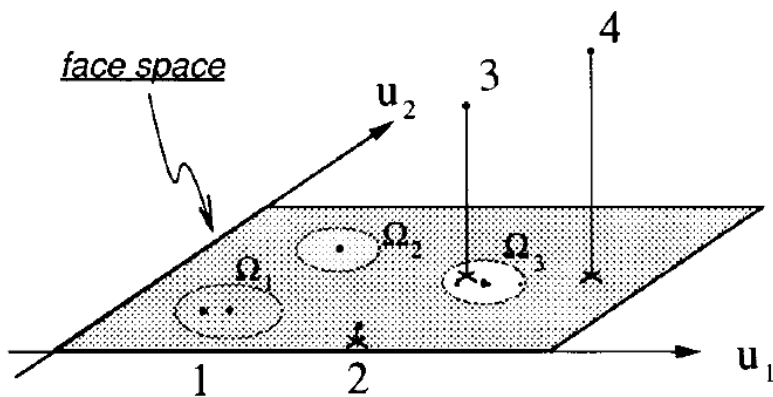
[Turk & Pentland, CVPR1991]



平均顔

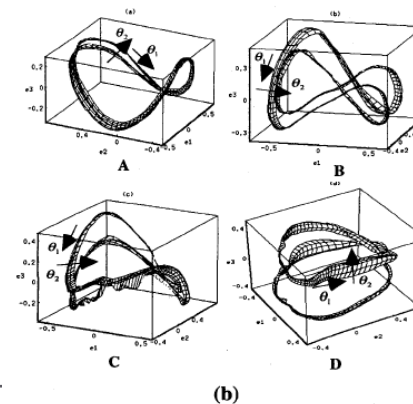
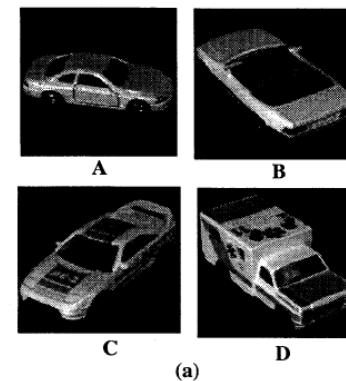
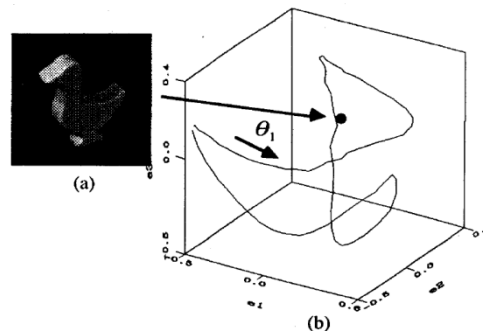
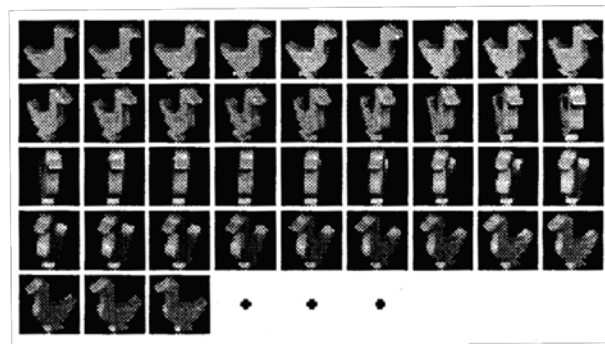
Eigenfaces

- 最初の7個の固有ベクトル
- 識別：
 - 入力画像を固有空間に投影
 - 最も近いクラスを求める



パラメトリック固有空間法 [Murase,1995]

- 各物体の様々な像から固有空間を構成
- 物体の姿勢, 光源の位置を固有空間上の最近傍点から推定 (平均角度誤差1.2度)
- 物体ごとの一連の見えの変化を多様体に変換 (補間し滑らかに)



統計的手法に基づく動画像からの 異常動作の検出

南里卓也、大津展之

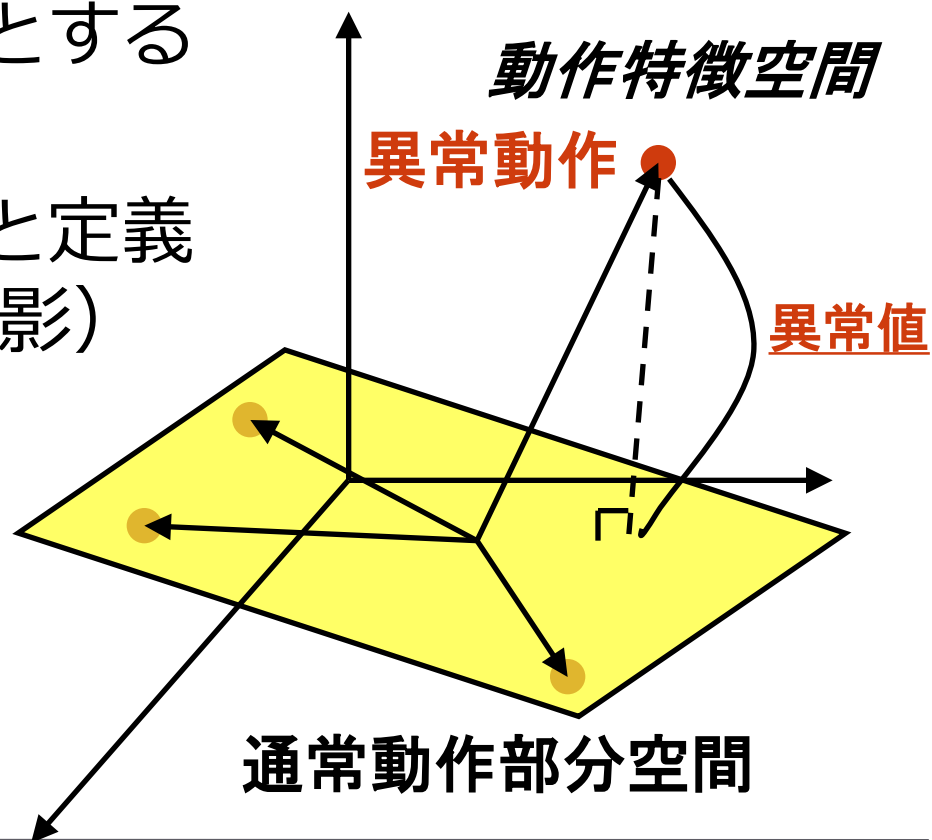
- 頻繁に起こる動作の部分空間をPCAで学習
- そこからの逸脱として、異常動作を検出



縦軸：
異常動作値

異常検知手法

- 動作特徴空間に
通常動作の部分空間を構成
- そこからの逸脱を**異常**とする
- 通常部分空間への
垂直距離として**異常値**と定義
(直交部分空間への射影)
- **部分空間法**

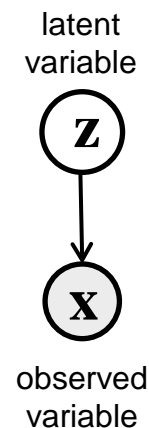


確率的バックグラウンド

- Probabilistic PCA

- 最尤推定で解くと、通常のPCAの解と一致
- 潜在変数 \mathbf{z} には回転の自由度がある

$$\begin{aligned}\mathbf{z} &\sim N(0, I_d), \quad p \geq d \geq 1 \\ \mathbf{x} | \mathbf{z} &\sim N(W\mathbf{z} + \mu, \sigma^2 I), \quad W \in \mathbf{R}^{p \times d}\end{aligned}$$



- 主成分分析と因子分析は基本的に同じ構造

- 因子分析は潜在空間上で回転を行い、解釈がしやすい軸を探す
- 観測データと構造のどちらの視点から見るかの違い

実装上のTips

- データを全部メモリに読み込む必要はない
 - 分散、平均だけ先に計算して固有値問題を解く
 - 順にデータを読み込んで射影する（オフセットに注意）
- データが高次元の場合は工夫が必要
 - 計算コストは基本的に次元数の3乗に比例
 - 普通に解けるのは一万次元くらいまで
 - 疎行列なら専用の解法がある（必要な数だけ上位の固有ベクトルを計算）
e.g. `scipy.sparse.linalg.eigs`
 - より一般にはBishopのPRML本など参照のこと

Locally linear embedding (LLE) [Roweis & Saul, 2000]

- PCAはデータの分布の非線形構造をつぶしてしまう
- LLEでは局所構造を保存した圧縮を行う
- ポイント：局所的には線形な構造を持っているとみなせる
 - 近傍サンプルの重みづけ和で表せる

$$\hat{\mathbf{x}}_i = \sum_{j \in N(i)} W_{ij} \mathbf{x}_j \quad \left(\sum_j W_{ij} = 1 \right)$$

j の近傍サンプル

LLE概要

- 1. 各サンプルの二乗誤差を最小とする W を求める
(解析的に計算できる)

$$\mathcal{E}_i = \left\| \mathbf{x}_i - \sum_{j \in N(i)} W_{ij} \mathbf{x}_j \right\|^2$$

- 2. 求まった W のもとで、同じ基準で誤差を最小とするように低次元のベクトル \mathbf{y} を設定する

$$\begin{aligned} & \sum_{i=1}^n \left\| \mathbf{y}_i - \sum_{j \in N(i)} W_{ij} \mathbf{y}_j \right\|^2 \\ &= \text{tr}(\mathbf{Y}^T \mathbf{M} \mathbf{Y}) \quad \text{ただし} \quad \mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W}), \quad \mathbf{Y}^T \mathbf{Y} = \mathbf{I} \\ & \quad \mathbf{Y} \text{ の第 } i \text{ 列が } \mathbf{y}_i \end{aligned}$$

解き方

- 学習サンプル数次元の固有値問題

- PCAでは

$\text{tr}(A^T C_X A)$ を $A^T A = I$ のもとで最大化 $\Rightarrow C_X$ の固有値の大きい順に固有ベクトルを選択

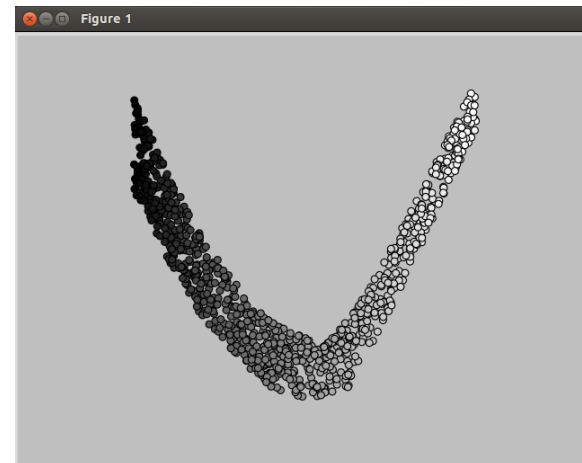
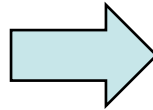
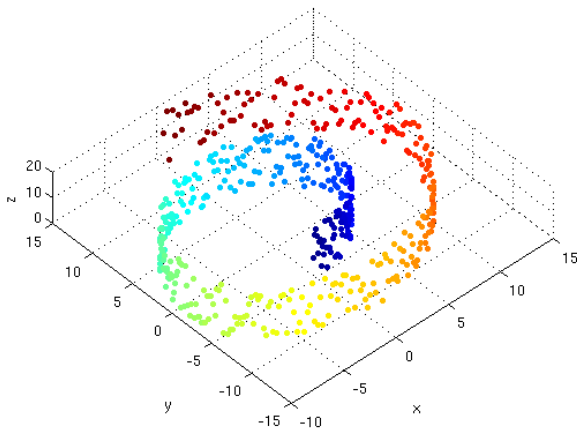
- LLEは

$\text{tr}(Y^T M Y)$ を $Y^T Y = I$ のもとで最小化 $\Rightarrow M$ の固有値の小さい順に固有ベクトルを選択

※ただし、原理的に最小の固有値は常にゼロ
(意味のないベクトル) となるので除外する

サンプル

```
>>> python lle.py
```



swiss roll のデータの隣接構造を崩さずに、2次元座標系へ埋め込みができてい
る（PCAでは不可能）

注意

- サンプル数次元の固有値問題 = 大変
- ただし、 $M = (I - W)^T (I - W)$ はスパースな行列になる（はず）

(古典的) 多次元尺度構成法: multi dimensional scaling (MDS)

- サンプル間の距離 (類似度) をできるだけ保存するように低次元への埋め込みを行う
- 距離が定義されていれば適用可能 (間隔尺度)

$s_{ii'}$ を二つのサンプル i, i' の類似度とする

(元の特徴空間での座標が分かっている場合、

$$s_{ii'} = (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_{i'} - \bar{\mathbf{x}}) \text{ などとして定義してもよい})$$

$$\sum_{i \neq i'} \left(s_{ii'} - (\mathbf{z}_i - \bar{\mathbf{z}})^T (\mathbf{z}_{i'} - \bar{\mathbf{z}}) \right)^2$$

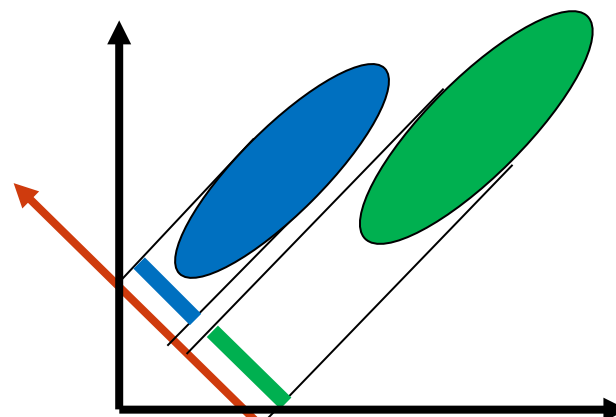
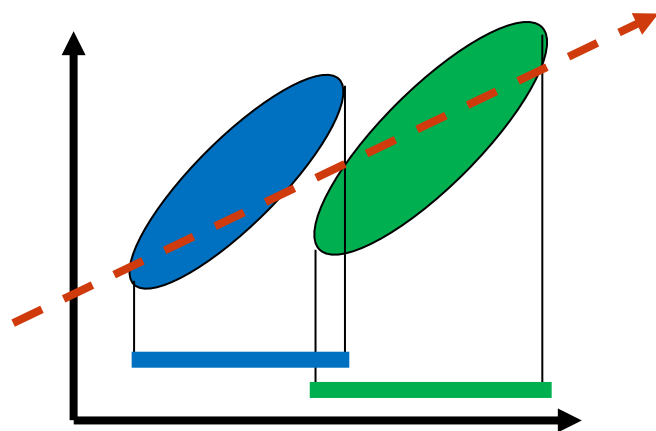
を最小とするように低次元の表現 \mathbf{z} へ各サンプルを配置する

線形判別分析：Fisher Discriminant Analysis (FDA)

- クラス（カテゴリ）のサンプルを最もよく分離する軸を見つける

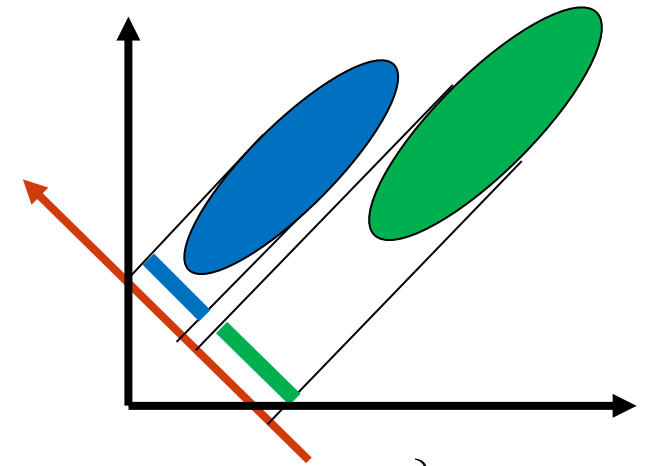
$$\text{線形射影} : z = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p = \mathbf{w}^T \mathbf{x}$$

- 同じクラス内のサンプルは互いに近くに集まり、異なるクラス同士のサンプルは遠く離れるように



FDA

- クラス内分散
 - クラスごとのサンプルの分散の平均



$$\begin{aligned}
 \underbrace{\sum_{i=1}^{N_c}}_{\text{クラス数}} \sum_{x \in \text{class}(i)} (\underbrace{\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \bar{\mathbf{x}}_i}_{\text{クラス}i\text{の平均ベクトル}}) (\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \bar{\mathbf{x}}_i)^T &= \mathbf{w}^T \left\{ \sum_{i=1}^{N_c} \sum_{x \in \text{class}(i)} (\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)^T \right\} \mathbf{w} \\
 &= \mathbf{w} \underline{C_W} \mathbf{w} \\
 &\quad \text{クラス内共分散行列}
 \end{aligned}$$

- クラス外分散
 - クラスの平均ベクトルの分散

$$\begin{aligned}
 \sum_{i=1}^{N_c} \underbrace{n_i}_{\text{クラス}i\text{のサンプル数}} (\underbrace{\mathbf{w}^T \bar{\mathbf{x}}_i - \mathbf{w}^T \bar{\mathbf{x}}}_{\text{全平均ベクトル}}) (\mathbf{w}^T \bar{\mathbf{x}}_i - \mathbf{w}^T \bar{\mathbf{x}})^T &= \mathbf{w}^T \left\{ \sum_{i=1}^{N_c} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \right\} \mathbf{w} \\
 &= \mathbf{w} \underline{C_B} \mathbf{w} \\
 &\quad \text{クラス外共分散行列}
 \end{aligned}$$

なお、全分散 = クラス内分散 + クラス外分散 となる ($C_X = C_W + C_B$)

FDA

- クラス内分散をできるだけ小さく、クラス外分散をできるだけ大きく → 比を最大化

$$J_{FDA} = \frac{\mathbf{w}C_B\mathbf{w}}{\mathbf{w}C_W\mathbf{w}} \quad \text{Fisher's discriminant criterion}$$

分子を固定（ $\mathbf{w}C_W\mathbf{w} = 1$ ）し、分母を最大化

$$J'_{FDA} = \mathbf{w}C_B\mathbf{w} - \lambda(\mathbf{w}C_W\mathbf{w} - 1)$$

\mathbf{w} で偏微分して整理すると

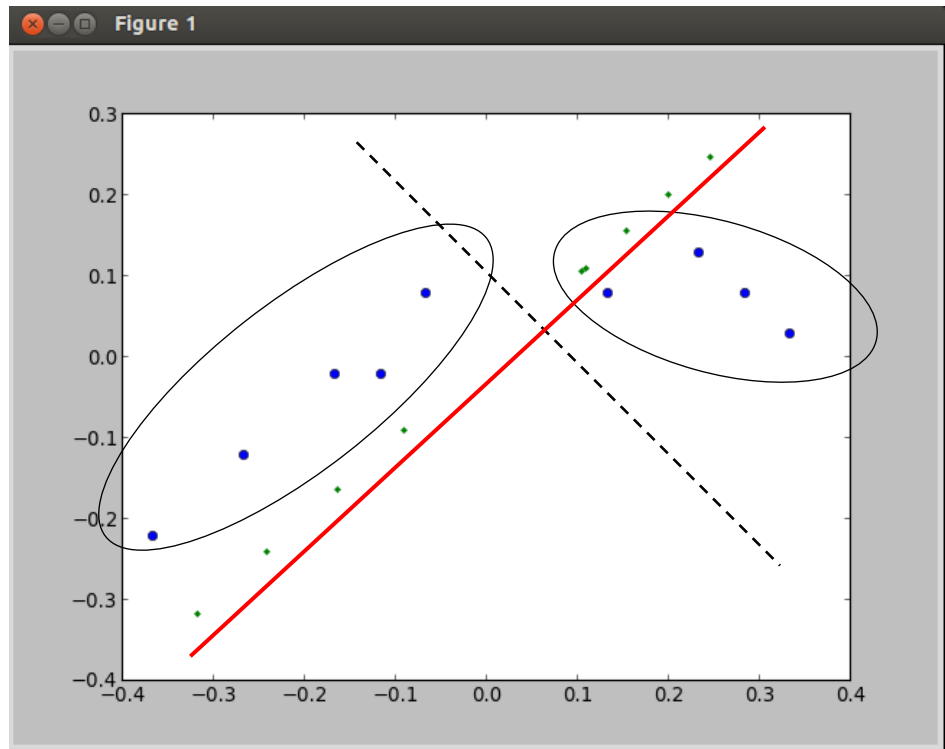
$$C_B\mathbf{w} = \lambda C_W\mathbf{w}$$

一般化固有値問題の解

固有値（フィッシャー基準の値）の大きい順に
固有ベクトルを用いる

サンプル

```
>>> python lda.py
```

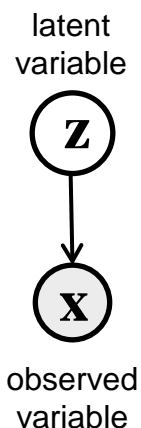


FDA：注意

- (クラス数 - 1)個しか軸（固有ベクトル）は求まらない
 - クラス内共分散行列のランクの問題
 - それ以上特徴が欲しい場合は、直交補空間に順次射影していくなど工夫が必要
 - あるいは、partial least squaresなど別の方法を用いる（後述）
- クラス外共分散行列のランクに注意
 - サンプル数が少ないと不安定になる

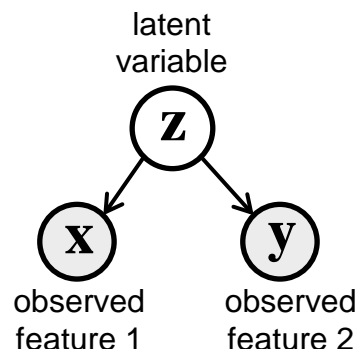
正準相関分析: Canonical Correlation Analysis (CCA)

- 二つの変量（量的データ）の間の潜在的な相関を発見する手法
- 対称な構造（どちらも互いに説明変数・目的変数）
- 主成分分析の二変量版



$$\begin{aligned}\mathbf{z} &\sim N(0, I_d), \quad p \geq d \geq 1 \\ \mathbf{x} | \mathbf{z} &\sim N(W\mathbf{z} + \mu, \sigma^2 I), \quad W \in \mathbf{R}^{p \times d}\end{aligned}$$

Probabilistic interpretation of PCA



$$\begin{aligned}\mathbf{z} &\sim N(0, I_d), \quad \min\{p, q\} \geq d \geq 1 \\ \mathbf{x} | \mathbf{z} &\sim N(W_x \mathbf{z} + \mu_x, \psi_x), \quad W_x \in \mathbf{R}^{p \times d} \\ \mathbf{y} | \mathbf{z} &\sim N(W_y \mathbf{z} + \mu_y, \psi_y), \quad W_y \in \mathbf{R}^{q \times d}\end{aligned}$$

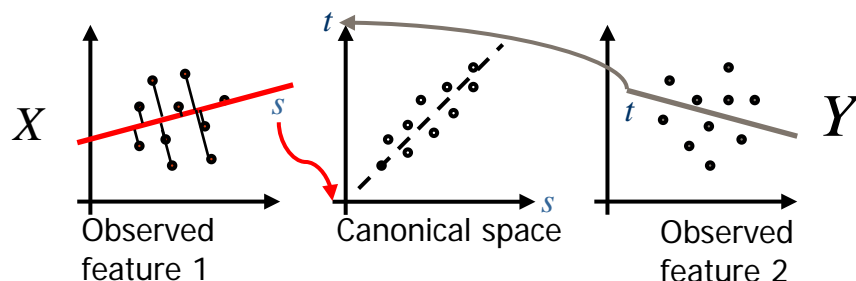
Probabilistic interpretation of CCA
[Bach and Jordan, 2005]

CCA

\mathbf{x} , \mathbf{y} : 2 種類の対応するデータ (e.g., 画像とテキストタグ)

線形変換 $s = \mathbf{a}^T (\mathbf{x} - \bar{\mathbf{x}})$, $t = \mathbf{b}^T (\mathbf{y} - \bar{\mathbf{y}})$ を、

s と t の相関が最大となるように決定する



ちなみに...

y をカテゴリラベルを数量化したベクトルにした場合、線形判別分析と一致する。

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \quad \mathbf{a}^T C_{XX} \mathbf{a} = 1, \mathbf{b}^T C_{YY} \mathbf{b} = 1$$

(導出は省略)

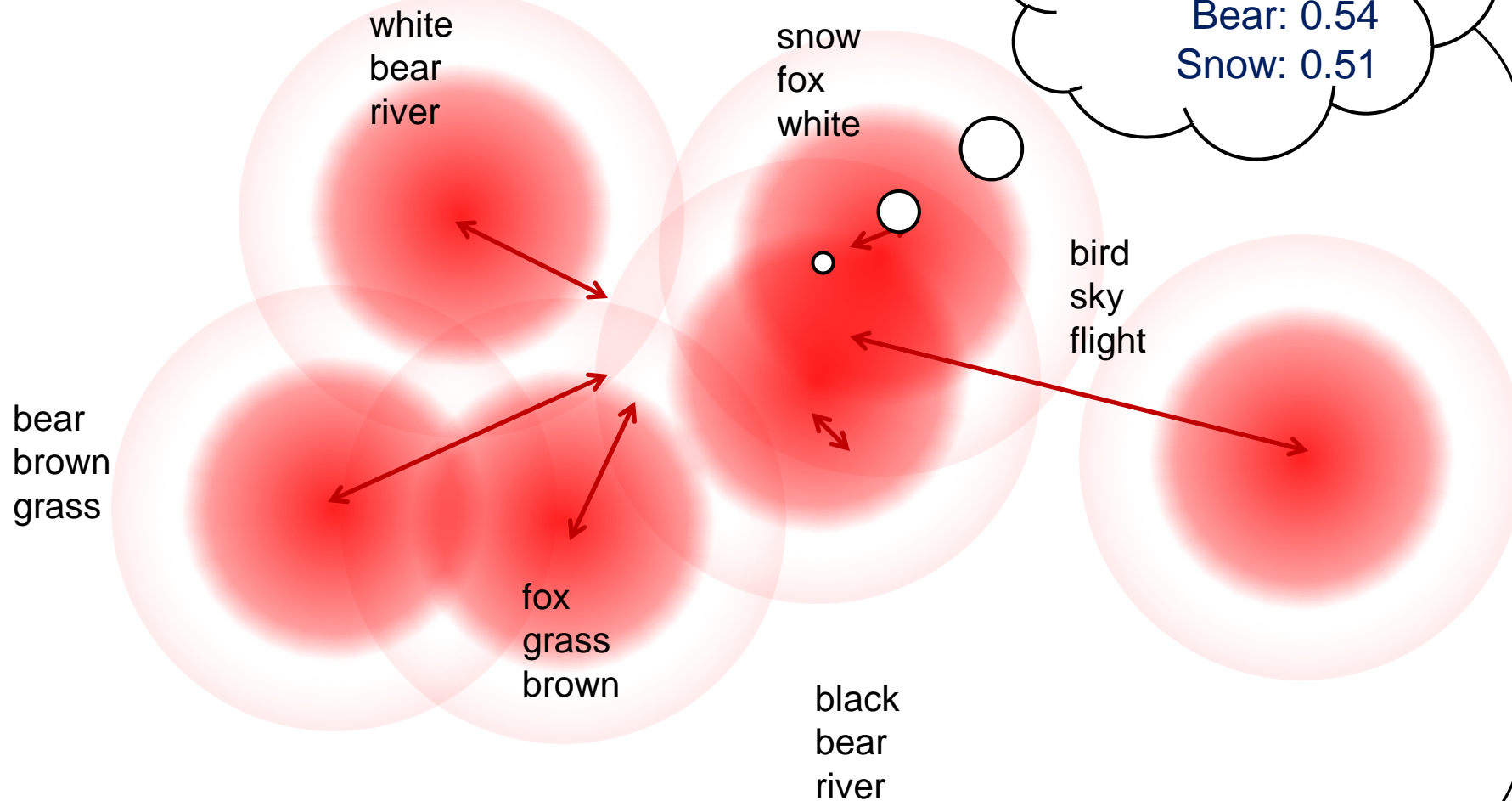
C : 共分散行列

λ : 正準相関係数

(例) 画像とテキストタグでCCA、次元圧縮



Fox: 0.90
White: 0.83
River: 0.54
Bear: 0.54
Snow: 0.51



類似手法との関係

A Unified Approach to PCA, PLS, MLR and CCA [Borga et al.]

- PCA

$$C_{XX}\mathbf{a} = \lambda\mathbf{a} \quad \mathbf{a}^T\mathbf{a} = 1$$

- PLS (partial least squares)

- 二変量間の共分散を最大化

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \quad \mathbf{a}^T\mathbf{a} = 1, \mathbf{b}^T\mathbf{b} = 1$$

- MLR (multiple linear regression)

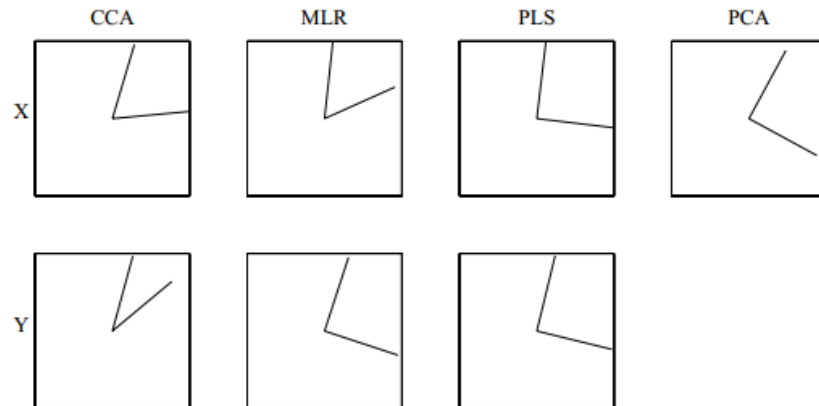
- 目的変数 y をできるだけ復元 (回帰)

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{XX} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \quad \mathbf{a}^T C_{XX} \mathbf{a} = 1, \mathbf{b}^T \mathbf{b} = 1$$

- CCA (canonical correlation analysis)

- 二変量間の相関を最大化

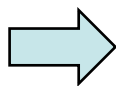
$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \quad \mathbf{a}^T C_{XX} \mathbf{a} = 1, \mathbf{b}^T C_{YY} \mathbf{b} = 1$$



数量化 I 類、II 類

- ダミー変数を用いた回帰分析、判別分析
 - 例)

売上	曜日		売上	日	月	火	水	木	金	土
10万円	火曜日		10万円	0	0	1	0	0	0	0
15万円	木曜日		15万円	0	0	0	0	1	0	0
8万円	日曜日		8万円	1	0	0	0	0	0	0



- 基本的に同じやり方でOKだが、データがスパースになりやすいので正規化（後の講義で解説予定）などに注意

対応分析、数量化Ⅲ類

- 見た目は異なるが実は同等の手法
- ダミー変数を用いた主成分分析（因子分析）と近い結果になる

	喫煙	飲酒	肺癌
被験者A	1	0	0
被験者B	1	1	1
被験者C	1	1	0

- クロス集計表のデータは、ダミー変数を用いた正準相関分析で（ある程度）解析可能

	喫煙	飲酒	肺癌
男性			
女性			
30代			

まとめ

- 再掲

- 目的変数がない場合

説明変数	手法
量的データ(比尺度)	主成分分析、因子分析
量的データ(間隔尺度)	クラスター分析、多次元尺度構成法、数量化Ⅳ類
質的データ	数量化Ⅲ類、対応分析

- 目的変数がある場合

目的変数	説明変数	手法
量的データ	量的データ	回帰分析、正準相関分析
	質的データ	数量化Ⅰ類
質的データ	量的データ	判別分析
	質的データ	数量化Ⅱ類

ダミー変数

ダミー変数