データサイエンス
第3・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：データ間の差異が測れなくなる
  - 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）

約束事

- 数式
  - 列ベクトルでデータが与えられることを前提

$$\mathbf{x} = \begin{pmatrix} 1.0 \\ 2.2 \\ -3.0 \\ 4.8 \end{pmatrix}$$

- コード
  - この講義では最初のうちは数式に揃えて列ベクトル前提とします
  - ただし、世の中の実際のコードは行ベクトル前提が普通
    - 主にメモリ配置の都合

$$\mathbf{x} = \begin{pmatrix} 1.0 & 2.2 & -3.0 & 4.8 \end{pmatrix}$$

準備：ベクトル、行列の微分

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_q \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

①ベクトル、行列をスカラーで微分

$$\frac{\partial \mathbf{x}}{\partial a} = \begin{pmatrix} \partial x_1 / \partial a \\ \vdots \\ \partial x_p / \partial a \end{pmatrix} \qquad \frac{\partial X}{\partial a} = \begin{pmatrix} \partial x_{11} / \partial a & \cdots & \partial x_{1n} / \partial a \\ \vdots & \ddots & \\ \partial x_{m1} / \partial a & \cdots & \partial x_{mn} / \partial a \end{pmatrix}$$

②スカラーをベクトル、行列で微分

$$\frac{\partial a}{\partial \mathbf{x}} = \begin{pmatrix} \partial a / \partial x_1 \\ \vdots \\ \partial a / \partial x_p \end{pmatrix} \qquad \frac{\partial a}{\partial X} = \begin{pmatrix} \partial a / \partial x_{11} & \cdots & \partial a / \partial x_{1n} \\ \vdots & \ddots & \\ \partial a / \partial x_{m1} & \cdots & \partial a / \partial x_{mn} \end{pmatrix}$$

③ベクトルをベクトルで微分

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \partial y_1 / \partial x_1 & \cdots & \partial y_q / \partial x_1 \\ \vdots & \ddots & \\ \partial y_1 / \partial x_p & \cdots & \partial y_q / \partial x_p \end{pmatrix}$$

主成分分析：Principal Component Analysis（PCA）

- p次元の特徴ベクトル $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ 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

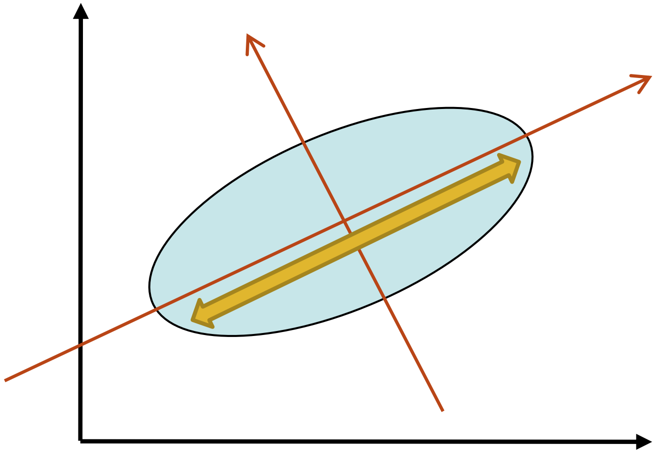
線形射影： $z = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p = \mathbf{a}^T \mathbf{x}$ （ただし $\mathbf{a}^T \mathbf{a} = 1$ ）

- データの分布を最もよく記述する軸は？

$\Rightarrow$ 分散最大基準

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

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

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

x の共分散行列

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

$\mathbf{a}^T\mathbf{a} = 1$ のもとで最大化

$\Updownarrow$

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

（ラグランジュの未定乗数法）

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

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

※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

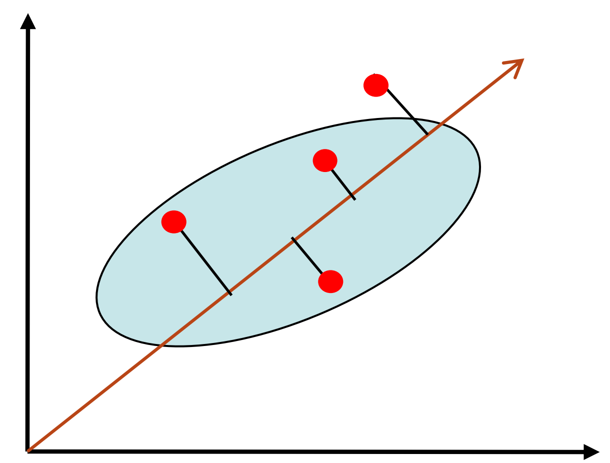
$$\hat{\mathbf{x}}_i = z_1\mathbf{a}_1 + z_2\mathbf{a}_2 + \cdots + z_m\mathbf{a}_m = \sum_{j=1}^{m}\mathbf{a}_j^T\mathbf{x}_i\mathbf{a}_j$$

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

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

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

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}_{\text{定数}} - \underbrace{\sum_{j=1}^{m}\mathbf{a}_j^T R_X\mathbf{a}_j}_{\text{結局こちらを最大化}}$$



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

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

PCA：つづき

- 複数の直交する軸が、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
  - 各軸上に射影されたデータは無相関

- 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：
  $$\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j$$
  - m番目の主成分までの累積寄与率：
  $$\sum_{j=1}^{m} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j$$

（固有値は降順にならんでいるものとする）

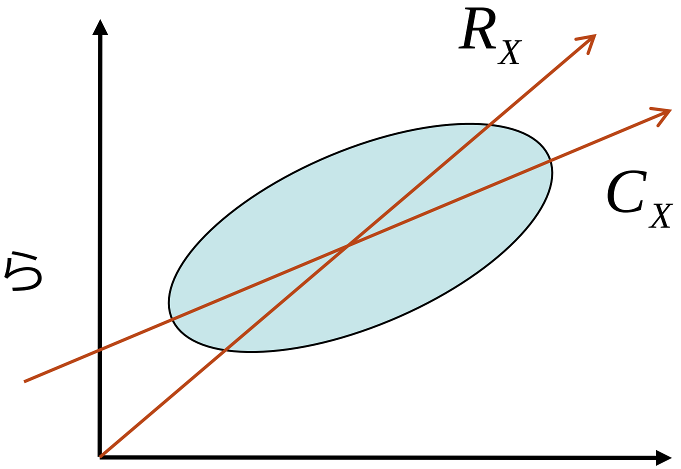
注意

- 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！）

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

  - 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（一致する）
  - 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

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

パターン認識への応用

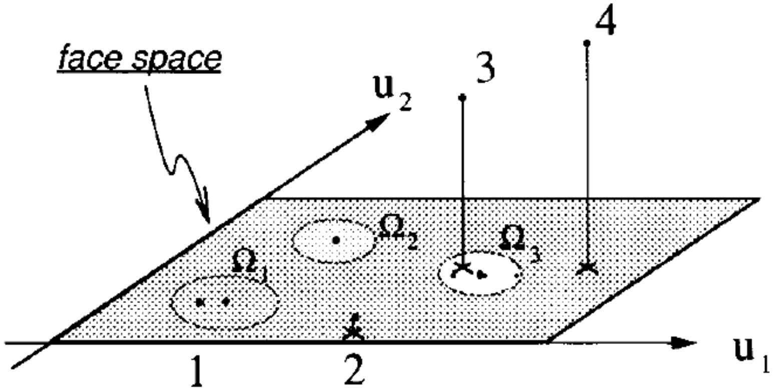
- 固有空間＝固有ベクトルが張る空間
- 学習サンプルから固有空間を求める
- 画像はベクトル1個（画素値）

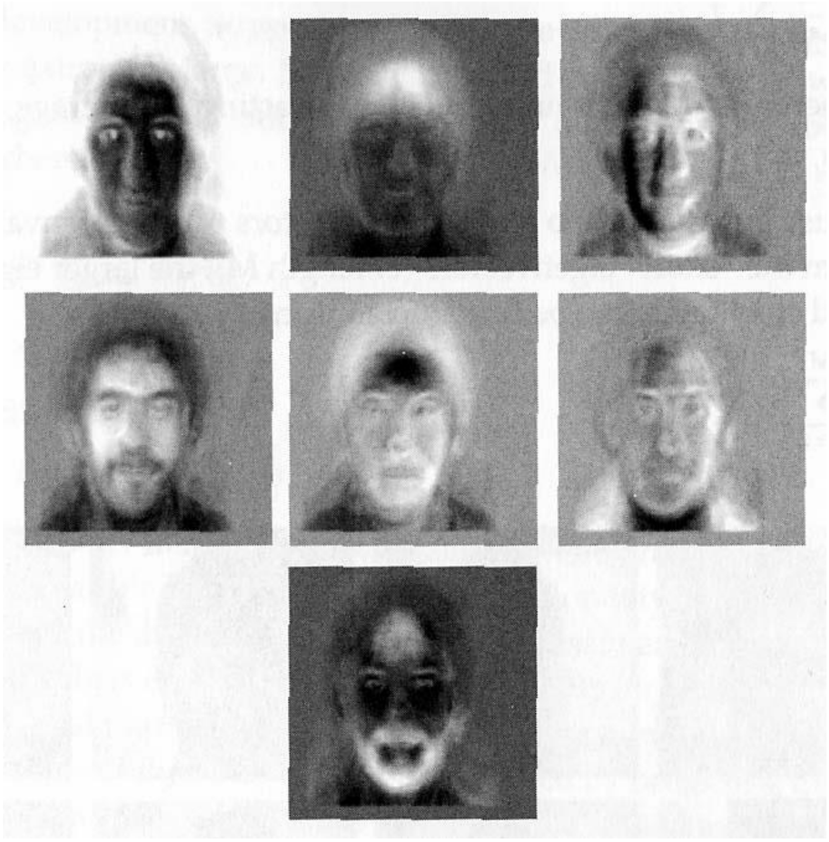
[Turk & Pentland, CVPR1991]

平均顔

Eigenfaces

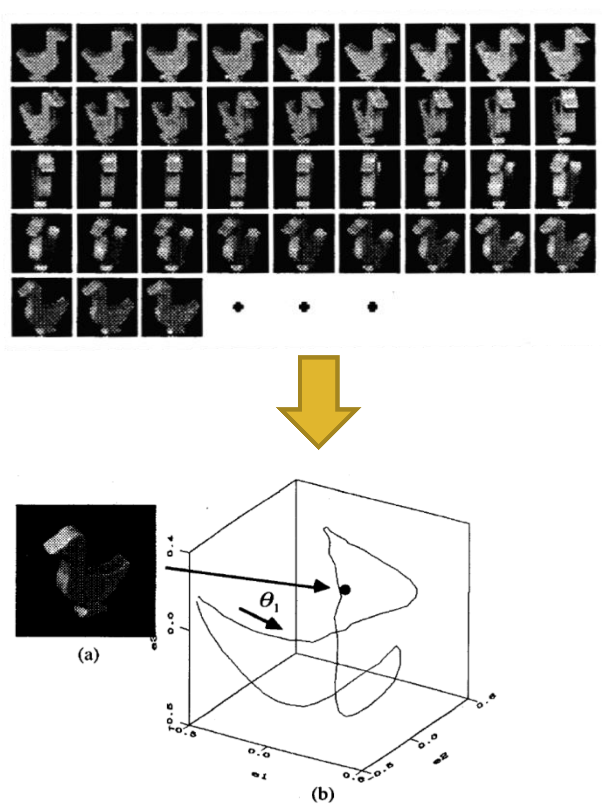
- 最初の7個の固有ベクトル
- 識別：
  - 入力画像を固有空間に投影
  - 最も近いクラスを求める

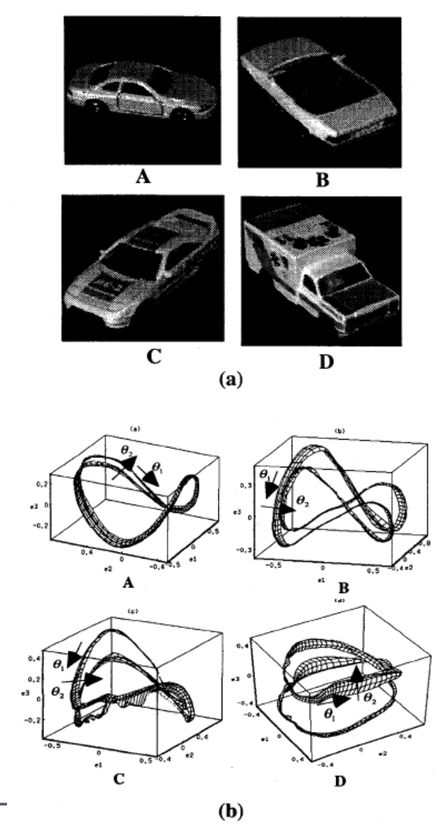




パラメトリック固有空間法 [Murase,1995]

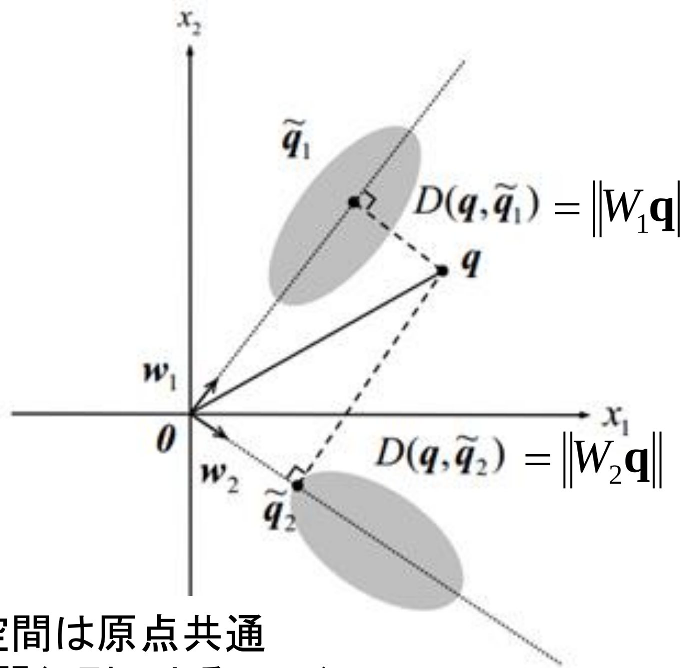
- 各物体の様々な像から固有空間を構成
- 物体の姿勢，光源の位置を固有空間上の最近傍点から推定（平均角度誤差1.2度)
- 物体ごとの一連の見えの変化を多様体に（補間し滑らかに）





部分空間法 (CLAFIC)

- 各クラスの成す部分空間 (PCAで学習）への近さを基準に識別
- 特徴が線形な構造を有している場合に有効
- 日本発の技術



距離最小基準: $\hat{C} = \arg\min_{C_i} \|W_i\mathbf{q}\|$

角度最小基準: $\hat{C} = \arg\min_{C_i} \frac{\|W_i\mathbf{q}\|}{\|\mathbf{q}\|}$

※部分空間は原点共通
（自己相関行列によるPCA）

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

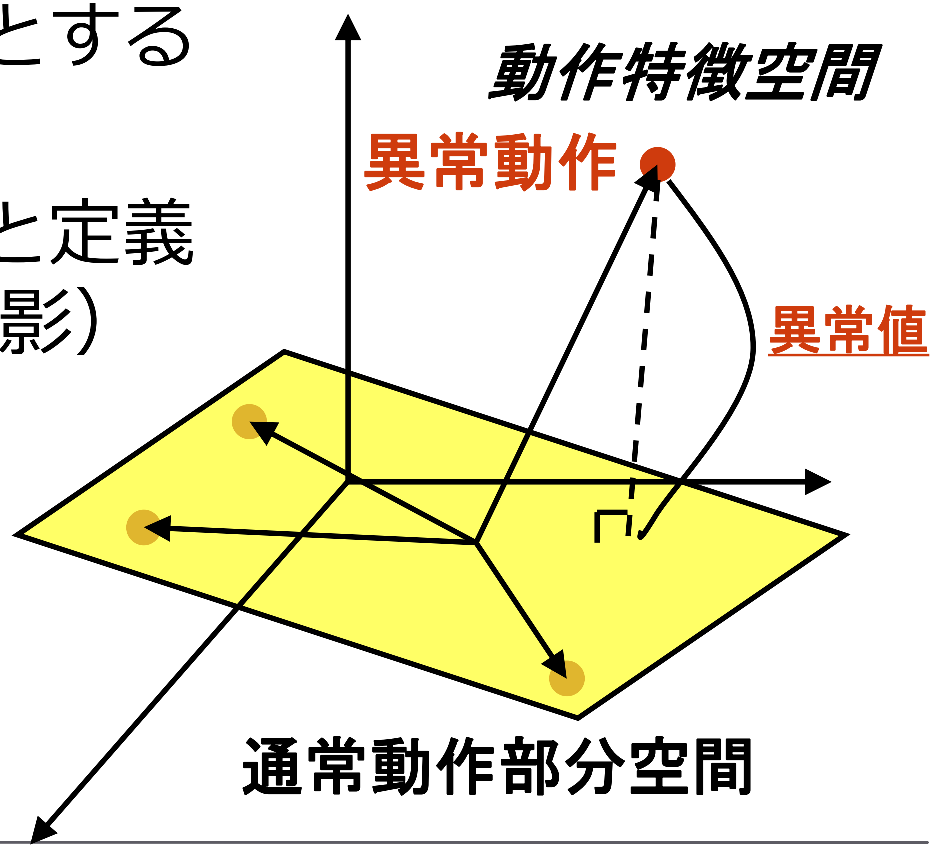
- 頻繁に起こる動作の部分空間をPCAで学習
- そこからの逸脱として，異常動作を検出

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）

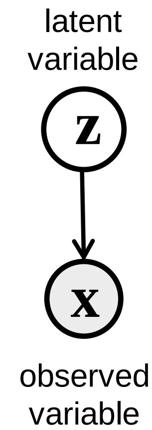
- 部分空間法



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 z には回転の自由度がある

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



- 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
  - 観測データと潜在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は次元数の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

Incremental PCA

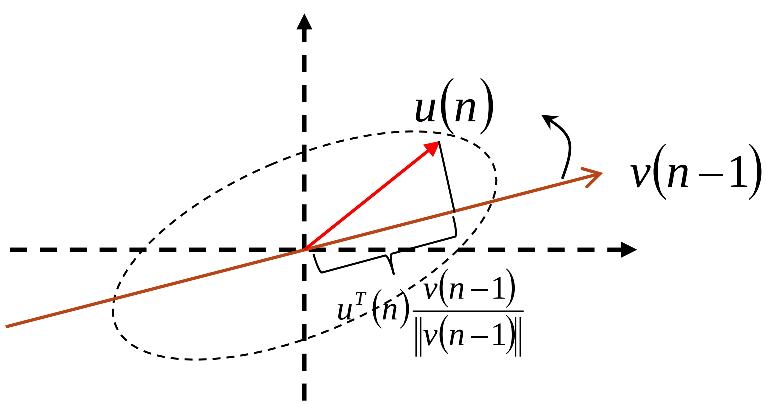
- Candid Covariance-free Incremental PCA
  [Went+, PAMI'03]
  - 逐次的に主成分ベクトルを求める手法のひとつ
  - 新規データ入力のたびに、主成分ベクトルをデータの方向へ引っ張る

n番目までのデータで
推定される主成分ベクトル

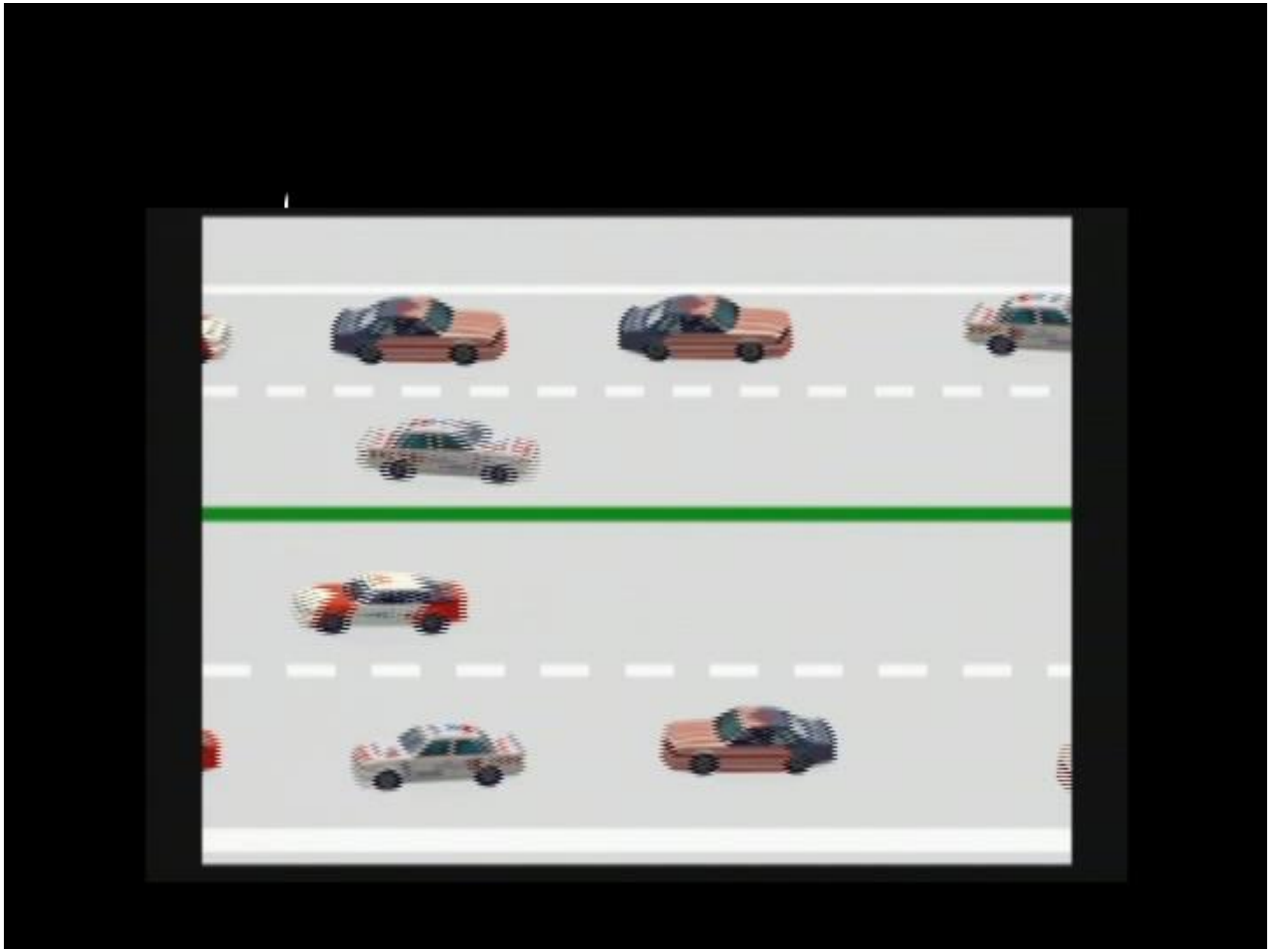
n番目のデータベクトル

$$v(n) = \frac{n-1}{n} v(n-1) + \frac{1}{n} u(n) \boxed{u^T(n) \frac{v(n-1)}{||v(n-1)||}}$$

n番目のデータの現在の主成分スコア

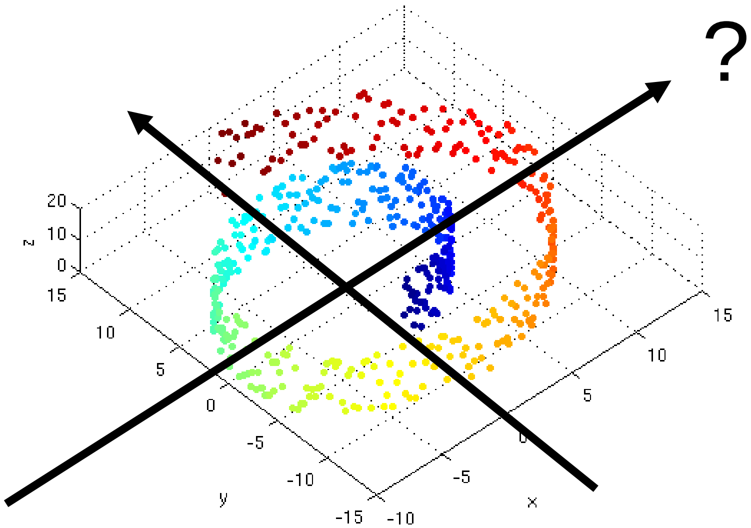


異常検出への応用



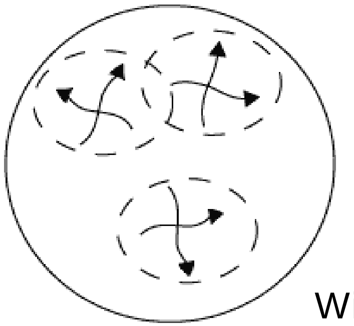
PCAの限界

- データの分布に非線形構造がある場合は対応できない



多様体学習

- 多様体
  - （一般的な意味での）空間を一般化した概念
  - 局所的にはユークリッド空間と見なせる点の集合（つまり微分可能）
  - 計量、接続が重要



Wikipediaより

- 実用上意味するところは要するに非線形次元圧縮
  - でも本来の言葉の定義はちゃんと理解しておこう
  - Isomap (MDS), LLE, Laplacian eigenmap, LPP

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

- データ間の距離（類似度）をできるだけ保存するように低次元へ埋め込みを行う
- 基本的には距離が定義されていることが前提（間隔尺度）

$s_{ii'}$ を二つのデータi,i’の類似度とする

（元の特徴量が比尺度である場合、

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

この時はPCAとほぼ等価）

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

を最小とするように低次元の表現 z へ各データを配置する

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：多様体は局所的には線形な構造を持っているとみなせる
  - 近傍データの重みづけ和で表せる

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

j の近傍データ

LLE概要

- 1. 各データ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)

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

$$\hat{W}_i = \frac{\sum_k C_{jk}^{-1}}{\sum_{lm} C_{lm}^{-1}}$$

$$where \quad C_{jk} = \left(\mathbf{x}_i - \mathbf{n}_j\right)^T \left(\mathbf{x}_i - \mathbf{n}_k\right)$$

$\mathbf{n}$は$\mathbf{x}_i$の近傍ベクトル

- 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yを設定する

$$\sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j \in N(i)} W_{ij} \mathbf{y}_j \right\|^2$$

$$= \mathrm{tr}\left(Y^T M Y\right)$$ ただし $M = (I - W)^T (I - W), \quad Y^T Y = I$

Y の第 i 列が $\mathbf{y}_i$

解き方

- 学習データ数（サンプルサイズ）次元の固有値問題

- PCAでは

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

- LLEは

  $\mathrm{tr}\left(Y^T M Y\right)$ を $Y^T Y = I$ のもとで最小化 ⇨ 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

注意

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

Laplacian eigenmaps [Belkin & Niyogi, 2001]

- 局所的な類似度構造を保存する埋め込みを学習
  - グラフに基づくLLEの一般化

- 1. 類似度行列 *W* を計算
  - ガウス類似度： $W_{i,j} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\delta} \right)$
  - k 最近傍類似度: $W_{i,j} = \begin{cases} 1 & \mathbf{x}_i \text{が} \mathbf{x}_j \text{の} k \text{最近傍に含まれるか、} \\ & \mathbf{x}_j \text{が} \mathbf{x}_i \text{の} k \text{最近傍に含まれる場合} \\ 0 & \text{それ以外の場合} \end{cases}$

- 2. 類似度で重み付けたデータ間の埋め込み距離を最小化

$$\sum_{i,j} W_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 = \mathrm{tr}\left(Y^T L Y\right) \quad L \equiv D - W \quad \text{（グラフラプラシアン）}$$

$$\text{ただし } Y^T D Y = I \qquad D \text{は} D_{ii} = \sum_j W_{ij} \text{なる対角行列}$$

LLE概要（再）

- 1. 各データ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)

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

- 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yを設定する

$$\sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j \in N(i)} W_{ij} \mathbf{y}_j \right\|^2$$

学習データしか埋め込めない

$$= \mathrm{tr}\left(Y^T M Y\right) \text{ ただし } M = (I - W)^T (I - W), \quad Y^T Y = I$$

Y の第 i 列が $\mathbf{y}_i$

Locality preserving projection (LPP)

[He and Niyogi, 2004]

- 局所的な類似度構造を保存する<u>線形射影</u>を学習

- 1. 類似度行列 *W* を計算
  - ガウス類似度： $$W_{i,j} = \exp\left(- \frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{2\delta} \right)$$
  - k 最近傍類似度: $$W_{i,j} = \begin{cases} 1 & \mathbf{x}_i \text{が} \mathbf{x}_j \text{の} k \text{最近傍に含まれるか、} \\ & \mathbf{x}_j \text{が} \mathbf{x}_i \text{の} k \text{最近傍に含まれる場合} \\ 0 & \text{それ以外の場合} \end{cases}$$

- 2. 類似度で重み付けたデータ間の埋め込み距離を最小化

$$J_{LPP} = \frac{1}{2} \sum_{i,j}^{N} W_{i,j} \left\| \mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j \right\|^2$$

LPP つづき

$$\frac{1}{2}\sum_{i,j}^{N} W_{i,j}\left\|\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j\right\|^2 = \sum_{i}^{N}\mathbf{a}^T\mathbf{x}_i D_{ii}\mathbf{x}_i^T\mathbf{a} - \sum_{i,j}^{N}\mathbf{a}^T\mathbf{x}_i W_{ij}\mathbf{x}_j^T\mathbf{a}$$

$$= \mathbf{a}^T X\left(D - W\right)X^T\mathbf{a} = \mathbf{a}^T XLX^T\mathbf{a}$$

$X = \left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\right)$

$D:$ （対角行列） $D_{ii} = \sum_{j}^{N} W_{ij}$ データ i にかかる重みの総和

$L = D - W$ **グラフラプラシアン行列**

上記を $\mathbf{a}^T XDX^T\mathbf{a} = 1$ の条件下で最小化（重み付の分散正規化）

$$XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}$$

の最小m固有値に対応する固有ベクトル

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

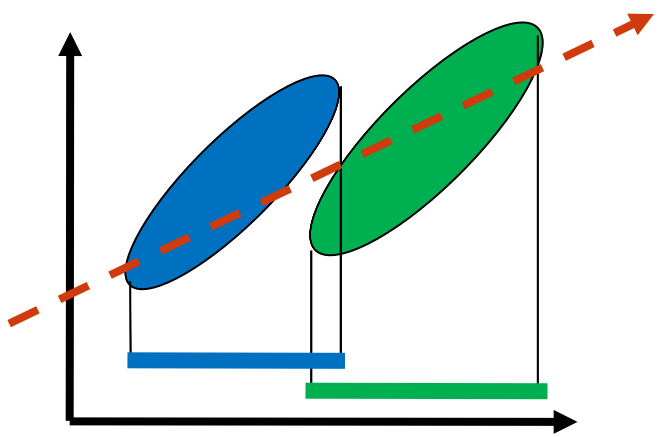
ダミー変数

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

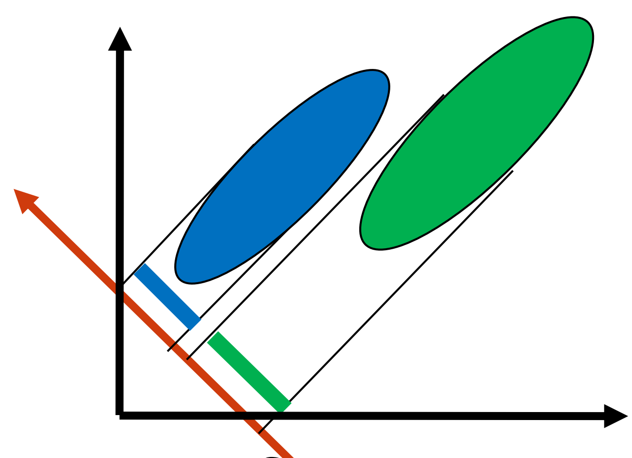
- クラス（カテゴリ）のサンプルを<u>最もよく分離する軸</u>を見つける

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

- 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

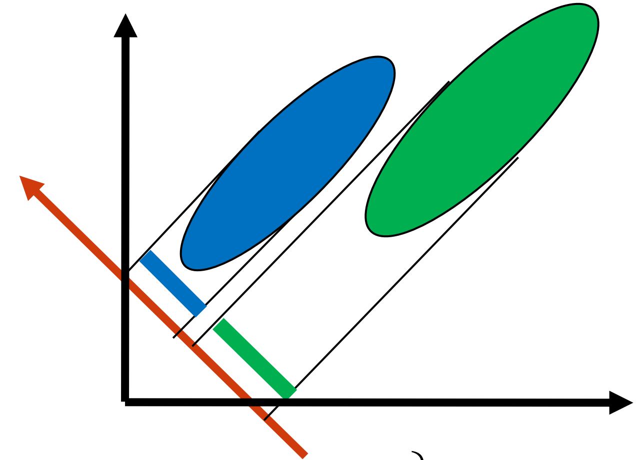


Overlap



Separate

FDA



- クラス内分散
  - クラスごとのデータの分散の平均

$$\sum_{i=1}^{N_c}\sum_{x\in class(i)}\left(\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\overline{\mathbf{x}}_i\right)\left(\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\overline{\mathbf{x}}_i\right)^T=\mathbf{w}^T\left\{\sum_{i=1}^{N_c}\sum_{x\in class(i)}\left(\mathbf{x}-\overline{\mathbf{x}}_i\right)\left(\mathbf{x}-\overline{\mathbf{x}}_i\right)^T\right\}\mathbf{w}$$

クラス数

クラスiの平均ベクトル

$$=\mathbf{w}^T\underline{C_W}\mathbf{w}$$

クラス内共分散行列

- クラス外分散
  - クラスの平均ベクトルの分散

$$\sum_{i=1}^{N_c}n_i\left(\mathbf{w}^T\overline{\mathbf{x}}_i-\mathbf{w}^T\overline{\mathbf{x}}\right)\left(\mathbf{w}^T\overline{\mathbf{x}}_i-\mathbf{w}^T\overline{\mathbf{x}}\right)^T=\mathbf{w}^T\left\{\sum_{i=1}^{N_c}n_i\left(\overline{\mathbf{x}}_i-\overline{\mathbf{x}}\right)\left(\overline{\mathbf{x}}_i-\overline{\mathbf{x}}\right)^T\right\}\mathbf{w}$$

クラスiのサンプル数

全平均ベクトル

$$=\mathbf{w}^T\underline{C_B}\mathbf{w}$$

クラス外共分散行列

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

$$J_{FDA} = \frac{\mathbf{w}^T C_B \mathbf{w}}{\mathbf{w}^T C_W \mathbf{w}}$$

Fisher’s discriminant criterion

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

$$J'_{FDA} = \mathbf{w}^T C_B \mathbf{w} - \lambda\left(\mathbf{w}^T C_W \mathbf{w} - 1\right)$$

w で偏微分して整理すると

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

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

FDA：注意

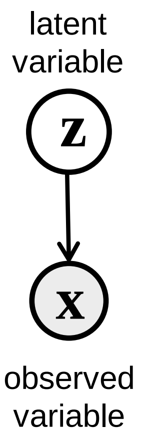
- (クラス数 – 1)個しか軸（固有ベクトル）は求まらない
  - クラス外共分散行列のランクの問題
  - それ以上特徴が欲しい場合は、直交補空間に順次射影していくなど工夫が必要
- クラス内共分散行列のランクに注意
  - サンプル数が少ないと不安定になる

$$C_W \to C_W + \alpha I$$

などとして正則化
（Regularized FDA)

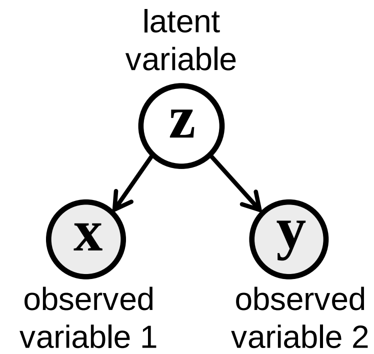
正準相関分析: Canonical Correlation Analysis (CCA)

- 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・目的変数の関係）
- 主成分分析の二変量版



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

Probabilistic interpretation of PCA



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

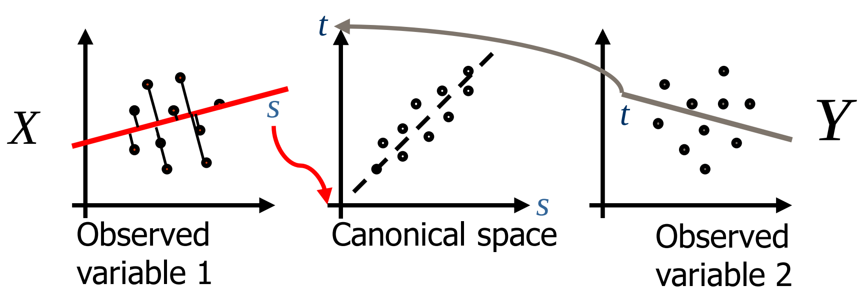
Probabilistic interpretation of CCA
[Bach and Jordan, 2005]

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。**

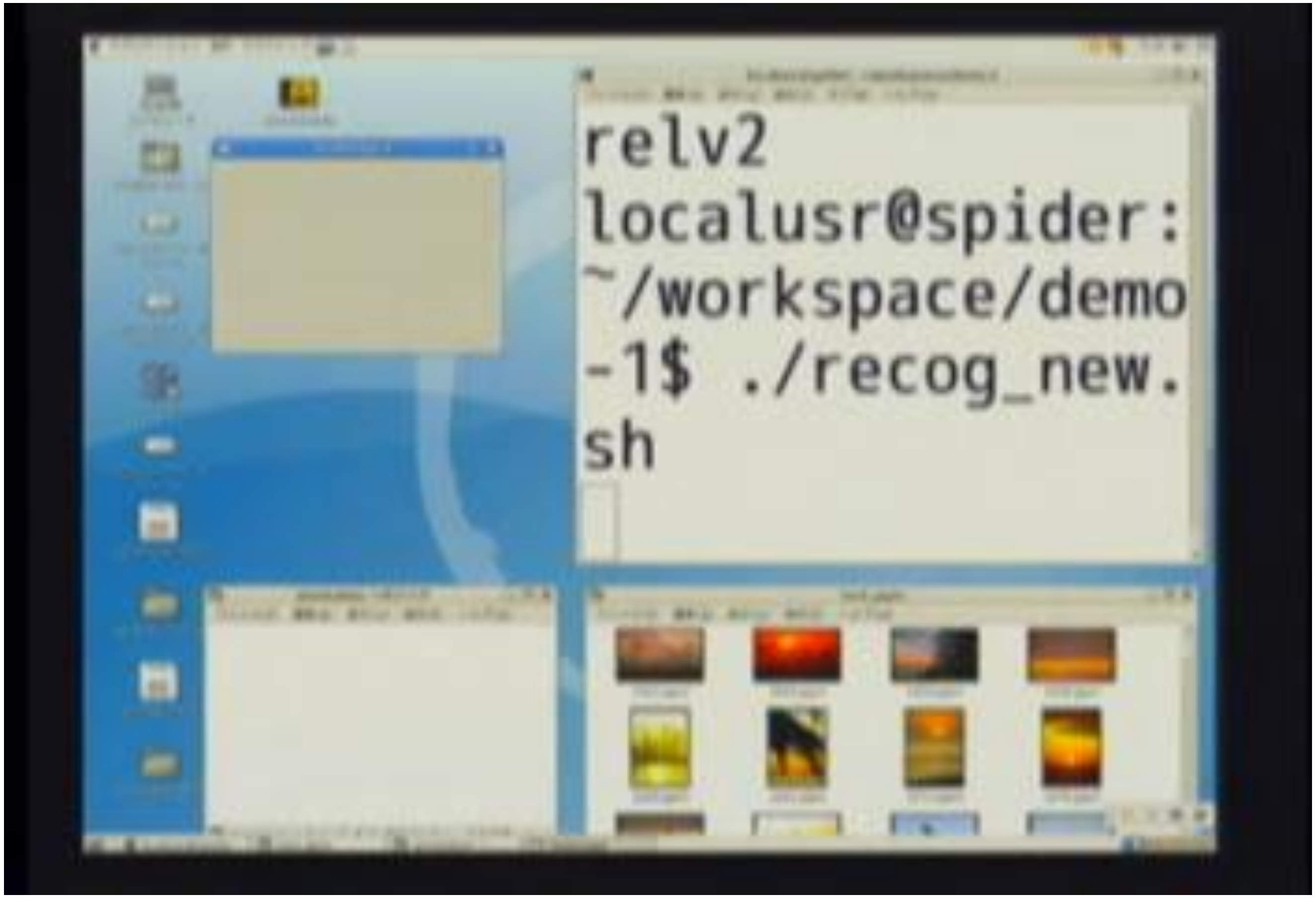
$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \quad \mathbf{a}^T C_{XX} \mathbf{a} = 1, \mathbf{b}^T C_{YY} \mathbf{b}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

(導出は省略)

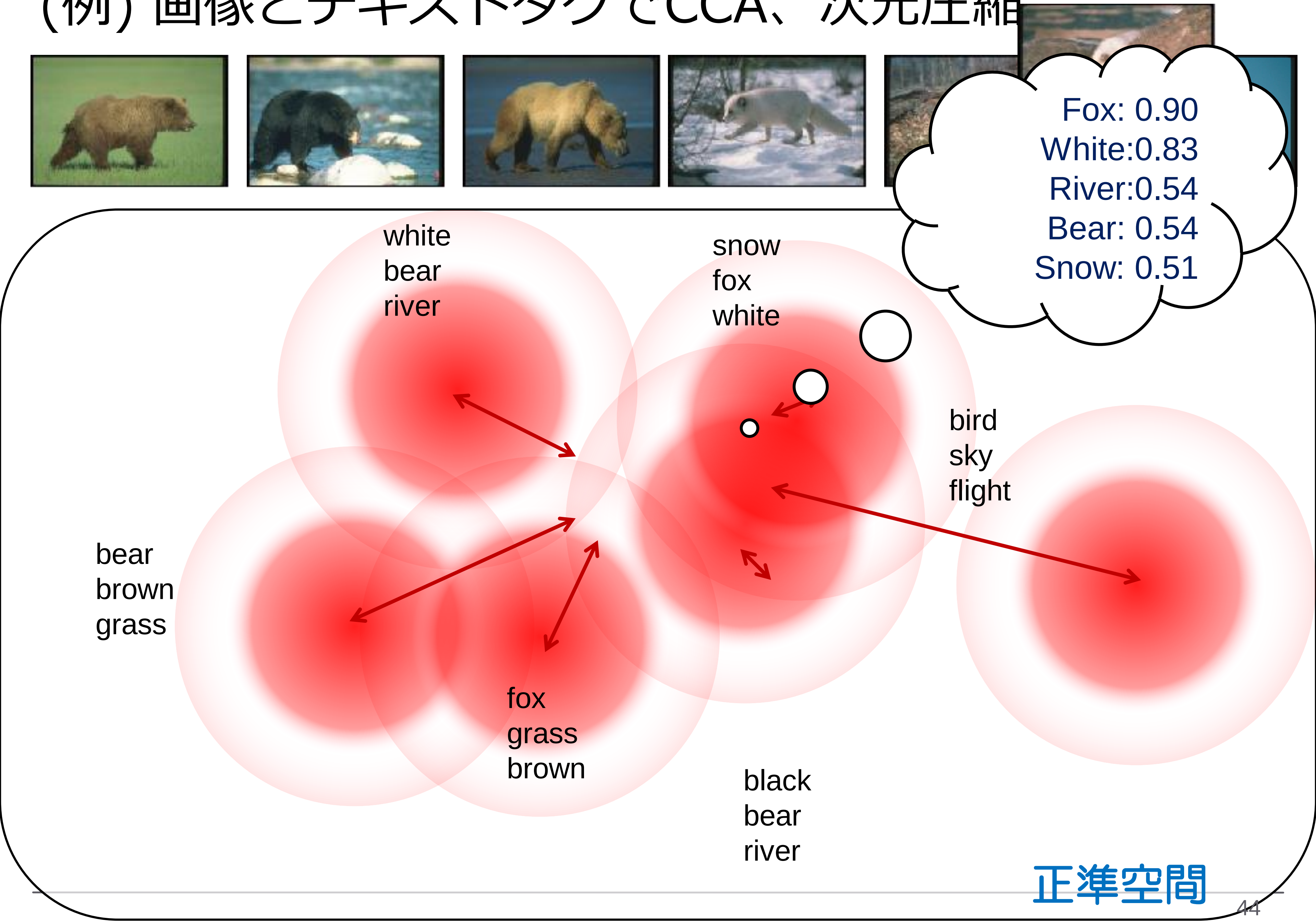
C： 共分散行列

λ： 正準相関係数

CCAを用いた画像認識

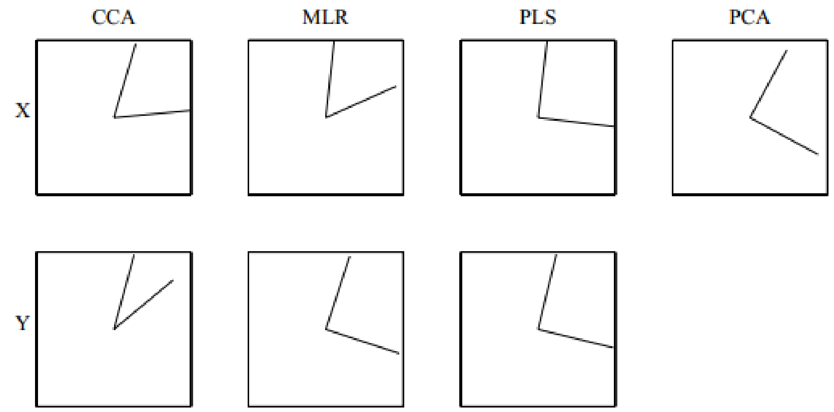


(例) 画像とテキストタグでCCA、次元圧縮



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 sqaures)
  - 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
10万円	火曜日
15万円	木曜日
8万円	日曜日

売上	日	月	火	水	木	金	土
10万円	0	0	1	0	0	0	0
15万円	0	0	0	0	1	0	0
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則化（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