１１章　次元削減

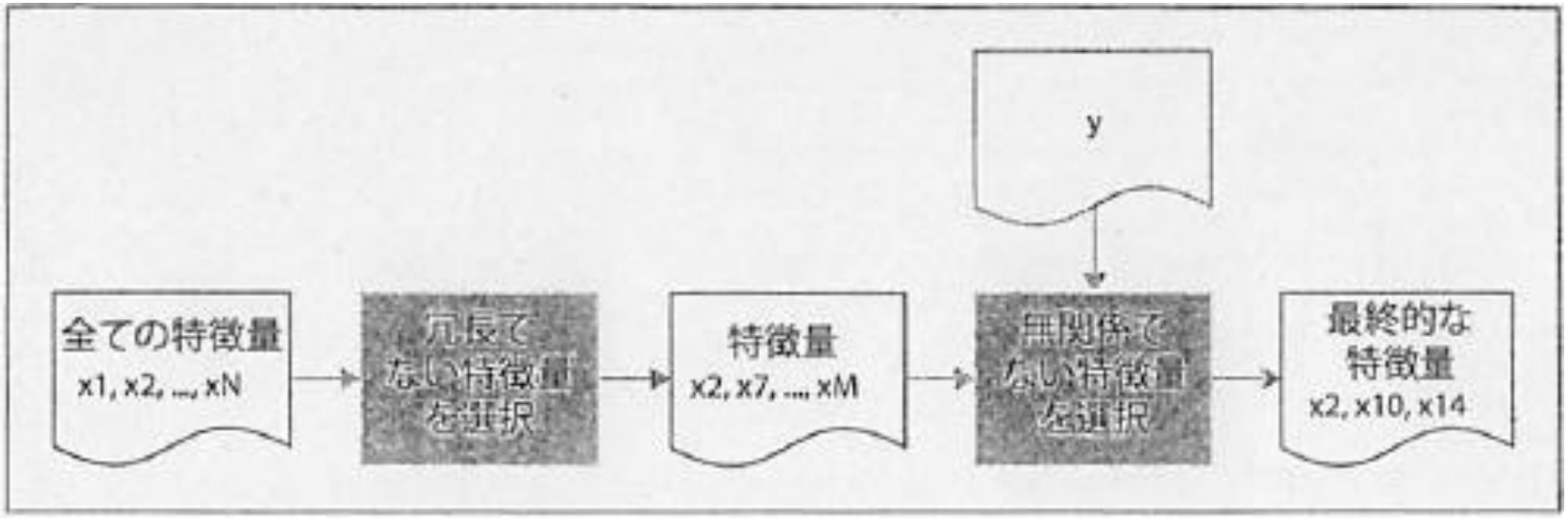
CAO　RUI

本章の目標：

- 無関係な特徴量や冗長な特徴量を削除する

特徴選択

- 冗長な特徴量をフィルター法を用いて検出する

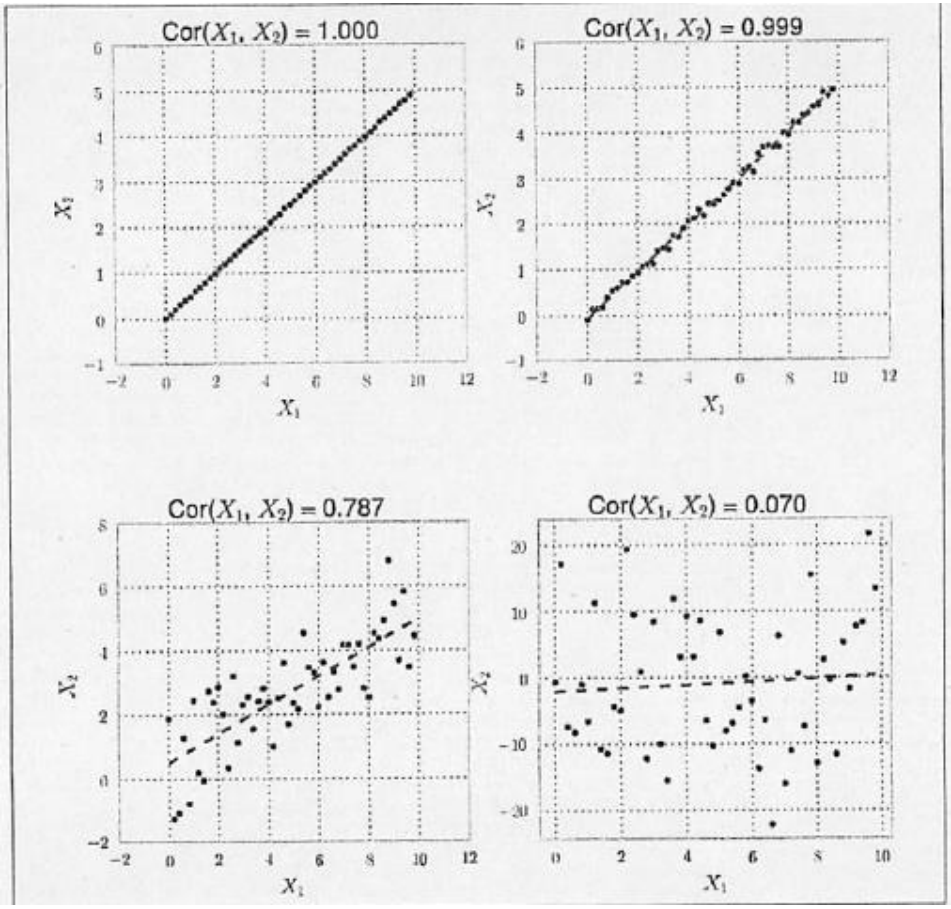


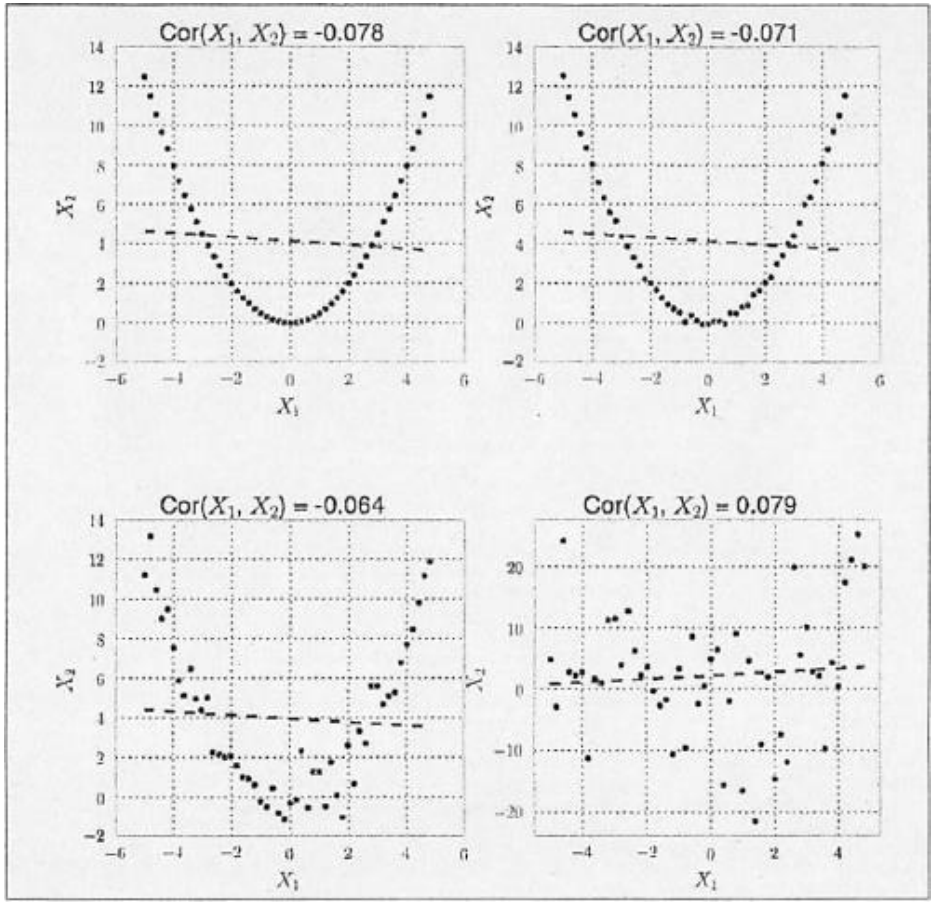
相関

- 目的：二つの特徴量間の関連性を見つける
- P値：対象とするデータが無関係なシステムから生成された確率

```
>> from scipy.stats import pearsonr
>> pearsonr([1,2,3], [1,2,3.1])
>> (0.9996222851612184 3, 0.017498096813278487)
>> pearsonr([1,2,3], [1,20,6])
>> (0.25383654128340477, 0.8366149366822740 5)
```

相関

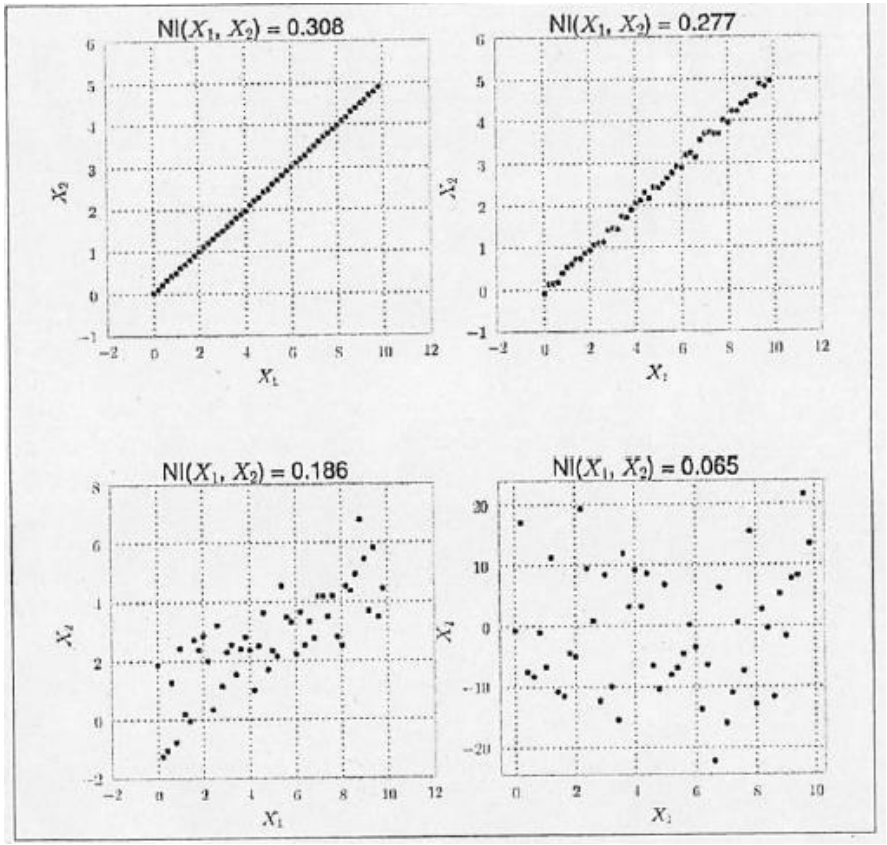


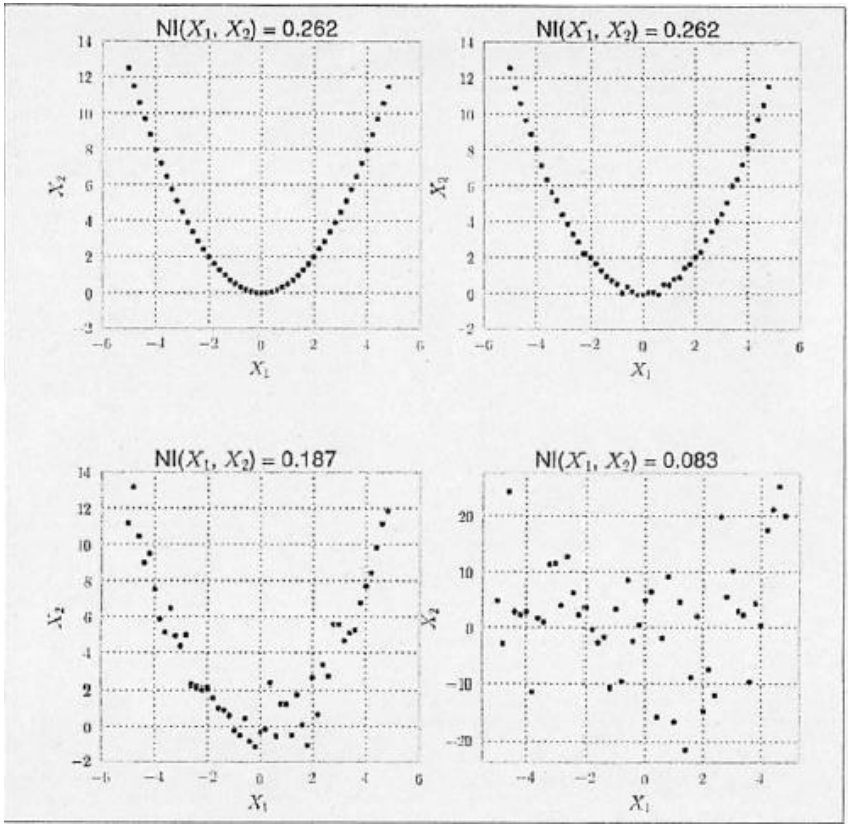


欠点：線形の関連性しか検出できません

相互情報量

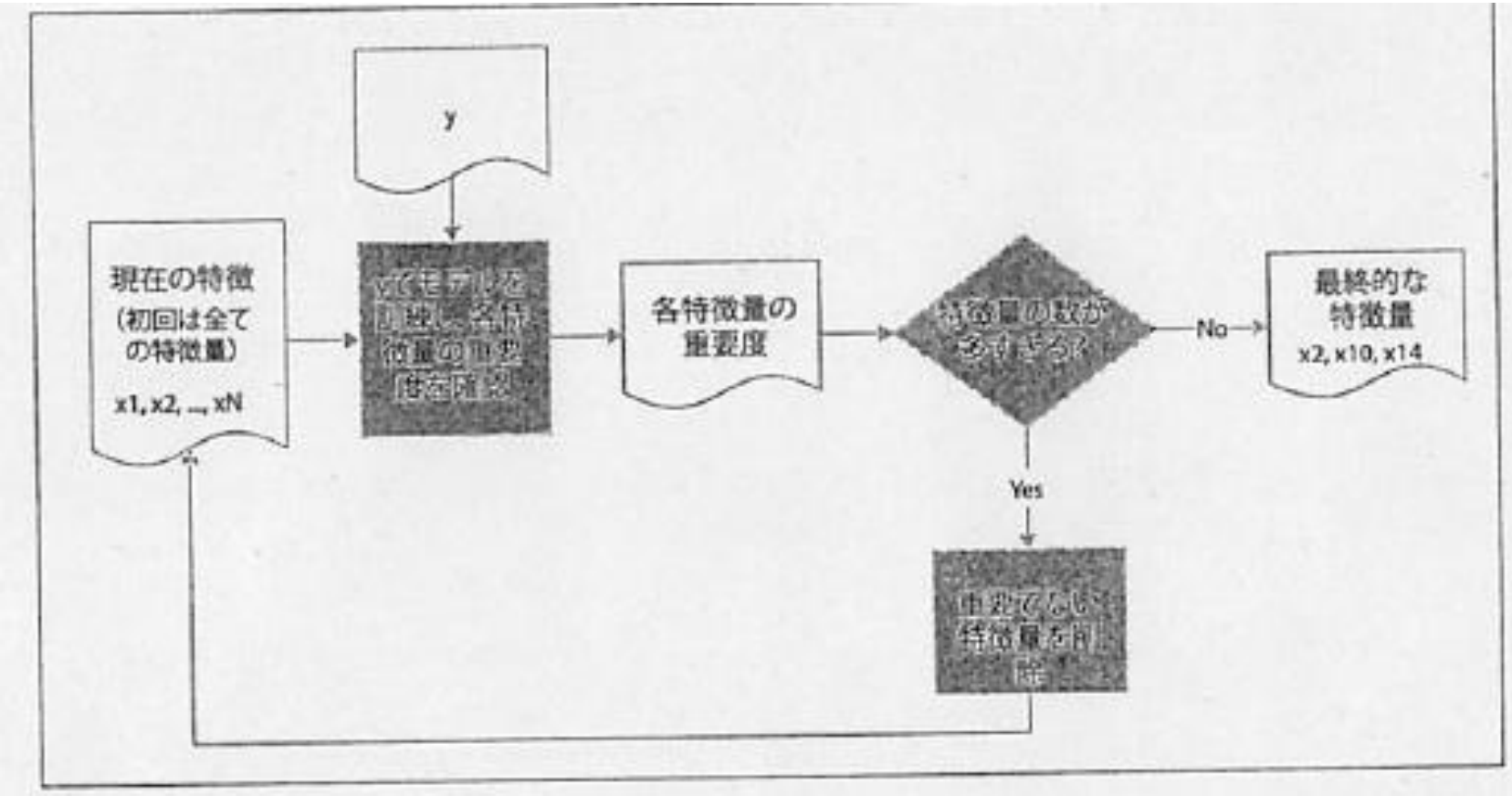
- 二つの特徴量がどれだけ共通する情報を持つかということを計算することです





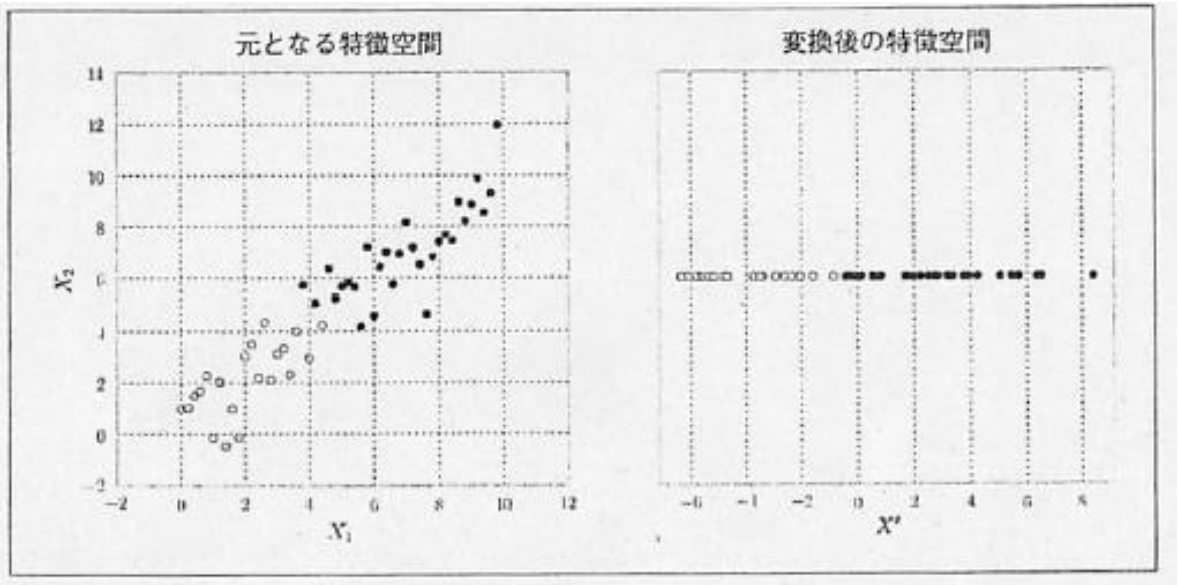
ラッパー法

- 目的：モデル自体にどの特徴量が有効であるか



主成分分析（PCA）

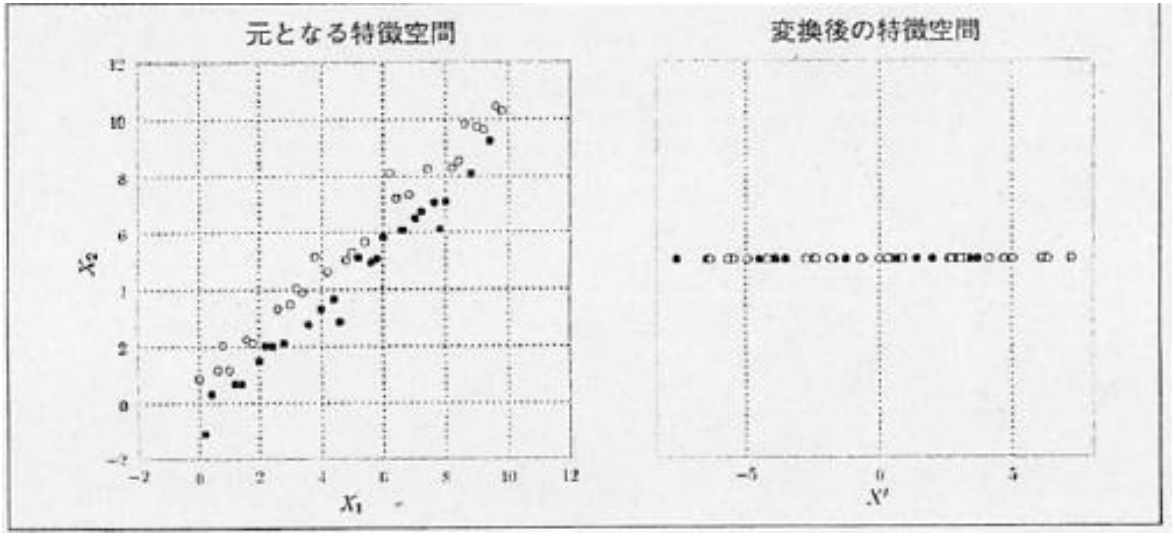
```
>>> x1 = np.arange(0, 10, .2)
>>> x2 = x1+np.random.normal(loc=0, scale=1, size=len(x1))
>>> X = np.c_[(x1, x2)] # c_[]で結合を行う
>>> good = (x1>5) | (x2>5) # データの一部を "good" なクラスとする
>>> bad = ~good
```

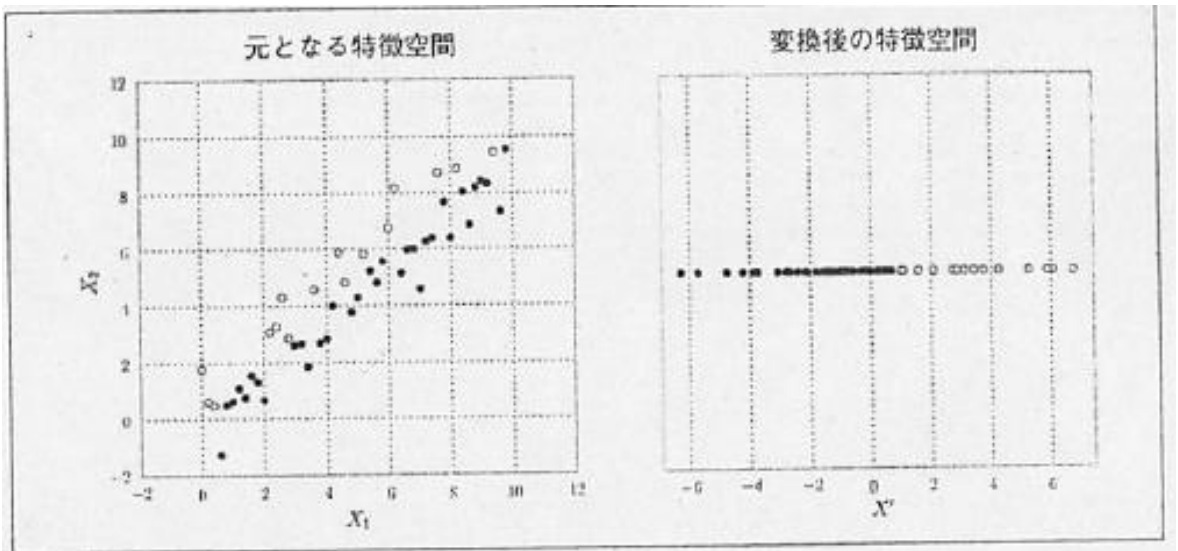



```
>>> print(pca.explained_variance_ratio_)
>>> [ 0.96393127]
```

PCAの限界とLDA

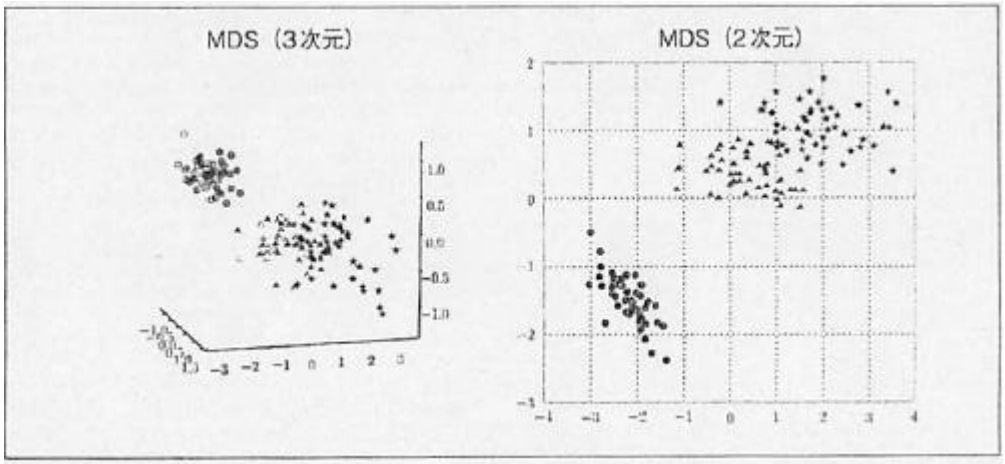
- >>> good = (x1>5) | (x2>5) → >>> good = x1>x2



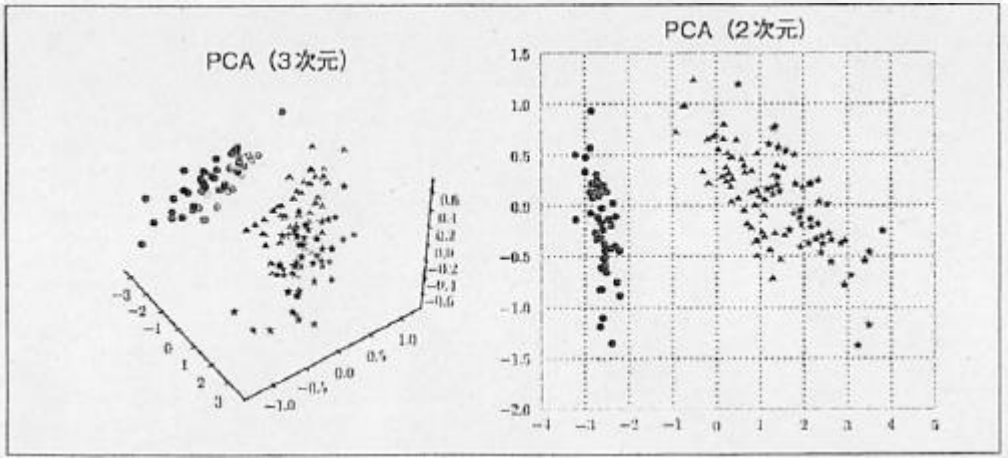


多次元尺度構成法（MDS）

- 目的：高次元からなるデータセットに対して、視覚的な印象を把握する



MDS



PCA