3.4　次元削減、特徴量抽出、多様体学習

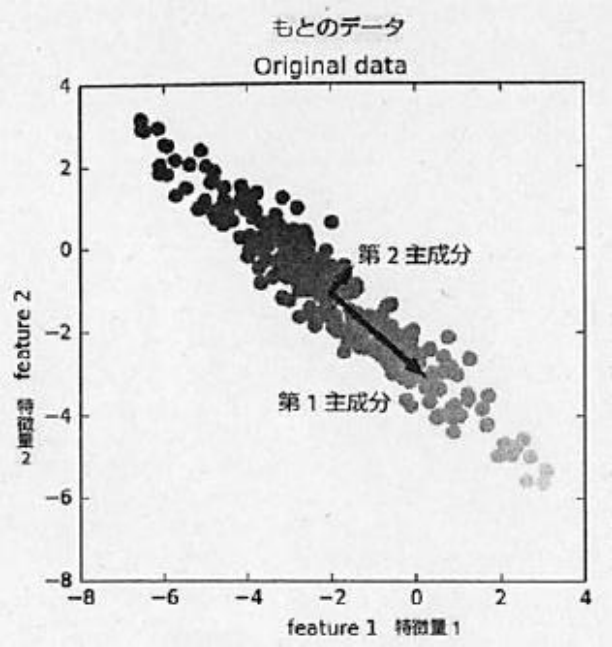
15t4034s
荘司　響之介

3.4.1　主成分分析

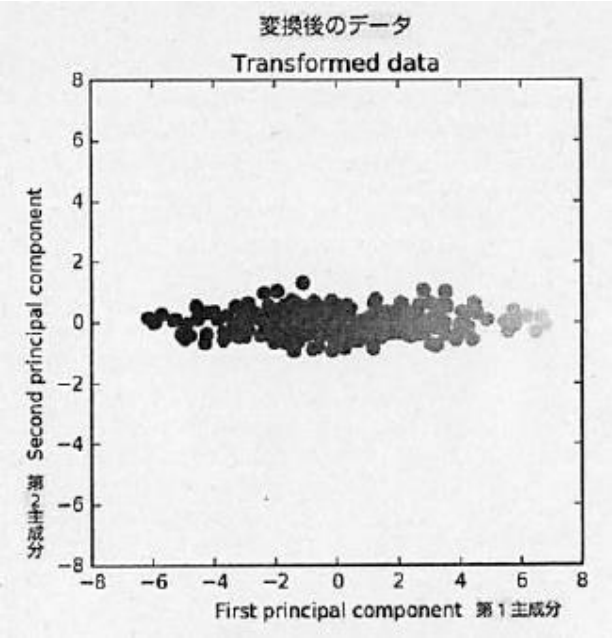
データセットの特徴量を相互に統計的に関連しないように回転する手法。

回転した後の特徴量から、重要な特徴量だけを抜き出す

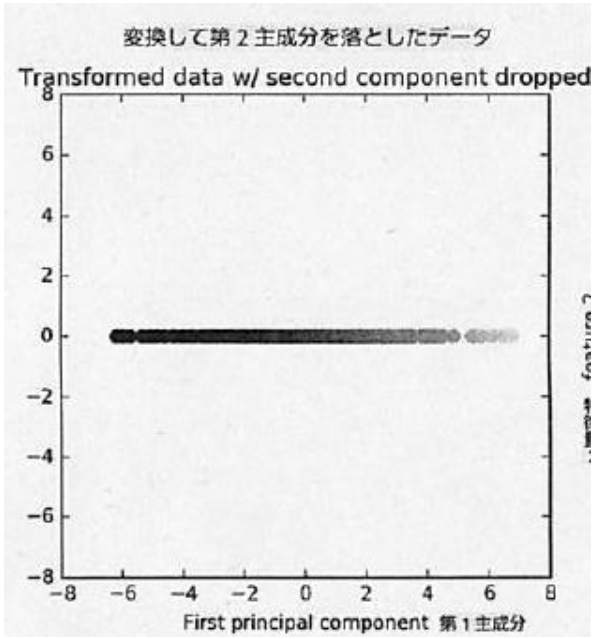
PCAを合成2次元データセットに適用した例を示す。



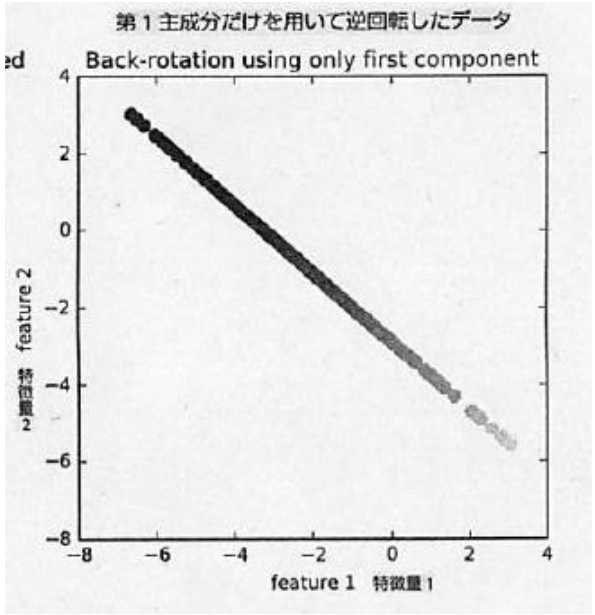
・最も分散が大きい方向を見つけ、それに「第1成分」というラベルを付ける。
・第1成分と直行する方向を「第2成分」とする。



・第 1 成分がx軸に、第2成分がy軸に沿うように回転させる。
・回転させる前にデータから平均値を引き、原点周辺にデータが来るようにする。



最も興味深い方向（第１成分）を見つけ、それ以外の方向（第2成分）は落とす。



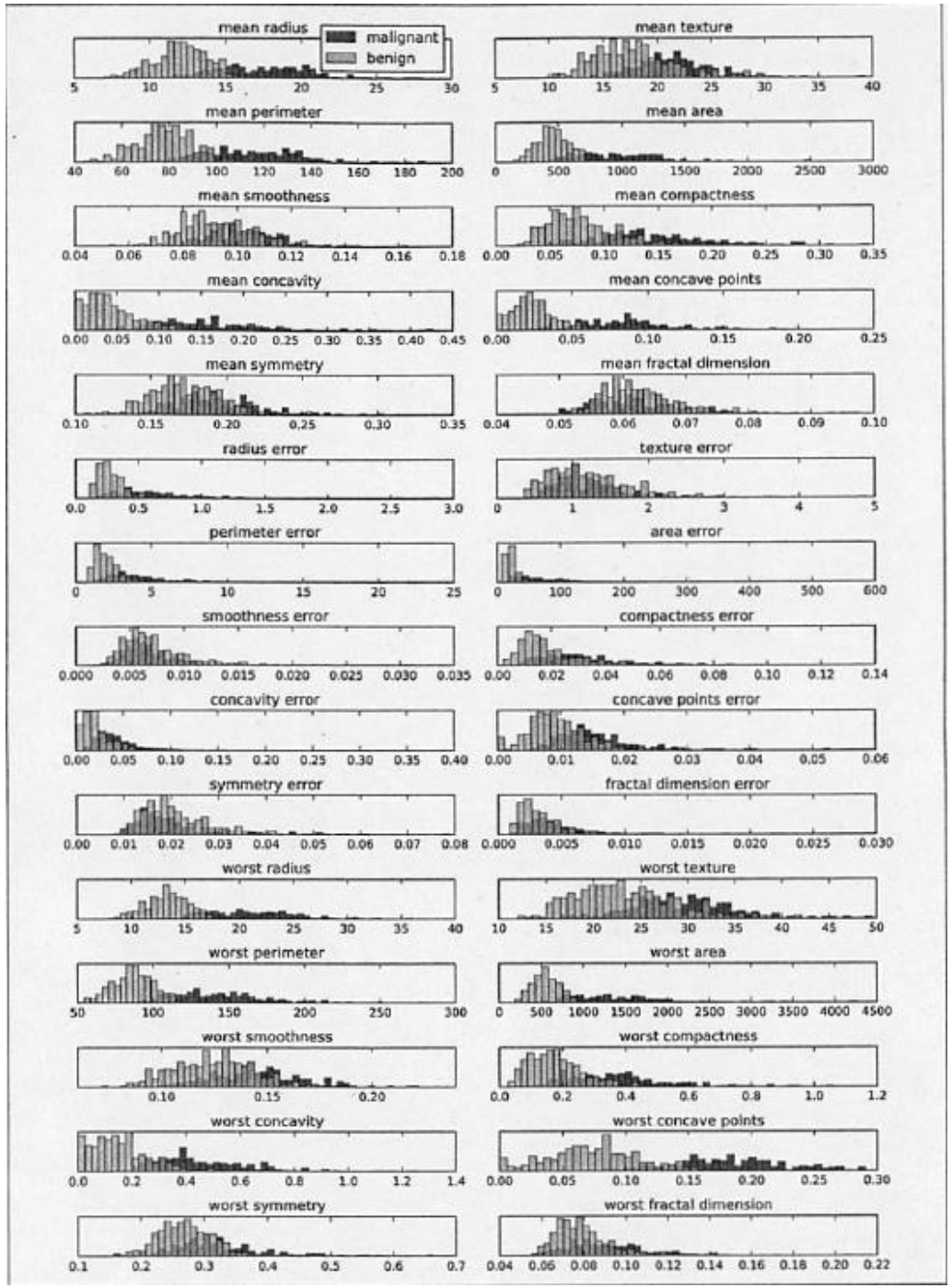
逆回転して平均を足し、データを元に戻す。

3.4.1.1　cancerデータセットのPCAによる可視化

2つより多い特徴量を持つデータ（cancerデータセット）の散布図を作ることは難しい。

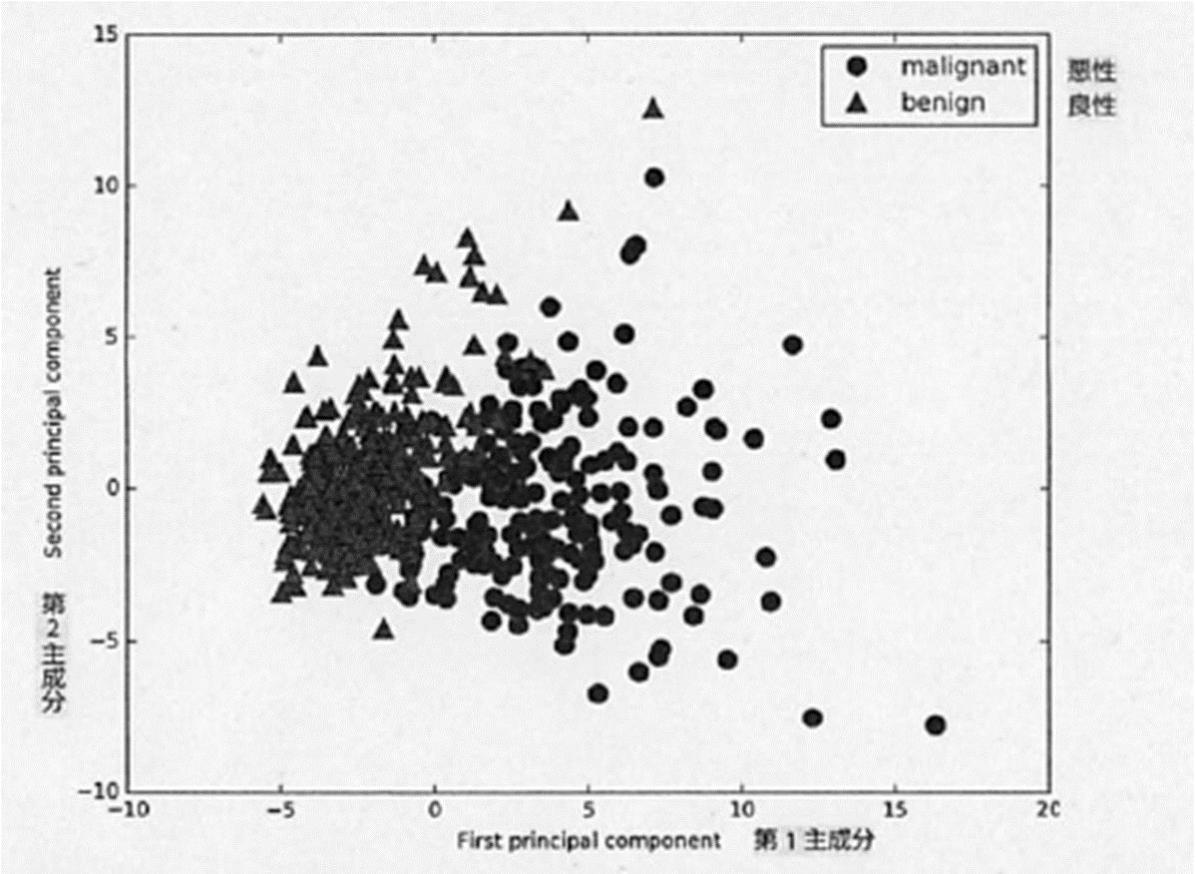
しかし、特徴量ごとに２つのクラスのヒストグラムを作成することで、単純に可視化することができる。

次ページに、作成したヒストグラムを示す。



例えば、「smoothness error」のヒストグラムは、
ほとんど重なっているのであまり情報がなさそうである。
一方、「worst concave points」のヒストグラムはほとんど
重なっていないので、情報が多い。

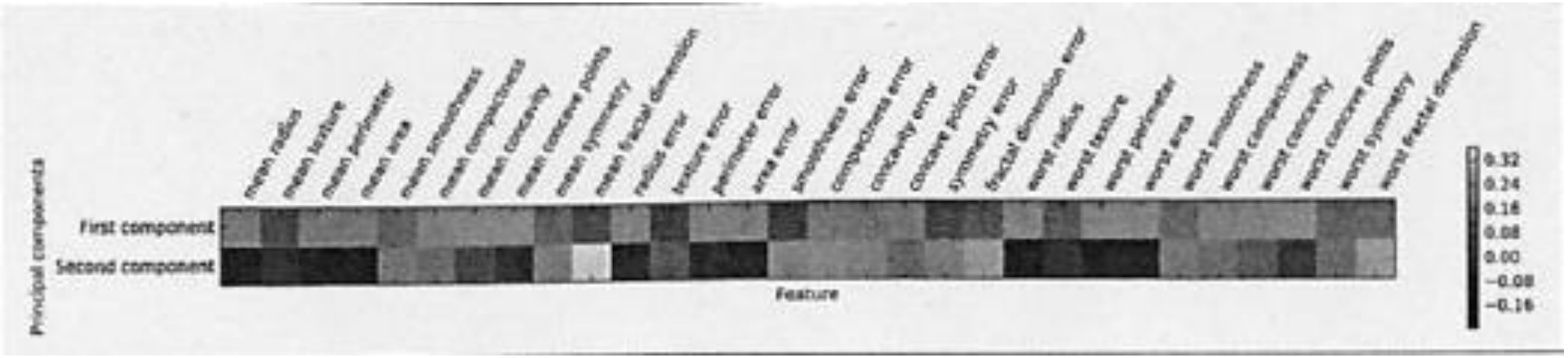
下図は、これをもとに２つの主成分を取り出して作成した散布図



主成分の詳細

```
[[ 0.219  0.104  0.228  0.221  0.143  0.239  0.258  0.261  0.138  0.064
   0.206  0.017  0.211  0.203  0.015  0.17   0.154  0.183  0.042  0.103
   0.228  0.104  0.237  0.225  0.128  0.21   0.229  0.251  0.123  0.132]
 [-0.234 -0.06  -0.215 -0.231  0.186  0.152  0.06  -0.035  0.19   0.367
  -0.106  0.09  -0.089 -0.152  0.204  0.233  0.197  0.13   0.184  0.28
  -0.22  -0.045 -0.2   -0.219  0.172  0.144  0.098 -0.008  0.142  0.275]]
```

行は、主成分に対応し、重要度ごとにソートされている。列は、特徴量に対応する。
係数をヒートマップで示したものが下図である。



3.4.1.2　固有顔による特徴量抽出

62人分の顔画像を用いてクラス分類をする。

・訓練データを変換せずに1-最近棒法クラス分類を使った場合

　精度：27%→あまり高くない

　理由

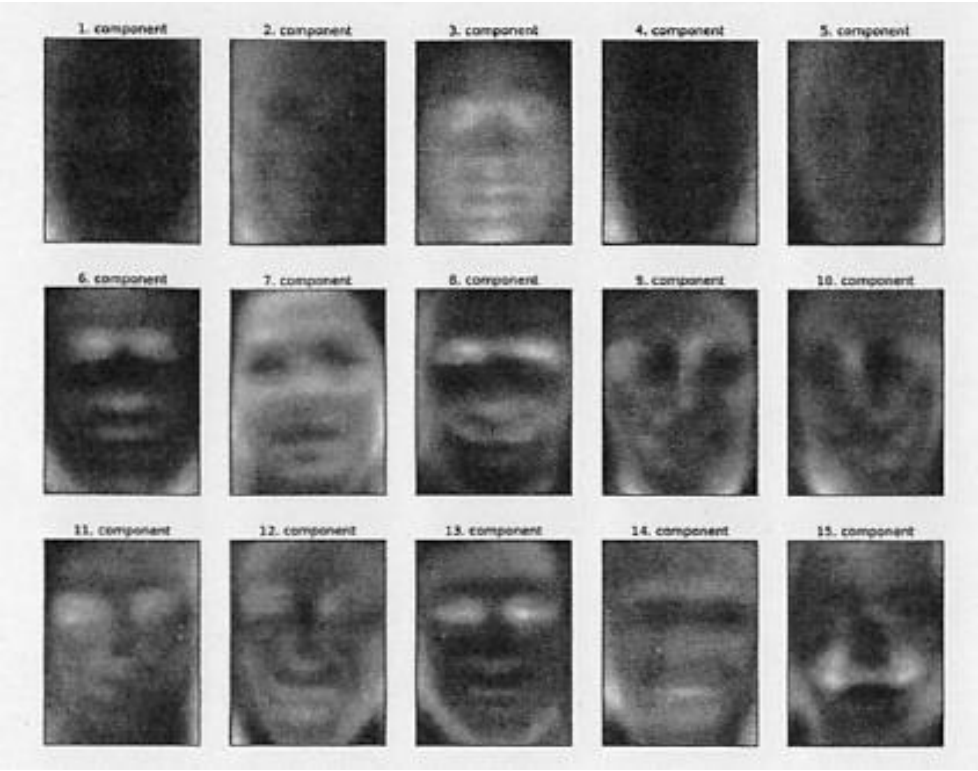
　ピクセル空間で距離を計算するのは、顔の近似度を測るのに適していない。例えば、1ピクセル右にずらすだけで、表現が全く変わってしまい、大きく変化したことになる。

PCAを利用

・PCAオブジェクトを訓練し、データを変換した場合

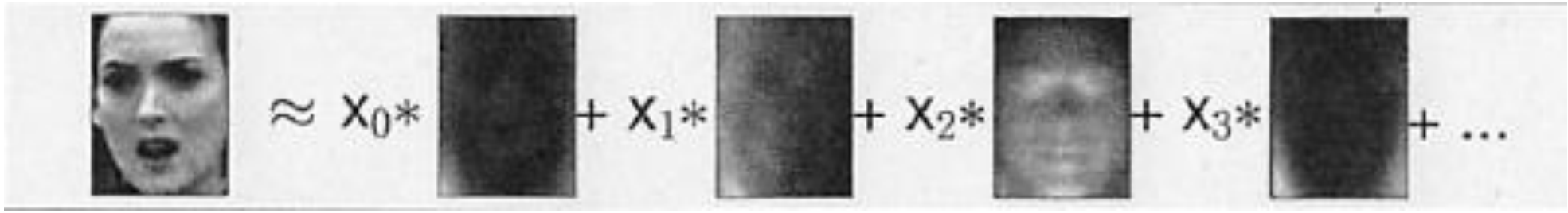
精度：36%→向上

主成分の方向は、入力空間の方向に対応するため、この空間の方向は画像になる。

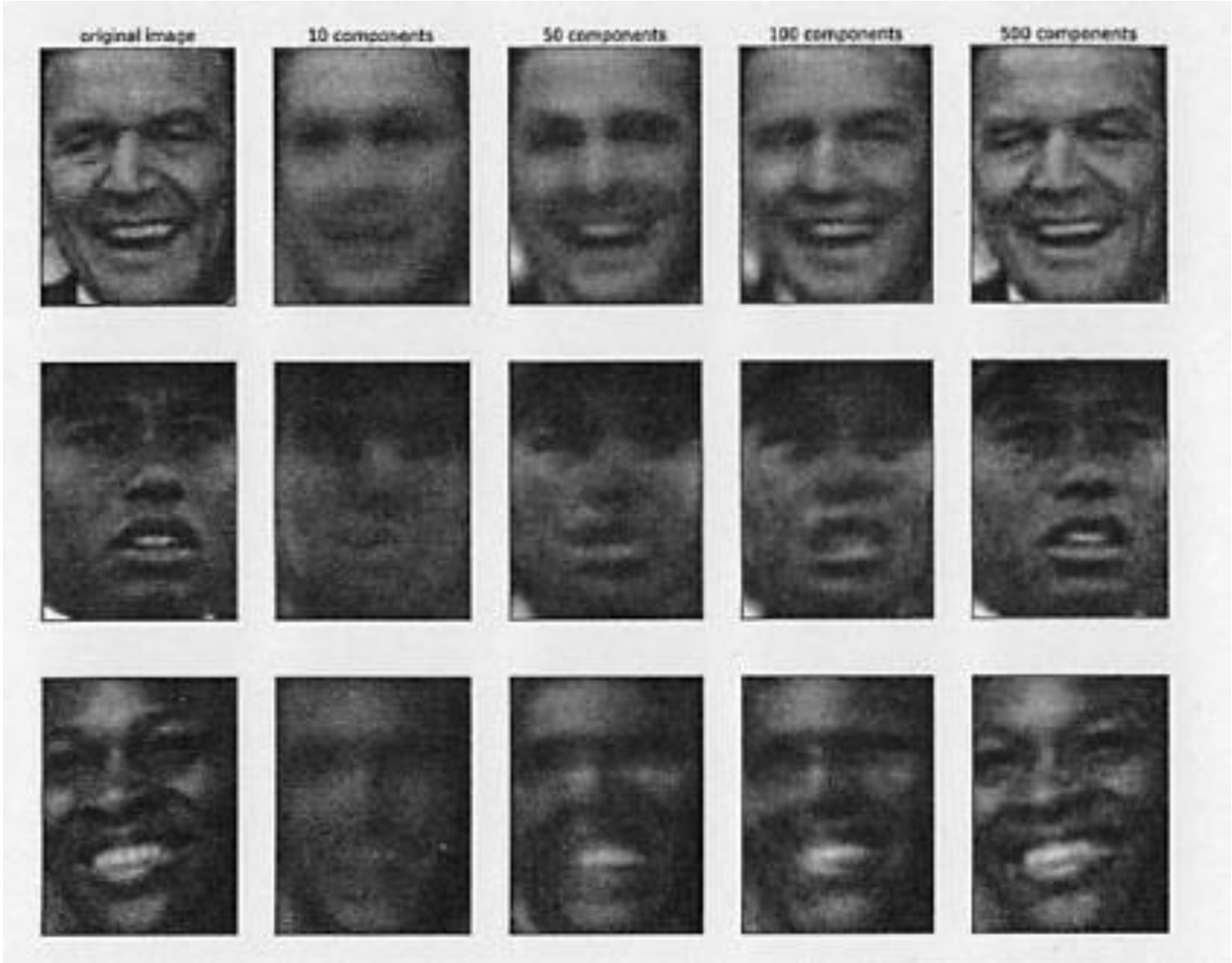


１つ１つが主成分である。

PCAは、テストデータポイントの主成分の重み付き和として表現する、一連の数字（PCAで回転後の新しい特徴量）を見つける手法だ、と解釈すると、以下のように表現できる。



主成分の一部だけを使って元画像を再現した図を次ページに示す。


original image
10 components
50 components
100 components
500 components

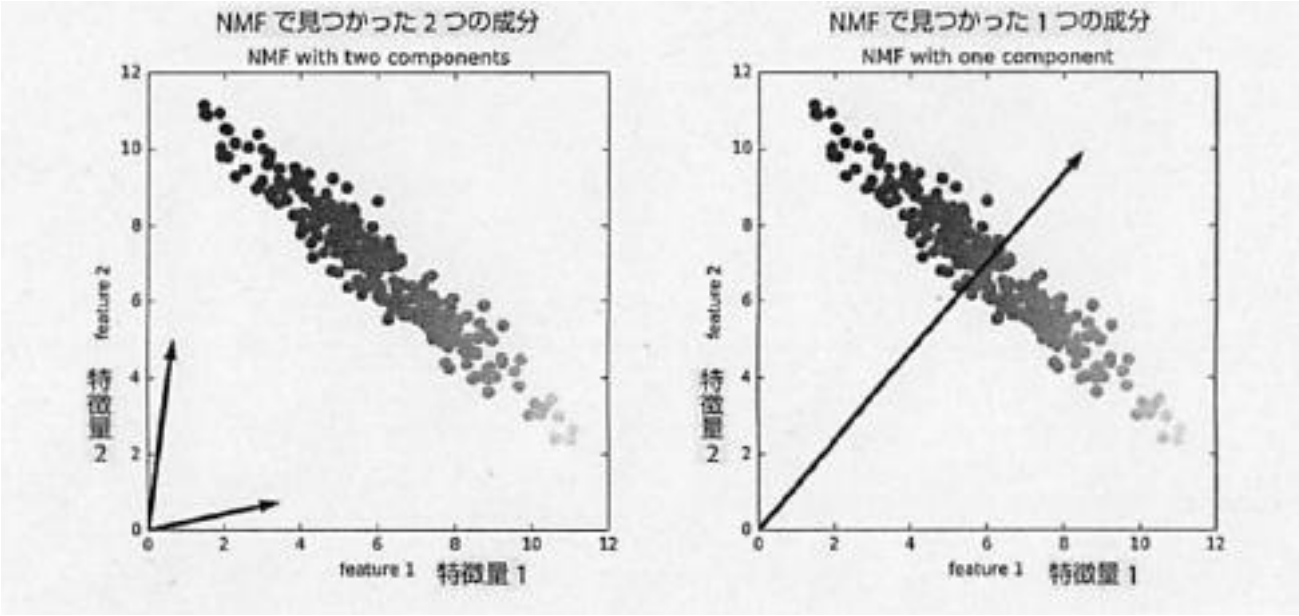
3.4.2　非負値行列因子分解（NMF）

有用な特徴量を抽出することを目的とする教師なし学習手法。

動作はPCAと似ているが、NMFでは係数と成分が非負であることが求められる。

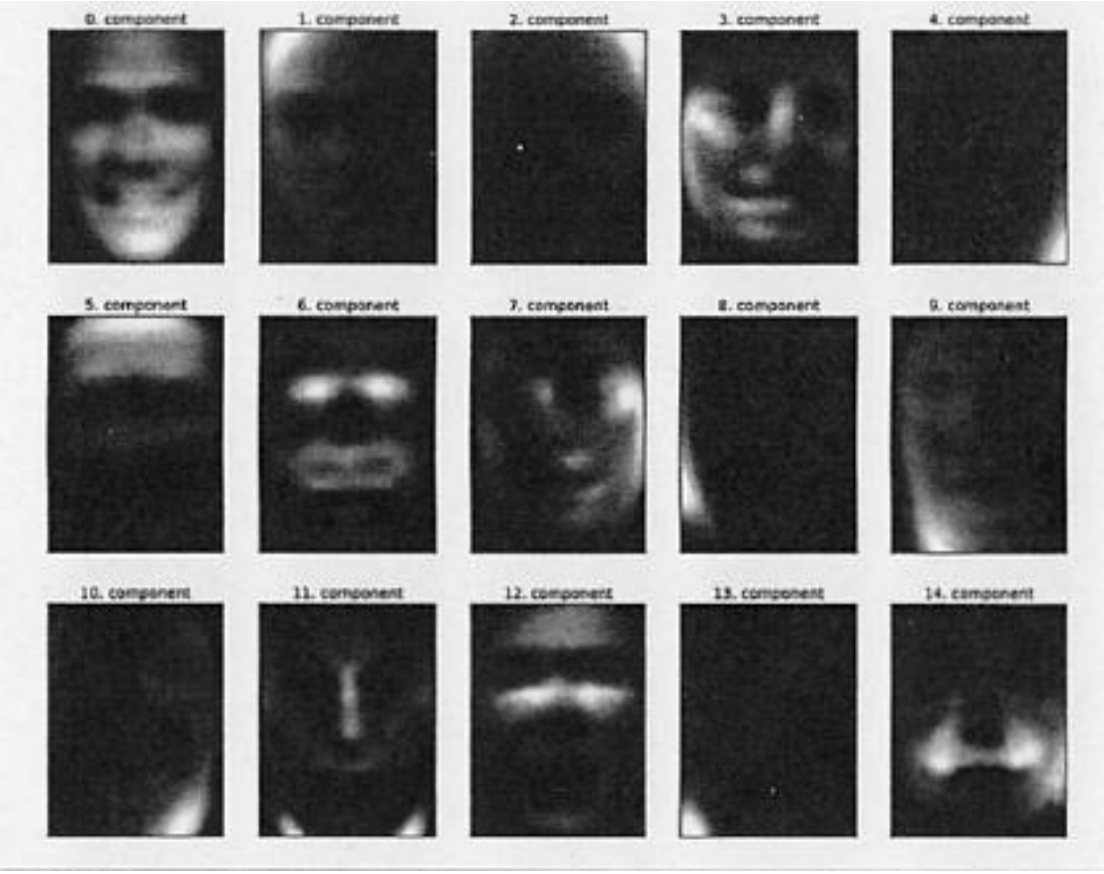
3.4.2.1　NMFの合成データへの適用

抽出された成分はデータへの方向であると考えることができる。
PCAとは異なり、成分の数が変わると、いくつかの成分がなくなるのではなく、全く別の成分集合が構成される。



3.4.2.2　NMFの顔画像への適用

NMFは、データ中から興味深いパターンを見つけるのに用いられる。
先ほど用いた顔画像データの最初の15成分は以下の図の様なった。



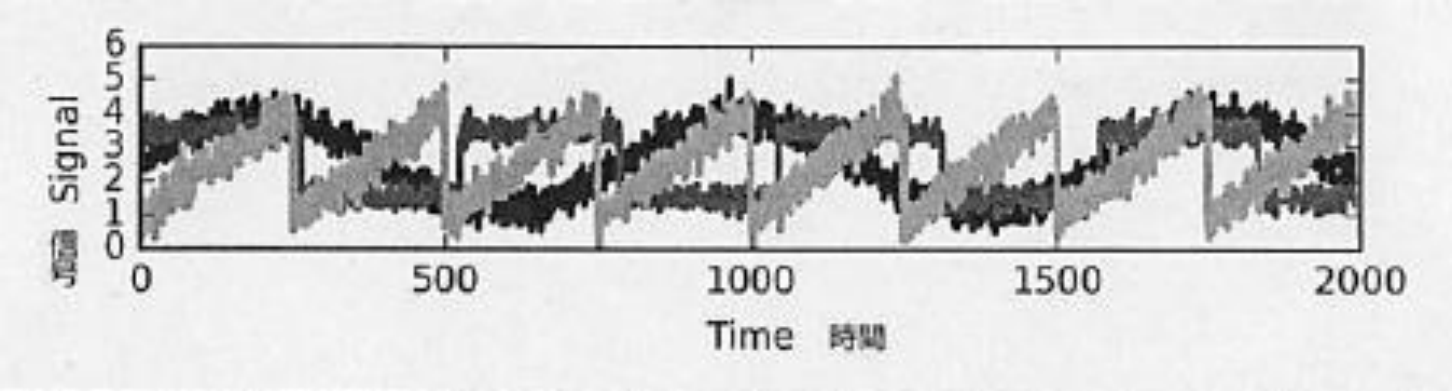
成分３は少し右を向いた顔、成分7は少し左を向いた顔を表していることがわかる。

成分3の係数が大きい顔画像

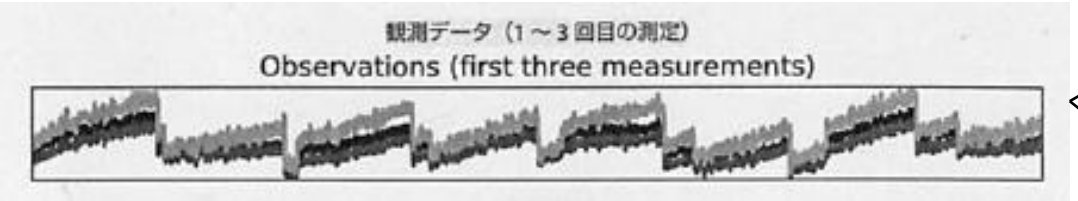
成分 7 の係数が大きい顔画像

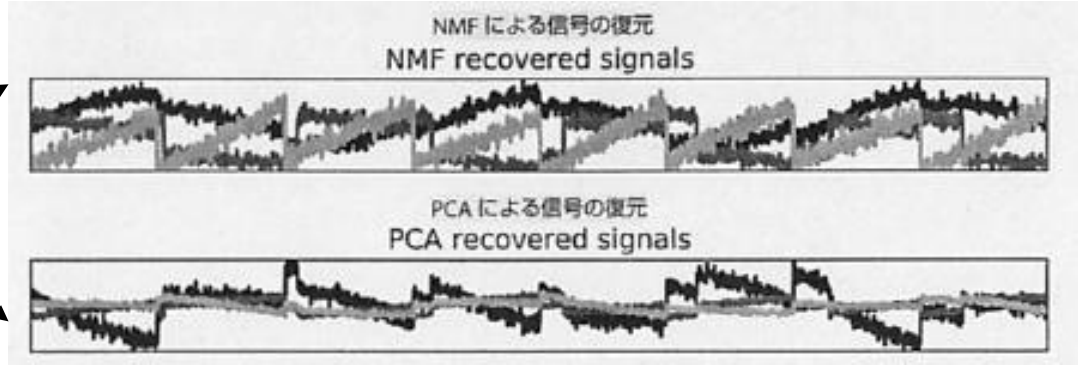
信号の復元

3つの信号源からの信号が組み合わされた信号に興味があるとする。



もとの信号は観測できず、この3つが混ざったものだけが観測できるとする。
NMFとPCAを用いてもとの信号を取り出したものを示す。



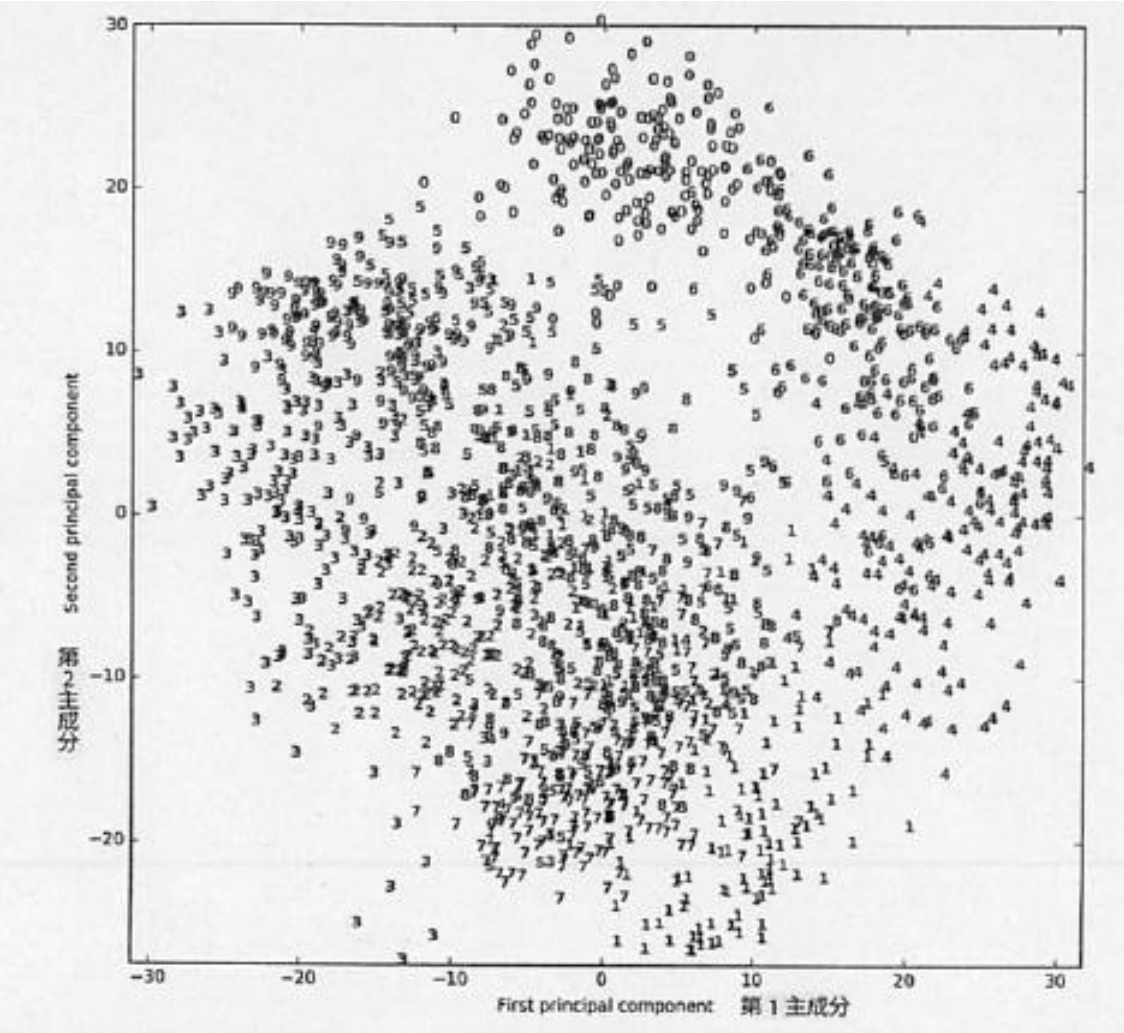


3.4.3 t-SNEを用いた多様体学習

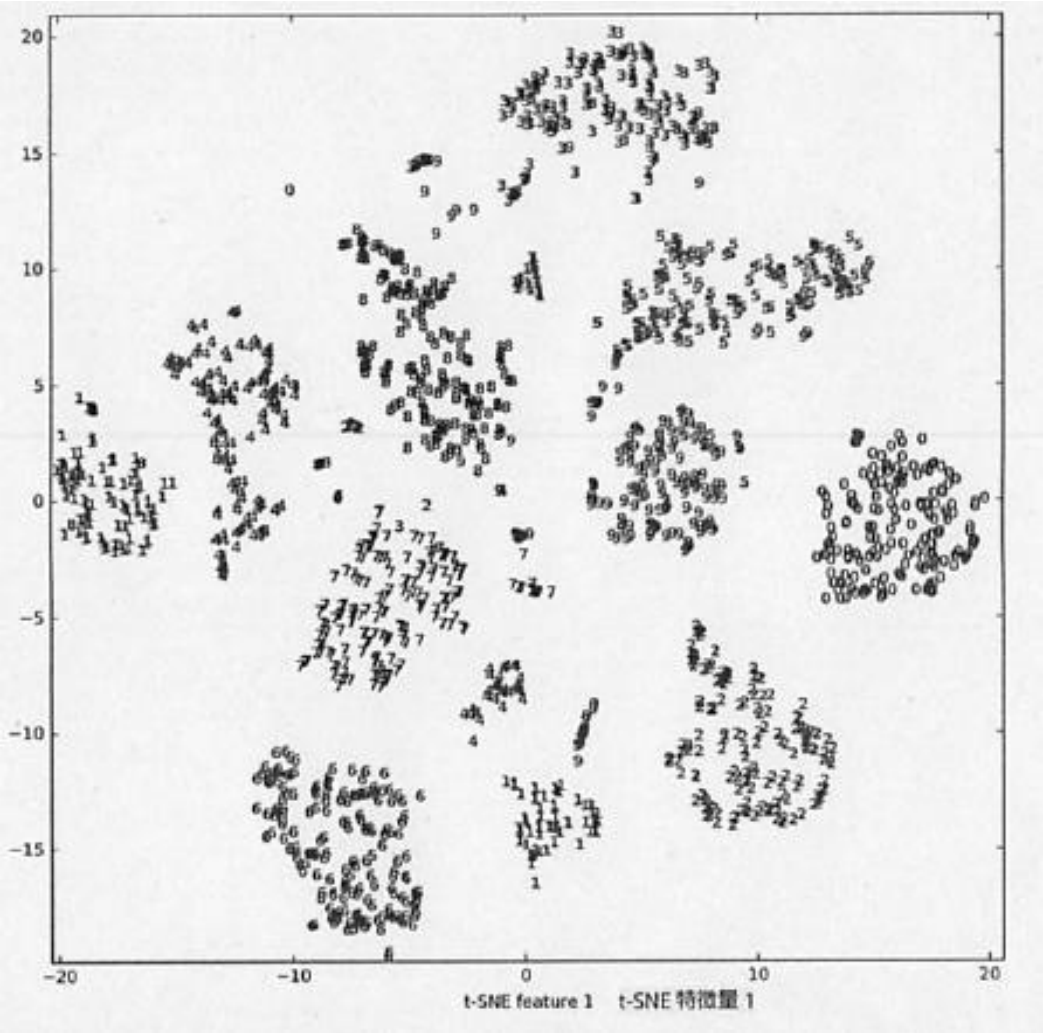
t-SNEは多様体学習アルゴリズムの一部である。

データポイントの距離を可能な限り維持する2次元表現を見つけようとする。まず最初にランダムな2次元表現を作り、そこから、もとの特徴空間で近いデータポイントを遠くに配置しようとする。つまり、どの点が近傍か示す情報を維持しようとする。

手書き数字データセットにPCAとt-SNEを適用させた結果を次ページに示す。



PCA



t-SNE

クラスが明確に分類されている