2020･05･18　ゼミ資料

Pythonではじめる教師なし学習
3章3.3節～6節

1116 17 9036

山口真哉

目的

次元削減アルゴリズムの
挙動に対する直観的な理解を得る

PCA

インクリメンタルPCA…　メモリに乗り切らないほど大きなデータセットに対して
データをメモリに乗るように切り分けて,
低次元のPCAを構築する.
切り分けるデータのサイズは自分で決められる.

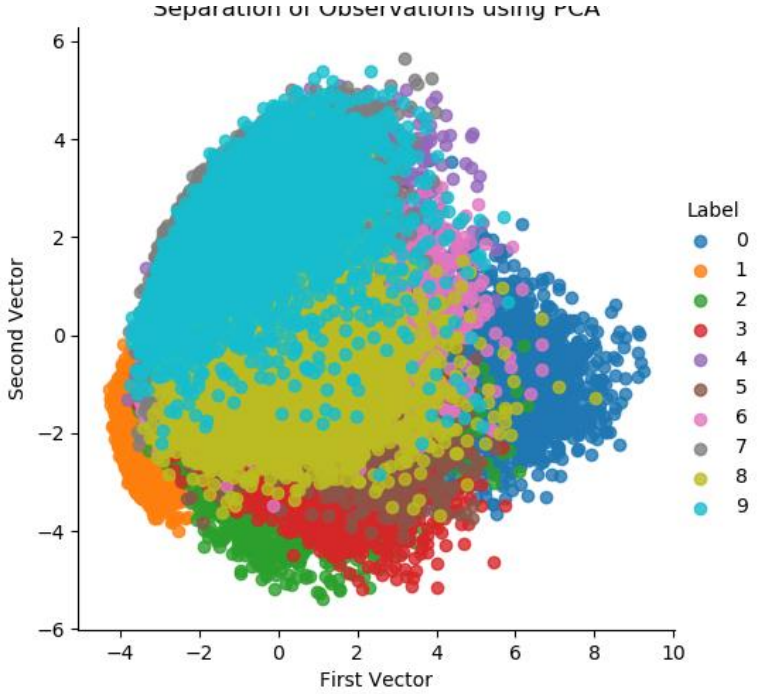
・このような時に有効　普通のPCAを実行してMLE（メモリ制限のエラー）が出るとき

・今回は ？　memory_profiler.memory_usage()
でPCAのメモリ使用量をチェックしたところ
513MBほどだったのでさすがに余裕そう.

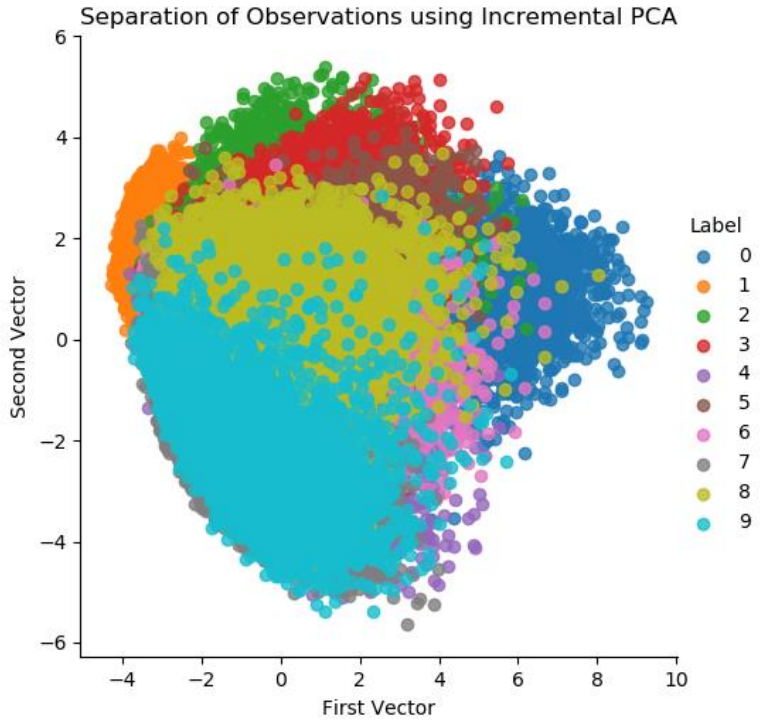
PCA

・比較

PCA



Incremental PCA



ハイパーパラメータは
元の次元数 = 784
(28ピクセル*28ピクセル)
だけ変更した.

ほぼ同じ結果

すごい！！

PCA

スパースPCA…　　大量のデータから有意な情報をうまく抜き出し，
ある程度の疎性（スパースさ）を保ったPCAを構築する.

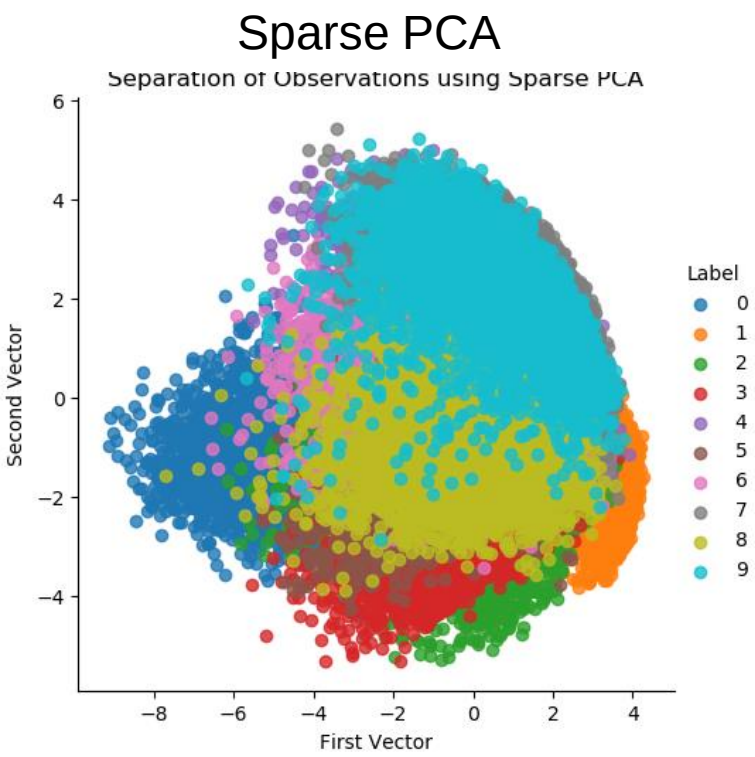
・このような時に有効　　各主成分の寄与率が小さいとき

・デメリット　　遅い

詳しくはここがわかりやすかった
https://stats.biopapyrus.jp/sparse-modeling/sparse-pca.html

PCA

・結果



ハイパーパラメータは

n_components = 100 (元の次元)
alpha = 0.0001 (スパースさ)
random_state = 2018 (シード値)

だけ変更.

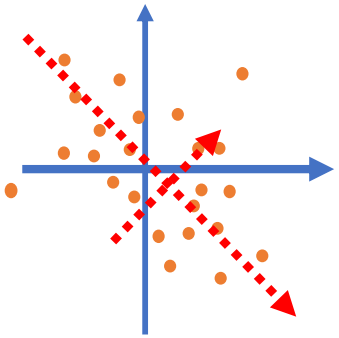
遅いのでサンプル数を50,000から10,000の変更

いろいろ試したが作者のような結果が得られなかったが,
うまく分離できていることがわかる.

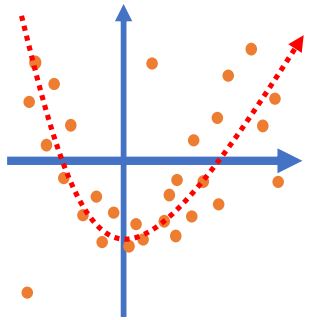
PCA

カーネルPCA…　今までのPCAは低次元空間に線形射影するが,
Kernel PCAは類似度関数を使用して,
非線形に射影して次元削減を行う.

・このような時に有効　　もとの特徴量集合が線形分離できないとき



線形分離できそう



線形分離できなさそう

PCA

今回はRBFカーネルを用いる.
RBFカーネルは以下の式で表される.

$$K(x, x') := \exp(\gamma ||x - x'||^2)$$

ただし,

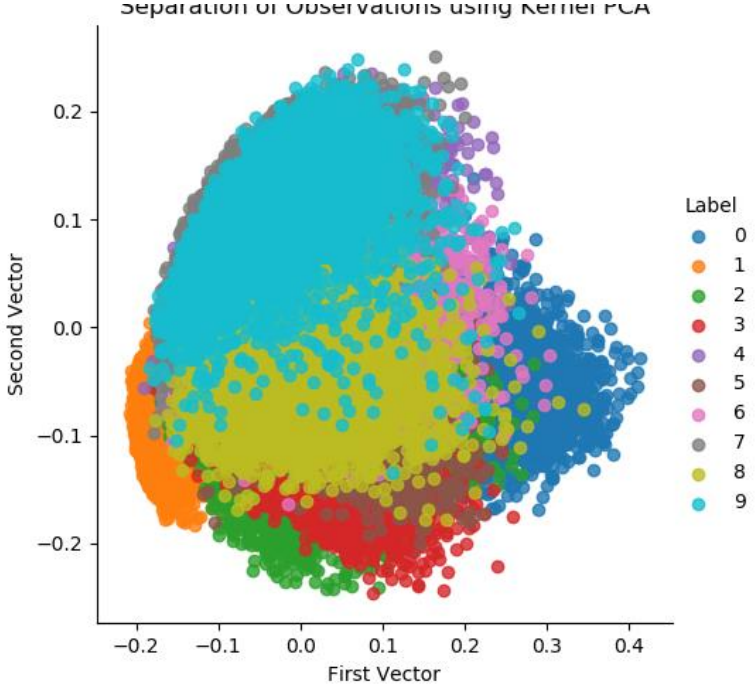
$||\cdot||$ はL2ノルム, $\gamma = \frac{1}{2\sigma^2}$

である

PCA

・結果

Kernel PCA



ハイパーパラメータは

n_components = 100 (元の次元)
Kernel = ‘rbf’ (RBF)
random_state = 2018 (シード値)

だけ変更.

遅いのでサンプル数を50,000から10,000の変更

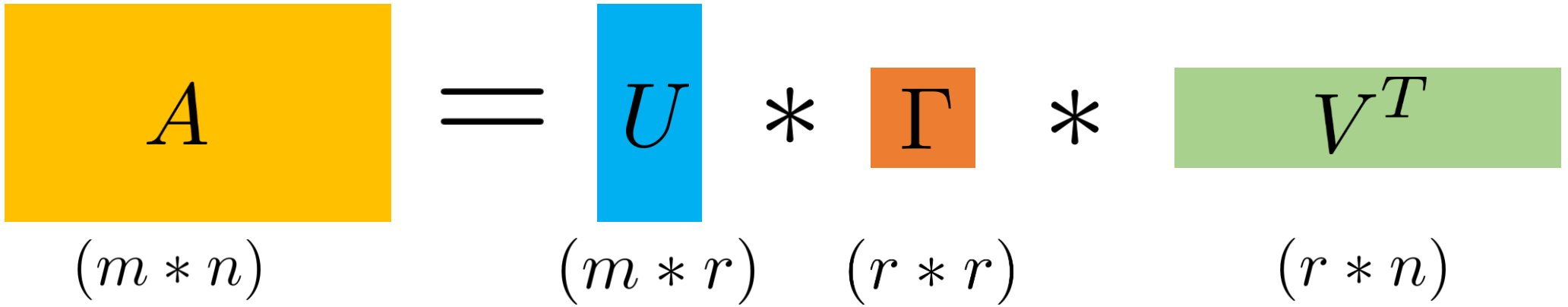
線形とほぼ一緒 →
RBFカーネルを用いても次元削減に効果がない.

特異値分解

特異値分解…　特徴量行列のRankより小さいRankの行列を作り
小さなRankの行列のベクトルの一部の線形結合
として元の行列を再構成できるようにする方法.

・PCAとの違い　PCAは共分散行列の固有値を用いるのに対し,
特異値分解は行列を分解する

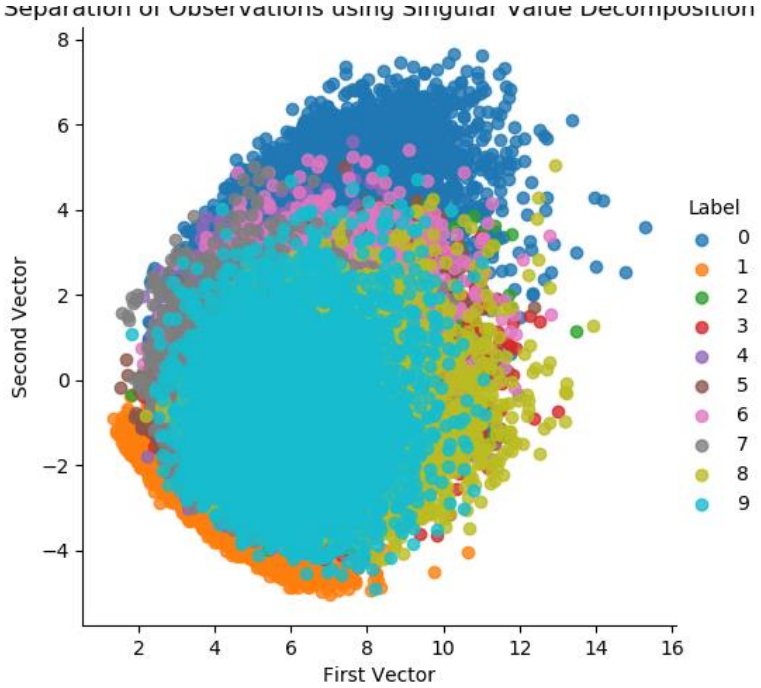
$A = U\Gamma V^T$ をしている様子



特異値分解

・結果

特異値分解



ハイパーパラメータは

n_components = 200 (元の次元)
random_state = 2018 (シード値)

だけ変更.

うまく分離できていて小さいランクの行列が
元の特徴量空間の重要な要素を捉えていることがわかる.

ランダム射影

ランダム射影…　Johnson-Lindenstrauss Lemmaに基づいた
ランダム行列による線形射影のアルゴリズム

詳しくはhttps://www.ipsj-kyushu.jp/page/ronbun/hinokuni/1007/B2/B2-3.pdf

・Johnson-Lindenstrauss Lemma(ジョンソン・リンデンシュトラウスの補題)

高次元空間を低次元空間に埋め込んだ場合，点間の距離がほぼ保存される.
言い換えると高次元から低次元に移しても元の特徴量の構造は保存される.

ランダム射影

ガウス型ランダム射影…　削減した特徴量空間で維持したい成分の数か
生成される次元数(eps)を指定できる.

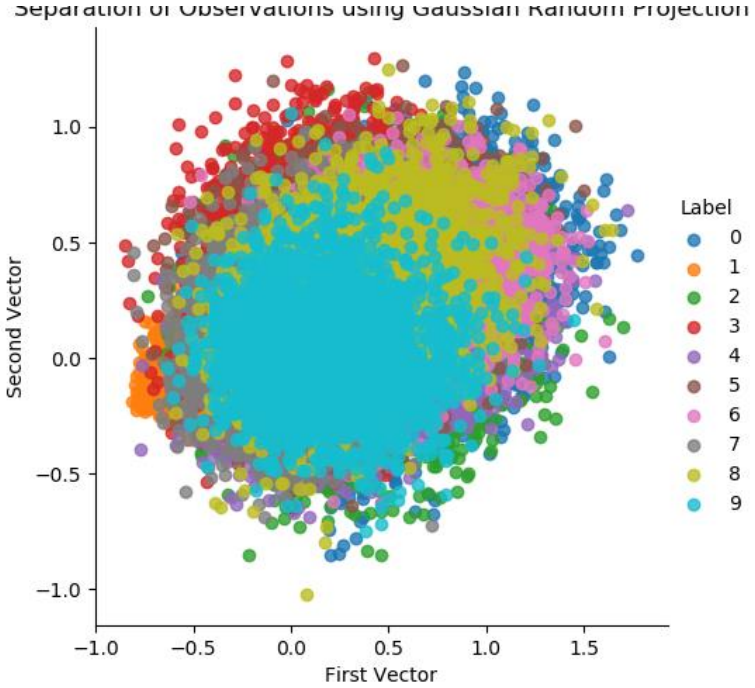
スパースランダム射影…

特徴量集合にある程度のスパースさを維持する.
ガウス型ランダム射影よりはるかに高速

ランダム射影

・結果

ガウス型ランダム射影



ハイパーパラメータは

eps = 0.5
random_state = 2018 (シード値)

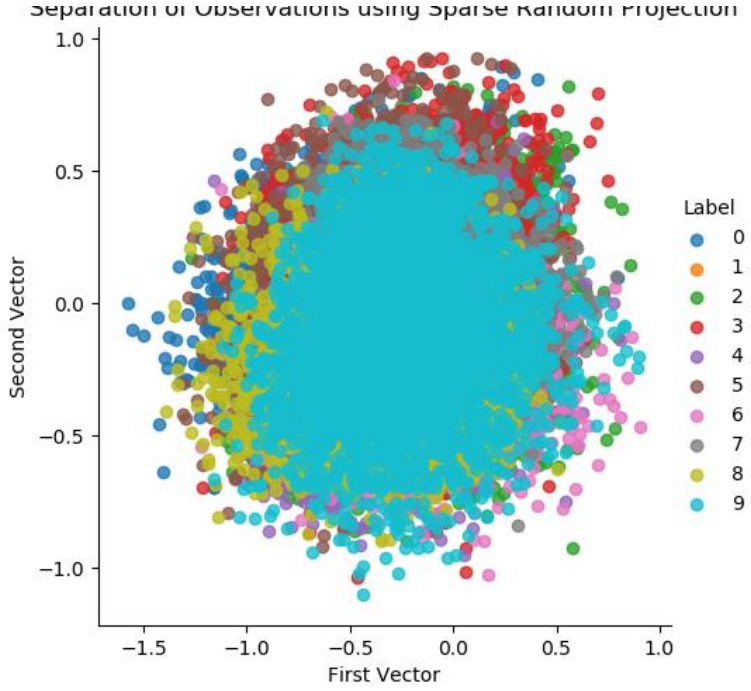
だけ変更.

線形射影ではあるがPCA系とは全く異なることがわかる.

ランダム射影

・結果

スパースランダム射影



ハイパーパラメータは

eps = 0.5
random_state = 2018 (シード値)

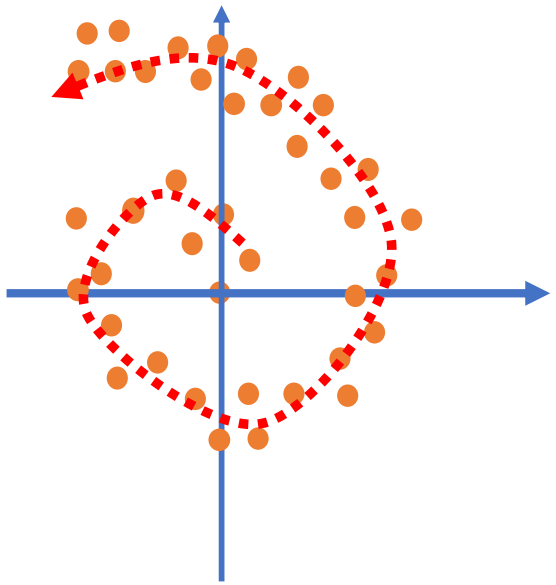
だけ変更.

ガウス型ランダム射影と似ているが少し異なる部分もある.

多様体学習

多様体学習…　データを高次元から低次元に線形射影するのではなく,
非線形に次元削減を行う.

例えば



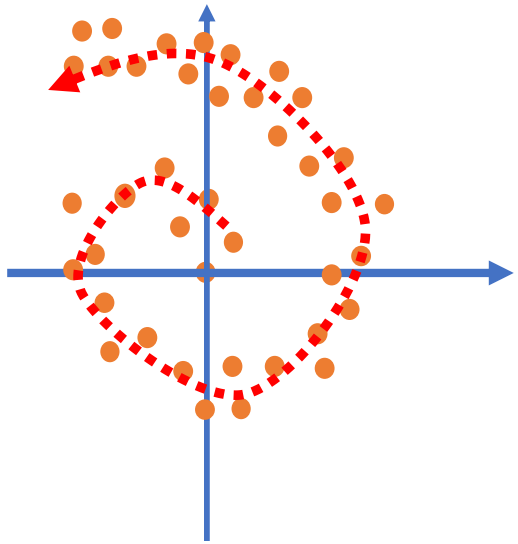
線形でやっても何もうれしくない

二次元に分布しているように見せかけて
一次元多様体上に分布している. → 多様体学習

詳しくは https://www.slideshare.net/kohta/risomap

多様体学習

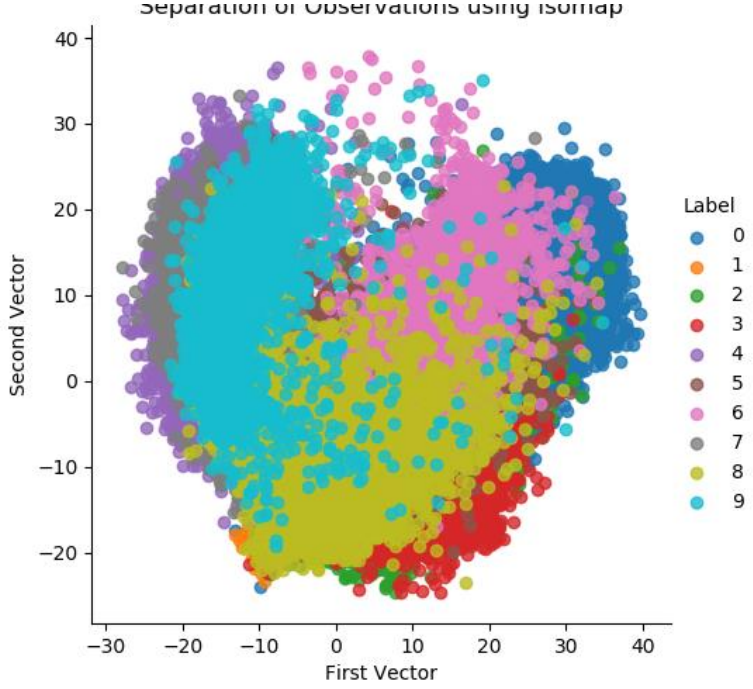
Isomap（等尺性マッピング）…　ユークリッド距離ではなく，曲線距離や測地線距離で
計算することで元の特徴量集合の低い次元の埋め込みを学習する．
言い換えると多様体上の近傍点に対する相対的な位置で
幾何構造を学習する．



多様体学習

・結果

Isomap



ハイパーパラメータは

n_neighbors = 5 (考慮する近傍点)
n_components = 10 (多様体の座標の数)

だけ変更.

これまでとは一風変わっている. (ハート型みたい)