主成分分析入門

高橋芳幸 (神戸大学大学院理学研究科)

目次

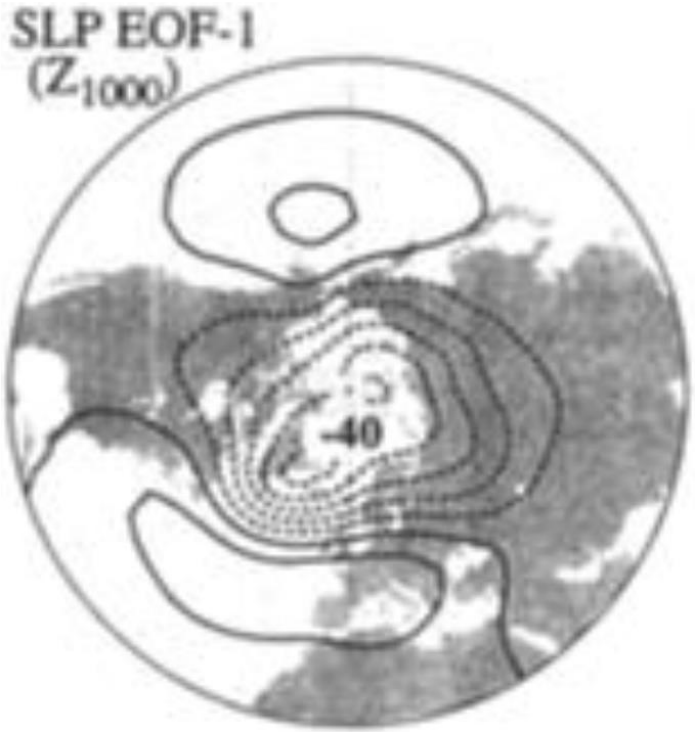
はじめに

- 主成分分析は, 多次元のデータの概要を把握するために用いられる手法.
  - 主成分分析 = Principal Component Analysis (PCA)
  - 多次元のデータから, ばらつきの大きな向き (軸) を探す
    - 結果としてデータの概要を把握することが容易になる (こともある)
- 利用例
  - 機械学習
    - 「次元削減」
    - 教師なし学習に分類されている.
  - 大気海洋
    - 「EOF 解析」(EOF = Empirical Orthogonal Function; EOF)
      - 気象学・海洋学の方言らしい
    - 大気構造の抽出に用いられている.
- 今回は, 主成分分析の基礎について紹介する.

大気の解析における例
北極振動

- 地球の北半球における冬季の高緯度大気の年々変動において卓越する構造
  - 北極域と中緯度の圧力がシーソーのように振動, のようにしばしば表現される.

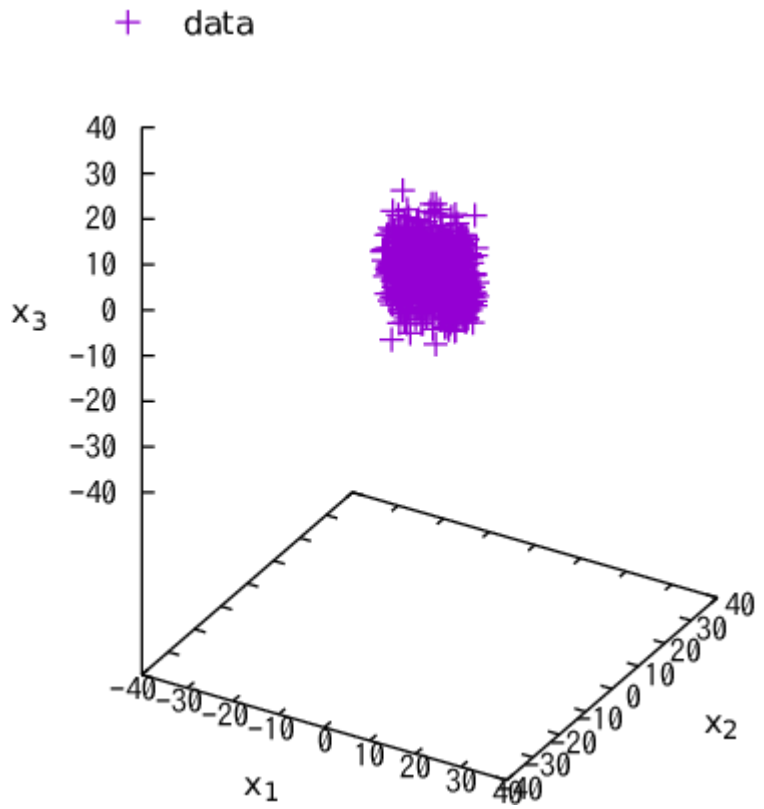
- 大気の話はこれだけ.
  - 「大気の変動に適用してみました」 な話は 1/12(水) 16:00 に神戸のセミナーで予定.
  - もし興味があれば連絡してください. 接続先をお知らせします.



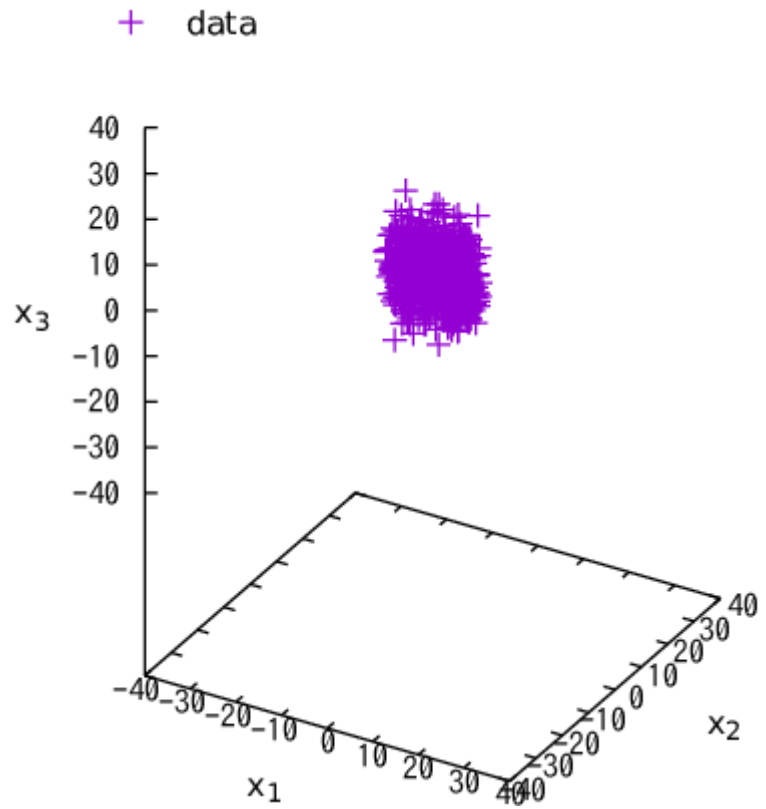
北半球の冬季の海面更生気圧 (を 1000 hPa 圧力面のジオポテンシャル高度の変動に変換したもの) における経験的直交関数 (EOF) 第一モード.
(Thompson and Wallace, 1998)

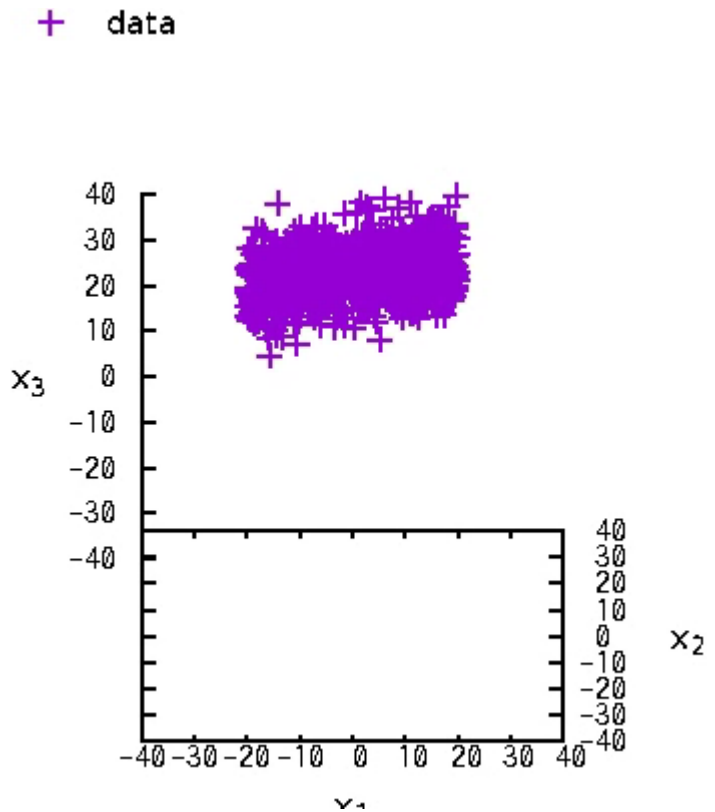
主成分分析－定性的な説明

- 3 変数の値の組 N 個
  $(x_{1,1}, x_{2,1}, x_{3,1}), (x_{1,2}, x_{2,2}, x_{3,2}),$
  $\cdots, (x_{1,N}, x_{2,N}, x_{3,N})$
  を考える.
  - ただし, 平均を引いてある.
  - 例えば,
    - あるクラスの生徒の算数と理科と英語の点数
    - 圧力, 温度, 比湿
    - …



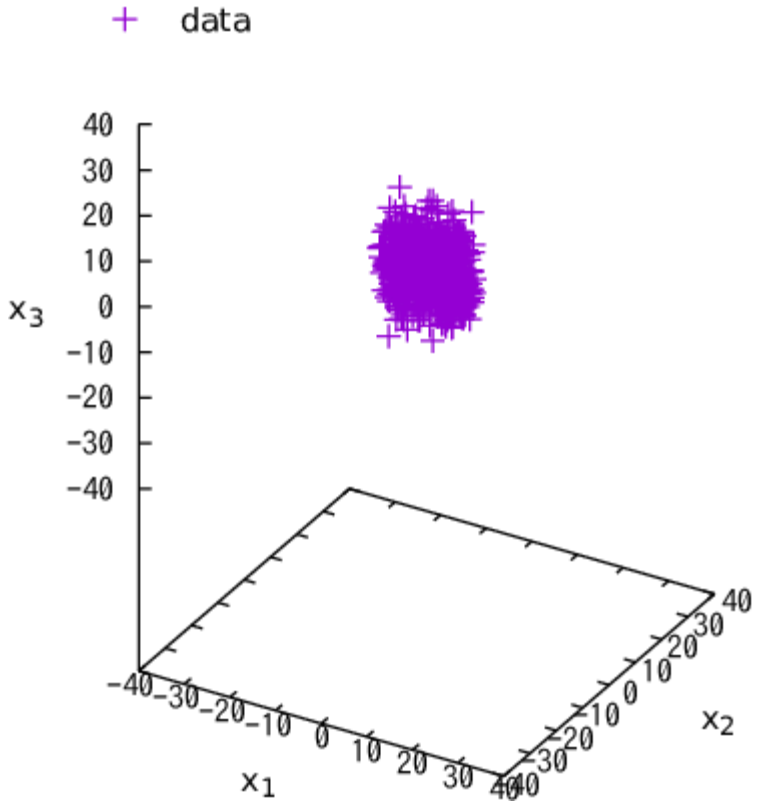
主成分分析 – 定性的な説明

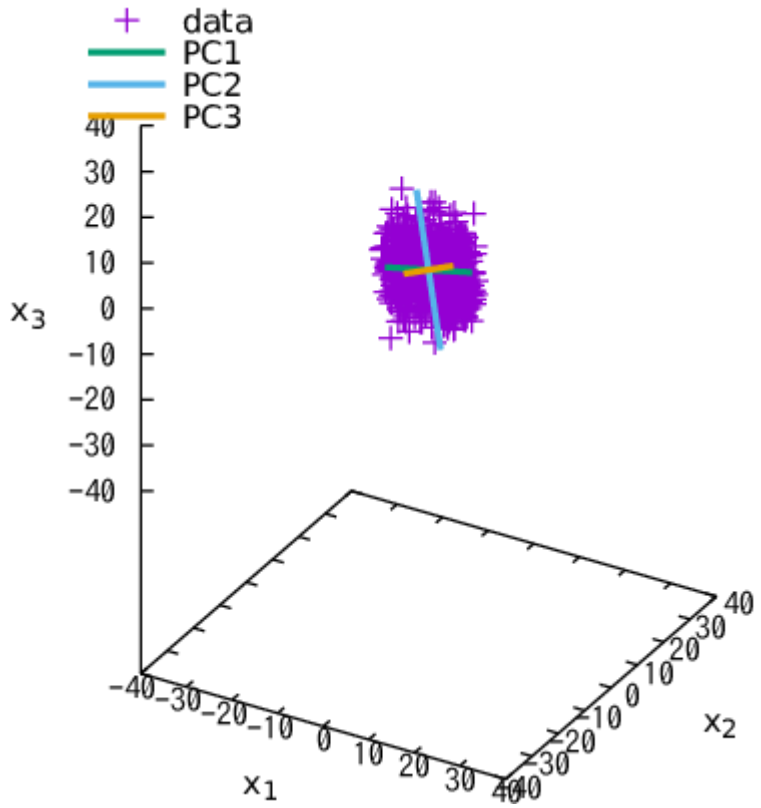


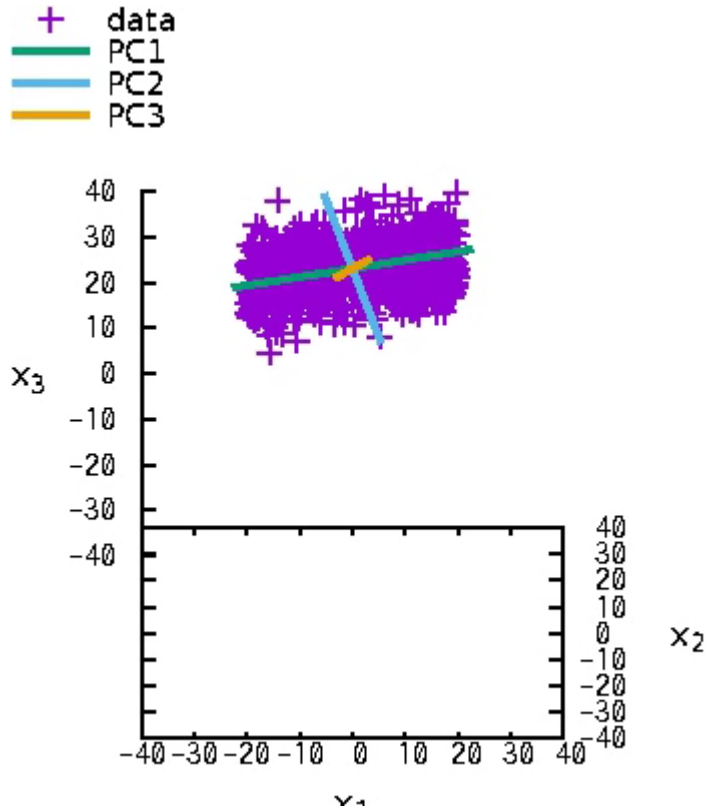


ぐるぐる回して様々な角度から確認しないと分布の概要がわからない.
(この例は 3 変数 (次元) なのでまだマシだが.)

主成分分析－定性的な説明

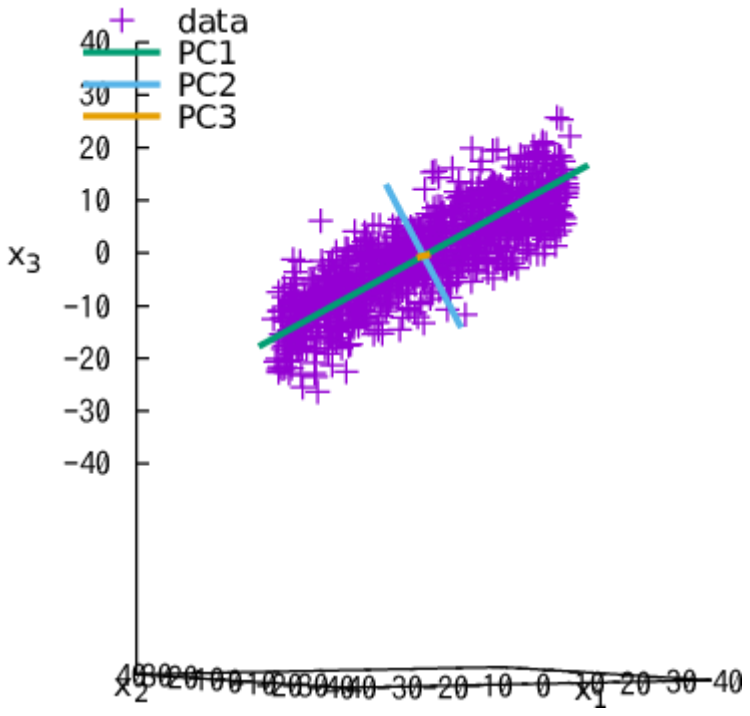


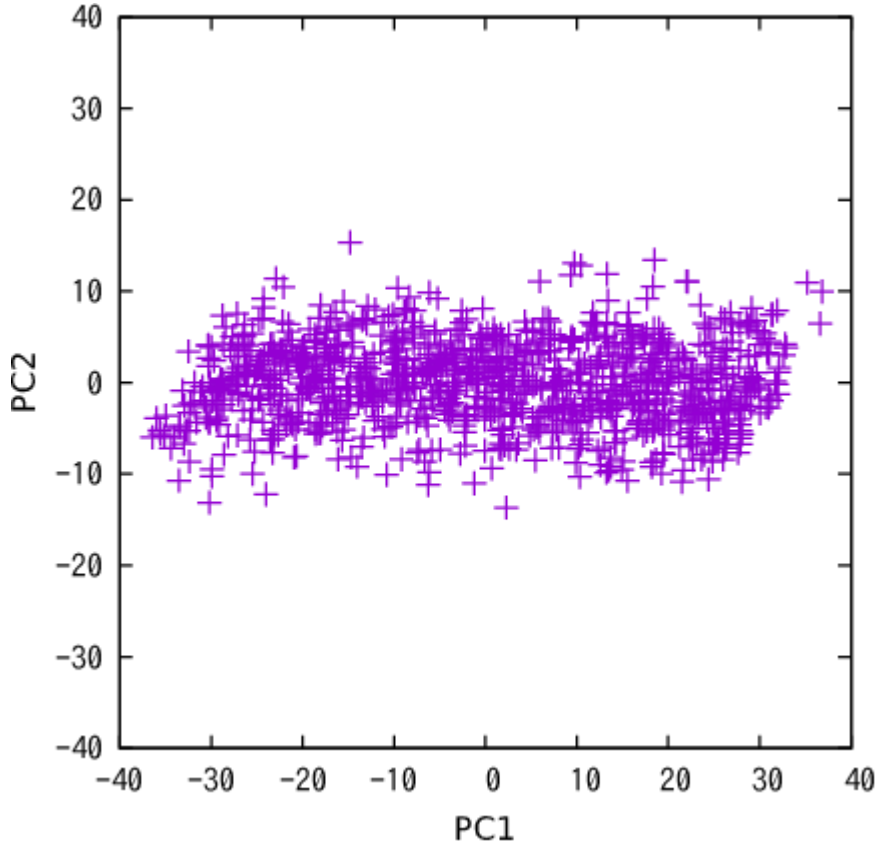




データがばらついている方向を探す.

主成分分析 – 定性的な説明





データのばらつきが最も大きい視点から分布を確認/描画.
そして, ばらつきの大きな軸に対して射影.
4 次元以上でも同様.

主成分分析 解説 (2 次元)

N 個のデータの組

$$(x_{1,1}, x_{1,2}), (x_{2,1}, x_{2,2}), \cdots, (x_{N,1}, x_{N,2})$$

において, それぞれのデータの平均 $\left(\overline{x_i} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} x_{n,i} \right)$ からの偏差,

$$(x'_{1,1}, x'_{1,2}), (x'_{2,1}, x'_{2,2}), \cdots, (x'_{N,1}, x'_{N,2})$$
$$= (x_{1,1} - \overline{x_1}, x_{1,2} - \overline{x_2}), (x_{2,1} - \overline{x_1}, x_{2,2} - \overline{x_2}), \cdots, (x_{N,1} - \overline{x_1}, x_{N,2} - \overline{x_2})$$

つまり,

$$\boldsymbol{x}'_n = (x'_{n,1}, x'_{n,2}) \quad (n = 1, 2, \cdots, N)$$

を考える.

主成分分析 解説 (2 次元)

このデータの, 原点を通る直線への射影を考える. 原点を通る直線の単位接線ベクトルを $\boldsymbol{a} = (a_1, a_2) \quad (|\boldsymbol{a}|^2 = a_1^2 + a_2^2 = 1)$ とすると, 射影は

$$z_n = \boldsymbol{x}'_n \cdot \boldsymbol{a} = x'_{n,1} a_1 + x'_{n,2} a_2$$

となる. ここで z_n は, 直線に射影された点 (直線への垂線と直線の交点) と原点との距離 (符号付き) であり, すべての点を射影すると, 元の二次元分布を一次元分布に変換したことになる.

主成分分析 解説 (2 次元)

ここで, 射影された点 z_n の分散が最も大きくなる直線 (の接線ベクトル) を求める. (このとき, 最も情報の損失が小さくなる.)

射影された点 z_n の分散は,

$$\sigma_z^2 = \frac{1}{N-1}\sum_{n=1}^{N}(z_n - \overline{z})^2 = \frac{1}{N-1}\sum_{n=1}^{N}\{(\boldsymbol{x}'_n - \overline{\boldsymbol{x}'})\cdot\boldsymbol{a}\}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}'_n\cdot\boldsymbol{a})^2$$

であるから, 求めたいのは, 分散 σ_z^2 が最大となる $\boldsymbol{a}$ である.

主成分分析 解説 (2 次元)

言い換えると, $|\boldsymbol{a}|^2 = a_1^2 + a_2^2 = 1$ の条件の下で $\dfrac{\partial \sigma_z^2}{\partial a_1} = \dfrac{\partial \sigma_z^2}{\partial a_2} = 0$ となる $\boldsymbol{a}$ を求めることである.

ラグランジュの未定乗数法を用いると, 定数 λ を用いて

$$F(\boldsymbol{a}, \lambda) = \sigma_z^2 - \lambda(a_1^2 + a_2^2 - 1)$$

を定義して,

$$\frac{\partial F}{\partial a_1} = 0, \quad \frac{\partial F}{\partial a_2} = 0, \quad \frac{\partial F}{\partial \lambda} = 0$$

を満たす $\boldsymbol{a}$ を求めれば良い.

主成分分析 解説 (2 次元)

計算してみると,

$$
\begin{aligned}
\frac{\partial F}{\partial a_1} &= 2\frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}'_n \cdot \boldsymbol{a})x'_{n,1} - 2\lambda a_1 \\
&= 2\frac{1}{N-1}\sum_{n=1}^{N}(x'_{n,1}a_1 + x'_{n,2}a_2)x'_{n,1} - 2\lambda a_1 \\
&= 2\frac{1}{N-1}\sum_{n=1}^{N}(x'_{n,1})^2 a_1 + 2\frac{1}{N-1}\sum_{n=1}^{N} x'_{n,1}x'_{n,2}a_2 - 2\lambda a_1 \\
&= 2\sigma^2_{x'_1}a_1 + 2C_{x'_1,x'_2}a_2 - 2\lambda a_1 = 0,
\end{aligned}
$$

主成分分析 解説 (2 次元)

同様にして計算すると,

$$\frac{\partial F}{\partial a_1} = 2\sigma_{x'_1}^2 a_1 + 2C_{x'_1,x'_2} a_2 - 2\lambda a_1 = 0,$$
$$\frac{\partial F}{\partial a_2} = 2C_{x'_1,x'_2} a_1 + 2\sigma_{x'_2}^2 a_2 - 2\lambda a_2 = 0,$$
$$\frac{\partial F}{\partial \lambda} = -(a_1^2 + a_2^2 - 1) = 0$$

主成分分析 解説 (2 次元)

整理すると,

$$\boldsymbol{V}\boldsymbol{a} = \lambda \boldsymbol{a}$$

$$\boldsymbol{V} = \begin{pmatrix} \sigma^2_{x'_1} & C_{x'_1,x'_2} \\ C_{x'_1,x'_2} & \sigma^2_{x'_2} \end{pmatrix}$$

$$a_1^2 + a_2^2 = 1$$

と書ける. $\boldsymbol{V}$ は分散と共分散を要素とする行列 (分散共分散行列) で, 問題は $\boldsymbol{V}$ の固有値問題に帰着する.

主成分分析 実装

- 今回は, 分散共分散行列の固有値問題を LAPACK の DSYEV サブルーチン (倍精度実対象行列の固有値, 固有ベクトルを計算) を用いて解く.
  - LAPACK - Linear Algebra PACKage
    - 数値線形代数計算のための Fortran ライブラリ
    - debian ならば liblapack-dev パッケージで利用可能
  - コンパイル例

```
$ gfortran  –I /usr/lib/x86_64-linux-gnu/lapack ¥
            –L /usr/lib/x86_64-linux-gnu/lapack ¥
            -o pca  pca.f90  -llapack
```

- なお, 分散共分散行列の固有値問題を解く代わりに, データを収めた行列 (サンプル数×変数の数) を特異値分解して特異値と特異ベクトルを求めることも等価である.
- 補足
  - 「主成分分析」「ワイン」などの単語を使って検索すると, Python や R を使った分析の解説をたくさん見つけることができる.

主成分分析例

主成分分析例

- あやめの種類
- (赤)ワイン成分

主成分分析例：あやめ
データの情報

- Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- 取得元
  - Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
  - https://archive.ics.uci.edu/ml/datasets/Iris
- 変数
  - がくの長さ (cm)
  - がくの幅 (cm)
  - 花弁の長さ (cm)
  - 花弁の幅 (cm)
  - 種類
    - Iris setosa (ヒオウギアヤメ)
    - Iris Versicolour (ブルーフラッグ)
    - Iris Virginica
- サンプル数：150

あやめの構造

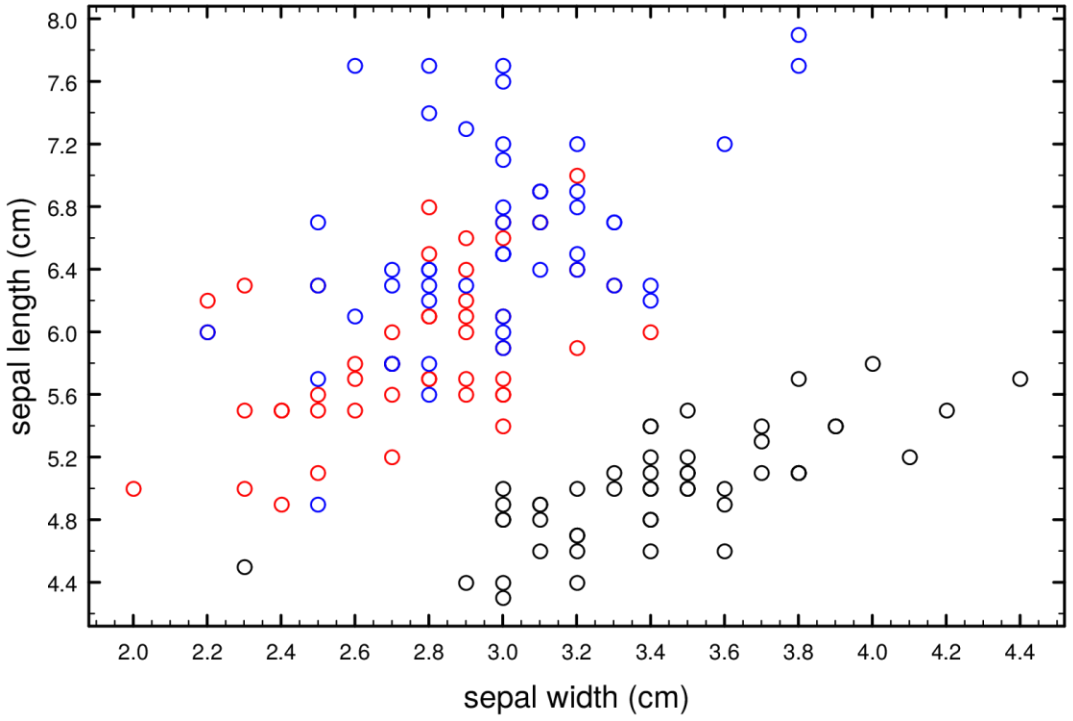
https://camo.qiitausercontent.com/a736bf6c242152ac2d4e365786b024040ca43247/68747470733a2f2f71696974612d696d6167652d73746f726e732e73332e61702d6e6f727468656173742d312e616d617a6f6e6177732e636f6d2f302f343336303838382f65333530623

主成分分析例：あやめ
分析の方針

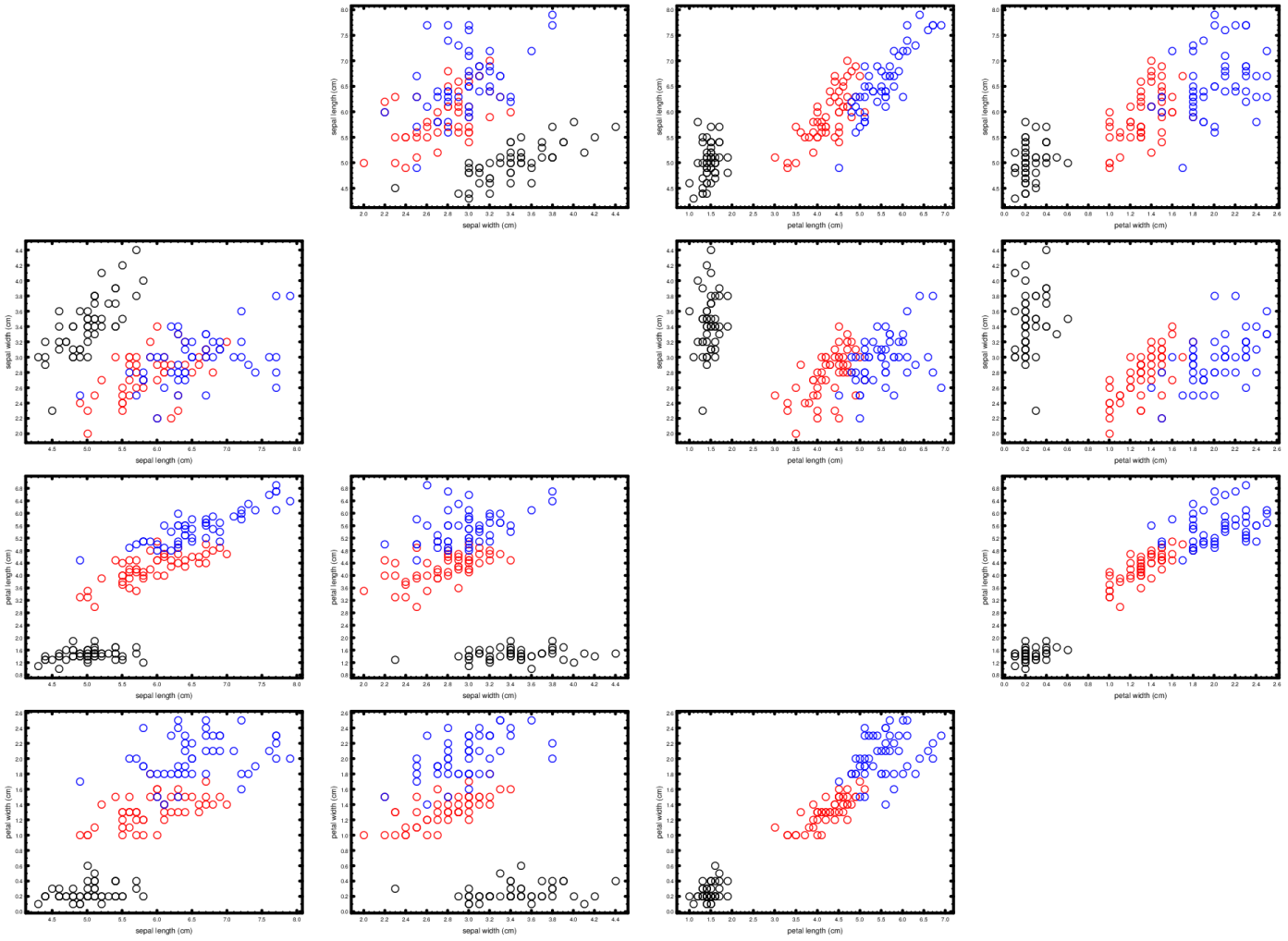
- がく, 花弁の長さと幅で, あやめの種類を分類できそうか?
  - がく, 花弁の長さと幅を「説明変数」として標準化して用いる.
    - 種類は「目的変数」扱い.

- 用語の説明 (機械学習用語)
  - 説明変数：予測に使う変数
  - 目的変数：予測する変数

データ概観 1

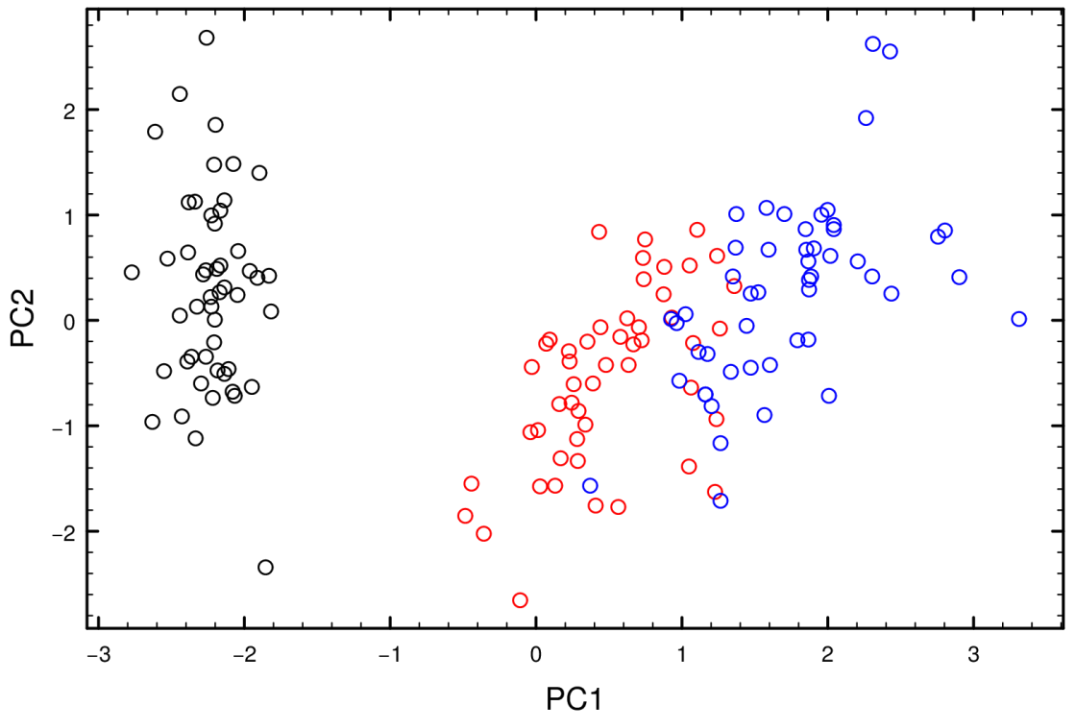

8.0
7.6
7.2
6.8
6.4
6.0
5.6
5.2
4.8
4.4
sepal length (cm)
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
sepal width (cm)

データ概観 2



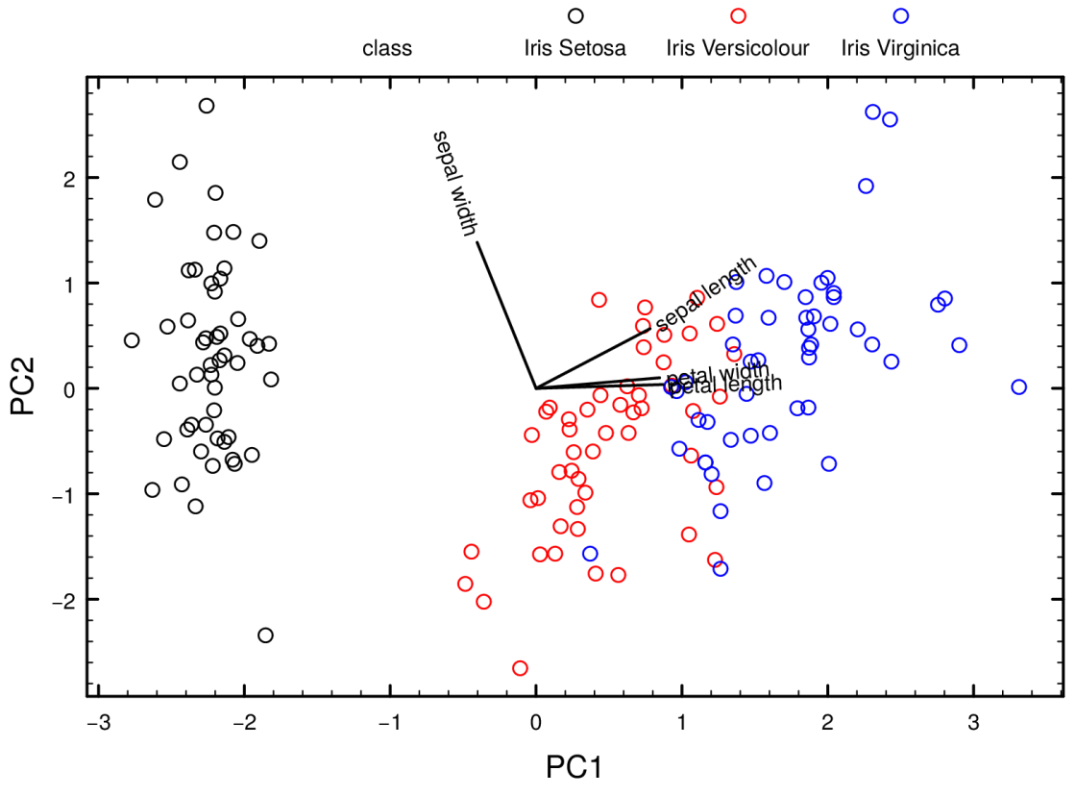
あやめの種類は,
花弁の長さや幅に
対して比較的わかり
やすく分布.
がくの幅にはほぼ
依らない.

主成分分析結果



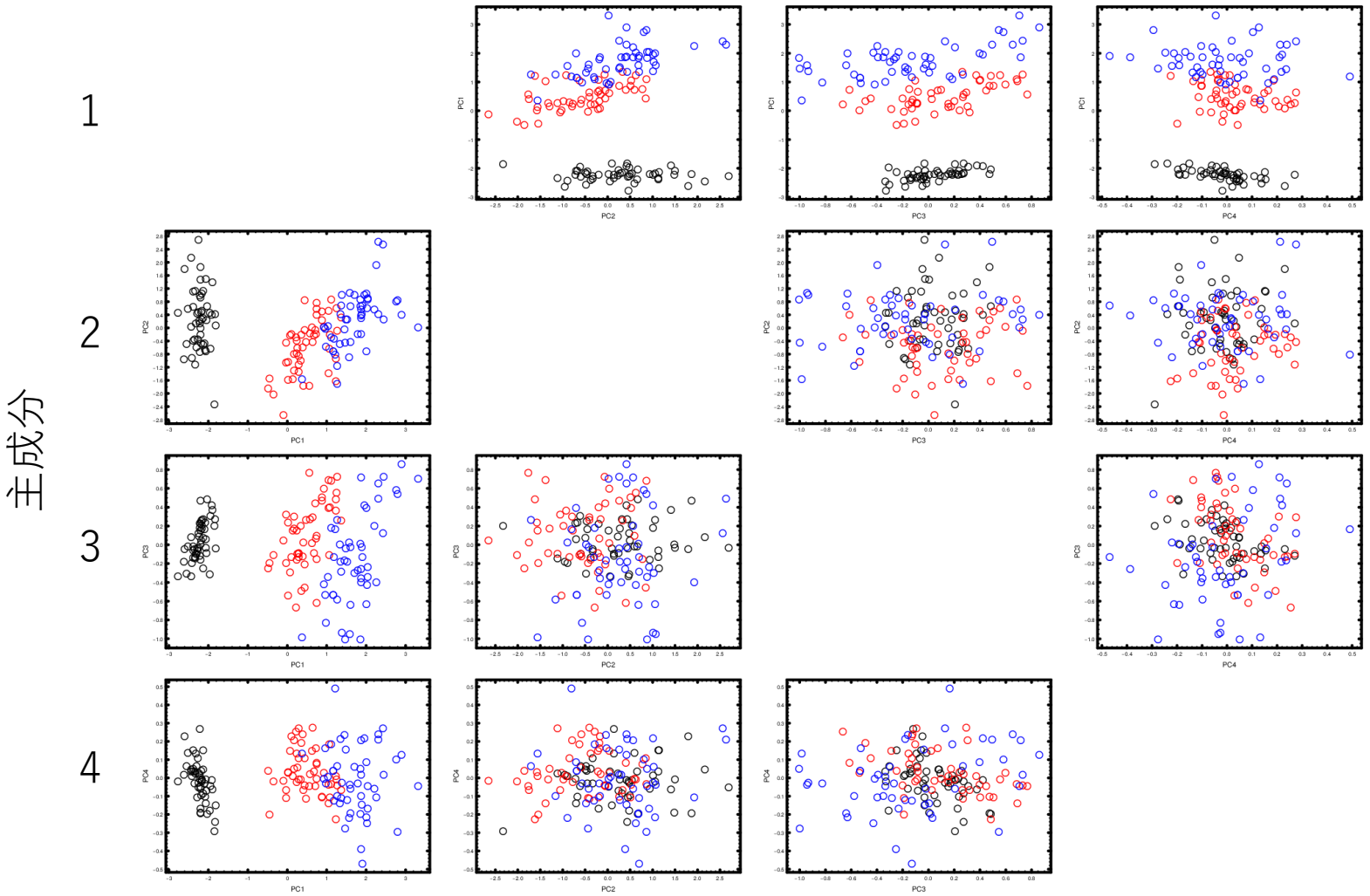
- あやめの種類は,
  - 第 1 主成分に対して比較的わかりやすく分布.
    - すべて綺麗に分かれるわけではなさそうだけど
  - 第 2 主成分にあまり依存しない.

主成分分析結果



- 花弁の長さ, 幅は第 1 主成分とほぼ平行
- がくの幅と第 1 主成分はほぼ直交
- 元データから予想できる結果(なので安心).

主成分分析結果



このデータは, 第 1 主成分に
対してあやめの種類が広く分布.
高次主成分に対しては,
あやめの種類はほぼ依らない.

主成分分析例：(赤)ワイン データの情報

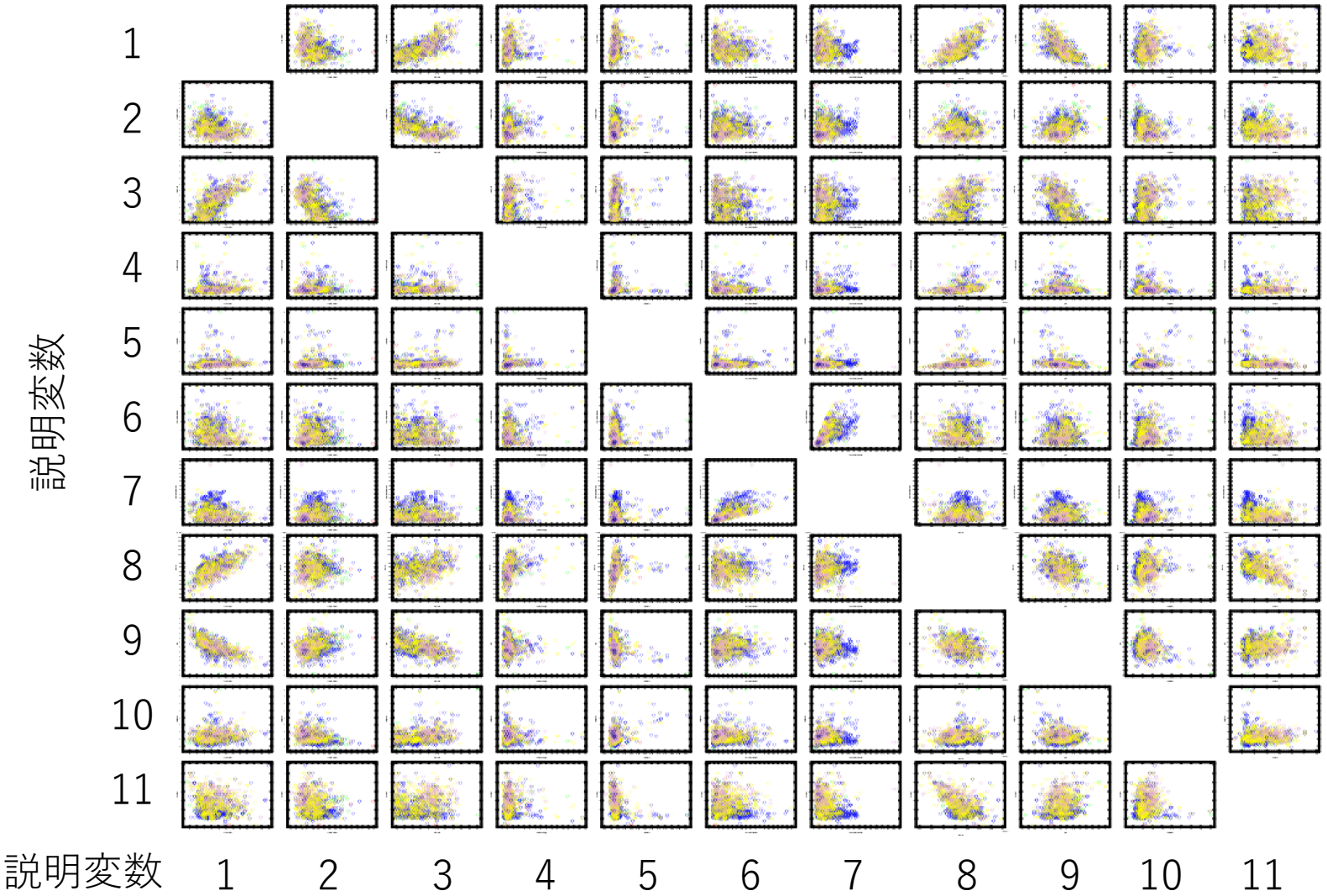
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- 取得元
  - https://archive.ics.uci.edu/ml/datasets/Wine+Quality
- ただし, 今回使うのは赤ワインのデータのみ.
- 参考ページ
  - https://tjo.hatenablog.com/entry/2015/11/26/190000
  - https://sudillap.hatenablog.com/entry/2013/04/27/203200

- 変数
  1. fixed acidity (酒石酸濃度)
  2. volatile acidity (酢酸濃度)
  3. citric acid (クエン酸濃度)
  4. residual sugar (残糖濃度)
  5. chlorides (塩化ナトリウム濃度)
  6. free sulfur dioxide (遊離SO2濃度)
  7. total sulfur dioxide (総SO2濃度)
  8. density (密度)
  9. pH
  10. sulphates (硫化カリウム濃度)
  11. alcohol (アルコール度数)
  12. quality (質; 0-10)
- サンプル数：1599
  - 今回使わない白ワインデータのサンプル数は 4898.

主成分分析例：(赤)ワイン
分析の方針

- ワインの成分で, ワインの質を分類できそうか?
  - ワインの成分「説明変数」として標準化して用いる.
    - 質は「目的変数」扱い.

データ概要



説明変数

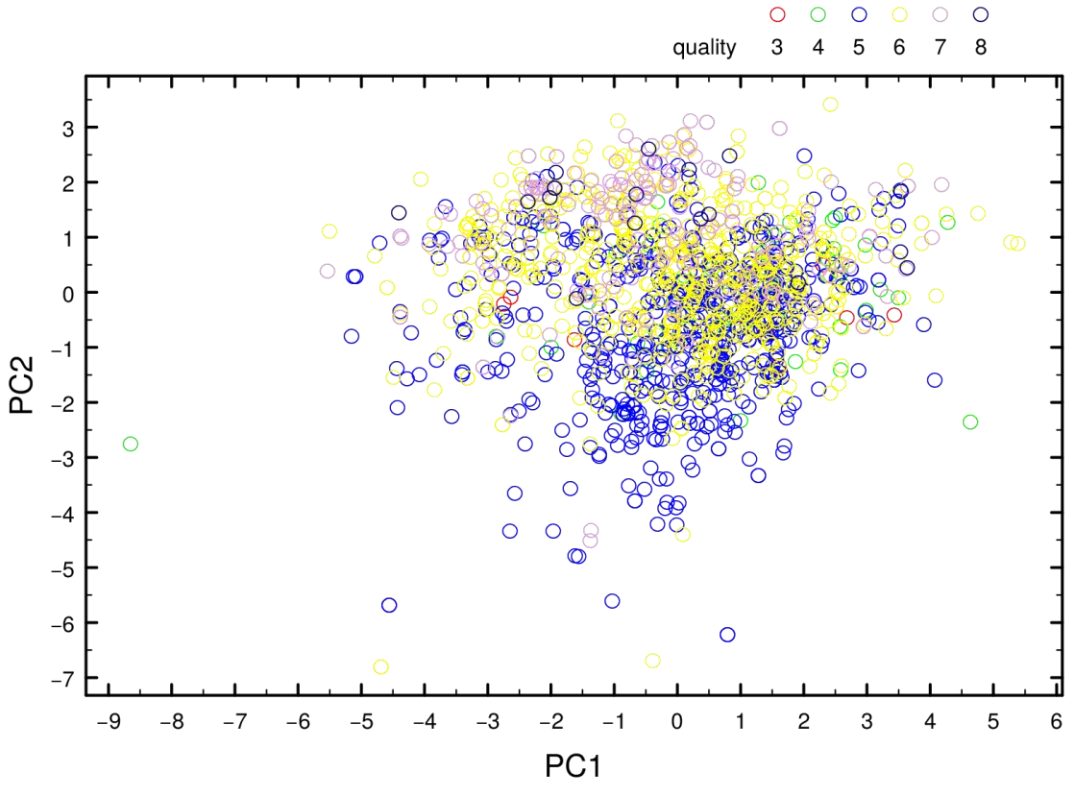
1. fixed acidity (酒石酸濃度)
2. volatile acidity (酢酸濃度)
3. citric acid (クエン酸濃度)
4. residual sugar (残糖濃度)
5. chlorides (塩化ナトリウム濃度)
6. free sulfur dioxide (遊離SO_2濃度)
7. total sulfur dioxide (総SO_2濃度)
8. density (密度)
9. pH
10. sulphates (硫化カリウム濃度)
11. alcohol (アルコール度数)

目的変数

quality (質; 0-10) [色付け]

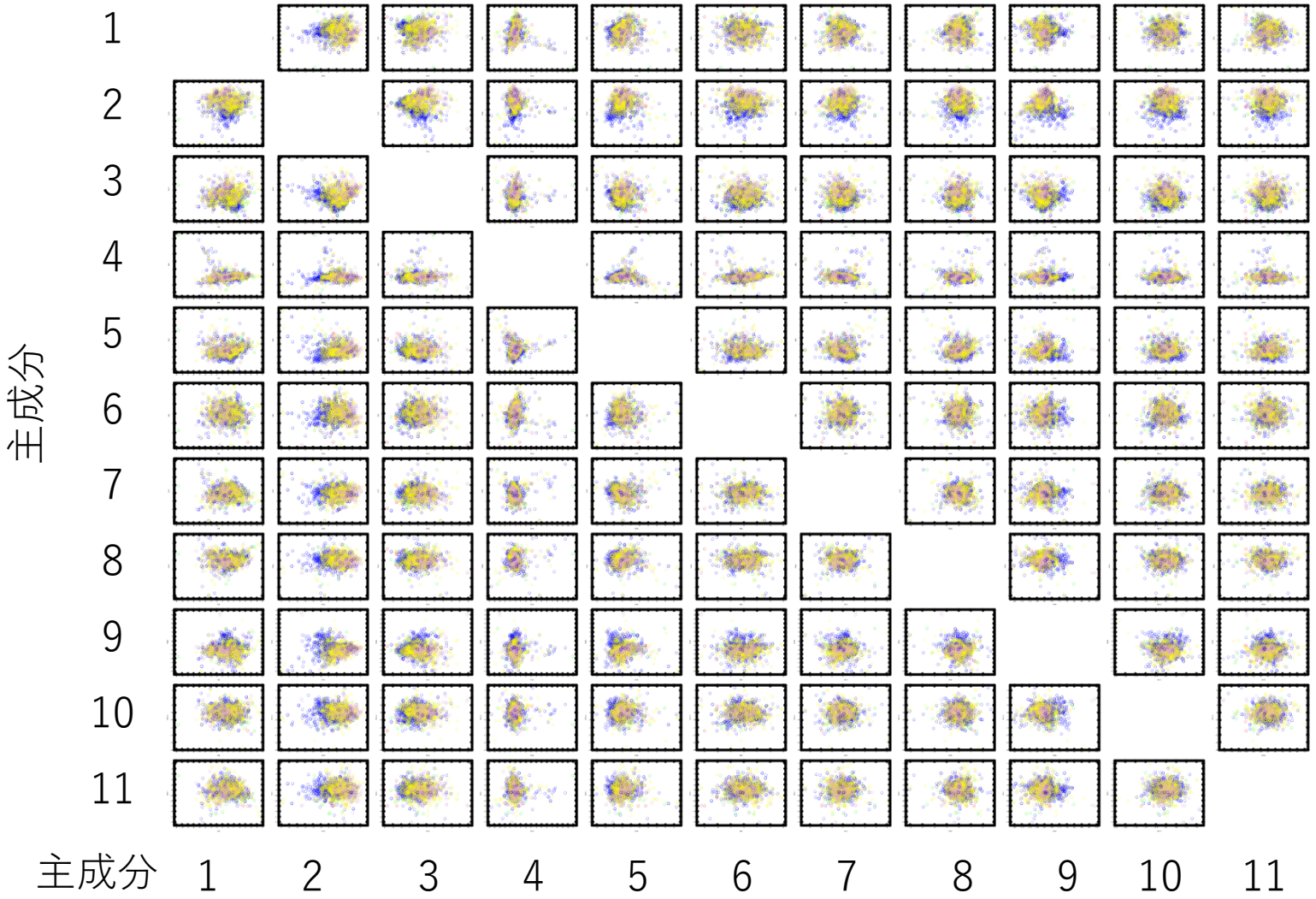
いくつか特徴はあるが, たくさんあってよくわからない

主成分分析結果



- ワインの質 (色) は, 第 2 主成分の方向に広がっているように見える.
  - 主成分第 1 モードの方向にはほぼ依らないように見える.
- しかし, 第 2 主成分に対してワインの質は重なりが大きい.

主成分分析結果

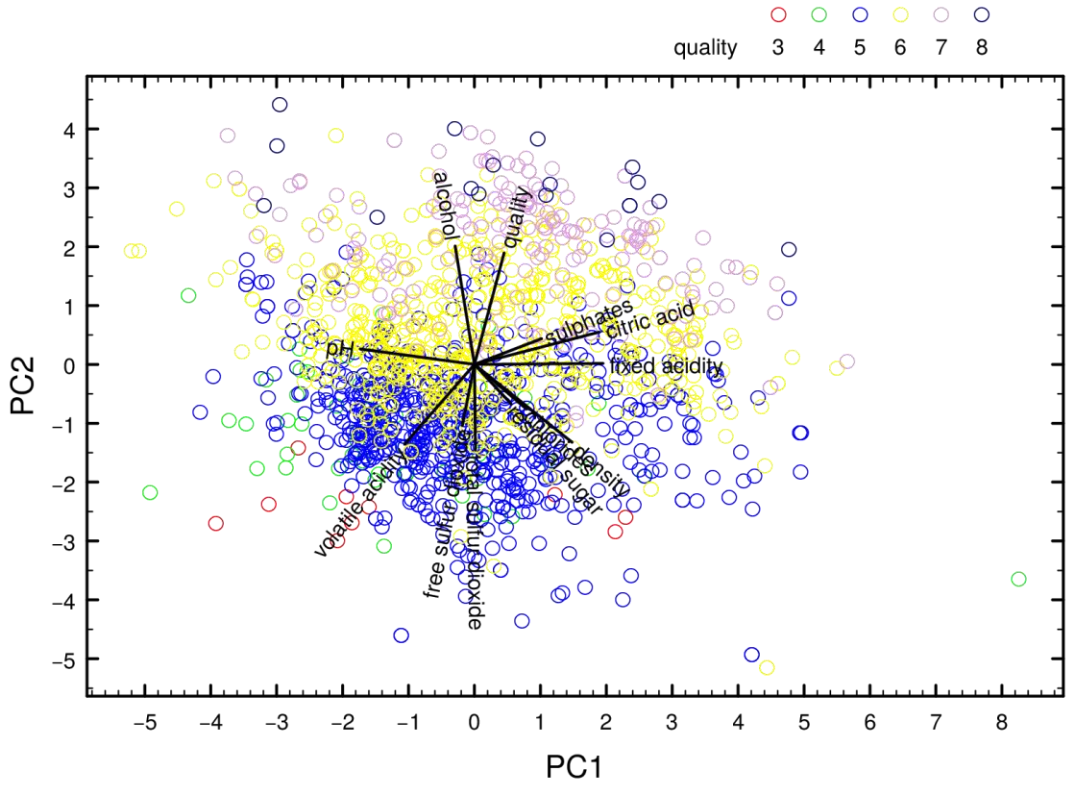


たくさんあってよくわからない

主成分分析例：(赤)ワイン
分析の方針

- ワインの質に貢献する成分は何か?
  - ワインの成分と質を「説明変数」として標準化して用いる.

主成分分析結果



- ワインの質の軸は, 第 2 主成分の方向.
- ワインの質の軸 (~第 2 主成分) に近い向きの軸
  - アルコール
  - 遊離SO_2濃度
  - 総SO_2濃度

… 結局酔えばいいってこと?
(ではないということだろう)

まとめ

- 主成分分析は, 多次元のデータの概要を把握するために用いられる手法であり, 機械学習や大気海洋の業界で用いられている.
  - 大気海洋では「EOF 解析」と呼ばれており, 北極振動などの発見に利用された.
- 主成分を求めるためには, データの分散共分散行列の固有値問題を解けばよい.
  - あやめと赤ワインのデータの分析例を示した.
- 主成分分析によって必ずしも求めたいものが得られるとは限らない. しかし, データの分布において主要な役割を果たす要因に関する手掛かりが得られる (こともある).

- EOF 解析 (大気データの解析例) については別のセミナー (1/12(水) 16:00) で紹介したい.