1　主成分分析

大量の変数を持つデータ群の場合，損失が少なくデータの関係の把握が行える主成分分析を学ぶ.

主成分分析の手順は，データの分散が最も大きくなる方向に軸をとり，これを第 1 主成分とする．次に分散が大きくなる方向に軸を取り、これを第 2 主成分とする．これらを元のデータの次元分だけくり返す行うことで，データの特徴を抽出し，データの次元を圧縮することができる.

下記の図が使用するデータである.

	コク	香り	酸味
S マルタ	-0.116248	0.116248	1.5275252
モーニング S	-1.276724	-1.245682	0.0727393
BOSS	1.0462287	-0.415227	0.8001323
FIRE	1.0462287	0.4152274	-0.654654
サンマルタ	1.0462287	1.2456822	1.5275252
BLACK 無糖	0.4649906	-0.415227	-0.654654
UCCB	-1.278724	1.2456822	-1.382047
ジョージア B	-1.278724	-1.245682	-1.382047
ROOT	-0.697486	-1.245682	0.0727393
WANDA 無糖	1.0462287	0.4152274	0.0727393

表 1　コーヒーのコク，香り，酸味を数値化したもの

学習データを

$$x_i = (x_i, ..., x_i d)^T (i = 1, ..., N)$$

とし，

データ行列を $X = (x_i, ..., x_N)^T$，原点を平均としたデータ行列を $\overline{X} = (x_i - \overline{x}, ..., x_N - \overline{x})^T$ としたとき，次のように定義された共分散行列

$$S = \frac{1}{N} X^{-T} X$$

に対して固有ベクトルを求める．

$$S u_i = \lambda_i u_i$$

この固有ベクトル u_i この固有ベクトル λ_i は式変形によって各ベクトルに対応する分散とみなせ，この分散が大きくなる固有ベクトルを用いて図のような平面や方向を決定できる．

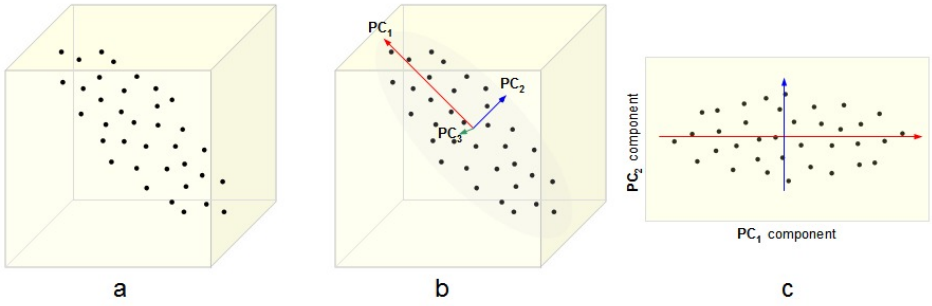

図 1　主成分分析の手順 [1]

今回は練習も兼ね，R を用いて主成分分析を行った．

データの読み込みを行う．

```
> data = read.csv('./caffee.csv',head=T,row.names=1)
```

主成分分析を行う．

```
> pca = prcomp(data, scale=T)
> pca
```

Standard deviations が分散の標準偏差で Rotation が固有ベクトルになっている．PC1, PC2 は第 1 主成分，第 2 主成分となる．

```
  Standard deviations:
  [1] 1.3182559 0.9086843 0.6606771

  Rotation:
             PC1          PC2         PC3
  コク 0.6579049 -0.02537522 -0.7526734
  香り 0.5192331  0.73919603  0.4289361
  酸味 0.5454888 -0.67301213  0.4994964
```

```
> summary(pca)
```

Propotion of Variance が各成分の分散の割合，Cumulative Proportion ではその列までの累積寄与率が書かれている

```
  Importance of components:
                            PC1    PC2    PC3
  Standard deviation     1.3183 0.9087 0.6607
  Proportion of Variance 0.5793 0.2752 0.1455
  Cumulative Proportion  0.5793 0.8545 1.0000
```

図を出力する．

```
1 > biplot(pca)
```

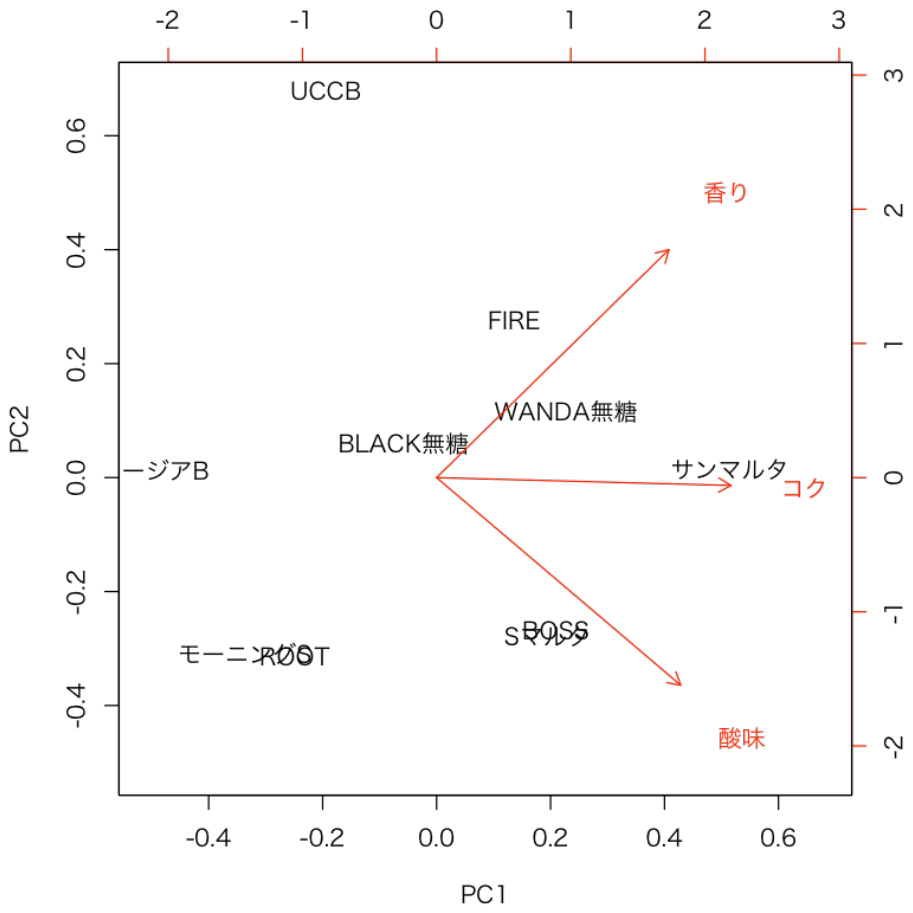

図 2　主成分分析の結果

図 2 は主成分分析の結果をプロットしたものである．

赤い矢印が要素の傾向を表している．これらの矢印の方向に近ければ近いほどその傾向が強い．この図から PC1(第 1 主成分) がコーヒーの総合的な美味しさを表していて，PC2(第 2 主成分) に関しては，香りと酸味に対応していると予想できた．

しかし，総合的に勘案した場合にそのように考えられるだけであって，どう解釈するかは人間の手にかかっているので注意が必要である．

今回は統計学方面に関する主成分分析が主であったが，機械学習でピクセルの画像データを扱う際，次元を削減することで計算量の増加を防いだり，未知のデータを予測する性能を向上させることができる．

参考文献

[1] 主成分分析とは
http://nbviewer.jupyter.org/github/contaconta/PCA_lecture/blob/master/PCA.ipynb

[2] 主成分分析を用いた次元削減、主成分ベクトルを用いた予測と線形回帰による予測の比較
http://dev.classmethod.jp/statistics/pythonscikit-learn-pca1/