11. Statistics II　統計II

- 共分散
- 相関係数
- 共分散行列/相関係数行列
- 共分散行列の固有値/固有ベクトル

csvファイルからのデータの読み込み(MATLAB Grader)

```
>> data=csvread('cars.csv',2,2);
```

- このファイルには406の車について７種類のデータ（燃費, 馬力　等）が含まれている

Output

```
data =

   1.0e+03 *

    0.0180    0.0080    0.3070    0.1300    3.5040    0.0120    0.0700
    0.0150    0.0080    0.3500    0.1650    3.6930    0.0115    0.0700
    0.0180    0.0080    0.3180    0.1500    3.4360    0.0110    0.0700
    0.0160    0.0080    0.3040    0.1500    3.4330    0.0120    0.0700
    0.0170    0.0080    0.3020    0.1400    3.4490    0.0105    0.0700
    0.0150    0.0080    0.4290    0.1980    4.3410    0.0100    0.0700
```

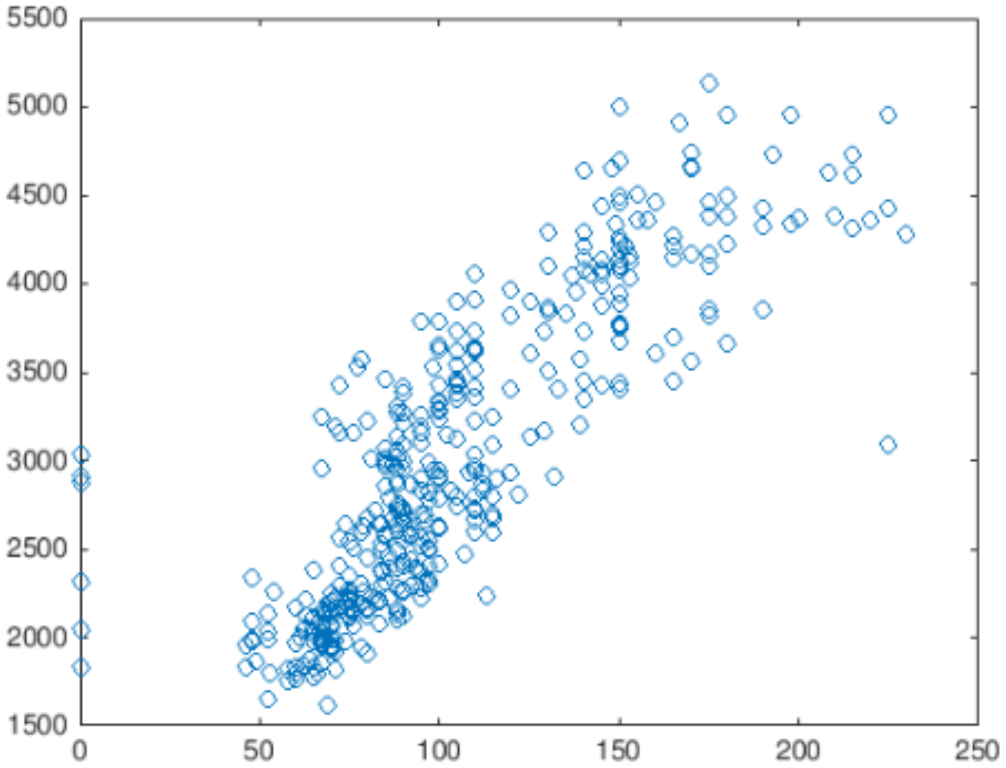
	A	B	C	D	E	F	G	H	I
1	Car	Origin	MPG	Cylinders	Displacem	Horsepow	Weight	Accelerat	Model
2	STRING	CAT	DOUBLE	INT	DOUBLE	DOUBLE	DOUBLE	DOUBLE	INT
3	Chevrolet	US	18	8	307	130	3504	12	70
4	Buick Sky	US	15	8	350	165	3693	11.5	70
5	Plymouth	US	18	8	318	150	3436	11	70
6	AMC Rebe	US	16	8	304	150	3433	12	70
7	Ford Torir	US	17	8	302	140	3449	10.5	70
8	Ford Gala	US	15	8	429	198	4341	10	70

共分散/相関係数

- 共分散：２つの変数の間に線形関係があるかどうかの指標
- 相関係数：２つの変数の間の線形関係の程度（強さ）を表す

例）406種類の車の馬力(horsepower)と重量(wight)を散布図にして表してみる

```
>> plot(data(:,4),data(:,5),'o')
```



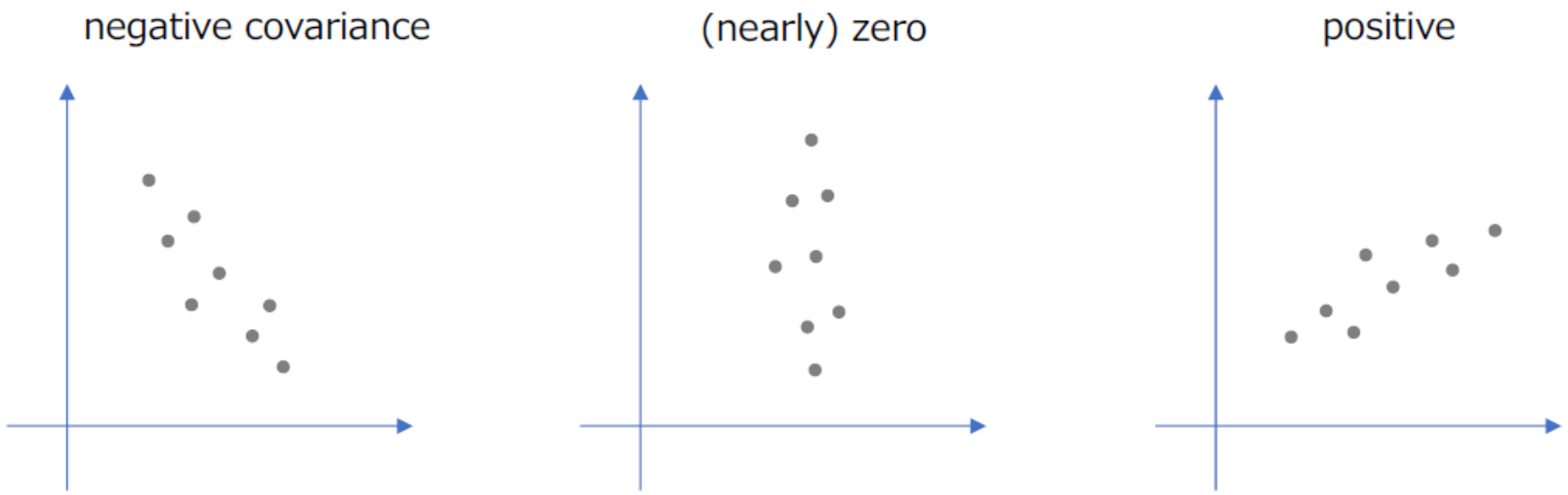
線形関係がある

共分散 covariance

- 共分散の定義

$$\mathrm{cov}(X, Y) = E[(X - E(X))(Y - E(Y)]$$

or $$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$



- ２つの変数が独立の場合は，共分散は単に分散となる

$$\mathrm{cov}(X, X) = E[(X - E(X))^2] = \mathrm{var}(X) = \sigma^2(X)$$

相関係数　correlation coefficient

- 定義　ピアソンの相関係数
  - 規格化された共分散と思えばいい

$$r(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

標準偏差

$$\sigma(X) = \sqrt{\mathrm{var}(X)} \quad \sigma(Y) = \sqrt{\mathrm{var}(Y)}$$

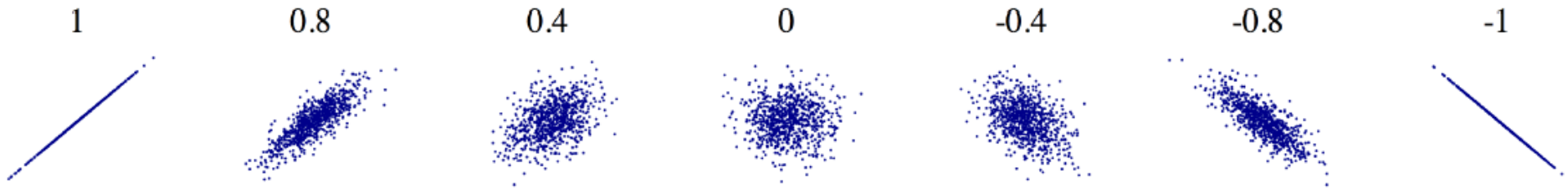
- 相関係数の計算にはcorrcoef関数を用いる

```
corrcoef(X,Y)
```

```
corrcoef(data(:,4),data(:,5))
```

```
ans =        X         Y
    X    1.0000    0.8408
    Y    0.8408    1.0000
```

- 相関係数は-1から１の範囲の値をとる
- corrcoef>0 正の相関 corrcoef<0 負の相関；　0 は相関なし



相関係数についての注意点

- 相関係数は因果関係を表すわけではない
  - 因果関係がなくても相関関係がある場合がある
  - チョコレート消費量とノーベル賞には相関関係があるが、因果関係があるとは限らない（チョコレート食べるとノーベル賞を取れるか？）
- dependenceは相関correlationと混同されるときがあるが,
  確率で定義され

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \Leftrightarrow \mathrm{P}(B) = \mathrm{P}(B \mid A)$$

- 相関係数は線形関係しか見ることができない,

下のような分布の場合はすべて相関係数は 0 になる.

共分散行列/相関行列

- ‘car.csv’には７つの変数がある
- ７つの変数について，そのうちの２つの変数の分散・相関係数を計算し，７×７の行列をつくることができる．これは共分散行列・相関行列と呼ばれる
- N×７行列のデータがあるとして

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$$

- そのデータの共分散行列は次のように定義される

$$\mathrm{cov}(\mathbf{X}) = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top = \frac{1}{N-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$$

ここでmはxの平均ベクトルであり　$\tilde{\mathbf{X}} = [\mathbf{x}_1 - \mathbf{m}, \ldots, \mathbf{x}_N - \mathbf{m}]^\top$

- 相関行列も同じよう方法で定義

共分散行列/相関行列

行列Xに対してcovやcorrcoef関数を用いると共分散行列や相関行列を計算することができる

```
ans =

   1.0e+05 *

    0.0007   -0.0001   -0.0067   -0.0025   -0.0560    0.0001    0.0002
   -0.0001    0.0000    0.0017    0.0006    0.0130   -0.0000   -0.0000
   -0.0067    0.0017    0.1101    0.0371    0.8287   -0.0016   -0.0015
   -0.0025    0.0006    0.0371    0.0164    0.2886   -0.0008   -0.0006
   -0.0560    0.0130    0.8287    0.2886    7.1742   -0.0102   -0.0100
    0.0001   -0.0000   -0.0016   -0.0008   -0.0102    0.0001    0.0000
    0.0002   -0.0000   -0.0015   -0.0006   -0.0100    0.0000    0.0001
```

計算結果は対称行列

```
ans =

    1.0000   -0.7356   -0.7643   -0.7267   -0.7875    0.4245    0.5862
   -0.7356    1.0000    0.9518    0.8235    0.8952   -0.5225   -0.3608
   -0.7643    0.9518    1.0000    0.8738    0.9325   -0.5580   -0.3817
   -0.7267    0.8235    0.8738    1.0000    0.8408   -0.6820   -0.4199
   -0.7875    0.8952    0.9325    0.8408    1.0000   -0.4301   -0.3154
    0.4245   -0.5225   -0.5580   -0.6820   -0.4301    1.0000    0.3020
    0.5862   -0.3608   -0.3817   -0.4199   -0.3154    0.3020    1.0000
```

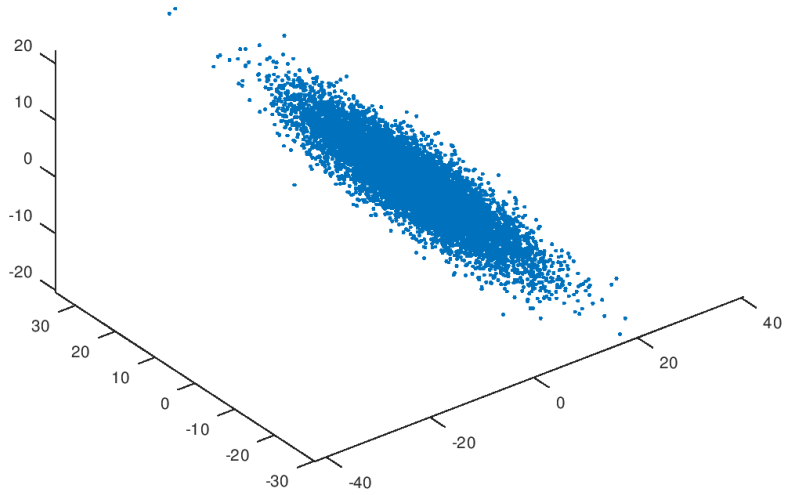
馬力と重量の相関係数

共分散行列の固有値/固有ベクトル(Octave)

- 共分散行列は空間内にデータがどのように分布するかを表す
- 共分散行列の固有ベクトルはどの方向にデータが広がっているかを表す
- 固有値はそのデータの広がりの幅を表す

例）授業のホームページよりe11files.zipをダウンロード
解凍して3d_ptdataを作業ディレクトリに置く

```
>> load('3d_ptdata.m')
>> size(X)
ans =

   10000       3

>> plot3(X(:,1),X(:,2),X(:,3),'.'); axis equal
```

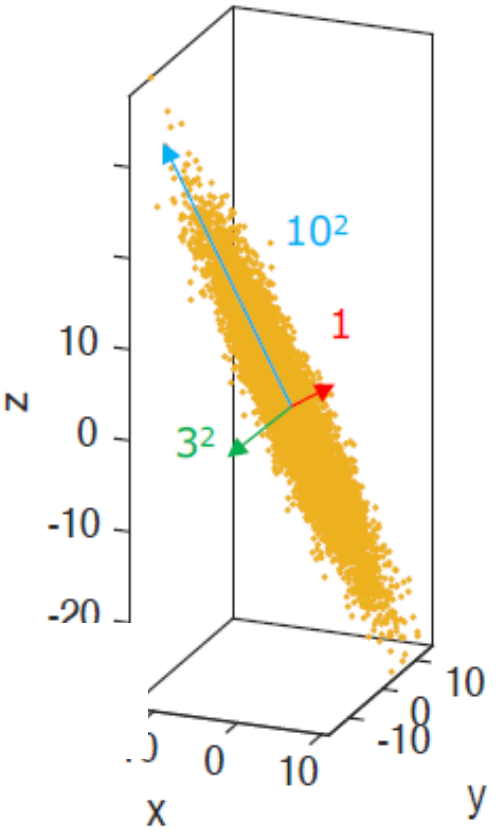


```
>> [V,W]=eig(cov(X))
V =

   0.867213  -0.208827  -0.452032
   0.496518   0.431157   0.753375
   0.037571  -0.877778   0.477591

W =

Diagonal Matrix

   0.98420         0         0
         0    9.05693         0
         0         0  103.15470
```

Xの共分散行列の固有ベクトルV, 固有値Wを求める

3つの直交軸
どの方向に広がりがあるか

広がりの幅（ばらつき）

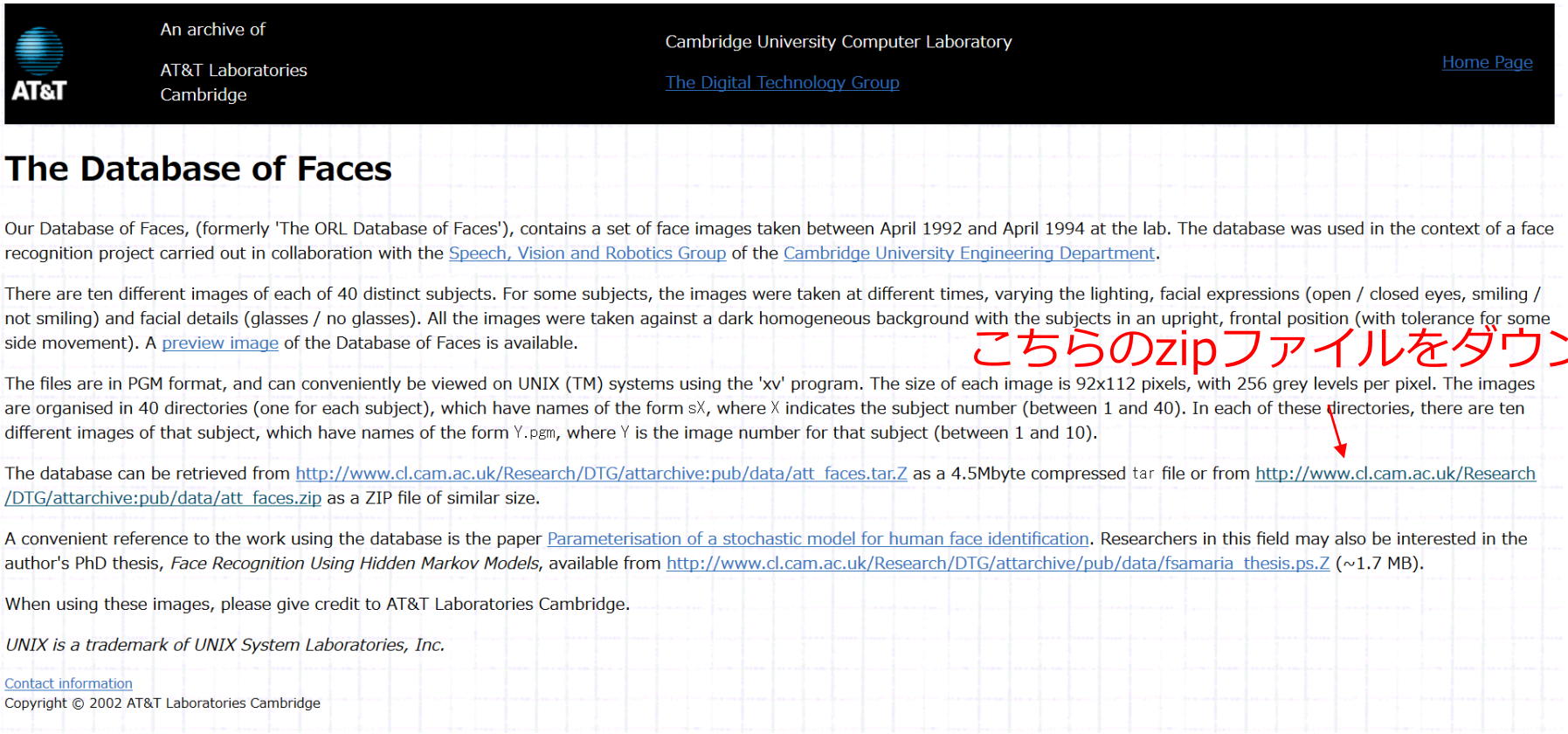


Exercise 11（準備）(Octaveの場合のみMATLAB Graderでは必要なし)

- 前のページの手法（主成分分析）はどのようなタイプのデータにも使える
- 今回の課題では顔写真データに対して主成分分析を行う.

準備

http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html からデータをダウンロード



AT&T

An archive of

AT&T Laboratories
Cambridge

Cambridge University Computer Laboratory

The Digital Technology Group

Home Page

The Database of Faces

Our Database of Faces, (formerly 'The ORL Database of Faces'), contains a set of face images taken between April 1992 and April 1994 at the lab. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.

There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). A preview image of the Database of Faces is available.

こちらのzipファイルをダウンロード

The files are in PGM format, and can conveniently be viewed on UNIX (TM) systems using the 'xv' program. The size of each image is 92x112 pixels, with 256 grey levels per pixel. The images are organised in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the form Y.pgm, where Y is the image number for that subject (between 1 and 10).

The database can be retrieved from http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.tar.Z as a 4.5Mbyte compressed tar file or from http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.zip as a ZIP file of similar size.

A convenient reference to the work using the database is the paper Parameterisation of a stochastic model for human face identification. Researchers in this field may also be interested in the author's PhD thesis, *Face Recognition Using Hidden Markov Models*, available from http://www.cl.cam.ac.uk/Research/DTG/attarchive/pub/data/fsamaria_thesis.ps.Z (~1.7 MB).

When using these images, please give credit to AT&T Laboratories Cambridge.

UNIX is a trademark of UNIX System Laboratories, Inc.

Contact information

- 作業ディレクトリにてzipファイルを解凍すると，att_facesディレクトリができる

別の場所に解凍した後でatt_facesディレクトリを作業ディレクトリに移動させてもよい

- 授業ホームページからダウンロードしたe11files.zip内にあるload_faces.mファイルを作業ディレクトリに移動して実行

```
>> load_faces
```

- 400人の顔写真のデータ（92×112ピクセル）が変数Xに格納され、Xは400×10304(=92×112)の行列になる
- 100番目の顔写真を表示する場合にはimshow関数を用いて

```
>> imshow(reshape(X(100,:),[112,92])/255);
```

100番目の顔写真のデータは行列Xの100列目に入っている

reshape関数で100列目を112×92行列に変形

/255はシステムによっては必要なし

- 共分散行列の最初の20個の固有値を求める.

  - ただしcov(X)で計算は（できないことはないが）しない．行列Xは400×10304の次元を持つので，その共分散行列は10304×10304行列になる．計算量が膨大になるのでこれはやらない.

  - svds関数を用いる.
    svds(A,n) Aの特異値とベクトルを大きい方からn個求める

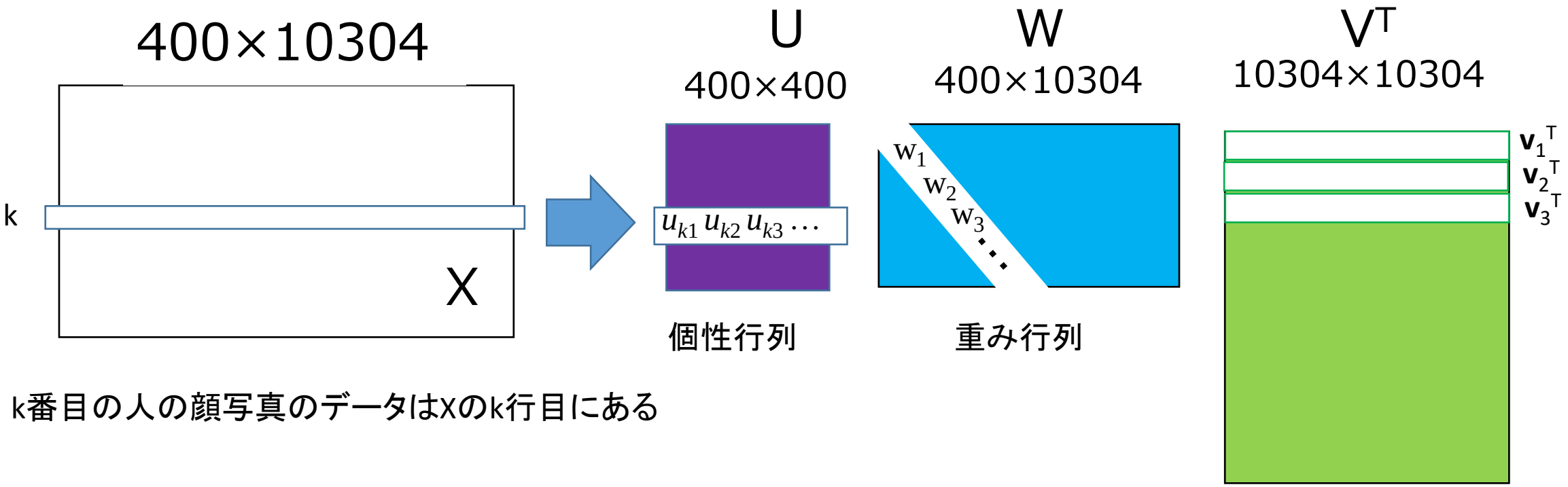
```
>> [U,W,V]=svds(X-mean(X),20);
```

  - 各人の顔写真から平均を引いたことで、その特徴が表されている

  - 計算結果の特異値，すなわち固有値の平方根（前回の講義資料を参照）を見ると，最初の数個は大きいがそれ以外は小さくなっている

  - データは低次元の空間にのみ存在する

特異値分解の解釈



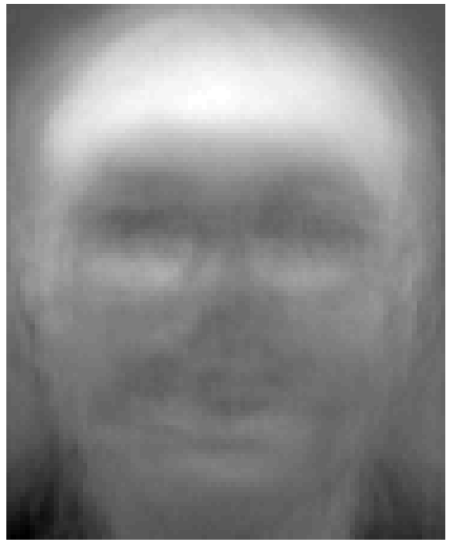
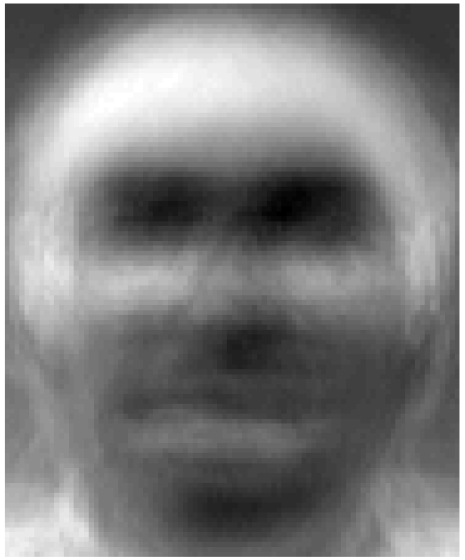
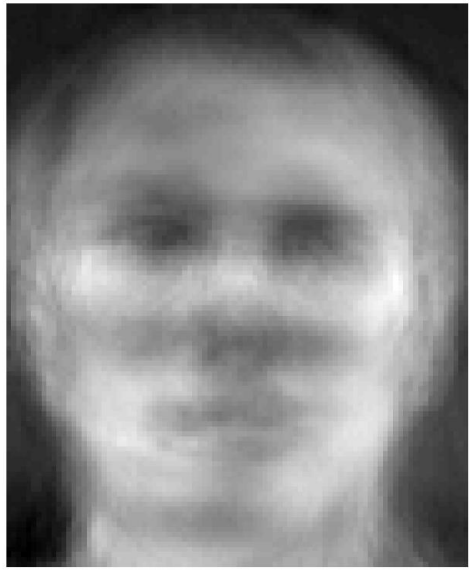
k番目の人の顔写真のデータはXのk行目にある

k番目の顔写真を分解したUWVから復元したい場合は

Uのk行目に，WとV^{T}を掛けていけばよい.

$$u_{k1}w_1\mathbf{v}_1{}^T + u_{k2}w_2\mathbf{v}_2{}^T + u_{k3}w_3\mathbf{v}_3{}^T + \ldots$$

固有顔viに個性(u)を重み(w)を付けて重ね合わせていくことで顔写真ができる


$$u_{k1}\mathrm{w}_1\mathbf{v}_1{}^{\mathrm{T}} + u_{k2}\mathrm{w}_2\mathbf{v}_2{}^{\mathrm{T}} + u_{k3}\mathrm{w}_3\mathbf{v}_3{}^{\mathrm{T}} + \ldots + \mathrm{mean}$$

元の顔が復元される

Exercise 11　(MATLAB Graderで解答提出してください)

1. 前述の特異値分解に基づき，20個の固有値を求めよ
2. V（固有顔）の上位4個のパターンを表示せよ
3. N番目の顔写真を特異値分解から得られた20個の固有顔を用いて復元し，元の画像と比較せよ