

11. 統計学II



東北大学 大学院工学研究科

嶋田 慶太

keita.shimada.c6@tohoku.ac.jp



TOHOKU
UNIVERSITY

- 復習
- 共分散と相関係数
- 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
- 課題

今週も

MATLAB Grader

を使います。



TOHOKU
UNIVERSITY

この授業での列ベクトルと行列の表現など(復習)

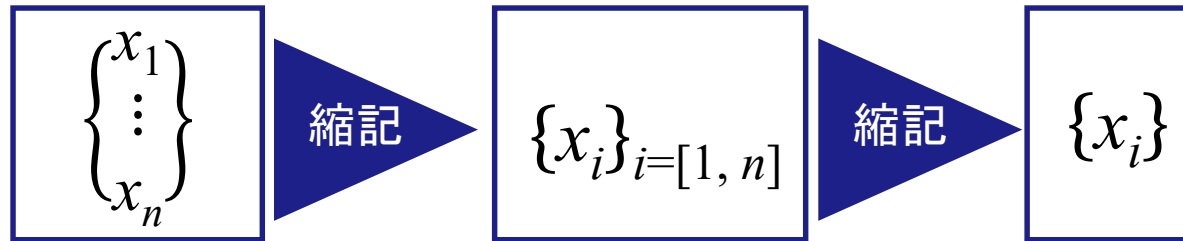


この授業での行列・列ベクトルの表現

列ベクトル

▶ 波括弧 $\{\}$ で表現

*世界的にこれがスタンダードというわけではないので
もし別の機会で使おうと思ったら、ルールを書いてください。



例えば

$\{1\}$ $\{0\}$

は全要素が
1, 0の列ベクトル

* \mathbf{x} のように太字もよく使う。

* 行ベクトルは $\{x_i\}^T$ のように列ベクトルの転置で表す

行列

▶ 大文字もしくは角括弧 $[\]$ で表現

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

例えば

$$[\{x_i\}\{y_i\}]$$

は i 行の成分が x_i, y_i の $n \times 2$ 行列



Octaveの統計関数(復習)

▶ 平均: mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{\{1\}^T \{x_i\}}{n}$$

▶ 分散: var 不偏分散と呼ばれる

$$V = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 = \frac{\{x_i - \mu\}^T \{x_i - \mu\}}{n-1}$$

▶ 標準偏差: std

$$\sigma = \sqrt{V}$$



csv ファイルの読み込み(Octave用)

▶ CSVファイル読み込み: csvread

*CSV: Comma Separated Value
カンマ区切り

授業のページからcars.csvをダウンロードし, 変数に読み込む

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

これは単なる変数名(AでもXでもよい)

406車種に関する数値データ
※csvreadは数値データしか読めず,
文字列は0と置かれる

```
>> data =
0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000
0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000    0.00000
0.00000    18.00000    8.00000    307.00000    130.00000    3504.00000    12.00000    70.00000    0.00000
0.00000    15.00000    8.00000    350.00000    165.00000    3693.00000    11.50000    70.00000    0.00000
0.00000    18.00000    8.00000    318.00000    150.00000    3436.00000    11.00000    70.00000    0.00000
```

Excel等で開いた場合

Car	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model	Origin
STRING	DOUBLE	INT	DOUBLE	DOUBLE	DOUBLE	DOUBLE	INT	CAT
Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	US
Buick Skylark 320	15	8	350	165	3693	11.5	70	US
Plymouth Satellite	18	8	318	150	3436	11	70	US

* The file copied from <https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>

(pounds)

(seconds for
0-60 mph (0-97 km/h))

今回の授業での csv の読み込み



7

▶ CSVファイル読み込み: csvread

The screenshot shows the MATLAB Grader web interface. The browser address bar displays the URL: <https://grader.mathworks.com/courses/7065-computer-aided-problem-solving-sample/problems/114951-solutions/new>. The page title is "MATLAB Grader" and the user name is "Shimada Keita".

On the left sidebar, the "CONTENTS" menu is open, showing a list of topics. The "統計学 II" (Statistics II) folder is expanded, and the "作業フィールド" (Assignment Field) item is highlighted with a red box.

The main content area displays the assignment "作業フィールド". It includes instructions in Japanese: "今回の講義はここで作業しよう！ (←こんなふうを書く今回がいつを指すかわからないから、みんなはこんな表現使わないでね)" and "csvread"を使って、csvファイルの読み込みをやってみよう！. It also mentions that downloading "cars.csv" is not necessary.

Below the instructions, there is a section titled "Your Script" with a text area containing the MATLAB code:

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

At the bottom right of the script area, there are links for "Reset" and "MATLAB Documentation".



TOHOKU
UNIVERSITY

共分散と相関係数



共分散と相関係数



共分散 (Covariance)

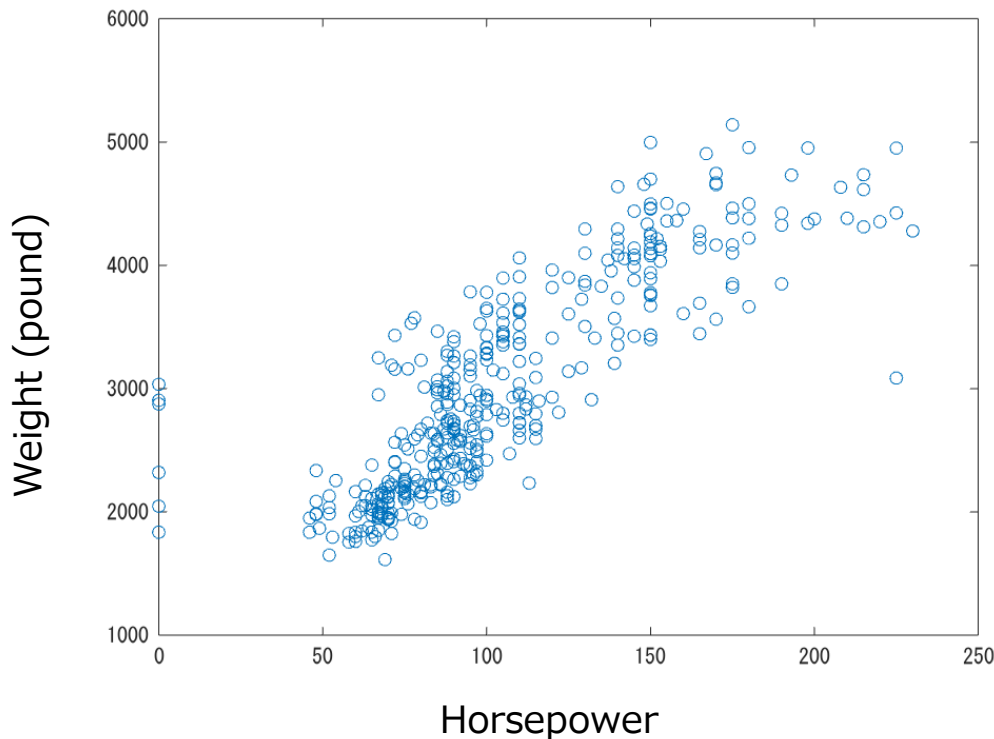


相関 (Correlation)

2つの変数の中の線形関係の程度を表す似た概念

Horsepower vs Weightの散布図 (scatter plot)

```
>> plot(data(3:408, 5), data(3:408, 6), 'o')
```



線形的な相関性が見える。

▶ 線形近似したい場合は第5回参照

どのくらい

線形相関性があるかを示す指標

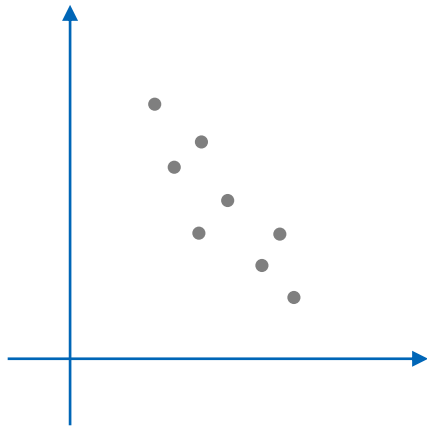


共分散 (Covariance of two variables)

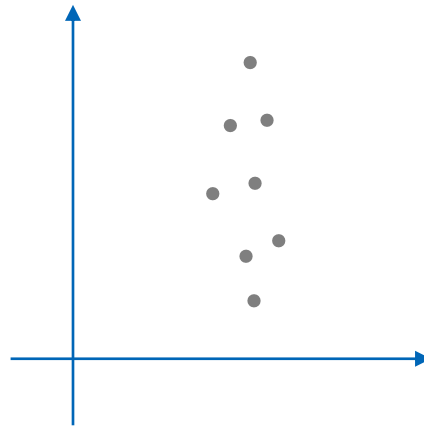
▶ 定義 $\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$

$$\text{cov}(\{x_i\}, \{y_i\}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\{y_i - \bar{y}\}^T \{x_i - \bar{x}\}}{n-1}$$

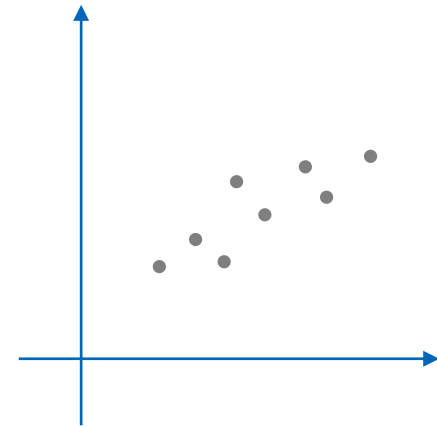
negative covariance



(nearly) zero



positive



▶ cov 共分散を求める組込み関数

▶ 分散との関係: 自身との共分散が分散

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



相関係数(Correlation coefficient)

▶ 定義 ピアソンの相関係数

共分散を標準偏差で規格化

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} \left(\begin{array}{l} \text{標準偏差} \\ \sigma(X) = \sqrt{\text{var}(X)} \\ \sigma(Y) = \sqrt{\text{var}(Y)} \end{array} \right)$$

$$r(\{x_i\}, \{y_i\}) = \frac{\{y_i - \bar{y}\}^T \{x_i - \bar{x}\}}{\|y_i - \bar{y}\| \|x_i - \bar{x}\|}$$

◀ 分子・分母の $(n-1)$ がキャンセルされ
ベクトルの方向余弦

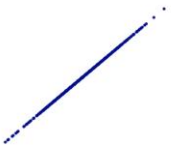
▶ corr 相関係数を求める組み込み関数

```
>> corr(data(3:408, 5), data(3:408, 6))
ans = 0.84081
```

相関係数は $[-1, 1]$ の値を取る

正の相関

1



0.8

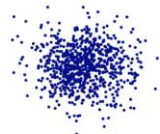


0.4



相関なし

0



-0.4

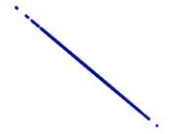


-0.8



負の相関

-1



相関に関する注意

▶ 相関関係は因果関係にあらず

例: ノーベル賞とチョコレートの消費量には相関はあっても因果があるとは言えない.

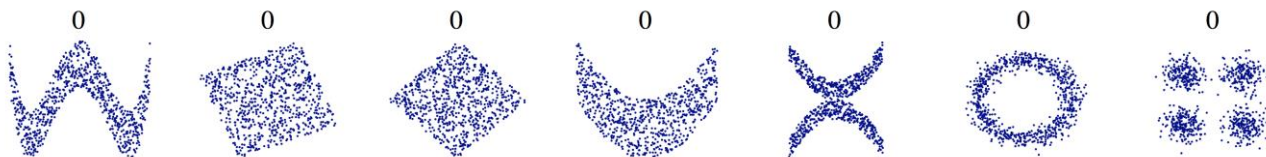
- ▶ チョコレートは嗜好品なので, チョコレートを食べられるほど「豊か」である「かもしれない」
- ▶ 主食をチョコレートに切り替えてもノーベル賞は多分取れない.

▶ 依存関係は相関関係も混同されがち

$$P(A \cap B) = P(A)P(B) \Leftrightarrow P(B) = P(B | A)$$

▶ 相関係数は線形的関係しか扱えない

これらの相関係数は0である.





TOHOKU
UNIVERSITY

共分散行列と相関係数行列

内積で共分散が求まるのだから、
行列を転置して掛合せば一発で全部求まる



共分散行列・相関行列

“car.csv” には 7 変数 406 種類のデータがある

$$\mathbf{X} = \left[\begin{array}{c} \left\{ \begin{array}{c} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{406,1} \end{array} \right\} \left\{ \begin{array}{c} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{406,2} \end{array} \right\} \cdots \left\{ \begin{array}{c} x_{1,7} \\ x_{2,7} \\ \vdots \\ x_{406,7} \end{array} \right\} \end{array} \right]$$

$$\widetilde{x}_{i,j} = x_{i,j} - \bar{x}_j$$

列ベクトル平均

\bar{x}_1

\bar{x}_2

\bar{x}_7

共分散および相関に関して、自身とを含め、 7×7 の行列を作れる。

$$\widetilde{\mathbf{X}} = \left[\begin{array}{c} \left\{ \begin{array}{c} x_{1,1} - \bar{x}_1 \\ x_{2,1} - \bar{x}_1 \\ \vdots \\ x_{406,1} - \bar{x}_1 \end{array} \right\} \left\{ \begin{array}{c} x_{1,2} - \bar{x}_2 \\ x_{2,2} - \bar{x}_2 \\ \vdots \\ x_{406,2} - \bar{x}_2 \end{array} \right\} \cdots \left\{ \begin{array}{c} x_{1,7} - \bar{x}_7 \\ x_{2,7} - \bar{x}_7 \\ \vdots \\ x_{406,7} - \bar{x}_7 \end{array} \right\} \end{array} \right] \quad \blacktriangleright \quad \text{cov}(\mathbf{X}) = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}$$



共分散行列・相関行列



COV と CORR

引数に行列を与えると,
各列ベクトルを変数とする共分散行列, 相関行列を計算する

```
>> X=data(3:408,2:8);
>> size(X)
ans =
```

```
406    7
```

```
>> cov(X)
ans =
```

```
7.0590e+01 -1.0581e+01 -6.7374e+02 -2.4739e+02 -5.6042e+03 9.9981e+00 1.8464e+01
-1.0581e+01 2.9315e+00 1.7098e+02 5.7130e+01 1.2983e+03 -2.5077e+00 -2.3155e+00
-6.7374e+02 1.7098e+02 1.1009e+04 3.7148e+03 8.2869e+04 -1.6412e+02 -1.5014e+02
-2.4739e+02 5.7130e+01 3.7148e+03 1.6419e+03 2.8858e+04 -7.7476e+01 -6.3788e+01
-5.6042e+03 1.2983e+03 8.2869e+04 2.8858e+04 7.1742e+05 -1.0212e+03 -1.0014e+03
9.9981e+00 -2.5077e+00 -1.6412e+02 -7.7476e+01 -1.0212e+03 7.8588e+00 3.1737e+00
1.8464e+01 -2.3155e+00 -1.5014e+02 -6.3788e+01 -1.0014e+03 3.1737e+00 1.4053e+01
```

```
>> corr(X)
ans =
```

```
1.00000 -0.73556 -0.76428 -0.72667 -0.78751 0.42449 0.58623
-0.73556 1.00000 0.95179 0.82347 0.89522 -0.52245 -0.36076
-0.76428 0.95179 1.00000 0.87376 0.93247 -0.55798 -0.38171
-0.72667 0.82347 0.87376 1.00000 0.84081 -0.68205 -0.41993
-0.78751 0.89522 0.93247 0.84081 1.00000 -0.43009 -0.31539
0.42449 -0.52245 -0.55798 -0.68205 -0.43009 1.00000 0.30199
0.58623 -0.36076 -0.38171 -0.41993 -0.31539 0.30199 1.00000
```

先に計算した
horsepower-weight correlation

MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model
-----	-----------	--------------	------------	--------	--------------	-------



TOHOKU
UNIVERSITY

共分散行列の固有値・固有ベクトル



分散共分散行列の固有値・固有ベクトル

データセットの準備

- ▶ e11files.zipをダウンロード
- ▶ 解凍して作業ディレクトリに置く

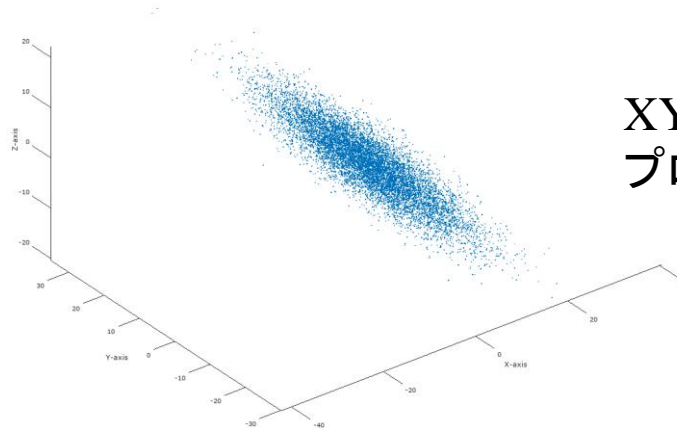
```
>> load('3d_ptdata.m')
>> size(X)
ans =
    10000         3

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

3d_ptdata.m の中身

```
1  # Created by Octave 4.2.1, Wed May 10 13:15:07 2017 JST <okatani@mbi2.local>
2  # name: X
3  # type: matrix
4  # rows: 10000
5  # columns: 3
6  4.012379437569767 -7.173648310016696 -8.447865860936579
7  -7.302062665623171 10.96863499616599 4.542533876048474
8  -0.7057253050448848 -1.668216964515125 3.441950474520851
9  -7.251964059975861 11.39873425020424 6.310863373678669
10 -1.125779124999649 -2.13018887890818 -1.496738894026304
11 -0.2817785572006778 3.238812971886579 -1.682407286765399
```

2行目: name: X ▶ 変数名がX
3行目: type: matrix ▶ 行列形式
4行目: rows: 10000 ▶ 10000行
5行目: columns: 3 ▶ 3列
のデータであることが分かる。



XYZ空間内で傾いた方向に
プロットが集中



分散共分散行列の固有値・固有ベクトル

分散共分散行列 $\text{cov}(X)$ を固有分解する

```
>> [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

```
>> Y=(X-mean(X))*V;  
>> plot3(Y(:,1),Y(:,2),Y(:,3),'.');  
>> axis equal;
```

- ▶ $\text{cov}(X)$: 対称行列 ($A = A^T$)
- ▶ 直交行列 ($A^{-1} = A^T$) で対角化可能



$$W = V^T \tilde{X}^T \tilde{X} V$$

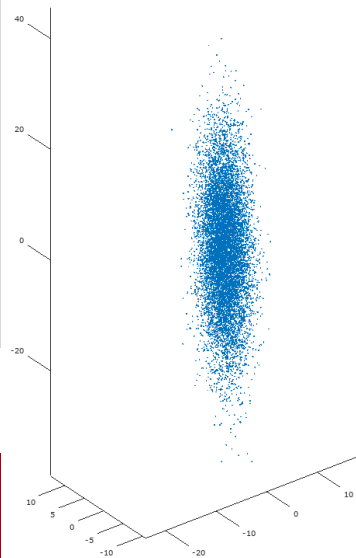
新しい座標系

$$Y = \tilde{X} V$$

と座標変換すると,

固有値

- ▶ ばらつく幅
- ▶ 大きいほど特徴が出る
(この例では新しいZ軸方向)



主成分分析(PCA)という



TOHOKU
UNIVERSITY

1
9

課題



主成分分析 (Exercises 11.1の準備)

例 100番目の画像の表示:

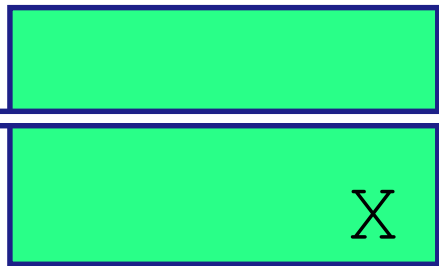
```
>> imshow(reshape(X(100,:), [112, 92]) / 255)
```

imshowで扱えるよう、
データを規格化

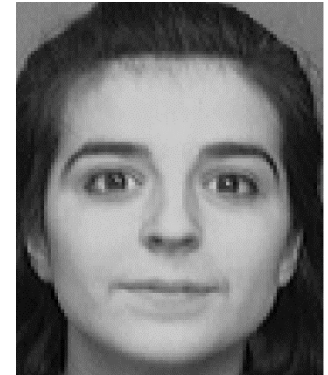
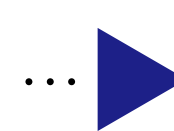
imshow: 画像表示の関数
(0~1の値で扱う)

reshape: 行列を変形させる関数
この場合Xの100列目を112×92行列に変形

各列に0~255の値のデータが格納されている.



1~112が1列目
113~224が2列目...
(転置して格納されている)



112行×92列にし
imshowで表示

スクリプト例: 1-10の写真を並べる

```
load_faces
for i = 1:10
    subplot(2, 5, i); imshow(reshape(X(i,:), [112, 92]) / 255)
endfor
```

▼i番目の写真の場所

主成分分析（Exercises 11.1の準備）

共分散行列 X の最初の20個の固有値を求めよう。



各顔データ点 ▶ 各画像 ($92 \times 112 = 10304$) 次元

顔データ個数: 400個 ▶ $\text{cov}(X)$ の計算 $400^2 \times 10304^2$ 回の掛け算を要する。

▲
一要素のための計算 ▲
10304 × 10304 行列

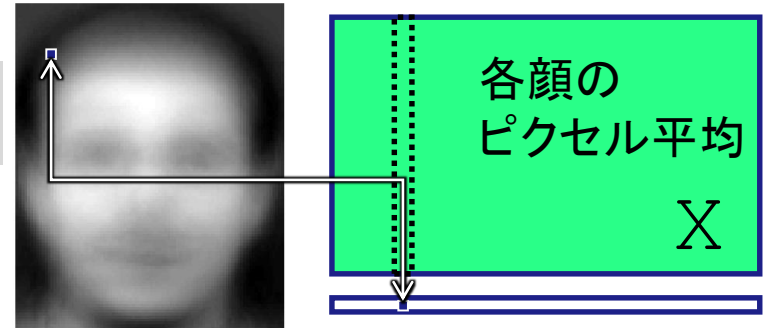
そのまま $\gg \text{cov}(X)$ してはいけない！ ...こともないのだが、今回はやめよう。

▶ `svds` 指定した個数の特異値とベクトルを大きい方から求める

```
 $\gg [U, W, V] = \text{svds}(X - \text{mean}(X), 20);$ 
```

先述の \tilde{X} 行列

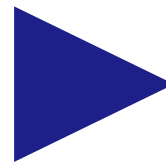
平均顔(?)



各人から平均を引いたもの ▶ 個性・特徴が残る(?)

最初の数個の特異値 (すなわち、固有値の平方根): 大

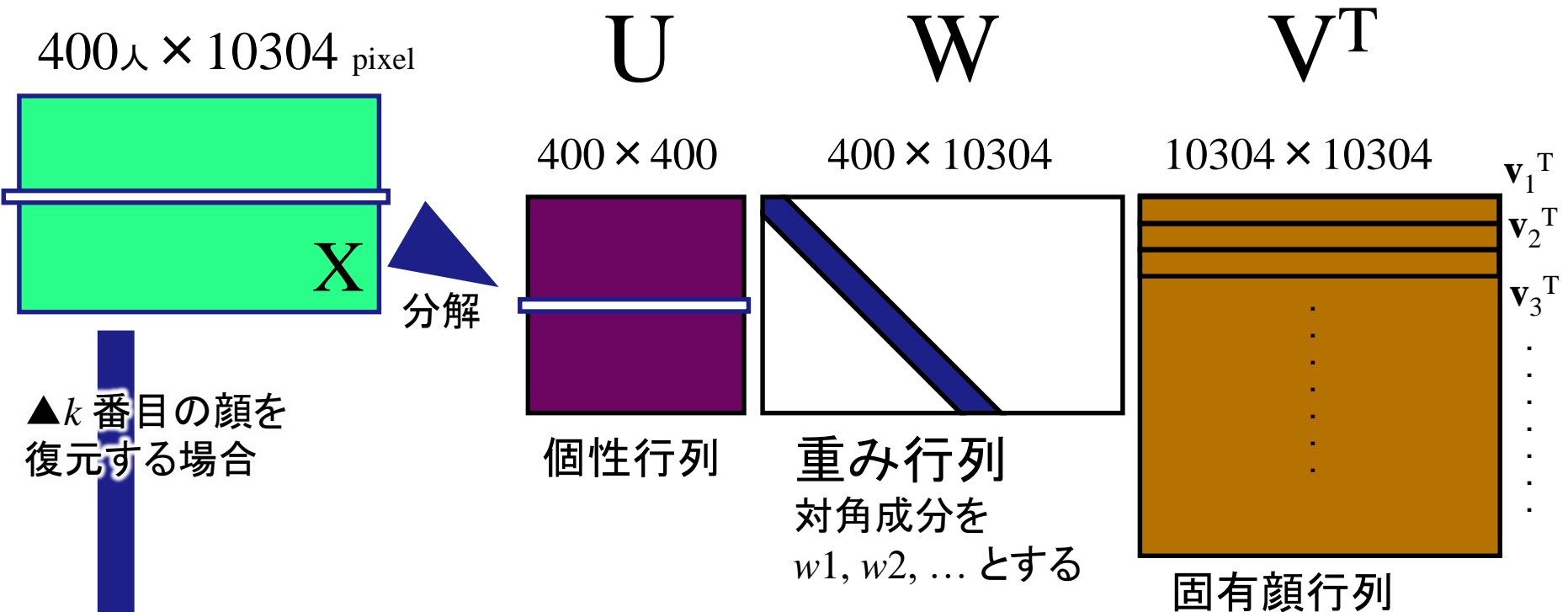
他の特異値: 小



データは
低次元の部分空間
に存在



今回の特異値分解の解釈



U の k 行目を取り出し, W と V^T を順次掛ける

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

u : 個人の「個性」に依存する重み
 w : ベクトルの重み(重要度ともいえる)
 v : 固有顔

固有顔 \mathbf{v}_i に個性の重みを付け,
どんどん重ね合わせる
「版画」のイメージ

今回の特異値分解の解釈



$$\text{mean} + u_k w_1 \mathbf{v}_1^T + u_{k2} w_2 \mathbf{v}_2^T + u_{k3} w_3 \mathbf{v}_3^T \dots + u_{k400} w_{400} \mathbf{v}_{400}^T$$

元の顔に戻る

400個の顔版画(V)に分解しており,
個性(U)と版の重要性(W)に応じ重ねると
元に戻せる, というようなイメージ



Exercise11

1. 前述の特異値分解に基づき, 20個の固有値をプロットせよ.
2. V (固有顔)の上位4個のパターンを表示せよ.
3. 31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, 元の画像と並べて表示せよ.

白黒の濃度の調整について

絵に含まれる最大値と最小値に合わせて白黒を割り振る場合

```
face=V(:,1); ◀ 扱いやすいように適当な変数に入れ替え  
imshow(reshape(face, [112, 92]), [min(face), max(face)])
```

imshow の引数として白黒の上限, 下限値を入れられる.