

# 情報管理

## 第5回：多次元データの可視化 ～主成分分析と線形判別分析～

# 今回の講義内容

今回は複数の情報からなるデータを可視化する方法を紹介します。

多次元データについて

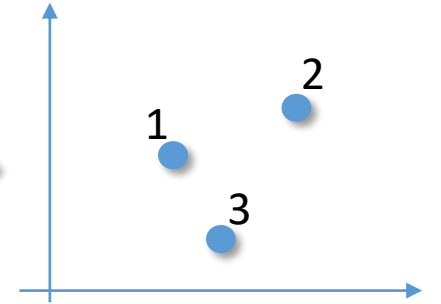
- 平均値ベクトルと分散共分散行列
- 標準化

多次元データの可視化

- 主成分分析
- 線形判別分析

	国語	数学	英語	理科	社会
1	80	65	85	58	70
2	85	95	80	97	88
3	50	40	20	43	60
			⋮		

可視化



特に主成分分析はよく用いられるデータ解析手法の一つです。

# 多次元データと統計量

# 多次元データ

例えば5教科テストのように各サンプル（生徒）について複数の情報からなるデータのことを、**多次元データ**，あるいは**多変量データ**と呼びます。

多次元データでは，各サンプルを多次元のベクトルとして扱います。

**次元数 = 5**

	国語	数学	英語	理科	社会
1	80	65	85	58	70
2	85	95	80	97	88
3	50	40	20	43	60
⋮					

**サンプル数**  
あるいは  
**データ数**

それぞれ5次元ベクトルと扱う

# 多次元データの統計量：平均と分散

多次元データの平均と分散に相当するのが、平均値ベクトルと分散共分散行列です。

**平均値ベクトル**：各次元の平均値を並べたベクトル

$$\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_D]^T = \left[ \frac{1}{N} \sum_{n=1}^N x_{n,1}, \frac{1}{N} \sum_{n=1}^N x_{n,2}, \dots, \frac{1}{N} \sum_{n=1}^N x_{n,D} \right]^T$$

	国語	数学	英語	理科	社会
1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$
2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$
3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$

N: サンプル数  
D: 次元数

**分散共分散行列**：以下の式で計算される行列

$$\begin{aligned} cov &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \\ &= \begin{bmatrix} \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,D} - \bar{x}_D) \\ \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,D} - \bar{x}_D) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,D} - \bar{x}_D) \end{bmatrix} \end{aligned}$$

# 分散共分散行列について

分散共分散行列の対角成分は、次元毎の分散です。  
非対角成分は、異なる次元間の相関を表します。

$$\begin{aligned}
 cov &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \\
 &= \begin{bmatrix} \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,1} - \bar{x}_1)(x_{n,D} - \bar{x}_D) \\ \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,2} - \bar{x}_2)(x_{n,D} - \bar{x}_D) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,1} - \bar{x}_1) & \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,2} - \bar{x}_2) & \dots & \frac{1}{N} \sum_{n=1}^N (x_{n,D} - \bar{x}_D)(x_{n,D} - \bar{x}_D) \end{bmatrix}
 \end{aligned}$$

1次元目と2次元目の相関  
D次元目の分散

# なぜ分散の表現に相関が必要なの？

以下の国語と英語の例を見てみましょう。

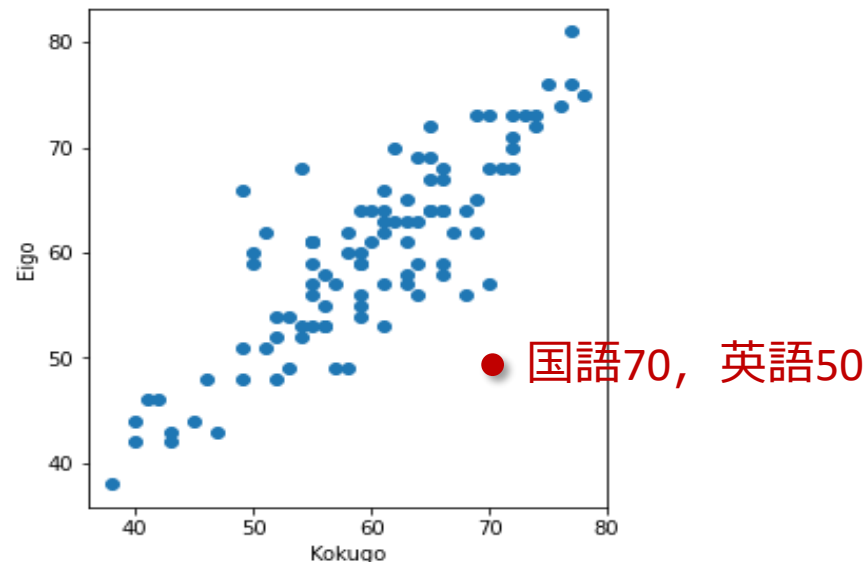
国語も英語も平均約 60，標準偏差（分散の平方根）は約10です。

よって $60 \pm 10 = 50 \sim 70$ 点は，典型的な点数の範囲と言えます。

では，国語が70点，英語が50点の学生は典型的でしょうか？

→ **No**。国語と英語を独立に考えると，平均±標準偏差の範囲内だが，実際の分布からはかけ離れている。

つまり，**分布の形状を表すためには，次元毎の分散だけでなく，相関の情報も必要。**



# 多次元データの標準化



# 標準化

多次元データの中には、次元毎の単位（スケール）が合っていないものがあります。

	身長 [cm]	体重 [kg]
1	172	70
2	167	62
3	170	68

スケールが合っていない次元を同一に扱うと、想定通りの分析ができない場合があります。

そこで、次元毎のスケールを合わせるため、次元毎の平均が0、標準偏差が1になるように正規化します。これを**標準化**と呼びます。

$$x_{norm} = \left[ \frac{x_1 - \bar{x}_1}{\sigma_1} \quad \frac{x_2 - \bar{x}_2}{\sigma_2} \quad \dots \quad \frac{x_D - \bar{x}_D}{\sigma_D} \right] \quad \sigma_D: D\text{次元目の標準偏差}$$

# 多次元データの可視化 1

## 主成分分析

# 目的

次元数が多いデータは、プロットが困難です。

重要な情報を保持したまま 2 次元に圧縮できれば、プロットが可能になります。

	国語	数学	英語	理科	社会
1	80	65	85	58	70
2	85	95	80	97	88
3	50	40	20	43	60



プロットが困難

⋮



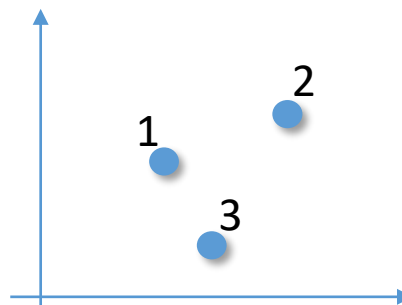
2次元に圧縮

	次元 1	次元 2
1	90	40
2	87	67
3	40	15

⋮




プロットが容易



# 次元圧縮のアプローチ

各次元の値に係数をかけて足し合わせることで、新たな次元を作成します。

	国語	数学	英語	理科	社会		次元 1	次元 2
1	80	65	85	58	70		90	40
2	85	95	80	97	88		87	67
3	50	40	20	43	60		40	15
			⋮				⋮	

$$\text{次元1} = a_1 * \text{国語} + b_1 * \text{数学} + c_1 * \text{英語} + d_1 * \text{理科} + e_1 * \text{社会}$$

$$\text{次元2} = a_2 * \text{国語} + b_2 * \text{数学} + c_2 * \text{英語} + d_2 * \text{理科} + e_2 * \text{社会}$$

では、重要な情報を保持するにはどのように各係数を決めれば良いでしょうか？

# そもそも重要な情報って何？

仮に、下の二つのテスト結果のうち、どちらか一方のテストの点だけを記録に残し、もう一方の点数は破棄するとします。

学生の能力を測ることがテストの目的とした場合、どちらのテストの点数を記録するのが望ましいでしょうか？

	テスト1	テスト2
1	90	90
2	70	91
3	60	90
4	49	90
5	72	91
6	56	91
	⋮	

おそらく多くの方はテスト1を記録するでしょう。

テスト2はどの生徒も似たような点数なので、能力差が測れないからです。

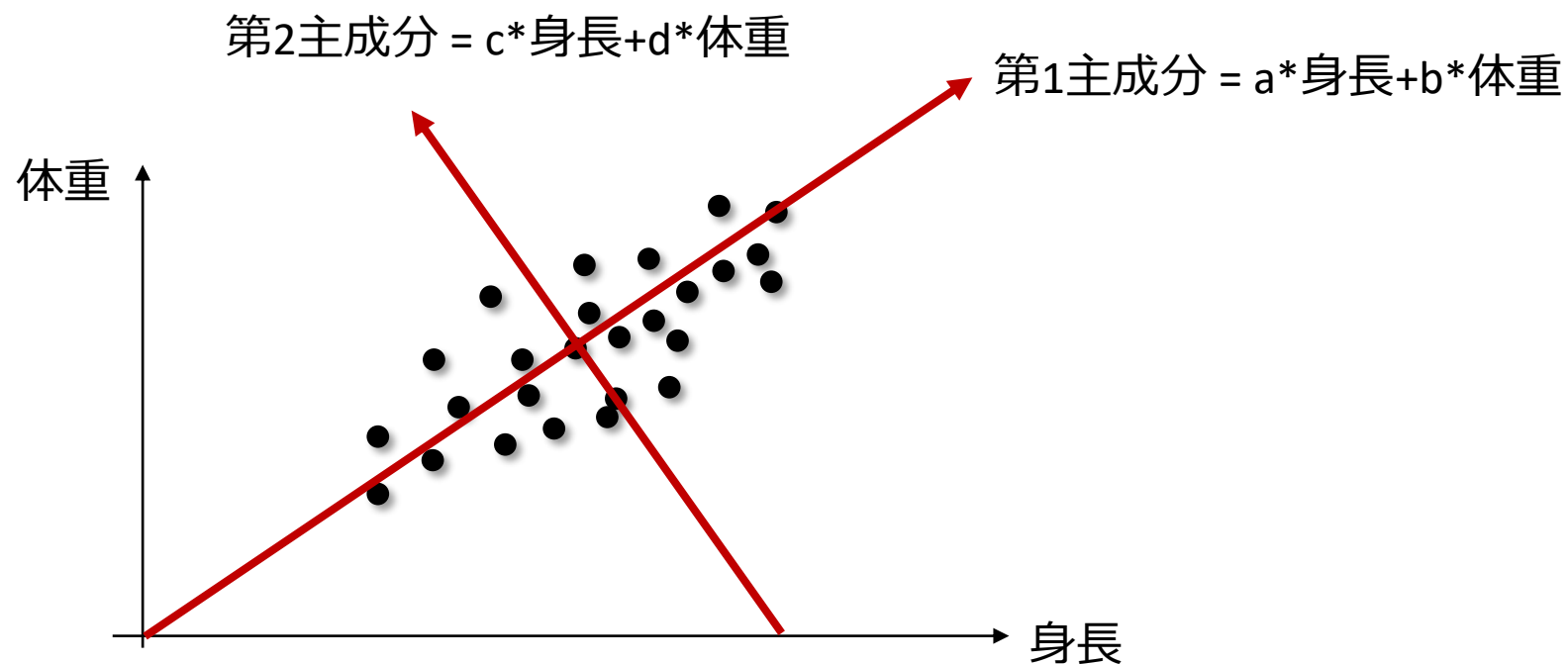
主成分分析では、サンプルによって多種多様な値を持つ情報 = 分散の大きい情報ほど重要であると考え、分散が最大となるように次元圧縮します。

# 2次元データに対する主成分分析の例

主成分分析は、元データの次元数と同じ数だけ新たな次元（軸）を作成します。

仮に元データが2次元だった場合、主成分分析も2個の軸を作成します。

このとき、主成分分析が作成したn番目の軸を**第n主成分**と呼びます。

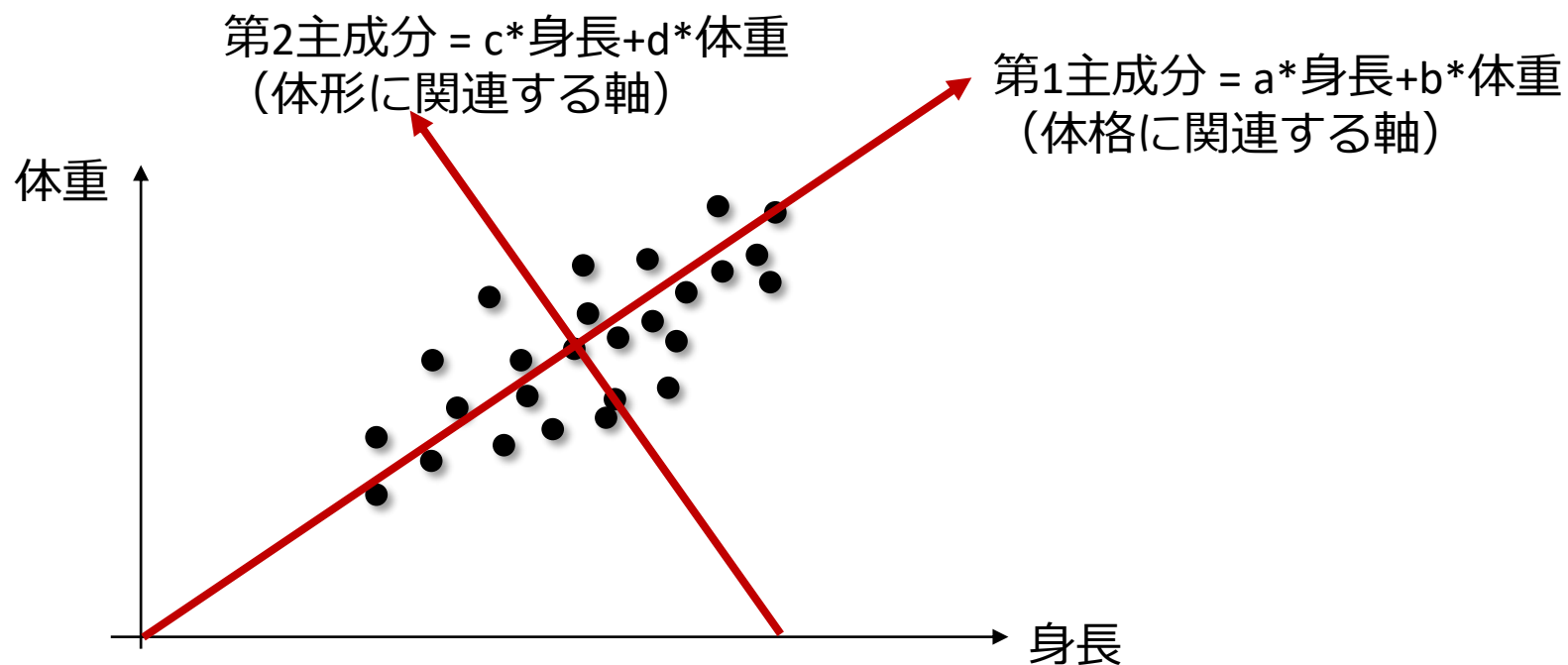


# 2次元データに対する主成分分析の例

最も分散が大きくなる方向へ、第1主成分の軸を作成します。

次に、第1主成分と直交し、かつ2番目に分散が大きくなる方向へ第2主成分の軸を作成します。（D次元データの場合、この処理を繰り返して第D主成分まで作成される）

つまり、番号が小さい主成分ほど分散が大きく、情報を多く保持していることになります。

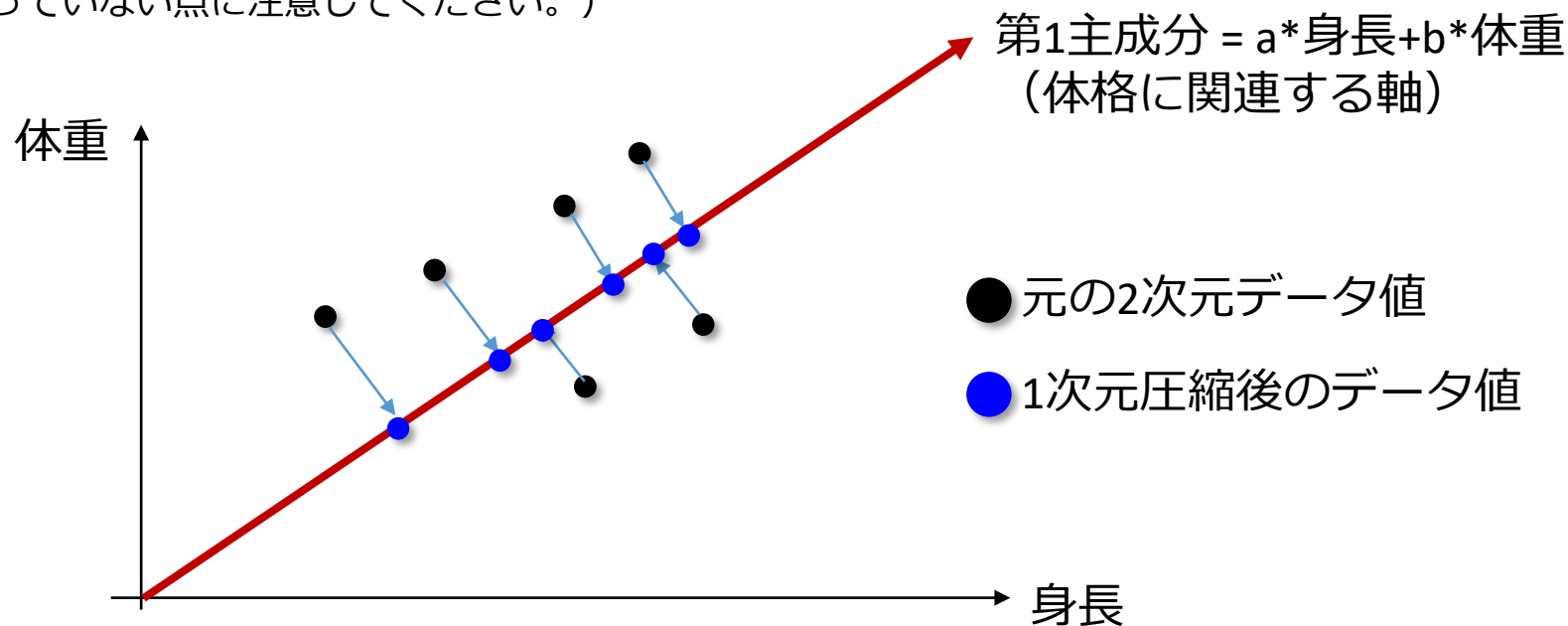


# 2次元データに対する主成分分析の例

第1主成分軸に向かって各2次元データを射影すると、2次元から1次元に圧縮したことになります。

いわば、身長と体重の2軸で表現されたデータが、新たにつくられた「体格」の軸のみで表現されるようになったということです。

(ここでは作成された軸を便宜上「体格」と呼んでいます。主成分分析は単に分散の大きい軸を作っているだけなので、実際は軸の意味付けなどは行っていない点に注意してください。)





# 主成分分析の係数の求め方

(導出はかなり長くなるので、割愛します。興味のある方は以下を読んでください。)

加納学, “主成分分析” <http://manabukano.brilliant-future.net/document/text-PCA.pdf>

1. 元のデータ  $x$  の分散共分散行列  $\text{cov}$  を計算する。
2. 分散共分散行列を固有値と固有ベクトルに分解する。
3. 固有値が、圧縮後の分散値に相当する。  
そこで、固有値が大きい順に固有ベクトルを並べ、それらを第1～第D主成分とする。
4. 二次元プロットをしたい場合は、第1主成分と第2主成分の軸に、各データを射影する。  
具体的には、1番目の固有ベクトルと元データの内積を第1主成分、2番目の固有ベクトルと元データの内積を第2主成分とする。

# 主成分分析を動かしてみよう

説明だけではいまいちピンとこないと思います。

05\_01\_pca.ipynbを動かして、主成分分析の動作を確認してみましょう。

(PCA: principal component analysis = 主成分分析の略)

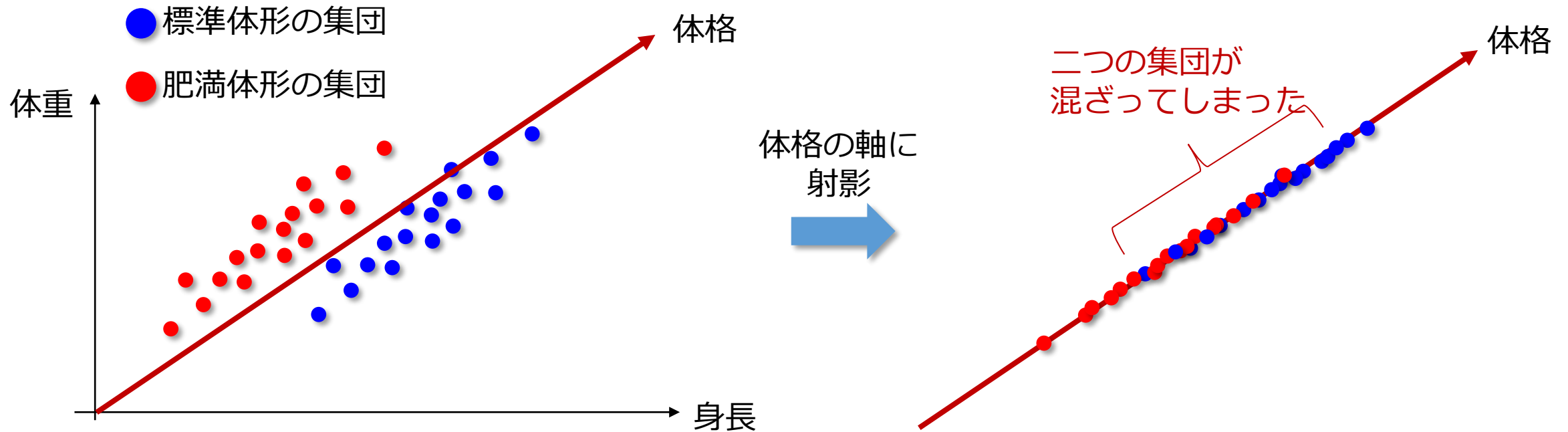
# 多次元データの可視化 2

## 線形判別分析

# 分散だけが重要な情報なのか？

主成分分析では分散最大化という基準で、身長と体重から「体格」という軸を作りだし、その軸へデータを射影することで次元圧縮をしていました。

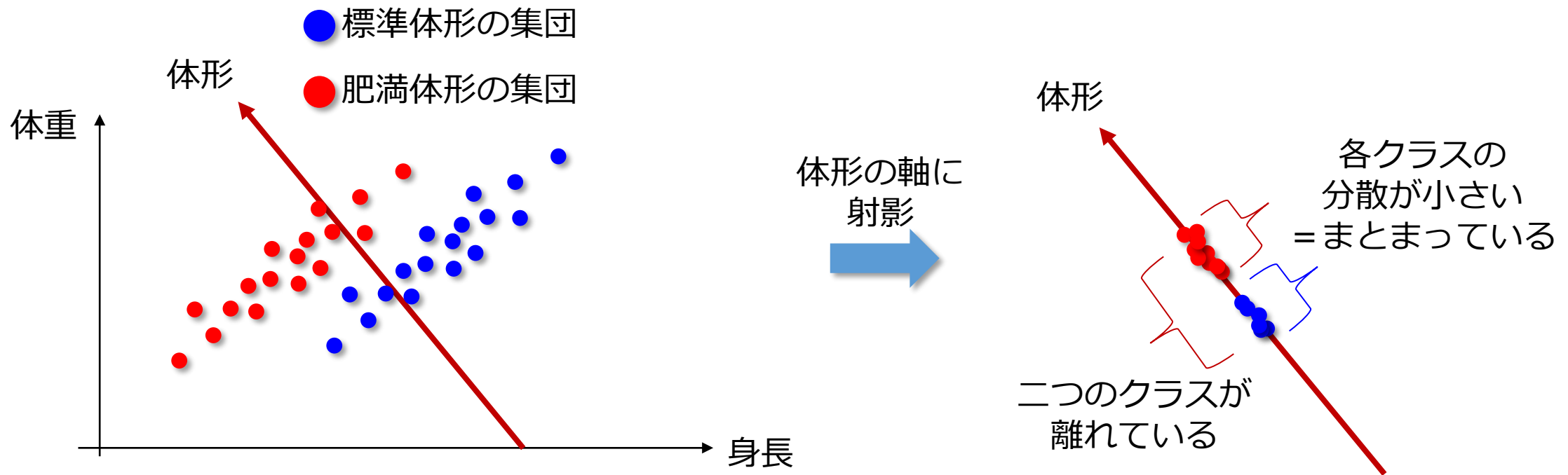
しかし実は入力データは標準体型の集団と肥満体型の集団で構成されており、二つの集団を見やすくすることが目的だった場合、必ずしも分散最大基準は正しいでしょうか？



# クラスの判別を目的とした圧縮基準

以下の条件を満たす場合、クラスの判別がしやすい圧縮になっていると言えます。

- 同一クラスが密集している = クラス内の分散が小さい
- 異なるクラスの分布が離れている = クラス間の分散が大きい



# クラス内分散共分散とクラス間分散共分散

クラス内分散共分散：小さいほど良い

$$S_{inner} = \sum_{c=1}^C \sum_{n=1}^{N^c} (\mathbf{x}_n^c - \bar{\mathbf{x}}^c)(\mathbf{x}_n^c - \bar{\mathbf{x}}^c)^T$$

クラスcに属するデータの  
分散共分散行列

$\mathbf{x}_n^c$ : クラスcに属するデータのnサンプル目  
 $\bar{\mathbf{x}}^c$ : クラスcに属するデータの平均値ベクトル  
 $C$ : クラスの数  
 $N^c$ : クラスcに属するデータのサンプル数

クラス間分散共分散：大きいほど良い

$$S_{intra} = \sum_{c=1}^C N^c (\bar{\mathbf{x}}^c - \bar{\mathbf{x}})(\bar{\mathbf{x}}^c - \bar{\mathbf{x}})^T$$

各クラスの平均値ベクトル $\bar{\mathbf{x}}^c$ を  
各データ $\mathbf{x}_n^c$ の代わりに  
用いた分散共分散

$\bar{\mathbf{x}}$ : データ全体の平均値ベクトル

最終的に、クラス間分散/クラス内分散が大きくなれば良い。

$$J = (S_{inner})^{-1} S_{intra}$$

# 線形判別分析の係数の求め方

1. 元のデータ  $x$  のクラス間分散/クラス内分散比  $J = (S_{inner})^{-1}S_{intra}$  を求める。
2.  $J$  を固有値と固有ベクトルに分解する。
3. 固有値が、圧縮後の分散値に相当する。  
そこで、固有値が大きい順に固有ベクトルを並べ、それらを第1～第D主成分とする。
4. 二次元プロットをしたい場合は、第1主成分と第2主成分の軸に、各データを射影する。  
具体的には、1番目の固有ベクトルと元データの内積を第1主成分、2番目の固有ベクトルと元データの内積を第2主成分とする。

主成分分析との違いは、固有値分解する対象が

- 主成分分析：データ全体の分散共分散行列
- 線形判別分析：クラス間分散/クラス内分散比行列

# 線形判別分析を動かしてみよう

05\_02\_lda.ipynbを動かして，線形判別分析の動作を確認してみましょう。

(LDA: linear discriminant analysis = 線形判別分析の略)



# おわりに

今回は、多次元データの扱い方について触れ、可視化方法として主成分分析と線形判別分析を紹介しました。

- 主成分分析：次元圧縮後のデータの分散を最大にする。
- 線形判別分析：次元圧縮後のデータのクラス間分散/クラス内分散比を最大にする。

これらの次元圧縮方法は、可視化だけではなく、パターン認識などにもよく使われます。

次元を減らせるため、過学習を軽減する効果があります。

# レポート課題

## 課題 1

第 5 回ファイル一式に含まれる, "car.csv" のデータに対して主成分分析を行い, 第 1 主成分と第 2 主成分の固有ベクトルを見ることで, 第 1 主成分と第 2 主成分がそれぞれどういった軸を作成したか考察せよ。

(例えば05\_01\_pca.ipynbでは「文系理系の偏り」や「総合的なテストの出来」という軸に解釈できる。というような形で記述すること。)

# レポート課題

## 課題 2

第5回ファイル一式に含まれる, "xyz.csv"のデータに対して線形判別分析を行え。

xyz.csvは, 二つのクラスに属するデータのx軸, y軸, z軸の値が記述された3次元データである。

2次元に圧縮してプロットし, 線形判別分析がうまく機能しているか否か答えよ。

「うまく機能している」と答えた場合はその理由を答えよ。

「うまく機能していない」と答えた場合はその理由に加えて, なぜうまく機能しないのか, どうすればうまく機能するか, について考察せよ。

5/24(火)は休講です

レポート提出期限: 5/31(火) AM10:30, ipynbファイルをhtmlファイルに変換して提出