今後の日程

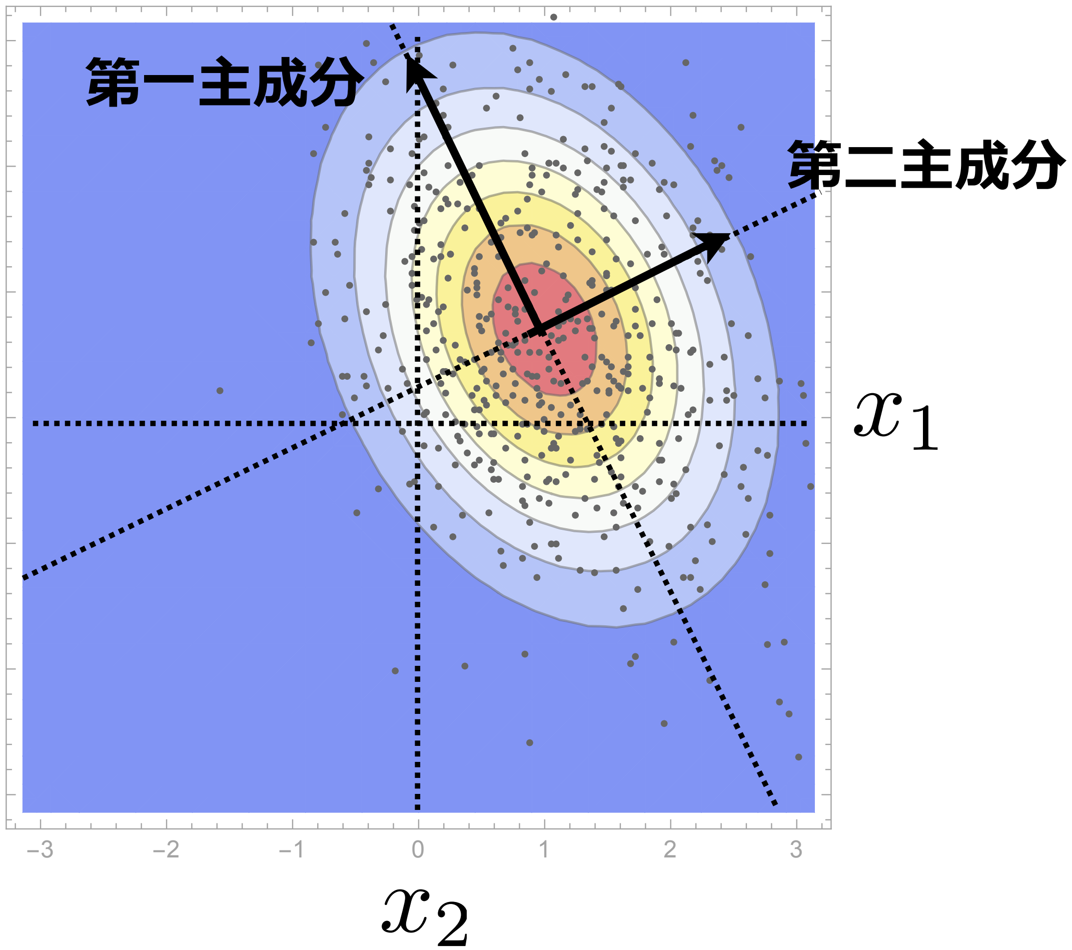
7/28 午前：固有値と２次形式　　午後：主軸変換と主成分分析
8/4　午前：数量化と特異値分解　午後：まとめ・アンケート
重要：本日配布のレポート試験の回答を8/5までに提出

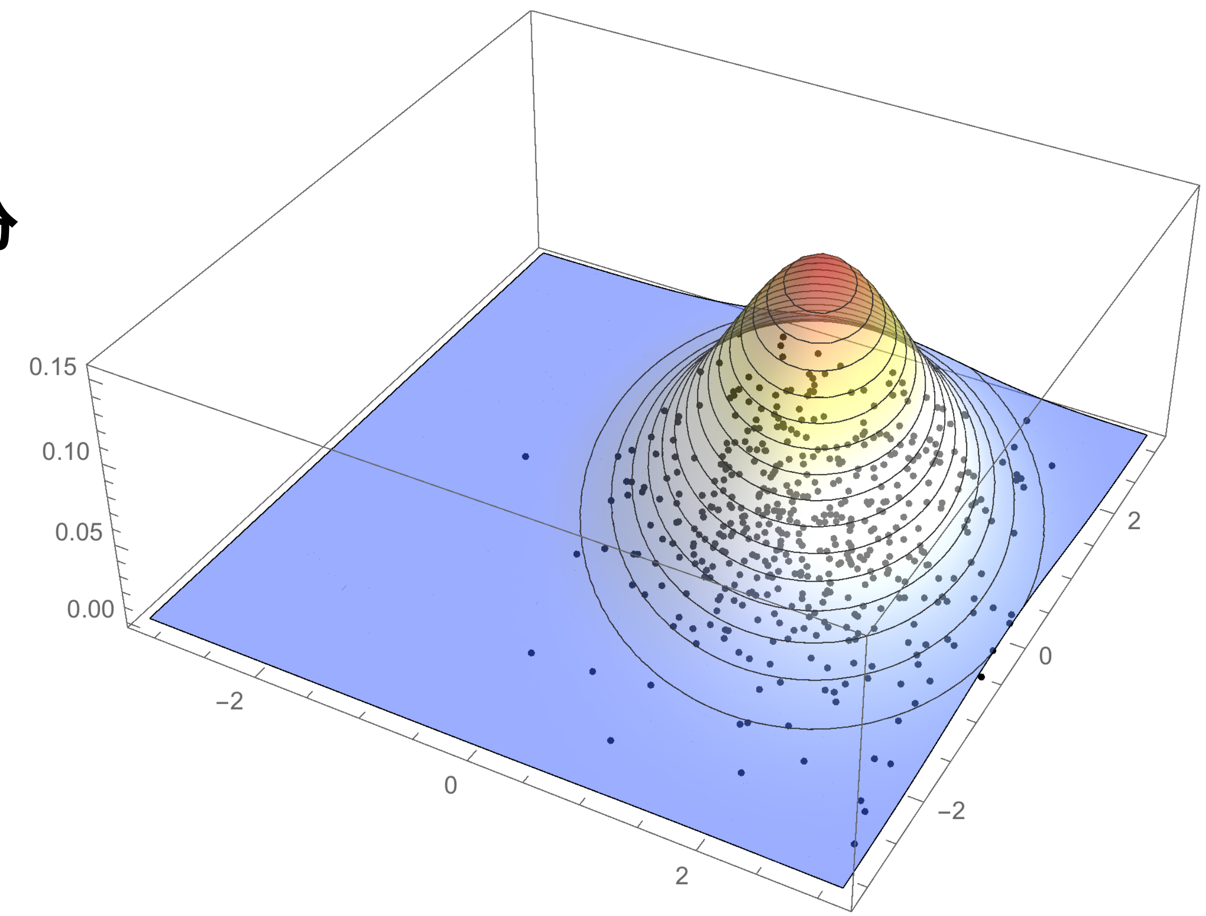
今日の話：主成分分析 (PCA)

[午前]
- 分散共分散行列
- 行列の固有値分解
- ２次形式の標準化

[午後]
- 主軸変換
- 主成分分析
- 主成分負荷量・寄与率・因子負荷量

主成分分析: Principal Component Analysis (PCA)





知りたいこと (教科書 p.6)

- 主成分により低い次元でデータを解釈できないか？
- それぞれの主成分の説明力はどれくらいか？
- 主成分により(x_1, x_2)の特徴付け・分類ができるか？

1.6　主成分分析とは

表 1.7 は，4 教科の試験の成績である．すべて量的変数と考える．

表 1.7　試験の成績のデータ

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

この（4 次元）データに基づいて知りたいことは次の通りである．

(1) 主成分の構成により低い次元でデータを解釈できないか．

(2) それぞれの主成分の説明力はどれくらいか．

(3) 科目や生徒の特徴付けおよび分類をどのようにできるか．

表 1.7 のデータを主成分分析で解析することにより次のことがわかる．

(1) 主要な主成分として次の第 1 主成分 z_1 と第 2 主成分 z_2 を得る．

$$z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$

$$z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$

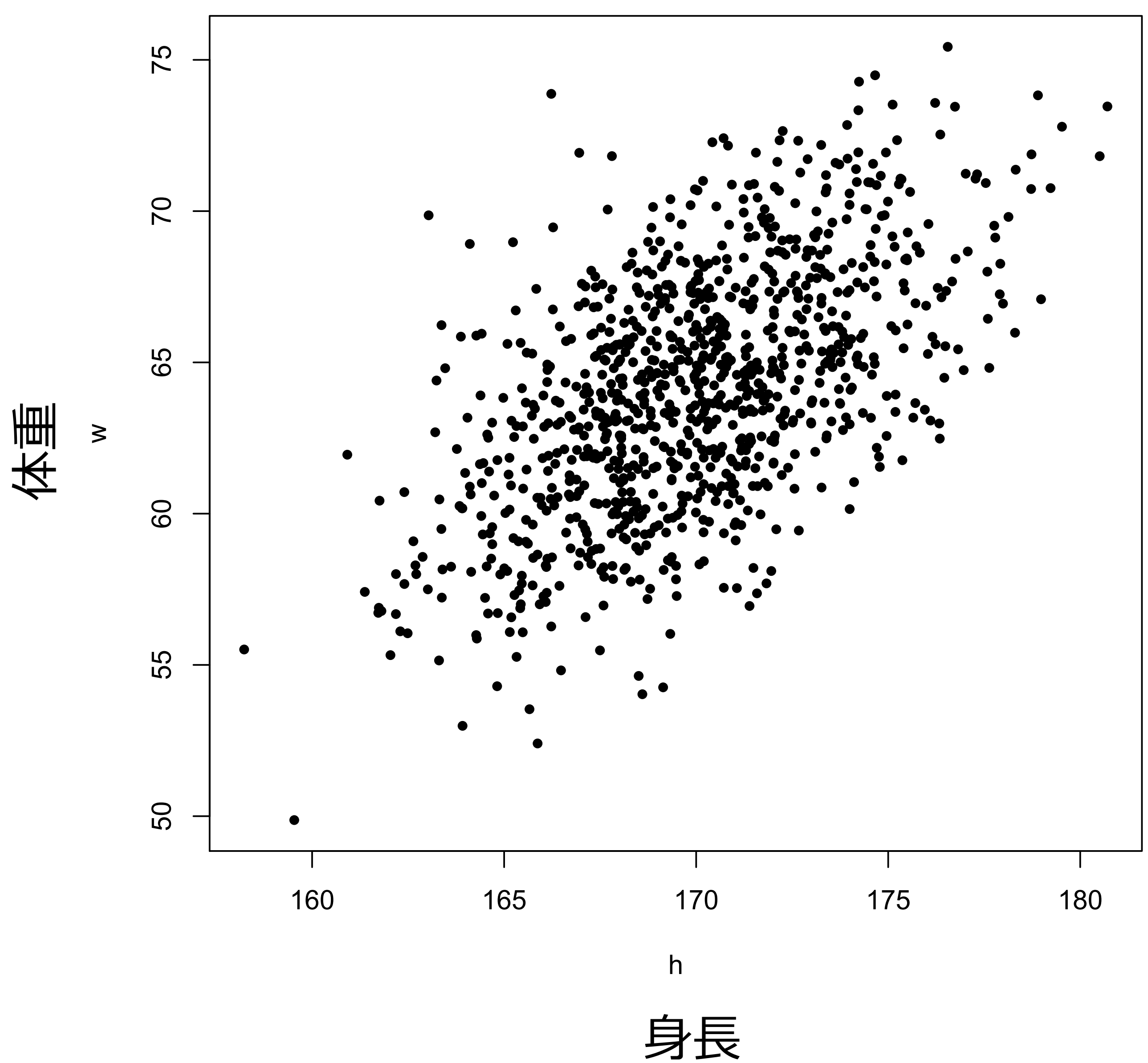
ここで，u_j は変数 x_j を標準化したものである．

(2) z_1 の寄与率は 0.680，z_2 の寄与率は 0.306 である．第 2 主成分までの累積寄与率は $0.680 + 0.306 = 0.986$ である．

(3) 係数の値より，z_1 は「総合的学力」，z_2 は「文系・理系の学力の違い」を表すと解釈できる．各生徒に対して z_1 と z_2 を計算することにより，生徒の特徴付けや分類ができる．

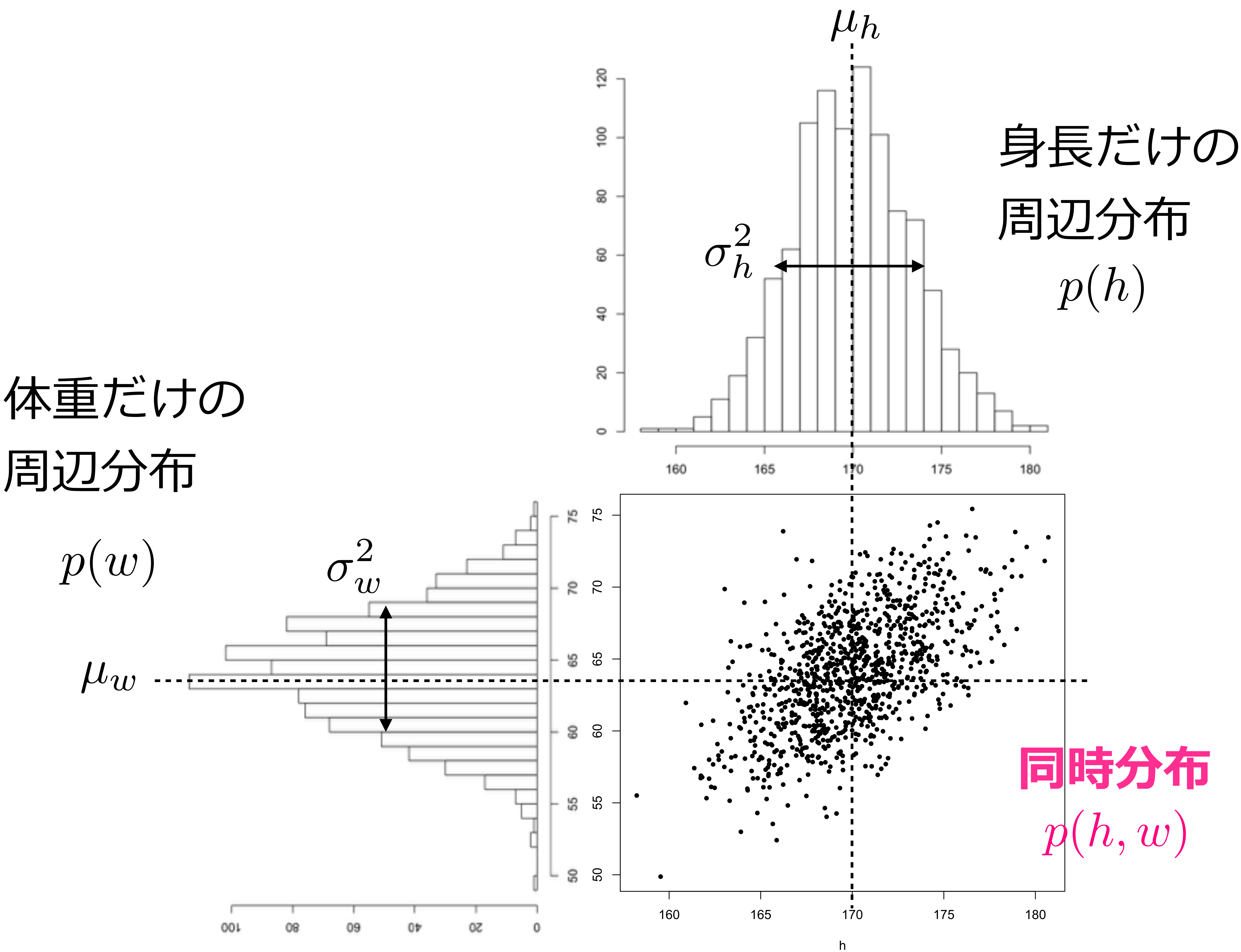
多変量のデータのイメージ (2変量の例)

散布図



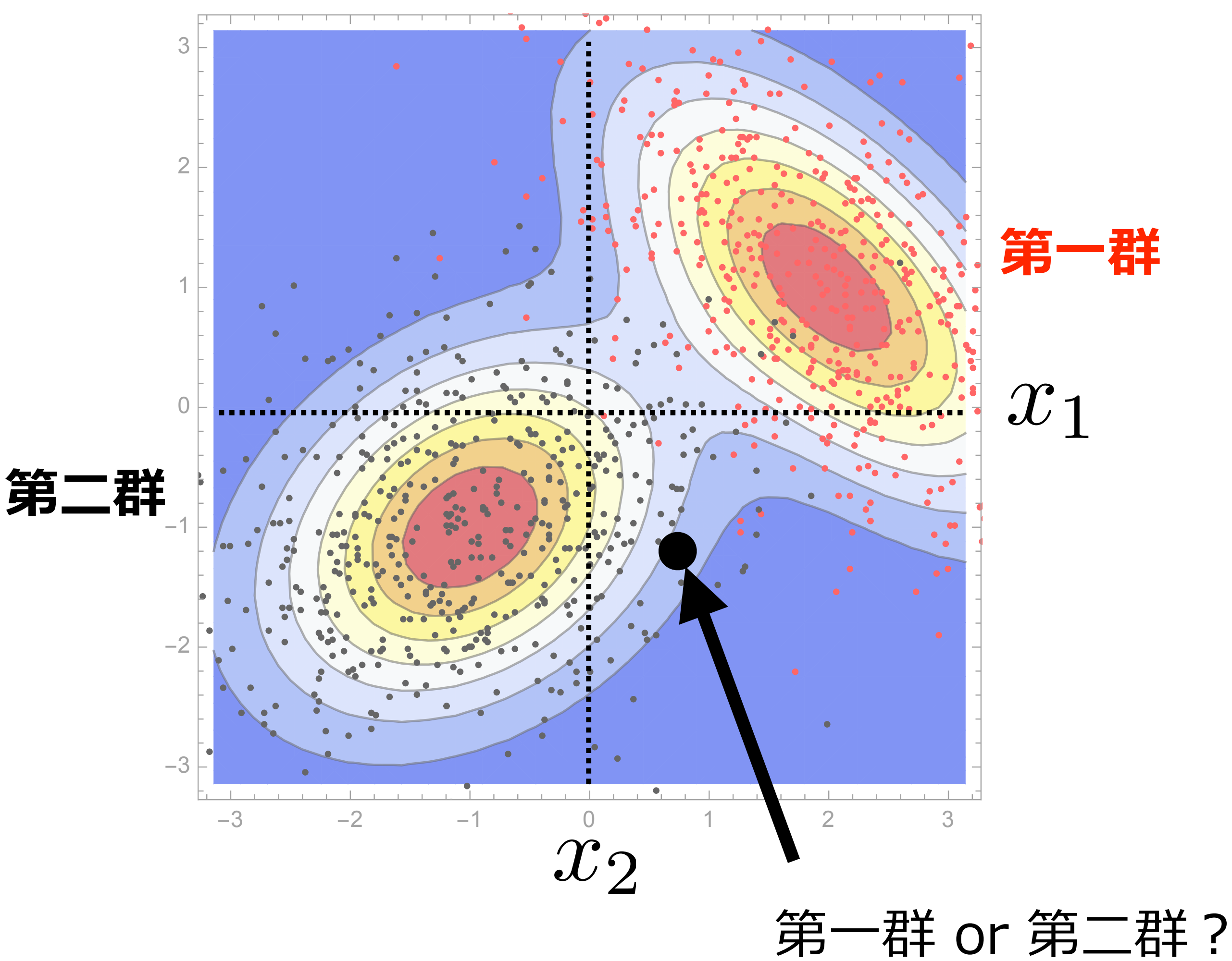
表形式データ

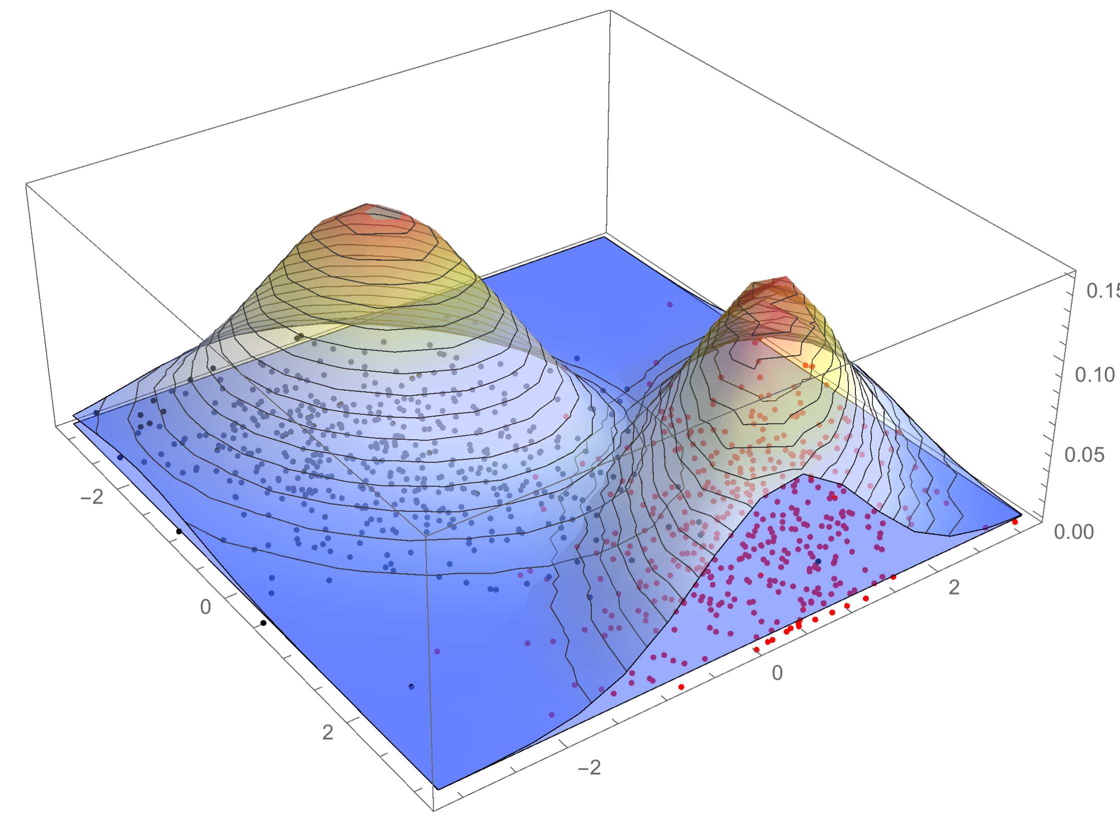
	身長 h	体重
1	174.0	64.1
2	169.6	65.4
3	168.4	67.4
4	171.7	63.4
5	172.1	72.3
6	167.0	63.4
7	167.0	62.5
:	:	:
999	172.7	64.9
1000	167.3	61.97



μ_h
σ_h^2
身長だけの
周辺分布
$p(h)$
体重だけの
周辺分布
$p(w)$
σ_w^2
μ_w
同時分布
$p(h, w)$
h
160
165
170
175
180
0
20
40
60
80
100
120
50
55
60
65
70
75

判別分析: Discriminant Analysis

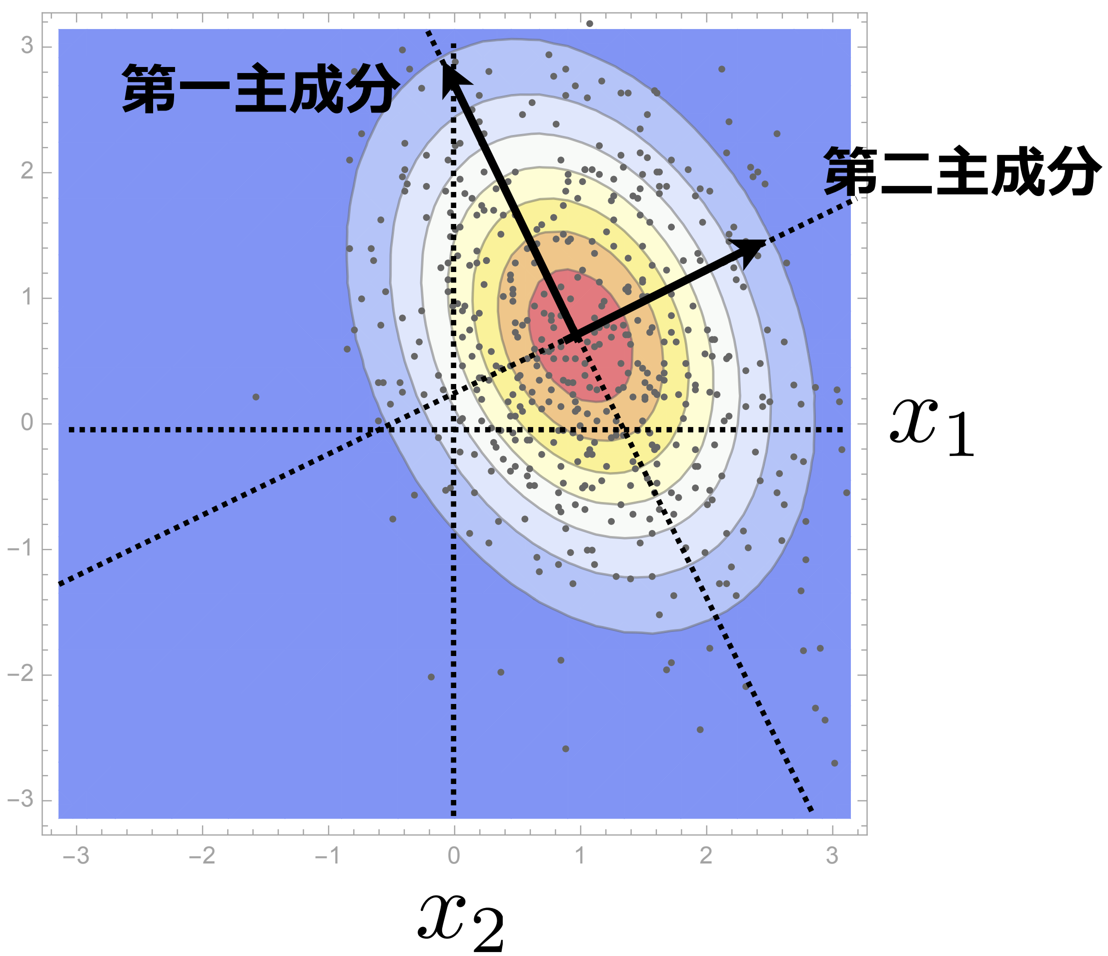


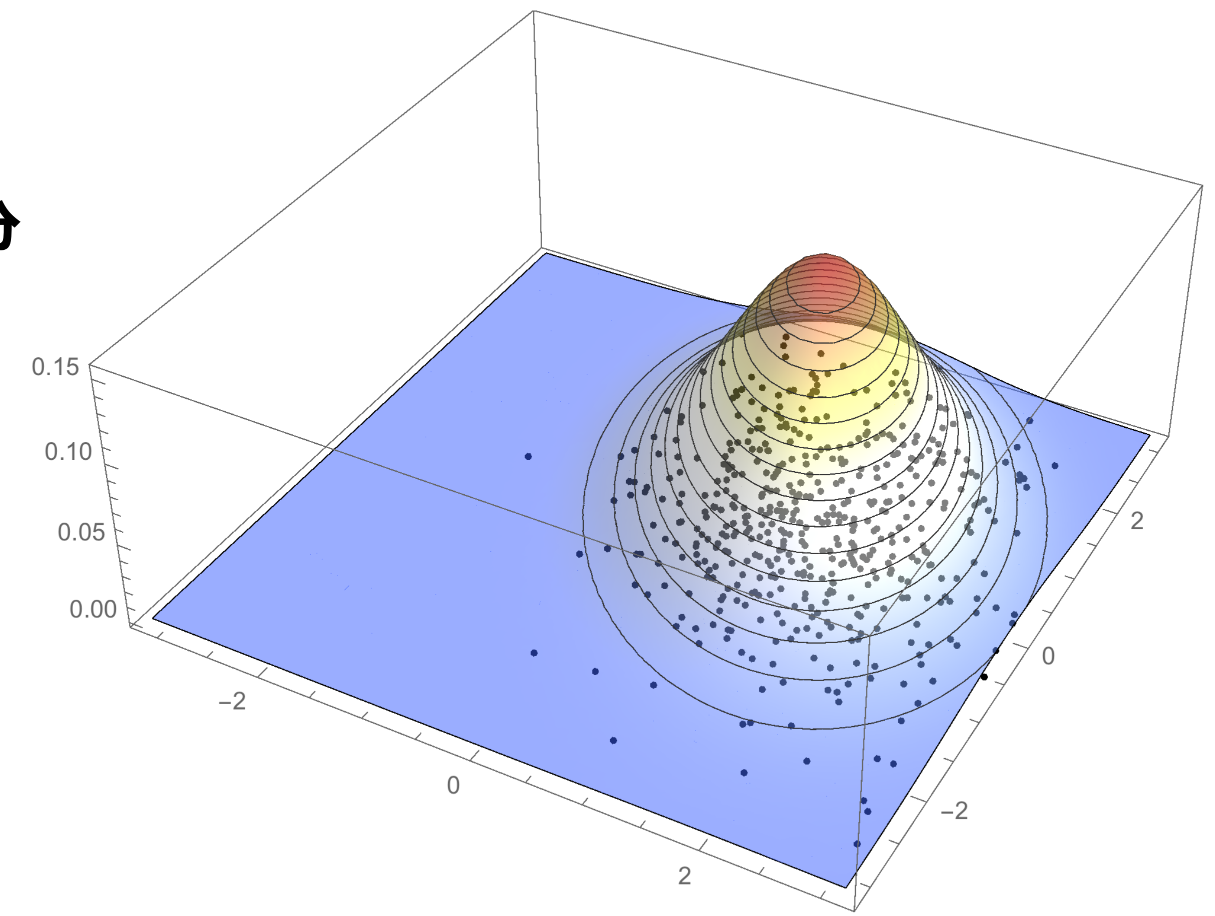


知りたいこと
(教科書 p.4)

- 検査値 (x_1, x_2) から第 1 群か第 2 群か判別できるか？
- 判別できるとすればその精度はどれくらいか？
- 例えば $x_1 = 0.7, x_2 = -1.1$ ならどう判別されるか？

主成分分析: Principal Component Analysis (PCA)





知りたいこと (教科書 p.6)

- 主成分により低い次元でデータを解釈できないか？
- それぞれの主成分の説明力はどれくらいか？
- 主成分により(x_1, x_2)の特徴付け・分類ができるか？

価格.com カメラ

デジタル一眼カメラ 人気ランキング

	売れ筋ランキング	注目ランキング	満足度ランキング
1位	 CANON EOS Kiss X7 ダブルズームキット **¥45,580**	 ニコン D500 ボディ **¥200,988**	 ニコン D810 24-120 VR レンズキット **¥301,781**
2位	 パナソニック LUMIX DMC-GF7W ダブルズームレンズキット **¥52,000**	 パナソニック LUMIX DMC-GF7W ダブルズームレンズキット **¥52,000**	 富士フイルム FUJIFILM X-Pro2 ボディ **¥158,044**
3位	 ニコン D500 ボディ **¥200,988**	 CANON EOS 5D Mark III ボディ **¥235,903**	 CANON EOS 8000D ボディ **¥73,799**

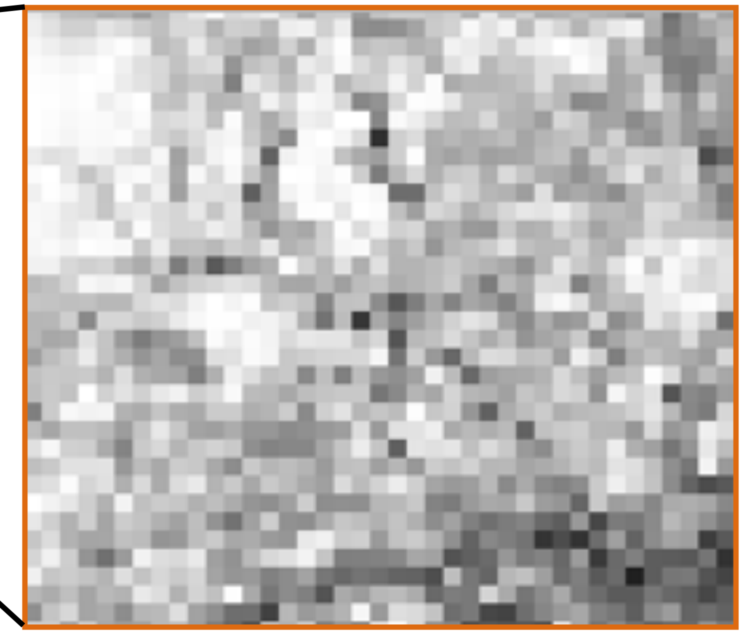
記録形式

記録フォーマット	DCF2.0
画像タイプ	JPEG、RAW（14bit、キヤノン独自）、RAW+JPEGラージ同時記録可能
記録画素数	L（ラージ）　：約1790万（5184×3456）画素 M（ミドル）　：約800万（3456×2304）画素 S1（スモール1）　：約450万（2592×1728）画素 S2（スモール2）　：約250万（1920×1280）画素 S3（スモール3）　：約35万（720×480）画素 RAW（ロウ）　：約1790万（5184×3456）画素

多変量としての「画像」

3456

5184



192	194	255	194	192	194	255	194	192	194	255	193	192	194	255	193
171	173	255	172	171	173	255	172	171	173	255	172	171	173	255	172
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
233	233	255	233	233	233	255	233	233	233	255	233	233	233	255	233
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239
239	239	255	239	239	239	255	239	239	239	255	239	239	239	255	239

$5184 \times 3456 = 17915904$

17915904個の要素を持つ
多変量ベクトルとして扱う

17915904個の要素を持つ多変量ベクトルとして扱う

次元削減問題 (p>>n問題)

p = 17915904

n = 1200枚あっても全然足らない…

(17915904次元空間に1200個の点がある状況)

でも、ほんとに17915904変量も要る？？

(固有ベクトルがほんとにp本要る？？)

固有ベクトル(主成分を与える基底)をひとつづつみてみよう…

固有顔 (Eigenface)

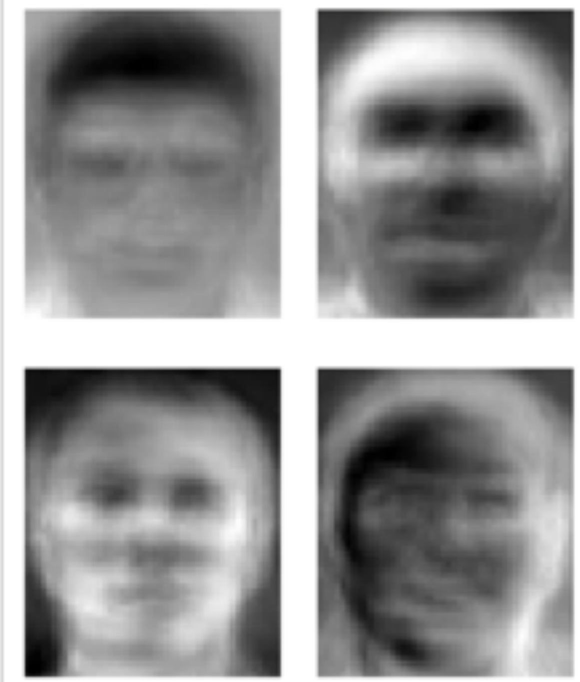
【例 6.13】 図 6.9 の上段左端は 50 人の顔画像（32×32 画素）をそれぞれ 24 通りの照明条件で撮影した，合計 1200 枚の顔画像の平均画像 $\boldsymbol{g}_0$ である．この 1200 枚の画像から計算した固有ベクトルから作られる基底の最初の 4 個 $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_4$ をその右に順に並べてある．

図 6.9　上段: 平均顔画像と第 1, 2, 3, 4 基底画像．中段と下段: 入力画像とその第 1, 2, 3, 4 近似．

固有顔

固有顔（英: **Eigenface**）とは、顔認識システムというコンピュータビジョンの応用で使われる固有ベクトルの集合である。固有顔を利用した顔認識は1987年、Matthew Turk と Alex Pentland が開発した。

この固有ベクトル群は、「人間の考えられる顔」の高次元ベクトル空間の確率分布についての共分散行列から得られる。

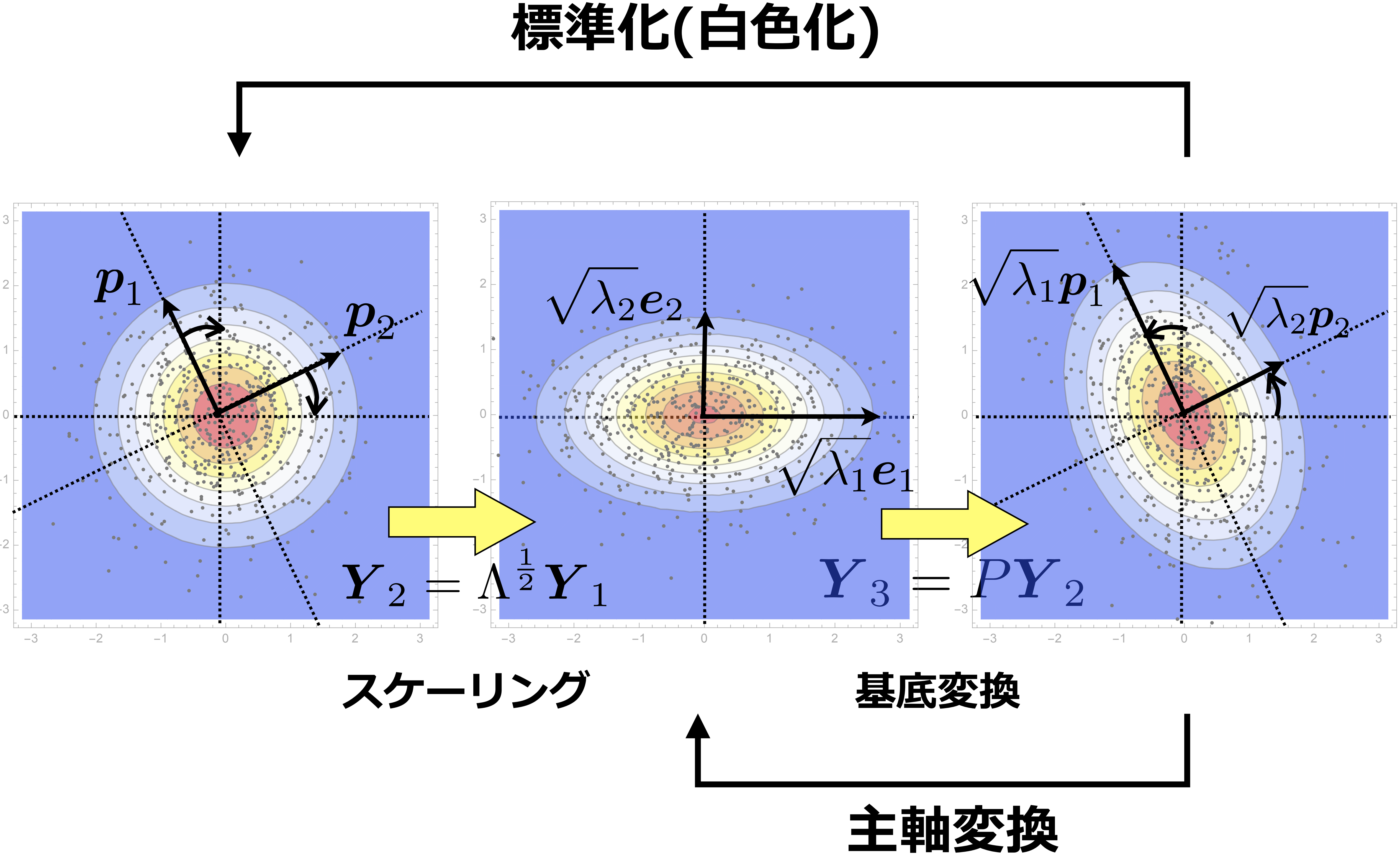


固有顔の例。AT&T研究所より

目次 [非表示]

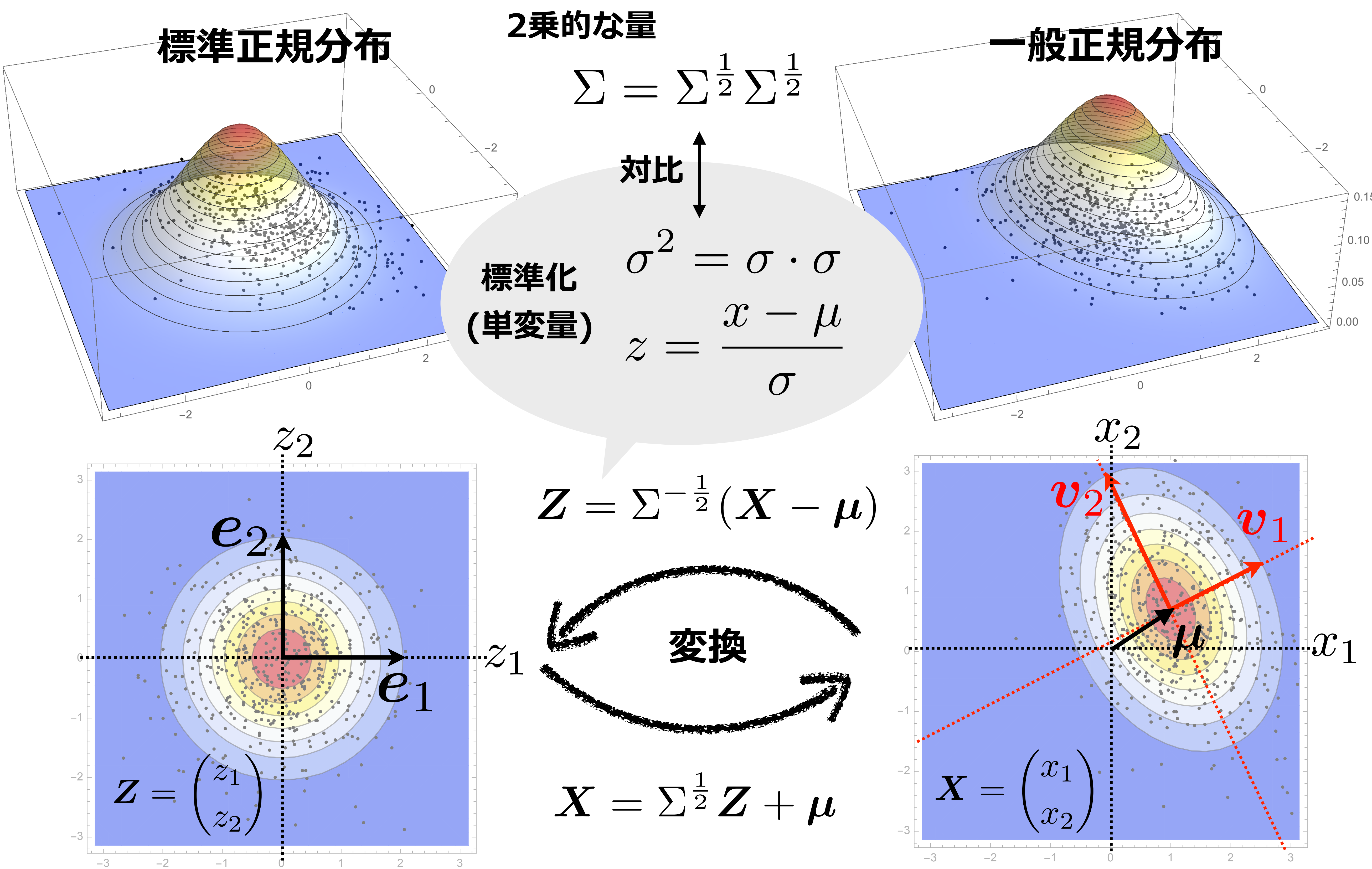
固有顔生成 [編集]

固有顔集合を生成するには、人間の顔のデジタル化された画像（同じ照明条件でなければならない）を多数集積し、目と鼻の位置を合わせる。そして、解像度を合わせるために再標本化する。このようなデータから主成分分析 (PCA) と呼ばれる手段で固有顔を抽出する。顔画像を固有顔に変換する過程は以下の通りである。



固有ベクトル方向に直交座標を取り直してから成分表示

基底の変換〜分散共分散行列〜正規分布

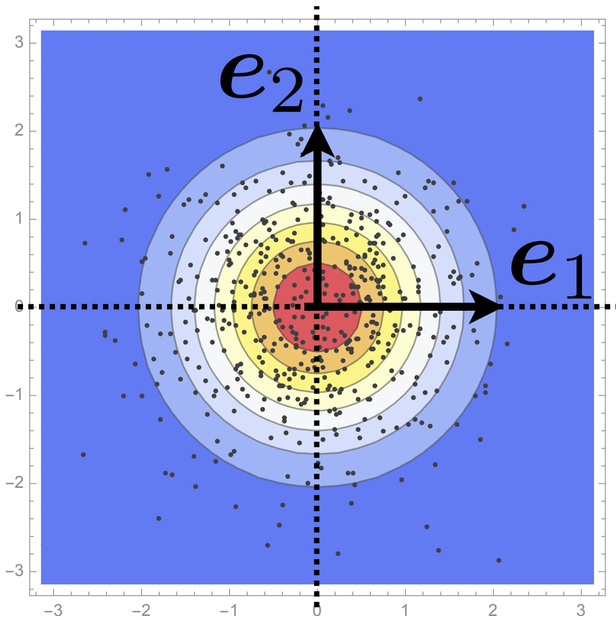


前回資料

標準化変換 (基底変換+平行移動)

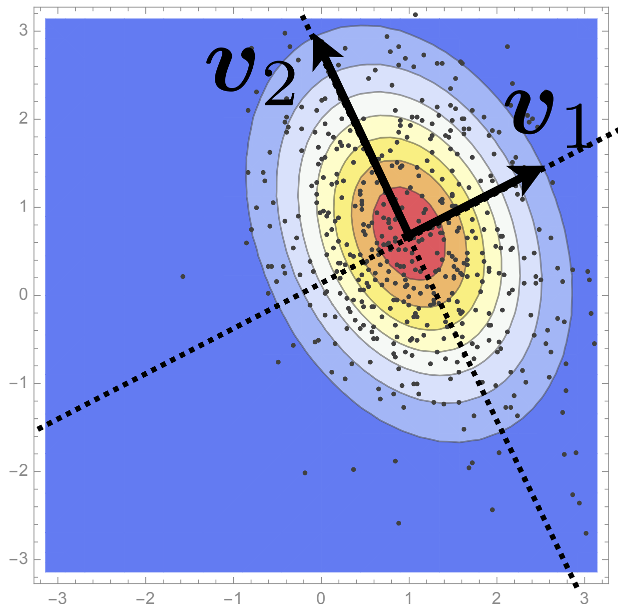
等確率面は球形

2D:等高線が真円形



等確率面は楕円体形

2D:等高線が楕円形



参考：確率変数の標準化 (単変量)

$$\sigma^2 = \sigma \cdot \sigma$$

$$Z := \frac{X - \mu}{\sigma}$$

Zの平均=0
Zの分散=1

確率変数の標準化 (多変量)

$$\boldsymbol{Z} := \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}$$

Zの平均ベクトル=0
Zの分散共分散行列=単位行列

分散共分散行列

多変量の分布の"散らばり"具合 (2次の統計量) を司る

$$\Sigma = \begin{bmatrix} \sigma_{x_1,x_1} & \sigma_{x_2,x_1} & \cdots & \sigma_{x_p,x_1} \\ \sigma_{x_1,x_2} & \sigma_{x_2,x_2} & \cdots & \sigma_{x_p,x_2} \\ \vdots & & & \\ \sigma_{x_1,x_p} & \sigma_{x_2,x_p} & \cdots & \sigma_{x_p,x_p} \end{bmatrix}$$

i, j 要素は x_i と x_j の共分散　σ_{x_i,x_j}

$$\sigma_{x_i,x_j} := \mathbb{E}\{(x_i - \mu_i)(x_j - \mu_j)\} \qquad \mu_i := \mathbb{E}\{x_i\}$$

→ 対角要素は分散　$\sigma_{x_i,x_i} = \mathbb{E}\{(x_i - \mu_i)^2\} = \sigma_{x_i}^2$

分散共分散行列：ベクトル・行列表記

$$\boldsymbol{X} := \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \boldsymbol{\mu} := \mathbb{E}\{\boldsymbol{X}\} = \begin{bmatrix} \mathbb{E}\{x_1\} \\ \mathbb{E}\{x_2\} \\ \vdots \\ \mathbb{E}\{x_p\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

とベクトル表記すれば

$$\Sigma = \mathbb{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\} = \mathbb{E}\left\{ \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 & \cdots & x_p - \mu_p \end{bmatrix} \right\}$$

$$= \mathbb{E}\left\{ \begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1) & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_p-\mu_p) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)(x_2-\mu_2) & \cdots & (x_2-\mu_2)(x_p-\mu_p) \\ \vdots & & & \\ (x_p-\mu_p)(x_1-\mu_1) & (x_p-\mu_p)(x_2-\mu_2) & \cdots & (x_p-\mu_p)(x_p-\mu_p) \end{bmatrix} \right\}$$

分散共分散行列の固有構造 (スペクトル分解)

$$\Sigma = P \Lambda P'$$

$$\begin{bmatrix} \sigma_{x_1,x_1} & \sigma_{x_2,x_1} & \cdots & \sigma_{x_p,x_1} \\ \sigma_{x_1,x_2} & \sigma_{x_2,x_2} & \cdots & \sigma_{x_p,x_2} \\ \vdots & & & \\ \sigma_{x_1,x_p} & \sigma_{x_2,x_p} & \cdots & \sigma_{x_p,x_p} \end{bmatrix} = \underset{\text{固有ベクトル}}{\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{p2} \\ \vdots & & & \\ p_{p1} & p_{p2} & \cdots & p_{pp} \end{bmatrix}} \underset{\text{固有値}}{\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}} \underset{\text{固有ベクトル(転置)}}{\begin{bmatrix} p_{11} & p_{21} & \cdots & p_{p1} \\ p_{12} & p_{22} & \cdots & p_{p2} \\ \vdots & & & \\ p_{1p} & p_{2p} & \cdots & p_{pp} \end{bmatrix}}$$

$$= \lambda_1 \begin{bmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{p1} \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{p1} \end{bmatrix} + \lambda_2 \begin{bmatrix} p_{12} \\ p_{22} \\ \vdots \\ p_{p2} \end{bmatrix} \begin{bmatrix} p_{12} & p_{22} & \cdots & p_{p2} \end{bmatrix} + \cdots + \lambda_p \begin{bmatrix} p_{1p} \\ p_{2p} \\ \vdots \\ p_{pp} \end{bmatrix} \begin{bmatrix} p_{1p} & p_{2p} & \cdots & p_{pp} \end{bmatrix}$$

- 対称行列なので固有ベクトルは直交する
- 正定値行列なので正定値な平方根行列が唯一存在

分散共分散行列の平方根行列

$$\Sigma^{\frac{1}{2}} = P \quad \Lambda^{\frac{1}{2}} \quad P'$$

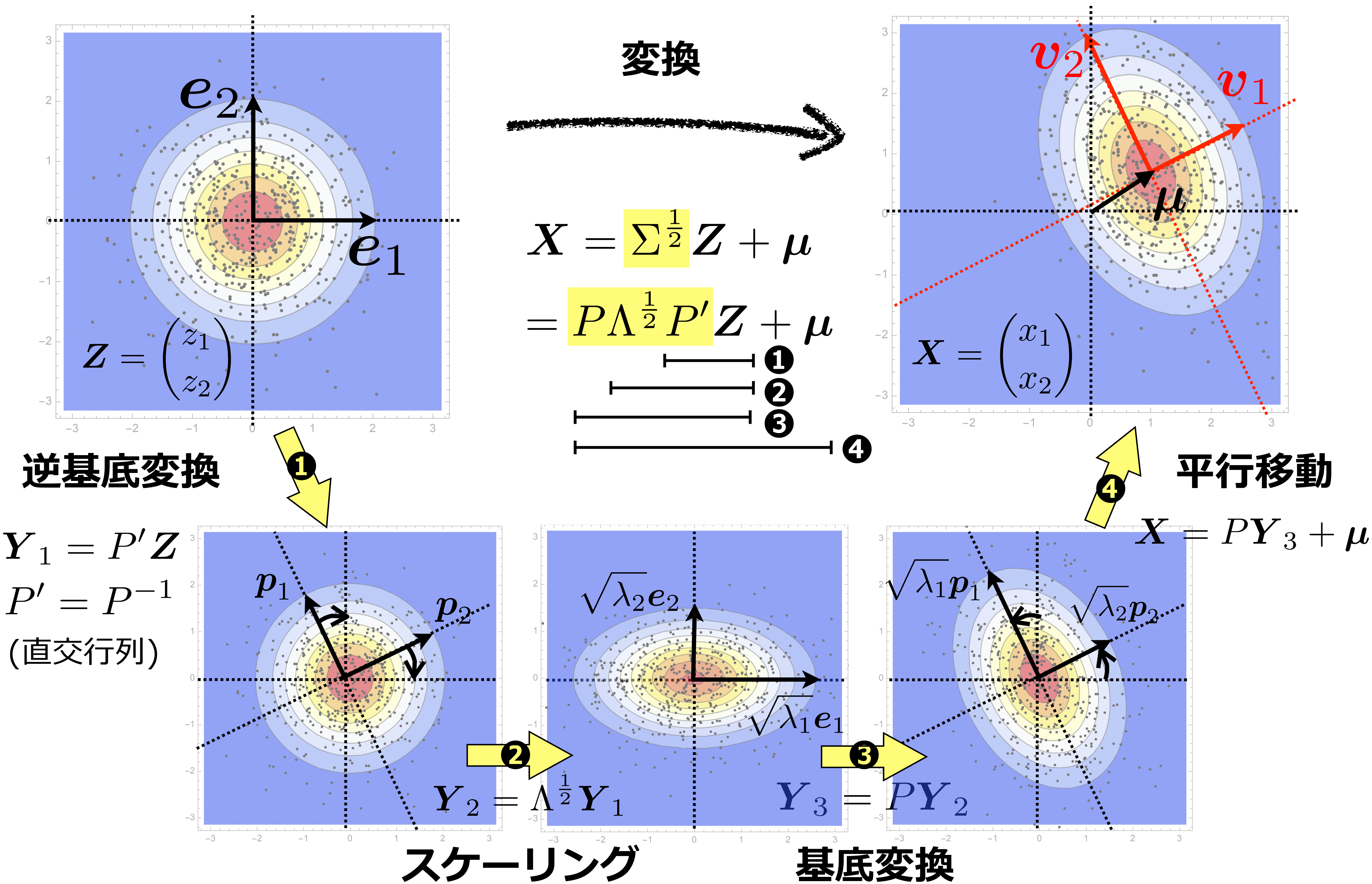
$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{p2} \\ \vdots & & & \\ p_{p1} & p_{p2} & \cdots & p_{pp} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{p1} \\ p_{12} & p_{22} & \cdots & p_{p2} \\ \vdots & & & \\ p_{1p} & p_{2p} & \cdots & p_{pp} \end{bmatrix}$$

- 二乗したらΣ

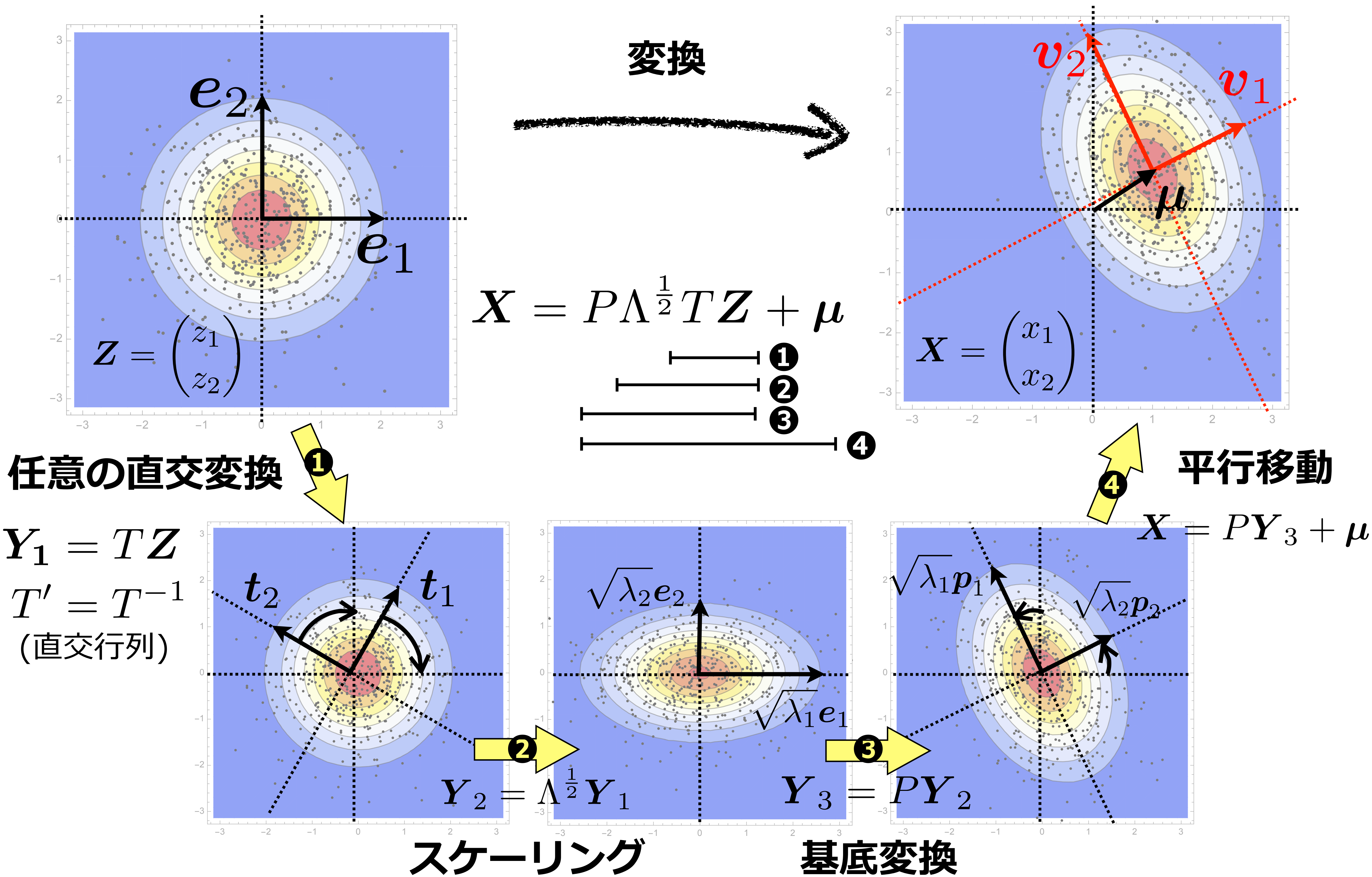
$$\Sigma^{1/2}\Sigma^{1/2} = P\Lambda^{1/2}P'P\Lambda^{1/2}P' = P\Lambda^{1/2}\Lambda^{1/2}P' = \Sigma$$

- 固有ベクトルはΣと同じ
- 正定値行列

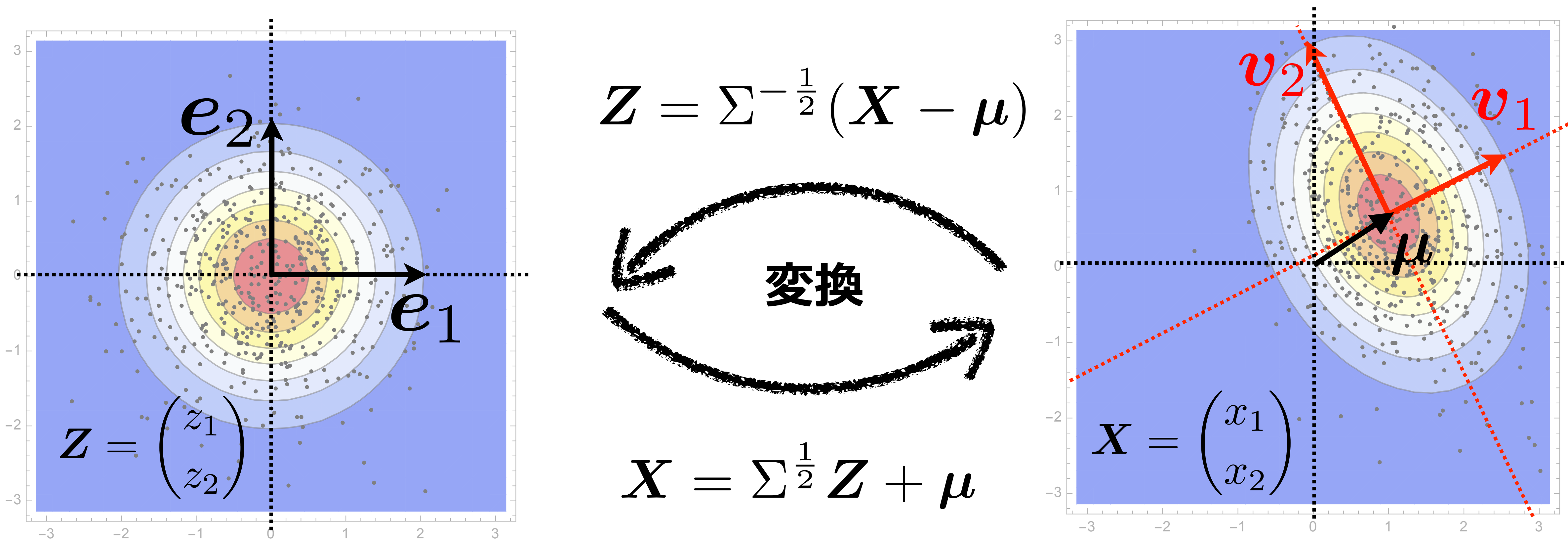
基底変換と固有値・固有ベクトル



注意：実は直交変換分の自由度がある



平方根行列だと対称+固有構造不変かつ逆変換も奇麗



いずれにしても言える事

$v_1 = \sqrt{\lambda_1} p_1$

$v_2 = \sqrt{\lambda_2} p_2$

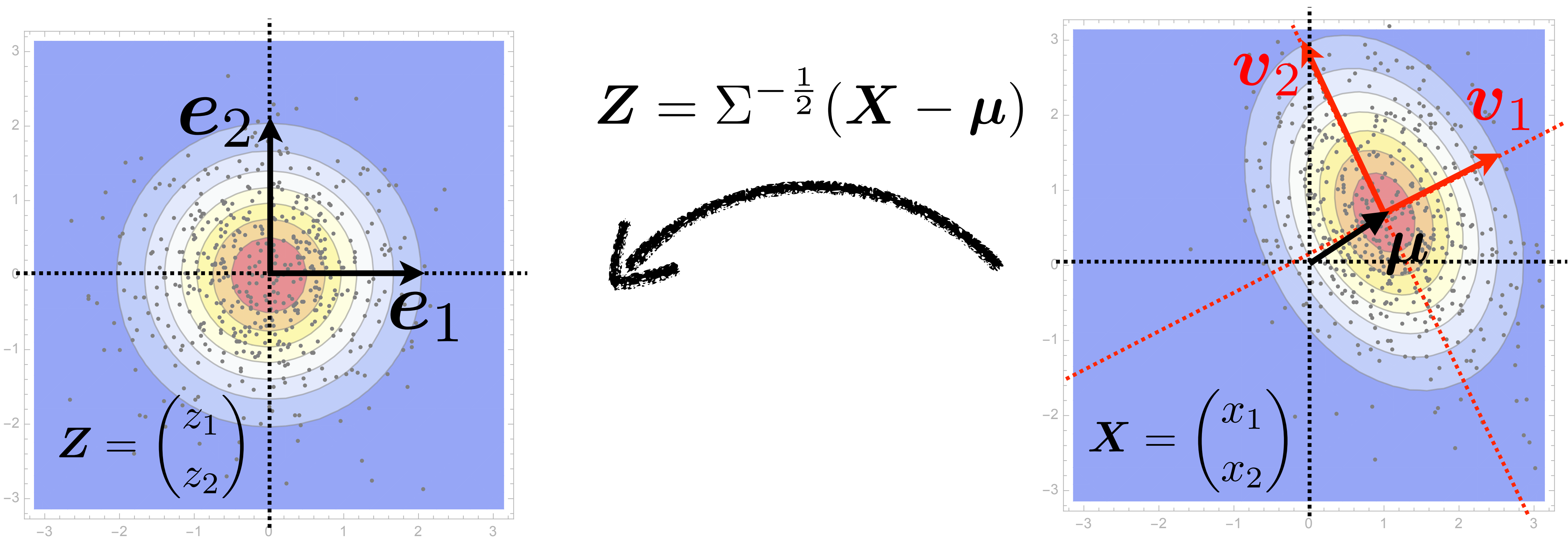
λ_i　Σ の固有値

p_i　Σ の固有ベクトル

同確率面(2Dでは等高線)の楕円体の長軸・短軸は
分散共分散行列の固有値・固有ベクトルで決まる！

等確率面の楕円体の式は？

逆変換した後に原点から等距離な点の集合



$$\|\boldsymbol{Z}\|^2 = \mathrm{Const.} \quad \Leftrightarrow \quad \|\Sigma^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})\|^2 = \mathrm{Const.}$$

マハラノビス距離

$$\|\Sigma^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})\|^2 = (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

正規分布の密度も一定 $\quad p(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right)$

前回資料

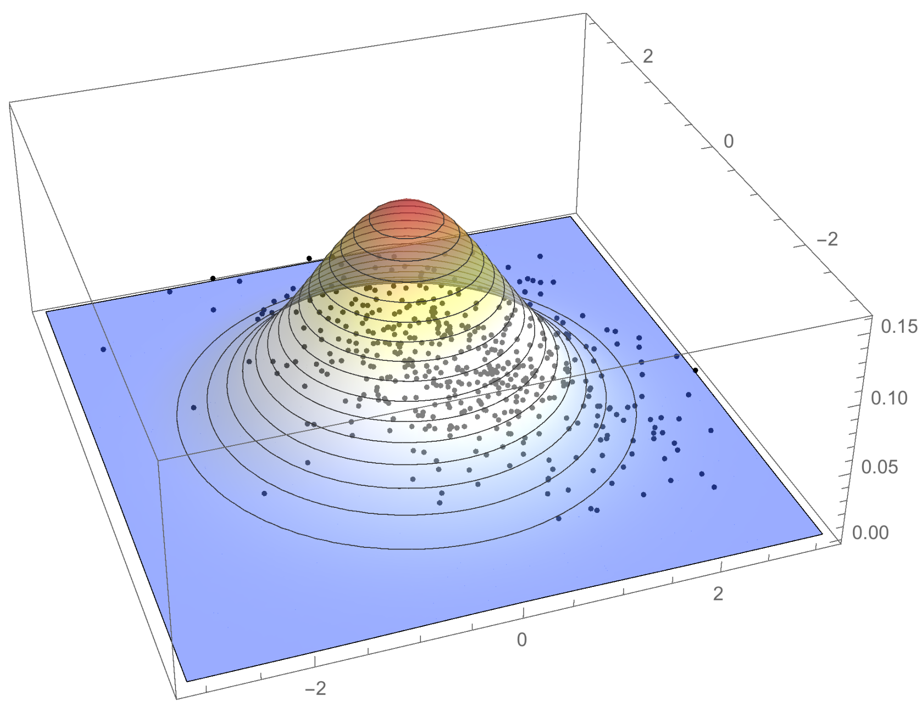
標準正規分布 vs 一般の多変量正規分布

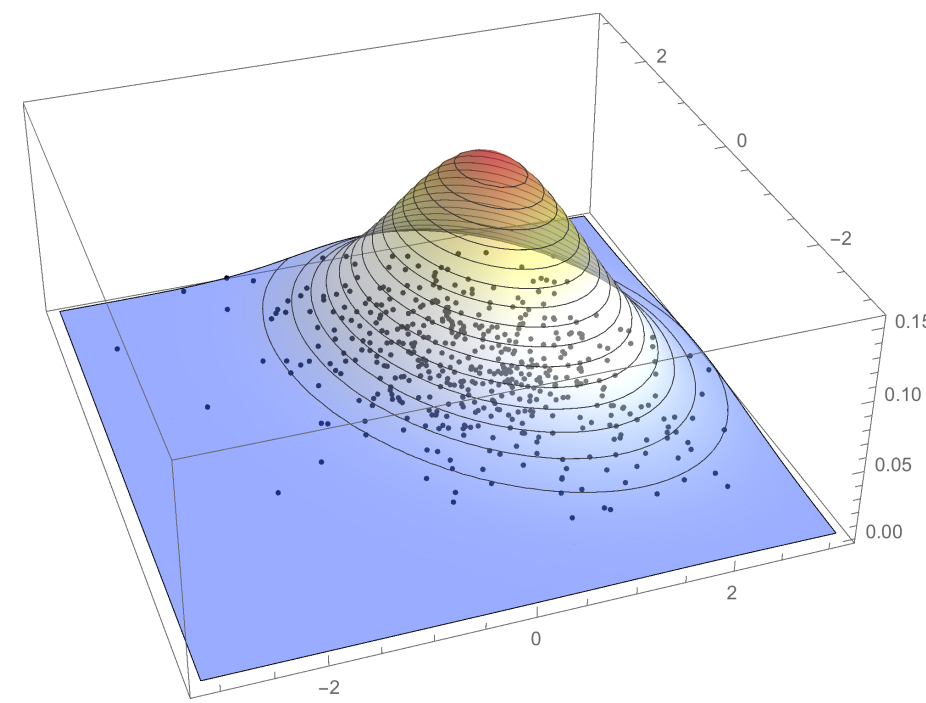
$N(\mathbf{0}, \boldsymbol{I})$

$$\frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}\right)$$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

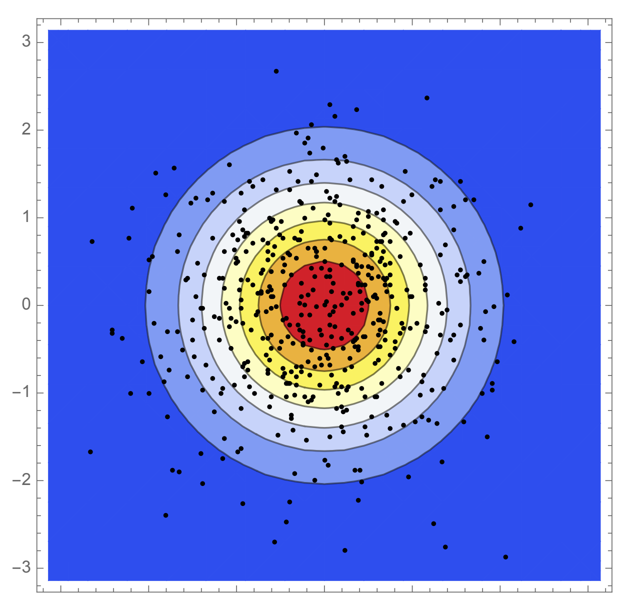
$$\frac{1}{(\sqrt{2\pi})^n\sqrt{\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$





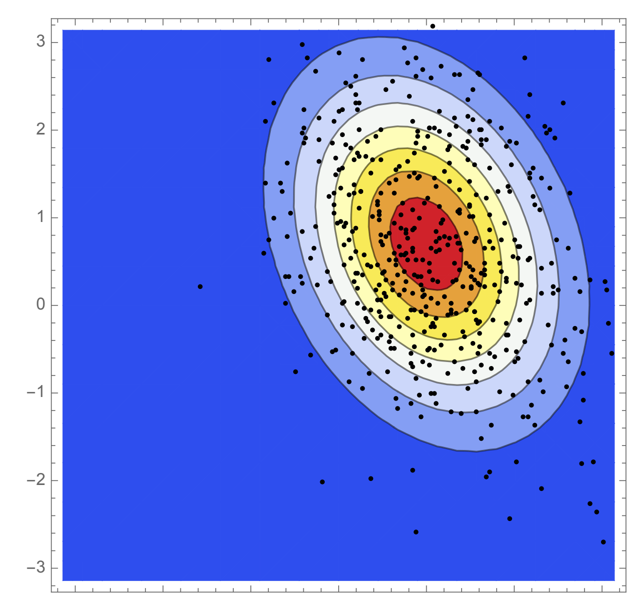
等確率面は球形

2D:等高線が真円形



等確率面は楕円体形

2D:等高線が楕円形



相関係数: 標準化変数の共分散

共分散 $\sigma_{x_i,x_j} := \mathbb{E}\{(x_i - \mu_i)(x_j - \mu_j)\}$

相関係数
$$r_{i,j} := \mathbb{E}\left\{\left(\frac{x_i - \mu_i}{\sigma_i}\right)\cdot\left(\frac{x_j - \mu_j}{\sigma_j}\right)\right\}$$
$$= \mathbb{E}\{z_i \cdot z_j\}$$

標準化 (平均0,分散1の確率変数に変換)

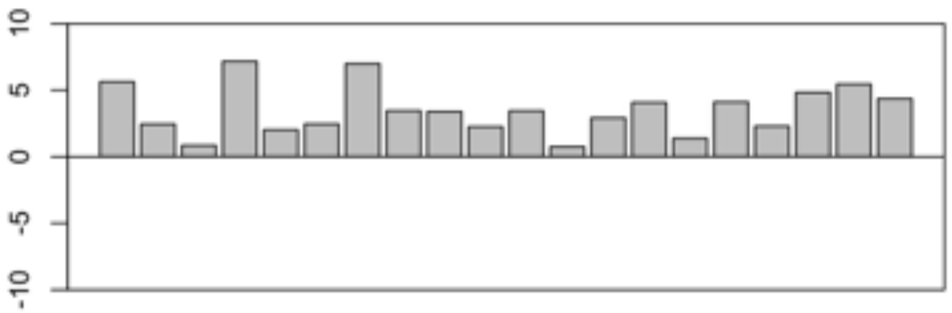
$$z_i = \frac{x_i - \mu_i}{\sigma_i} \Rightarrow \begin{array}{l} \mathbb{E}\{z_i\} = 0 \\ \mathrm{var}\{z_i\} = \mathbb{E}\{(z_i - 0)^2\} = 1 \end{array}$$

相関係数の直感的イメージ

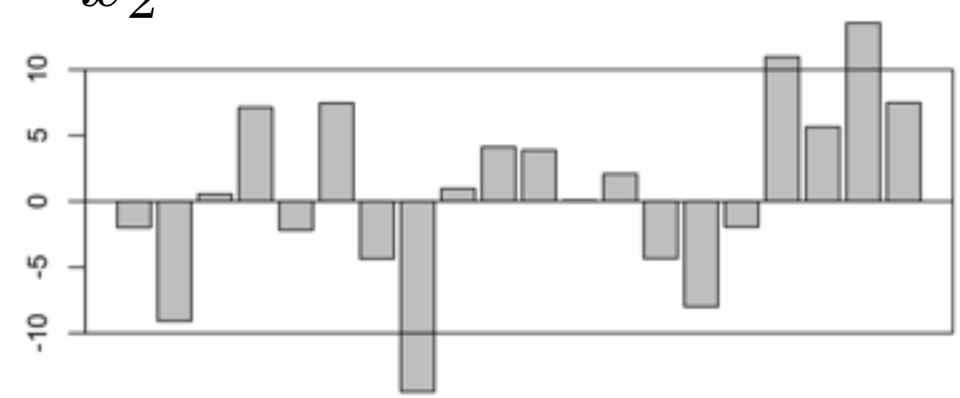
相関係数　$r_{i,j} := \mathbb{E}\left\{\left(\frac{x_i - \mu_i}{\sigma_i}\right)\cdot\left(\frac{x_j - \mu_j}{\sigma_j}\right)\right\}$

$$= \mathbb{E}\{z_i \cdot z_j\}$$

x_1

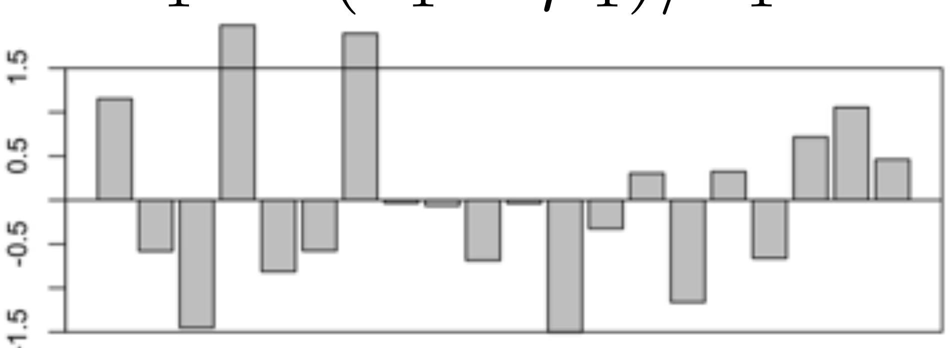


x_2

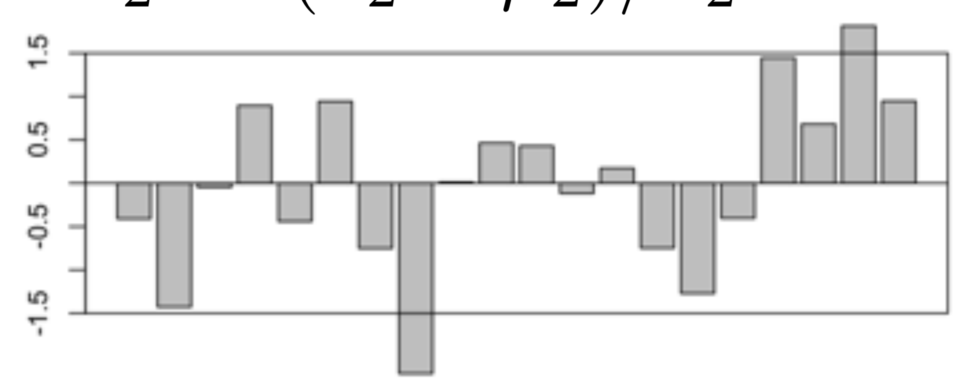


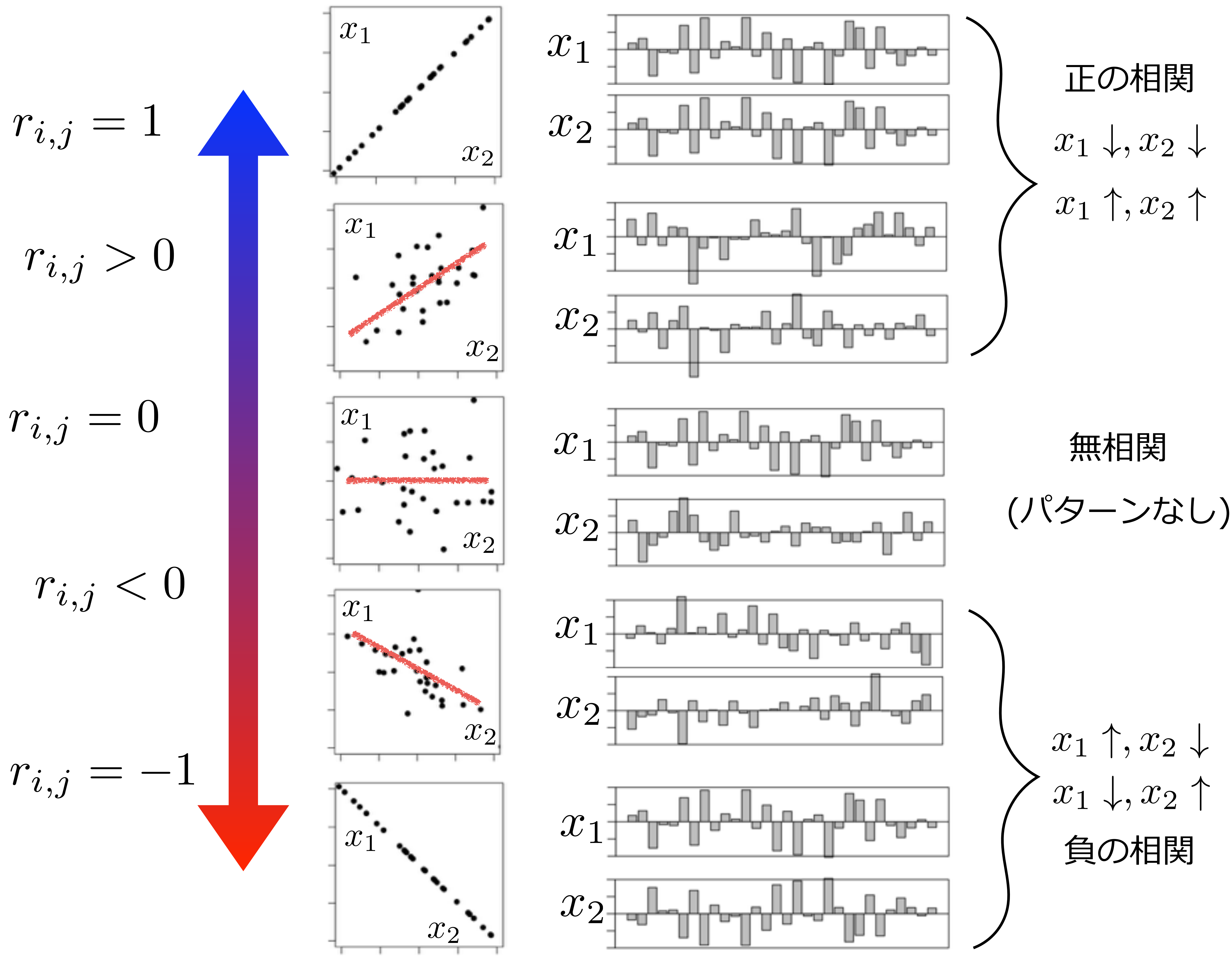
標準化　分布の真ん中(平均)とブレ幅(分散)を整えた後の変動の類似

$z_1 := (x_1 - \mu_1)/\sigma_1$



$z_2 := (x_2 - \mu_2)/\sigma_2$




$r_{i,j} = 1$
$r_{i,j} > 0$
$r_{i,j} = 0$
$r_{i,j} < 0$
$r_{i,j} = -1$
x_1
x_2
正の相関
$x_1 \downarrow, x_2 \downarrow$
$x_1 \uparrow, x_2 \uparrow$
無相関
(パターンなし)
$x_1 \uparrow, x_2 \downarrow$
$x_1 \downarrow, x_2 \uparrow$
負の相関

相関行列との関係：ベクトル・行列表記

$$
\underbrace{\begin{bmatrix} \sigma_{x_1,x_1} & \sigma_{x_2,x_1} & \cdots & \sigma_{x_p,x_1} \\ \sigma_{x_1,x_2} & \sigma_{x_2,x_2} & \cdots & \sigma_{x_p,x_2} \\ \vdots & & & \\ \sigma_{x_1,x_p} & \sigma_{x_2,x_p} & \cdots & \sigma_{x_p,x_p} \end{bmatrix}}_{\Sigma}
=
\underbrace{\begin{bmatrix} \sigma_{x_1} & 0 & \cdots & 0 \\ 0 & \sigma_{x_2} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sigma_{x_p} \end{bmatrix}}_{D}
\underbrace{\begin{bmatrix} \frac{\sigma_{x_1,x_1}}{\sigma_{x_1}\sigma_{x_1}} & \frac{\sigma_{x_2,x_1}}{\sigma_{x_2}\sigma_{x_1}} & \cdots & \frac{\sigma_{x_p,x_1}}{\sigma_{x_p}\sigma_{x_1}} \\ \frac{\sigma_{x_1,x_2}}{\sigma_{x_1}\sigma_{x_2}} & \frac{\sigma_{x_2,x_2}}{\sigma_{x_2}\sigma_{x_2}} & \cdots & \frac{\sigma_{x_p,x_2}}{\sigma_{x_p}\sigma_{x_2}} \\ \vdots & & & \\ \frac{\sigma_{x_1,x_p}}{\sigma_{x_1}\sigma_{x_p}} & \frac{\sigma_{x_2,x_p}}{\sigma_{x_2}\sigma_{x_p}} & \cdots & \frac{\sigma_{x_p,x_p}}{\sigma_{x_p}\sigma_{x_p}} \end{bmatrix}}_{R}
\underbrace{\begin{bmatrix} \sigma_{x_1} & 0 & \cdots & 0 \\ 0 & \sigma_{x_2} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sigma_{x_p} \end{bmatrix}}_{D}
$$

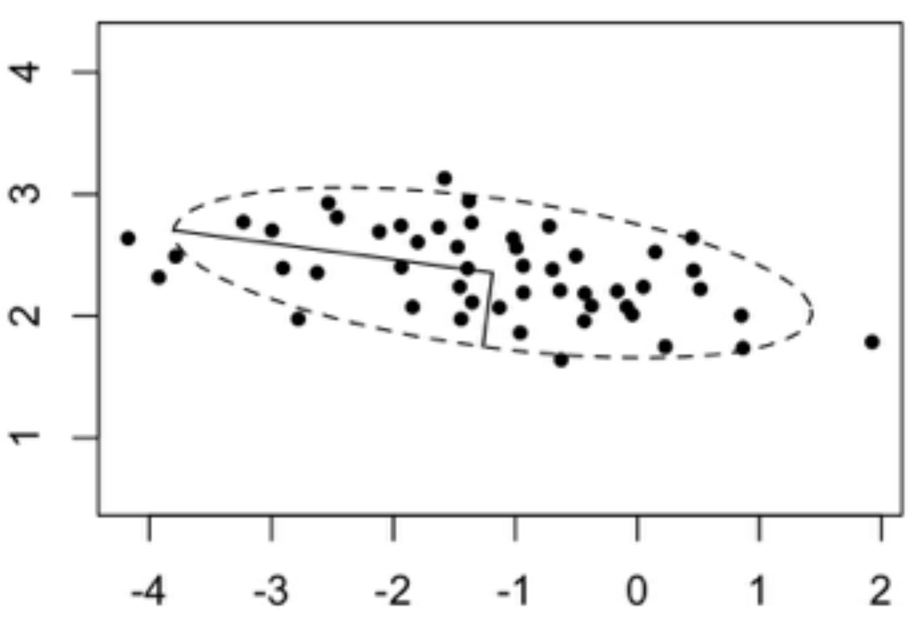
分散共分散行列　　**相関行列**

$R = P\Lambda P'$ (固有値分解)

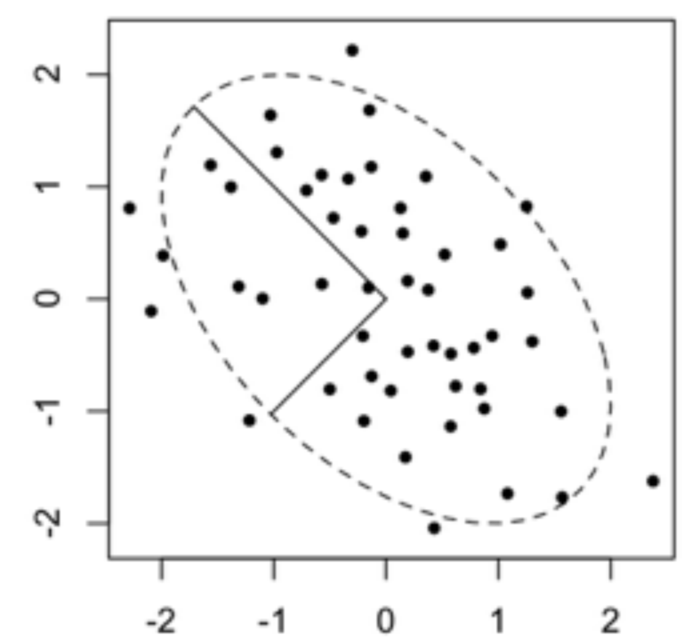
$\Sigma = DP\Lambda P'D' = (DP)\Lambda(DP)'$

固有構造は変化する

$(DP)(DP)' \neq I$



各々の変数を標準化
(× 多変量の標準化)



主成分分析

❶教科書の流儀　　変量を各々標準化してから解析

⇒ **相関行列の固有値問題に帰着**

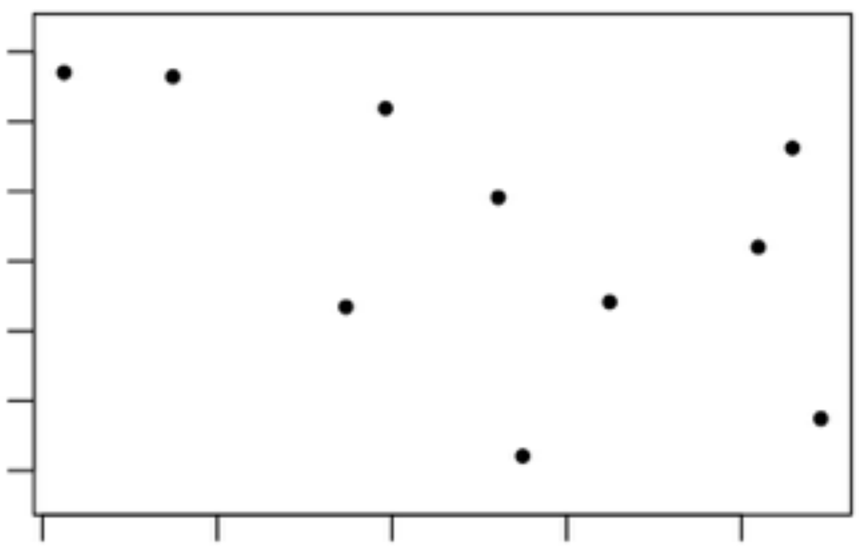
❷他の流儀　　標準化はしないで解析

⇒ **分散共分散行列の固有値問題に帰着**

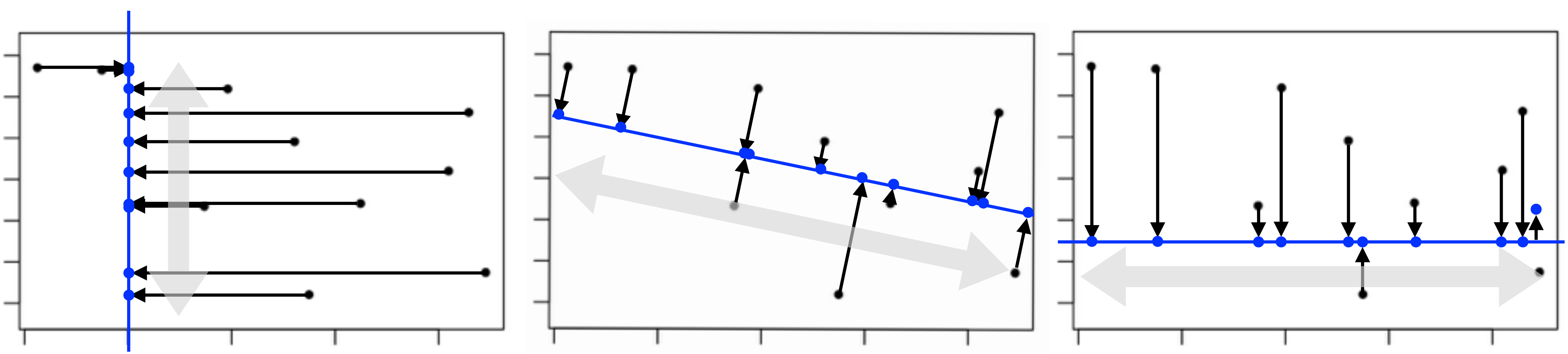
- 結果は異なるが、計算の要領は同じ
- 変量で単位が違う(身長cm,体重kg)場合で主成分を解釈するなら❶
- データそのままの低次元表現を求めるなら❷

主成分分析

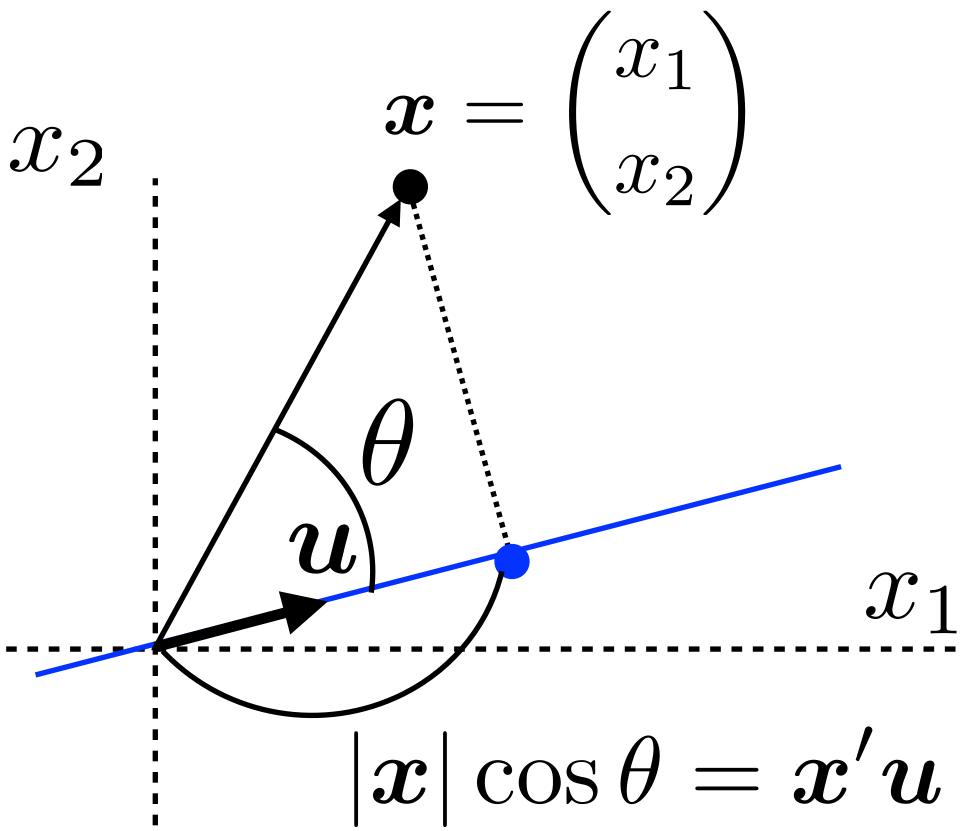
多次元データの持っている情報をできるだけ損なわずにより低次元のデータに情報を縮約する方法



射影すると分散が最大になる軸を探す



主成分分析



$$|\boldsymbol{x}| \cos\theta = \boldsymbol{x}'\boldsymbol{u}$$

$$\|\boldsymbol{u}\|^2 = 1$$

$\boldsymbol{u}$ 方向の軸に射影した際の分散

$$\begin{aligned} \mathrm{var}(\boldsymbol{u}'\boldsymbol{x}) &= \mathbb{E}\left\{(\boldsymbol{u}'\boldsymbol{x} - \boldsymbol{u}'\boldsymbol{\mu})^2\right\} \\ &= \boldsymbol{u}'\mathbb{E}\left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'\right\}\boldsymbol{u} \\ &= \boldsymbol{u}'\Sigma\boldsymbol{u} \end{aligned}$$

最大にするにはLagrange未定乗数法より

$$\max_{\boldsymbol{u}}\{\mathrm{var}(\boldsymbol{u}'\boldsymbol{x}) \mid \|\boldsymbol{u}\|^2 = 1\}$$

$$\Leftrightarrow\ \max_{\boldsymbol{u}}\{\boldsymbol{u}'\Sigma\boldsymbol{u} - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1)\}$$

$$\Leftrightarrow\ \Sigma\boldsymbol{u} = \lambda\boldsymbol{u}$$ (微分=0より)

なので、**最大の固有値の固有ベクトル**が解

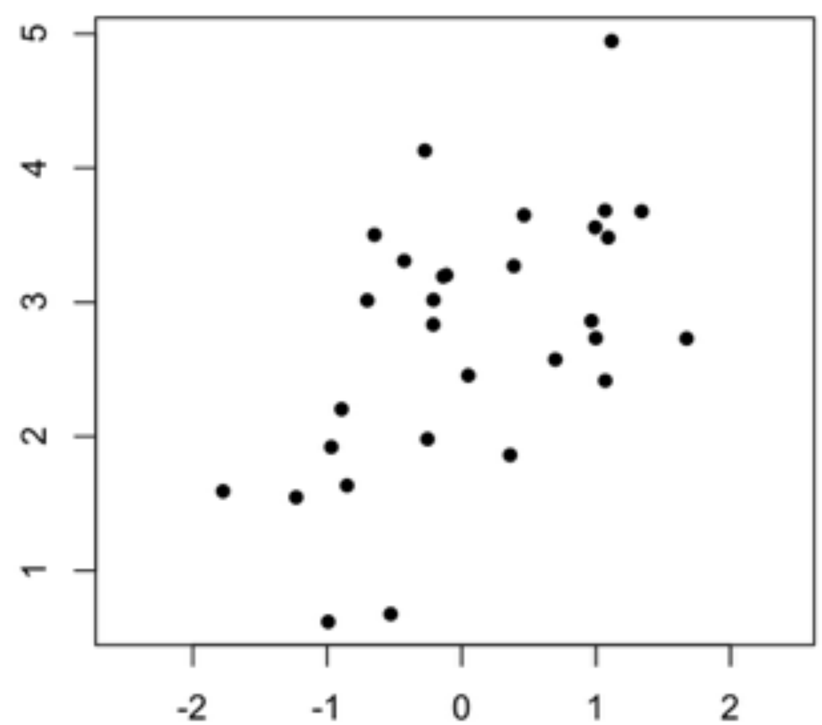
分散の最大値は**最大の固有値**

蛇足: 線形回帰 vs 直交回帰

最小二乗回帰

Ordinary Least Squares (OLS)

yのみ誤差混入

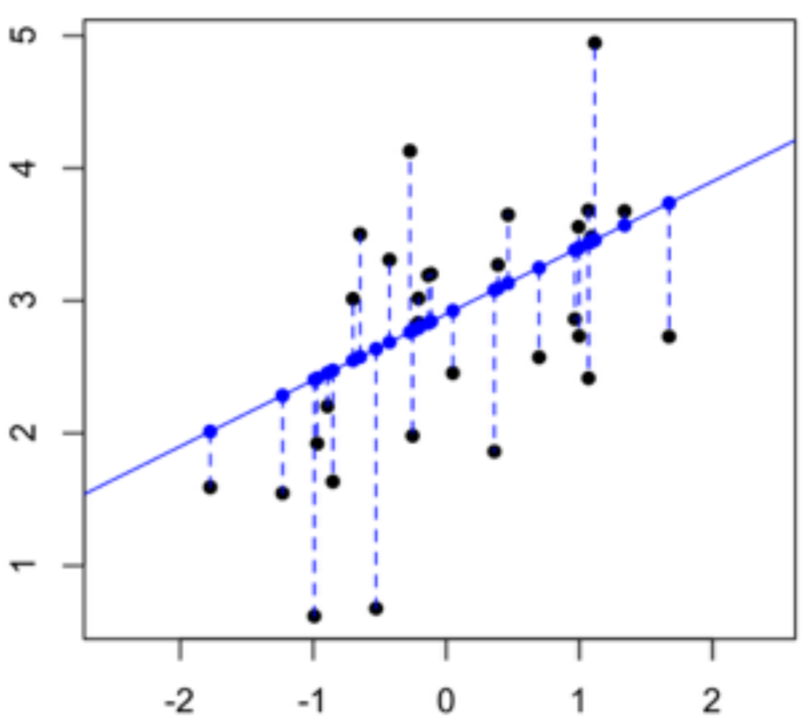


直交回帰

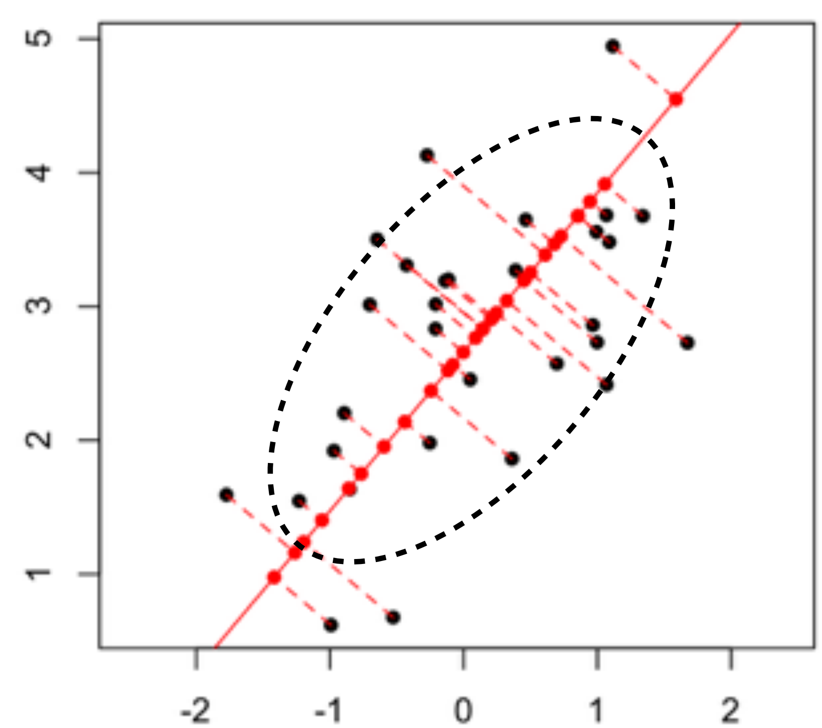
Total least squares (TLS)

x,yともに誤差混入

最小二乗推定量



第1主成分方向

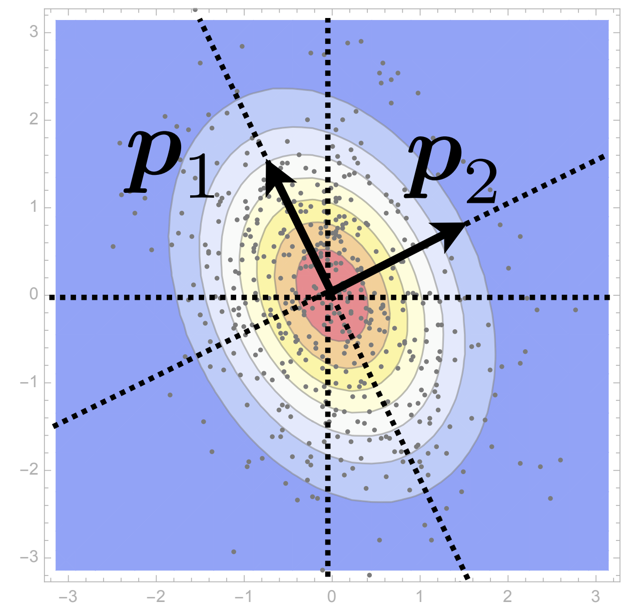


第k主成分とΣの固有値・固有ベクトル

第二主成分は第一主成分に直交する方向から探索

第三主成分は第一主成分,第二主成分に直交する方向から探索

$$\vdots$$



結局、**第k主成分**は**k番目の固有値の固有ベクトル**

$$z_j = \boldsymbol{p}_k \boldsymbol{x}_j \qquad \text{標本 } \boldsymbol{x}_j \text{ の第 } k \text{ 主成分}$$

標本の第k主成分

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}'_k \boldsymbol{x}_1 \\ \boldsymbol{p}'_k \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{p}'_k \boldsymbol{x}_n \end{bmatrix}$$

標本点 x の上位k主成分表示

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 & \cdots & \boldsymbol{p}_k \\ | & | & & | \end{bmatrix}' \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

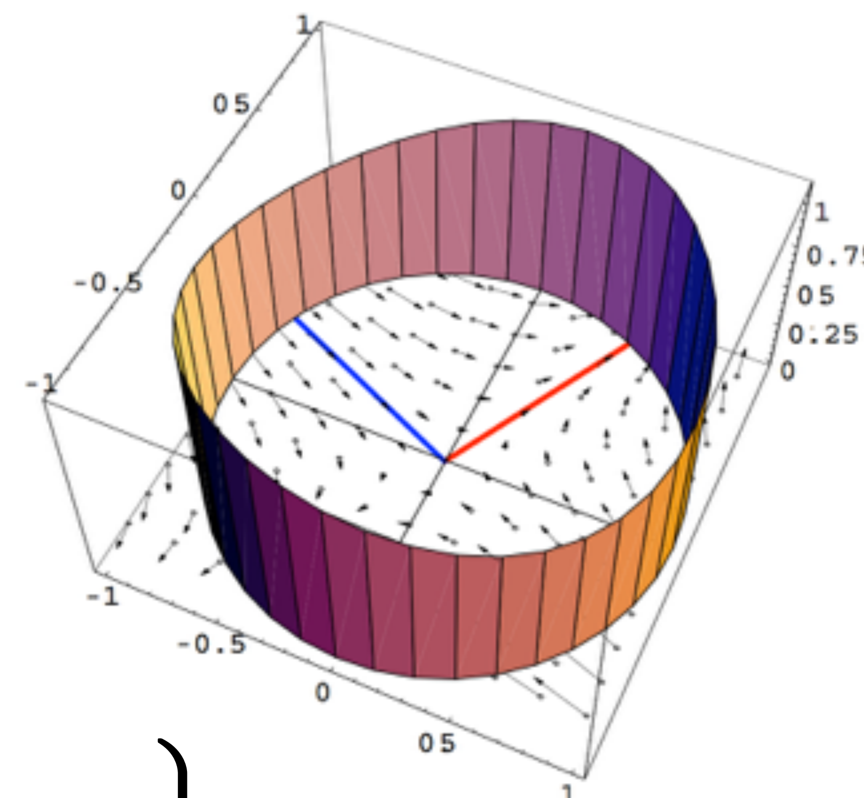
Rayleigh-Ritzの定理と二次形式の最大・最小

対称行列 A に対して、Rayleigh商 $\dfrac{\boldsymbol{u}' A \boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}}$

固有値　$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$

正規直交固有ベクトル　$\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$

$\|\boldsymbol{u}\|^2 = 1$ 上での $\boldsymbol{u}' A \boldsymbol{u}$ の値



$$\lambda_1 = \max_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\boldsymbol{u}' A \boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}}$$

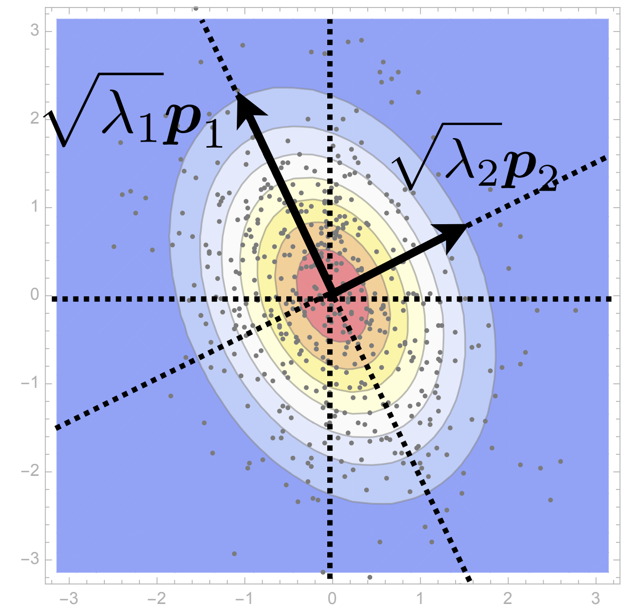
$$\lambda_n = \min_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\boldsymbol{u}' A \boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}}$$

$$\lambda_k = \max_{\boldsymbol{u} \neq \boldsymbol{0}} \left\{ \frac{\boldsymbol{u}' A \boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}} \;\middle|\; \boldsymbol{u} \perp \boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{k-1} \right\}$$

主成分と分散共分散行列の固有構造

第二主成分は第一主成分に直交する方向から探索

結局、第k主成分はk番目の固有値の固有ベクトル



$$\boldsymbol{v}_1 = \sqrt{\lambda_1}\boldsymbol{p}_1 \qquad \lambda_i \quad \Sigma \text{ の固有値}$$

$$\boldsymbol{v}_2 = \sqrt{\lambda_2}\boldsymbol{p}_2 \qquad \boldsymbol{p}_i \quad \Sigma \text{ の固有ベクトル}$$

- $\boldsymbol{v}_1$ や $\boldsymbol{v}_2$ を**主成分負荷量**と呼ぶ
- $\boldsymbol{p}_i$ 方向の軸に射影した際の分散の値は $\lambda_i \Leftrightarrow \mathrm{var}(p_i'\boldsymbol{x}) = \lambda_i$
- $\lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 = \mathrm{tr}(\Sigma)$
- i 主成分の分散の全体に占める割合 $\lambda_i / \sum_{i=1}^{p} \lambda_i$ を**寄与率**と呼ぶ
- 各々の標本 $\boldsymbol{x}_j$ の主成分軸への射影 $\boldsymbol{p}_i'\boldsymbol{x}_j$ を**主成分得点**と呼ぶ

因子負荷量 (Factor Loading)

以下の関係より

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & & & \\ p_{p1} & p_{p2} & \cdots & p_{pp} \end{bmatrix}' \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$\sigma_{y_i}^2 = \mathrm{var}(\boldsymbol{p}_i, \boldsymbol{x}) = \lambda_i$$

$$D_y = \mathrm{diag}(1/\sigma_{y_i}) = \Lambda^{-1/2}$$

$$D_x = \mathrm{diag}(1/\sigma_{x_i})$$

因子負荷量：元の各々の変量 x_i と主成分 y_j との相関係数

$$R = \mathbb{E}\{D_x(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{y} - \mathbb{E}\{\boldsymbol{y}\})'D_y'\}$$

$$r(x_i, y_j) = \frac{\sqrt{\lambda_j} p_{ij}}{\sigma_{x_i}}$$

$$= D_x \mathbb{E}\{(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'\} P D_y = D_x P \Lambda P' P \Lambda^{-1/2} = D_x P \Lambda^{1/2}$$

$$\begin{bmatrix} r(x_1, y_1) & r(x_1, y_2) & \cdots & r(x_1, y_p) \\ r(x_2, y_1) & r(x_1, y_2) & \cdots & r(x_2, y_p) \\ \vdots & & & \\ r(x_p, y_1) & r(x_1, y_2) & \cdots & r(x_p, y_p) \end{bmatrix} = \begin{bmatrix} 1/\sigma_{x_1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{x_2} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1/\sigma_{x_p} \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & & & \\ p_{p1} & p_{p2} & \cdots & p_{pp} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix}$$

※各変量 x_i が元から標準化してあれば主成分負荷量に等しい

主成分分析：教科書の具体例

表 9.1　試験の成績のデータ

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

表 9.2　因子負荷量

	国語 x_1	英語 x_2	数学 x_3	理科 x_4
z_1	$r_{z_1x_1} = 0.804$	$r_{z_1x_2} = 0.842$	$r_{z_1x_3} = 0.838$	$r_{z_1x_4} = 0.814$
z_2	$r_{z_2x_1} = 0.583$	$r_{z_2x_2} = 0.524$	$r_{z_2x_3} = -0.531$	$r_{z_2x_4} = -0.570$
z_3	$r_{z_3x_1} = -0.114$	$r_{z_3x_2} = 0.123$	$r_{z_3x_3} = -0.115$	$r_{z_3x_4} = 0.104$
z_4	$r_{z_4x_1} = 0.035$	$r_{z_4x_2} = -0.034$	$r_{z_4x_3} = -0.036$	$r_{z_4x_4} = 0.038$

表 9.3　標準化した値（$u_1 \sim u_4$）と主成分得点（z_1, z_2）

生徒 No.	標準化した値				主成分得点	
	u_1	u_2	u_3	u_4	z_1	z_2
1	0.956	0.694	−0.015	−0.033	0.796	0.857
2	0.224	0.509	0.552	0.856	1.072	−0.348
3	−1.190	−0.972	−1.459	−1.367	−2.491	0.319
4	−0.215	−0.278	1.582	1.467	1.280	−1.763
5	1.444	1.528	−0.325	−0.311	1.166	1.802
6	−1.337	−1.435	−1.149	−1.033	−2.477	−0.299
7	−0.800	−0.509	−0.170	0.189	−0.643	−0.679
8	0.566	0.093	−0.789	−1.200	−0.669	1.341
9	−0.751	−0.926	0.448	0.189	−0.518	−1.148
10	1.102	1.296	1.325	1.244	2.485	−0.084

第 1 主成分の寄与率 $= \lambda_1/p = 0.680$

第 2 主成分の寄与率 $= \lambda_2/p = 0.306$

第 3 主成分の寄与率 $= \lambda_3/p = 0.013$

第 4 主成分の寄与率 $= \lambda_4/p = 0.001$

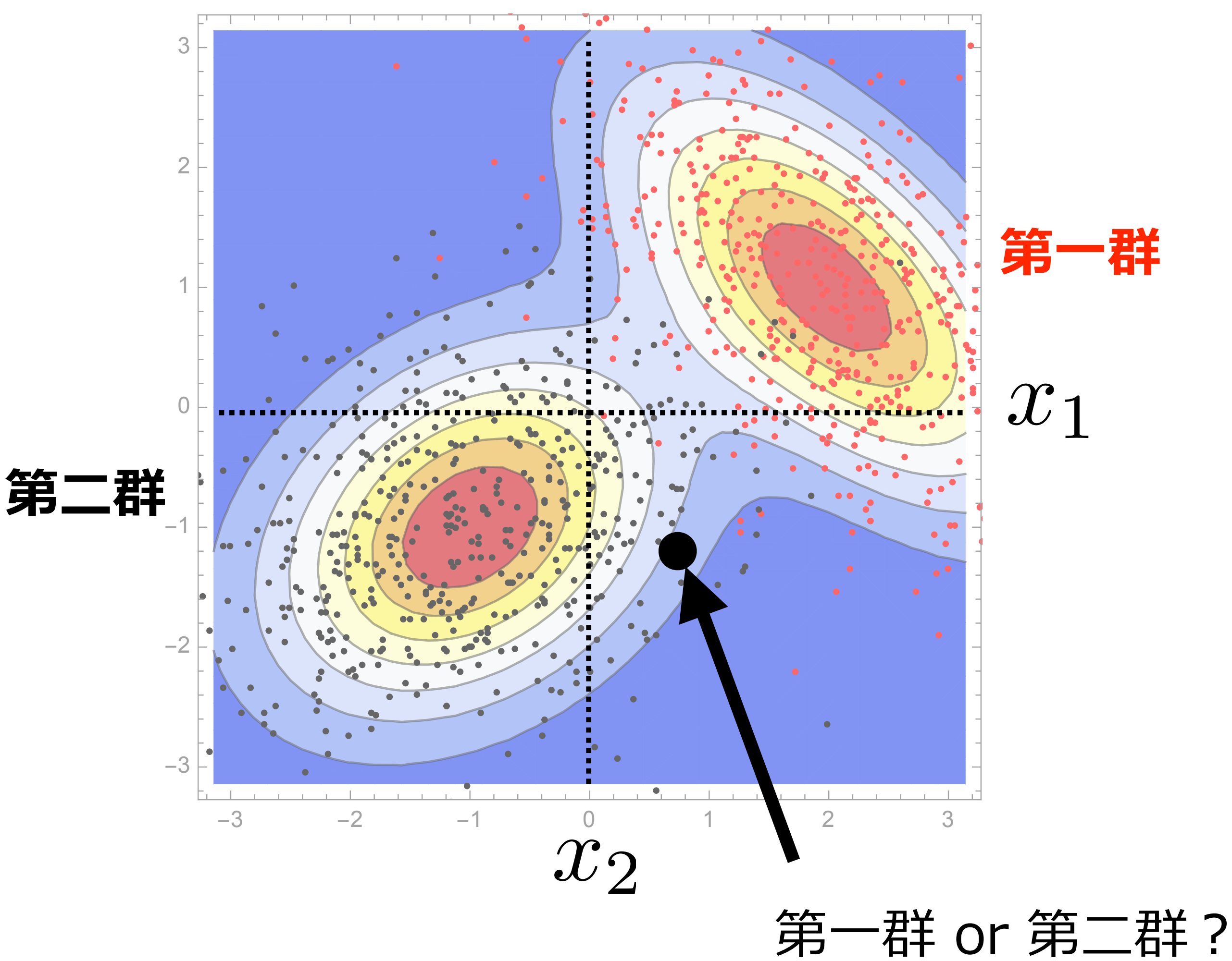
$$z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$

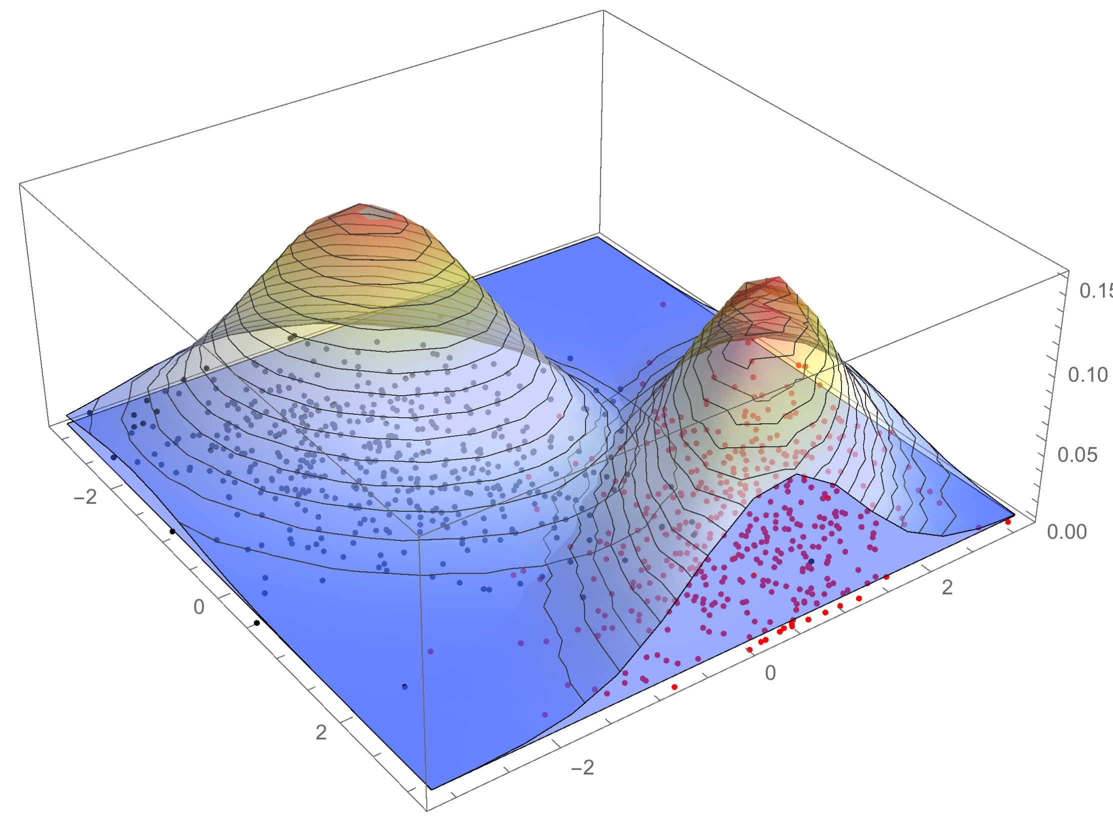
$$z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$

$$z_3 = -0.499u_1 + 0.539u_2 - 0.504u_3 + 0.455u_4$$

$$z_4 = 0.485u_1 - 0.474u_2 - 0.506u_3 + 0.533u_4$$

判別分析: Discriminant Analysis

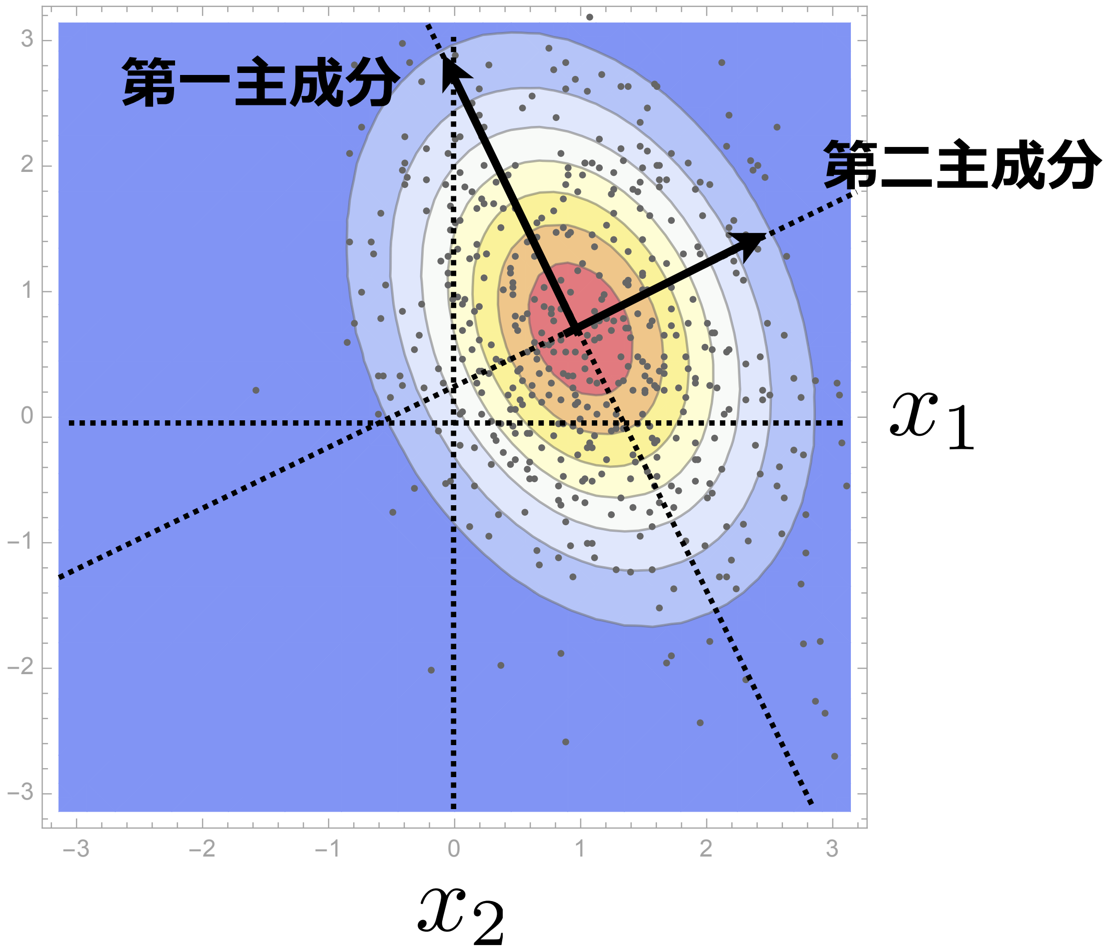


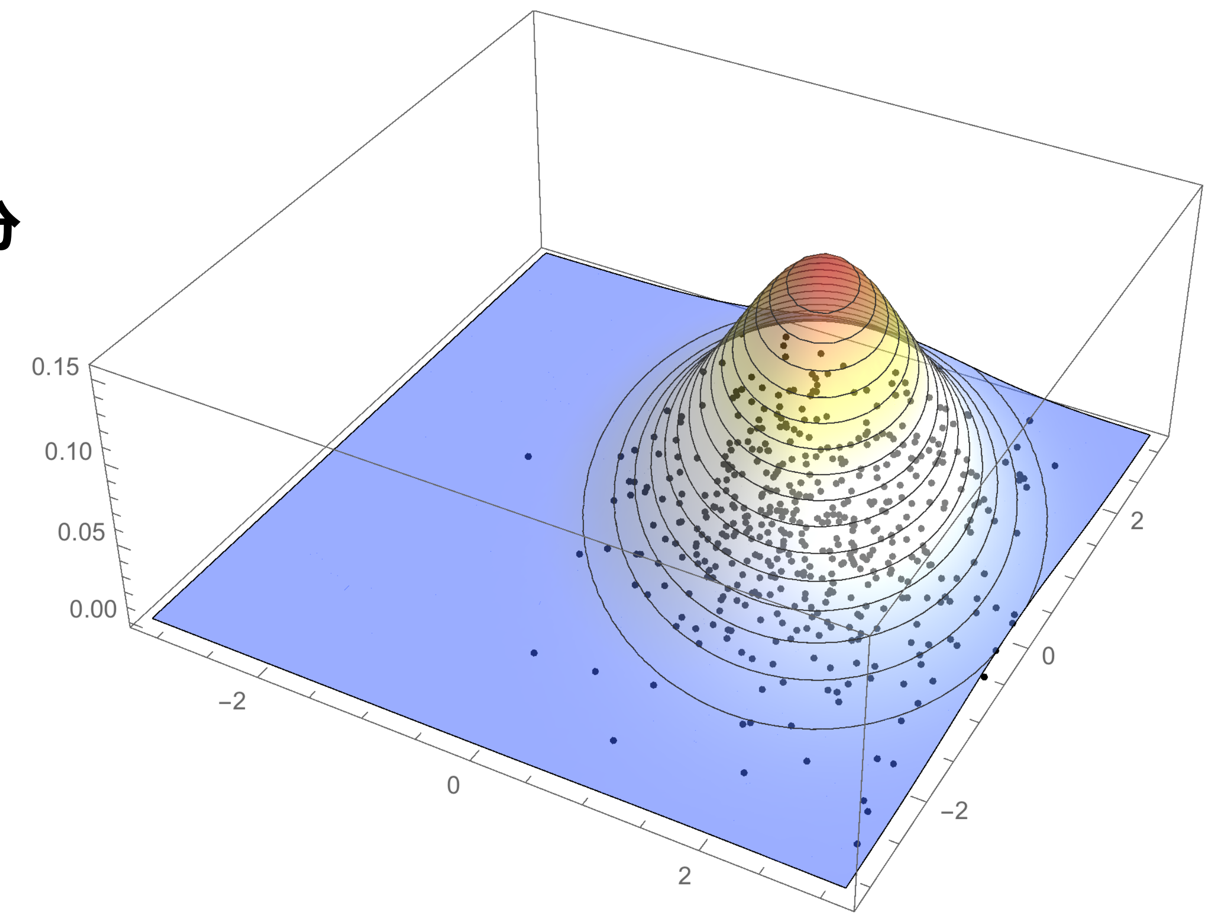


知りたいこと
(教科書 p.4)

- 検査値 (x_1, x_2) から第 1 群か第 2 群か判別できるか？
- 判別できるとすればその精度はどれくらいか？
- 例えば $x_1 = 0.7, x_2 = -1.1$ ならどう判別されるか？

主成分分析: Principal Component Analysis (PCA)





知りたいこと (教科書 p.6)

- 主成分により低い次元でデータを解釈できないか？
- それぞれの主成分の説明力はどれくらいか？
- 主成分により(x_1, x_2)の特徴付け・分類ができるか？

1.6　主成分分析とは

表 1.7 は，4 教科の試験の成績である．すべて量的変数と考える．

表 1.7　試験の成績のデータ

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

この（4 次元）データに基づいて知りたいことは次の通りである．

(1) 主成分の構成により低い次元でデータを解釈できないか．

(2) それぞれの主成分の説明力はどれくらいか．

(3) 科目や生徒の特徴付けおよび分類をどのようにできるか．

表 1.7 のデータを主成分分析で解析することにより次のことがわかる.

(1) 主要な主成分として次の第 1 主成分 z_1 と第 2 主成分 z_2 を得る.

$$z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$

$$z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$

ここで, u_j は変数 x_j を標準化したものである.

(2) z_1 の寄与率は 0.680, z_2 の寄与率は 0.306 である. 第 2 主成分までの累積寄与率は $0.680 + 0.306 = 0.986$ である.

(3) 係数の値より, z_1 は「総合的学力」, z_2 は「文系・理系の学力の違い」を表すと解釈できる. 各生徒に対して z_1 と z_2 を計算することにより, 生徒の特徴付けや分類ができる.