主成分分析
Principal Component Analysis

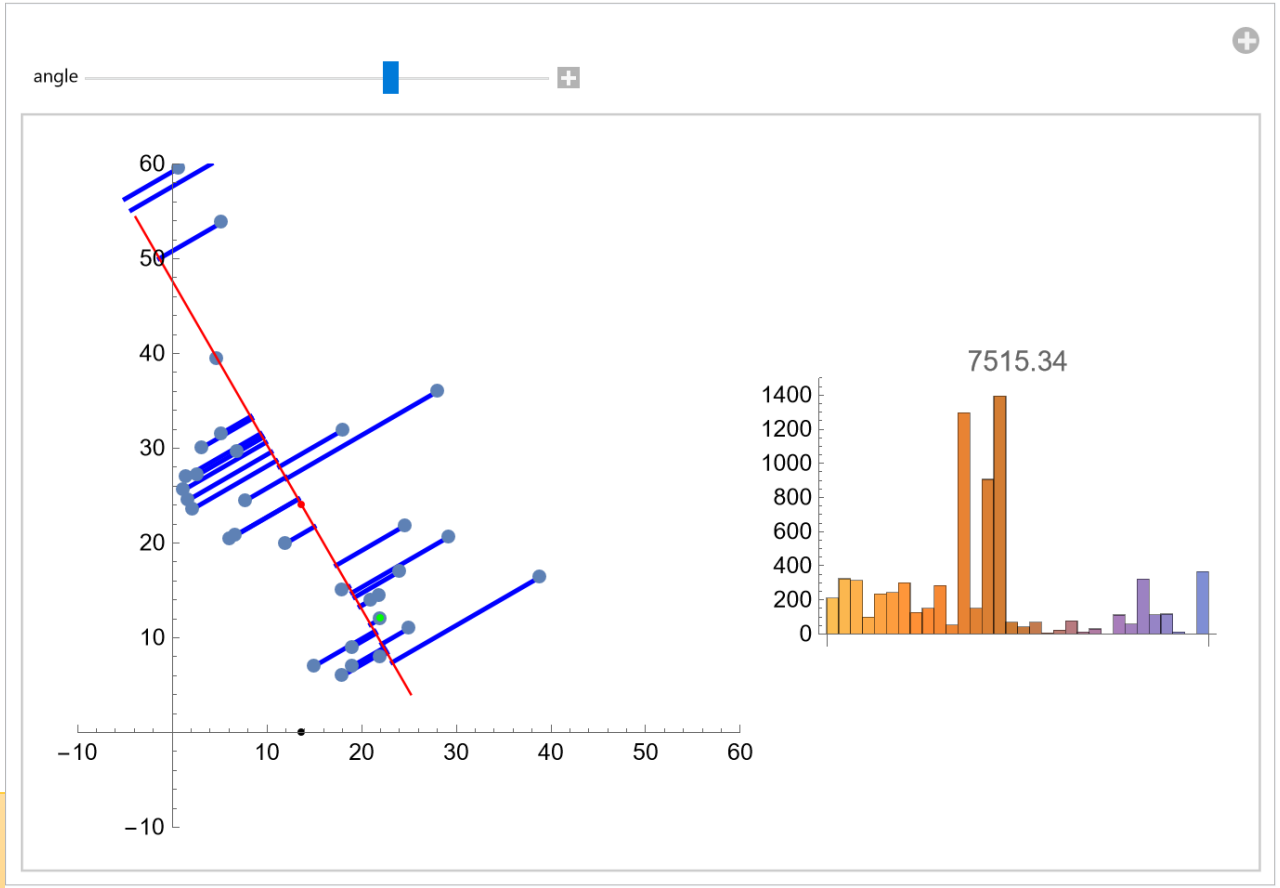
2024年2月10日

学習院大学経済学部教授

白田　由香利

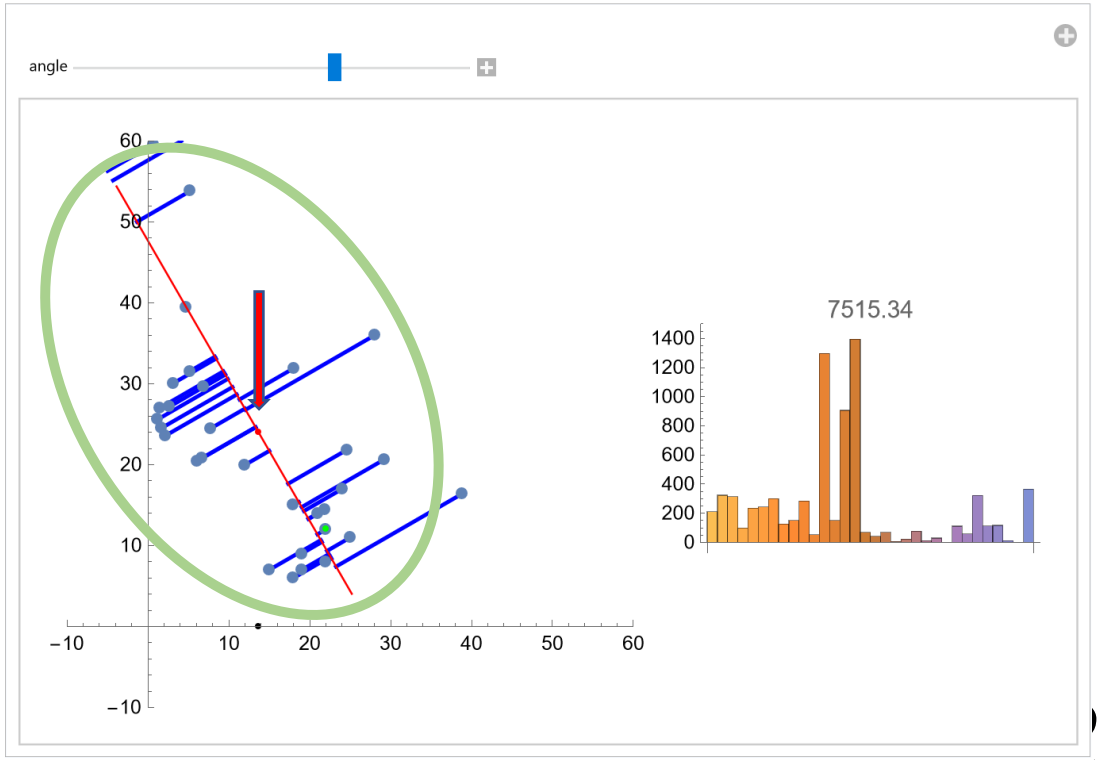
PCA:　次元圧縮技術のひとつ

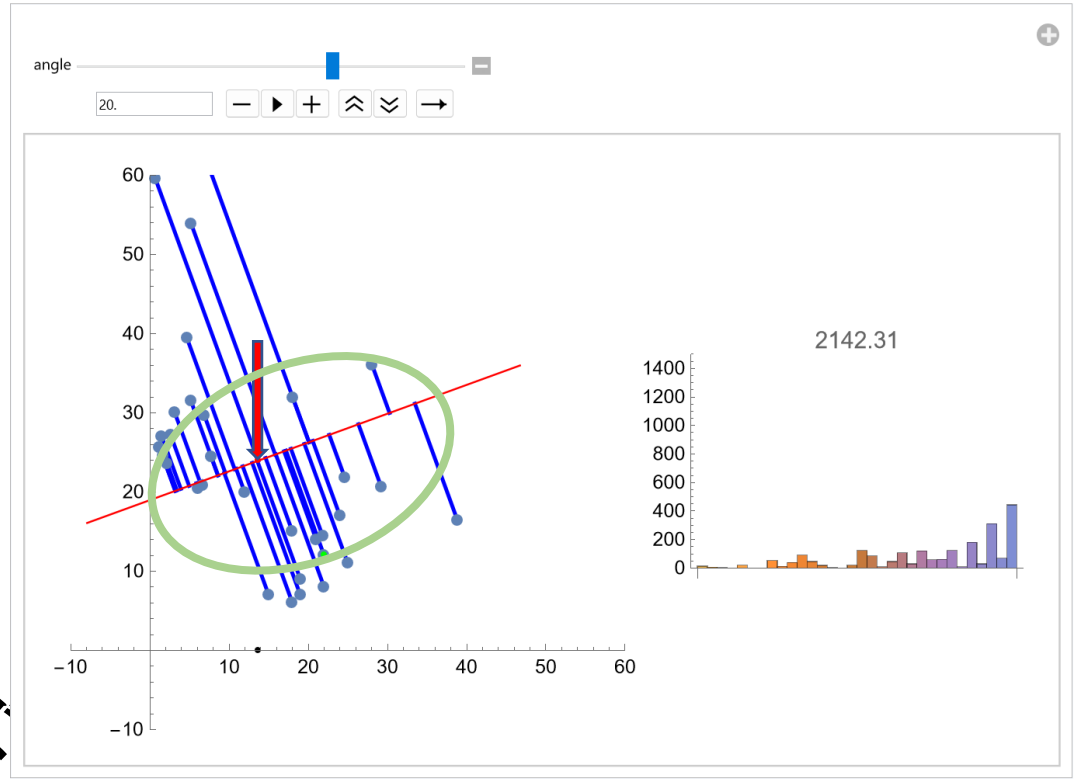
- 2次元データ
  →1次元に圧縮する例
- (x,y)のデータが20個以上
- どちらの方向に分布が伸びているか→z軸
- 垂線の足と平均の距離
  →z値



グラフィクス教材
www-cc.gakushuin.ac.jp/~20010570/VDStat/

PCA:　次元圧縮技術のひとつ
z値の分散を最大化する方向が主成分軸



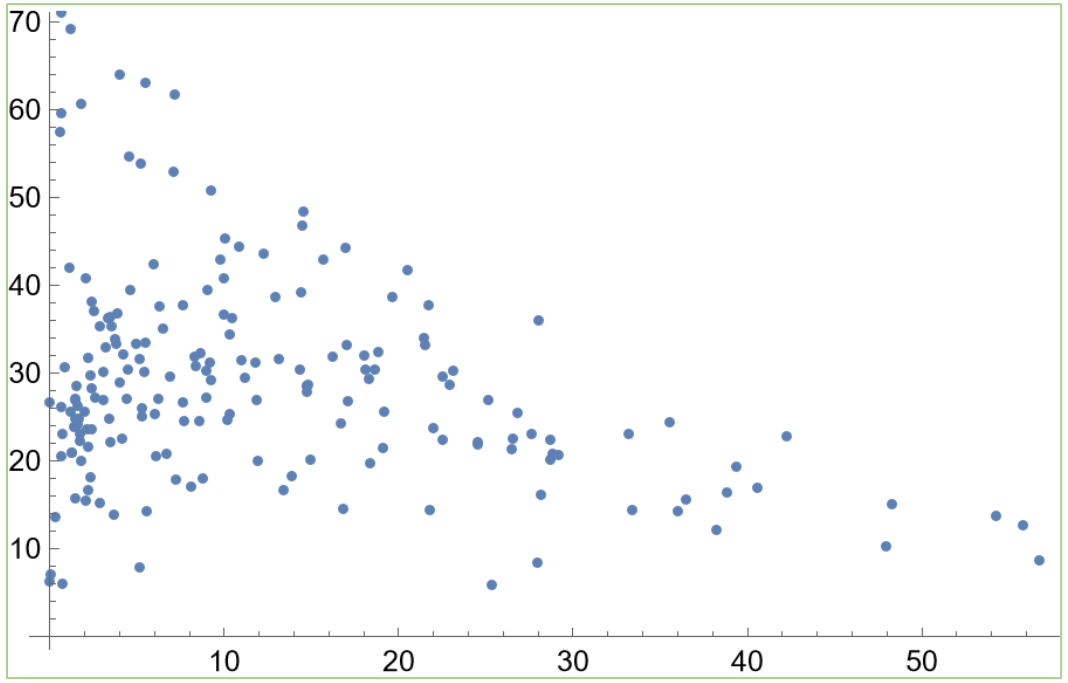


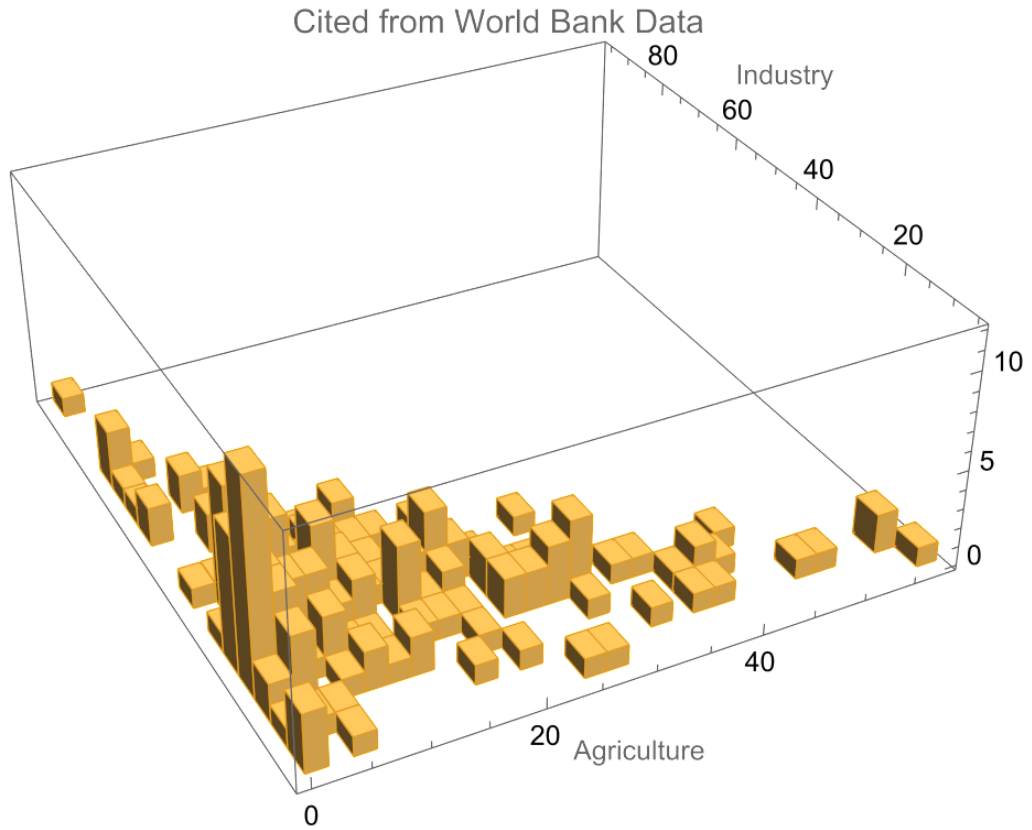
垂線の足と重心との距離がそのデータの z 値

Z 値の分散のようす（右図棒グラフ）をみて，総和が最大の方向で止める.

PCAは分散共分散行列から計算可能（ぐるぐる回すよりも正確）

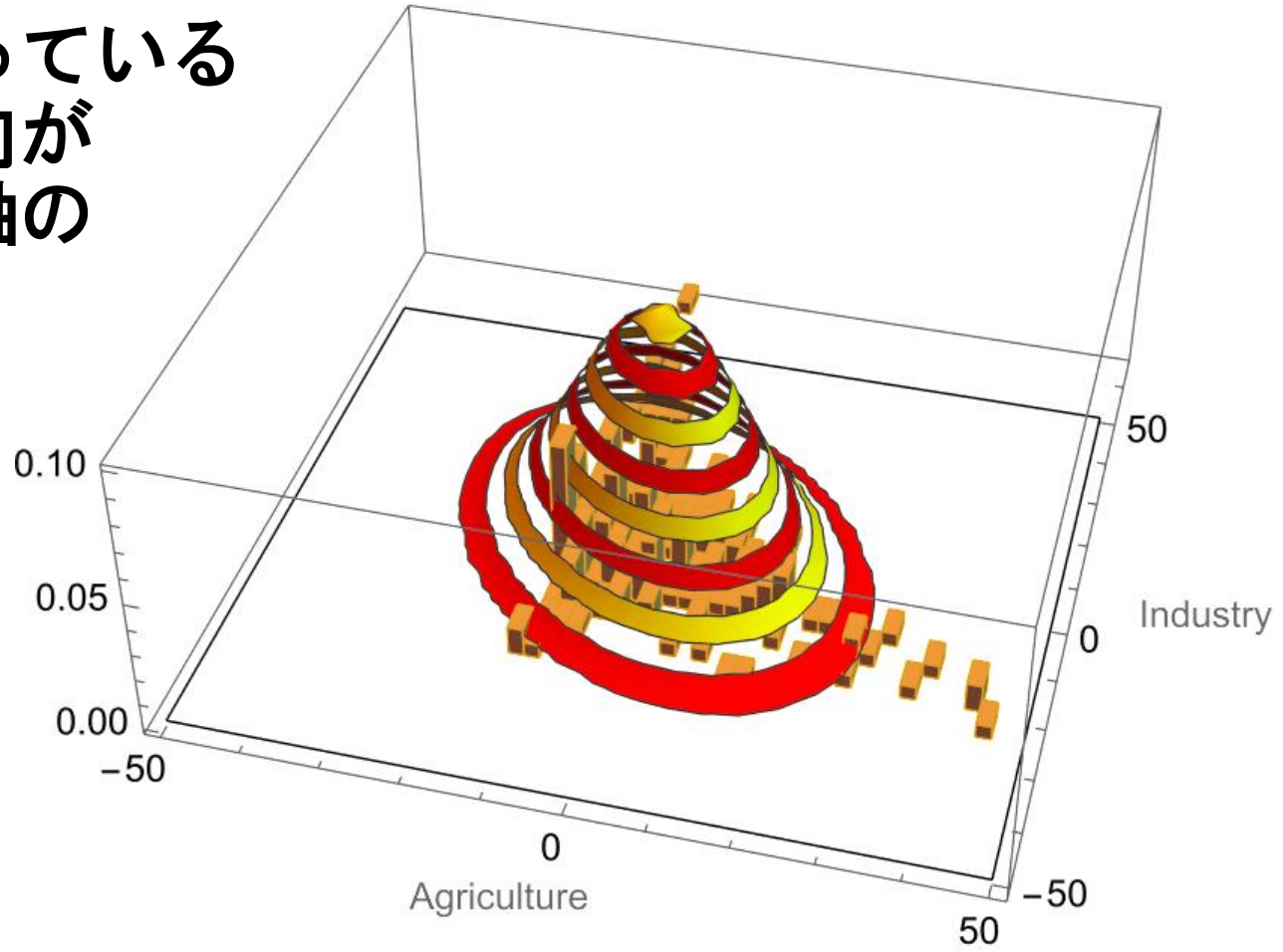
- 2次元データ，3次元ヒストグラムでは右図のようになる

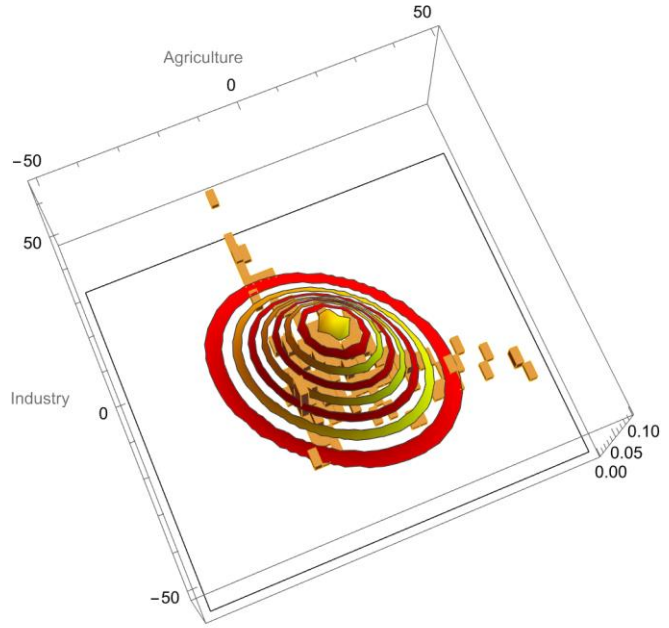




2次元ガウス分布形状のお帽子を被せる

- 引っ張っている力の方向が主成分軸の方向

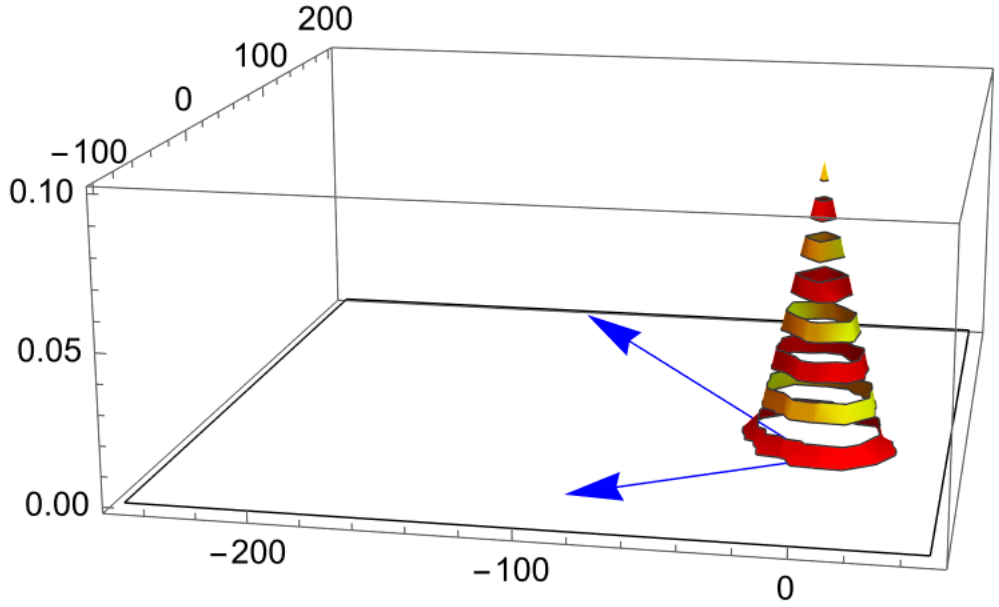


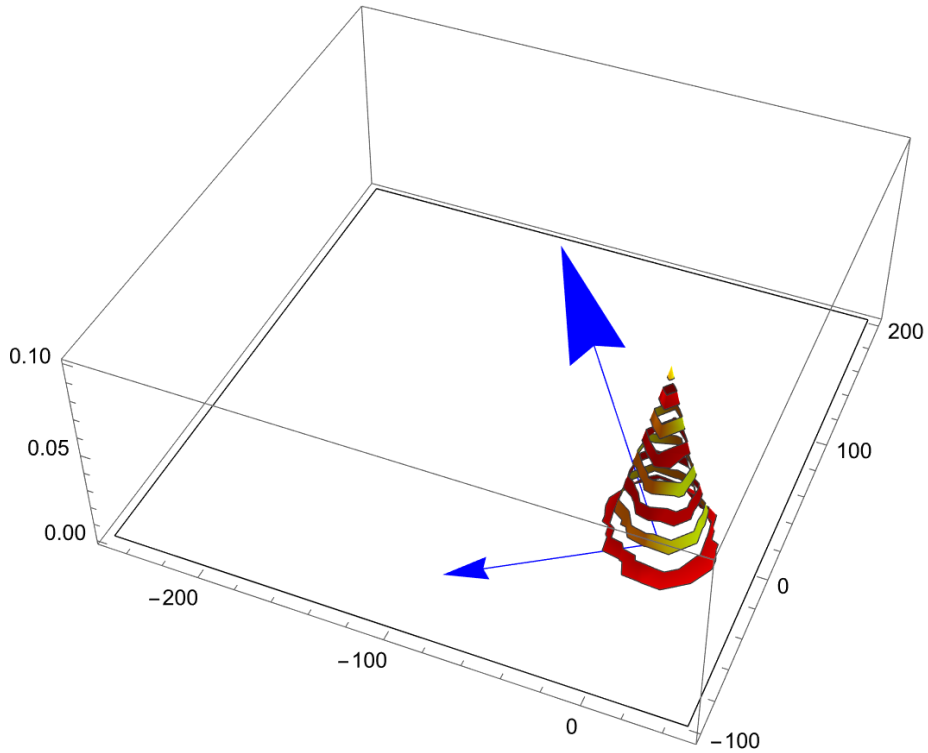


2次元ガウス分布形状のお帽子を被せる

- データから**分散共分散行列**を求める.
  その行列の固有ベクトルと固有値を求める
- 第1主成分軸と第2主成分軸
- 軸の方向→固有ベクトルで求まる
- 力の強さ→固有値で求まる

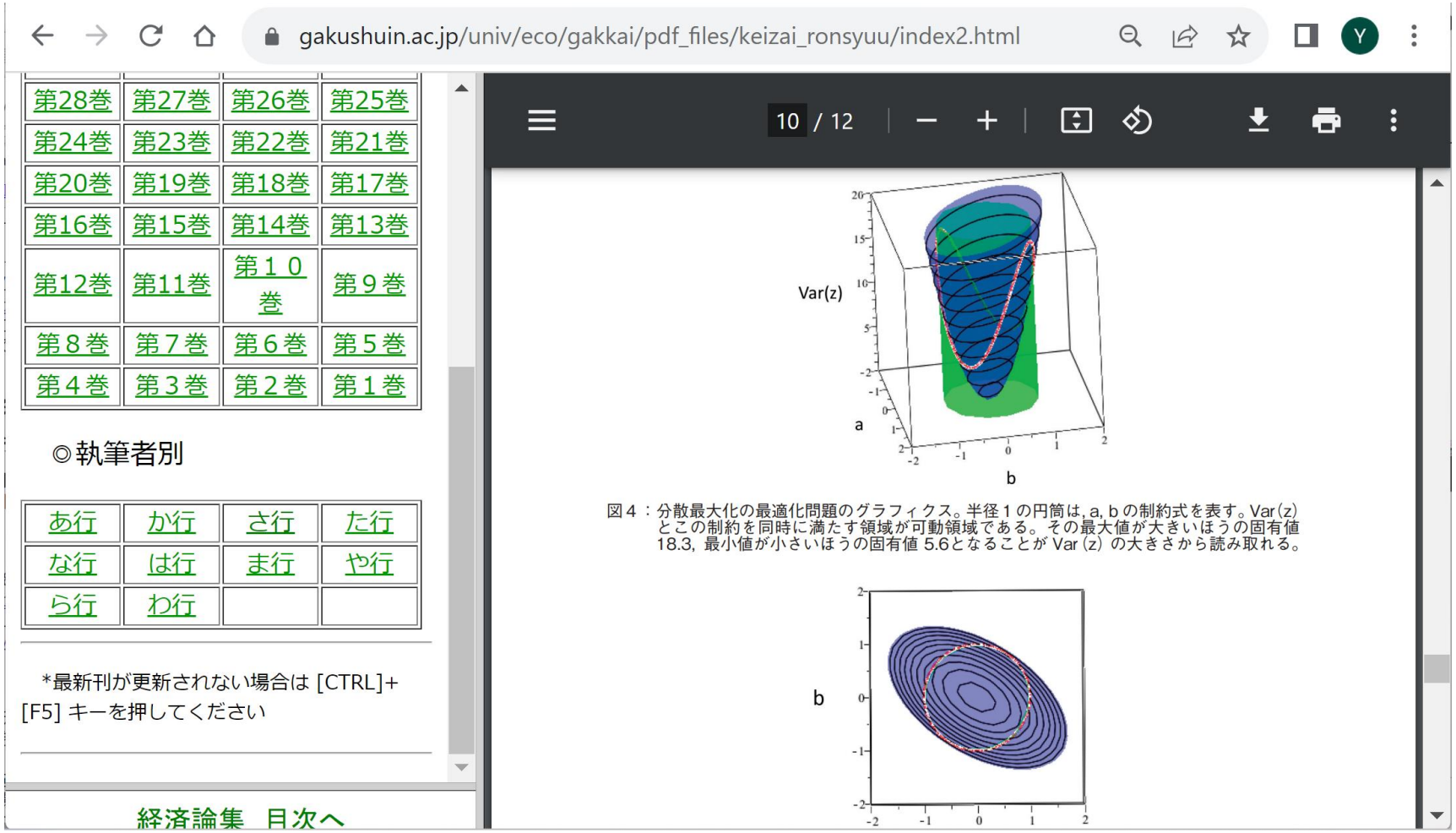
$$\begin{bmatrix} 146.966 & -57.7582 \\ -57.7582 & 194.497 \end{bmatrix}$$





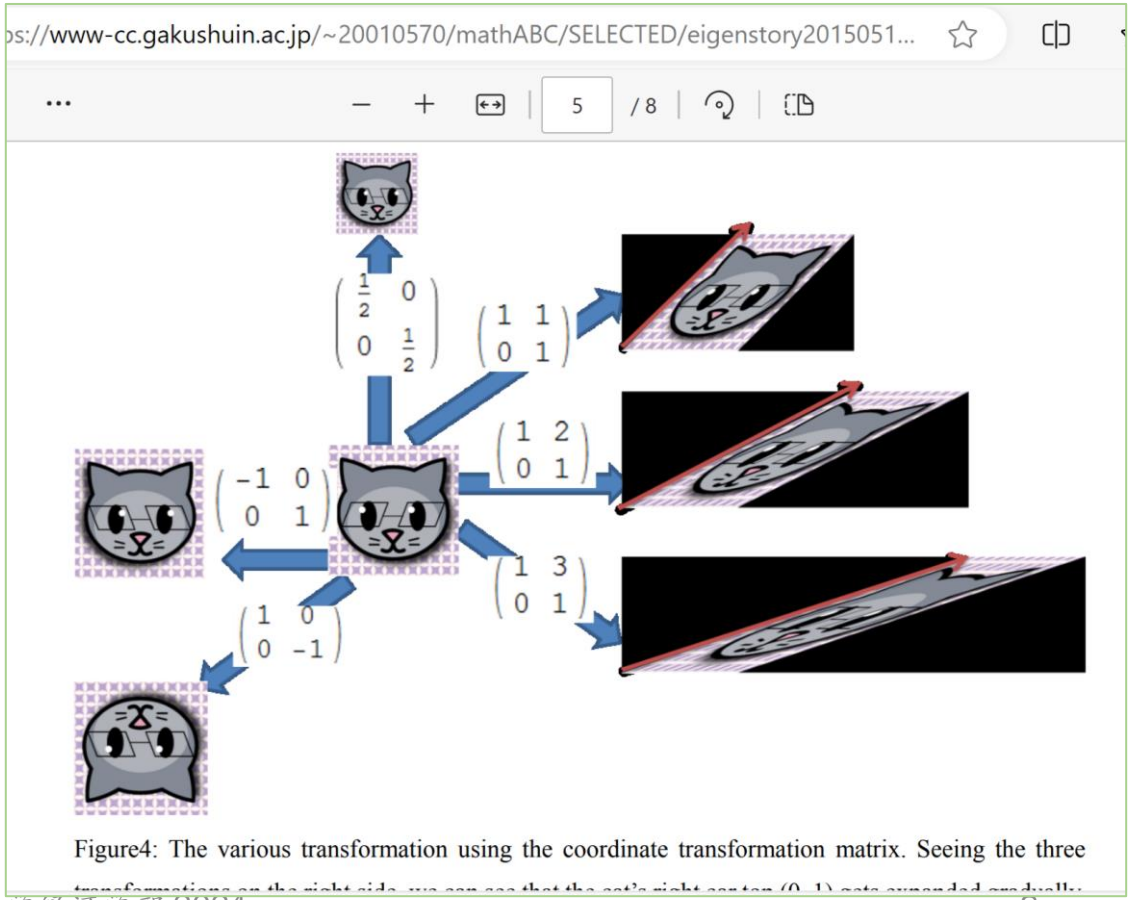
分散最大化問題が固有方程式の問題になるのか．その鮮やかな変換は以下の論文に説明あります．

- 白田　由香利：固有値の概念の教授法 ―経営学科に適した線型代数の教授法，『学習院大学経済論集』第50巻　第１号（2013年４月）



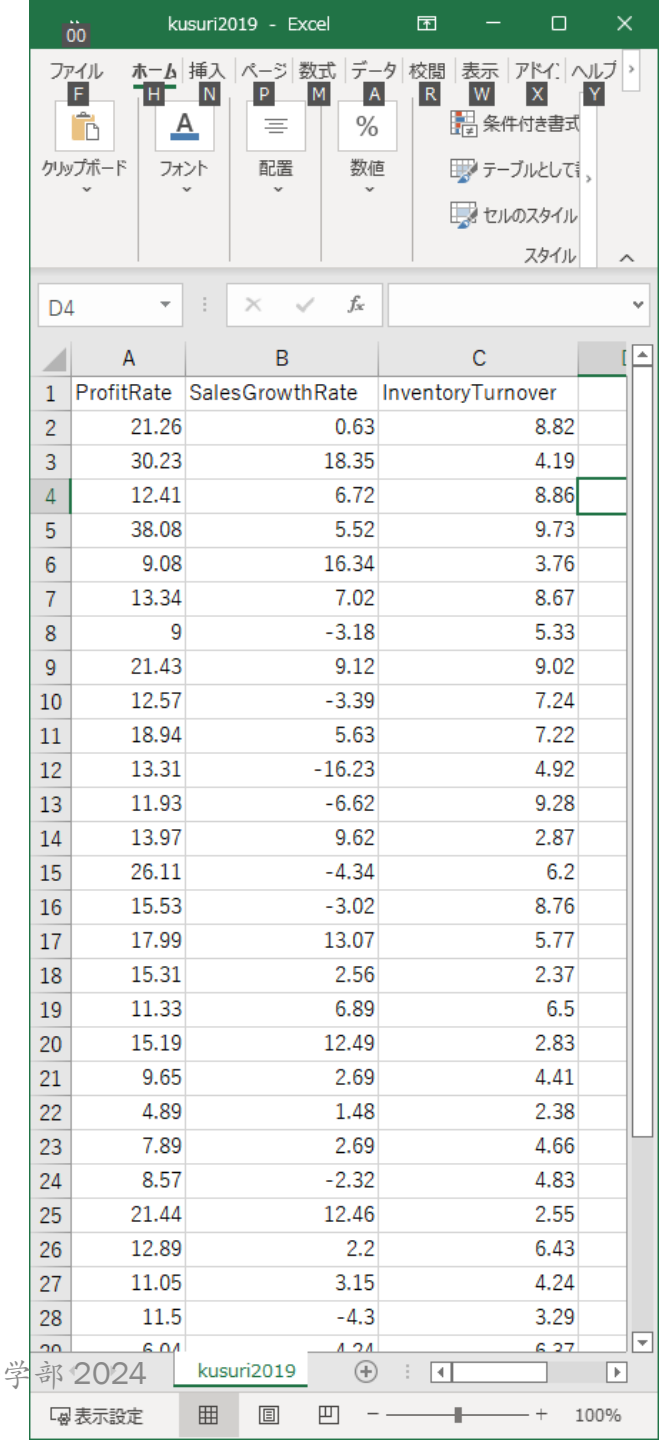
固有値って何？
行列の固有値の意味を理解するためには

- Y. Shirota: “A fable story for understanding eigenvalues titled “Enlargement Factors of the Magnification Machine are Eigenvalues”(クリックしてください)
- 行列を拡大マシンだと解釈すると，拡大率が固有値
- 固有値は，行列の基底を変えても変わりません.
- 異なる基底の犬の国と猫の国の間で，拡大マシンを輸出しようとすると変換が必要
- インドネシア語バージョンもあります.



製薬会社2019年度
経営指標データ　34社

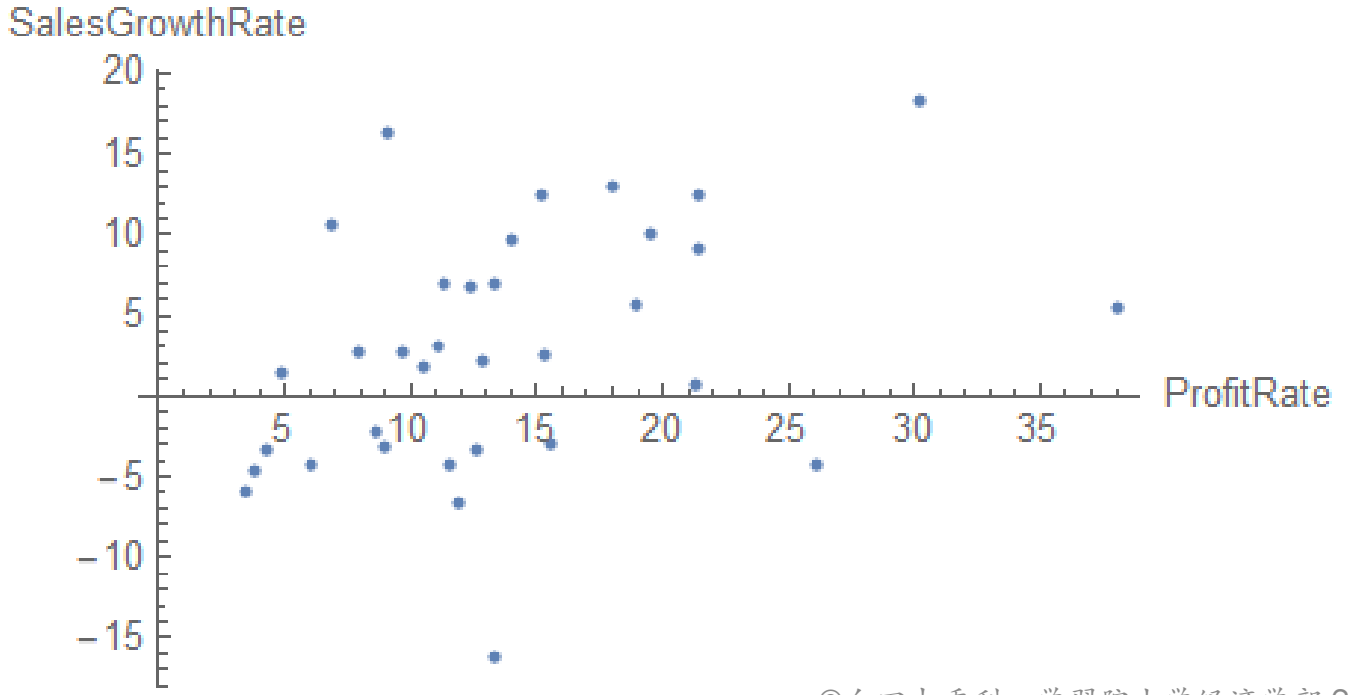
- 売上高営業利益利益率
- 売上高成長率
- 在庫回転率



	A	B	C
1	ProfitRate	SalesGrowthRate	InventoryTurnover
2	21.26	0.63	8.82
3	30.23	18.35	4.19
4	12.41	6.72	8.86
5	38.08	5.52	9.73
6	9.08	16.34	3.76
7	13.34	7.02	8.67
8	9	-3.18	5.33
9	21.43	9.12	9.02
10	12.57	-3.39	7.24
11	18.94	5.63	7.22
12	13.31	-16.23	4.92
13	11.93	-6.62	9.28
14	13.97	9.62	2.87
15	26.11	-4.34	6.2
16	15.53	-3.02	8.76
17	17.99	13.07	5.77
18	15.31	2.56	2.37
19	11.33	6.89	6.5
20	15.19	12.49	2.83
21	9.65	2.69	4.41
22	4.89	1.48	2.38
23	7.89	2.69	4.66
24	8.57	-2.32	4.83
25	21.44	12.46	2.55
26	12.89	2.2	6.43
27	11.05	3.15	4.24
28	11.5	-4.3	3.29

利益率～売上高成長率　散布図

- どちらの方向に分布が偏っていますか？
- それを教えてくれるのがPCA
  （**主成分分析**, Principal Component Analysis)



PCAは、
分散共分散行列の固有値と固有ベクトル

- EXCELでやってみましょう。以下では不偏分散を使って下さい
- 問1：与えられたデータのProfitRateの分散を求めよ
- 問2：与えられたデータのSalesGrowthRateの分散を求めよ
- 問3：与えられたデータのProfitRate とSalesGrowthRateの共分散を求めよ

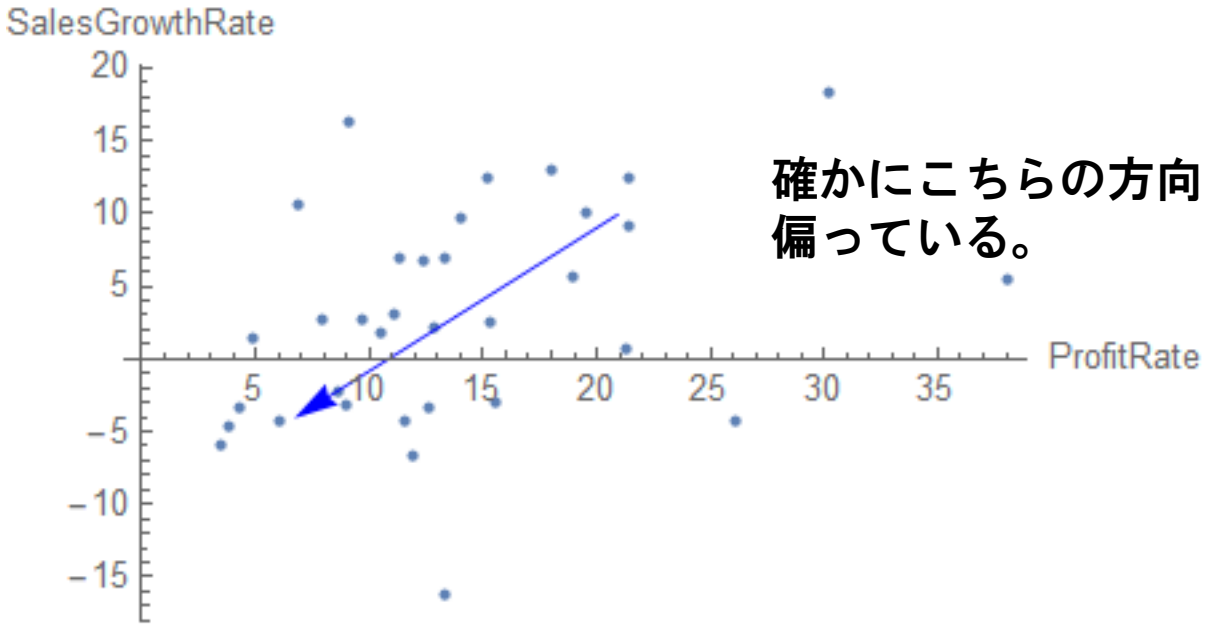
- 分散共分散行列 $\begin{pmatrix} 58.0442 & 21.405 \\ 21.405 & 57.4246 \end{pmatrix}$

第1主成分軸　PC1軸
第1固有ベクトル

- 第1固有ベクトル $\begin{pmatrix} -0.712205 \\ -0.701972 \end{pmatrix}$

問4：上記ベクトルの各要素の2乗の和（距離）はいくつか？

原点からの距離が１になるような値にしてあることを確認せよ。平方和

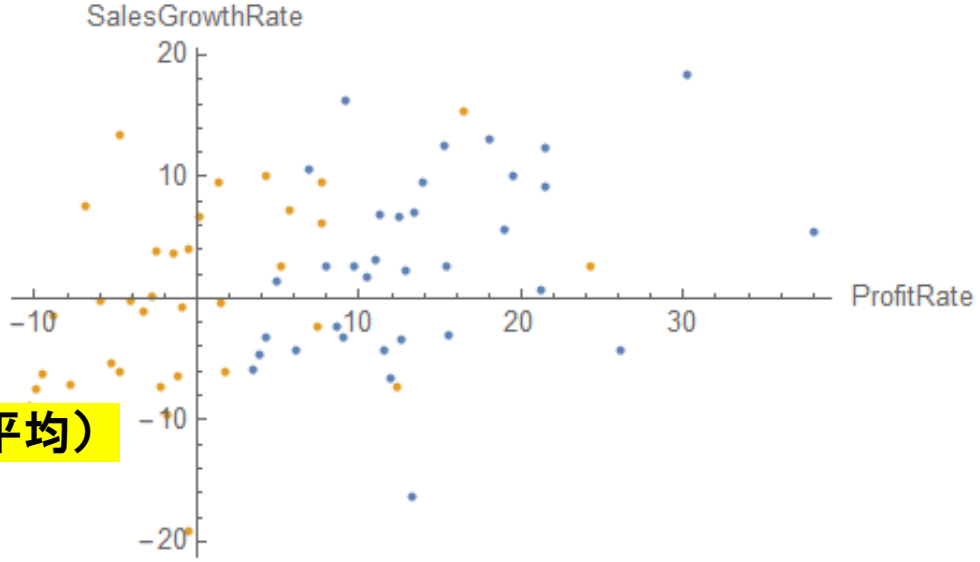


確かにこちらの方向に分布が
偏っている。

このケースのPC1の経営的意味解釈としては、
２つの指標の平均を取っている。
軸の向きの取り方によって正負は
変わるが、解釈としては同じ。
人間が正負の意味を逆にすればよい。

各社のPC１の値を求める

- データの重心(平均)を原点とする。青の点群が黄色の点群に移動
- 問5：EXCELでデータの重心のベクトルを計算せよ。X-meanと呼ぶことにする。



PC1の値＝固有ベクトル１×（その会社のデータ　ー　平均）

$$\begin{pmatrix} -0.712205 \\ -0.701972 \end{pmatrix}$$

- 固有ベクトルのバランス比で足し合わせて、その会社のPC1値を計算する

$$\begin{pmatrix} 7.45853 & -2.30265 \\ 16.4285 & 15.4174 \\ -1.39147 & 3.78735 \\ 24.2785 & 2.58735 \\ -4.72147 & 13.4074 \\ -0.461471 & 4.08735 \\ -4.80147 & -6.11265 \\ 7.62853 & 6.18735 \\ -1.23147 & -6.32265 \\ 5.13853 & 2.69735 \\ -0.491471 & -19.1626 \\ -1.87147 & -9.55265 \\ 0.168529 & 6.68735 \\ 12.3085 & -7.27265 \\ 1.72853 & -5.95265 \\ 4.18853 & 10.1374 \\ 1.50853 & -0.372647 \\ -2.47147 & 3.95735 \\ 1.38853 & 9.55735 \\ -4.15147 & -0.242647 \\ -8.91147 & -1.45265 \\ -5.91147 & -0.242647 \\ -5.23147 & -5.25265 \\ 7.63853 & 9.52735 \\ -0.911471 & -0.732647 \\ -2.75147 & 0.217353 \\ -2.30147 & -7.23265 \\ -7.76147 & -7.17265 \\ 5.67853 & 7.20735 \\ -9.99147 & -7.51265 \\ -9.52147 & -6.19265 \\ -10.3615 & -8.87265 \\ -3.35147 & -1.15265 \\ -6.94147 & 7.64735 \end{pmatrix} \times \begin{pmatrix} -0.712205 \\ -0.701972 \end{pmatrix} = \begin{pmatrix} -3.69561 \\ -22.523 \\ -1.6676 \\ -19.1075 \\ -6.04893 \\ -2.54055 \\ 7.71054 \\ -9.77642 \\ 5.31538 \\ -5.55315 \\ 13.8017 \\ 8.03856 \\ -4.81436 \\ -3.661 \\ 2.94752 \\ -10.0992 \\ -0.812794 \\ -1.01776 \\ -7.69791 \\ 3.12703 \\ 7.36651 \\ 4.38051 \\ 7.41309 \\ -12.1281 \\ 1.16345 \\ 1.80703 \\ 6.71623 \\ 10.5628 \\ -9.10363 \\ 12.3896 \\ 11.1283 \\ 13.6078 \\ 3.19606 \\ -0.424479 \end{pmatrix}$$

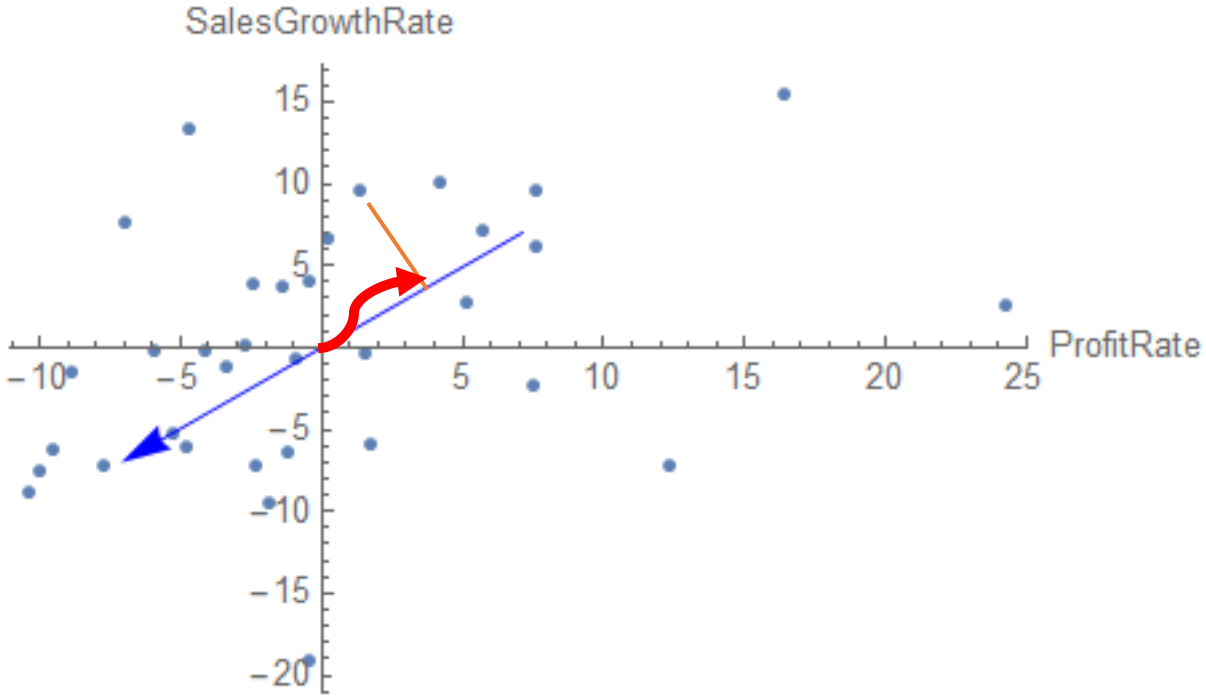
皆さんがパイソンで計算した第1固有ベクトルは $\begin{pmatrix} 0.712205 \\ 0.701972 \end{pmatrix}$

そのため、PC1の値の正負が逆になっている。

ツールによって軸の正方向の取り方は逆になることもある．人間が解釈を変える．強い→弱い

PC1値
原点からPC1軸への垂線の足への距離(正負あり）

- 以下の例の会社はPC１値がマイナスである(原点が０であるから）
- PC1軸は
  ProfiteRateの少なさ
  とSalesGrowthRate
  の少なさを
  表わすので、PC1値
  が負の方が、優れて
  いる会社であると
  言える。



- 正負は平均（原点）から見ての相対的位置を述べている。

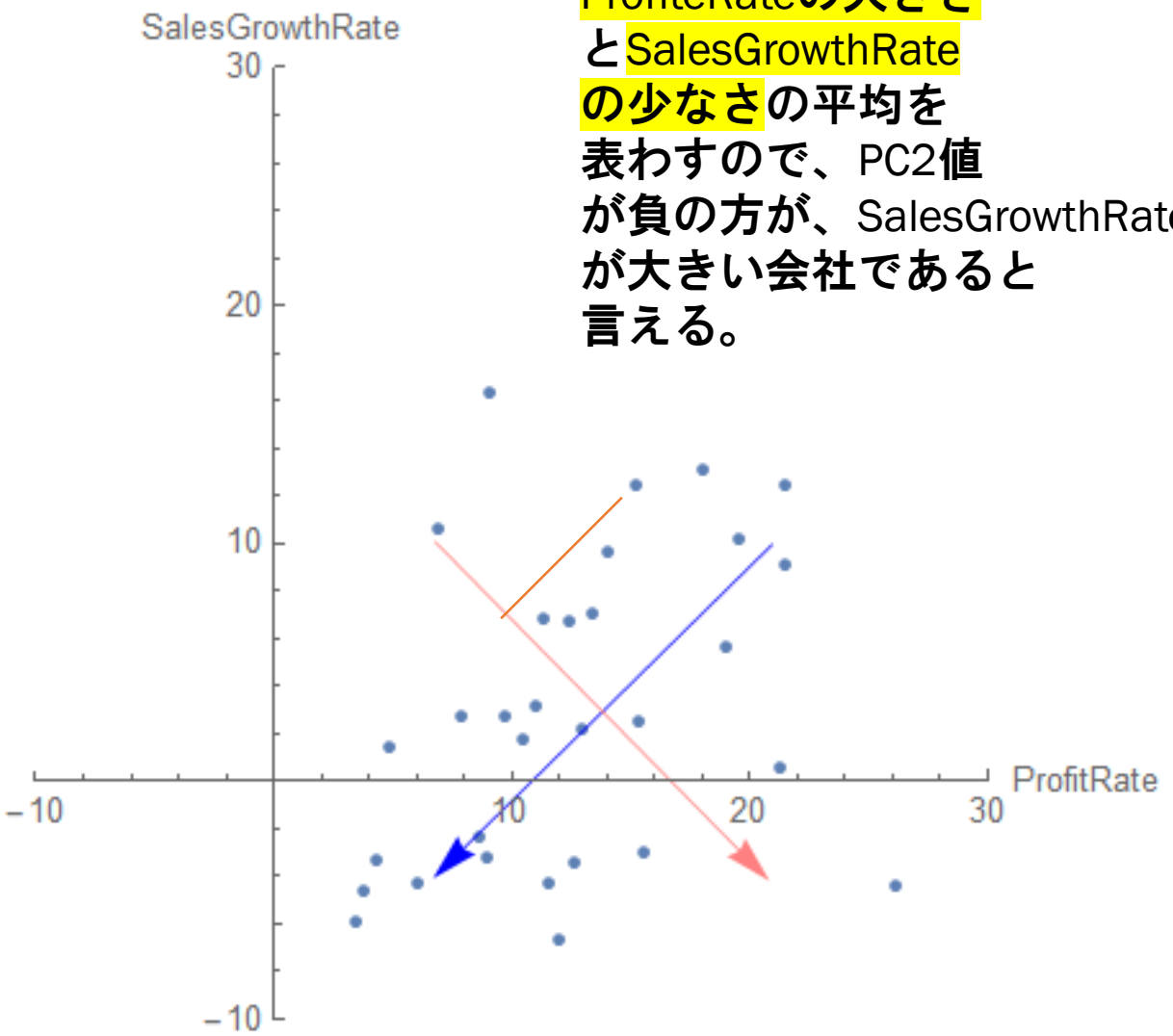
第2主成分軸とは？

$$\begin{pmatrix} 0.701972 \\ -0.712205 \end{pmatrix}$$

- PC1軸とPC2軸は必ず直交
- 問6：コサインの計算をしてPC1軸とPC2軸が直交することを確認せよ。

$$\begin{pmatrix} -0.712205 & 0.701972 \\ -0.701972 & -0.712205 \end{pmatrix}$$

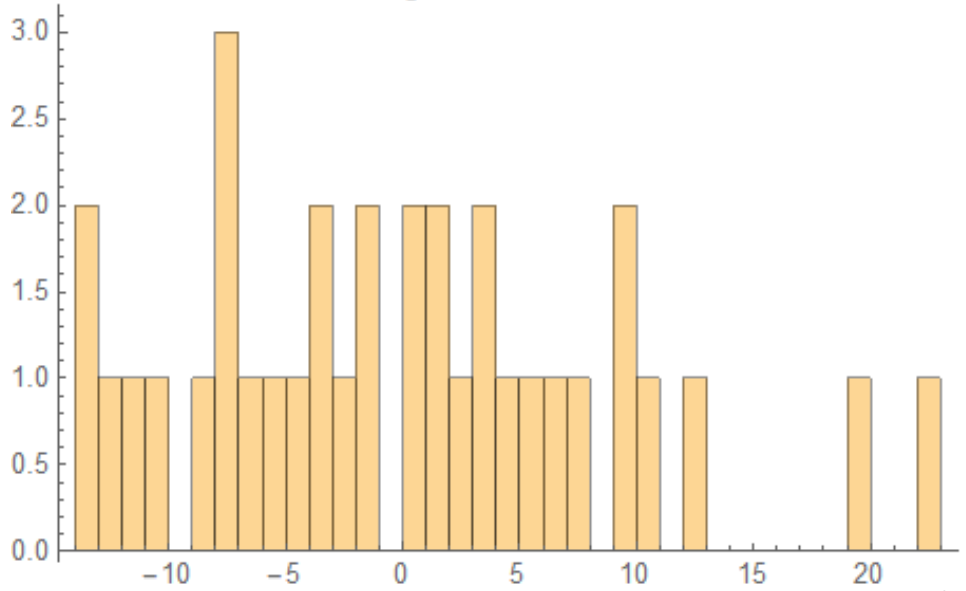
-0.71*0.70+(-0.70)*(-0.71)=0

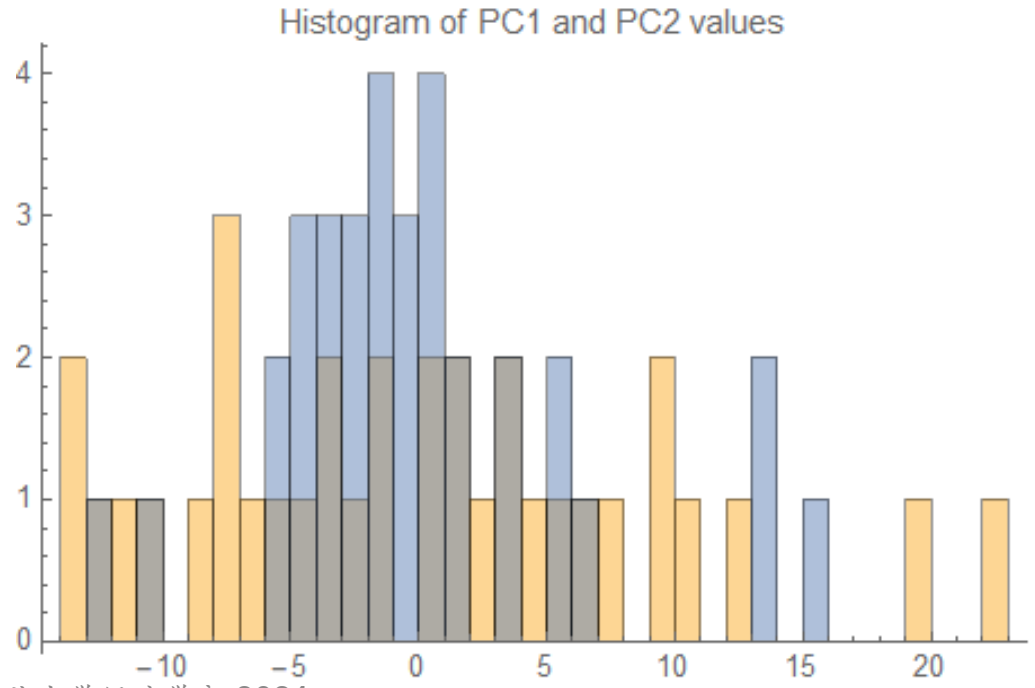


PC2軸は
ProfiteRateの大きさ
とSalesGrowthRate
の少なさの平均を
表わすので、PC2値
が負の方が、SalesGrowthRate
が大きい会社であると
言える。

PC1軸はPC1値の分散が最大化する方向

- PC1値のヒストグラムと
  PC2値のヒストグラム。どちらが分散が大きいですか？
- 問7：EXCELでPC1値の不偏分散とPC2値の不偏分散を求めよ。



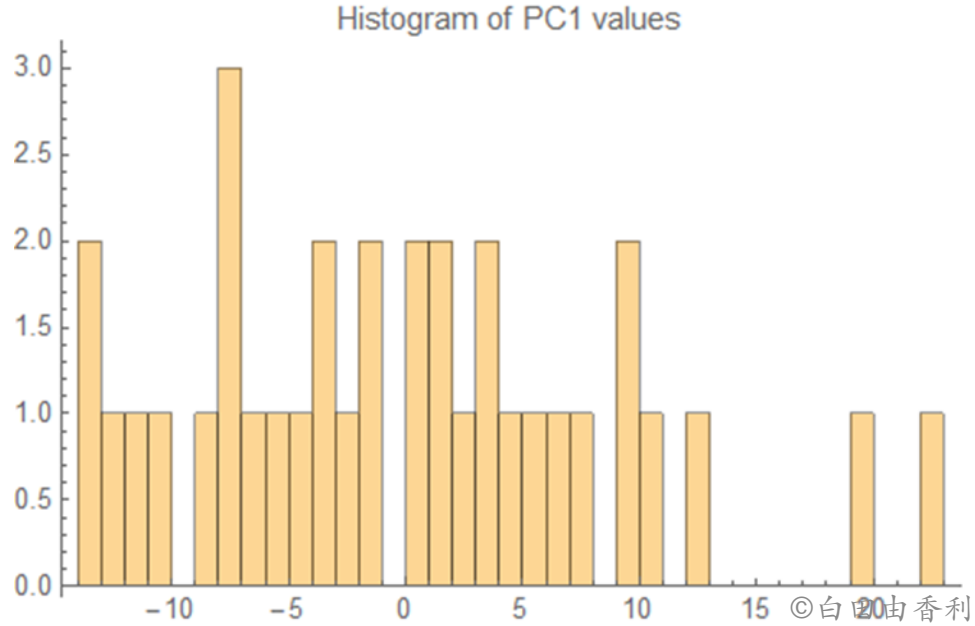


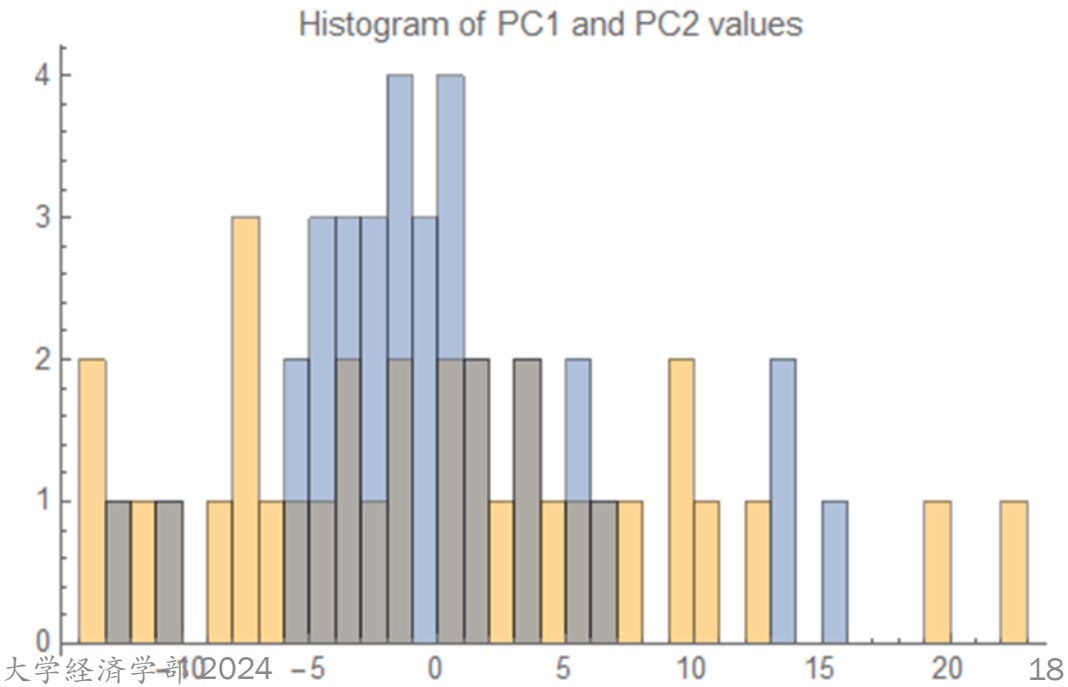
固有値はPC1値、PC２値の分散の大きさを表わす

- $\begin{pmatrix} 79.1416 \\ 36.3272 \end{pmatrix}$

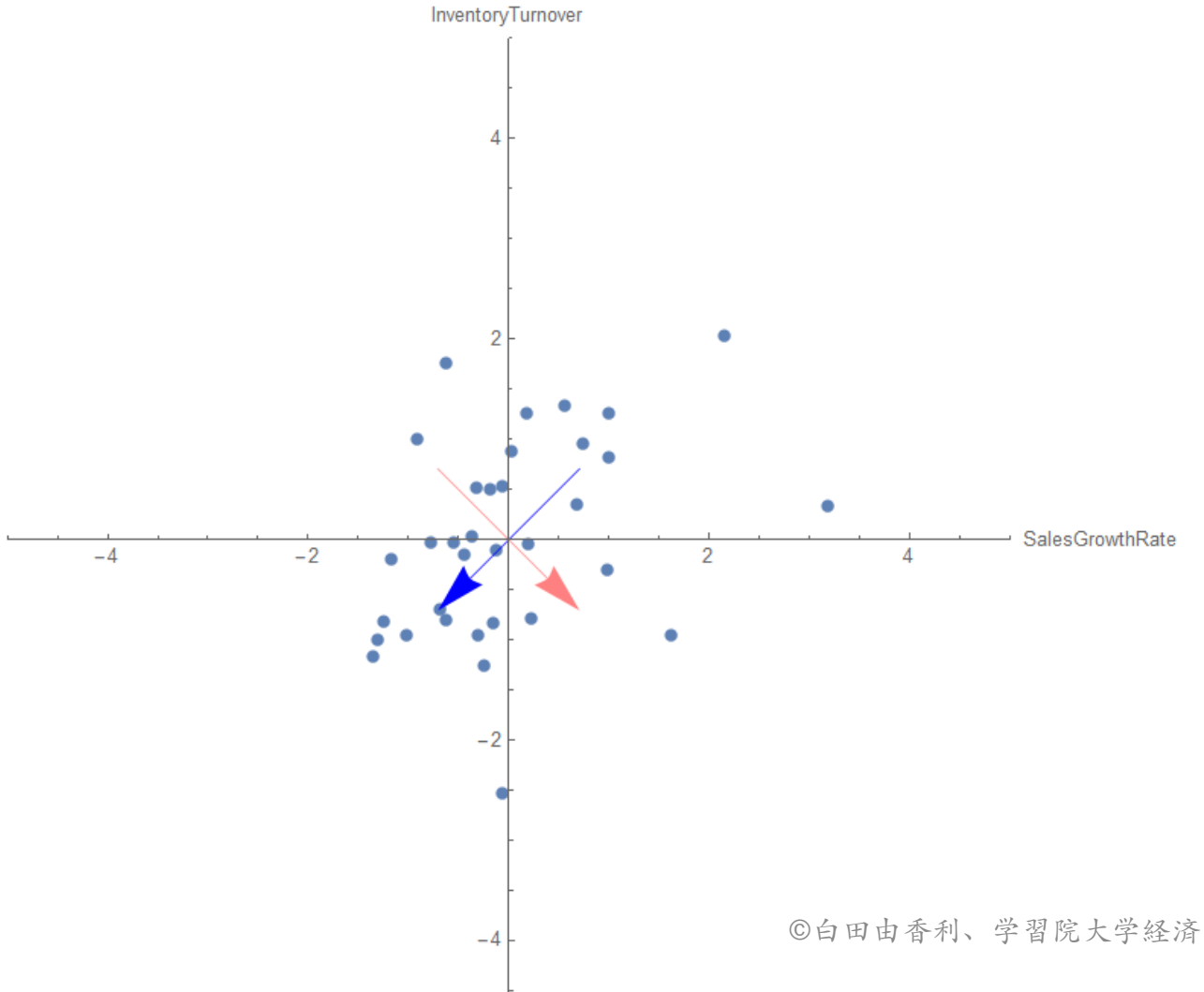
  PC1の分散のほうが大きい

- **固有ベクトル：どちらの方向に分布が伸びているか**
- **固有値：分布がどれ位広がっているのか**





課題8：製薬会社34社のデータを各指標に関して、標準化してから固有ベクトル



$$\begin{pmatrix} 1. & 0.370754 \\ 0.370754 & 1. \end{pmatrix}$$

$$\begin{pmatrix} -0.707107 & 0.707107 \\ -0.707107 & -0.707107 \end{pmatrix}$$

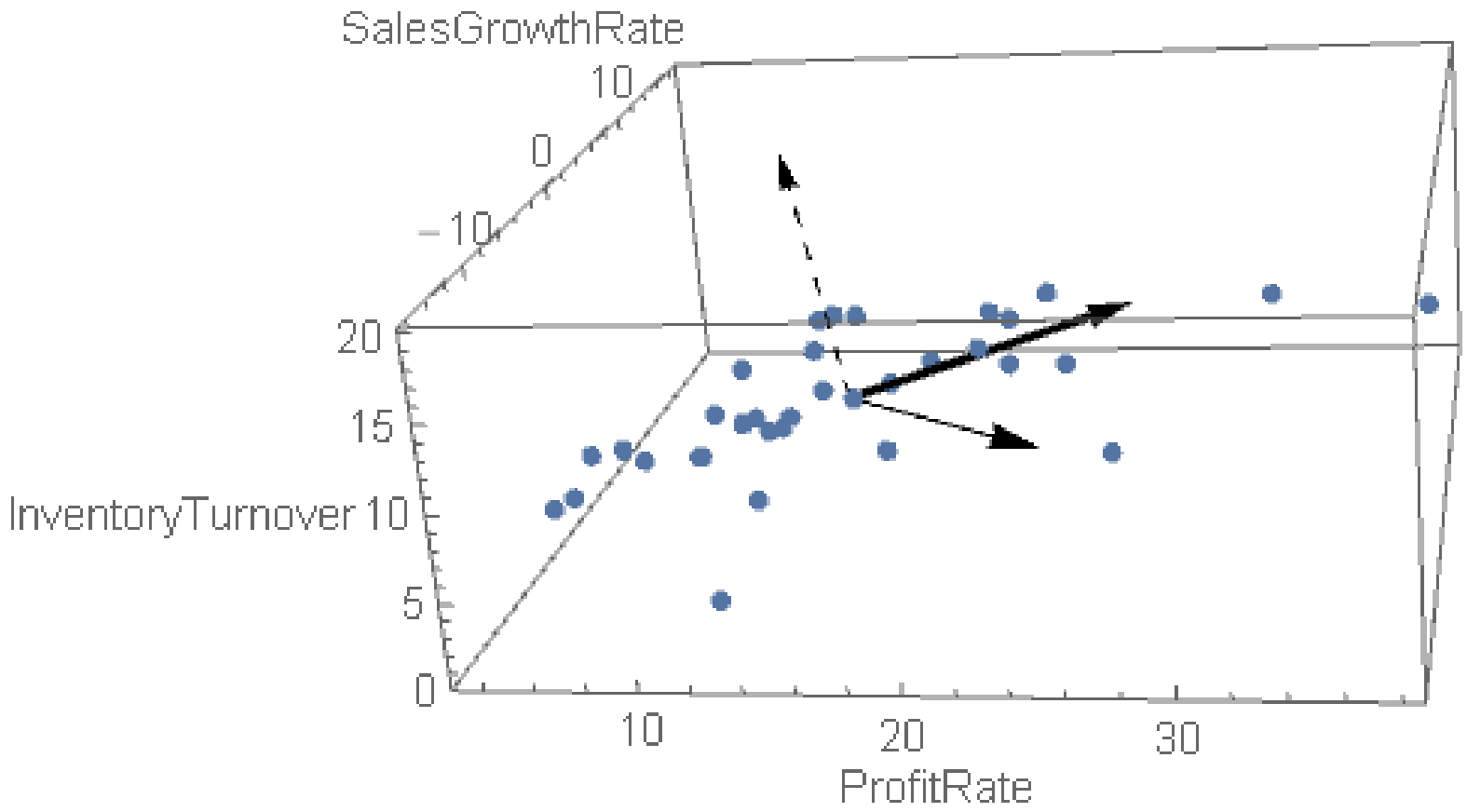
同じになる

$$\begin{pmatrix} 1.37075 \\ 0.629246 \end{pmatrix}$$

３つの経営指標

・入力されたデータそのままで散布図

PC1軸、 PC2軸、 PC3軸　直交する3軸

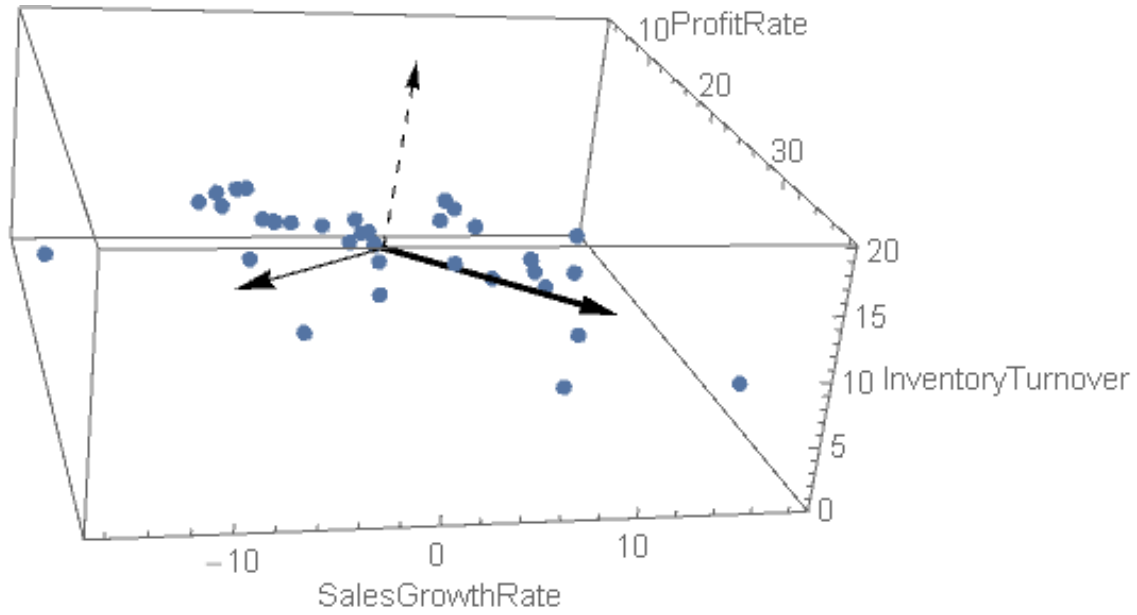


共分散行列、固有ベクトル、固有値

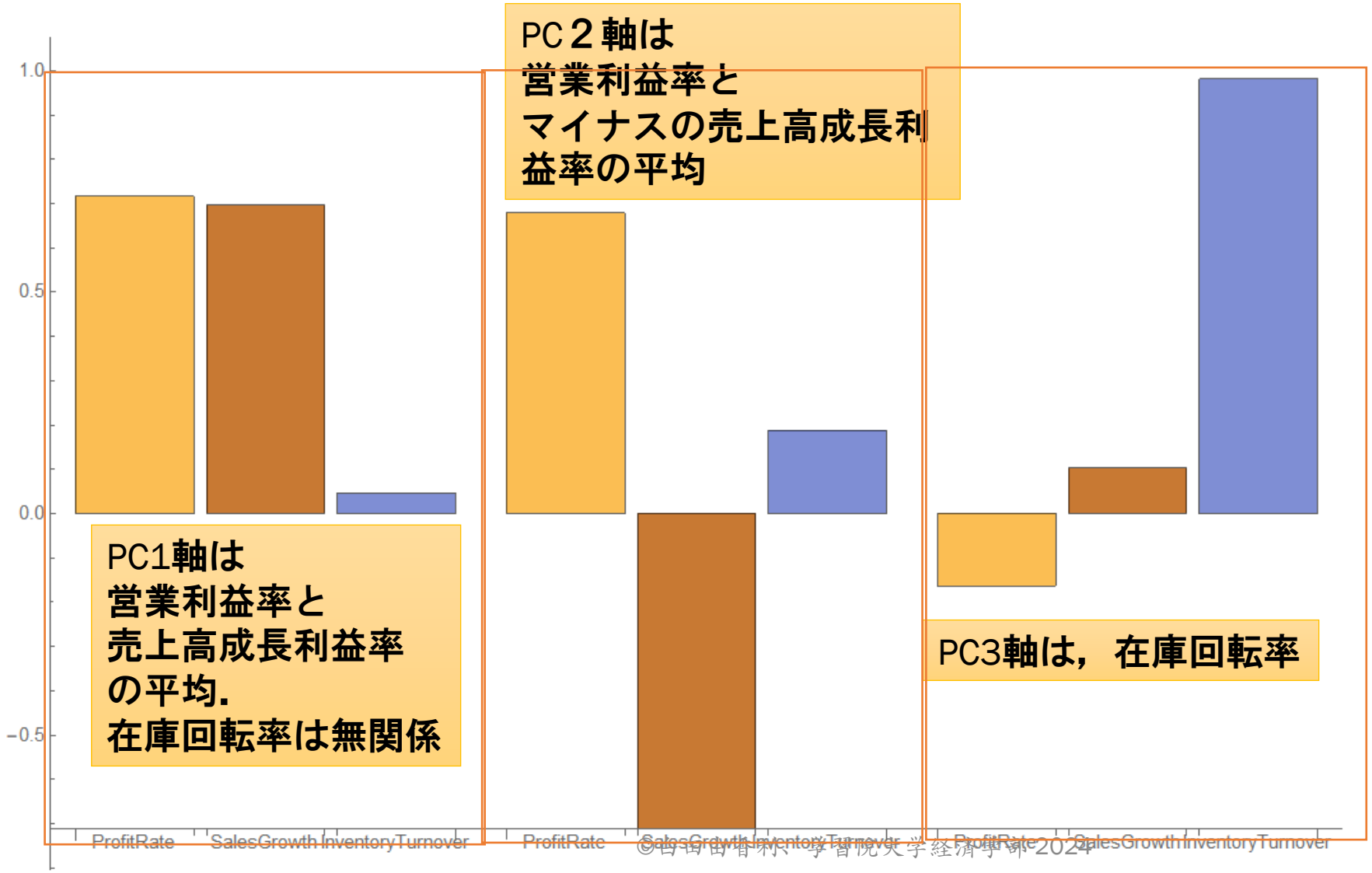
$$\begin{pmatrix} 58.0442 & 21.405 & 6.76006 \\ 21.405 & 57.4246 & -2.09129 \\ 6.76006 & -2.09129 & 5.08141 \end{pmatrix}$$

$$\begin{pmatrix} 0.716093 & 0.678753 & -0.162805 \\ 0.696514 & -0.710093 & 0.103138 \\ 0.0456015 & 0.187252 & 0.981253 \end{pmatrix}$$

$$\begin{pmatrix} 79.2944 \\ 37.5159 \\ 3.74 \end{pmatrix}$$

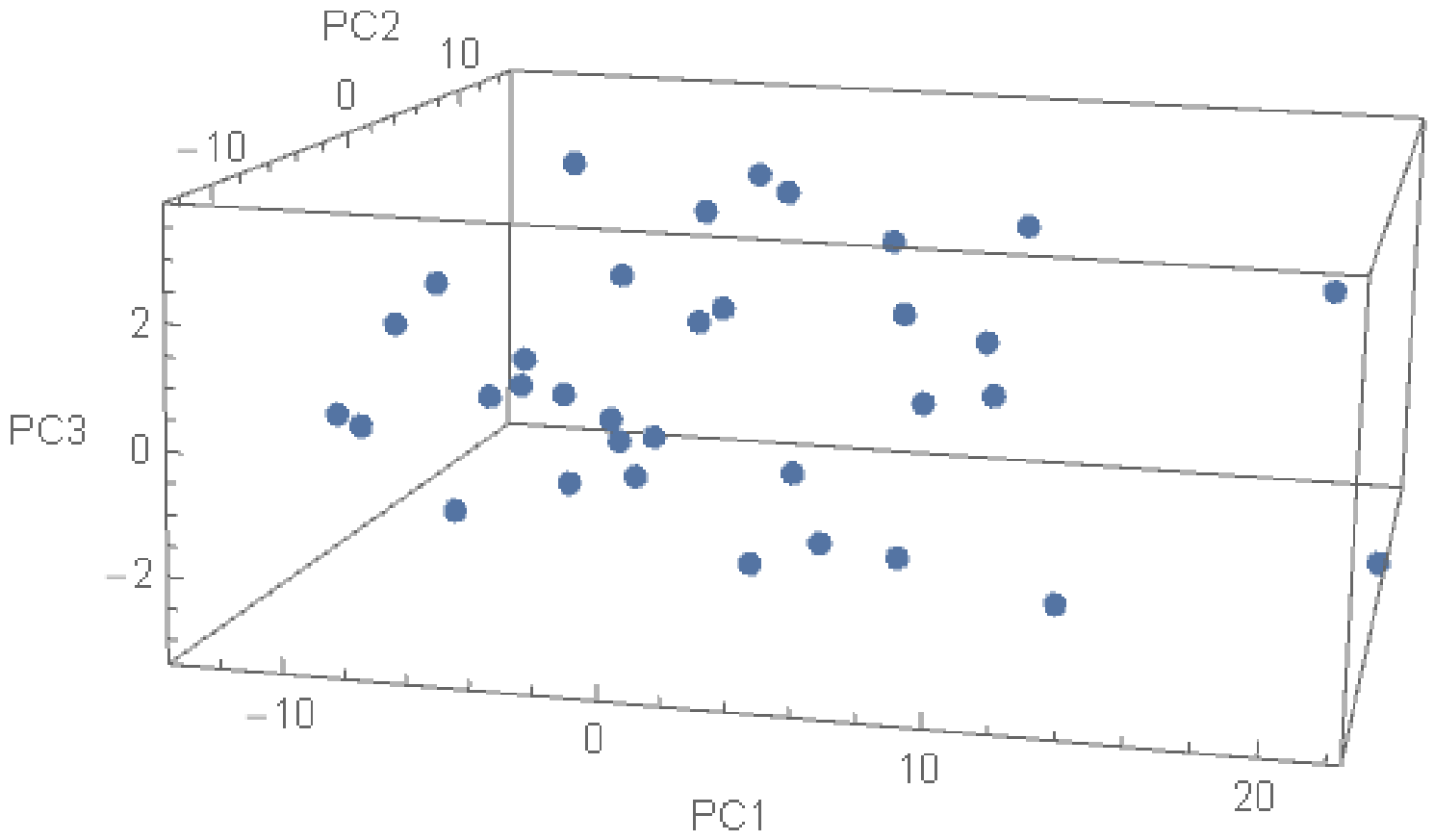


主成分軸の解釈
⇒バランス（固有ベクトルの要素）を見て考える



PC1値,PC2値,PC3値を計算する。そしてそのPC1,2,3の値でプロット

- ３主成分軸でグラフを描く

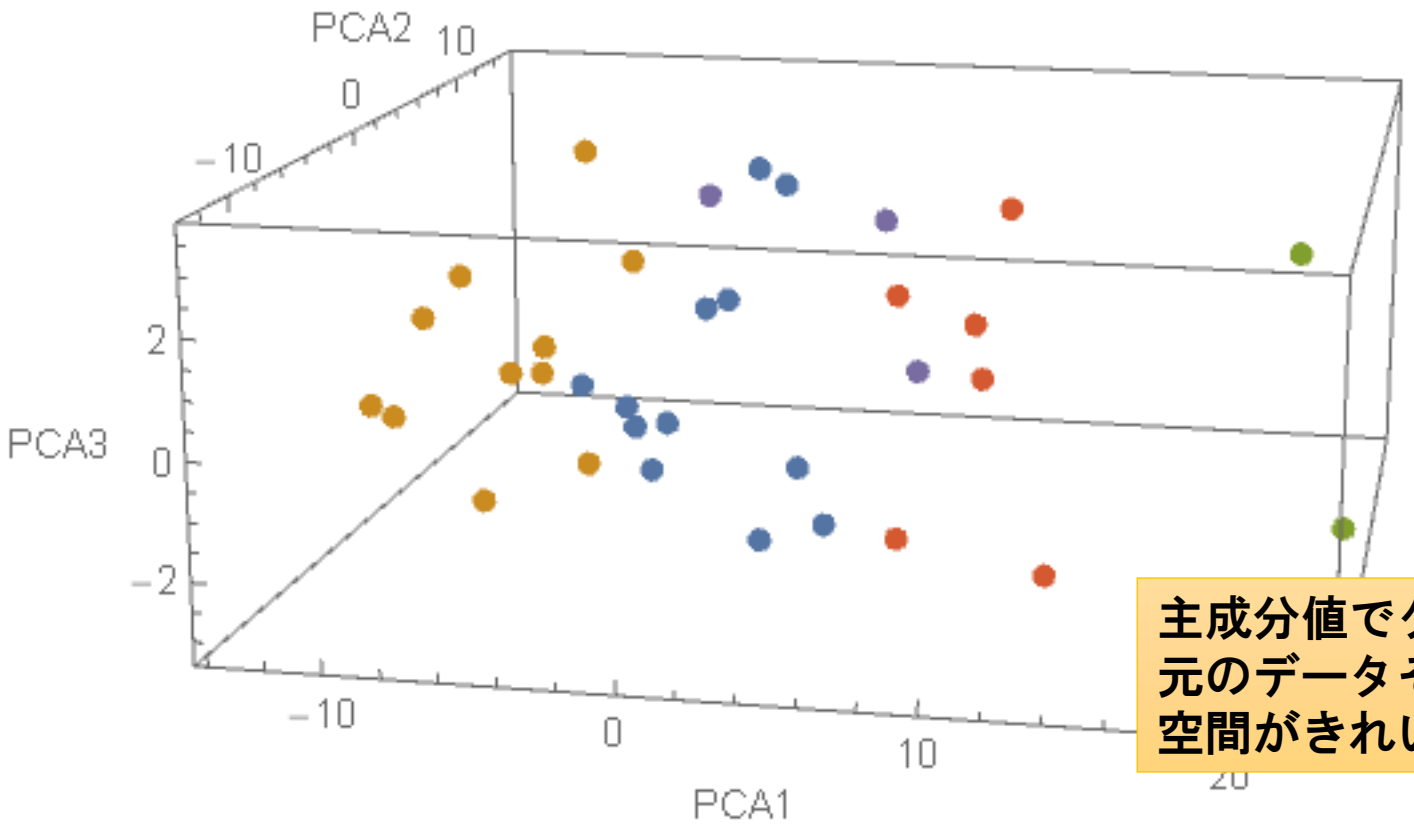


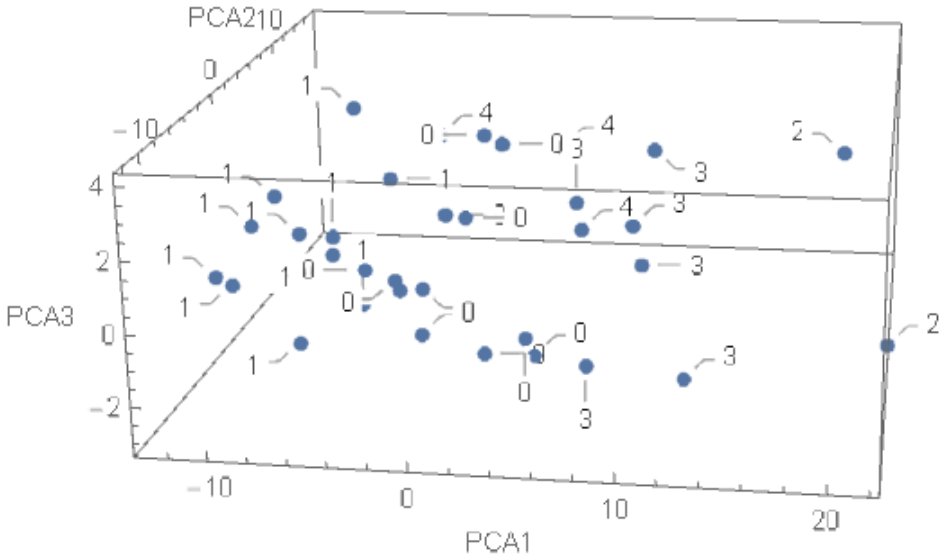
データを使ってクラスタリングを行う

- **教師なしデータ**
- K-Means

PC1,2,3の値での散布図にクラスタリング k=5の結果を色で示す

- クラスターごとに空間分割されている

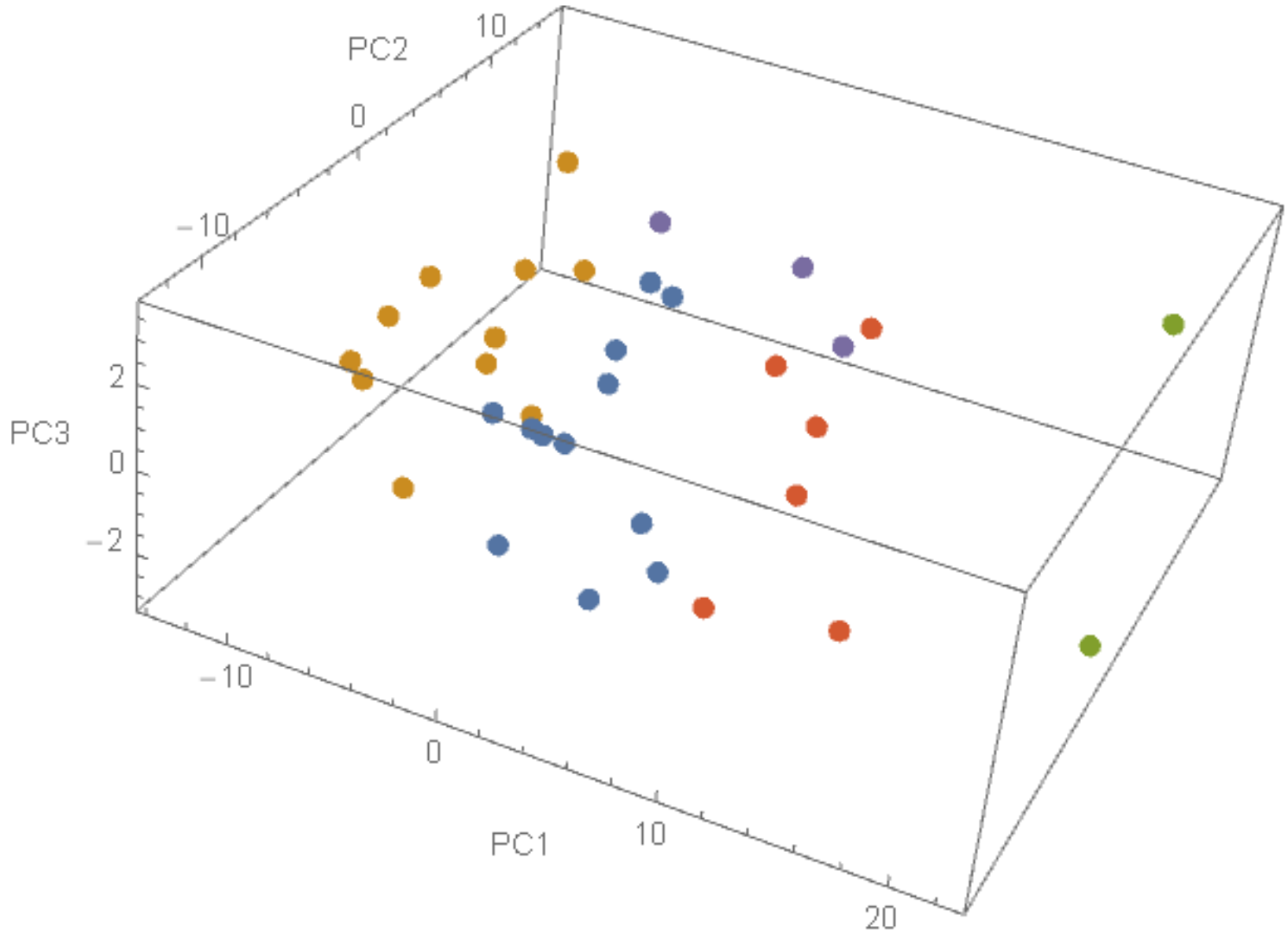




主成分値でクラスタリングしなくても，
元のデータそのものでクラスタリングした結果であっても
空間がきれいに分かれている

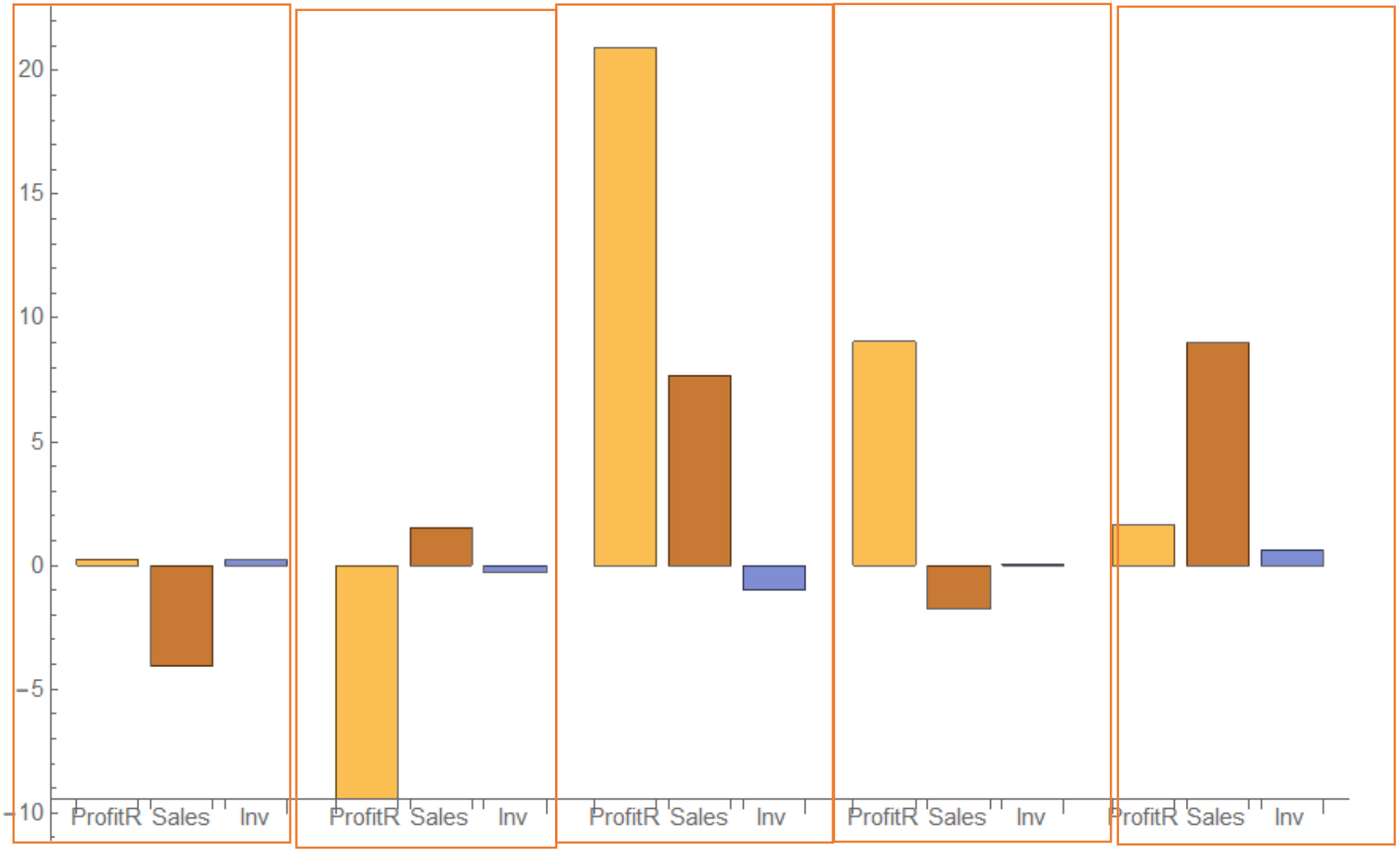
どうしてか？

- 直交する方向に軸をとったから



各クラスターはどのような特徴をもつか クラスタ毎に３つの経営指標の平均値

- 第3クラスタが最優秀
- ついで、クラスタ４、クラスタ５
- 最もよくない企業のクラスタは２
- 課題9：クラスタ毎に３つの経営指標の平均値をEXCELで求めよ



まとめ

- PCAはデータの分布の概要が分かる.
- 他の教材PDF
- https://shirotaabc.sakura.ne.jp/hosomichi/PDF/PCAcluster.pdf
- https://shirotaabc.sakura.ne.jp/hosomichi/PDF/PCA.pdf
- https://shirotaabc.sakura.ne.jp/hosomichi/PDF/PCAmulticolinearity.pdf