データ解析

第5回: 因子分析

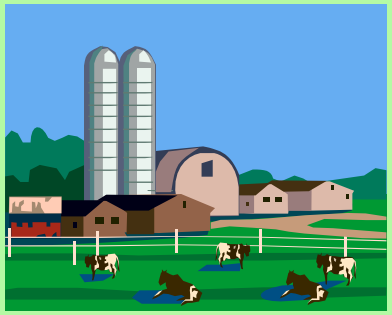
渡辺澄夫

見つけた
要因は
真実か
幻か

観測データ

回帰・判別分析

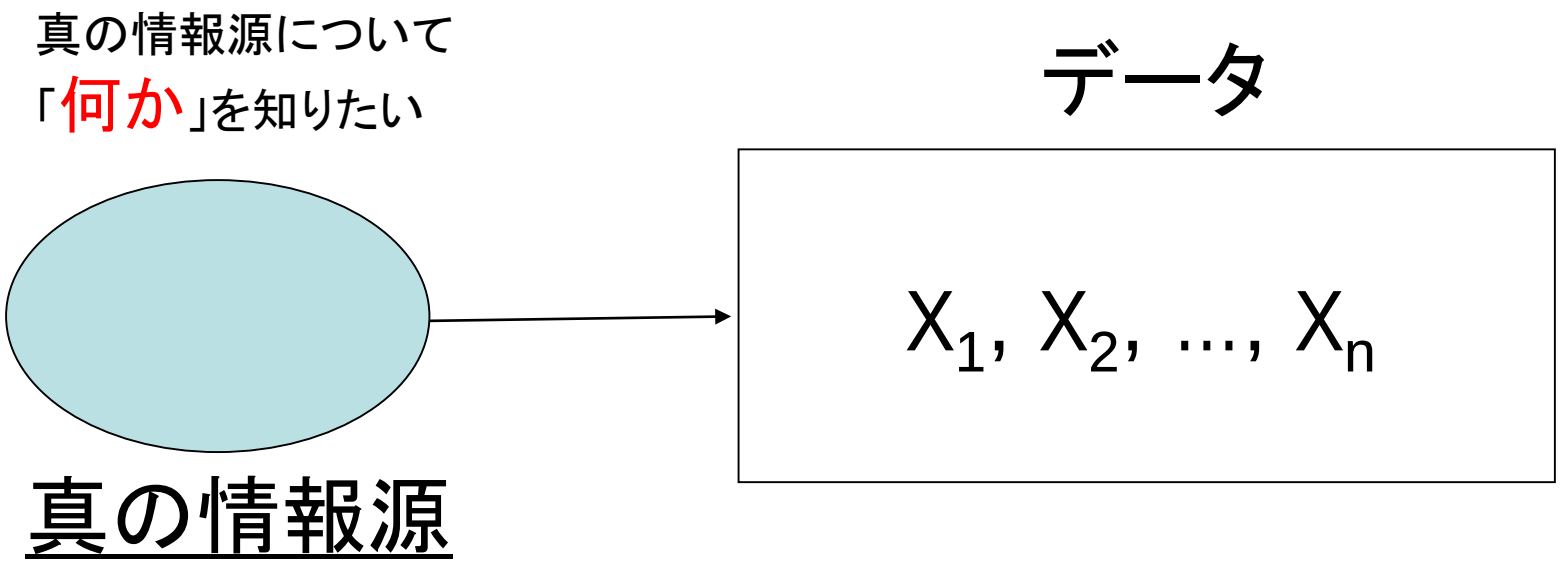
解析方法



主成分・因子
・クラスタ分析

ラベルを持たないデータ

主成分分析ではデータの情報をできるだけ残して少ない次元の空間上に表現する方法を考えた。因子分析ではデータを生成した要因が少ない個数であると考えて推測する問題を考える。



データの背後に
少数の要因があるか

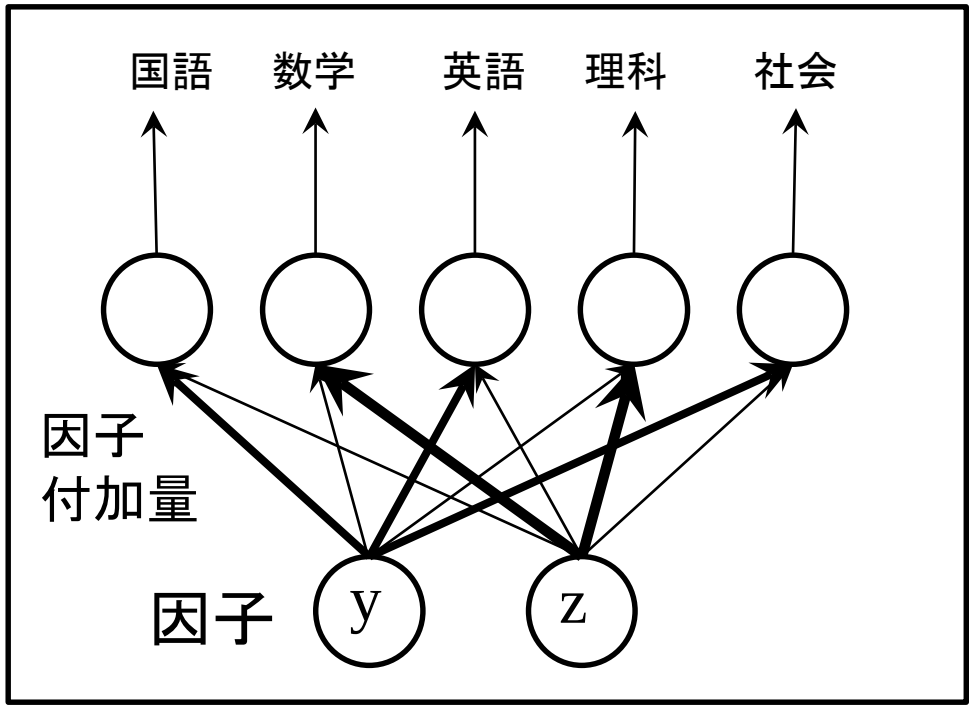
注意：要因といっても、「原因と結果」における
原因というわけではなく、少数個の変数で
説明できるかという意味です。

因子分析とは

（例）

「たくさんの中学生の5科目の
成績を調べ、隠れている
要因である(y,z) を見つけて、
そこからの影響の強さを
調べたら図のようになった。
(y,z) は何ですか」

という問題を考えるために
因子(y,z)と因子負荷量の
両方を求める方法を作りたい。



(※)この例は仮想例であり
実データではありません。

（国、数、英、理、社）の値と
因子(y,z) は個人によって変動する。
因子負荷量の値は個人によらない。

因子分析のモデル

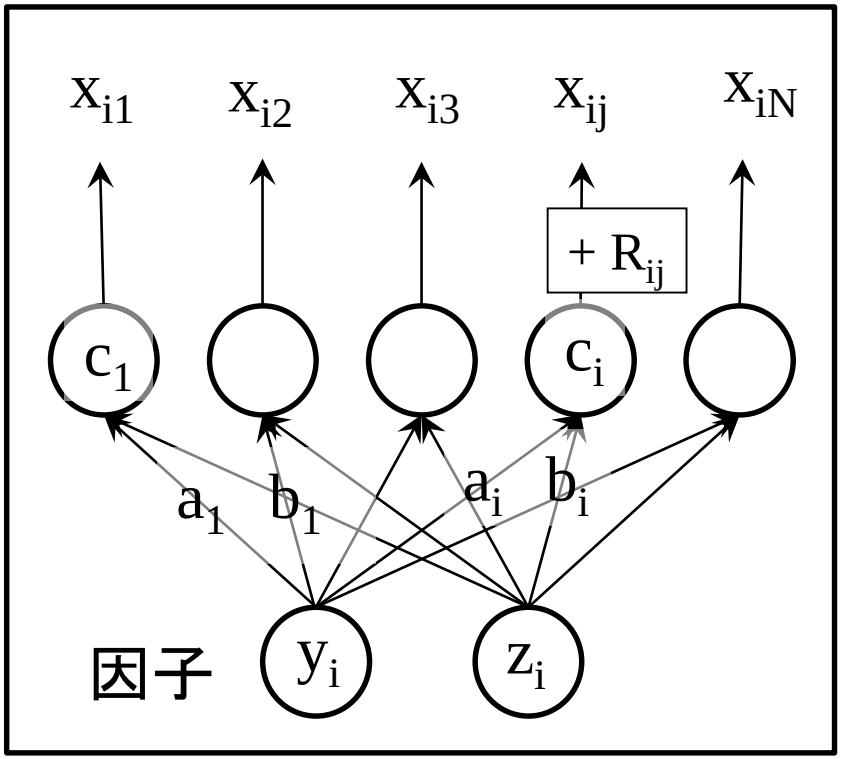
因子分析に用いられるモデル

データ $X_i = (X_{i1}, X_{i2}, \ldots, X_{iN}) \in \mathbf{R}^N$ $(i=1,2,\ldots,n)$ が次のモデルに従うモデルを考える。

$$X_{ij} = a_j Y_i + b_j Z_i + c_j + R_{ij}$$

ここで

(1) パラメータ (a_j, b_j, c_j) $(j=1,2\ldots,N)$
(2) 潜在変数 (Y_i, Z_i) $(i=1,2,\ldots,n)$
は平均0分散1の正規分布に従う。
(3) R_{ij}は平均0分散(1/s) の
正規分布に従う。



(注意）実世界のデータが有限個の要因で説明できることはほとんどないが、このモデルで実世界を推測したときに何がわかるかを考察する。これはモデルであって実世界ではない。

因子分析のモデル

統計モデル

$$p(x|y,z,a,b,c,s) = \prod_{j=1}^{N} (s/2\pi)^{1/2} \exp\{ -(s/2)(x_j - a_j y - b_j z - c_j)^2 \}$$

因子 (y,z) の事前分布を $p(y,z) = (1/2\pi) \exp\{ -(1/2)(y^2 + z^2)\}$ とし

パラメータ (a,b,c,s) の事前分布を一様分布とする。このとき

$$p(y,z,a,b,c,s|x) = (1/Z)\, p(x|y,z,a,b,c,s)\, p(y,z).$$

ここで Z は x だけの関数であり (y,z,a,b,c,s) に依存しない。

方法: (事後確率最大化法)
データ $\{X_i;\ i=1,2,\ldots,n\}$ が与えられた時の (y_i,z_i,a,b,c,s) の条件つき確率密度関数(に比例する量)

$$L = \prod_{i=1}^{n} p(X_i|y_i,z_i,a,b,c,s)\, p(y_i,z_i)$$

を最大にする (y_i,z_i,a,b,c,s) を推定量とする。

因子分析と回帰分析

事後確率に比例する量 L は

$$L = \prod_{i=1}^{n} (s/2\pi)^{N/2} \exp\{ - \sum_{j=1}^{N} (s/2)(X_{ij} - a_j y_i - b_j z_i - c_j)^2 - (1/2)(y_i^2 + z_i^2) \}$$

(注意）もしも $\{(y_i,z_i) ; i=1,2,\ldots,n\}$ の値がわかっているときには、回帰分析と同じであり、L を最大にするパラメータを直接解くことができるが因子分析では y_i,z_i,a,b,c,s のすべてが不明なので、これを直接解くことは難しい。 この対数 log L を最大化する方法として最急上昇法を適用することにする (-log L についての最急降下法と同じ）。

（注意） (a,b,c,s,y,z) についての最大化を考えたいのであるが

(a,b) をθ回転して(y,z) を(θ)回転しても L は同じ

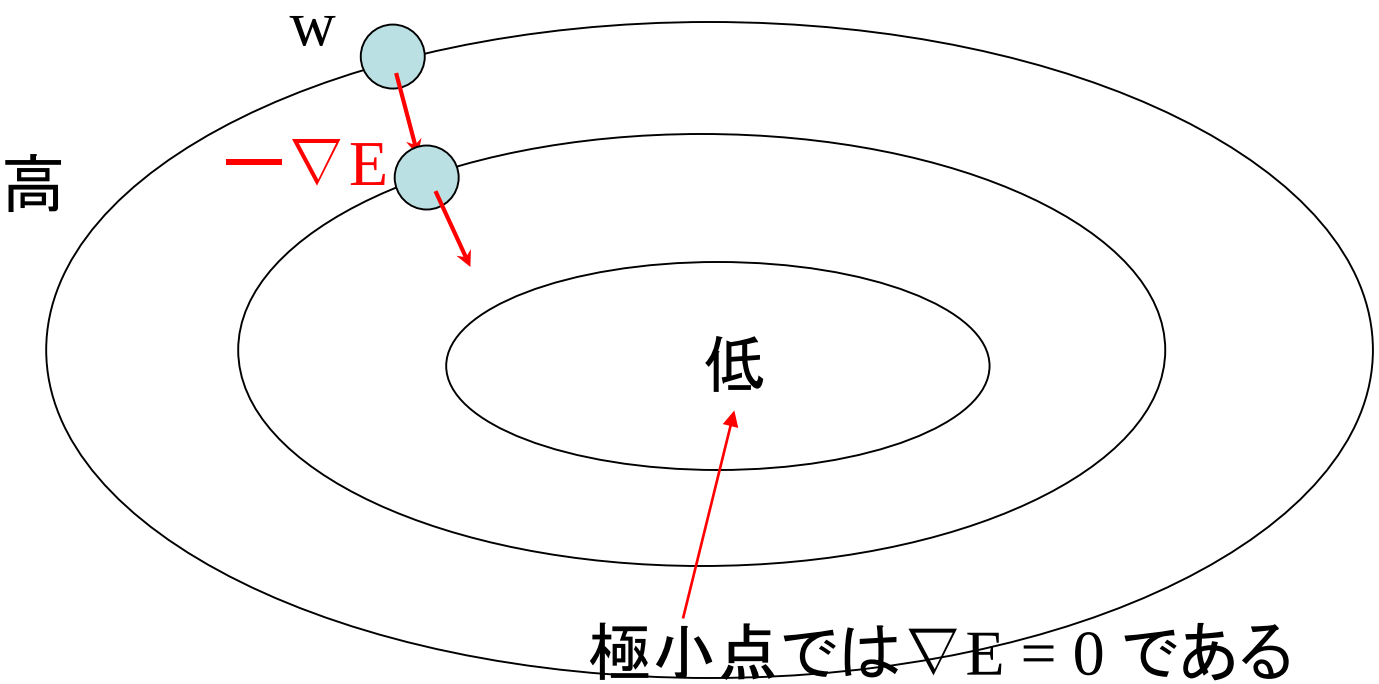
であるから、最大化する値はユニークには定まらない。

準備： 最急降下法

一般に $E(w) = E(w_1, w_2, \ldots, w_d)$ を最小にする w^* を探す方法を考えます。

$$\nabla E(w) = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d} \right)$$

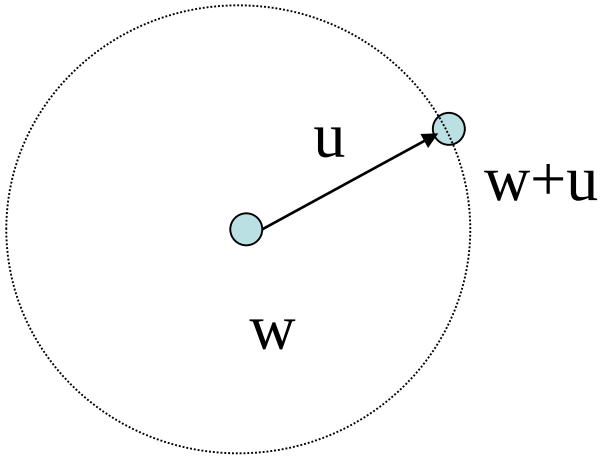
$= \mathrm{grad}\, E$ とも書く



なぜ∇E

テーラー展開 : ||u|| が十分に小さいとき

$$E(w+u) = E(w) + u \cdot \nabla E(w) + O(u^2)$$



$\| u \|$ = 一定のとき

E(w+u) が最小になるのは

$u \propto -\nabla E(w)$ のとき

(注意)ー∇E は等高線と直交している

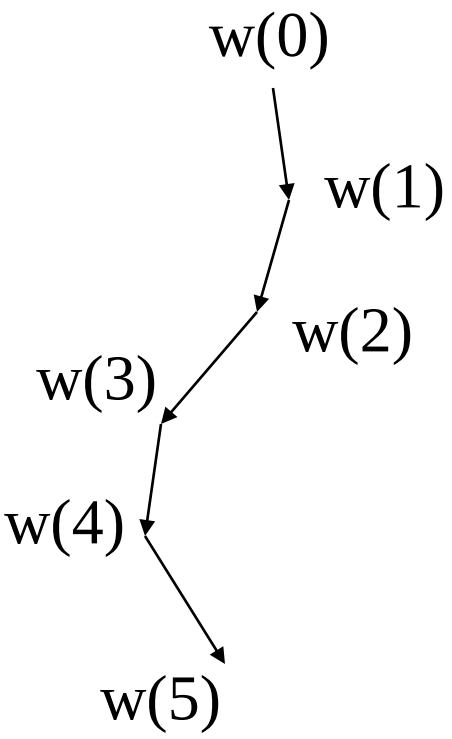
最急降下法

常微分方程式 $\dfrac{dw}{dt} = -\nabla E(w)$

離散化　t=1,2,3,…　　$\eta>0$ は定数

$$w(t+1) - w(t) = -\eta \nabla E(w(t))$$

$\eta>0$　:十分小さな定数



アルゴリズム

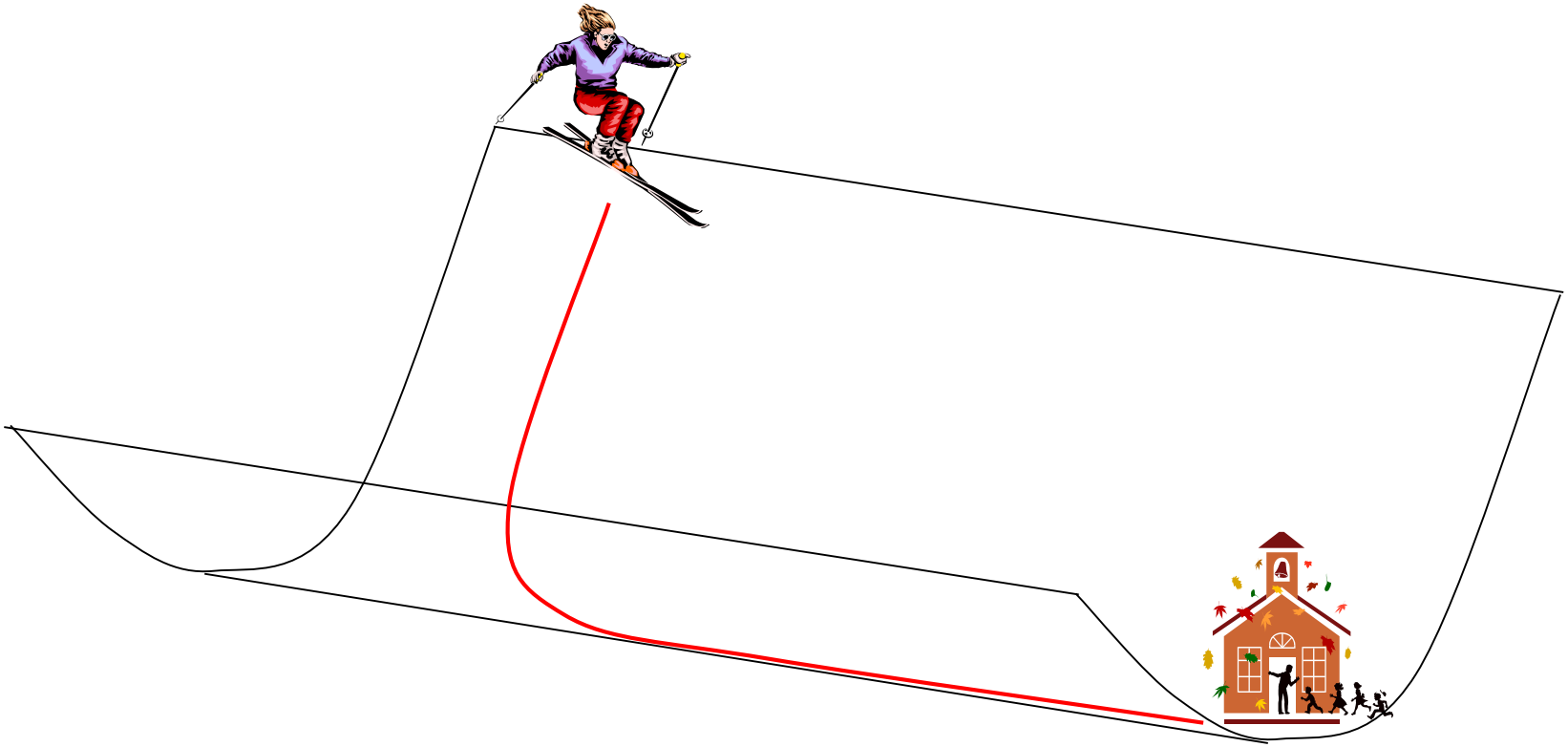
(1) 出発点 w(1) をランダムに初期化

(2) $w(n+1) = w(n) - \eta \nabla E(w(n))$

(3) n=1,2,3,…を繰り返す

E(w(n)) はだんだん小さくなってゆく
ただし極小解を抜け出すことはできない。

最急降下法は効率的か



最急降下法は局所的には効率的ですが大局的には効率的ではありません。これの改善のために様々な方法があります。最急降下法を途中で止めるほうが予測誤差を小さくすることもあります。

最急上昇法による最適化

最急上昇法で log L を最大化する。

$$\log L = \Sigma_i (-1/2)\{\Sigma_j s(x_{ij} - a_j y_i - b_j z_i - c_j)^2 + (y_i^2 + z_i^2)\} - (Nn/2)\log s$$

最急上昇法

$$\Delta a_j = \partial \log L / \partial a_j = s\, \Sigma_i \{(x_{ij} - a_j y_i - b_j z_i - c_j)\ y_i\}$$

$$\Delta b_j = \partial \log L / \partial b_j = s\, \Sigma_i \{(x_{ij} - a_j y_i - b_j z_i - c_j)\ z_i\}$$

$$\Delta c_j = \partial \log L / \partial c_j = s\, \Sigma_i \{(x_{ij} - a_j y_i - b_j z_i - c_j)\}$$

$$\Delta y_i = \partial \log L / \partial y_i = \{s\, \Sigma_j (x_{ij} - a_j y_i - b_j z_i - c_j)\, a_j - y_i\}$$

$$\Delta z_i = \partial \log L / \partial z_i = \{s\, \Sigma_j (x_{ij} - a_j y_i - b_j z_i - c_j)\, b_j - z_i\}$$

s については毎回最適化: $s = (1/nN)\, \Sigma_i \{\Sigma_j (x_{ij} - a_j y_i - b_j z_i - c_j)^2\}$

問1

log L が $\{a_j, b_j\}$ の関数として下記のように与えられているという。
最急上昇法で$\{a_j,b_j\}$ を求めるための式を導出せよ。

$$\boxed{\log L = \sum_{i=1}^{n} (-1/2)\{\sum_{j=1}^{N} (x_{ij} - a_j b_j y_i - b_j^2 z_i)^2\}}$$

$\Delta a_j =$

$\Delta b_j =$

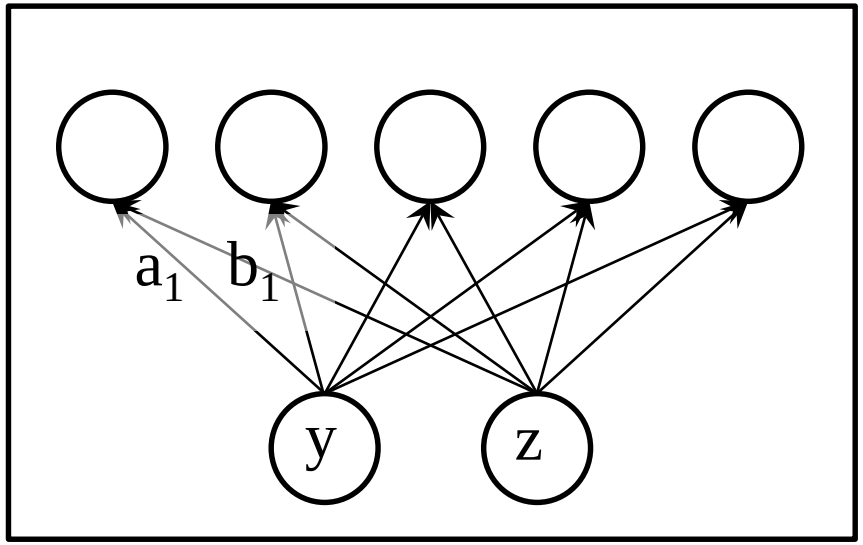
(a,b)の回転の問題

$\{(a_j,b_j);j=1,2,\ldots,N\}$ を回転したものは数式の上では等価であるため

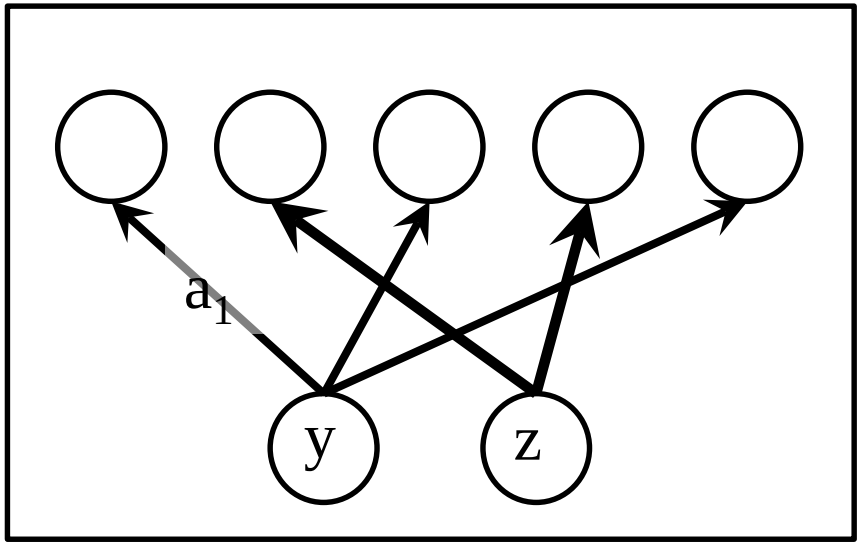
最適化後に、因子(y,z) の解釈を容易にするための方法が工夫される

ことが多い。「解釈が容易」とは･･･

「解釈が難しい」



「解釈が容易」

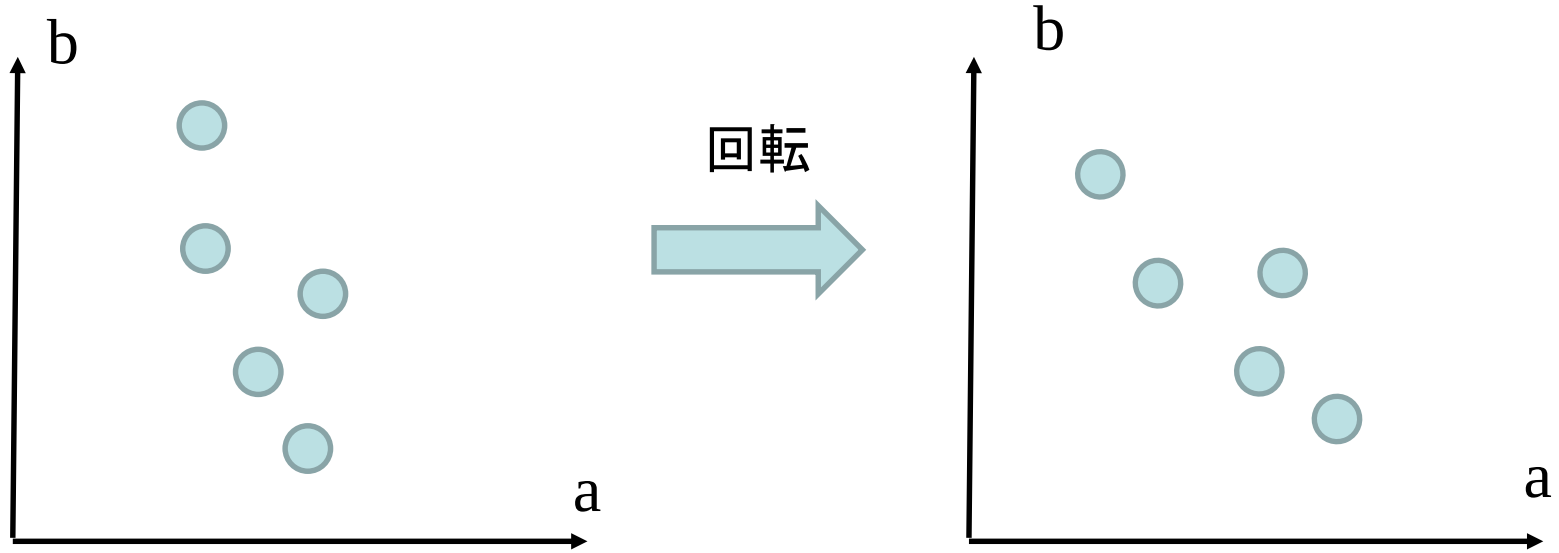


(a,b)の回転の問題つづき

(a,b) $=\{(a_j,b_j)\}$ の解釈しやすさを表す指標にもいろいろなものが提案されている。a^2 の分散と b^2 の分散の和

$$\sum_{j=1}^{N} \{ V(a_j^2) + V(b_j^2) \}$$

を最大にするように回転すると、解釈が比較的容易になると思われるのでこの方法をバリマックス法という。(a,b) を回転すると(y,z)も回転。３変数以上になると回転が難しくなる。



解き方の例

$$A_j(\theta) = a_j \cos(\theta) - b_j \sin(\theta) \qquad (j=1,2,\ldots N)$$
$$B_j(\theta) = a_j \sin(\theta) + b_j \cos(\theta)$$

$$\boxed{\begin{aligned} V(\theta) = \sum_{j=1}^{N} [\, & A_j(\theta)^4 - \{ (1/N)\Sigma_k A_k(\theta)^2 \}^2 \\ & + B_j(\theta)^4 - \{ (1/N)\Sigma_k B_k(\theta)^2 \}^2 \,] \end{aligned}}$$

これを最急上昇法で最大化する。

$$\boxed{\begin{aligned} \Delta\theta = \Sigma [\, & 4A_j(\theta)^3 A_j(\theta)' - \{(1/N)\, \Sigma_k A_k(\theta)^2 \}\{(1/N)\, \Sigma_k A_k(\theta)\, A_k(\theta)'\} \\ & + [\, 4B_j(\theta)^3 B_j(\theta)' - \{(1/N)\, \Sigma_k B_k(\theta)^2 \}\{(1/N)\, \Sigma_k B_k(\theta)\, B_k(\theta)'\} \,] \end{aligned}}$$

因子分析のアルゴリズム

１. データ $\{X_i \in \mathbf{R}^N;\ i=1,2,\ldots,n\}$ を得る。必要だと考えたら標準化して
各次元ごとに平均0分散1にする。

2. 次の値を最大化する (y_i,z_i,a,b,c,s) を求める。

$$\log L = \sum_{i=1}^{n} (-1/2s)\{ \sum_{j=1}^{N} (x_{ij} - a_j y_i - b_j z_i - c_j)^2 + (y_i^2 + z_i^2)/s \}$$

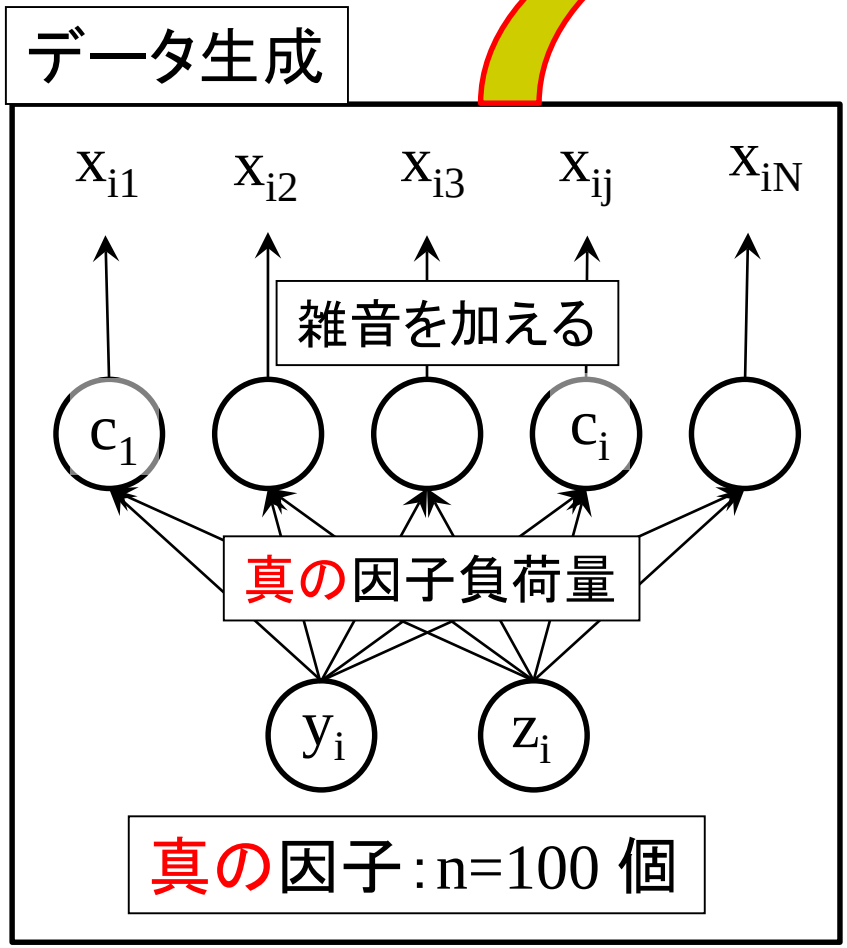
３. 回転して (a,b) を理解しやすくする。

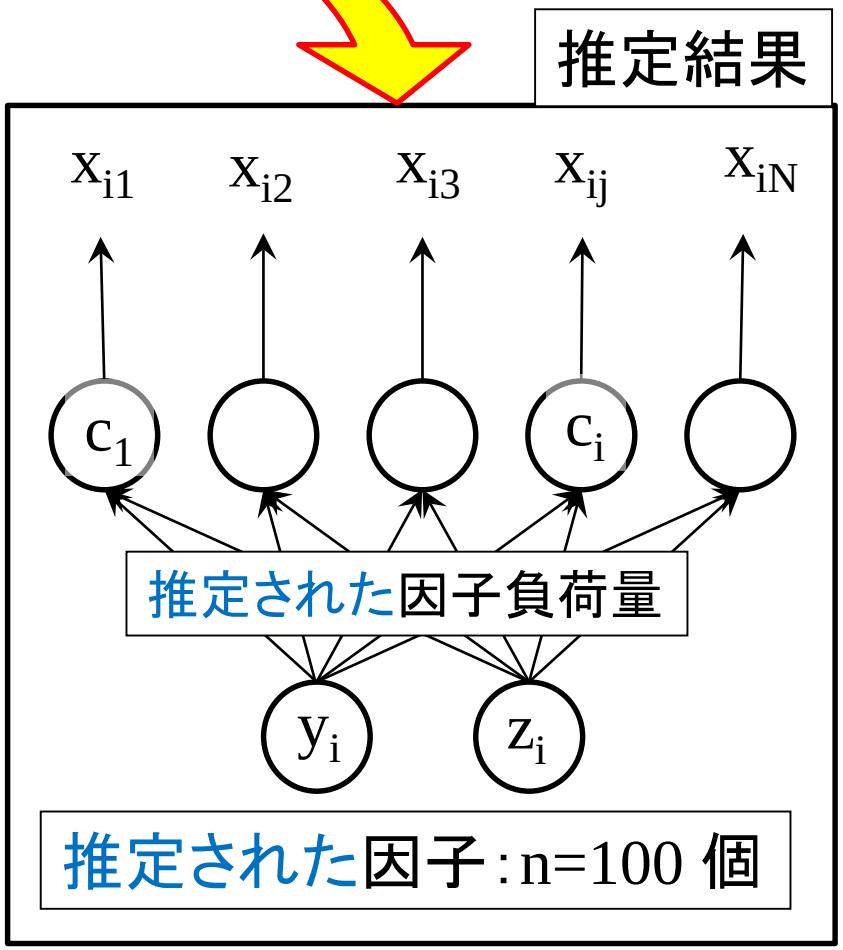
４. 因子負荷量 を表すベクトル $a \in \mathbf{R}^n$, $b \in \mathbf{R}^N$ を表示して意味を考える。

５. 因子$\{ (y_i,z_i)\}$ を2次元空間に表示して意味を考える。

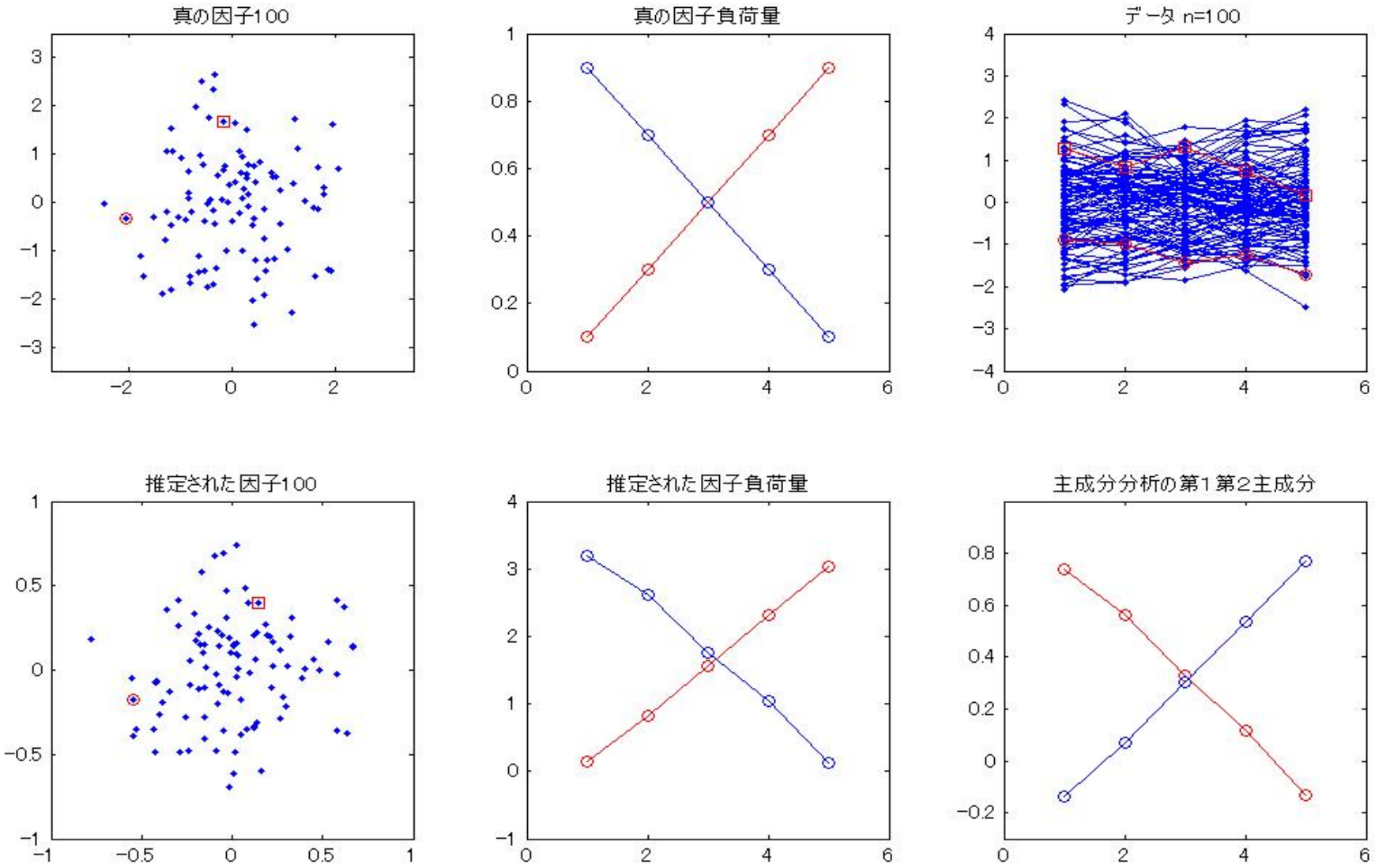
人工データによる例

観測データ: 5次元 n=100



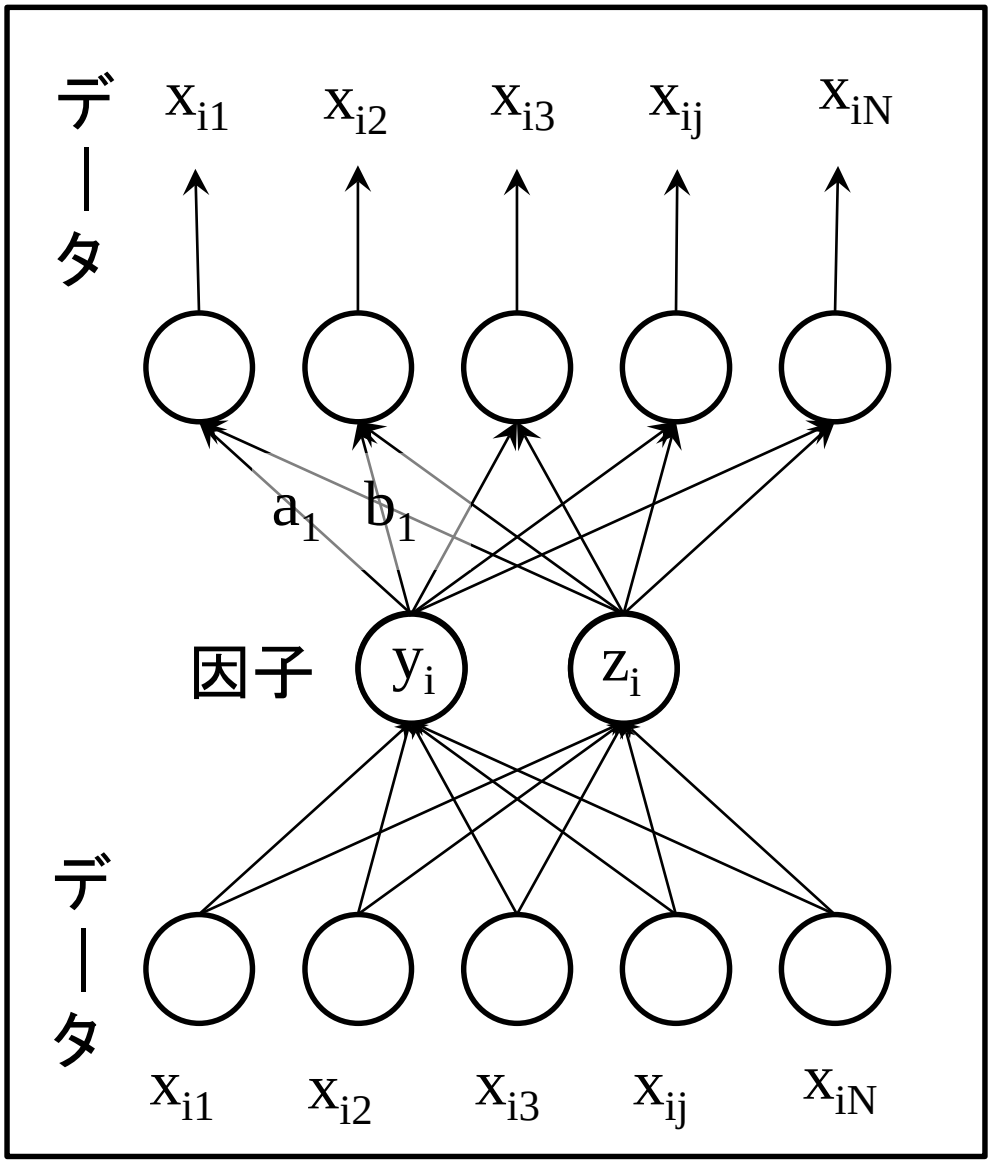


人工データによる例


真の因子100
真の因子負荷量
データ n=100
推定された因子100
推定された因子負荷量
主成分分析の第1第2主成分

人工データによる例

参考:因子分析と主成分分析



$\{ s_j \}$ が一定で $c_j=0$ のとき主成分分析を用いて(a,b)を決めることができる。

主成分分析は値の不定性はないので考えやすい。

(a,b)が決まっても(y,z)を定めるには同様の回転手法が必要になる。

主成分分析では固有ベクトルは直交する。因子分析では回転して解釈しやすいものを探す。

実世界とモデル

いろいろな因子分析法

○　得られたデータに対して

$$X_{ij} = a_j Y_i + b_j Z_i + c_j + R_{ij}$$

を用いる推測法として、cj=0 の場合に、相関行列 $S = (1/n)\Sigma_i X_i X_i^T$
ベクトル a と b　からできる行列 $(a,b)(a,b)^T$と R の分散共分散
行列 (1/s)I から

$$S = (a,b)(a,b)^T + (1/s)I$$

を満たす (a,b) と s を求める方法がある。

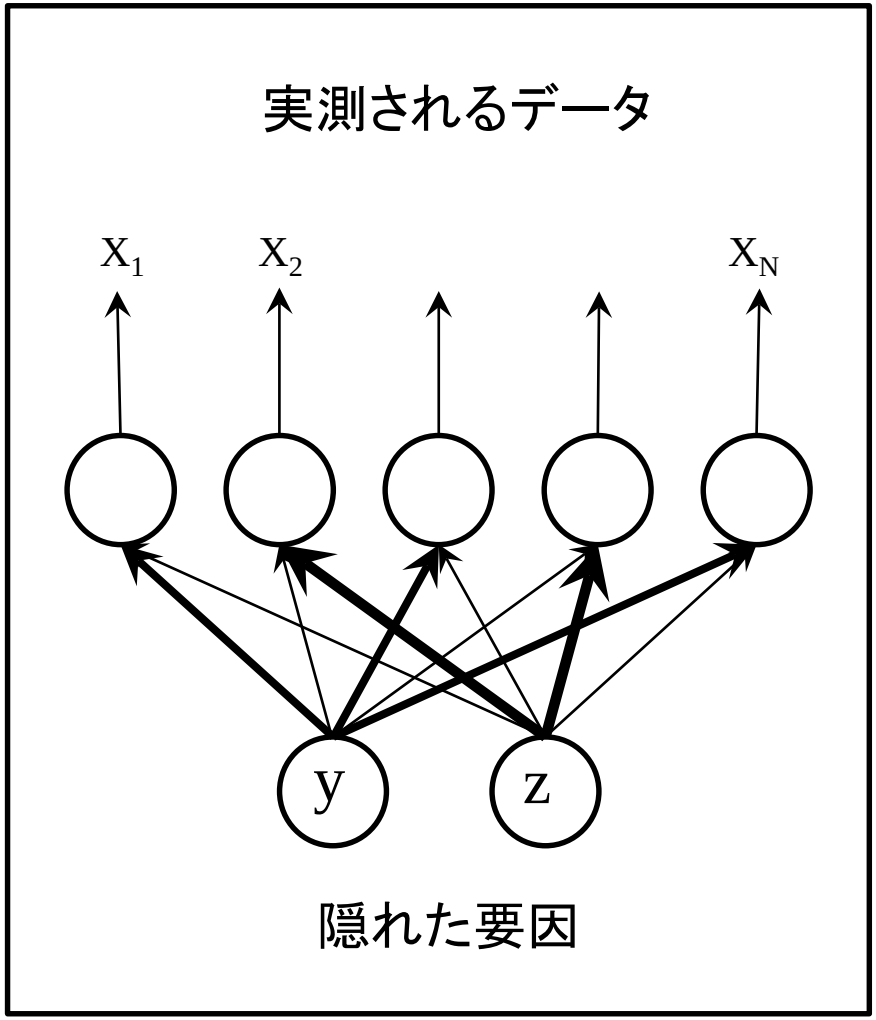
○　回転を行なう場合の「解釈のしやすさ」についてもいろいろな提案があり、直交回転だけでなく斜交回転行なうものも提案されている。

因子分析の方法さまざま

「因子分析」と呼ばれる方法は
データ $(X_1,X_2,\ldots,X_N)$ に対して
要因となる (y,z) を見出そうとする
ものである。

ここでは事後確率最大化法を用いたが、「因子分析」のために設定される条件や推測法にはとても多くのヒューリスティクスがある。

統計的推測を行う場合にヒューリスティクスを用いてはならないということはないが、あまりにヒューリスティクスばかりを用いると得られた結果をどのように評価したらよいかわからなくなるという問題がある。



統計的推測は正しいのだろうか

しかしながらよく考えなおしてみると、そもそも「統計モデル」というものは実世界ではなく、人間が準備したモデルに過ぎない。我々は実世界そのものを知ることはなくデータがあるだけだからである。

？？？　統計モデルとヒューリスティクスには異なる点があるのだろうか。

☆　この講義の後半で説明することになるが、統計モデルを用いる場合には、推測の結果が「実世界に対してどの程度に間違っているか」を理論的に知る方法がある。これと比較して、ヒューリスティクスを用いる方法は、そのようなことが困難であることが多い。

統計モデルとヒューリスティクスの違いは、「正しいか正しくないか」の違いではなく、推測の結果の正しさを知る手段があるかないかの違いだと考えると、第一歩の理解としては正しいだろうと思います。

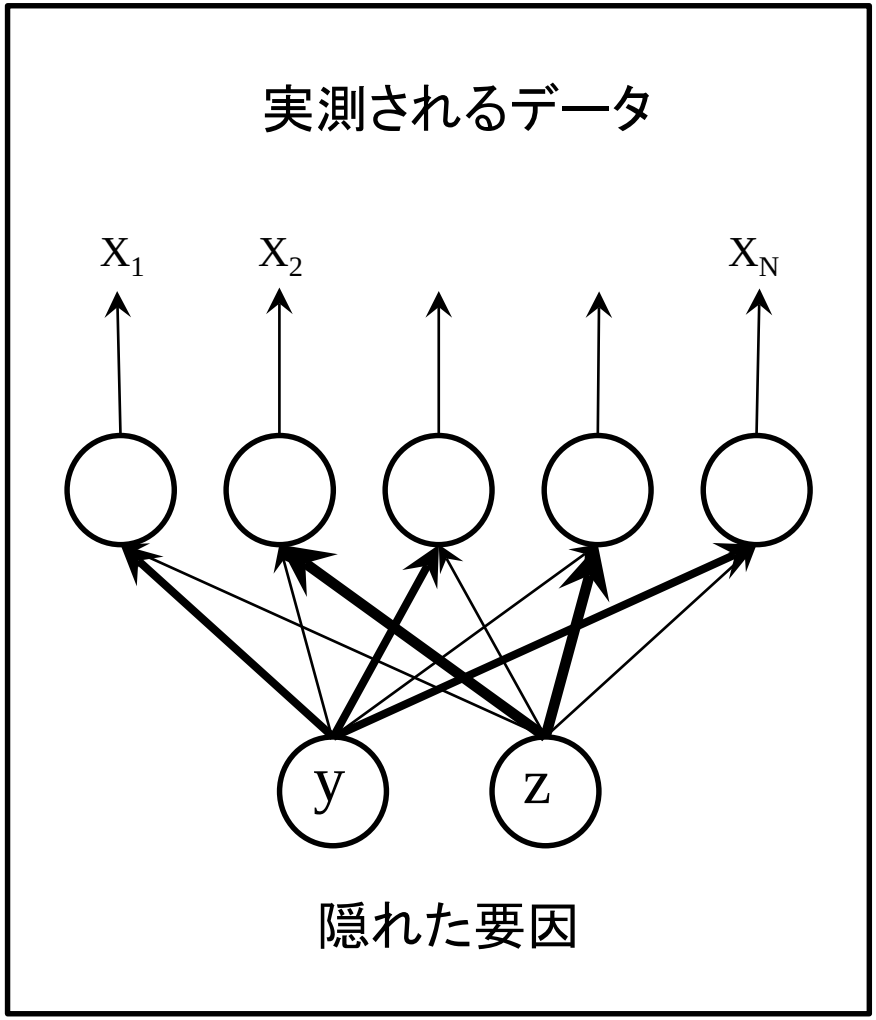
因子分析と似た方法

独立成分解析

因子分析とよくにた方法で研究されているものとして、例えば、(y,z) は正規分布に従うのではなく、他の分布に従っていて、かつ、Y と Z は独立である場合を考えるものがあります。

非負値行列分解

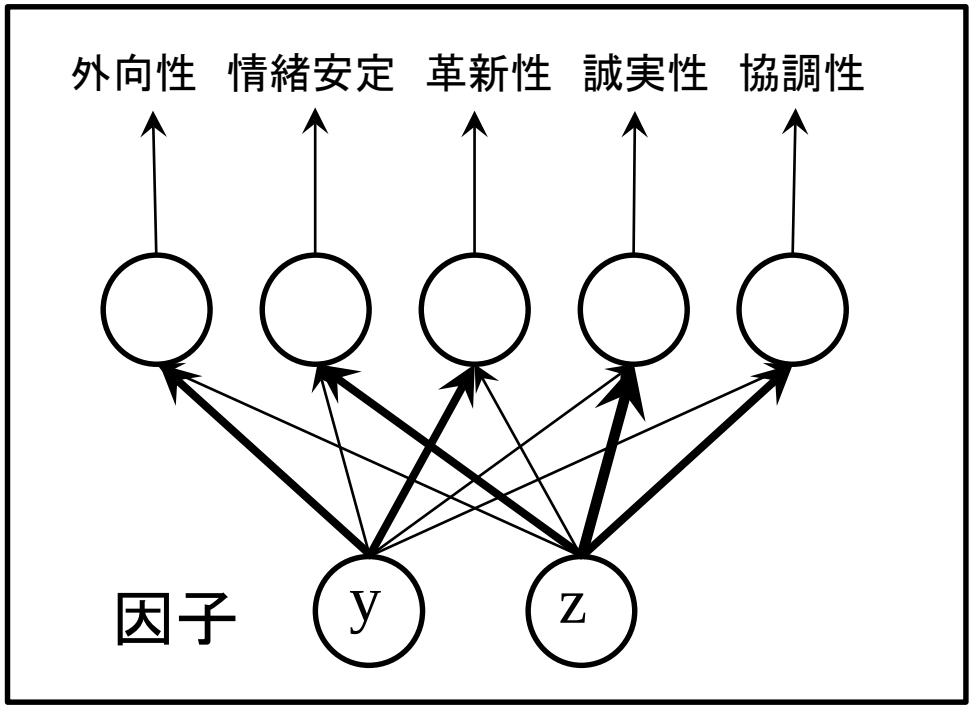
もしも因子負荷行列も因子もどちらも負の値にならないことが分かっている場合には、非負であるという条件をつけることで (Y,Z) を求めようとする方法があります。



例１

（例）よくある例

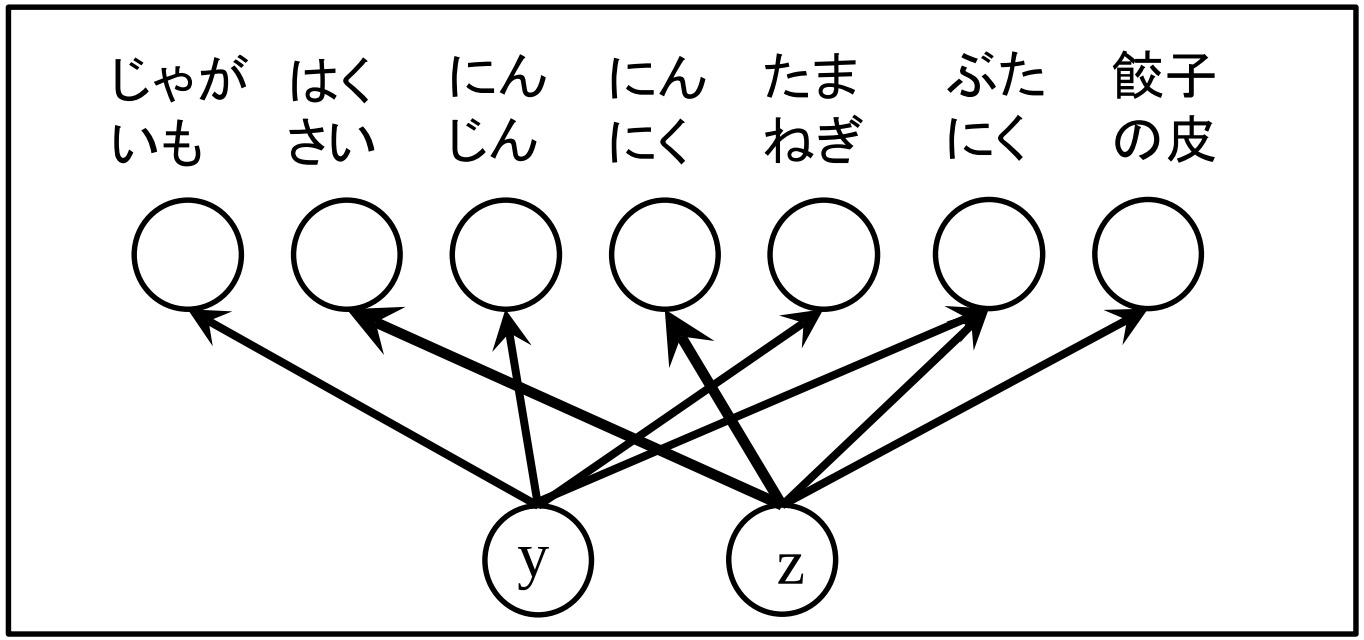
たくさんの人に性格についての質問を行なって5個の特徴について数値データを得た。この5個の特徴を決めている因子が２個あるのではないかと思って解析をしたら因子から5個への結合の強さが図のようになった。(y,z) は何ですか。



(※）この例は仮想例であり実データではありません。

問2

ある家庭で一ヶ月の買い物の種目について因子分析を行なったところ下記のようになった。因子であるｙとｚは何か。



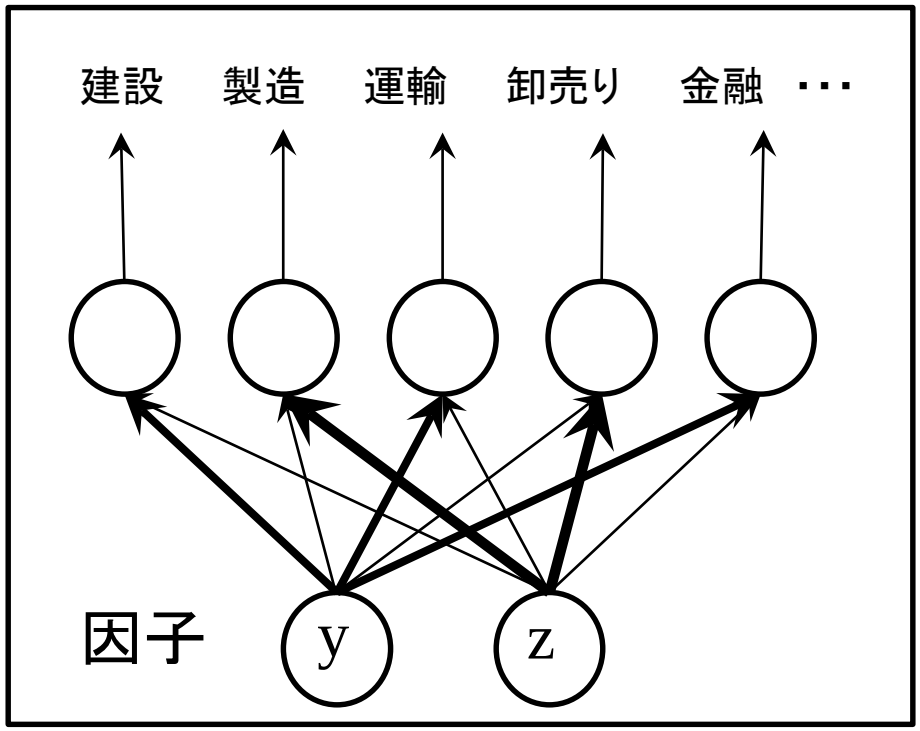
これは因子とは何かを考えるための例です。
正解かどうかは問題ではないので気楽に考えましょう。

(注) 因子が何であるかを考えるためには考察対象の分野を理解している必要があります。

実際のデータ

例2

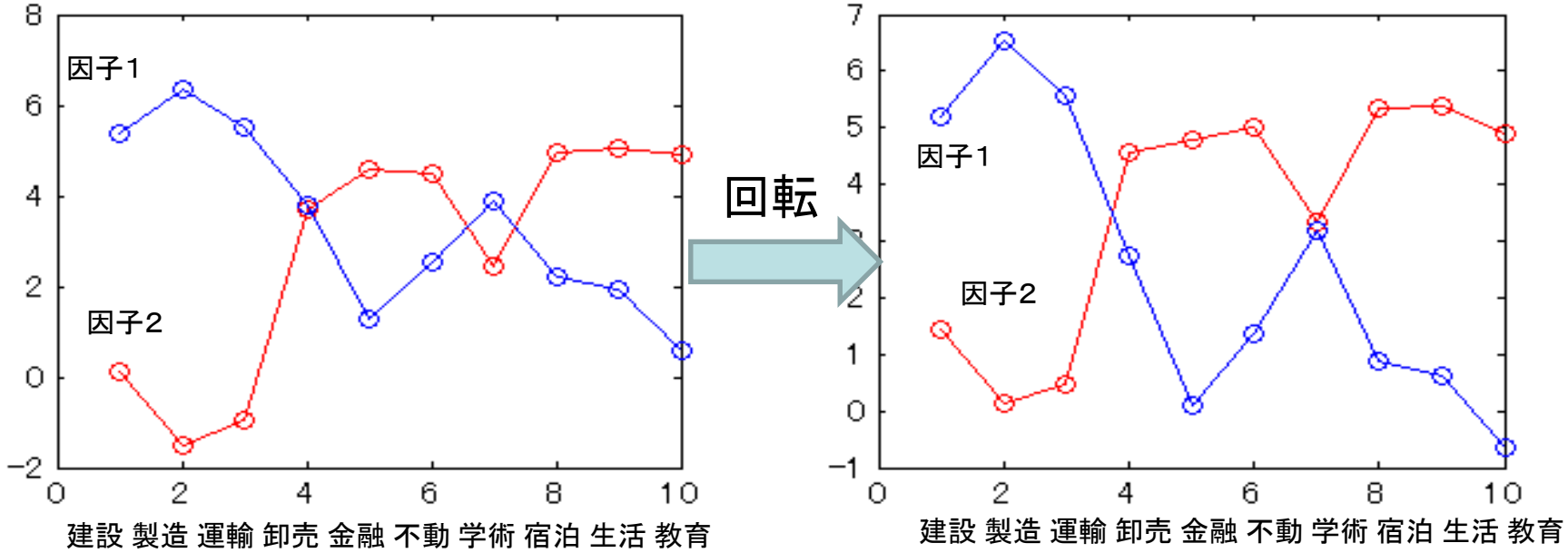
「街の単位」を「◎町〇丁目」とする。
横浜市にある1070個の街で下記の
業種で働く人の人数を調査して
10次元のベクトルを1070個得た。

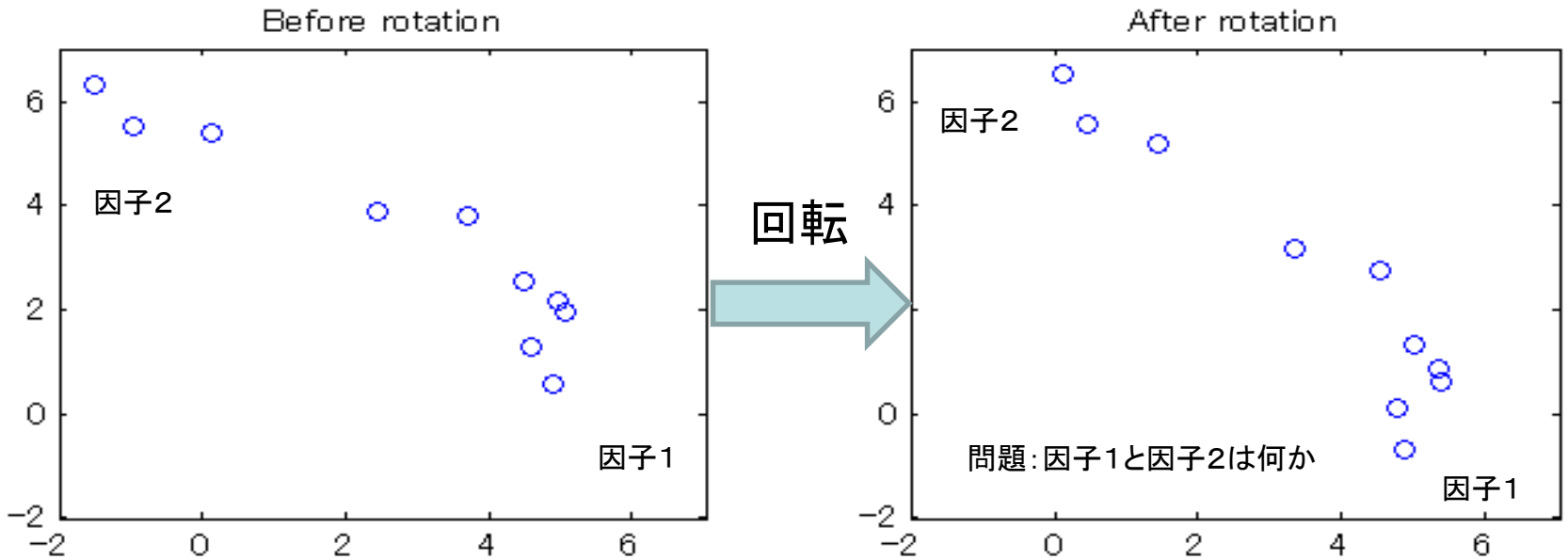


1 建設業
2 製造業
3 運輸郵便業
4 卸売小売業
5 金融保険業
6 不動産物品賃貸業
7 学術研究専門・技術サービス業
8 宿泊飲食サービス業
9 生活関連サービス娯楽業
10 教育学習支援業

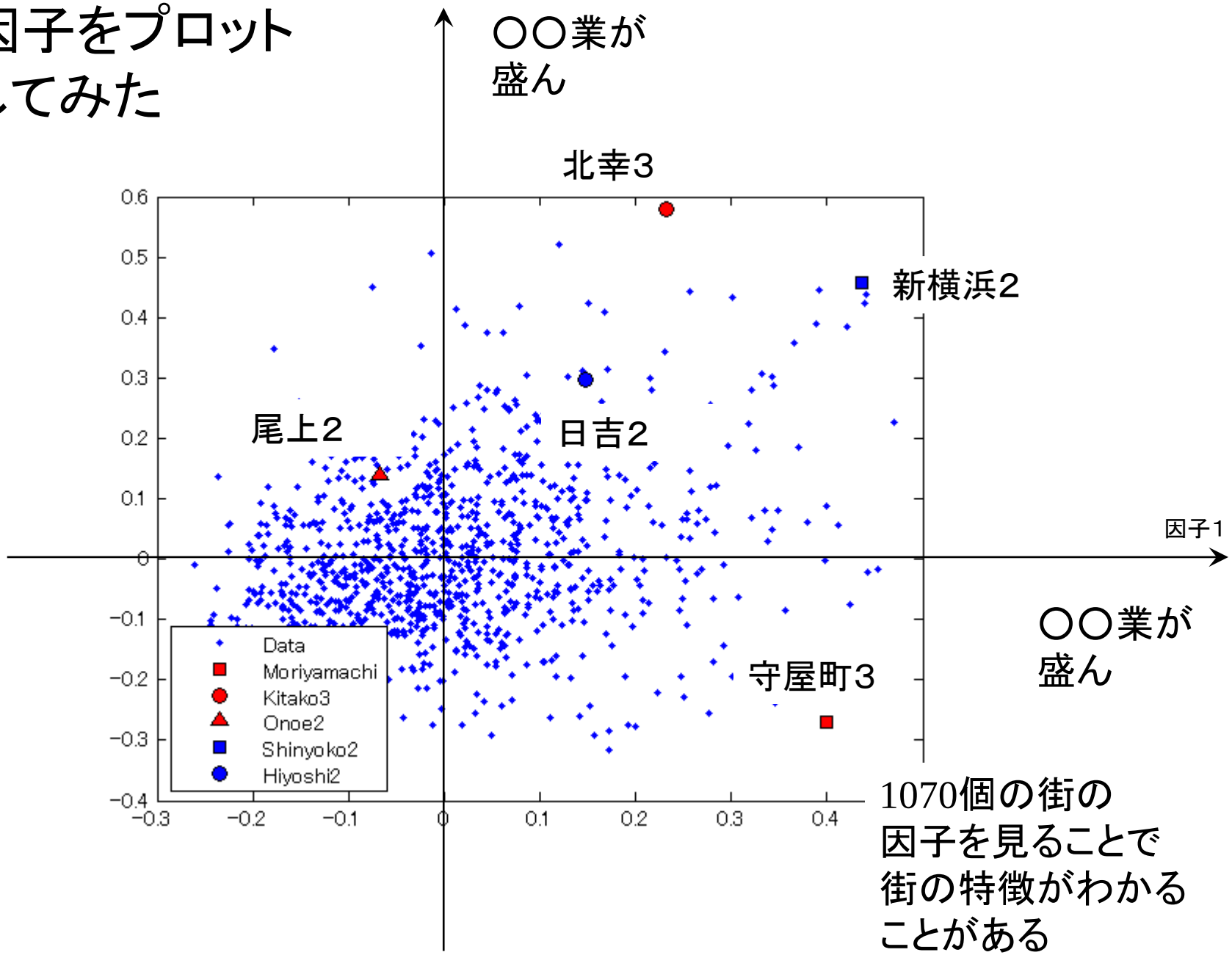
政府統計の総合窓口（e－Stat）を使わせて頂きました。
http://www.e-stat.go.jp/estat/html/spec.html
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

因子負荷量を書いてみた



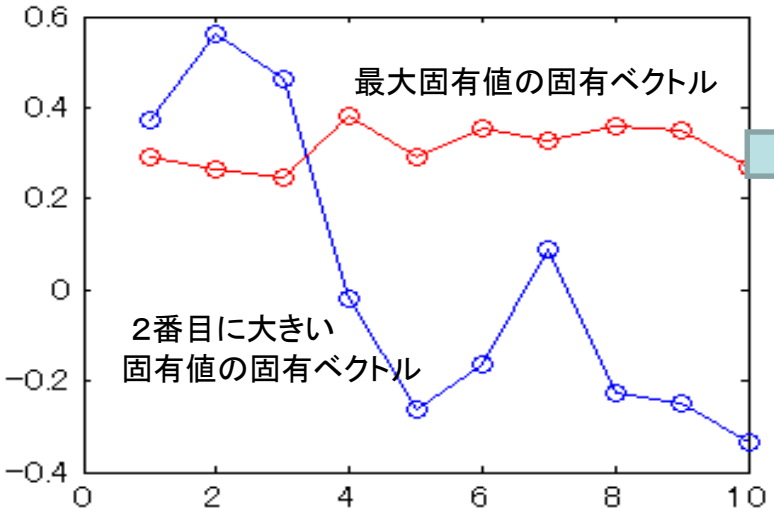


因子をプロットしてみた



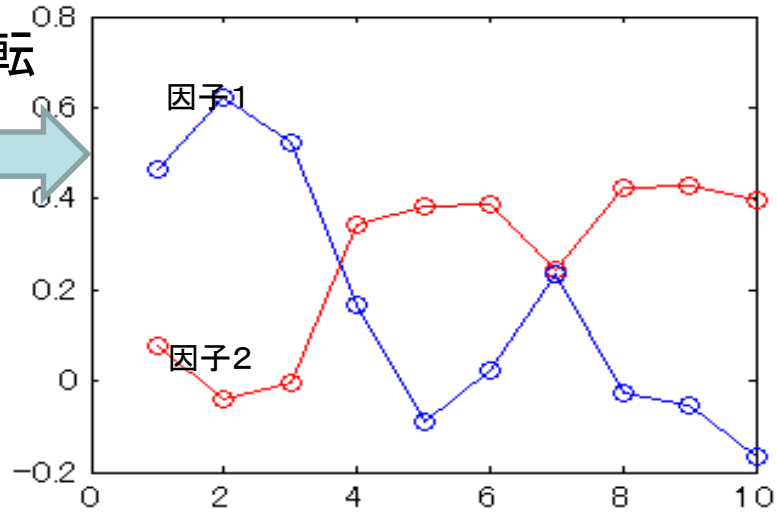
1070個の街の
因子を見ることで
街の特徴がわかる
ことがある

参考：主成分分析から得られたベクトル

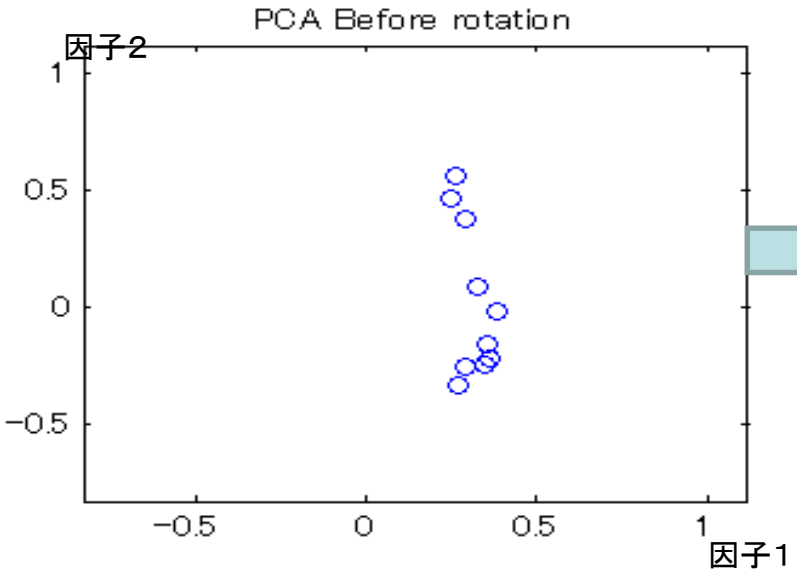


建設 製造 運輸 卸売 金融 不動 学術 宿泊 生活 教育

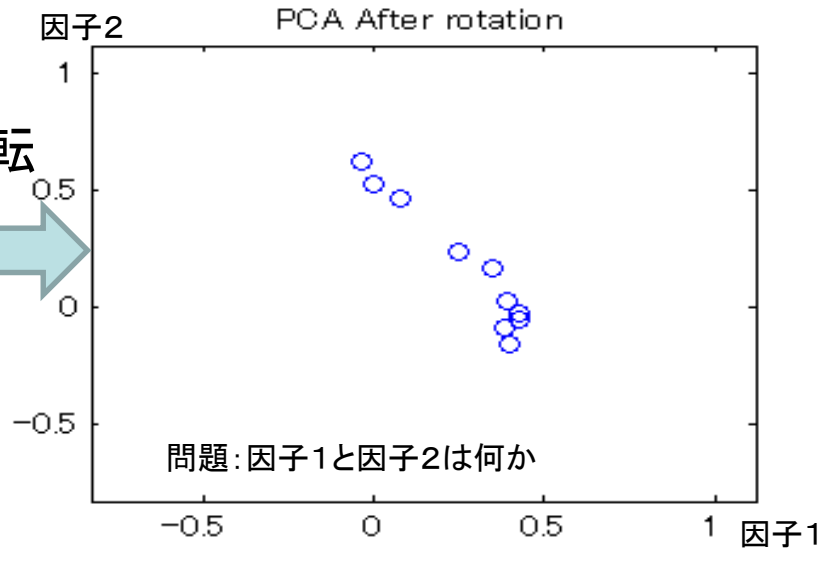
回転



建設 製造 運輸 卸売 金融 不動 学術 宿泊 生活 教育



回転



注意

○ 応用上、因子分析は主成分分析と似た目的で用いられることが多く、
また解析の結果も似ていることが多い。

○ しかし主成分分析は x の情報をできるだけ取り出す部分空間の発見で
あるのに対して、因子分析は「y→x」となる要因 y を見つけることで
あって統計学的な意味は異なる。

○ 主成分分析では固有ベクトルは互いに直交する。
因子分析では因子は解釈のしやすさを基準にして回転する。

○ 因子は通常は2個で解析される。3個以上でもできるが、回転
の最適化などの問題の複雑度が増す。

○ 因子の意味は 因子負荷量を見て x との関係から考察する。
適切な言語化ができるとよいのであるが。

○ 因子は、因果関係における原因を表しているわけではない。

問3

横浜市にある1070個の街の業種に関して
得られた二つの因子が何であるかを考察せよ。