データ解析

第4回: 主成分分析

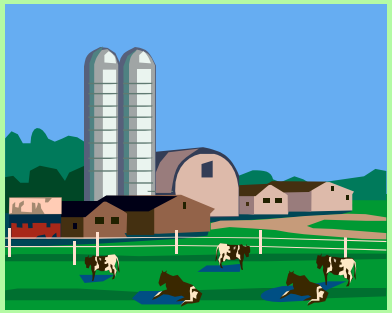
渡辺澄夫

名前のない世界へ

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

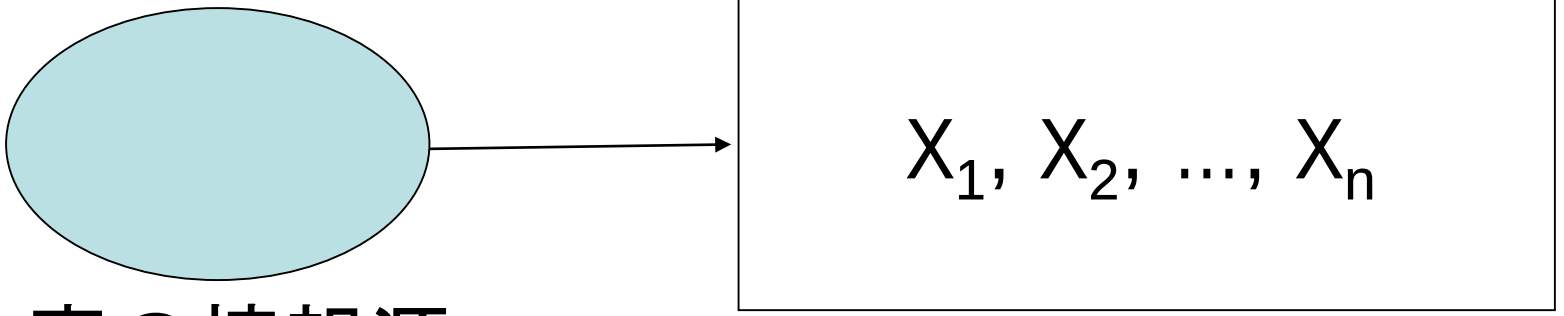
データの背後に何を想像するか

ラベルがないデータがあるとき、その背後に存在する構造として
人間が想像するものには、いくつかの典型的なものがある。

ラベルを持たないデータ

回帰分析・判別分析では、情報 X に対するラベル Y が用意されていて、推論「X→Y」を推測した。世の中にはラベルのないデータが圧倒的に多くそのデータについての構造を抽出したいという要望がたくさん存在する。

真の情報源について
「何か」を知りたい



真の情報源

データの背後にある構造としてよく考察されるもの

ラベルを持たないデータがたくさんあるとき、そのデータの背後に存在する構造として想像されることが多いものがいくつかある。

1. データは低次元線形空間の上にある。→ 主成分分析
2. データを少ない個数の要因で説明する。→ 因子分析
3. データは幾つかのクラスタに分けられる。→ クラスタ分析

この講義では、1、2、3について学びます。この他にも例えば次の問題を考えたい人もあるかもしれませんが、この講義では学びません。

4. データ文字列を発生した確率文法を推定したい。→ 確率文法の推定
5. データの背後に人間の意志があるかどうか判定したい。→ できるの？

いつか再び出会う

ラベルを持たないデータがたくさんあるとき、そのデータの背後に存在する構造を推測することは、今日の人工知能やデータサイエンスの主要な研究対象です。

1. データは多様体の上にある。→自己組織化写像、砂時計型深層学習
2. データを少数の要因で説明できる。→ 独立成分解析、非負値行列分解
3. データは幾つかのクラスタに分けられる。→ 変分ベイズ法、階層クラスタ法
4. データを発生した確率文法を推定したい。→ 構文解析変分ベイズ法
5. データの背後に人間の意志があるかどうか判定したい。→ できるの？

上記のように高度なものはこの講義ではやりませんが、みなさんは将来、研究や実務で出会うかもしれないですね。

主成分分析とは

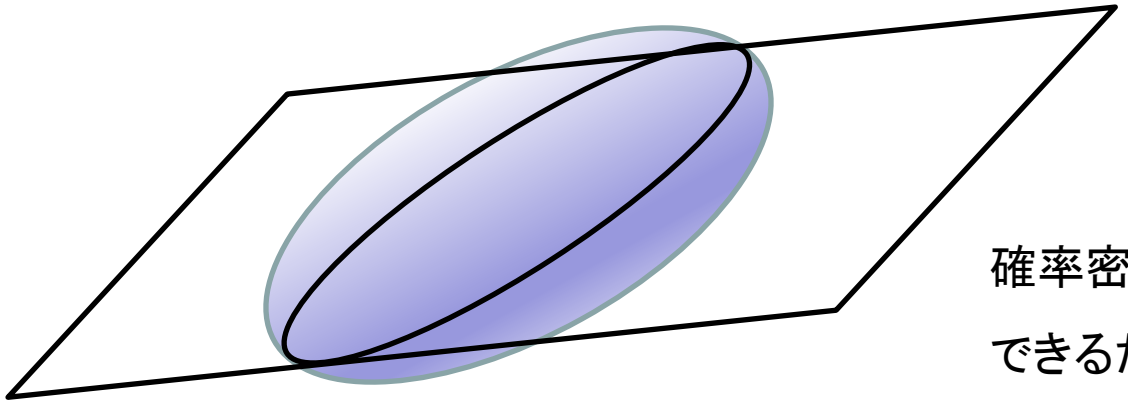
N次元データが、ほぼ低次元の線形空間上にあるとき、
その線形空間を抽出してみよう。

主成分分析で何をしたいか

問題．N次元実ユークリッド空間に値をとる確率変数 X が確率密度関数 q(x) に従うとし、k を 1 以上N未満の自然数とする。N × N 行列 A で条件「 rank A ≦ k 」を満たすものに対して

$$E(A) = \int || x - Ax ||^2 \, q(x) \, dx$$

を最小にする A を求めたい。



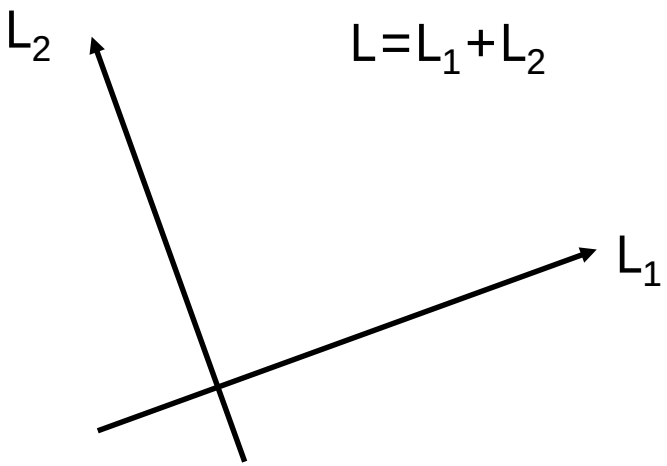
確率密度関数　q(x) の情報を
できるだけ残せるような低次元の
部分空間を抽出したい。

準備：直交補空間

N 実ユークリッド空間 L を考える。L に内積 (,) が定義されている。L の部分空間 L_1 が与えられたとき

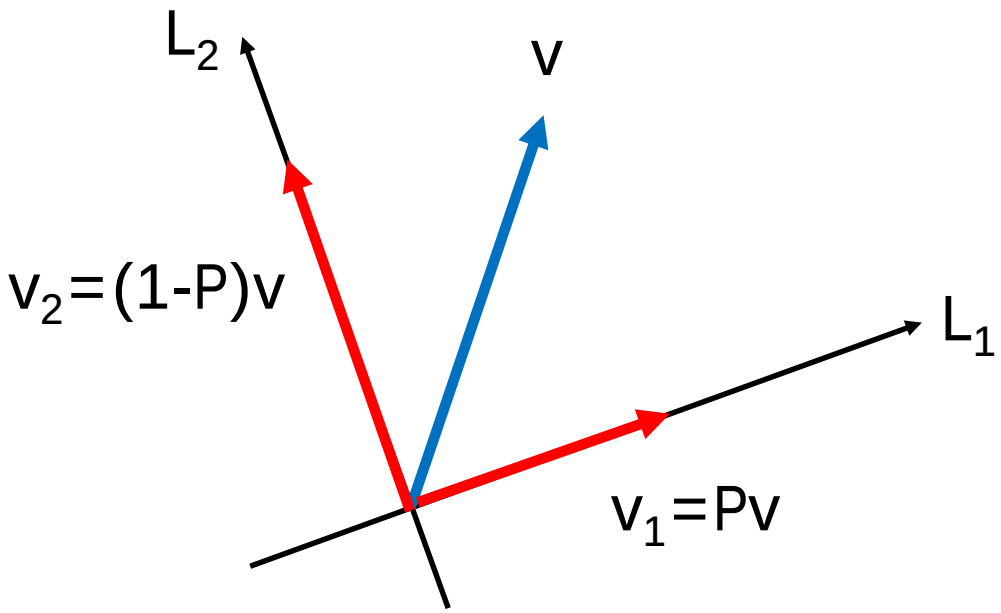
$$L_2=\{\ v\in L\ ;\ (v,u)=0\ (\forall u\in L_1)\ \}$$

は L の部分空間である。L_2 をL_1の直交補空間という。このとき L_2 の直交補空間は L_1 である。$L=L_1(+)L_2$ を直和分解という。



準備：射影

直和分解 $L=L_1(+)L_2$ が与えられたとき、L の任意の元 v はある$v_1 \in L_1$ とある$v_2 \in L_2$ がそれぞれユニークに存在して $v=v_1+v_2$ と書ける。ここで v から v_1 への写像と v からv_2 への写像は線形写像である。$v_1=Pv$ と書き P を L_1 への直交射影という。このとき $v_2=(1-P)v$ が成り立つ。(1-P)は L_2への直交射影である。定義から $P=P^T$, $P^2=P$ が成り立つ。

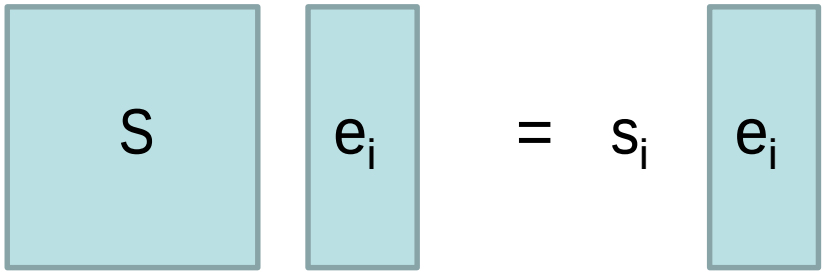


準備：対称行列と固有値

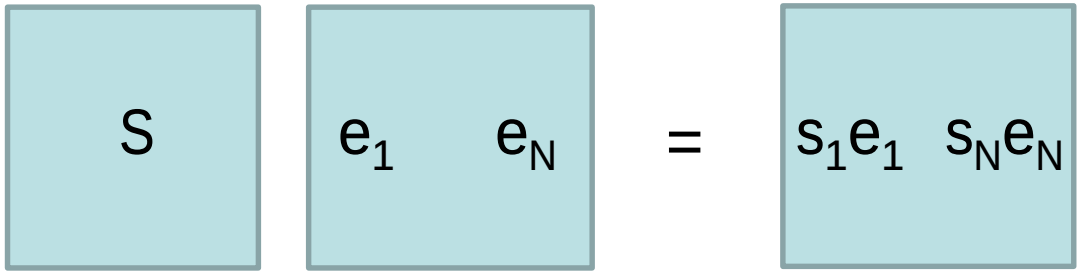
N × N 実行列 S が $S=S^T$ を満たすとき S を対称行列という。対称行列 S が与えられたとき、実数の固有値 $\{s_1,s_2,\ldots,s_N\}$ と正規直交基底$\{e_1,e_2,..,e_N\}$（$(e_i,e_j)=\delta_{ij}$ を満たすもの）が存在して

$$S\, e_i = s_i\, e_i \qquad (1 \leqq i \leqq N)$$

が成り立つ。これを絵で書くと



これを$\{e_i\}$ について並べて書くと

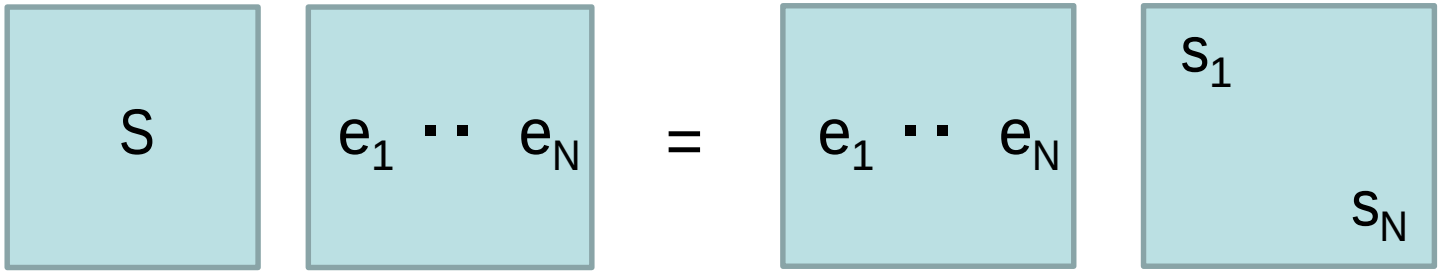


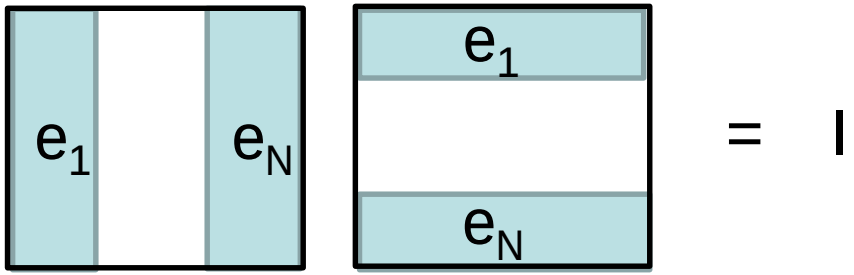
準備：対称行列の対角化

これより、$e_1,e_2,..,e_N$ を横に並べた行列を $U=[e_1,e_2,...,e_N]$ と書き、対角成分が $s_1,s_2,...,s_N$ であり非対角成分が 0 の行列を $D=D(s_1,s_2,...,s_N)$ とかくと

$$S\,[e_1,e_2,..,e_N]=[e_1,e_2,..,e_N]\,D(s_1,s_2,...,s_N)$$

従って、$S = U D U^T$ が成り立つ。ここで U が直交行列であることから $U^T=U^{-1}$ であることを用いた。（「対称行列は直交行列で対角化できる」という）。





準備:トレースと射影

1. $\{ e_1, e_2, .., e_N \}$ を一般の正規直交基底とする。このとき任意のN×N 行列 A について次式が成立する。

$$tr(A) = \Sigma_i (e_i, Ae_i).$$

任意の直交射影 P について $(e_i, Pe_i) = || Pe_i ||^2$ から $0 \leqq (e_i, Pe_i) \leqq 1$.

また直和分解 PL(+)(1-P)L を考え、PL および (1-P)L の正規直交基底をとることにより

$$tr(P) = \Sigma_i (e_i, Pe_i) = \text{rank } P.$$

2. 正規直交基底 $\{ e_1, e_2, .., e_N \}$ が与えられたとき、$\{ e_1, e_2, .., e_k \}$ によって生成される部分空間への直交射影 P は

$$P\, v = \Sigma_{i \leqq k}\, e_i\, (e_i, v)$$

である。

最適な線形写像は直交射影である

補題1. 条件「 rank A ≦ k 」を満たす A で E(A) を最小にする A は直交射影である。

(証明) A のランクは k 以下だから、A の値域 L_1 は k 次元以下の線形空間。L_2 をL_1の直交補空間として直和分解 $L=L_1(+)L_2$ を考える。

$$x = x1+x2$$

$$x - Ax = (x_1-Ax) + x_2$$

$$|| x - Ax ||^2 = ||x_1-Ax||^2+||x_2||^2$$

従って

$$E(A) = \int || x - Ax ||^2 q(x)\, dx = \int \{ ||x_1-Ax||^2+||x_2||^2 \} q(x)\, dx$$

これが最小になるのは、任意の x について $Ax=x_1$ がなりたつとき、つまり A が直交射影のときである。(証明終)

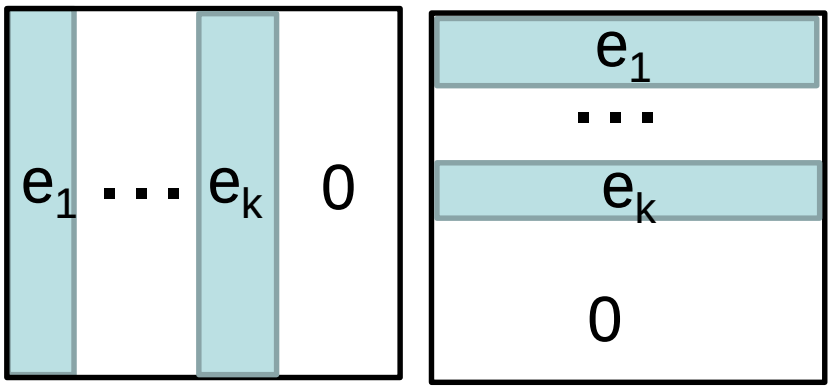
主成分分析の定理

定理. N×N行列 A で条件「 rank A ≦ k 」を満たすものに対して

$$E(A) = \int || x - Ax ||^2 \, q(x) \, dx$$

と定義する。対称行列 S を $S = \int x \, x^T \, q(x) \, dx$ によって定義する。

S は正定値行列なので固有値はすべて正である。S の固有値を $0 < s_N < s_{N-1} < \cdots < s_1$ とし、固有値 s_i に対応する固有ベクトルを e_i とする ($\{e_i\}$ として正規直交基底を取る)。E(A) を最小にする A は $\{e_1, e_2, \ldots, e_k\}$ によって生成される線形部分空間への正射影 $A\,x = \Sigma_{i \leqq k} \, e_i \, (e_i, x)$ である。

$A =$ 

(注意)$\{e_1, e_2, \ldots, e_k\}$ の順番については確定しない。なお、固有値の中に等しいものがあり $s_k = s_{k+1}$ の場合にはどちらの固有ベクトルを用いてもよい。

証明

（証明）$C= \int ||x||^2\, q(x)\, dx$ と書く。補題1からAは射影であるから $A^2=A$.

$$\int (x,A^2x)\, q(x)\, dx = \int (x,Ax)\, q(x)\, dx = \int \mathrm{tr}(A\, xx^T)\, q(x)\, dx = \mathrm{tr}(AS)$$

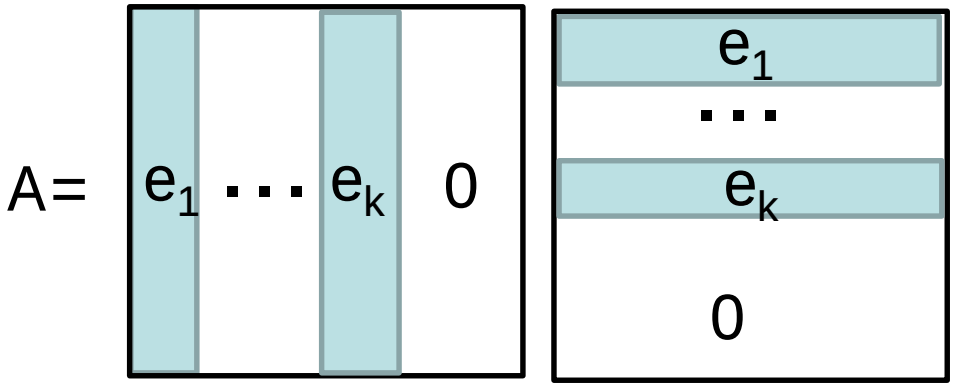
であるから $E(A)= C - \mathrm{tr}(AS) = C - \Sigma_i\, s_i\, (e_i, Ae_i) = C - \Sigma_i\, s_i\, a_i$. ここで $a_i=(e_i,Ae_i)$ とかいた。さて A は射影であるから $0 \leqq a_i \leqq 1$, $\Sigma_i\, a_i \leqq k$ が成り立つ。従ってE(A)の最小化は $\Sigma_i\, s_i\, a_i$ の最大化と同じである（$\{a_i\}$ が変数）。

$$\begin{aligned} \Sigma_i\, s_i\, a_i &\leqq \Sigma_{i \leqq k}\, s_i\, a_i + s_{k+1}\, (\Sigma_{i > k}\, a_i) \\ &\leqq \Sigma_{i \leqq k}\, s_i\, a_i + s_{k+1}\, (k - \Sigma_{i \leqq k}\, a_i) = k s_{k+1} + \Sigma_{i \leqq k}\, (s_i - s_{k+1})\, a_i \end{aligned}$$

であり $(s_i-s_{k+1})>0$ だから、最後の式の最大値は $\Sigma_{i \leqq k}\, s_i$ である。特別な場合として「a_i = if(i≦k)1, else 0 」を考えれば上記の等号が満たされる。そのとき以外では等号は成立しない。(証明終）

主成分分析は貪欲 (greedy) 計算でよい

証明した定理から、rank A ≦ k において最適な A は次のものであることがわかった（ただし$\{e_1,e_2,\ldots,e_k\}$の順番は確定しない）。



ここで k=1,2,3,…の順番で E(A) を最小化する A を求める問題を考えて一個ずつ$\{e_1,e_2,\ldots,e_k\}$を最適化していっても結果は同じであるから、一回固有値問題を解いておけば、k が変わるごとに固有値問題を解きなおす必要はない。

データからの推定

主成分分析は任意の分布に対して適用できる･･･。

データを用いて推測する

実世界では q(x) はわからないのでデータ $\{X_i ; i=1,2,\ldots,n\}$ を用いて

$$E(A) = (1/n) \Sigma_i \| X_i - AX_i \|^2$$

を最小にする A を探すことになる。これを主成分分析という。

データの数 n が無限大に近づくとき、対数の法則

$$(1/n) \Sigma_i \| X_i - AX_i \|^2 \rightarrow \int \| x - Ax \|^2 q(x)\, dx$$

が成り立ち、これを最小にする A も確率収束する。

主成分分析は座標不変ではない

主成分分析はどのようなデータに対しても適用できるが解析結果の
適切な解釈を行うためには、データの性質を予め確認する必要がある。

情報 $x=(x_1,x_2)$ に対して座標軸ごとの定数倍 (ax_1,bx_2) を考えると
主成分分析の結果は、(a,b) の値によって異なるものになる。
たとえば（身長、体重）に主成分分析を行うとき、単位として何を使うか
によって主成分分析の結果は異なることになる。

この問題が起きないようにするには、各軸ごとに平均 m_j と標準偏差 s_j
を求めて $(x_j-a_j)/s_j$ を使うとよいかもしれない。

また、データが正規分布のような山形から外れている場合には
そのような形になるように前処理を施してから解析するほうがよい
可能性がある。

解析した後、考察に用いる量は

1. 各データ x_i について 第一主成分(e_1,x_i), 第2主成分(e_2,x_i), ...を計算して散布図上に図示してみる。これは高次元空間にあって直接的に見ることができないデータの様子の可視化である。

2. 因子負荷量 $(s_1/\Sigma s_i, s_2/\Sigma s_i, \ldots)$ を計算すると、各主成分が全体の中でどの程度の重みで役割を果たしているかが数量的にわかる。

3. 各主成分の意味を考えるには固有ベクトルの x の要素との関係の強さを見る。考察のもとになる情報は x の各成分であるから、その間の関係の強さから主成分の意味を考える。うまく主成分に名前をつけられると良いのであるが。

主成分分析の手順

実世界では q(x) はわからないのでデータ $\{X_i ; i=1,2,\ldots,n\}$を用いて

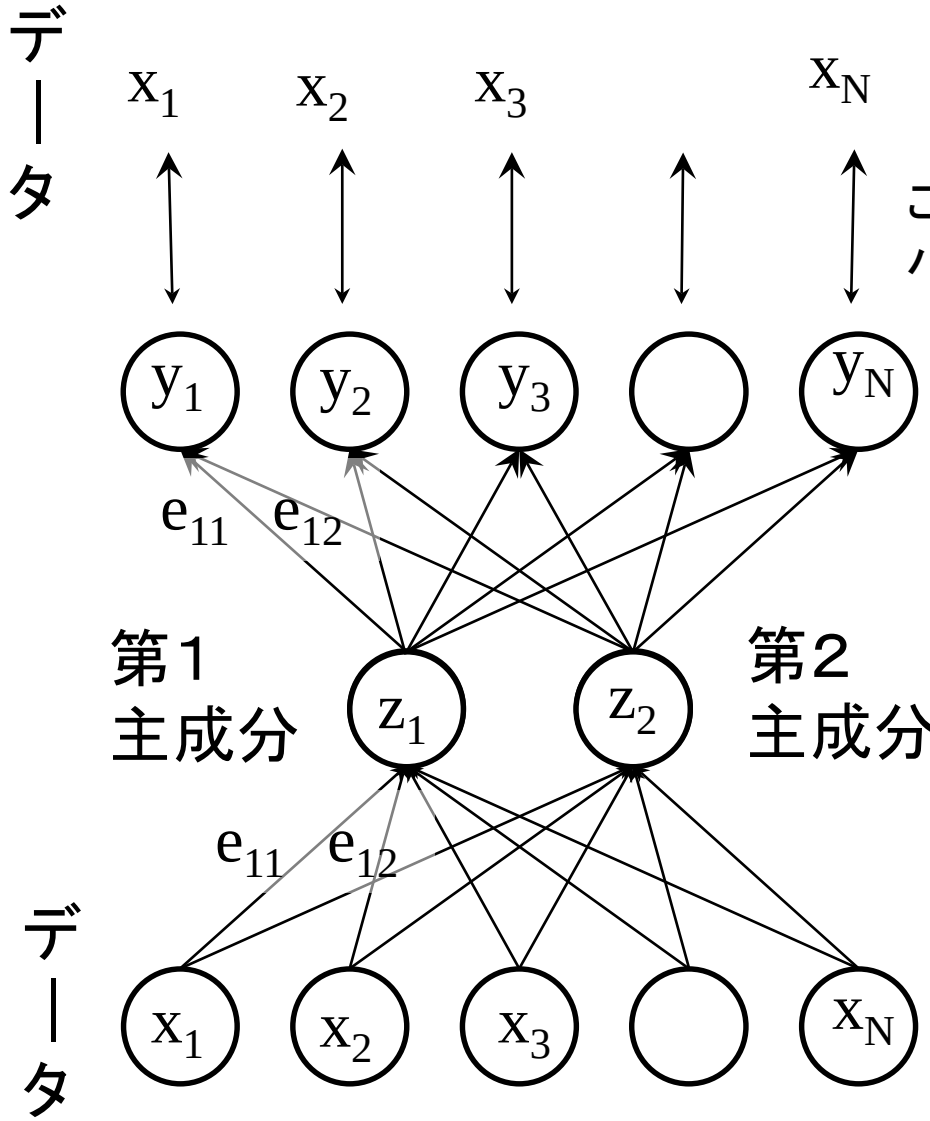
$$E(A) = (1/n) \Sigma_i || X_i - AX_i ||^2$$

を最小にする A を探すことになる。これを主成分分析という。

主成分分析の手順:

0. データ$\{X_i ; i=1,2,\ldots,n\}$に前処理を行う。
1. $S^* = (1/n) \Sigma_i X_i (X_i)^T$ を計算して推測値 S^* を求める。
2. S^* の固有値($s_1 > s_2 > \cdots > s_n > 0$) と対応する固有ベクトルからなる正規直交規定$\{e_1,e_2,\ldots,e_k\}$ を見つける。
3. 固有値から因子負荷量 $(s_1/\Sigma s_i, s_2/\Sigma s_i, \ldots)$ を計算。
4. 各データ x_i について 主成分第1(e_1,x_i), 第2(e_2,x_i), ...を計算。
5. 固有ベクトルを見て各主成分の意味を考える。

主成分分析の構造



ここの誤差が最小になるように
パラメータを決めた

固有値 s_1 に対する
$e_1=(e_{11},e_{12},\ldots,e_{1N})$

固有値 s_1 に対する
$e_2=(e_{21},e_{22},\ldots,e_{2N})$

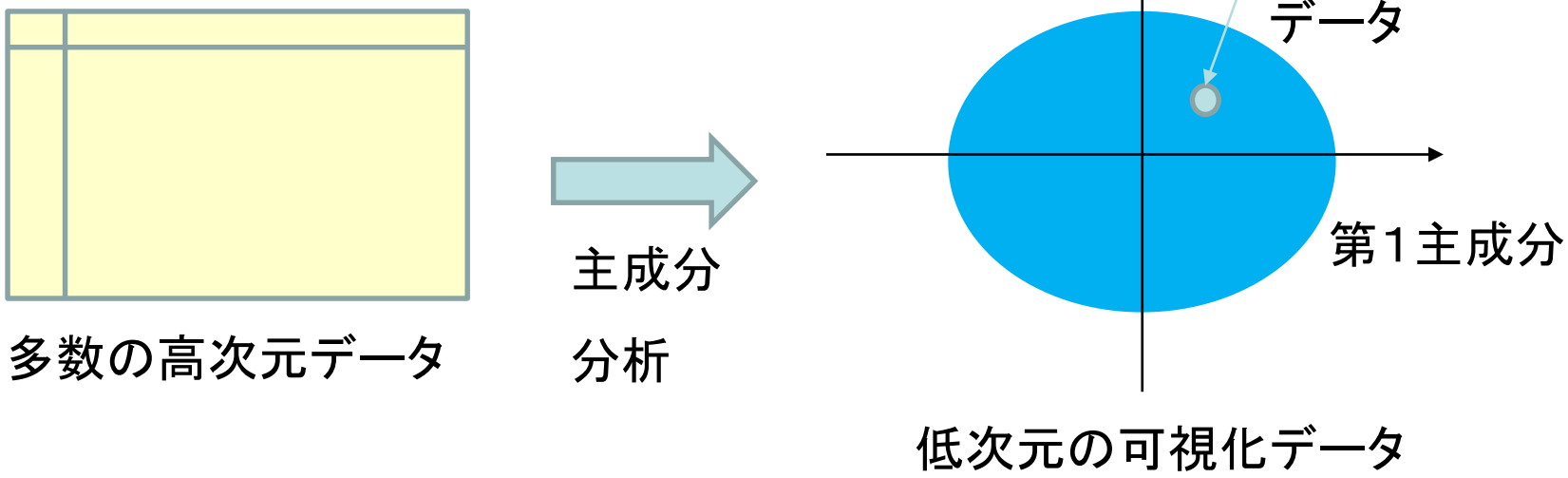
主成分分析は x を次で近似:
$y = e_1(e_1,x) + e_2(e_2,x)$

第1主成分 $z_1=(e_1,x)$

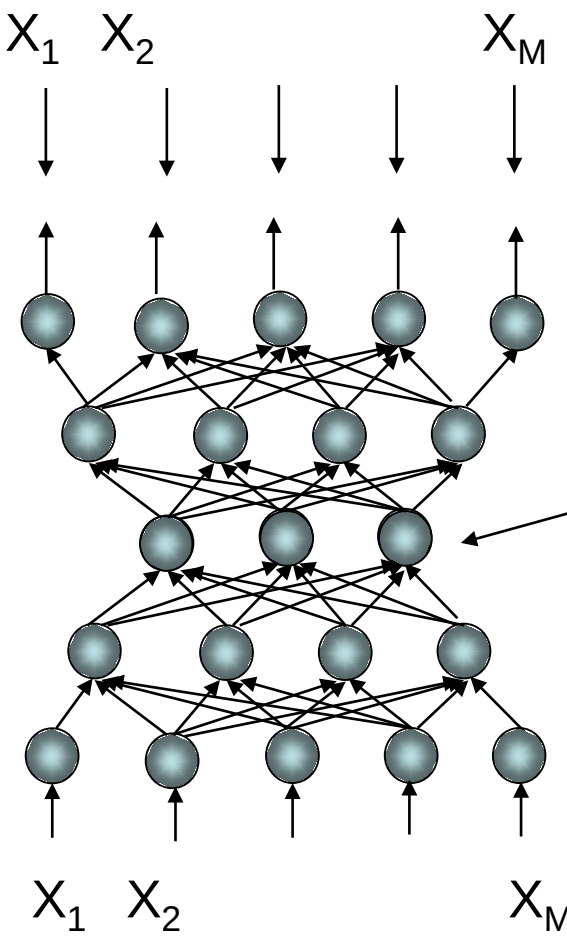
第2主成分 $z_2=(e_2,x)$

主成分分析は解析の第１段階

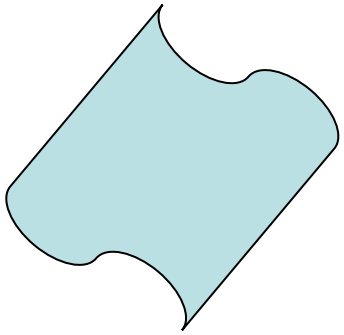
ラベルのないデータがあったとき、主成分分析はそのデータのおおよその性質を調べる目的で、第１段階として用いられることが多いと思います。
すなわち、主成分分析はゴールではなく、スタートであって、主成分分析の結果からどのような仮説を立て何をしていくかの計画を立てることがデータサイエンティストの仕事であろうと思います。



参考：砂時計型ニューラルネット



M 次元空間内の3次元多様体



関数 f(g(x,w),u) でw とu を最適化。
中間の次元を小さくする。

入力よりも少ない個数の中間ユニットを
設定することで、入力が作る低次元
多様体の表現が中間ユニットに
作られる・・・と期待される。

（Auto-Encoder には別の型もあります）

実際の例

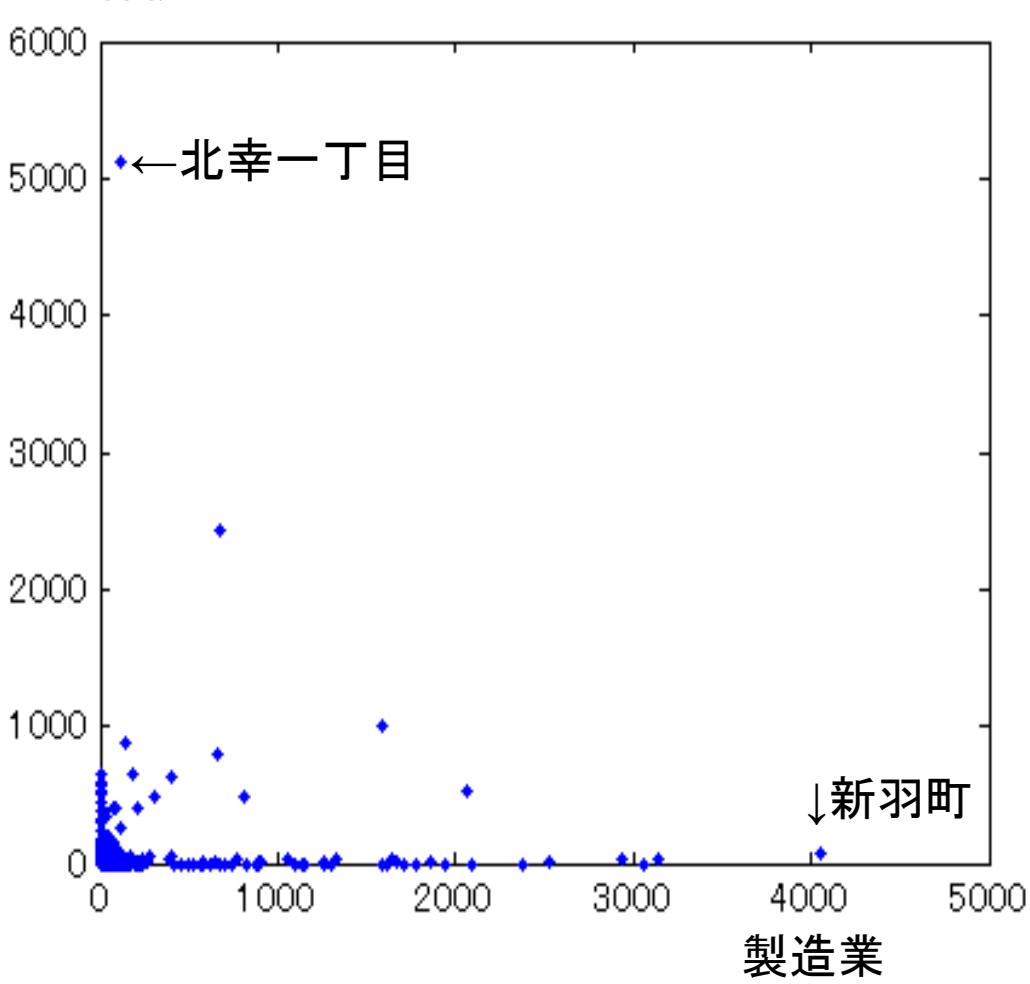
実際の例(1)

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

横浜市の n=1070 の「〇町〇丁目」について下記の業種で働く人の数を調べた。

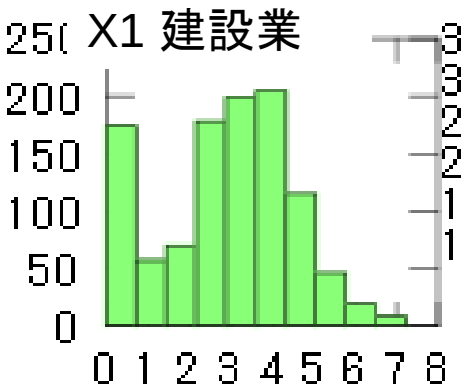
X1 建設業
X2 製造業
X3 運輸郵便業
X4 卸売小売業
X5 金融保険業
X6 不動産業
X7 学術技術業
X8 宿泊業
X9 生活関連業
X10 教育業

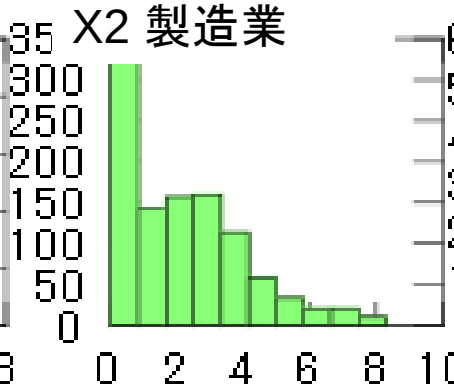
町の規模によって働く人の数は大きく異なるのでlog(1+X_k) と

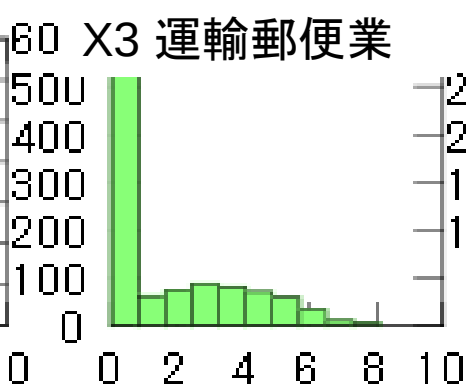


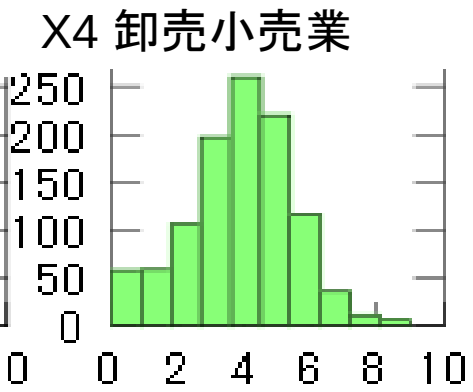
実際の例(2)

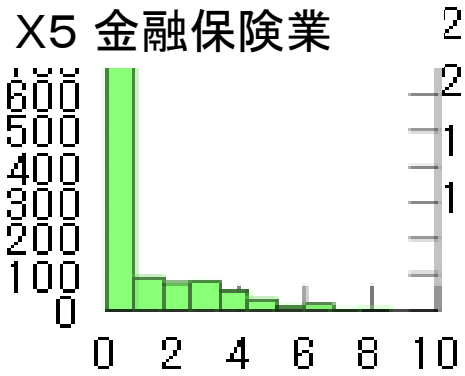
$\log(1+X_k)$ のデータのヒストグラム

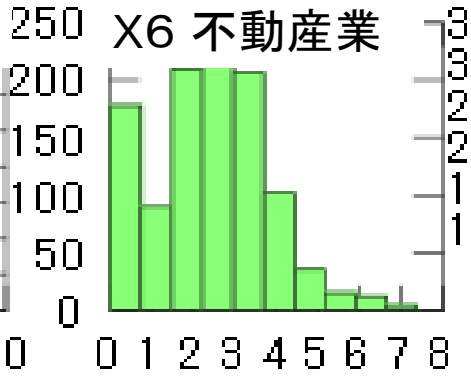


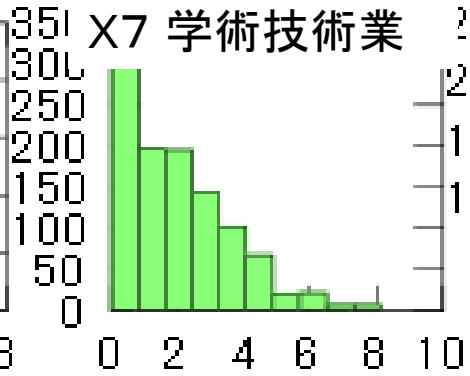


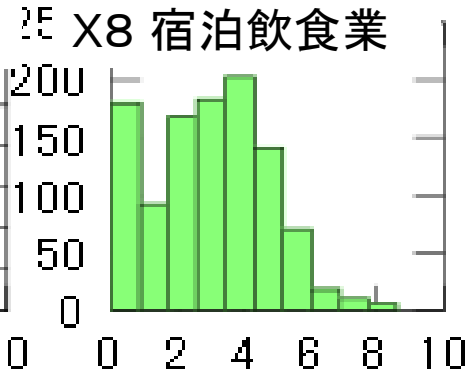


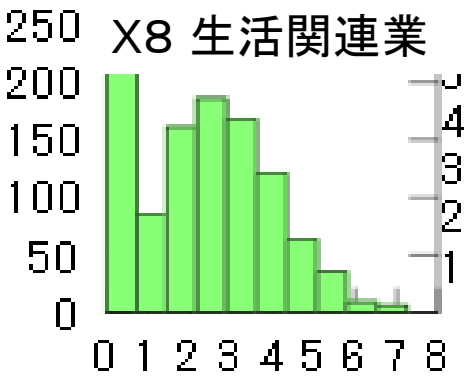


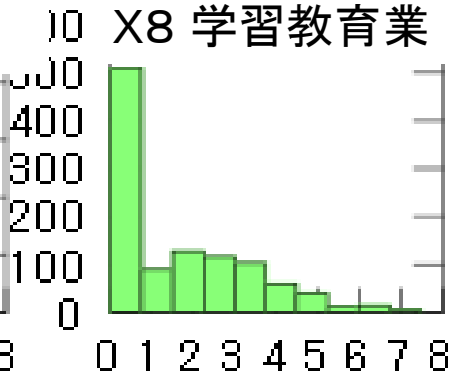






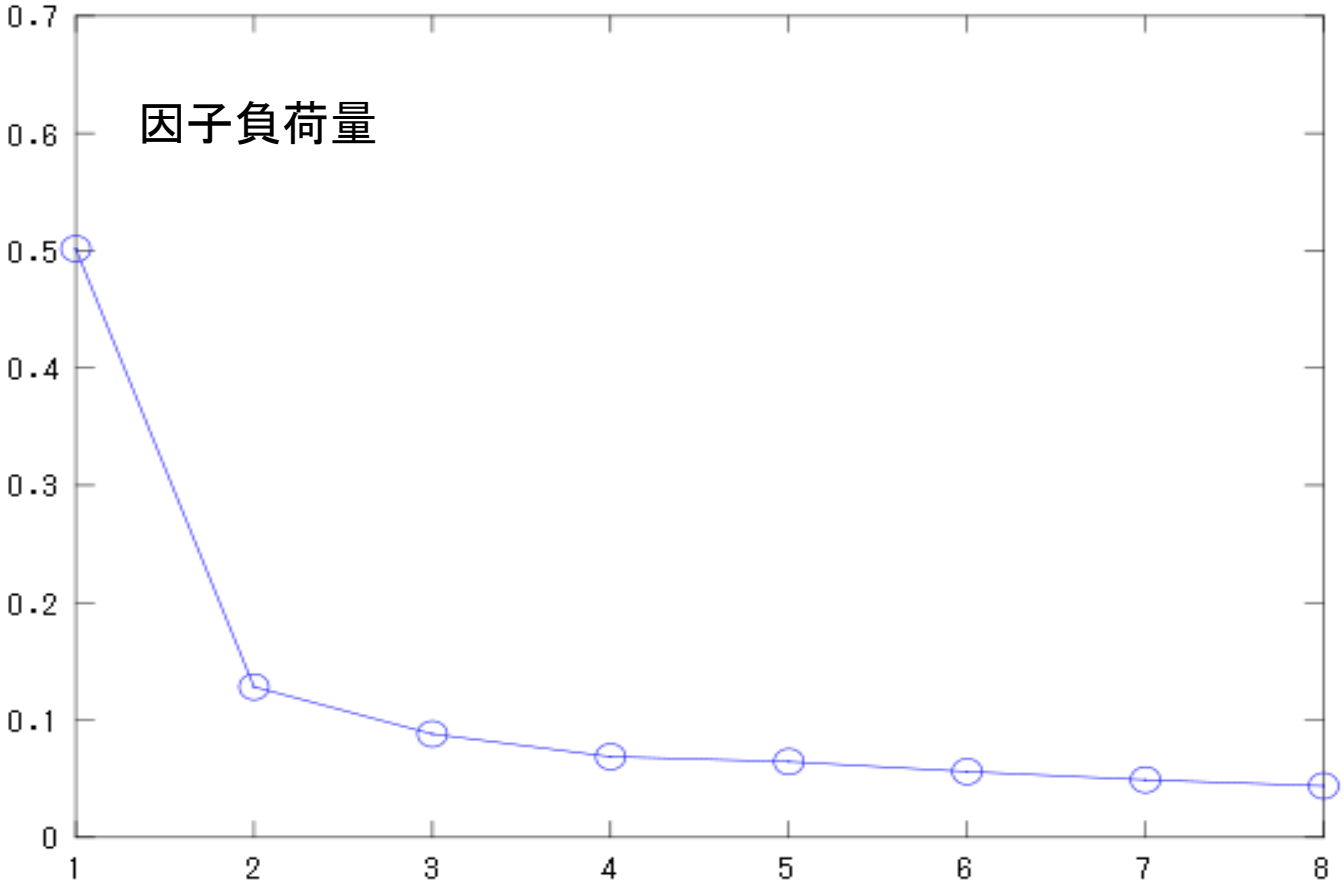






因子負荷量

固有値の相対的な大きさを表したもの。各主成分がどの程度の重要性を持つかがわかる。本当にｋ次元の線形空間上にあるときには正の固有値はk個。



固有ベクトルの様子

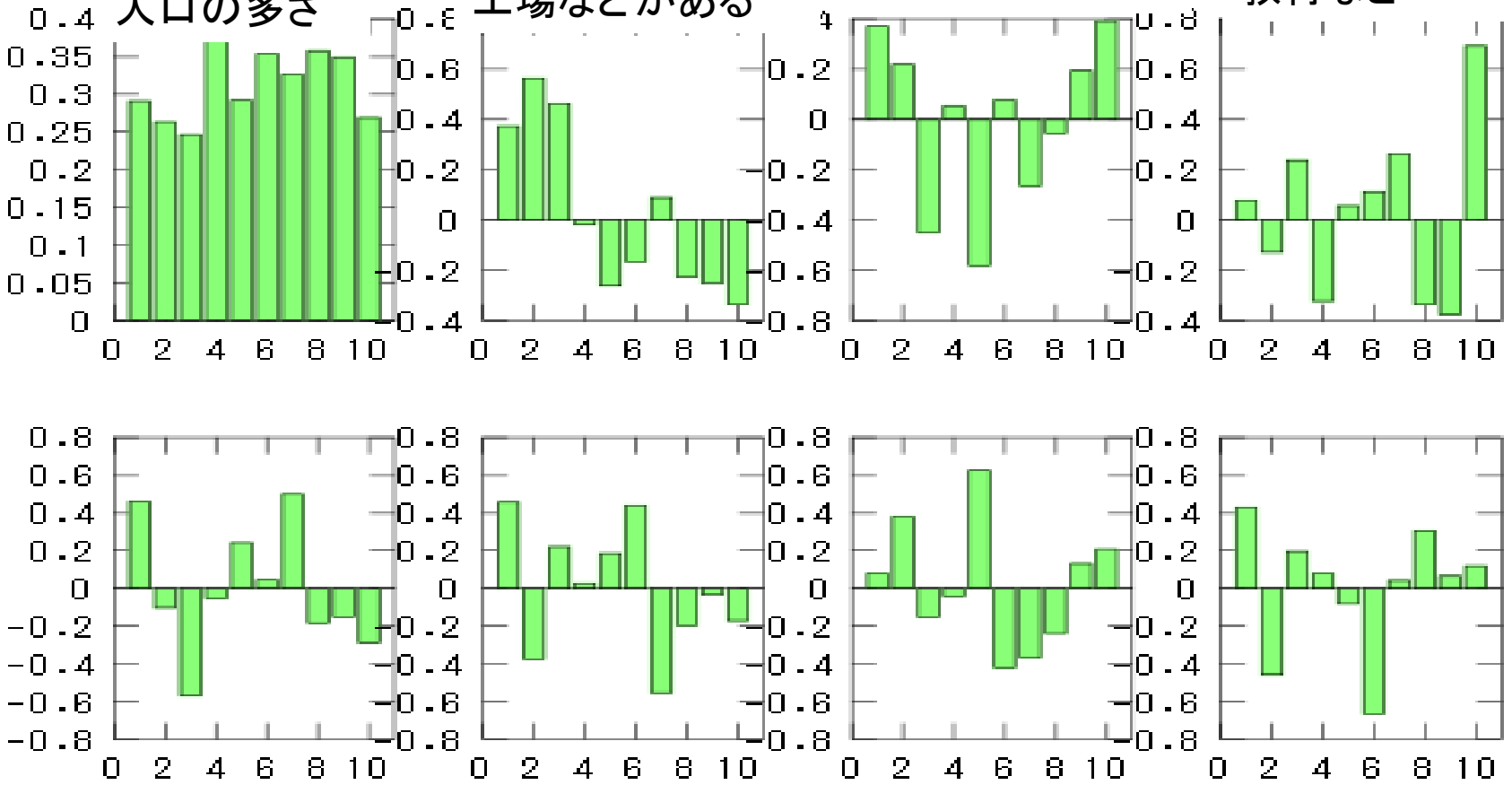
固有ベクトルの意味は、x の各成分との関係から解析者が考えるしかない。

第1主成分
人口の多さ

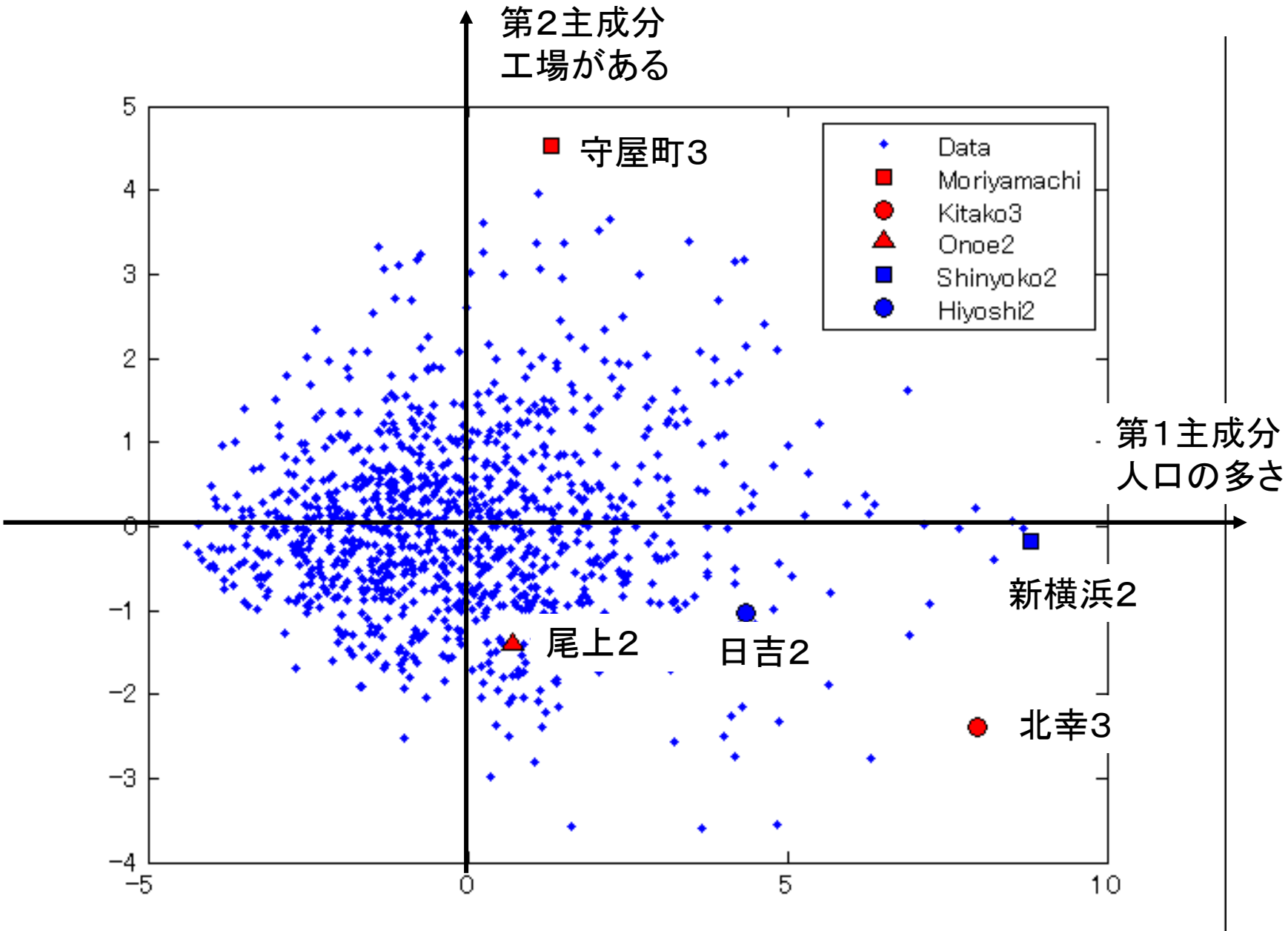
第2主成分
工場などがある

第3主成分(-)
金融・運輸など

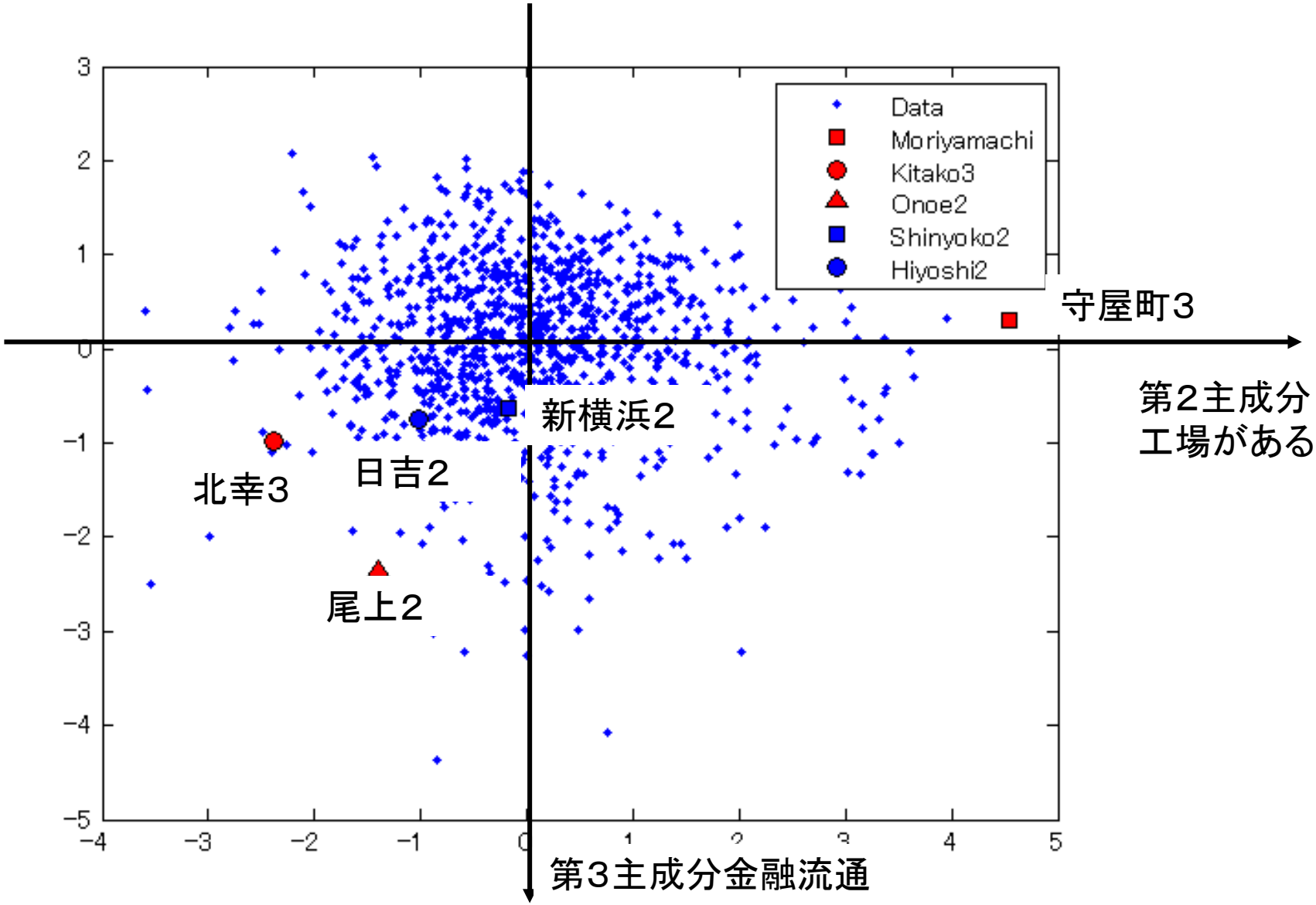
第4主成分
教育など



第1第2主成分の散布図



第2第3主成分の散布図



ラベルなしデータでよく現れる問題

1．主成分の意味を読み取るにはどうしたらよいですか。

　主成分の意味は x との関係から読み取ることになります。データ解析者がそのデータについての知識がない場合には、データそのものについての専門家との話し合いが必要になることもあります。

2．得られた結果は正しいのですか。

　主成分分析はどちらかといえば発見的な方法であり、得られた主成分や固有ベクトルがどの程度に真のものに近いかについては数学的に高度な問題です。もしも、発見できたものについての検定を行いたい場合には、別の統計モデルで帰無仮説と対立仮説を作って性質のわかりやすい形で行うほうがよいかもしれません。

3．誰も気づかなかったことが発見できますか。

　主成分の中に誰も気づかなかったものが含まれている可能性はありますが、そのことに気づくことができるためには、固有ベクトルで表された情報を言語化できる必要があり、言語化できるものは予め人間が知っていることに限られるかもしれません。「まったく知らないことに気づくことができる？」