主成分分析(PCA)とは

2009.11.4

「互いに相関のある多種類の変数を，互いに無相関な少数個の変数に要約する．」

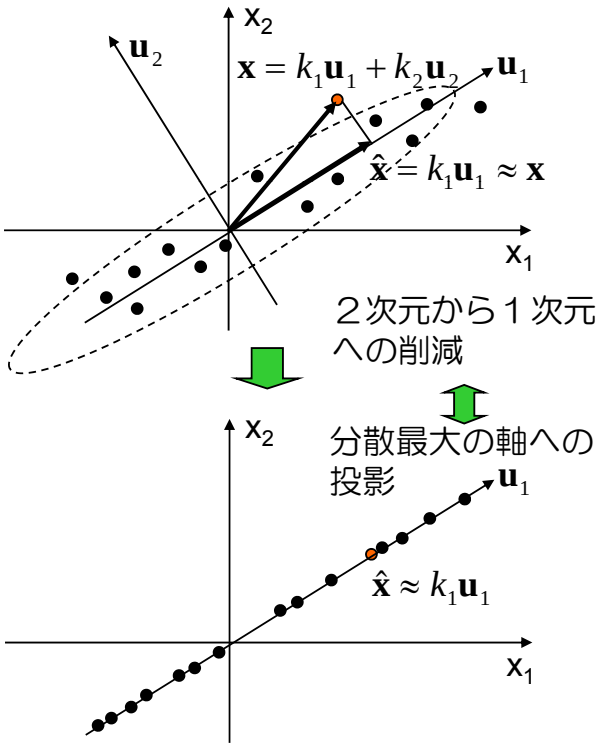
⇩

少ない次元数で解析，圧縮などを行う．

$\mathbf{u}_1$：サンプルの分散が最大の方向
（単位ベクトル）

$\mathbf{u}_2$：２番目に分散が大きい方向
（単位ベクトル）
（２変数の場合は，分散が最小の方向）

平均ベクトルが０の２変数分布の場合

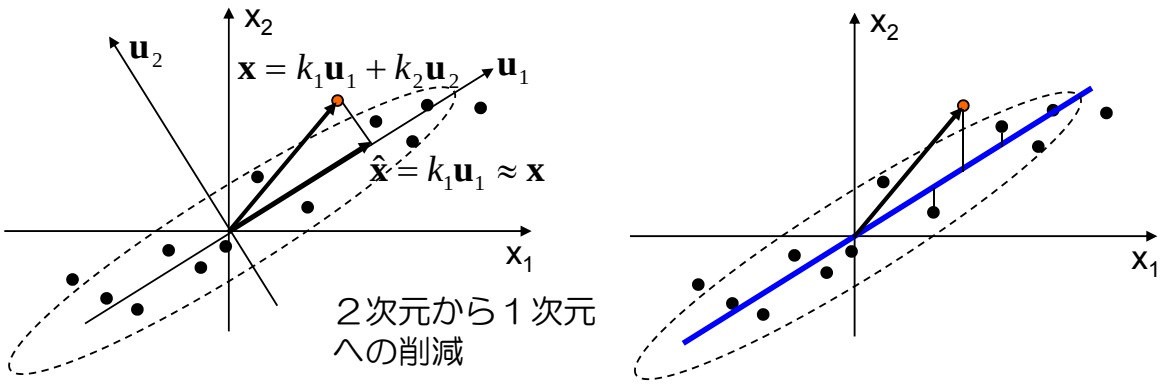


回帰分析とはどこが違う？

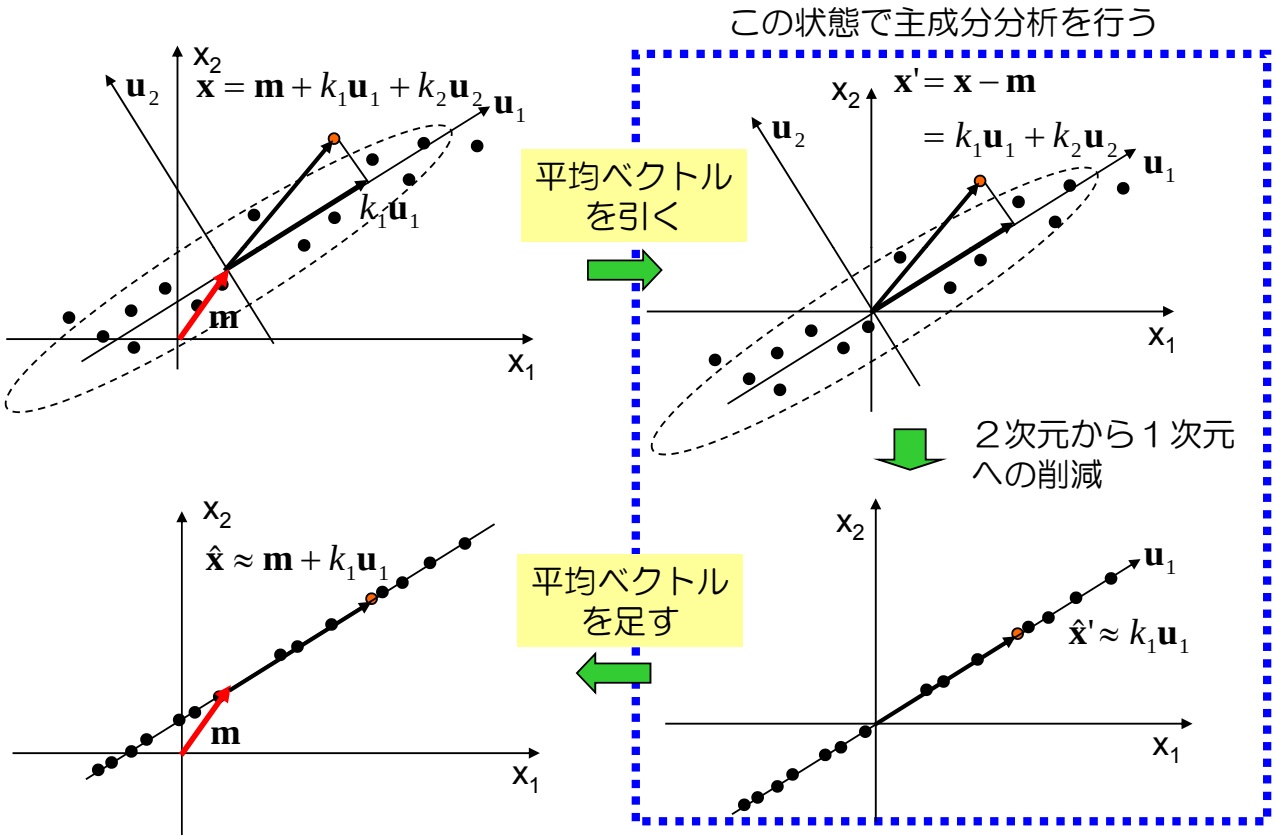
主成分分析：
複数個の変数から互いに無相関な変数（もとの変数の線形結合で表現される）に集約すること．
（回帰分析のような主従関係はない）

回帰分析：
１個以上の説明変数を用いて1個（以上）の目的変数を近似的な関数で表現すること．

２変数に対する１次回帰分析との違い

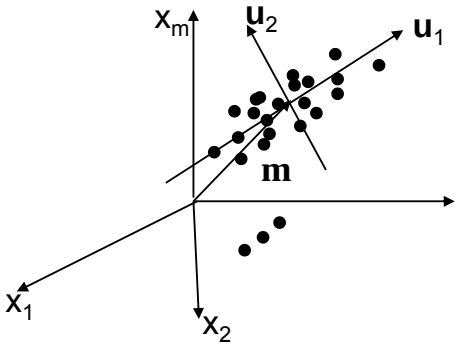


平均が0ベクトルでないときの手順



高次元の主成分分析

m次元から２次元への要約



$$\mathbf{x} = \mathbf{m} + k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m$$

$$\hat{\mathbf{x}} = \mathbf{m} + k_1\mathbf{u}_1 + k_2\mathbf{u}_2$$

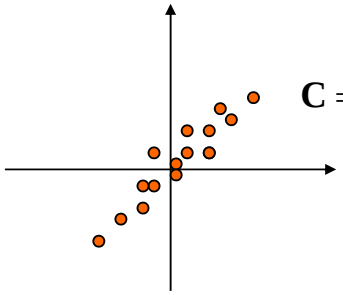
共分散行列の復習（２変数の場合）

共分散行列　　各変数とも，平均を０にしてから
相関を計算して得られる行列

$$\mathbf{C}=\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)(x_{i,2}-m_2) \\ \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)(x_{i,2}-m_2) & \frac{1}{n}\sum_{i=1}^{n}(x_{i,2}-m_2)^2 \end{bmatrix}$$

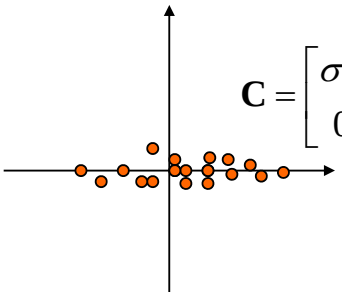
$$=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$$

(例１）２変数間に相関がある場合



$$\mathbf{C}=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$$

(例２）２変数間に相関がない場合



$$\mathbf{C}=\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}$$

主成分の方向$\mathbf{u}_1,\mathbf{u}_2$を求める

方針

２変数間で相関が０になる方向，
すなわち共分散行列が対角行列になる
方向を求めればよい．

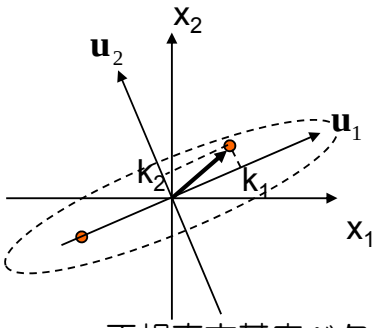
上記の条件を満足する，正規直交基底
ベクトルを$\mathbf{u}_1,\mathbf{u}_2$とする．

この２つの基底ベクトルを用いた座標
変換は以下のように表される．

$$\begin{bmatrix}k_1\\ k_2\end{bmatrix}=\begin{bmatrix}u_{11} & u_{12}\\ u_{21} & u_{22}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\quad\cdots(1)$$

またはベクトル表現で

$$\mathbf{k}=\mathbf{U}^T\mathbf{x}\quad\text{ただし}\quad\mathbf{U}=\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}$$



u_1,u_2: 正規直交基底ベクトル

Uで変換した後の共分散行列$\mathbf{C}_k$は

$$\mathbf{C}_k=\langle\mathbf{k}\mathbf{k}^T\rangle=\langle\mathbf{U}^T\mathbf{x}(\mathbf{U}^T\mathbf{x})^T\rangle$$

$$=\mathbf{U}^T\langle\mathbf{x}\mathbf{x}^T\rangle\mathbf{U}=\mathbf{U}^T\mathbf{C}\mathbf{U}\quad\cdots(2)$$

両辺にUを左からかけて
左右入れ替えると

$$\mathbf{U}\mathbf{U}^T\mathbf{C}\mathbf{U}=\mathbf{U}\mathbf{C}_k\quad\cdots(3)$$

主成分の方向を求める（つづき）

$$\mathbf{U}\mathbf{U}^T\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{C}_k \quad \cdots(3)$$

Uの正規直交性より

$$\mathbf{U}^T = \mathbf{U}^{-1} \quad \cdots(4)$$

よって（３）式は

$$\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{C}_k$$

共分散の対角化が**U**によってなされるとすると

$$\mathbf{C}_k = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \cdots(5)$$

成分に分けて表すと

$$\mathbf{C}[\mathbf{u}_1 \quad \mathbf{u}_2] = [\mathbf{u}_1 \quad \mathbf{u}_2]\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

または

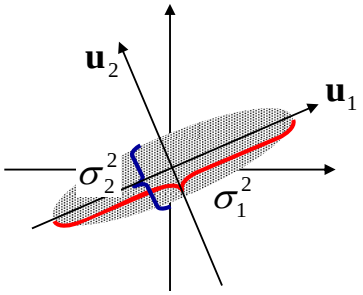
$$\mathbf{C}\mathbf{u}_1 = \sigma_1^2\mathbf{u}_1, \mathbf{C}\mathbf{u}_2 = \sigma_2^2\mathbf{u}_2$$

これは，もとのデータの共分散行列**C**に対する固有値問題に他ならない．

すなわち，u_1, u_2は行列**C**の固有ベクトル，σ_1^2, σ_2^2は固有値として求められる．

Cは実対称行列 $\Rightarrow \sigma_i^2 \geq 0 \quad (for\ \ all\ \ i)$

固有値σ_i^2は対角化されたデータの各変数の分散を与える．



主成分の方向を求める――一般の高次元データ―

２次元での議論をそのままm次元へ拡張すればよい

サンプルデータ：

$$\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = \begin{bmatrix} x_{1,1} & & x_{n,1} \\ \vdots & \cdots & \vdots \\ x_{1,m} & & x_{n,m} \end{bmatrix}$$

サンプルデータの共分散行列：

$$\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

対角化する基底ベクトルのセット：

$$\mathbf{U} = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_m] = \begin{bmatrix} u_{1,1} & & u_{m,1} \\ \vdots & \cdots & \vdots \\ u_{1,m} & & u_{m,m} \end{bmatrix}$$

座標変換後の共分散行列：

$$\mathbf{C}_k = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{bmatrix}$$

固有値問題の表現：

$$\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{C}_k$$

または

$$\mathbf{C}\mathbf{u}_i = \sigma_i^2\mathbf{u}_i, \quad for\ \ i = 1,...,m$$

主成分の方向を求める－計算例－

共分散行列が以下の式で表されるサンプルの集合を考える.

$$\mathbf{C}=\begin{bmatrix}1 & \alpha\\ \alpha & 1\end{bmatrix}$$

ただし平均ベクトルは0とする.

問題１：αが以下の３種類の場合について,
サンプルの分布を模式的に描きなさい.

(i) $\alpha=0$
(ii) $\alpha=0.5$
(iii) $\alpha=1.0$

問題２：$0<\alpha<1$の場合について，主成分を計算で求めなさい.

ヒント
①固有値問題 $\mathbf{Cu}=\lambda\mathbf{u}\Leftrightarrow(\mathbf{C}-\lambda\mathbf{I})\mathbf{u}=\mathbf{0}\cdots(1)$
を解くことと同値である.
②正規性の条件$\|\mathbf{u}\|=\sqrt{u_1^2+u_2^2}=1$を用いる.

低次元主成分による近似と誤差

もとの変数と主成分ベクトルの係数との間には以下の関係がある.

$$\mathbf{k}=\mathbf{U}^T\mathbf{x}$$

また，Uの正規直交性より

$$\mathbf{Uk}=\mathbf{UU}^T\mathbf{x}\Rightarrow\mathbf{x}=\mathbf{Uk}$$

または

$$\mathbf{x}=\sum_{i=1}^{m}k_i\mathbf{u}_i$$

いま，m＝２とし，１次元主成分による近似とそれによる誤差を考える.

あるj番目のサンプルについて，もとデータおよび低次元（１次元）による近似表現は以下のように書ける.

$$\mathbf{x}^{(j)}=k_1^{(j)}\mathbf{u}_1+k_2^{(j)}\mathbf{u}_2$$

$$\hat{\mathbf{x}}^{(j)}=k_1^{(j)}\mathbf{u}_1$$

誤差ベクトルは

$$\begin{aligned}\mathbf{e}^{(j)}&=\mathbf{x}^{(j)}-\hat{\mathbf{x}}^{(j)}\\&=(k_1{}^{(j)}\mathbf{u}_1+k_2{}^{(j)}\mathbf{u}_2)-k_1{}^{(j)}\mathbf{u}_1\\&=k_2{}^{(j)}\mathbf{u}_2\end{aligned}$$

誤差ベクトルの大きさ（ノルムの2乗）は

$$|\mathbf{e}^{(j)}|^2=|k_2{}^{(j)}\mathbf{u}_2|^2=k_2^{(j)2}$$

サンプル全体での誤差の平均は

$$E=<\mathbf{e}>=\frac{1}{n}\sum_{j=1}^{n}|\mathbf{e}^{(j)}|^2=\frac{1}{n}\sum_{j=1}^{n}k_2^{(j)2}=\sigma_2^2$$

すなわち誤差は，用いなかった第2主成分の残差（分散）に等しい.

低次元主成分による近似と誤差（つづき）

より一般に，m次元データに対するr次元主成分による近似とそれによる誤差を考える.

もとのデータおよび近似データを

$$\mathbf{x}=\sum_{i=1}^{m}k_i\mathbf{u}_i,\quad \hat{\mathbf{x}}=\sum_{i=1}^{r}k_i\mathbf{u}_i \qquad r<m$$

と書く．誤差ベクトルは

$$\begin{aligned}\mathbf{e}^{(j)}&=\mathbf{x}^{(j)}-\hat{\mathbf{x}}^{(j)}\\&=\sum_{i=1}^{m}k_i^{(j)}\mathbf{u}_i-\sum_{i=1}^{r}k_i^{(j)}\mathbf{u}_i\\&=\sum_{i=r+1}^{m}k_i^{(j)}\mathbf{u}_i\end{aligned}$$

であり，その2ノルムは

$$\begin{aligned}|\mathbf{e}^{(j)}|^2&=k_{r+1}^{(j)2}+k_{r+2}^{(j)2}+\cdots+k_m^{(j)2}\\&=\sum_{i=r+1}^{m}k_i^{(j)2}\end{aligned}$$

このとき，サンプル全体での誤差の平均は

$$\begin{aligned}E(r)&\equiv<|\mathbf{e}|^2>\\&=\frac{1}{n}\sum_{j=1}^{n}|\mathbf{e}^{(j)}|^2\\&=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=r+1}^{m}k_i^{(j)2}\\&=\sum_{i=r+1}^{m}\frac{1}{n}\sum_{j=1}^{n}k_i^{(j)2}\\&=\sum_{i=r+1}^{m}\sigma_i^2\end{aligned}$$

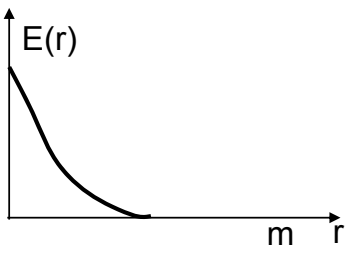
すなわち，誤差は，用いなかった主成分の残差（分散）の和に等しい.

近似による誤差と累積寄与率

主成分を，分散の大きい順に番号付けしたものならば，誤差

$$E(r)=\sum_{i=r+1}^{m}<k_i^2>=\sum_{i=r+1}^{m}\sigma_i^2$$

は，r に関して単調減少関数となる



逆に，はじめのr個の成分でどのくらい正確に，もとの分布を表せるかの尺度として，以下に示す累積寄与率がある.

$$\text{累積寄与率(r)}=\frac{\sum_{i=1}^{r}\sigma_i^2}{\sum_{i=1}^{m}\sigma_i^2}$$