主成分分析(PCA)とは

「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．」

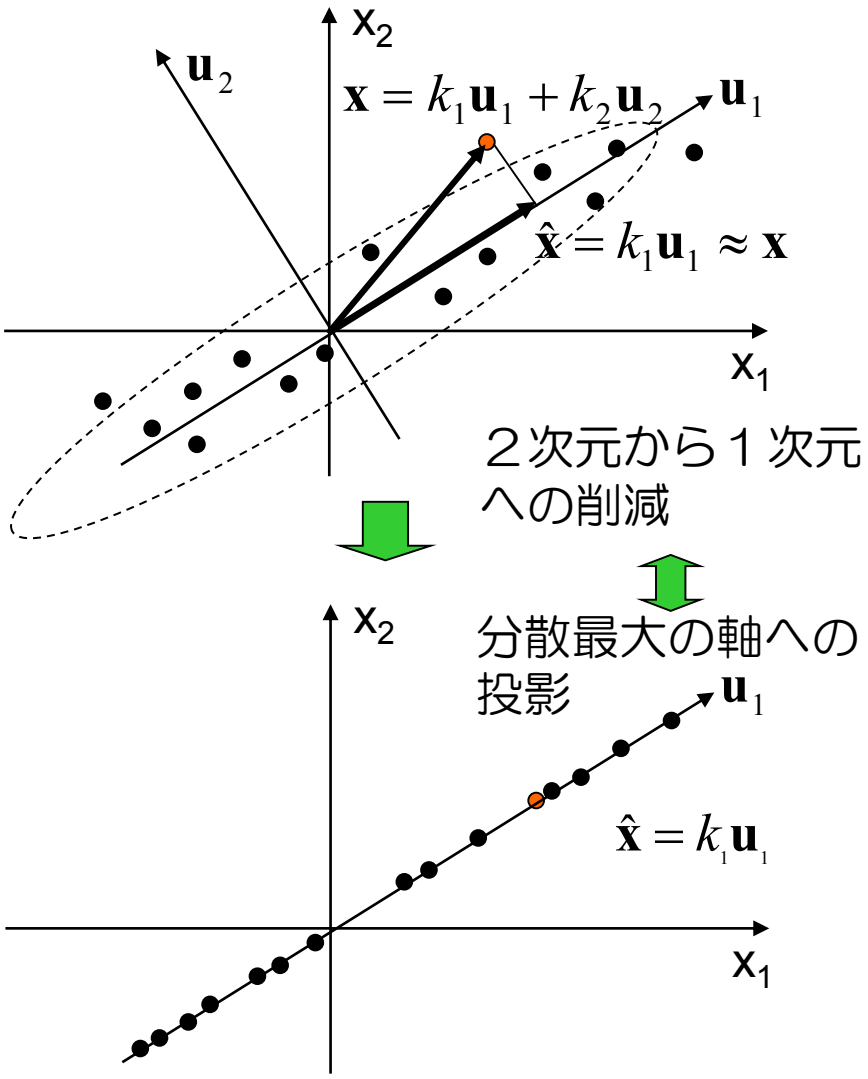
⇩

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

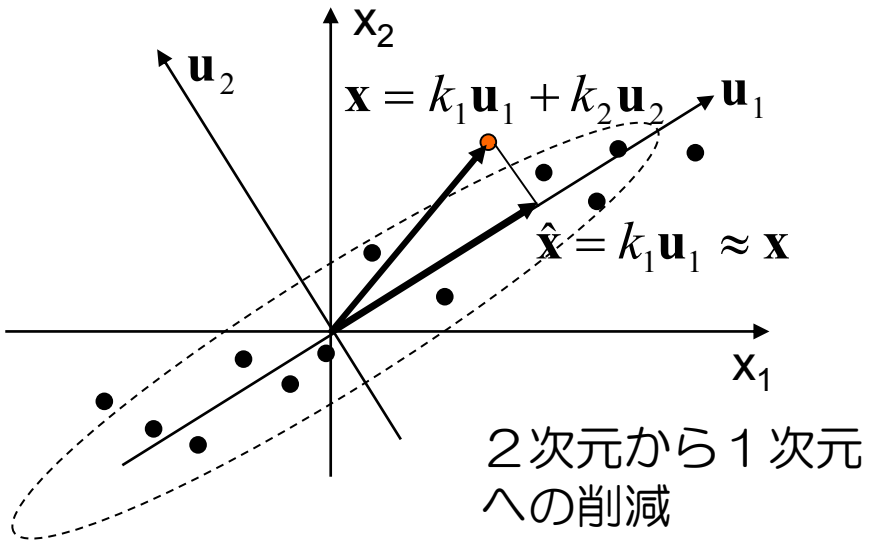


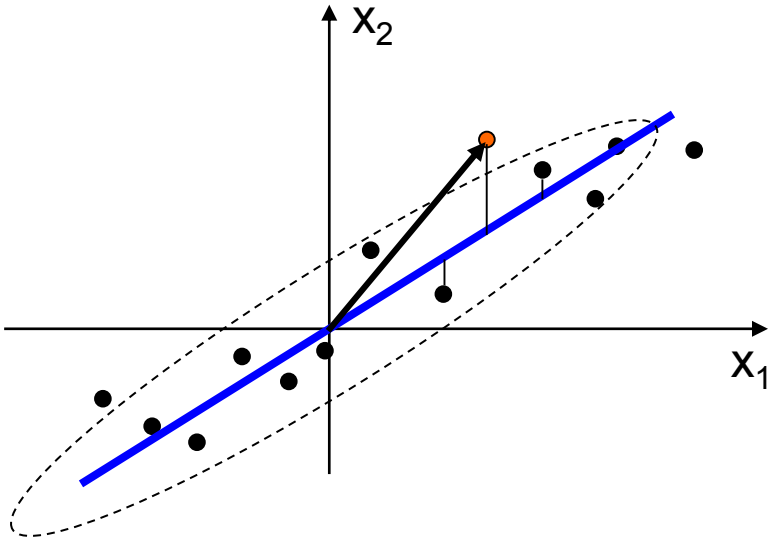
回帰分析とはどこが違う？

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）

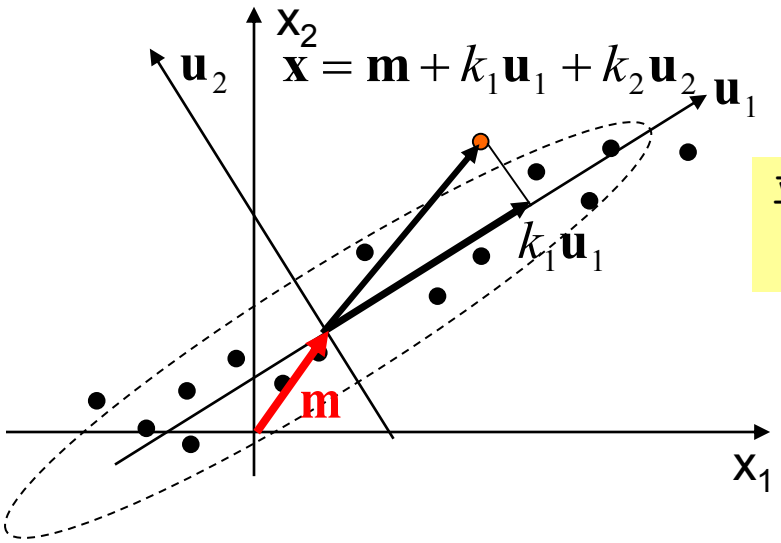
回帰分析：
１個以上の説明変数を用いて1個（以上）の目的変数を近似的な関数で表現すること.

２変数に対する１次回帰分析との違い



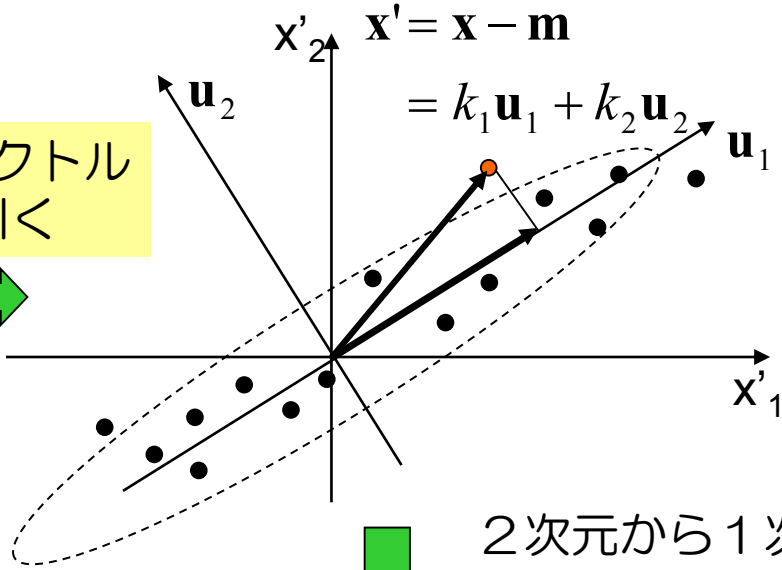


平均が0ベクトルでないときの手順

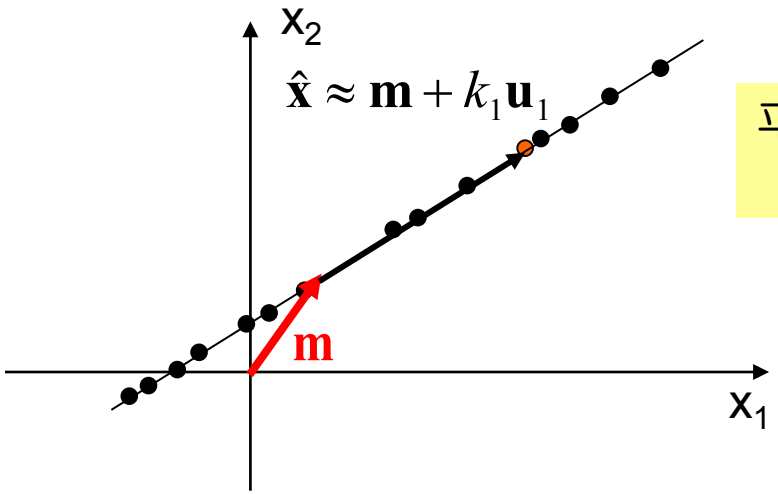


この状態で主成分分析を行う

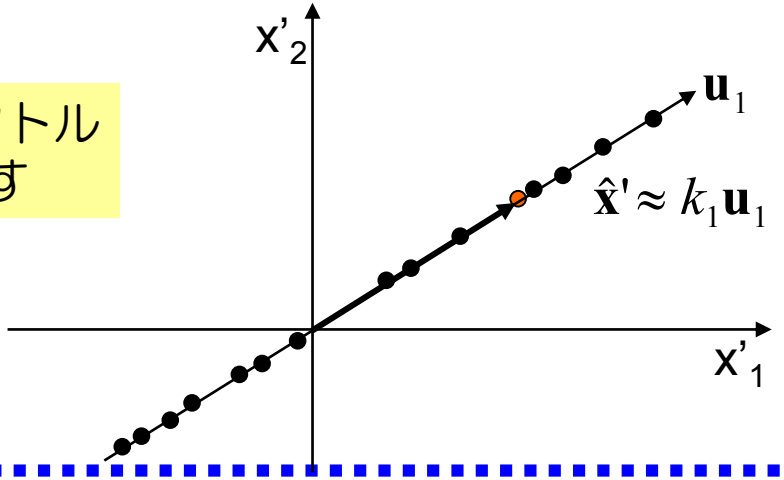
平均ベクトル
を引く



２次元から１次元
への削減

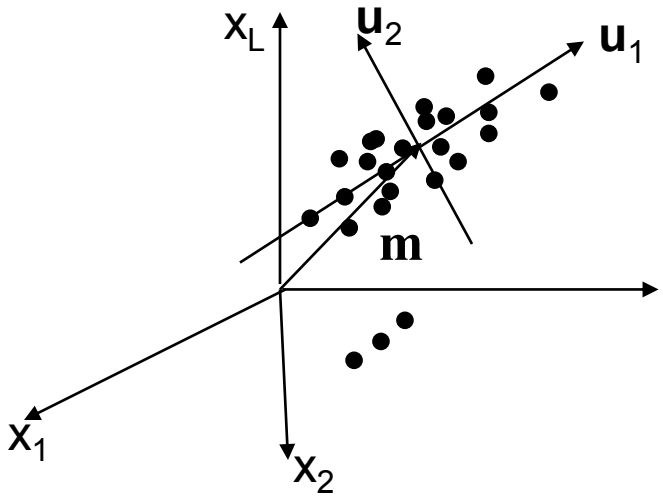


平均ベクトル
を足す



高次元の主成分分析

n次元から２次元への要約



$$\mathbf{x} = \mathbf{m} + \mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_L$$

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

共分散行列（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

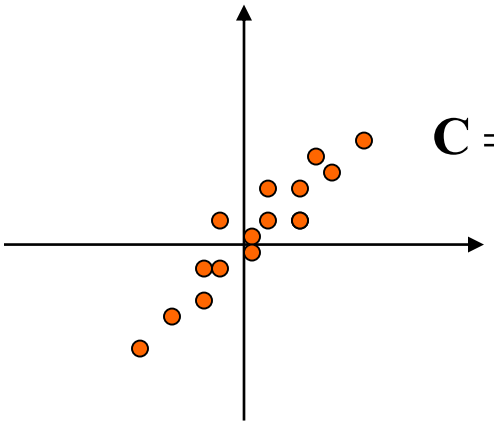
$$\mathbf{C}=\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)(x_{i,2}-m_2) \\ \frac{1}{n}\sum_{i=1}^{n}(x_{i,1}-m_1)(x_{i,2}-m_2) & \frac{1}{n}\sum_{i=1}^{n}(x_{i,2}-m_2)^2 \end{bmatrix}$$

集合平均を<>で表す

$$=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T=\begin{bmatrix} <(x_1-m_1)^2> & <(x_1-m_1)(x_2-m_2)> \\ <(x_1-m_1)(x_2-m_2)> & <(x_2-m_2)^2> \end{bmatrix}$$

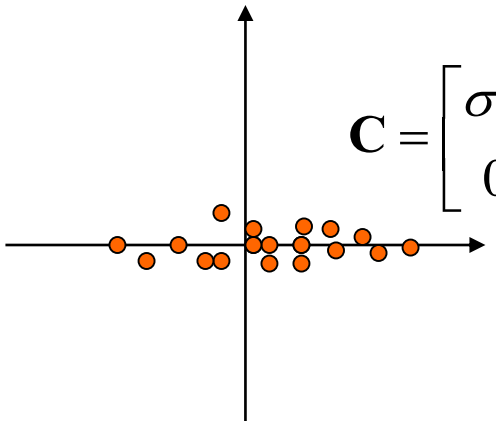
$$=\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

（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

今後の数式展開は平均を0にする処理が済んでいるものとして進める。
ダッシュ(')付きの記号にすべきだが、煩わしいので省略する。

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.

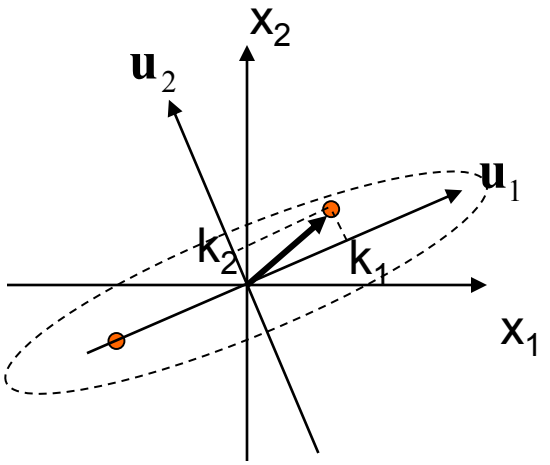
上記の条件を満足する，正規直交基底ベクトルを$\mathbf{u}_1$,$\mathbf{u}_2$とする.

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

$$\mathbf{C}_k = \left\langle \mathbf{k}\mathbf{k}^T \right\rangle = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}_i\mathbf{k}_i^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}^T\mathbf{x}_i(\mathbf{U}^T\mathbf{x}_i)^T$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}^T\mathbf{x}_i\mathbf{x}_i^T(\mathbf{U}^T)^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{U}$$

$$= \mathbf{U}^T\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{U}$$

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

主成分の方向を求める（つづき）

両辺にUを左からかけて
左右入れ替えると

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

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

１列目の式　２列目の式

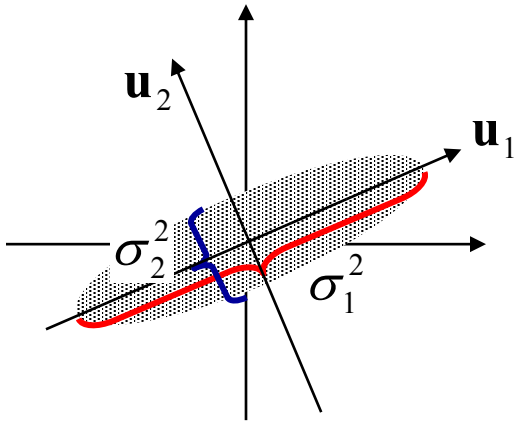
$$\mathbf{C}\mathbf{u}_1 = \sigma_1^2\mathbf{u}_1, \mathbf{C}\mathbf{u}_2 = \sigma_2^2\mathbf{u}_2$$

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.

$\mathbf{C}$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L次元へ拡張すればよい

サンプルデータ：

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

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：

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_L \end{bmatrix} = \begin{bmatrix} u_{1,1} & & u_{L,1} \\ \vdots & \cdots & \vdots \\ u_{1,L} & & u_{L,L} \end{bmatrix}$$

座標変換後の共分散行列：

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

固有値問題の表現：

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

または

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

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

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

または

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

いま，L＝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

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

と書く．誤差ベクトルは

$$\mathbf{e}^{(j)}=\mathbf{x}^{(j)}-\hat{\mathbf{x}}^{(j)}$$

$$=\sum_{i=1}^{L}k_i^{(j)}\mathbf{u}_i-\sum_{i=1}^{r}k_i^{(j)}\mathbf{u}_i$$

$$=\sum_{i=r+1}^{L}k_i^{(j)}\mathbf{u}_i$$

であり，その2ノルムは

$$|\mathbf{e}^{(j)}|^2=k_{r+1}^{(j)^2}+k_{r+2}^{(j)^2}+\cdots+k_L^{(j)^2}$$

$$=\sum_{i=r+1}^{L}k_i^{(j)^2}$$

このとき，サンプル全体での誤差の平均は

$$E(r)\equiv<|\mathbf{e}|^2>$$

$$=\frac{1}{n}\sum_{j=1}^{n}|\mathbf{e}^{(j)}|^2$$

$$=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=r+1}^{L}k_i^{(j)^2}$$

$$=\sum_{i=r+1}^{L}\frac{1}{n}\sum_{j=1}^{n}k_i^{(j)^2}$$

$$=\sum_{i=r+1}^{L}\sigma_i^2$$

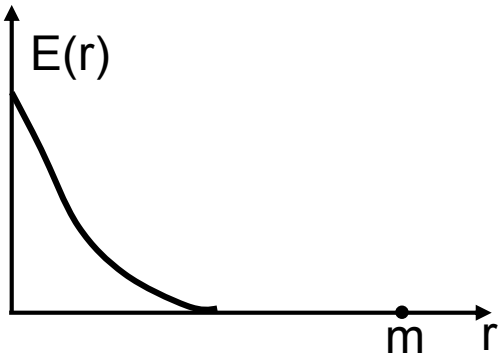
すなわち，誤差は，用いなかった主成分の残差（分散）の和に等しい.

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

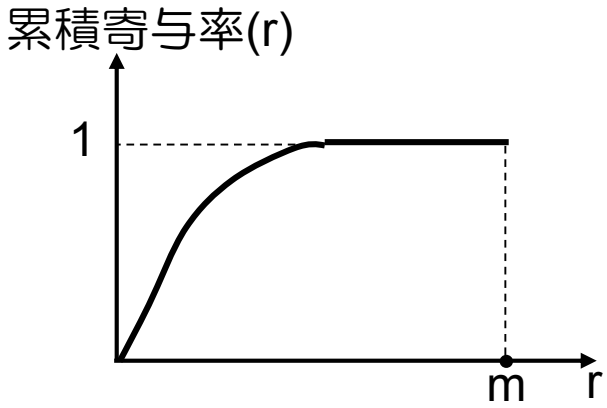
$$E(r) = \sum_{i=r+1}^{L} < k_i^2 > = \sum_{i=r+1}^{M} \sigma_i^2$$

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.

累積寄与率(r)＝



主成分の方向を求める－課題の答え－

$$\mathbf{Cu} = \lambda\mathbf{u} \Leftrightarrow (\mathbf{C} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0} \cdots (1)$$

が自明でない解(**u=0**以外の解）を
もつためには，

$$|\mathbf{C} - \lambda\mathbf{I}| = 0 \quad \cdots (2)$$

が必要十分条件となる．実際に計算すると

$$|\mathbf{C} - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & \alpha \\ \alpha & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - \alpha^2 = 0 \quad \cdots (3)$$

この式をλについて解くと

$$\lambda = 1 \pm \alpha \quad \cdots (4)$$

を得る．いま，便宜的に，２つの解を
以下のように表そう．

$$\lambda_1 = 1 + \alpha$$
$$\lambda_2 = 1 - \alpha$$

２つの解のうち，まずλ_1を式(1)に代入すると

$$u_1 + \alpha u_1 = (1+\alpha)u_1 \cdots$$

より

$$u_1 = u_2 \cdots (6)$$

を得る．正規性の条件，

$$u_1^2 + u_2^2 = 1 \cdots (7)$$

と連立させて解くと

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \cdots (8)$$

を得る．同様に，$\lambda_2 = 1 - \alpha$に対しては

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \cdots (9)$$

を得る．

主成分の方向を求める－計算例－

固有値問題の２つの解を整理してみる.

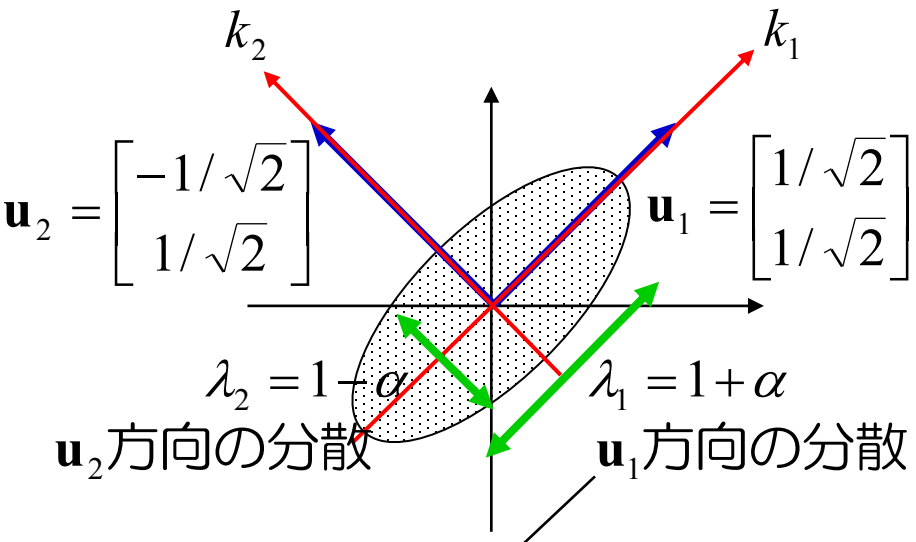
固有値　：$\lambda_1 = 1+\alpha$

固有ベクトル：$\mathbf{u}_1 = \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

固有値　：$\lambda_2 = 1-\alpha$

固有ベクトル：$\mathbf{u}_2 = \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

$0 < \alpha$の場合，$\lambda_1 = 1+\alpha > 1-\alpha = \lambda_2$
であるから，λ_1に対応した$\mathbf{u}_1$が
分散の大きい軸を表し，
λ_2に対応した$\mathbf{u}_2$が
分散の小さい軸を表す.

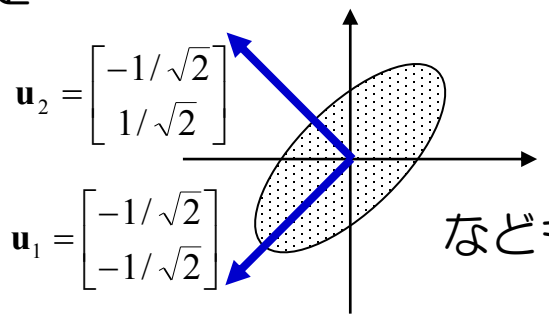


$$\mathbf{C}_k = \begin{bmatrix} 1+\alpha & 0 \\ 0 & 1-\alpha \end{bmatrix}$$

$\mathbf{u}_2$方向の分散

共分散は0.

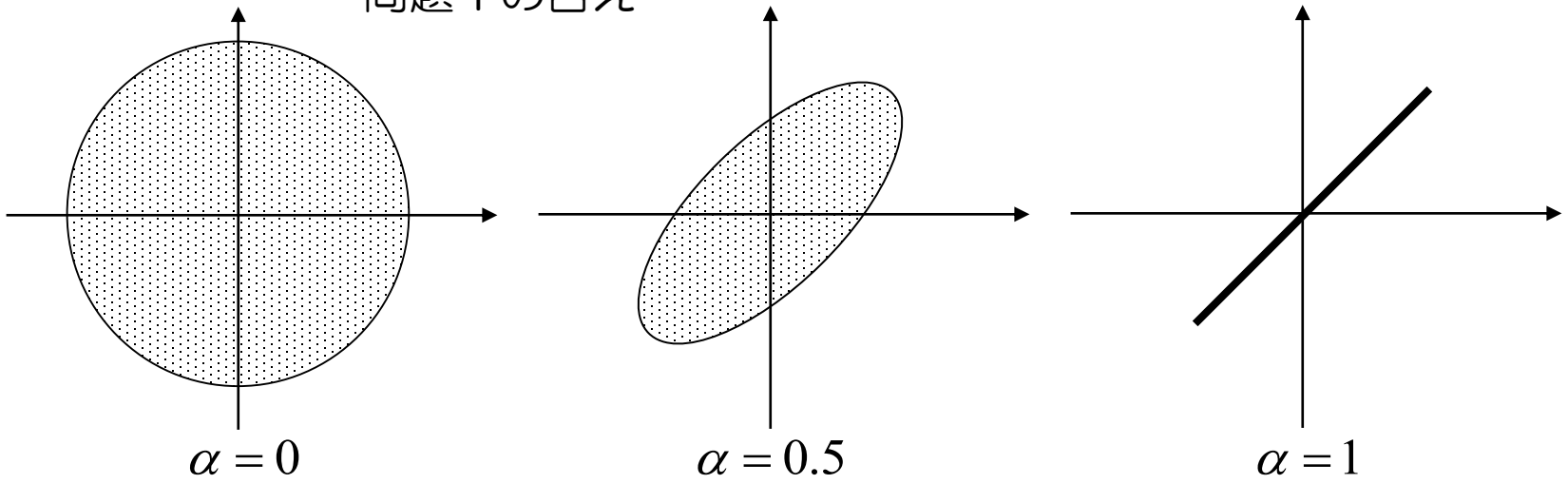
補足



などもOK.

主成分の方向を求める－計算例－

問題１の答え



右の分布は分散が
２変数で等しくない
ので間違い.