多変量解析

２００９年度後期　羽石　秀昭（はねいし）

講義内容：
- 統計解析の基礎
- 重回帰分析 (Multiple Regression Analysis)
  ＋情報量基準 (information Criterion)
- 主成分分析 (Principal Component Analysis)
- 判別分析 (Discriminant Analysis)
- 検定・分散分析

参考書籍：
１．基本統計学 宮川公男 著
　　有斐閣　 ２，６７８円
２．情報量基準による統計解析入門　鈴木儀一郎　著
　　講談社　２，７１８円
３．情報量統計学　坂元，石黒，北川　著
　　共立出版　３，７６０円
４．多変量解析法 奥野，久米，芳賀，吉澤 著
　　日科技連 ２，８００円
５．多変量解析概論 塩谷 著
　　朝倉書店 ３，７０８円

講義日程
１０月　７，１４，２１，２８
１１月　４，１１，１８，２５（中間テスト）
１２月　２（休講），９，１６
１月　１３，２０，２７
２月　３（期末テスト）

回帰分析とはどのようなものか　～単回帰について～

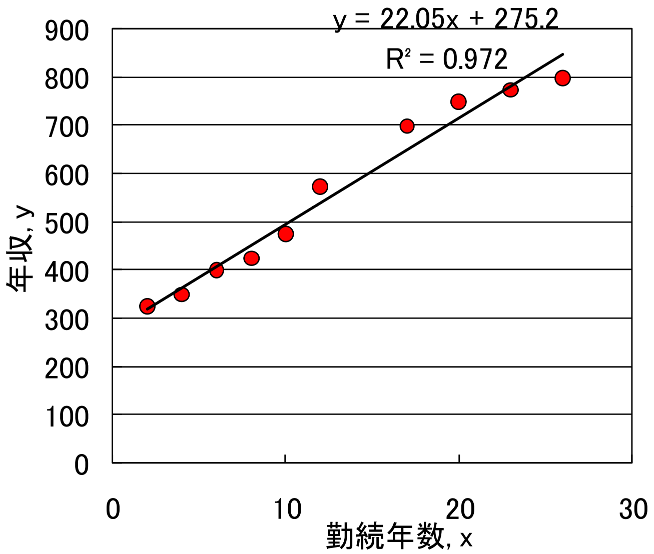
１つの変数ｘから，１つの変数ｙを推定する.

例）勤続年数と年収の関係を分析する.
直線で関係式を表現する.

$$y = ax + b$$

x：説明変数
y：目的変数

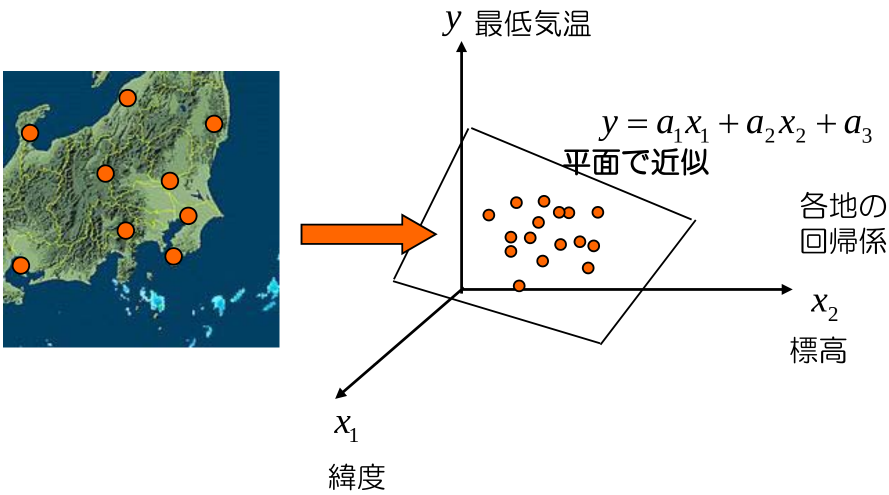
勤続年数, x	年収, y
2	325
4	350
6	400
8	425
10	475
12	575
17	700
20	750
23	775
26	800



回帰分析とはどのようなものか ～重回帰について～

２つ以上の変数ｘ$_1$，ｘ$_2$，．．．から，１つの変数ｙを推定する.

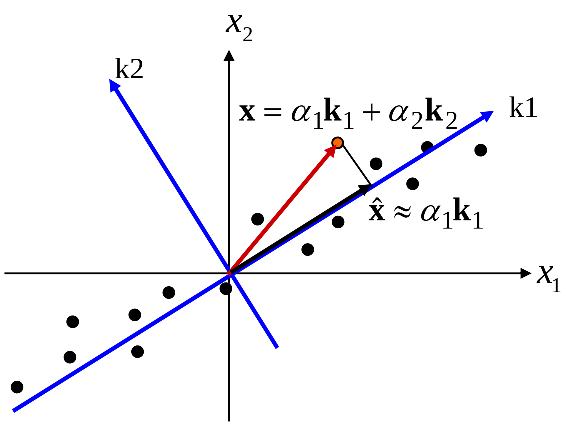
例） 最低気温（ｙ）と緯度（ｘ$_1$），標高（ｘ$_2$）の関係



各地のデータ（サンプル）から回帰係数a_1,a_2,a_3を決定する.

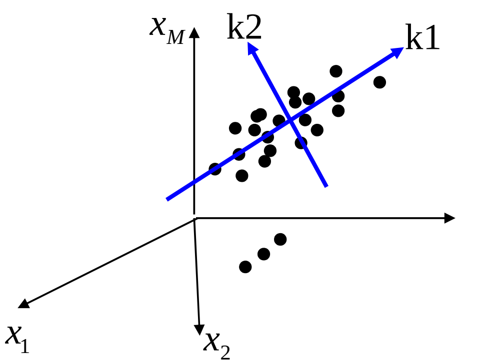
主成分分析とはどのようなものか

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.



$\mathbf{k}_1$：サンプルの分散が最大の方向
$\mathbf{k}_2$：２番目に分散が大きい方向

M次元空間の場合も同様：



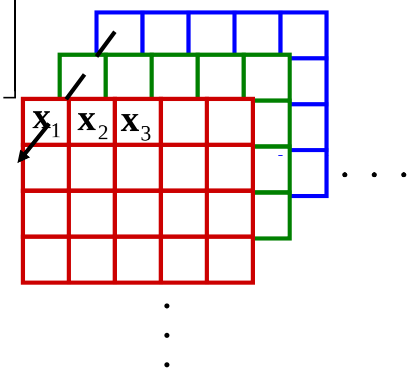
本授業での表記の約束

原則として
・ベクトル量は太字 $\mathbf{X}$
・スカラー量は細字 x
で表す

～応用事例　RGBカラー画像を２バンドで表す～

オリジナル画像

$$\mathbf{x}_1 = \begin{bmatrix} R_1 \\ G_1 \\ B_1 \end{bmatrix}$$



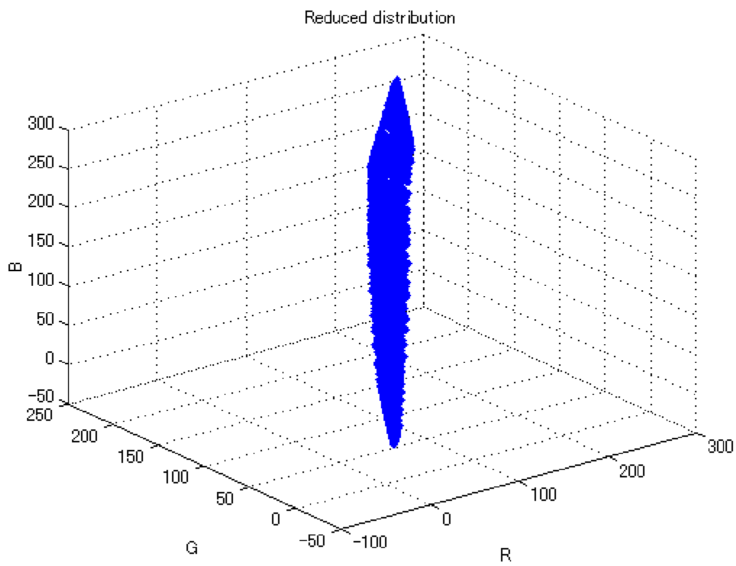
デモソフトで表示

RGB空間での画素値の分布

Program name:PCAdemoRGB.m

～応用事例　RGBカラー画像を２バンドで表す～（つづき）

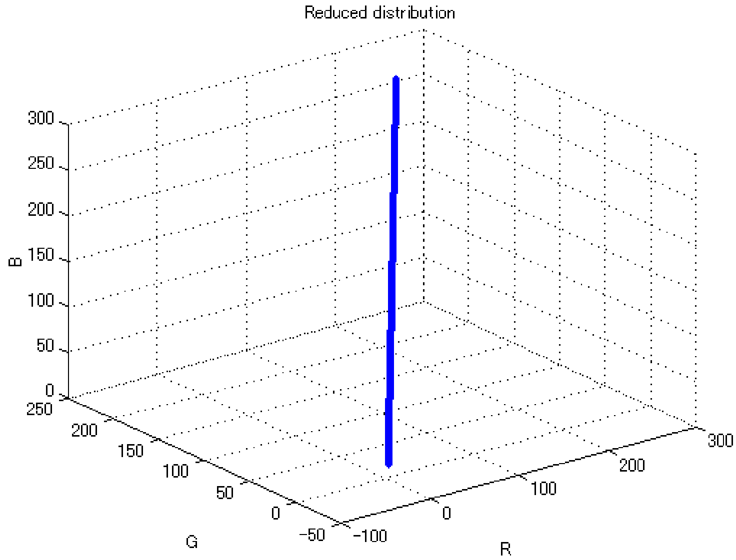
第１および第２主成分のみ



RGB空間での画素値の分布

～応用事例　RGBカラー画像を１バンドで表す～

第１主成分のみ



RGB空間での画素値の分布

～応用事例　RGBカラー画像を１or ２バンドで表す～

オリジナルカラー画像

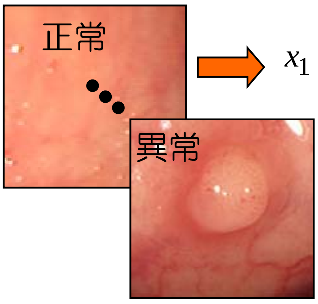
第１および第２主成分のみ

第１主成分のみ

判別分析とはどのようなものか

例）内視鏡画像からの自動診断

診断のついている
画像群



$x_1 \quad x_2$

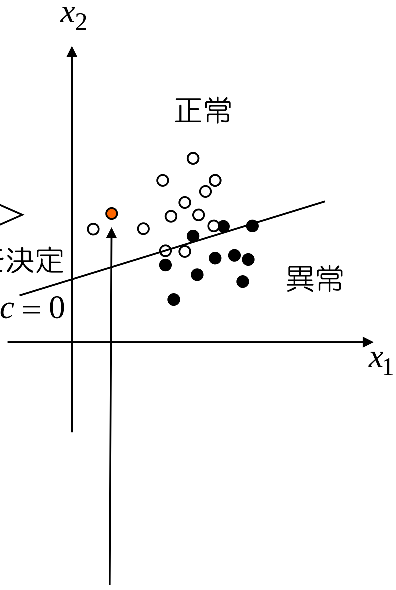
$x_1 \quad x_2$

画像から特徴量x_1，x_2
（色，形など）を抽出

プロット

判別関数を決定

$$ax_1 + bx_2 + c = 0$$



新しい画像がきたとき：
① 特徴量を算出
② 判別関数により，正常，異常を判断.

１変数の統計量・変数の標準化

<u>１変数の統計量</u>

ｎ個のサンプル（標本）の観測値xが
$x_1, x_2, \ldots, x_n$
と得られているとする.

・平均（１次の統計量）
mean

・分散（２次の統計量）（母集団の分散の推定値ではなく，サンプル自体の分散）
variance

・標準偏差（分散の平方根.
ただし正の値のみを扱う）
standard deviation

<u>変数の標準化（基準化，正規化：normalization ）</u>

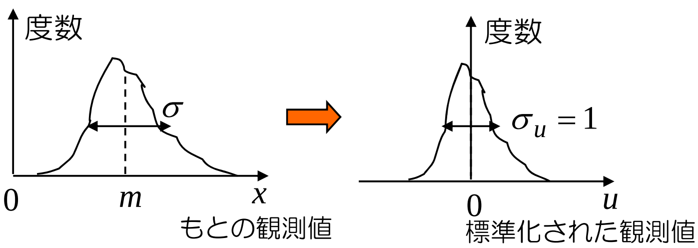
観測値$\{x_i\}$を以下の式により変換することを
標準化という.

$$u_i = \frac{x_i - m}{\sigma}$$

標準化されて得られる変数u_iは平均が０，
標準偏差が１である.

各自導出のこと

たとえば度数分布で考えると



２変数間の相関・共分散

１つのサンプルにつき２つの観測値(x_{1i},x_{2i})が得られるものとする.
それぞれの平均値が,

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}$$

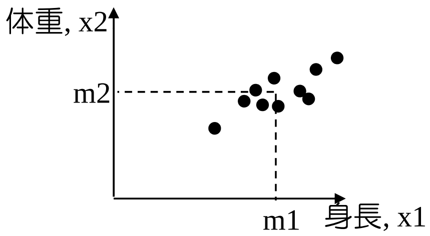
のとき,

$$\sigma_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{1i}-m_1)(x_{2i}-m_2)$$

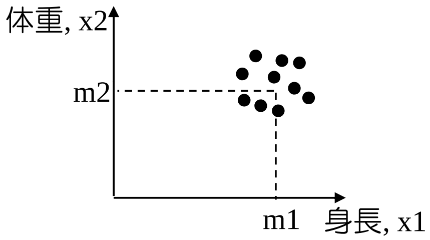
を２つの変数の共分散(covariance)という.

x_1の変化のしかたとx_2の変化のしかたに
相関があれば，共分散の絶対値は大きくなる.
相関がまったくなければ共分散は０となる.

例）１０人の体重と身長の関係



もし，以下のようなら共分散は小さい



手計算による演習

例題１　２変数をもつ３つのサンプル,
(1,1),(2,2),(3,3)をグラフにプロットしなさい.
また，各変数の平均と，共分散を求めなさい.

例題２　２変数をもつ４つのサンプル,
(1,1),(3,1),(1,3),(3,3)をグラフにプロットしなさい．また,
各変数の平均と，共分散を求めなさい.

多変数の統計量

サンプル（標本）データ

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix}, \ldots, \mathbf{x}_n = \begin{bmatrix} x_{1n} \\ x_{2n} \end{bmatrix}$$

あるいは，まとめて

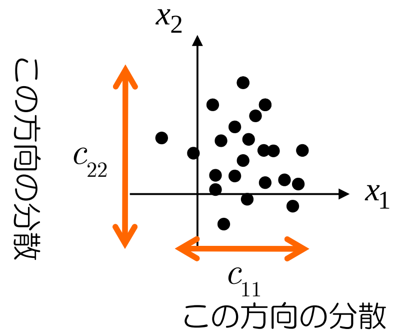
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}$$

本授業での表記の約束

原則として
・ベクトルは太字小文字
・行列は太字大文字
で表す

平均ベクトル

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n) = \begin{bmatrix} (1/n)\sum_{i=1}^{n} x_{1i} \\ (1/n)\sum_{i=1}^{n} x_{2i} \end{bmatrix}$$



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

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - m_1)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - m_1)(x_{2i} - m_2) \\ \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - m_1)(x_{2i} - m_2) & \frac{1}{n}\sum_{i=1}^{n}(x_{2i} - m_2)^2 \end{bmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t$$

相関係数

相関係数

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{1i} - m_1)(x_{2i} - m_2)}{\sigma_1\sigma_2}$$

ただし

$$\sigma_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{1i} - m_1)^2}, \sigma_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - m_2)^2}$$

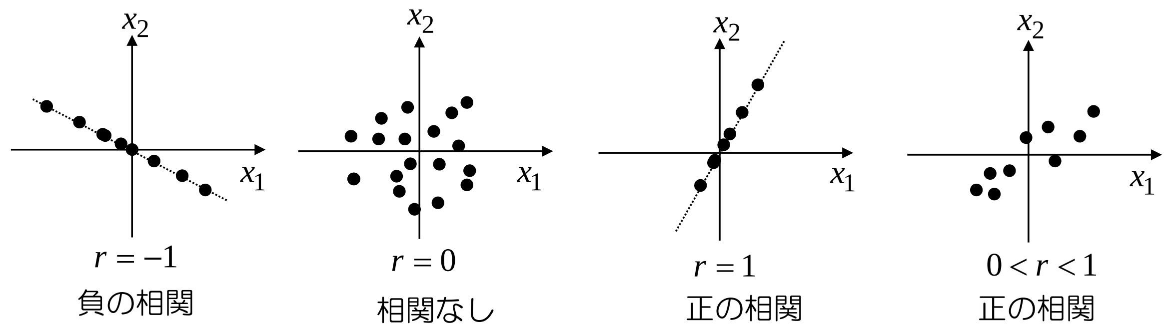
標準化した変数の相関

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{1i} - m_1)}{\sigma_1}\cdot\frac{(x_{2i} - m_2)}{\sigma_2} = \frac{1}{n}\sum_{i=1}^{n} u_{1i}\cdot u_{2i}$$

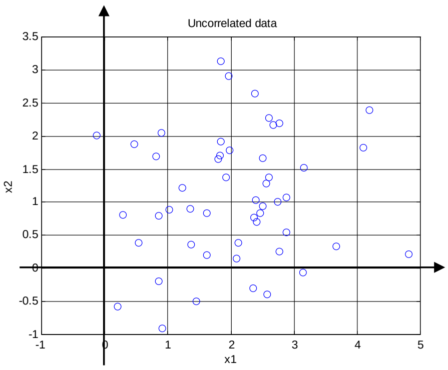
検証

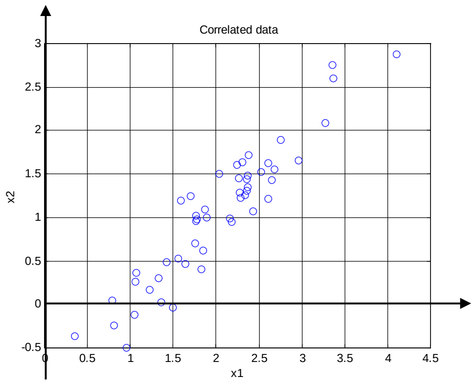
$x_{2i} = a x_{1i} + b$
ただし $a \neq 0, b$ は定数．
のとき， r を計算せよ．

相関係数の取りうる範囲： $-1 \leq r \leq 1$



共分散行列の例





共分散行列

$$\begin{bmatrix} 1.0964 & 0.1011 \\ 0.1011 & 0.8924 \end{bmatrix}$$

相関係数　$r = -0.224$

共分散行列

$$\begin{bmatrix} 0.5149 & 0.5100 \\ 0.5100 & 0.5225 \end{bmatrix}$$

相関係数　$r = 0.922$

多変数の共分散行列

一般にｄ変数の場合の共分散行列は

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

(k,l)成分はk番目の変数とl番目の変数の間の共分散を意味する.

$$c_{kl} = \sum_{i=1}^{n}(x_{ki}-m_k)(x_{li}-m_l)$$

もし，すべての変数が互いに無相関なら
共分散行列は対角行列になる.
（対角要素は，各変数の分散を表す）

$$\mathbf{C} = \frac{1}{n}\begin{bmatrix} c_{11} & 0 & \cdots & 0 \\ 0 & c_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_{dd} \end{bmatrix}$$