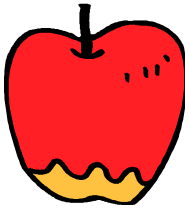
多変量解析

第9回　判別分析

萩原•篠田
情報理工学部

多変量情報の解析法

りんごを
買う



大きさ

色
つや
形
重さ
たたいた音

さまざまな
特性データ
から
総合判断

1特性評価ごとの単純合計ではない

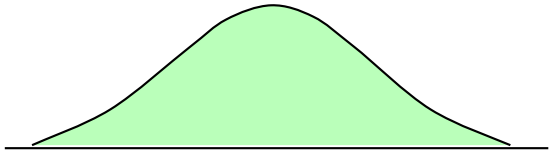
総合判断の
基準
（評価因子）
各特性の関わり

多変量解析

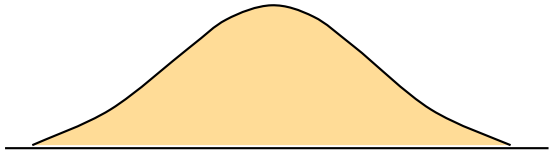
判別分析

りんご

固さ、甘さを測定



固さの分布

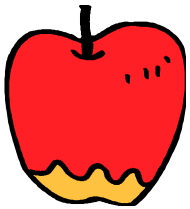


甘さの分布

同じ産地から出荷されたりんご？

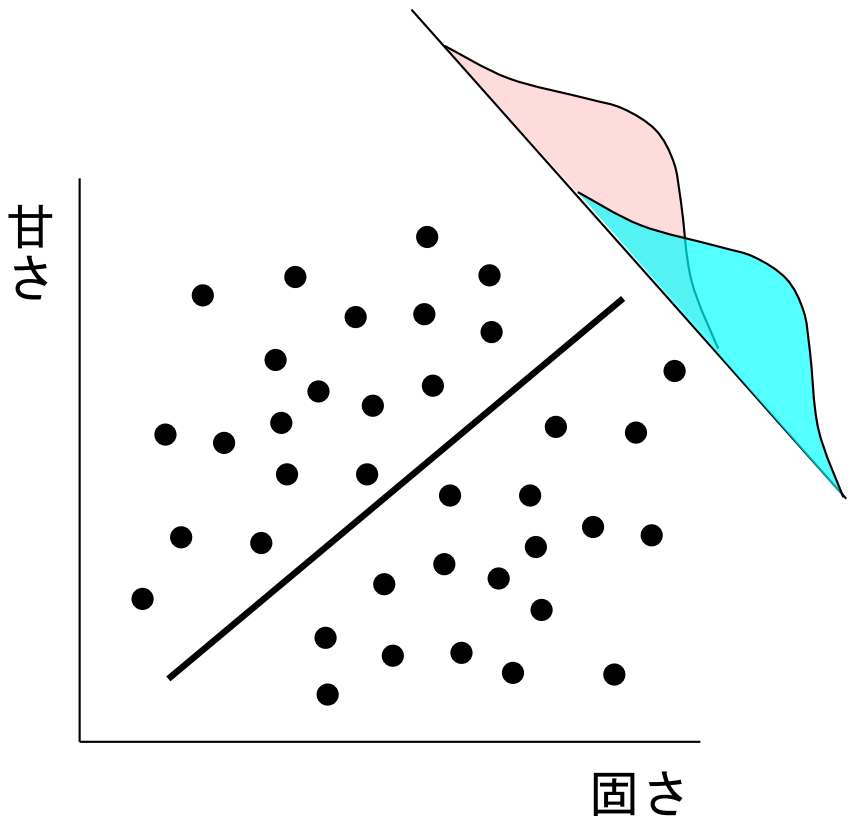
判別分析法

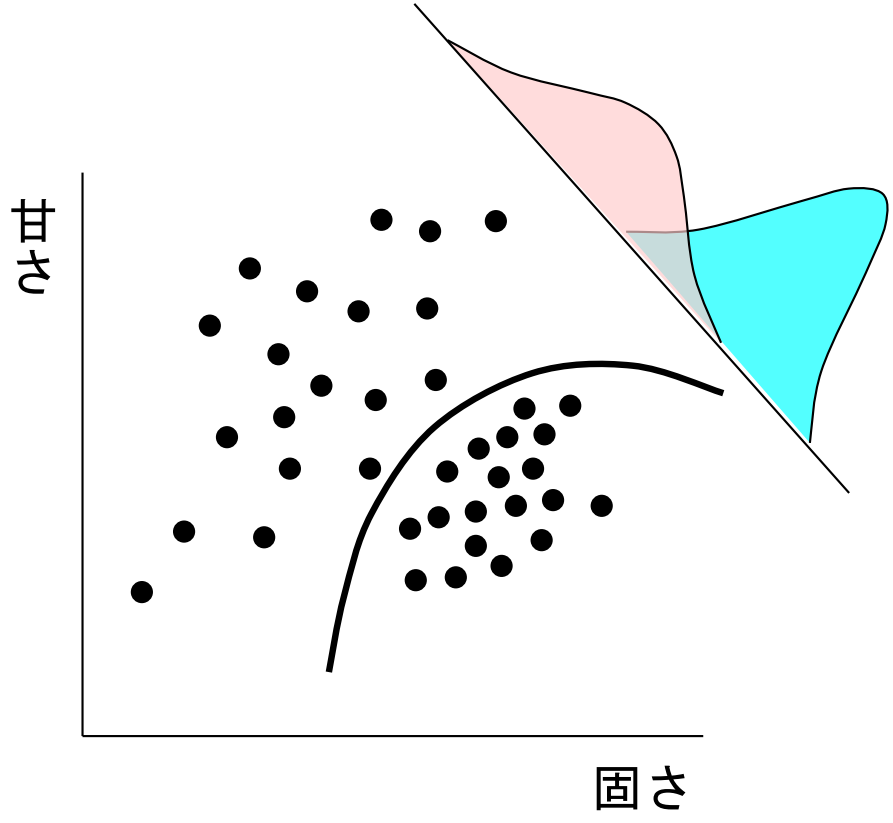
同じ産地から出荷されたりんご？



判別分析の手法

「固さ」と「甘さ」の組み合わせで産地が判別
→ 線引きの方法は？





2群の分散が等しい場合

線型判別関数

最もよく分離する直線を仮定し
その直線のどちら側に来るかを判別

2群の分散が等しくない場合

マハラノビスの距離

各グループの分布状態を考慮し
各グループの中心からの距離で判別

判別分析

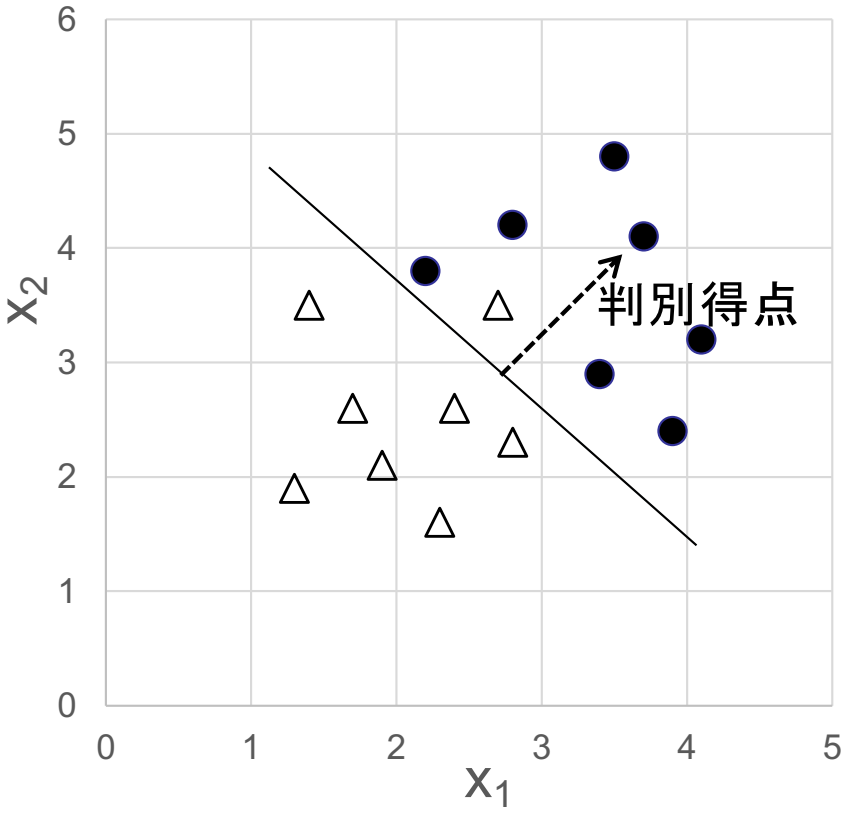
前立腺疾患を腫瘍マーカー1とマーカー2から予測可能か？

質的変数　　量的変数

患者 No.	前立腺疾患 y	マーカー1 x_1	マーカー2 x_2
1	前立腺ガン	3.4	2.9
2	前立腺ガン	3.9	2.4
3	前立腺ガン	2.2	3.8
4	前立腺ガン	3.5	4.8
5	前立腺ガン	4.1	3.2
6	前立腺ガン	3.7	4.1
7	前立腺ガン	2.8	4.2
8	前立腺肥大症	1.4	3.5
9	前立腺肥大症	2.4	2.6
10	前立腺肥大症	2.8	2.3
11	前立腺肥大症	1.7	2.6
12	前立腺肥大症	2.3	1.6
13	前立腺肥大症	1.9	2.1
14	前立腺肥大症	2.7	3.5
15	前立腺肥大症	1.3	1.9

keywords

線形判別関数、判別得点、
マハラノビスの距離、標準化



（例題）線型判別関数

疾患 ＜ G1前立腺ガン / G2前立腺肥大症

腫瘍マーカーで判定 マーカーA マーカーB

表　グループG1疾患とグループG2疾患のマーカーA、マーカーB

◆ グループG1

説明変数 / 患者	マーカーA	マーカーB
	x_1	x_2
1	3.4	2.9
2	3.9	2.4
3	2.2	3.8
4	3.5	4.8
5	4.1	3.2
6	3.7	4.1
7	2.8	4.2
G1平均	3.371	3.629

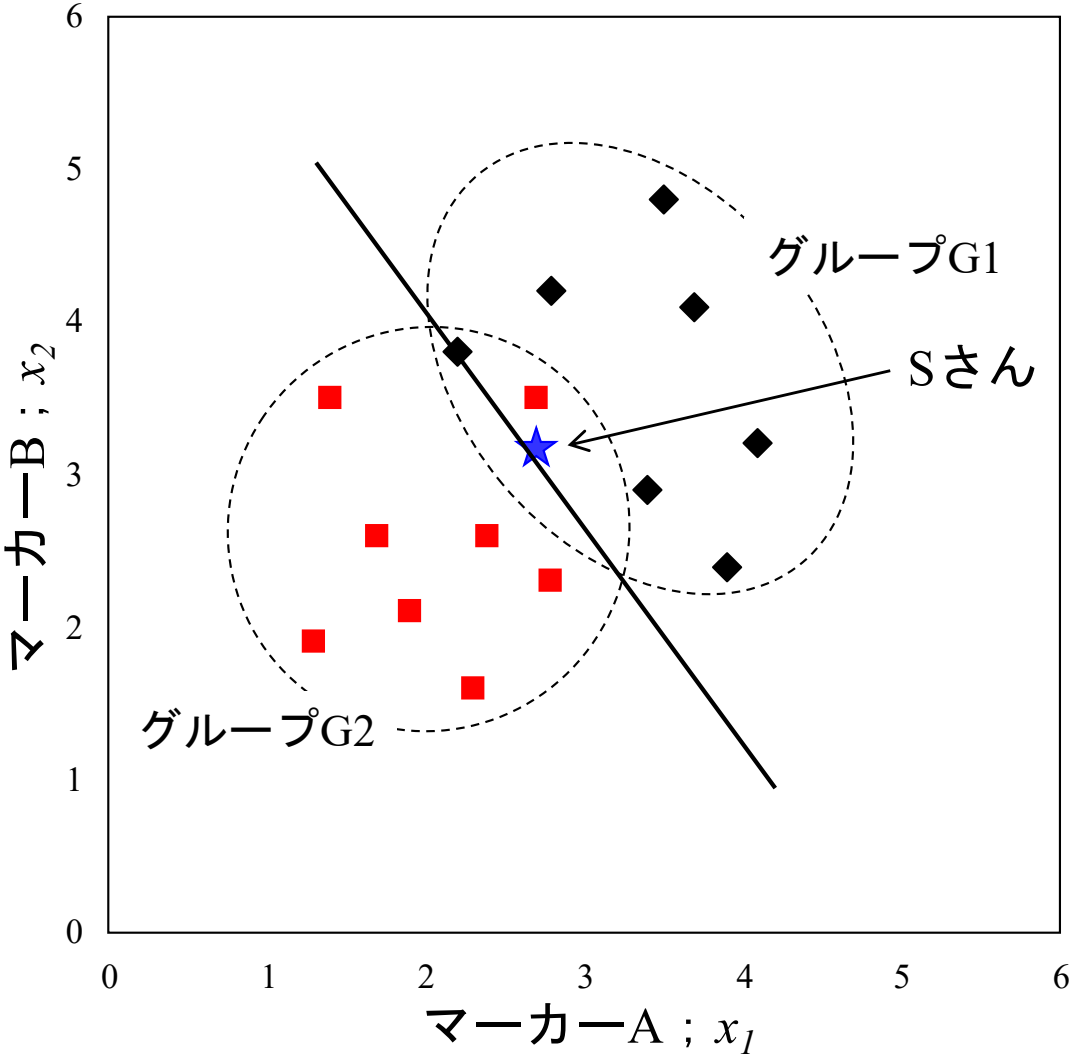
	マーカーA	マーカーB
	x_1	x_2
全平均	2.673	3.033

■ グループG2

説明変数 / 患者	マーカーA	マーカーB
	x_1	x_2
1	1.4	3.5
2	2.4	2.6
3	2.8	2.3
4	1.7	2.6
5	2.3	1.6
6	1.9	2.1
7	2.7	3.5
8	1.3	1.9
G2平均	2.063	2.513

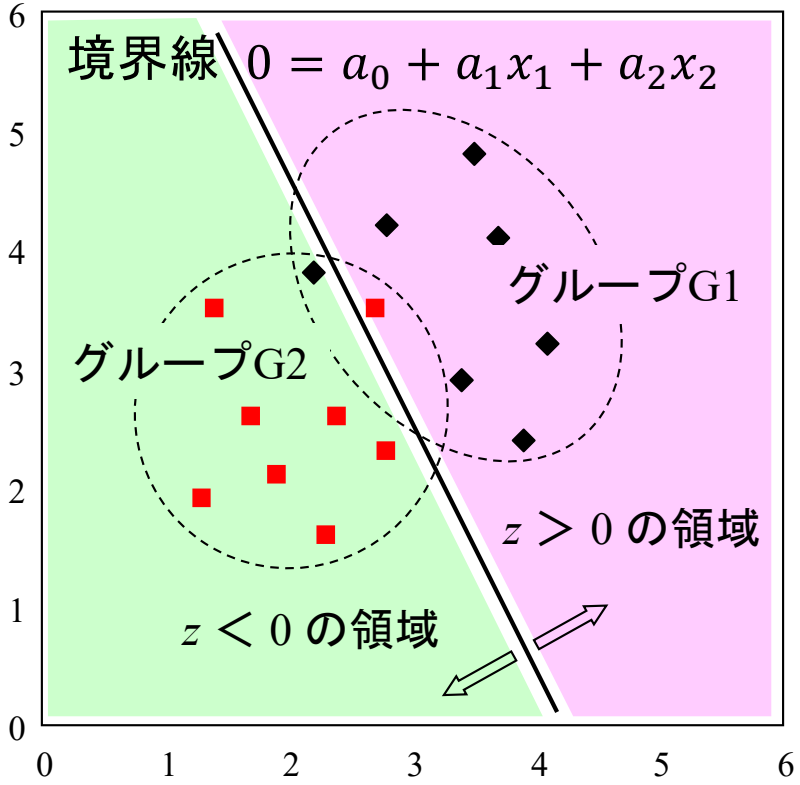
★ 16番目の被験者Sさんのマーカーの値は
マーカーA；x_1= 2.7、マーカーB；x_2= 3.1
Sさんはどちらのグループに属するのか

線形判別：線形判別関数で定義される判別得点の正負で判別



判別得点 zを線形判別関数で定義

$$z = a_0 + a_1x_1 + a_2x_2$$



点(p, q)と境界線の距離 ヘッセの標準形 $\dfrac{|a_0 + a_1p + a_2q|}{\sqrt{a_1^2 + a_2^2}}$

判別得点
$= 直線までの距離 \times \sqrt{a_1^2 + a_2^2}$

判別得点

	グループG1の判別得点	グループG2の判別得点
	$z_1^{(1)}=3.4a_1+2.9a_2+a_0$ $z_2^{(1)}=3.9a_1+2.4a_2+a_0$ $z_3^{(1)}=2.2a_1+3.8a_2+a_0$ $z_4^{(1)}=3.5a_1+4.8a_2+a_0$ $z_5^{(1)}=4.1a_1+3.2a_2+a_0$ $z_6^{(1)}=3.7a_1+4.1a_2+a_0$ $z_7^{(1)}=2.8a_1+4.2a_2+a_0$	$z_1^{(2)}=1.4a_1+3.5a_2+a_0$ $z_2^{(2)}=2.4a_1+2.6a_2+a_0$ $z_3^{(2)}=2.8a_1+2.3a_2+a_0$ $z_4^{(2)}=1.7a_1+2.6a_2+a_0$ $z_5^{(2)}=2.3a_1+1.6a_2+a_0$ $z_6^{(2)}=1.9a_1+2.1a_2+a_0$ $z_7^{(2)}=2.7a_1+3.5a_2+a_0$ $z_8^{(2)}=1.3a_1+1.9a_2+a_0$
グループ内平均	$\bar{z}^{(1)} = \bar{x}_1^{(1)}a_1 + \bar{x}_2^{(1)}a_2 + a_0$ $= 3.371a_1 + 3.629a_2 + a_0$	$\bar{z}^{(2)} = \bar{x}_1^{(2)}a_1 + \bar{x}_2^{(2)}a_2 + a_0$ $= 2.063a_1 + 2.513a_2 + a_0$
全平均	$\bar{z} = \bar{x}_1 a_1 + \bar{x}_2 a_2 + a_0 = 2.673a_1 + 3.033a_2 + a_0$	

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(z_j^{(i)}-\bar{z}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{z}^{(i)}-\bar{z}\right)^2 + \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(z_j^{(i)}-\bar{z}^{(i)}\right)^2$$

全変動SS_T　　群間変動SS_B　　群内変動SS_W

2群の分離を良くするには全変動に占める群間変動の割合（相関比）が大きくなるようにすればよい

線型判別関数
$z = a_0 + a_1x_1 + a_2x_2$
の係数(a_1, a_2)の決定

⟺

$F(a_1, a_2) = \dfrac{SS_B}{SS_T}$ の最大値を与える(a_1, a_2)の決定

$F(a_1, a_2) = \frac{SS_B}{SS_T}$ の最大値を与える(a_1, a_2)

$$\begin{cases} \dfrac{\partial F}{\partial a_1} = \dfrac{1}{SS_T}\left(\dfrac{\partial SS_B}{\partial a_1} - \dfrac{SS_B}{SS_T}\dfrac{\partial SS_T}{\partial a_1}\right) = 0 \\ \dfrac{\partial F}{\partial a_2} = \dfrac{1}{SS_T}\left(\dfrac{\partial SS_B}{\partial a_2} - \dfrac{SS_B}{SS_T}\dfrac{\partial SS_T}{\partial a_2}\right) = 0 \end{cases}$$

どちらからも同じ$\frac{a_1}{a_2}$の方程式が得られ

それを解くと $\dfrac{a_1}{a_2} = 1.606$

境界線 $0 = a_0 + a_1 x_1 + a_2 x_2$

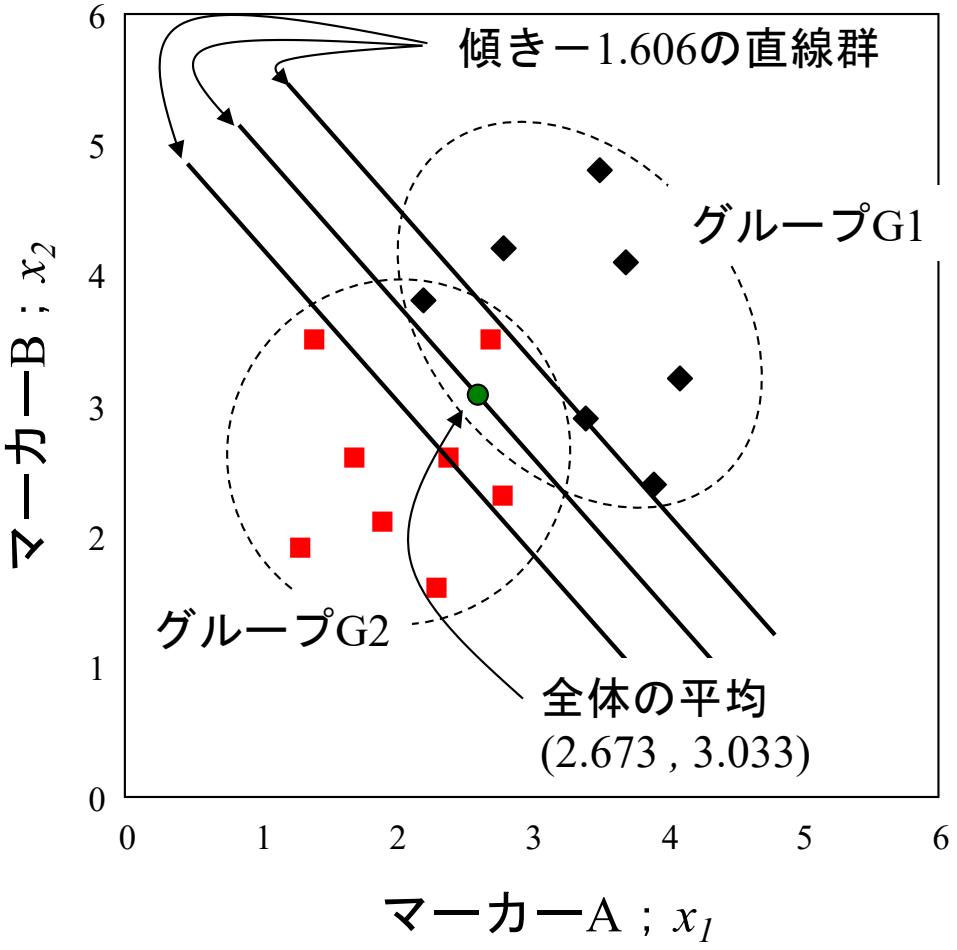
変形すると $0 = \dfrac{a_0}{a_2} + \dfrac{a_1}{a_2} x_1 + x_2$

境界線は全平均 (2.673 , 3.033) を通る

代入して $\dfrac{a_0}{a_2} = -1.606 x_1 - x_2 = -7.325$

最良の境界線は $0 = -7.325 + 1.606 x_1 + x_2$

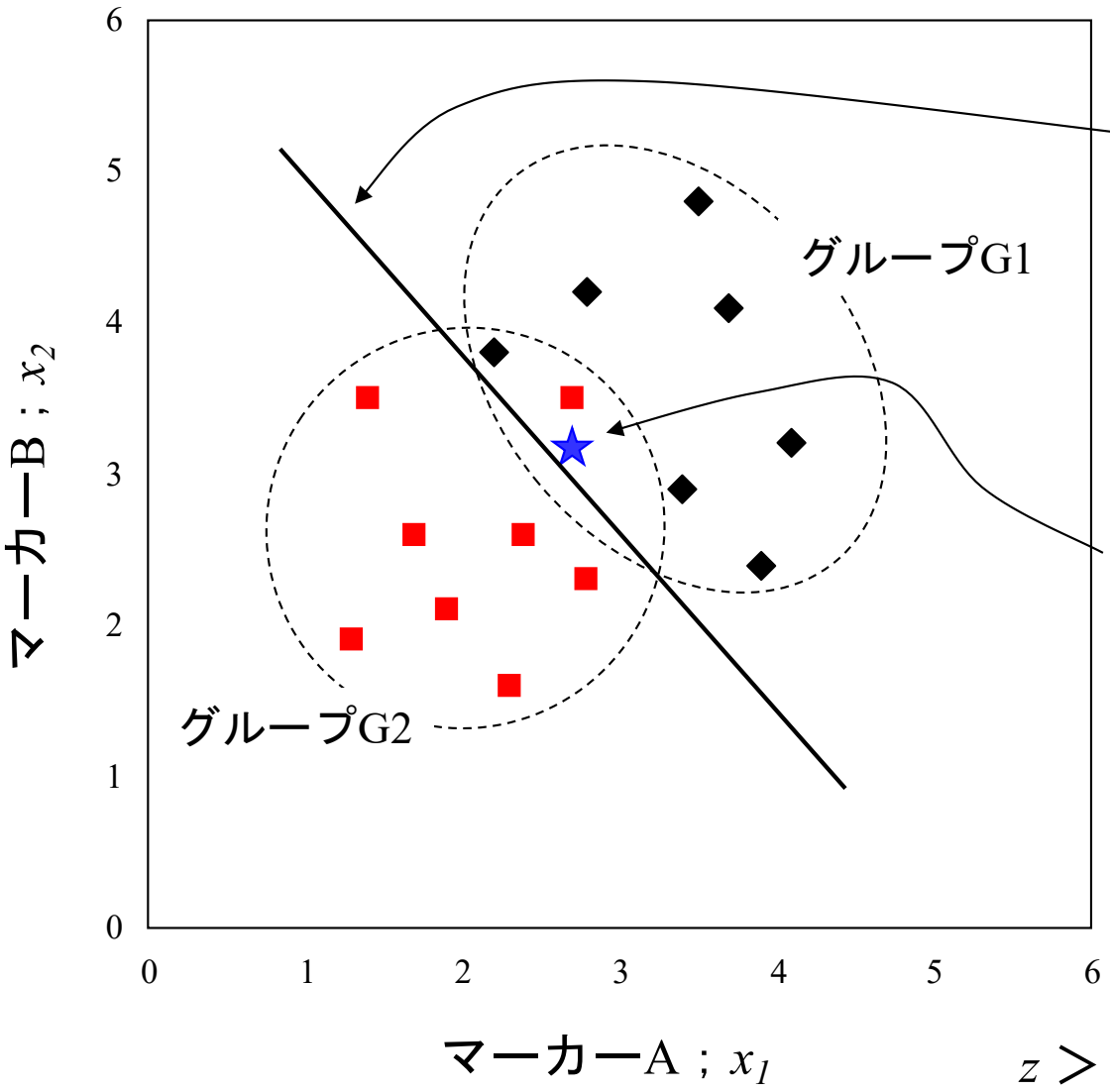
線型判別関数は $z = -7.325 + 1.606 x_1 + x_2$



$$u(x) = y(x) \cdot z(x)$$
$$u'(x) = \frac{du}{dx} = \frac{dy}{dx} \cdot z + y \cdot \frac{dz}{dx} = y'(x) \cdot z(x) + y(x) \cdot z'(x)$$

$$v(x) = \frac{1}{y(x)} = \{y(x)\}^{-1}$$
$$v'(x) = \frac{dv}{dx} = \frac{dv}{dy} \cdot \frac{dy}{dx} = (-1) \cdot \{y(x)\}^{-2} \cdot y'(x) = -\frac{y'(x)}{\{y(x)\}^2} = -\{v(x)\}^2 \cdot y'(x)$$

$$w(x) = \frac{z(x)}{y(x)} = z(x) \cdot \{y(x)\}^{-1}$$
$$w'(x) = \frac{dw}{dx} = \frac{dz}{dx} \cdot \frac{1}{y} + z \cdot \frac{d}{dx}\left(\frac{1}{y}\right) = \frac{z'(x)}{y(x)} + z(x) \cdot \frac{-y'(x)}{\{y(x)\}^2} = \frac{z'(x) - w(x) \cdot y'(x)}{y(x)}$$



$$0 = 1.606x_1 + x_2 - 7.325$$

線型判別関数に代入

$$1.606 \times 2.7 + 3.1 - 7.325 = 0.111$$

$z > 0$：グループG1

$z < 0$：グループG2

Sさんはグループ G1に属すると判別される

別解

G_i の j 番目データの判別得点　$z_j^{(i)} = a_0 + x_{1j}^{(i)} a_1 + x_{2j}^{(i)} a_2$

G_i の判別得点の平均　$\bar{z}^{(i)} = a_0 + \bar{x}_1^{(i)} a_1 + \bar{x}_2^{(i)} a_2$

判別得点の全平均　$\bar{z} = a_0 + \bar{x}_1 a_1 + \bar{x}_2 a_2$

全変動 SS_T = 群間変動 SS_B + 群内変動 SS_W

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(z_j^{(i)} - \bar{z}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left\{\left(x_{1j}^{(i)} - \bar{x}_1\right)a_1 + \left(x_{2j}^{(i)} - \bar{x}_2\right)a_2\right\}^2$$

$$SS_B = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{z}^{(i)} - \bar{z}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left\{\left(\bar{x}_1^{(i)} - \bar{x}_1\right)a_1 + \left(\bar{x}_2^{(i)} - \bar{x}_2\right)a_2\right\}^2$$

$$SS_W = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(z_j^{(i)} - \bar{z}^{(i)}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left\{\left(x_{1j}^{(i)} - \bar{x}_1^{(i)}\right)a_1 + \left(x_{2j}^{(i)} - \bar{x}_2^{(i)}\right)a_2\right\}^2$$

2群の分離を良くするには群間変動SS_B/全変動SS_T（=相関比）を最大にする

$F(a_1, a_2) = \dfrac{SS_B}{SS_T}$ の最大化　⟺　$G(a_1, a_2) = \dfrac{SS_W}{SS_B}$ の最小化

$$F(a_1, a_2) = \frac{SS_B}{SS_T} = \frac{SS_B}{SS_B + SS_W} = \frac{1}{1 + \dfrac{SS_W}{SS_B}} = \frac{1}{1 + G(a_1, a_2)}$$

別解

ベクトル表現 $\boldsymbol{x}_j{}^{(i)} = \begin{pmatrix} x_{1j}{}^{(i)} \\ x_{2j}{}^{(i)} \end{pmatrix}$ $\bar{\boldsymbol{x}}^{(i)} = \begin{pmatrix} \bar{x}_1^{(i)} \\ \bar{x}_2^{(i)} \end{pmatrix}$ $\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix}$ $\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$

$$SS_B = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{z}^{(i)} - \bar{z}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left\{\left(\bar{x}_1^{(i)} - \bar{x}_1\right)a_1 + \left(\bar{x}_2^{(i)} - \bar{x}_2\right)a_2\right\}^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\boldsymbol{a}'\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)'\boldsymbol{a}$$

$$SS_W = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(z_j^{(i)} - \bar{z}^{(i)}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left\{\left(x_{1j}^{(i)} - \bar{x}_1^{(i)}\right)a_1 + \left(x_{2j}^{(i)} - \bar{x}_2^{(i)}\right)a_2\right\}^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\boldsymbol{a}'\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)'\boldsymbol{a}$$

偏微分=0

$$\frac{\partial G}{\partial \boldsymbol{a}} = \frac{1}{SS_B}\left(\frac{\partial SS_W}{\partial \boldsymbol{a}} - \frac{SS_W}{SS_B}\frac{\partial SS_B}{\partial \boldsymbol{a}}\right) = \begin{pmatrix} \dfrac{\partial G}{\partial a_1} \\ \dfrac{\partial G}{\partial a_2} \end{pmatrix} = \frac{1}{SS_B}\left\{\begin{pmatrix} \dfrac{\partial SS_W}{\partial a_1} \\ \dfrac{\partial SS_W}{\partial a_2} \end{pmatrix} - \frac{SS_W}{SS_B}\begin{pmatrix} \dfrac{\partial SS_B}{\partial a_1} \\ \dfrac{\partial SS_B}{\partial a_2} \end{pmatrix}\right\} = 0$$

$$\therefore \ \frac{\partial SS_W}{\partial \boldsymbol{a}} = G\frac{\partial SS_B}{\partial \boldsymbol{a}} \quad \Rightarrow \quad \boxed{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)'\boldsymbol{a} = G\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)'\boldsymbol{a}}$$

$$\frac{\partial SS_W}{\partial \boldsymbol{a}} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{x}_j{}^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)'\boldsymbol{a}$$

$$\frac{\partial SS_W}{\partial a_1} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(x_{1j}^{(i)} - \bar{x}_1^{(i)}\right)\left\{\left(x_{1j}^{(i)} - \bar{x}_1^{(i)}\right)a_1 + \left(x_{2j}^{(i)} - \bar{x}_2^{(i)}\right)a_2\right\}$$

$$\frac{\partial SS_W}{\partial a_2} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(x_{2j}^{(i)} - \bar{x}_2^{(i)}\right)\left\{\left(x_{1j}^{(i)} - \bar{x}_1^{(i)}\right)a_1 + \left(x_{2j}^{(i)} - \bar{x}_2^{(i)}\right)a_2\right\}$$

$$\frac{\partial SS_B}{\partial \boldsymbol{a}} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)'\boldsymbol{a}$$

$$\frac{\partial SS_B}{\partial a_1} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{x}_1^{(i)} - \bar{x}_1\right)\left\{\left(\bar{x}_1^{(i)} - \bar{x}_1\right)a_1 + \left(\bar{x}_2^{(i)} - \bar{x}_2\right)a_2\right\}$$

$$\frac{\partial SS_B}{\partial a_2} = 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{x}_2^{(i)} - \bar{x}_2\right)\left\{\left(\bar{x}_1^{(i)} - \bar{x}_1\right)a_1 + \left(\bar{x}_2^{(i)} - \bar{x}_2\right)a_2\right\}$$

別解

前ページより $\boxed{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)}\right)'\boldsymbol{a}=G\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}}\right)'\boldsymbol{a}}$

左辺

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)}\right)'\boldsymbol{a}=\sum_{i=1}^{2}\sum_{j=1}^{n_i}\begin{pmatrix}\left(x_{1j}^{(i)}-\bar{x}_1^{(i)}\right)^2 & \left(x_{1j}^{(i)}-\bar{x}_1^{(i)}\right)\left(x_{2j}^{(i)}-\bar{x}_2^{(i)}\right)\\ \left(x_{2j}^{(i)}-\bar{x}_1^{(i)}\right)\left(x_{1j}^{(i)}-\bar{x}_1^{(i)}\right) & \left(x_{2j}^{(i)}-\bar{x}_2^{(i)}\right)^2\end{pmatrix}\begin{pmatrix}a_1\\a_2\end{pmatrix}$$

$$=\sum_{i=1}^{2}(n_i-1)\begin{pmatrix}Var\left(x_1^{(i)}\right) & Cov\left(x_1^{(i)},x_2^{(i)}\right)\\ Cov\left(x_2^{(i)},x_1^{(i)}\right) & Var\left(x_2^{(i)}\right)\end{pmatrix}\begin{pmatrix}a_1\\a_2\end{pmatrix}=\boldsymbol{S_W a}$$

右辺

$$G\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}}\right)'\boldsymbol{a}=G\left\{n_1\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}\right)'+n_2\left(\bar{\boldsymbol{x}}^{(2)}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(2)}-\bar{\boldsymbol{x}}\right)'\right\}\boldsymbol{a}$$

$$=G\left(\frac{n_1n_2}{n_1+n_2}\right)^2\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right)\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right)'\boldsymbol{a}=k\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right)$$

ただし

$$\bar{\boldsymbol{x}}=\frac{n_1\bar{\boldsymbol{x}}^{(1)}+n_2\bar{\boldsymbol{x}}^{(2)}}{n_1+n_2}$$

$$k=G\left(\frac{n_1n_2}{n_1+n_2}\right)^2\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right)'\boldsymbol{a}=G\left(\frac{n_1n_2}{n_1+n_2}\right)^2\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\left\{\left(\bar{x}_1^{(1)}-\bar{x}_1^{(2)}\right)a_1+\left(\bar{x}_2^{(1)}-\bar{x}_2^{(2)}\right)a_2\right\}$$

以上より

$$\boldsymbol{S_W a}=k\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right)$$

$$\begin{pmatrix}a_1\\a_2\end{pmatrix}=k\boldsymbol{S_W}^{-1}\begin{pmatrix}\bar{x}_1^{(1)}-\bar{x}_1^{(2)}\\ \bar{x}_2^{(1)}-\bar{x}_2^{(2)}\end{pmatrix}$$

$z=a_0+a_1x_1+a_2x_2=a_2\left(\frac{a_0}{a_2}+\frac{a_1}{a_2}x_1+x_2\right)$の正負で判別

→ a_1/a_2が求まればよい → ここでは $k=1$として

$$\boxed{\begin{pmatrix}a_1\\a_2\end{pmatrix}=\boldsymbol{S_W}^{-1}\begin{pmatrix}\bar{x}_1^{(1)}-\bar{x}_1^{(2)}\\ \bar{x}_2^{(1)}-\bar{x}_2^{(2)}\end{pmatrix}}$$

別解

G1

説明変数 / 患者	マーカーA $x_{1j}^{(1)}$	マーカーB $x_{2j}^{(1)}$
1	3.4	2.9
2	3.9	2.4
3	2.2	3.8
4	3.5	4.8
5	4.1	3.2
6	3.7	4.1
7	2.8	4.2
$\bar{x}^{(1)}$	3.371	3.629
Var	0.4390	0.6957
Cov	-0.2007	

G2

説明変数 / 患者	マーカーA $x_{1j}^{(2)}$	マーカーB $x_{2j}^{(2)}$
1	1.4	3.5
2	2.4	2.6
3	2.8	2.3
4	1.7	2.6
5	2.3	1.6
6	1.9	2.1
7	2.7	3.5
8	1.3	1.9
$\bar{x}^{(2)}$	2.063	2.513
Var	0.3284	0.4842
Cov	0.01911	

	マーカーA $\bar{x}_1$	マーカーB $\bar{x}_2$
$\bar{x}$	2.673	3.033

$$\boldsymbol{S_W} = \sum_{i=1}^{2}(n_i - 1)\begin{pmatrix} Var\left(x_1^{(i)}\right) & Cov\left(x_1^{(i)}, x_2^{(i)}\right) \\ Cov\left(x_2^{(i)}, x_1^{(i)}\right) & Var\left(x_2^{(i)}\right) \end{pmatrix} = 6\begin{pmatrix} 0.4390 & -0.2007 \\ -0.2007 & 0.6957 \end{pmatrix} + 7\begin{pmatrix} 0.3284 & 0.01911 \\ 0.01911 & 0.4842 \end{pmatrix}$$

$$= \begin{pmatrix} 4.9330 & -1.0705 \\ -1.0705 & 7.5630 \end{pmatrix}$$

$$\begin{pmatrix} \bar{x}_1^{(1)} - \bar{x}_1^{(2)} \\ \bar{x}_2^{(1)} - \bar{x}_2^{(2)} \end{pmatrix} = \begin{pmatrix} 3.371 - 2.063 \\ 3.629 - 2.513 \end{pmatrix} = \begin{pmatrix} 1.308 \\ 1.116 \end{pmatrix}$$

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \boldsymbol{S_W}^{-1}\begin{pmatrix} \bar{x}_1^{(1)} - \bar{x}_1^{(2)} \\ \bar{x}_2^{(1)} - \bar{x}_2^{(2)} \end{pmatrix} = \begin{pmatrix} 0.20914 & 0.02960 \\ 0.02960 & 0.13641 \end{pmatrix}\begin{pmatrix} 1.308 \\ 1.116 \end{pmatrix} = \begin{pmatrix} 0.3068 \\ 0.1910 \end{pmatrix}$$

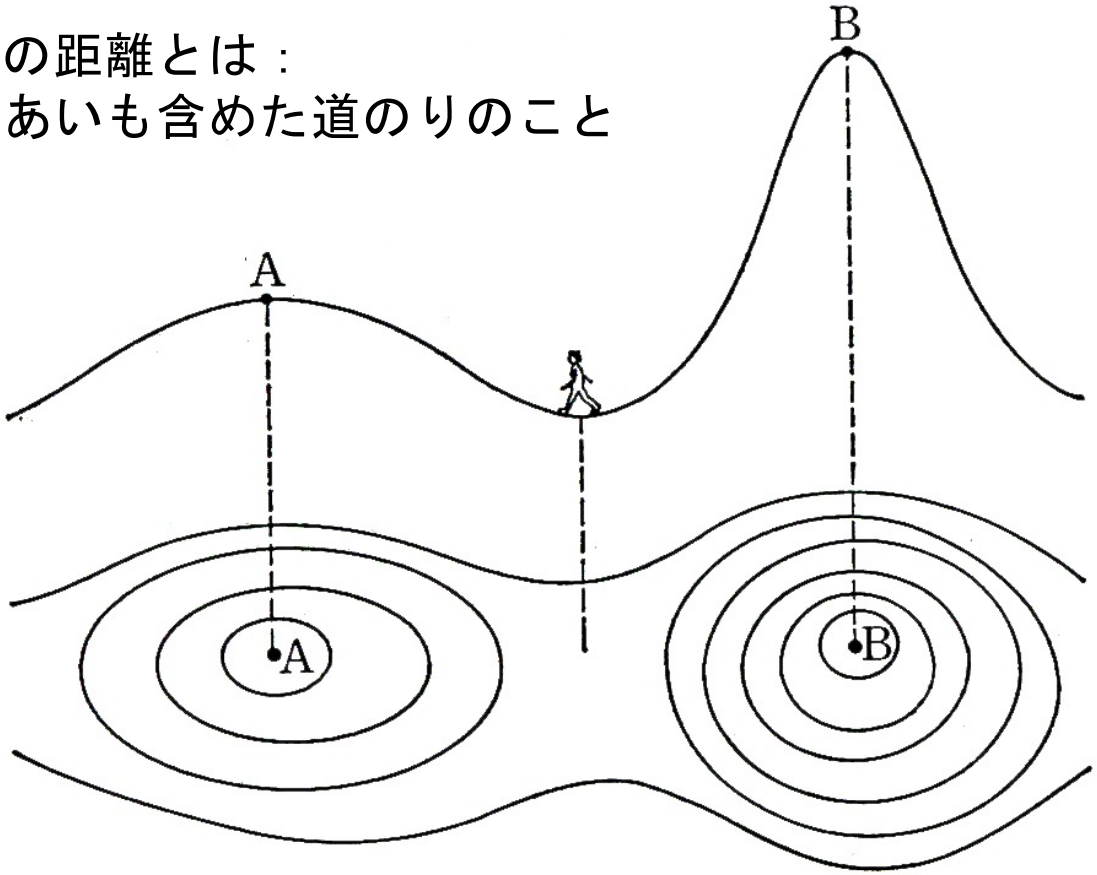
$$\frac{a_1}{a_2} = \frac{0.3068}{0.1910} = 1.606$$

マハラノビスの距離

マハラノビスの距離による判別 ― 各グループの分布状態を考慮し
各グループの中心からの距離で判別

山登りの場合：距離は近くても急斜面であると道のりは長く感じられる
：距離は遠くてもなだらかであると道のりは短く感じられる

マハラノビスの距離とは：
“傾き”ぐあいも含めた道のりのこと



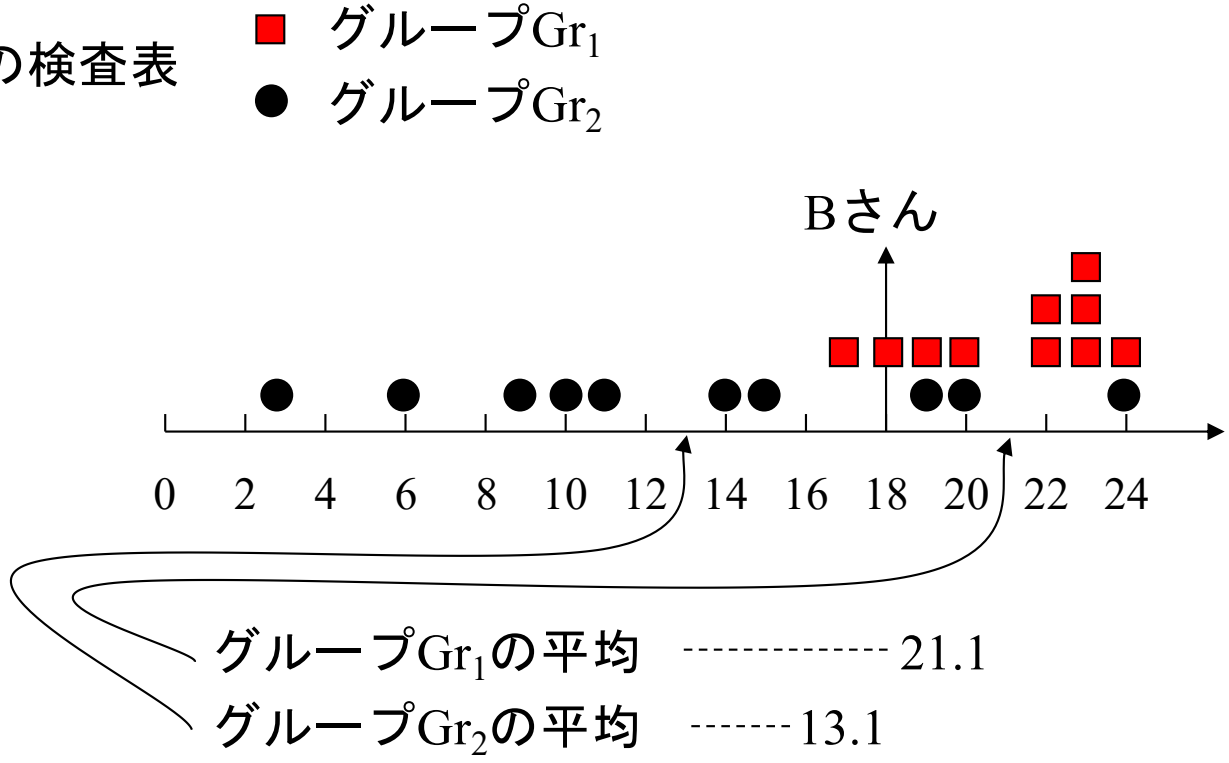
マハラノビスの距離による判別分析

1 変量のマハラノビスの距離

Bさんの検査結果は18

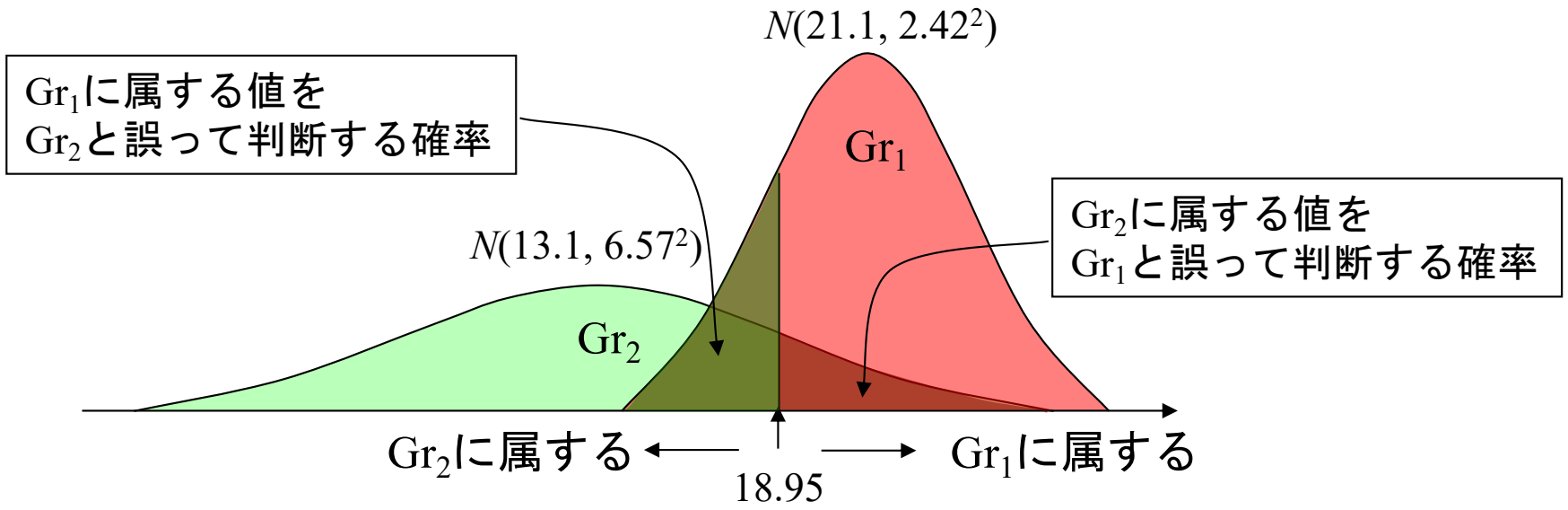
グループGr_1とグループGr_2の検査表

グループGr_1	グループGr_2
22	24
20	19
23	11
23	6
17	9
24	10
23	3
18	15
22	14
19	20



Bさんの グループGr_1の平均との差 21.1－18 = 3.1
グループGr_2の平均との差 18－13.1 = 4.9

BさんはグループGr_1により近いと思われる。しかし…



はじめBさんはグループGr_1に入るのではないかと判断された
しかし分散を考慮に入れたマハラノビスの距離では…

マハラノビスの距離

$$D^2 = \frac{(x-\bar{x})^2}{s^2}$$

$$D_1{}^2 = \frac{(x-\bar{x}_1)^2}{s_1{}^2} = \frac{(x-21.1)^2}{2.42^2}$$

$$D_2{}^2 = \frac{(x-\bar{x}_2)^2}{s_2{}^2} = \frac{(x-13.1)^2}{6.57^2}$$

Bさんの場合は $x = 18$

$$D_1{}^2 = 1.28^2$$
$$D_2{}^2 = 0.75^2$$

F検定による等分散性の検定：
Gr_1の分散 $\neq$ Gr_2の分散　を確認済

$D_1 > D_2$　BさんはGr_1よりもGr_2に近いと判断され，Gr_2に属すると判別される

2 変数のマハラノビスの距離

1 変数 x の場合

x の平均 $\bar{x}$

x の分散 s^2

マハラノビスの距離

$$D^2 = \frac{(x-\bar{x})^2}{s^2}$$

2 変数 x_1, x_2 の場合

$$D^2 = \frac{(x-\bar{x})^2}{s^2} = (x-\bar{x}) \cdot \frac{1}{s^2} \cdot (x-\bar{x}) = \underbrace{(x-\bar{x})}_{\text{ベクトルに}} \cdot \underbrace{(s^2)^{-1}}_{\text{分散共分散行列に}} \cdot \underbrace{(x-\bar{x})}_{\text{ベクトルに}}$$

2変数のマハラノビスの距離

$$D^2 = (x_1 - \bar{x}_1, x_2 - \bar{x}_2) \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{pmatrix}$$

分散共分散行列
の逆行列

線型判別関数の例題で用いた

G1とG2のマーカーAとマーカーBのデータについて具体的に計算

G1

説明変数 / 患者	マーカーA $x_{1j}^{(1)}$	マーカーB $x_{2j}^{(1)}$
1	3.4	2.9
2	3.9	2.4
3	2.2	3.8
4	3.5	4.8
5	4.1	3.2
6	3.7	4.1
7	2.8	4.2
$\bar{x}^{(1)}$	3.371	3.629
Var	0.4390	0.6957
Cov	-0.2007	

G1のマハラノビスの距離${D_1}^2$

$$\begin{pmatrix} \bar{x}_1^{(1)} \\ \bar{x}_2^{(1)} \end{pmatrix} = \begin{pmatrix} 3.371 \\ 3.629 \end{pmatrix}$$

$$\begin{pmatrix} s_{11}^{(1)} & s_{12}^{(1)} \\ s_{21}^{(1)} & s_{22}^{(1)} \end{pmatrix} = \begin{pmatrix} 0.439 & -0.201 \\ -0.201 & 0.696 \end{pmatrix}$$

$$\begin{pmatrix} s_{11}^{(1)} & s_{12}^{(1)} \\ s_{21}^{(1)} & s_{22}^{(1)} \end{pmatrix}^{-1} = \begin{pmatrix} 2.625 & 0.758 \\ 0.758 & 1.656 \end{pmatrix}$$

$$\begin{aligned} {D_1}^2 &= \left(x_1 - \bar{x}_1^{(1)}, x_2 - \bar{x}_2^{(1)}\right) \begin{pmatrix} s_{11}^{(1)} & s_{12}^{(1)} \\ s_{21}^{(1)} & s_{22}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \bar{x}_1^{(1)} \\ x_2 - \bar{x}_2^{(1)} \end{pmatrix} \\ &= (x_1 - 3.371, x_2 - 3.629) \begin{pmatrix} 2.625 & 0.758 \\ 0.758 & 1.656 \end{pmatrix} \begin{pmatrix} x_1 - 3.371 \\ x_2 - 3.629 \end{pmatrix} \\ &= 2.625{x_1}^2 + 1.656{x_2}^2 + 2\times 0.758 x_1 x_2 - 23.200 x_1 - 17.128 x_2 + 70.182 \end{aligned}$$

線型判別関数の例題で用いた

G1とG2のマーカーAとマーカーBのデータについて具体的に計算

G2

説明変数 / 患者	マーカーA $x_{1j}^{(2)}$	マーカーB $x_{2j}^{(2)}$
1	1.4	3.5
2	2.4	2.6
3	2.8	2.3
4	1.7	2.6
5	2.3	1.6
6	1.9	2.1
7	2.7	3.5
8	1.3	1.9
$\bar{x}^{(2)}$	2.063	2.513
Var	0.3284	0.4842
Cov	0.01911	

G2のマハラノビスの距離${D_2}^2$

$$\begin{pmatrix} \bar{x}_1^{(2)} \\ \bar{x}_2^{(2)} \end{pmatrix} = \begin{pmatrix} 2.063 \\ 2.513 \end{pmatrix}$$

$$\begin{pmatrix} s_{11}^{(2)} & s_{12}^{(2)} \\ s_{21}^{(2)} & s_{22}^{(2)} \end{pmatrix} = \begin{pmatrix} 0.328 & 0.0191 \\ 0.0191 & 0.484 \end{pmatrix}$$

$$\begin{pmatrix} s_{11}^{(2)} & s_{12}^{(2)} \\ s_{21}^{(2)} & s_{22}^{(2)} \end{pmatrix}^{-1} = \begin{pmatrix} 3.056 & -0.120 \\ -0.120 & 2.071 \end{pmatrix}$$

$$\begin{aligned} {D_2}^2 &= \left(x_1 - \bar{x}_1^{(2)}, x_2 - \bar{x}_2^{(2)}\right) \begin{pmatrix} s_{11}^{(2)} & s_{12}^{(2)} \\ s_{21}^{(2)} & s_{22}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \bar{x}_1^{(2)} \\ x_2 - \bar{x}_2^{(2)} \end{pmatrix} \\ &= (x_1 - 2.063, x_2 - 2.513) \begin{pmatrix} 3.056 & -0.120 \\ -0.120 & 2.071 \end{pmatrix} \begin{pmatrix} x_1 - 2.063 \\ x_2 - 2.513 \end{pmatrix} \\ &= 3.056{x_1}^2 + 2.071{x_2}^2 - 2\times 0.120 x_1 x_2 - 12.005 x_1 - 9.913 x_2 + 24.839 \end{aligned}$$

マハラノビスの距離による境界線

線型判別関数による判別分析の場合
(x_1, x_2)平面を1本の直線

$$0 = 1.606x_1 + x_2 - 7.325$$

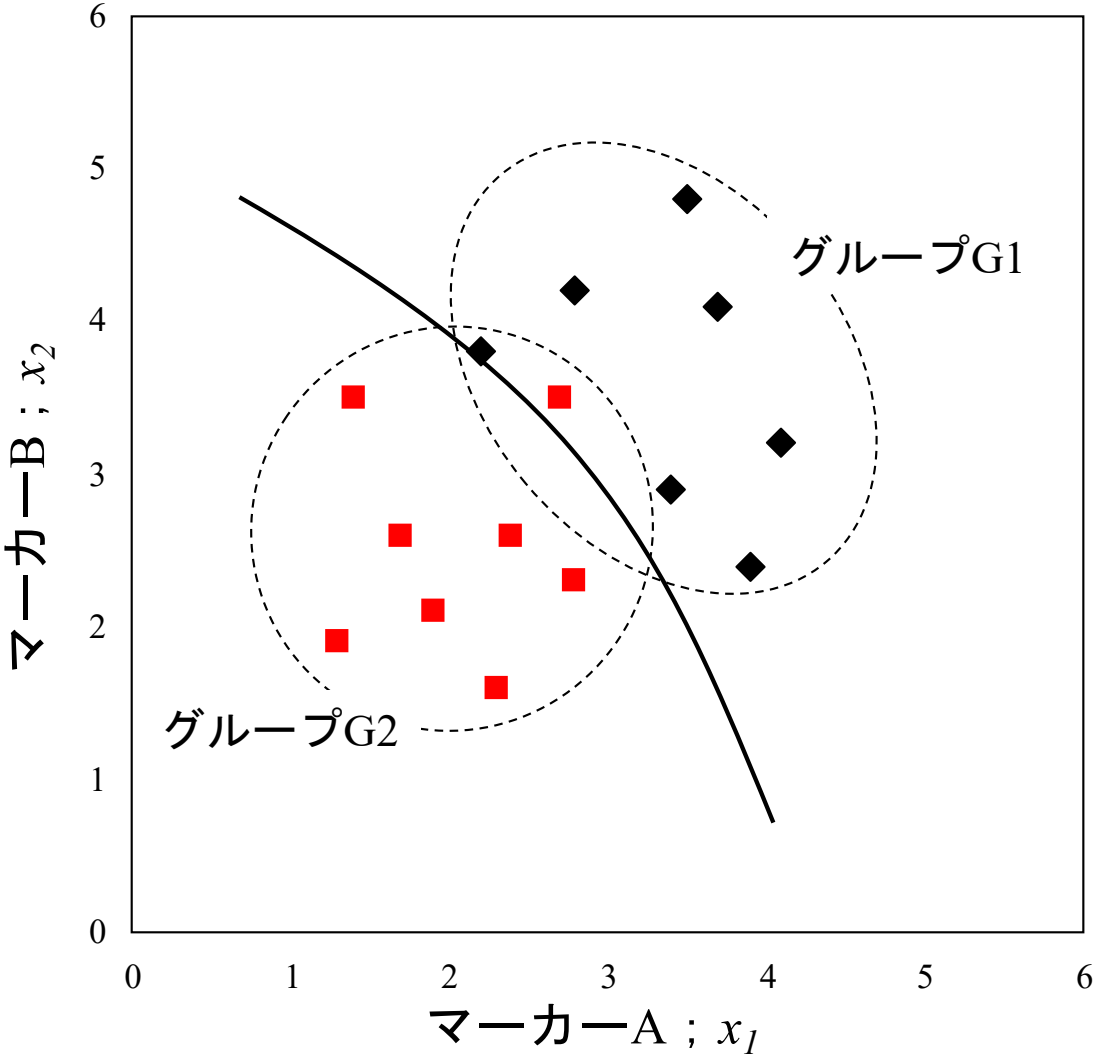
で2つの領域に分けた

同様に

マハラノビスの距離の場合の境界線は

$$0 = {D_1}^2 - {D_2}^2$$

で2つの領域に分ける



$${D_1}^2 - {D_2}^2 = 2.625{x_1}^2 + 1.656{x_2}^2 + 2\times 0.758x_1x_2 - 23.200x_1 - 17.128x_2 + 70.182$$
$$-(3.056{x_1}^2 + 2.071{x_2}^2 - 2\times 0.120x_1x_2 - 12.005x_1 - 9.913x_2 + 24.839)$$

マハラノビスの距離による判別分析の境界線は

$$0 = -0.431{x_1}^2 - 0.415{x_2}^2 + 1.756x_1x_2 - 11.195x_1 - 7.215x_2 + 45.344$$

正答率と誤判別率

グループG1に属しているのに
グループG2と誤って判別

グループG2に属しているのに
グループG1と誤って判別

線型判別関数の正答率と誤判別率

グループG1

No.	判別得点	
1	1.035	正
2	1.338	正
3	0.008	正
4	3.096	正
5	2.460	正
6	2.717	正
7	1.372	正

グループG2

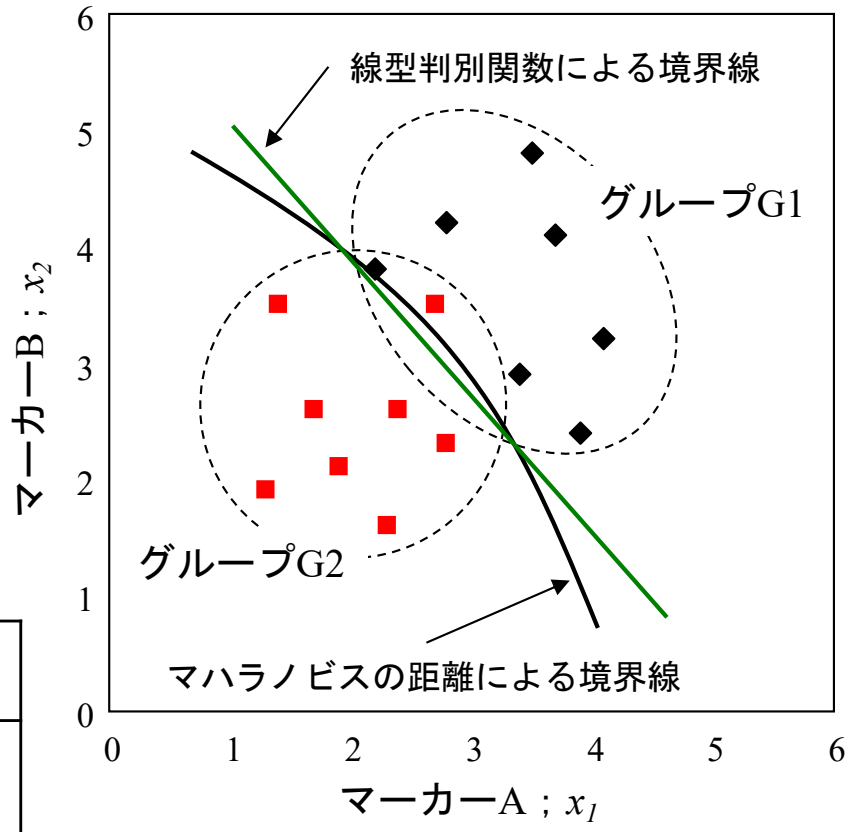
No.	判別得点	
1	-1.577	正
2	-0.871	正
3	-0.528	正
4	-1.995	正
5	-2.031	正
6	-2.174	正
7	0.511	誤
8	-3.337	正

正答率 $= \frac{7}{7} = 1$

誤判別率 $= \frac{0}{7} = 0$

正答率 $= \frac{7}{8} = 0.875$

誤判別率 $= \frac{1}{8} = 0.125$



マハラノビスの距離による
正答率と誤判別率も同様に
求めることができる

F検定による等分散性の検定：
G1の分散 = G2の分散　を確認済

判別分析

①判別分析の2種類の手法を図を用いて説明せよ。

②線型判別関数の導出において2群の分離をできるだけ良くするための考え方を式を用いて説明せよ。