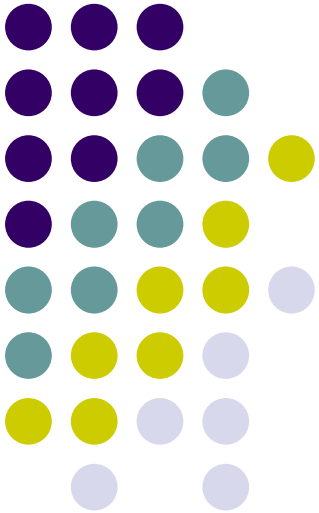
第4章　多変量解析

4. 外的基準が分類の場合の分析法
   4.1 判別分析
   4.2 数量化Ⅱ類
5. 数量化Ⅲ類



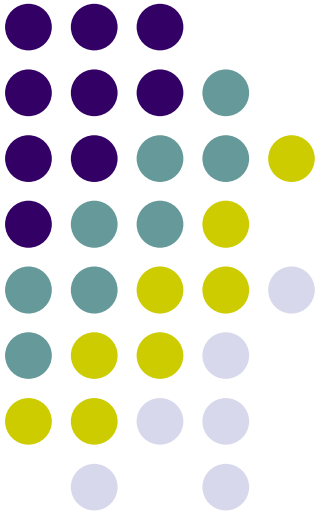
堀田 敬介

分析法の分類

目的	外的基準(目的変数)		説明要因(説明変数)	分析法
予測	あり	量的	量的	重回帰分析
			質的(カテゴリー)	数量化Ⅰ類
分類，判別		質的(カテゴリー)	量的	**判別分析**
			質的(カテゴリー)	**数量化Ⅱ類**
要因抽出，パターン分類	なし		量的	主成分分析
				因子分析
			質的(カテゴリー)	**数量化Ⅲ類**
				数量化Ⅳ類

判別分析

目的　　：　判別
外的基準　：　質的(カテゴリー)
説明要因　：　量的



判別分析とは？

説明要因χ：量的
点数（英語）
点数（国語）

学生	1	2	3	4	5	6	7								15	16	17	18	19	20
点数(英語)	45	40	60	50	35	62	43	60		53	50	65	60	35	45	44	50	55	49	53
点数(国語)	55	47	45	58	47	56	40	78	40	47	49	62	49	70	60	58	55	61	54	57
A大学合否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否	否

外的基準 y：質的
合格か不合格か

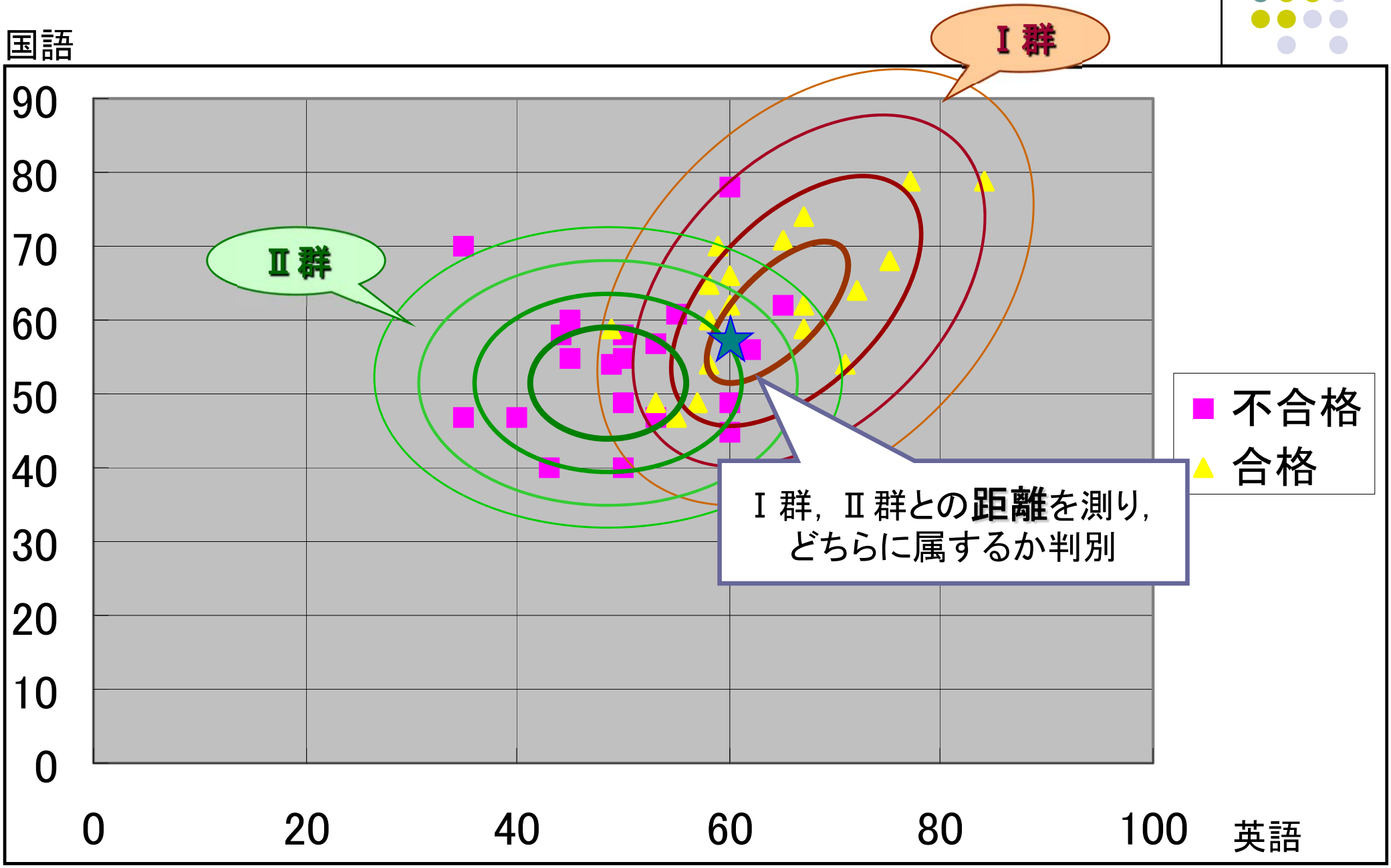
学生	21	22	23	24	25	26	27	28	29	30							37	38	39	40
点数(英語)	65	58	75	67	58	77	55	84	58	72							60	59	53	60
点数(国語)	71	60	68	62	54	79	47	79	65	64							66	70	49	62
A大学合否	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合	合

α君の点数	
点数（英語）	60
点数（国語）	57

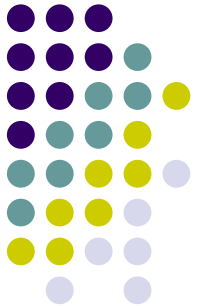
僕は**合格**するかなぁ？

判別（分類）したい!

判別分析とは？


国語
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
英語
Ⅰ群
Ⅱ群
不合格
合格
Ⅰ群，Ⅱ群との距離を測り，
どちらに属するか判別

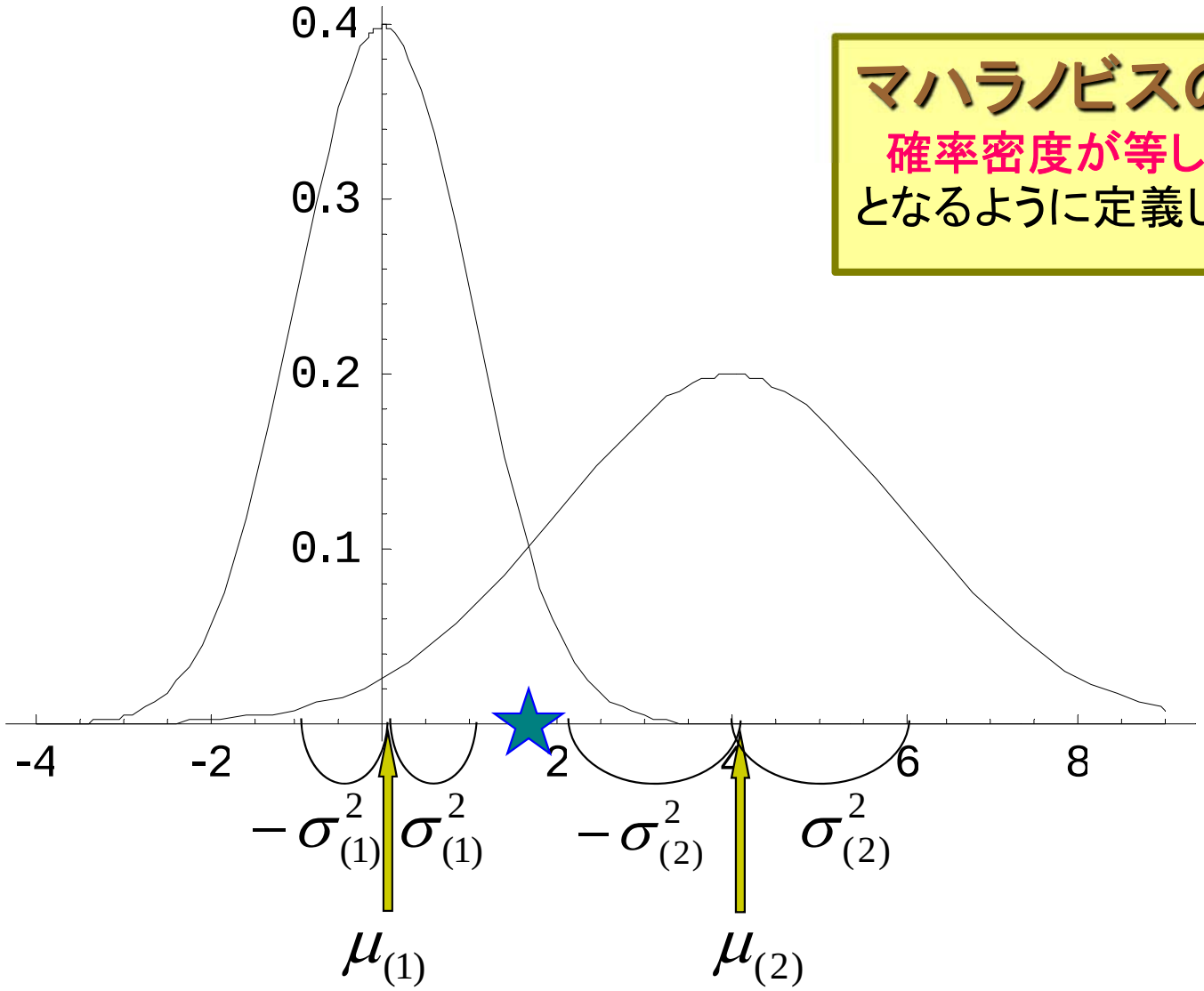
判別分析とは？



【判別分析の適用例】

- 集団検診：年齢，血圧，眼底所見→ある病気にかかっているか？
- 年代測定：出土化石の分析測定→ある年代よりも前か後か？
- 政党支持：ある政党の支持層と非支持層の比較分析
- 品質管理：生産プロセスなどから，ある製品の良・不良判定
- 客層分析：購買データなどから，固定客・非固定客の判別

１変量の判別〔マハラノビスの汎距離による〕

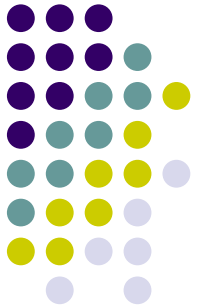


マハラノビスの汎距離

確率密度が等しい点は等距離

となるように定義した距離

１変量の判別〔マハラノビスの汎距離による〕



Ⅰ群の

母平均 $\mu_{(1)} \approx \bar{x}^{(1)} = \frac{1}{n_{(1)}} \sum_{i=1}^{n_{(1)}} x_i^{(1)}$ ，母分散 $\sigma_{(1)}^2 \approx \bar{s}_{(1)}^2 = \frac{1}{n_{(1)}-1} \sum_{i=1}^{n_{(1)}} (x_i^{(1)} - \bar{x}^{(1)})$

Ⅰ群からの距離 $D_{(1)}^2 = \frac{(x-\mu_{(1)})^2}{\sigma_{(1)}^2}$

母平均・母分散は既知とは限らないが，標本数は十分大きいと仮定し，推定値を代入．

Ⅱ群の

母平均 $\mu_{(2)} \approx \bar{x}^{(2)} = \frac{1}{n_{(2)}} \sum_{i=1}^{n_{(2)}} x_i^{(2)}$，母分散 $\sigma_{(2)}^2 \approx \bar{s}_{(2)}^2 = \frac{1}{n_{(2)}-1} \sum_{i=1}^{n_{(2)}} (x_i^{(2)} - \bar{x}^{(2)})$

Ⅱ群からの距離 $D_{(2)}^2 = \frac{(x-\mu_{(2)})^2}{\sigma_{(2)}^2}$

判定

$D_{(1)}^2 < D_{(2)}^2$ ⇒ xはⅠ群に属す

$D_{(1)}^2 > D_{(2)}^2$ ⇒ xはⅡ群に属す

１変量の判別〔マハラノビスの汎距離による〕

母分散が等しい（$\sigma_{(1)}^2 = \sigma_{(2)}^2$）場合

$$D_{(2)}^2 - D_{(1)}^2 = \frac{-2(\mu_{(2)} - \mu_{(1)})}{\sigma_{(1)}^2}\left\{x - \frac{\mu_{(2)} + \mu_{(1)}}{2}\right\}$$

$\mu_{(2)} > \mu_{(1)}$ なら

$D_{(2)}^2 > D_{(1)}^2 \Leftrightarrow x < \dfrac{\mu_{(2)} + \mu_{(1)}}{2} = \bar{\mu}$ ⇒ **xはⅠ群に属す**

$D_{(2)}^2 < D_{(1)}^2 \Leftrightarrow x > \dfrac{\mu_{(2)} + \mu_{(1)}}{2} = \bar{\mu}$ ⇒ **xはⅡ群に属す**

母分散が等しくない（$\sigma_{(1)}^2 \neq \sigma_{(2)}^2$）場合

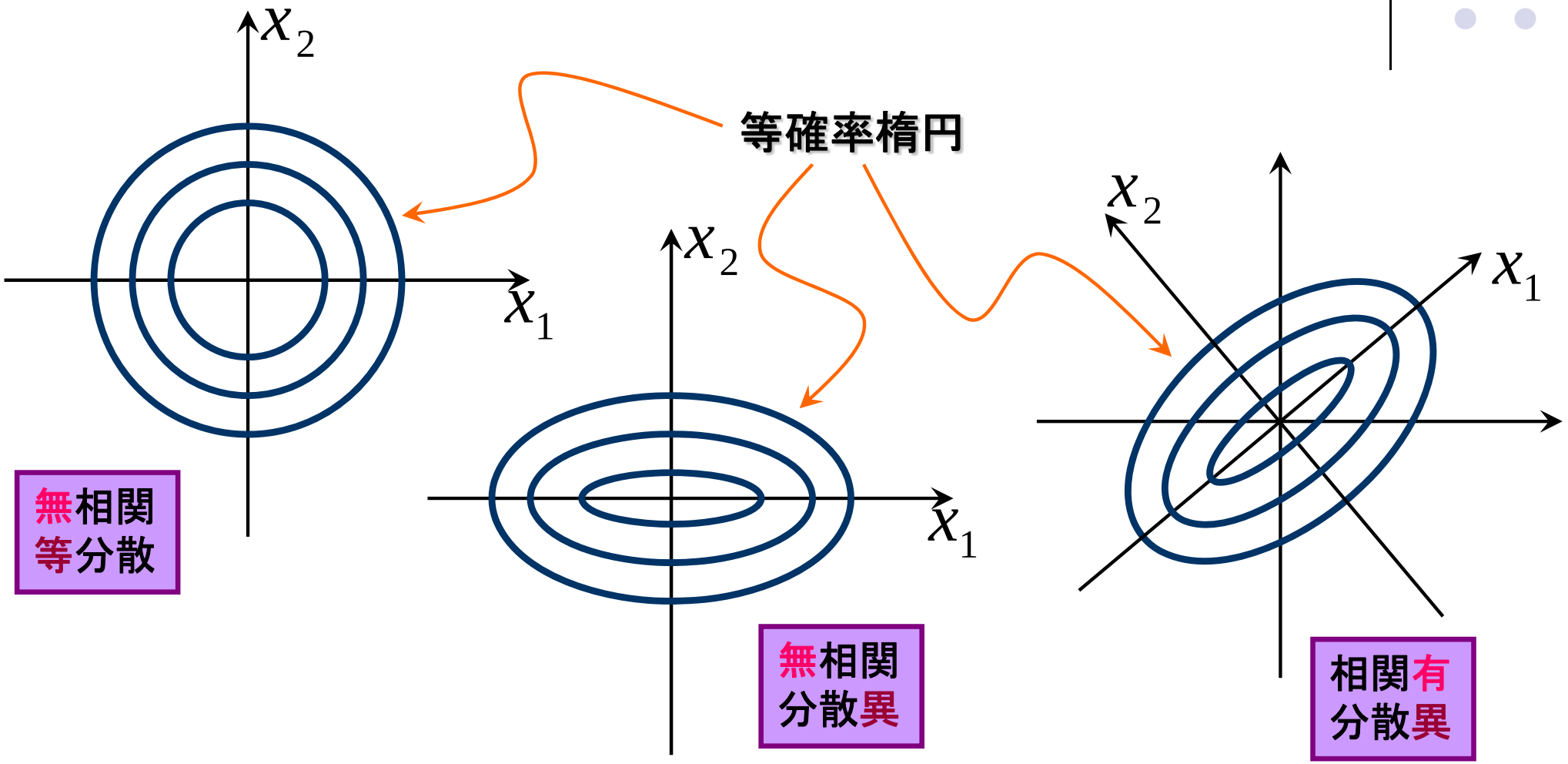
$D_{(2)}^2 = D_{(1)}^2$ となる境界値を c とすると（ただし，$\mu_{(2)} > \mu_{(1)}$ とする）

$$c = \frac{\mu_{(1)}\sigma_{(2)} + \mu_{(2)}\sigma_{(1)}}{\sigma_{(1)} + \sigma_{(2)}}$$

$x < c$ ⇒ **xはⅠ群に属す**

$x > c$ ⇒ **xはⅡ群に属す**

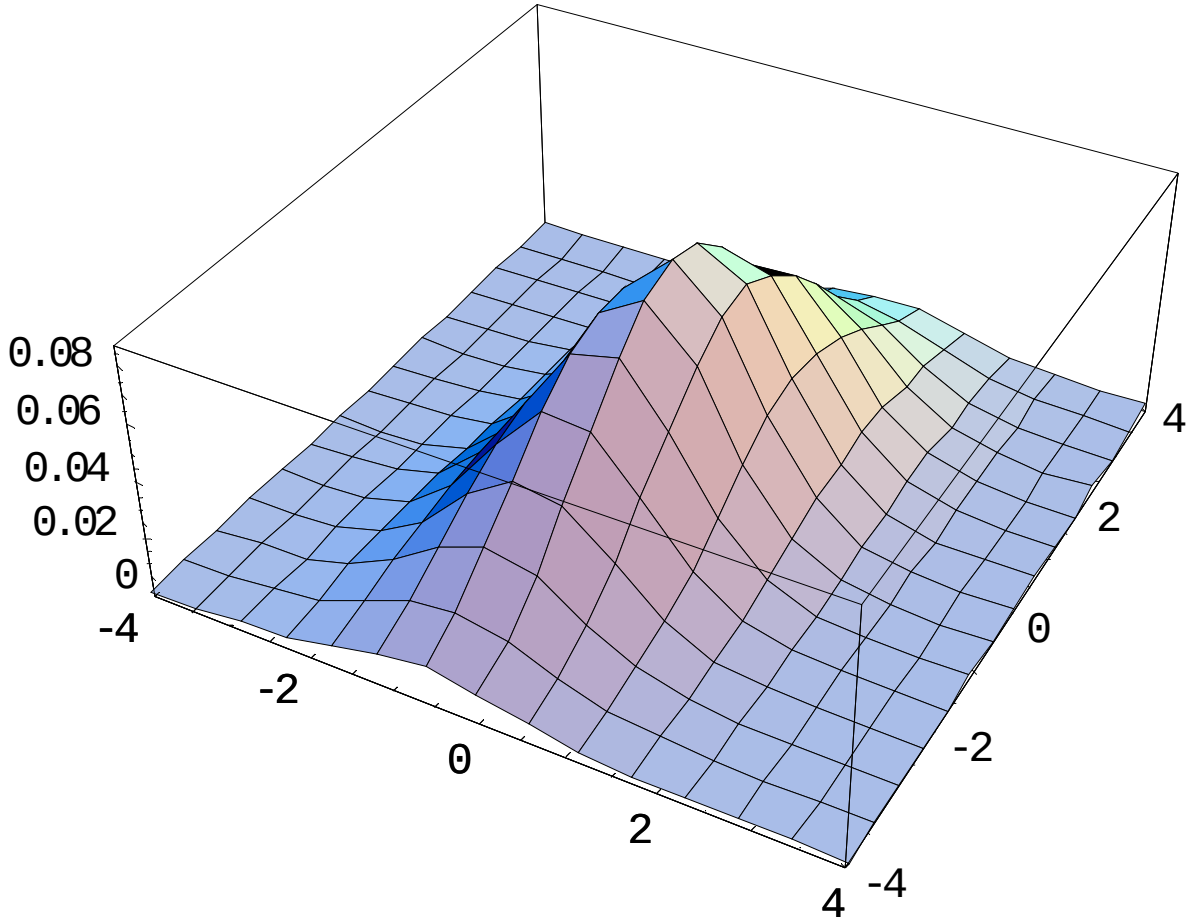
2変量以上の判別〔マハラノビスの汎距離〕



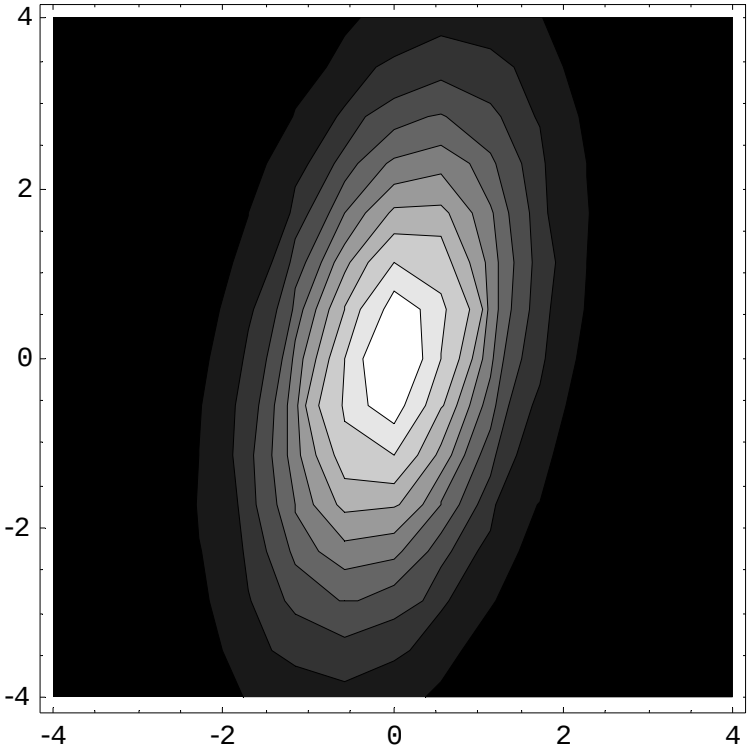
マハラノビスの汎距離

確率密度が等しい点は等距離となるように定義した距離

2変量以上の判別〔マハラノビスの汎距離〕

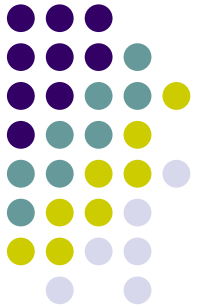


分散が異なり，相関のある2変量についての2次元正規分布



その等高線プロット

2変量以上の判別〔マハラノビスの汎距離〕



Ⅰ群・Ⅱ群各々の変数 j の母平均を

$$\mu_{(k)}^{j} \approx \bar{x}_{(k)}^{j} = \frac{1}{n_{(k)}} \sum_{i=1}^{n_{(k)}} (x_{(k)}^{j})_i \ (k = 1,2)$$

とし $\mu_{(k)} = [\mu_{(k)}^{1}, \cdots, \mu_{(k)}^{n(k)}]^T \ (k = 1,2)$ と表す

母平均・母分散は既知とは限らないが，標本数は十分大きいと仮定し，推定値を代入．

Ⅰ群・Ⅱ群の変数i, j の共分散を

$\sigma_{ij(k)} \approx \bar{s}_{ij(k)} \ (k = 1,2)$ とし，分散共分散行列を

$$\sum_k = \begin{bmatrix} \sigma_{11(k)} & \sigma_{12(k)} & \cdots & \sigma_{1n_{(k)}(k)} \\ \sigma_{21(k)} & \sigma_{22(k)} & \cdots & \sigma_{1n_{(k)}(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_{(k)}1(k)} & \sigma_{n_{(k)}2(k)} & \cdots & \sigma_{n_{(k)}n_{(k)}(k)} \end{bmatrix} (k = 1,2)$$ と表す

注: $\sigma_{ii(k)} = \sigma_{i(k)}^2$ は，変数 i の分散

2変量以上の判別〔マハラノビスの汎距離〕

各群と対象点χの距離（マハラノビスの汎距離）

$$D^2_{(k)} = (x-\mu_{(k)})^T \sum_k^{-1} (x-\mu_{(k)}) \ (k=1,2)$$

判定

$D^2_{(1)} < D^2_{(2)}$ ⇒ χはⅠ群に属す

$D^2_{(1)} > D^2_{(2)}$ ⇒ χはⅡ群に属す

２変量以上の判別〔線形判別関数〕

分散共分散行列が等しい（ $\Sigma_1^{-1} = \Sigma_2^{-1} = \Sigma^{-1}$ ）場合

$$D_{(2)}^2 - D_{(1)}^2$$

$$= (x-\mu_{(2)})^T \Sigma^{-1}(x-\mu_{(2)}) - (x-\mu_{(2)})^T \Sigma^{-1}(x-\mu_{(2)})$$

$$= 2(x-\mu)^T \Sigma^{-1}(\mu_{(1)} - \mu_{(2)})$$

ここで $a = \Sigma^{-1}(\mu_1 - \mu_2)$ とおくと，$z = (x-\mu)^T a$ となる

判定

$z > 0$ ⇒ **χはⅠ群に属す**

$z < 0$ ⇒ **χはⅡ群に属す**

線形判別関数

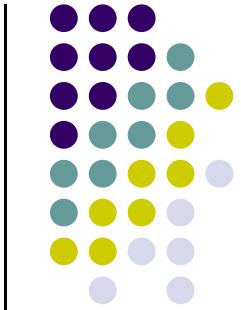
注：

$$z = (x_1 - \mu_1)a_1 + \ldots + (x_n - \mu_n)a_n$$

$$= a_0 + a_1 x_1 + \cdots + a_n x_n$$

ただし，$(a_0 = -\mu^T a)$

例題：α君は合格するか？



α君の点数＝ $\begin{pmatrix} 60 \\ 57 \end{pmatrix}$

Ⅰ群〔不合格者〕

$$\mu_1 = \begin{pmatrix} 50.2 \\ 54.4 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 73.54 & 12.60 \\ 12.60 & 89.20 \end{pmatrix} \quad \Sigma_1^{-1} = \begin{pmatrix} 0.0139 & -0.0020 \\ -0.0020 & 0.0115 \end{pmatrix}$$

Ⅱ群〔合格者〕

$$\mu_2 = \begin{pmatrix} 63.5 \\ 62 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 79.73 & 58.63 \\ 58.63 & 95.47 \end{pmatrix} \quad \Sigma_2^{-1} = \begin{pmatrix} 0.0229 & -0.0140 \\ -0.0140 & 0.0191 \end{pmatrix}$$

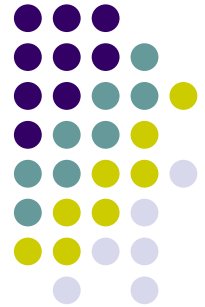
$$D_{(1)}^2 = \begin{pmatrix} 60 \\ 57 \end{pmatrix} - \begin{pmatrix} 50.2 \\ 54.4 \end{pmatrix}^T \begin{pmatrix} 0.0139 & -0.0020 \\ -0.0020 & 0.0115 \end{pmatrix} \begin{pmatrix} 60 \\ 57 \end{pmatrix} - \begin{pmatrix} 50.2 \\ 54.4 \end{pmatrix} = 1.316$$

$$\vee$$

$$D_{(2)}^2 = \begin{pmatrix} 60 \\ 57 \end{pmatrix} - \begin{pmatrix} 63.5 \\ 62 \end{pmatrix}^T \begin{pmatrix} 0.0229 & -0.0140 \\ -0.0140 & 0.0191 \end{pmatrix} \begin{pmatrix} 60 \\ 57 \end{pmatrix} - \begin{pmatrix} 63.5 \\ 62 \end{pmatrix} = 0.265$$

α君は**Ⅱ群に属す**，即ち，**合格する**であろう

補足：等分散性の検定〔1変量の場合〕



- Ⅰ群・Ⅱ群の確率分布が各々 $N(\mu_{(1)},\sigma^2_{(1)}), N(\mu_{(2)},\sigma^2_{(2)})$
  - 帰無仮説：$\sigma^2_{(1)}=\sigma^2_{(2)}$ ，対立仮説：$\sigma^2_{(1)}\neq\sigma^2_{(2)}$
  - 統計量 $F=\dfrac{\bar{s}^2_{(1)}}{\bar{s}^2_{(2)}}$ が自由度$(n_{(1)}-1, n_{(2)}-1)$のF分布に従う
  - 有意水準αについて

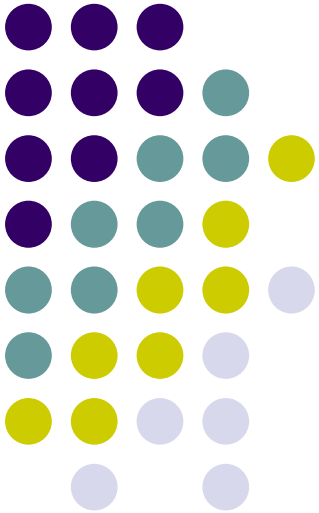
$$F > F^{n_1-1}_{n_2-1}\left(\frac{\alpha}{2}\right) \text{ or } F < F^{n_1-1}_{n_2-1}\left(1-\frac{\alpha}{2}\right)$$

なら，帰無仮説を棄却し，両群の分散は等しくないと言える

  - p変量の場合は別の統計量が自由度 $p(p+1)/2$ の χ^2分布に従うことを用いる

数量化Ⅱ類

目的　　：　判別
外的基準　：　質的(カテゴリー)
説明要因　：　質的(カテゴリー)



数量化Ⅱ類とは？

外的基準 y：質的
セ・リーグかパ・リーグか

説明要因χ：質的
本拠地，親会社業績，
62年度成績

球団名	リーグ	本拠地	親会社業種	62年度成績
巨人	セ・リーグ	東京	新聞	A
中日	セ・リーグ	名古屋	新聞	A
広島	セ・リーグ	広島	市	A
ヤクルト	セ・リーグ	東京	メーカー	B
大洋	セ・リーグ	横浜	市	B
阪神	セ・リーグ	大阪	電鉄	C
西部	パ・リーグ	東京	電鉄	A
阪急	パ・リーグ	大阪	電鉄	A
日ハム	パ・リーグ	東京	メーカー	B
南海	パ・リーグ	大阪	電鉄	B
ロッテ	パ・リーグ	川崎	メーカー	C
近鉄	パ・リーグ	大阪	電鉄	C

セ・リーグとパ・リーグを区分する特徴を知りたい！

数量化Ⅱ類とは？

説明要因x：質的
3つの説明要因に
ダミー変数を導入

サンプル		外的基準	説明要因（アイテム）									
球団		リーグ	本拠地			親会社業種				62年度成績		
No.	球団名		1	2	3	1	2	3	4	1	2	3
1	巨人	1	1	0	0	1	0	0	0	1	0	0
2	中日	1	0	0	1	1	0	0	0	1	0	0
3	広島	1	0	0	1	0	0	1	0	1	0	0
4	ヤクルト	1	1	0	0	0	0	0	1	0	1	0
5	大洋	1	1	0	0	0	0	1	0	0	1	0
6	阪神	1	0	1	0	0	1	0	0	0	0	1
7	西部	2	1	0	0	0	1	0	0	1	0	0
8	阪急	2	0	1	0	0	1	0	0	1	0	0
9	日ハム	2	1	0	0	0	0	0	1	0	1	0
10	南海	2	0	1	0	0	1	0	0	0	1	0
11	ロッテ	2	1	0	0	0	0	0	1	0	0	1
12	近鉄	2	0	1	0	0	1	0	0	0	0	1

ダミー変数
を導入し，
数量化Ⅱ類
で分析!

ダミー変数と予測式

- ダミー変数

セリーグ　巨人　本拠地　東京

$$\delta_{il}(jk)=\begin{cases}1 : i\text{ 群のサンプル }l\text{ が，アイテム }j\text{ のカテゴリー }k\text{ に反応}\\ 0 : o.\,w.\end{cases}$$

横浜，川崎，名古屋，大阪，広島

- 予測式

ダミー変数 $\delta_{\cdot\cdot}(jk)$ の係数（アイテム・カテゴリーの評点）

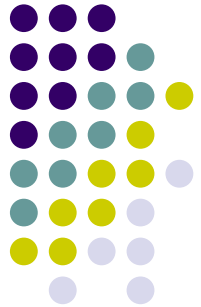
$$Y_{il}=\sum_{j=1}^{R}\sum_{k=1}^{c_j}a_{jk}\delta_{il}(jk)$$ ：i 群のサンプル l の評点

全変動に対し，K個の群の**群間変動を最大に**するように a_{jk} を定める

（**相関比** = 群間分散 / 全分散 **を最大にする**ようa_{jk}を定める）

サンプル			説明要因（アイテム）								
		j	1			…	R				
		k	1	…	c_1	…	1	…	c_R		
l	i										
1	1		1	…	0	…	0	…	1		
:			:		:						
n_1			0	…	0		0	…	1		
:	:		:				:				
1	K		0	…	1	…	0	…	0		
:											
n_K			0	…	1		1	…	0		

全変動･群間変動，相関比



- 全変動･群間変動

$$\sum_{i=1}^{K}\sum_{l=1}^{n_i}(Y_{il}-\bar{Y}_{\cdot\cdot})^2=\sum_{i=1}^{K}n_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2+\sum_{i=1}^{K}\sum_{l=1}^{n_i}(Y_{il}-\bar{Y}_{i\cdot})^2$$

全変動 S_T　　群間変動 S_B　　群内変動 S_W

$$\frac{1}{n}\sum_{i=1}^{K}\sum_{l=1}^{n_i}(Y_{il}-\bar{Y}_{\cdot\cdot})^2=\frac{1}{n}\sum_{i=1}^{K}n_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2+\frac{1}{n}\sum_{i=1}^{K}\sum_{l=1}^{n_i}(Y_{il}-\bar{Y}_{i\cdot})^2$$

全分散 σ_T^2　　群間分散 σ_B^2　　群内分散 σ_W^2

- 相関比

$$\eta^2=\frac{\sigma_B^2}{\sigma_T^2}=\frac{\bar{s}_B^2}{\bar{s}_T^2}$$

相関比 = 群間分散 / 全分散
を**最大にする**よう a_{jk} を定める

全変動･群間変動，相関比

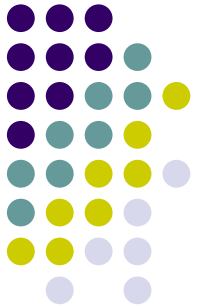
- 全変動･群間変動

$$S_T = \sum_{i=1}^{K}\sum_{l=1}^{n_i}(Y_{il} - \bar{Y}_{..})^2 = \sum_{j=1}^{R}\sum_{k=1}^{c_j}\sum_{u=1}^{R}\sum_{v=1}^{c_u} t(jk,uv)a_{jk}a_{uv}$$

$$S_B = \sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \sum_{j=1}^{R}\sum_{k=1}^{c_j}\sum_{u=1}^{R}\sum_{v=1}^{c_u} b(jk,uv)a_{jk}a_{uv}$$

$$\begin{cases} t(jk,uv) = f(jk,uv) - \dfrac{n_{jk}n_{uv}}{n} \\ b(jk,uv) = \displaystyle\sum_{i=1}^{K} \frac{g^i(jk)g^i(uv)}{n_i} - \frac{n_{jk}n_{uv}}{n} \end{cases}$$

全変動・群間変動，相関比



- 全変動・群間変動

$f(jk,uv)$ ：アイテム j カテゴリー k とアイテム u カテゴリー v の両方に反応したサンプルの数

$g^{i}(jk)$ ：第 i 群でアイテム j カテゴリー k に反応したサンプル数

n_{jk} ：アイテム j カテゴリー k に反応したサンプル数

n ：全反応数 【サンプル数ではないことに注意】

- 相関比

$$\eta^2 = \frac{\sum_{j=1}^{R}\sum_{k=1}^{c_j}\sum_{u=1}^{R}\sum_{v=1}^{c_u} b(jk,uv)a_{jk}a_{uv}}{\sum_{j=1}^{R}\sum_{k=1}^{c_j}\sum_{u=1}^{R}\sum_{v=1}^{c_u} t(jk,uv)a_{jk}a_{uv}}$$

この値を**最大にする**ように a_{jk} を決定する！

全変動・群間変動，相関比

もとの表

サンプル	外的基準		説明要因（アイテム）								
		j	1			…	R		
		k	1	…	c_1	…	1	…	c_R
l	i								
1	1		1	…	0	…	0	…	1
:			:		:				
n_1			0	…	0		0	…	1
:	:		:				:		
1	K		0	…	1	…	0	…	0
:									
n_K			0	…	1		1	…	0

群毎の集計表

外的基準	説明要因（アイテム）							
	1			…	R			
i	1	…	c_1	…	1	…	c_R	
1	8	…	7	…	9	…	1	36
:	:				:			:
K	5	…	2	…	11	…	3	19
合計	21	…	17	…	31	…		96

n_{jk}　$g^i(jk)$　n_i　n

アイテム，カテゴリーに関するクロス集計表

			説明要因（アイテム）								
			1			…	R		
			1	…	c_1	…	1	…	c_R
説明要因（アイテム）	1	1	5	…	1	…	3	…	4
		…	:		:				
		c_1	1	…	2		4	…	2
	:	:	:				:		
	R	1	7	…	2	…	6	…	1
		…							
		c_R	2	…	3		1	…	5

$f(jk,uv)$

全変動・群間変動，相関比

- 予測式の係数を求めたいが…

各サンプルは各アイテムに対し，必ず**ただ一つだけに反応**する（一つだけが1となる）とすると，各アイテム内のカテゴリー変数に**一次従属性**が成り立つ

⇒

各アイテム内のカテゴリー変数のうち一つは，他が決まると決まり**一意に定まらない！**

⇒

標準化を行う

$$a_{j1} = 0 \ (j = 1, \ldots, R)$$

or

$$\sum_{k=1}^{c_j} n_{jk} a_{jk} = 0 \ (j = 1, \ldots, R)$$

$a_{j1} = 0 \ (j = 1, \ldots, R)$ ⇒

相関比

$$\eta^2 = \frac{\sum_{j=1}^{R} \sum_{k=2}^{c_j} \sum_{u=1}^{R} \sum_{v=2}^{c_u} b(jk, uv) a_{jk} a_{uv}}{\sum_{j=1}^{R} \sum_{k=2}^{c_j} \sum_{u=1}^{R} \sum_{v=2}^{c_u} t(jk, uv) a_{jk} a_{uv}}$$

全変動・群間変動，相関比

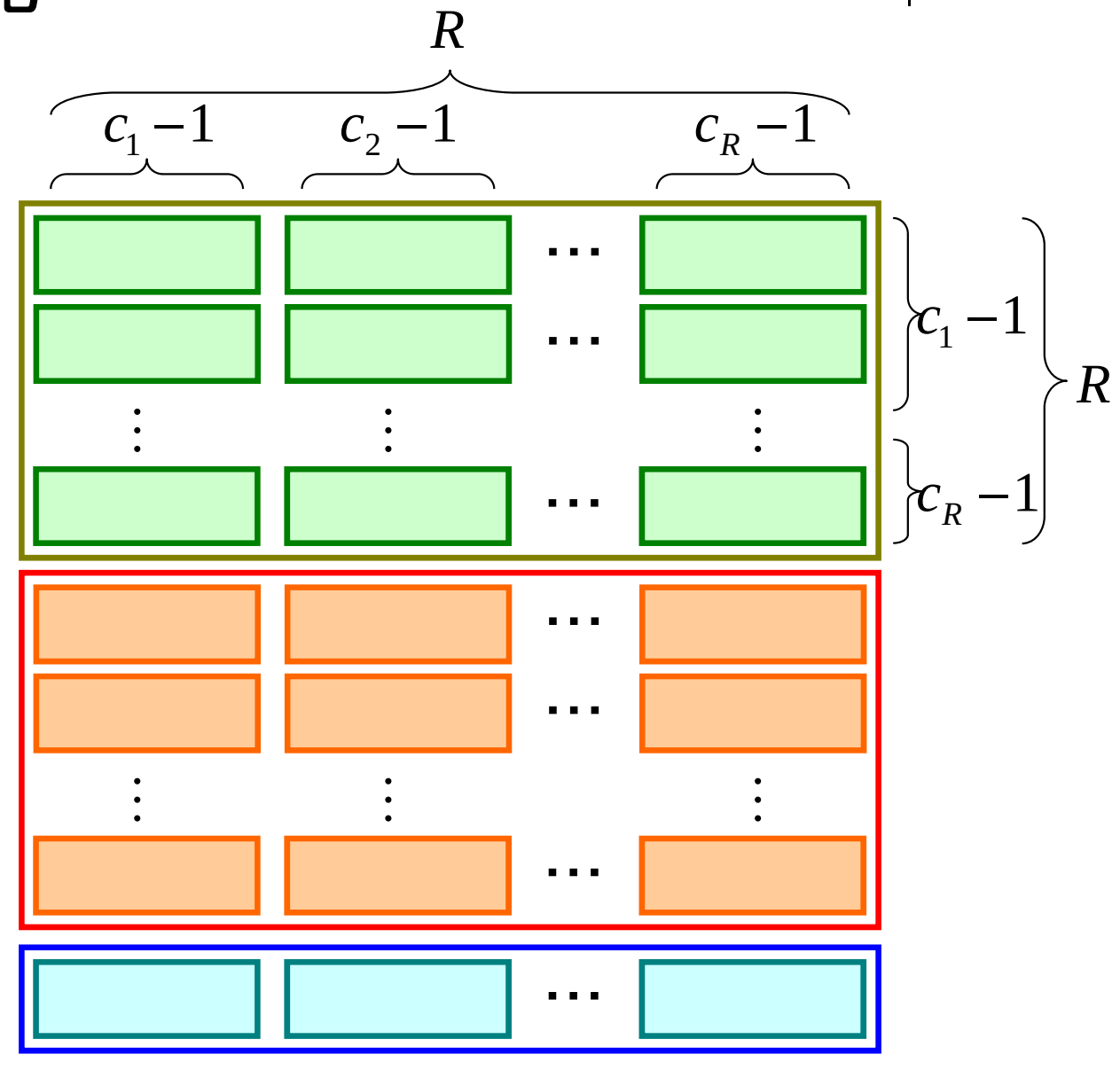
- 相関比の行列表記

$$\eta^2 = \frac{a^T B a}{a^T T a}$$

$$B = [b(jk,uv)] =$$

$$T = [t(jk,uv)] =$$

$$a^T = [a_{jk}]^T =$$



全変動・群間変動，相関比

- 予測式の係数を決定しよう！

$$\sum_{j=1}^{R}\sum_{k=2}^{c_j}\{b(jk,uv)-\eta^2 t(jk,uv)\}a_{jk}=0$$

$$(u=1,\ldots,R,\ \mathrm{v}=1,\ldots,c_u)$$

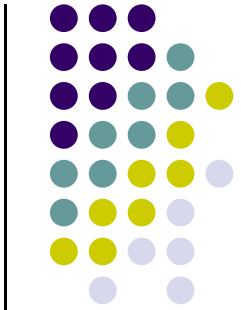
$a_{..}$で偏微分し=0

$$\Leftrightarrow (B-\eta^2 T)a=0 \Leftrightarrow \left|B-\eta^2 T\right|=0$$

一般化固有値問題

最大固有値 η^2 に対する固有ベクトル a を求めると，その**固有ベクトル**が**予測式の係数**である

補足：一般化固有値問題



- 一般化固有値問題

$$(B-\eta^2 T)a=0$$ ただし，B：対称行列，T：対称正定値行列

T：対称正定値行列よりCholesky分解し，

$$T=LL^T$$

これより，$A=L^{-1}BL^{-T}$ とおくと，

$$(B-\eta^2 T)a=0$$
$$\Leftrightarrow (LAL^T-\eta^2 LL^T)a=0$$
$$\Leftrightarrow Ab=\eta^2 b \quad (a=L^{-T}b)$$

となる．よって，行列Aの（最大）固有値・固有ベクトルを求めればよい．
ただし，固有ベクトルは大きさ1に基準化する．

例題:リーグの特徴抽出

サンプル		外的基準	説明要因(アイテム)									
球団		リーグ	本拠地			親会社業種				62年度成績		
No.	球団名		1	2	3	1	2	3	4	1	2	3
1	巨人	1	1	0	0	1	0	0	0	1	0	0
2	中日	1	0	0	1	1	0	0	0	1	0	0
3	広島	1	0	0	1	0	0	1	0	1	0	0
4	ヤクルト	1	1	0	0	0	0	0	1	0	1	0
5	大洋	1	1	0	0	0	0	1	0	0	1	0
6	阪神	1	0	1	0	0	1	0	0	0	0	1
7	西部	2	1	0	0	0	1	0	0	1	0	0
8	阪急	2	0	1	0	0	1	0	0	1	0	0
9	日ハム	2	1	0	0	0	0	0	1	0	1	0
10	南海	2	0	1	0	0	1	0	0	0	1	0
11	ロッテ	2	1	0	0	0	0	0	1	0	0	1
12	近鉄	2	0	1	0	0	1	0	0	0	0	1

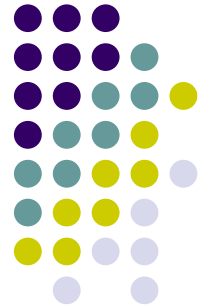
	説明要因(アイテム)										
リーグ	本拠地			親会社業種				62年度成績			
	1	2	3	1	2	3	4	1	2	3	
セ・リーグ	3	1	2	2	1	2	1	3	2	1	18
パ・リーグ	3	3	0	0	4	0	2	2	2	2	18
合計	6	4	2	2	5	2	3	5	4	3	36

n_{jk} $g^{i}(jk)$ n n_{i}

	本拠地			親会社業種				62年度成績		
	東京近郊	大阪	その他	新聞社	電鉄	市	メーカー	A	B	C
東京近郊	6	0	0	1	1	1	3	2	3	1
大阪	0	4	0	0	4	0	0	1	1	2
その他	0	0	2	1	0	1	0	2	0	0
新聞社	1	0	1	2	0	0	0	2	0	0
電鉄	1	4	0	0	5	0	0	2	1	1
市	1	0	1	0	0	2	0	1	1	0
メーカー	3	0	0	0	0	0	3	0	2	1
A	2	1	2	2	2	1	0	5	0	0
B	3	1	0	0	1	1	2	0	4	0
C	1	2	0	0	2	0	1	0	0	3

$f(jk,uv)$

例題:リーグの特徴抽出



- 相関比 $\eta^2 = \dfrac{a^T B a}{a^T T a}$
- 一般化固有値問題 $(B - \eta^2 T)a = 0$

$B = [b(jk,uv)] =$

1	2	3	1	2	3	4	1	2	3
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.1	−0.1	−0.1	0.2	−0.1	0.1	−0.1	0.0	0.1
0.0	−0.1	0.1	0.1	−0.2	0.1	−0.1	0.1	0.0	−0.1
0.0	−0.1	0.1	0.1	−0.2	0.1	−0.1	0.1	0.0	−0.1
0.0	0.2	−0.2	−0.2	0.3	−0.2	0.1	−0.1	0.0	0.1
0.0	−0.1	0.1	0.1	−0.2	0.1	−0.1	0.1	0.0	−0.1
0.0	0.1	−0.1	−0.1	0.1	−0.1	0.0	−0.0	0.0	0.0
0.0	−0.1	0.1	0.1	−0.1	0.1	−0.0	0.0	0.0	−0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.1	−0.1	−0.1	0.1	−0.1	0.0	−0.0	0.0	0.0

$T = [t(jk,uv)] =$

1	2	3	1	2	3	4	1	2	3
5.0	−0.7	−0.3	0.7	0.2	0.7	2.5	1.2	2.3	0.5
−0.7	3.6	−0.2	−0.2	3.4	−0.2	−0.3	0.4	0.6	1.7
−0.3	−0.2	1.9	0.9	−0.3	0.9	−0.2	1.7	−0.2	−0.2
0.7	−0.2	0.9	1.9	−0.3	−0.1	−0.2	1.7	−0.2	−0.2
0.2	3.4	−0.3	−0.3	4.3	−0.3	−0.4	1.3	0.4	1.6
0.7	−0.2	0.9	−0.1	−0.3	1.9	−0.2	0.7	0.8	−0.2
2.5	−0.3	−0.2	−0.2	−0.4	−0.2	2.8	−0.4	1.7	0.8
1.2	0.4	1.7	1.7	1.3	0.7	−0.4	4.3	−0.6	−0.4
2.3	0.6	−0.2	−0.2	0.4	0.8	1.7	−0.6	3.6	−0.3
0.5	1.7	−0.2	−0.2	1.6	−0.2	0.8	−0.4	−0.3	2.8

例題:リーグの特徴抽出

- 一般化固有値問題を解いて…

$$\eta^2 = 0.1809, -0.104096, \cdots$$

最大固有値

最大固有値に対する固有ベクトル

$$a = \begin{bmatrix} 0 \\ 0.026 \\ -0.356 \\ \hdashline 0 \\ 0.439 \\ -0.037 \\ 0.485 \\ \hdashline 0 \\ -0.318 \\ -0.298 \end{bmatrix}$$

カテゴリー内の和が0になるように基準化

カテゴリースコア

$$a = \begin{bmatrix} 0.110 \\ 0.136 \\ -0.246 \\ \hdashline -0.222 \\ 0.217 \\ -0.259 \\ 0.263 \\ \hdashline 0.205 \\ -0.113 \\ -0.092 \end{bmatrix}$$

本拠地
- 東京近郊
- 大阪
- その他

親会社業種
- 新聞
- 電鉄
- 市
- メーカー

62年度成績
- A
- B
- C

例題:リーグの特徴抽出

- 結果

		カテゴリースコア	範囲 Range	偏相関係数
本拠地	東京近郊	0.110	0.3813	0.1243
	大阪	0.136		
	その他	−0.246		
親会社業種	新聞社	−0.222	0.5224	0.6253
	電鉄	0.217		
	市	−0.259		
	メーカー	0.263		
62年度成績	A	0.205	0.3183	0.2936
	B	−0.113		
	C	−0.092		

このアイテムが「セ・リーグ」「パ・リーグ」の違いに一番寄与している

重相関係数 $R=0.727$, $R^2=0.528$

例題:リーグの特徴抽出

- 結果

	予測値Y	外的基準	本拠地	親会社業種	62年度成績
巨人	0.093794	1	0.110038	−0.22157	0.205327
中日	−0.26191	1	−0.24567	−0.22157	0.205327
広島	−0.29935	1	−0.24567	−0.25901	0.205327
ヤクルト	0.260463	1	0.110038	0.263366	−0.11294
大洋	−0.26191	1	0.110038	−0.25901	−0.11294
阪神	0.260463	1	0.135632	0.217215	−0.09238
西部	0.53258	2	0.110038	0.217215	0.205327
阪急	0.558174	2	0.135632	0.217215	0.205327
日ハム	0.260463	2	0.110038	0.263366	−0.11294
南海	0.239906	2	0.135632	0.217215	−0.11294
ロッテ	0.281019	2	0.110038	0.263366	−0.09238
近鉄	0.260463	2	0.135632	0.217215	−0.09238

外的基準と要因アイテム間の相関行列

	外的基準	本拠地	親会社業種	62年度成績
外的基準	1.0000	0.4643	0.6954	−0.1624
本拠地	0.4643	1.0000	0.6490	−0.5314
親会社業種	0.6954	0.6490	1.0000	−0.4778
62年度成績	−0.1624	−0.5314	−0.4778	1.0000

例題:リーグの特徴抽出

- 結果

セ・リーグと判別

パ・リーグと判別

広島	−0.29935
中日	−0.26191
大洋	−0.26191
巨人	0.093794
南海	0.239906
阪神	0.260463
近鉄	0.260463
ヤクルト	0.260463
日ハム	0.260463
ロッテ	0.281019
西部	0.53258
阪急	0.558174

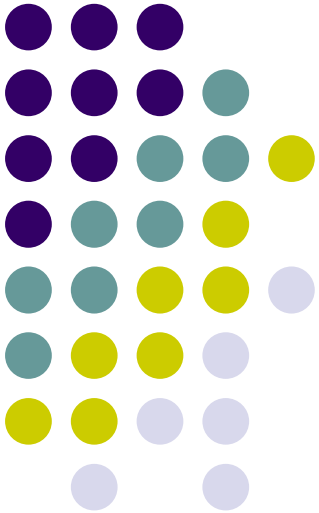
誤

誤

誤

数量化Ⅲ類

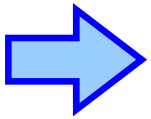
目的　　：　パターン分類
外的基準　：　なし
説明要因　：　質的(カテゴリー)



数量化Ⅲ類とは？

外的基準 なし
説明要因x：質的

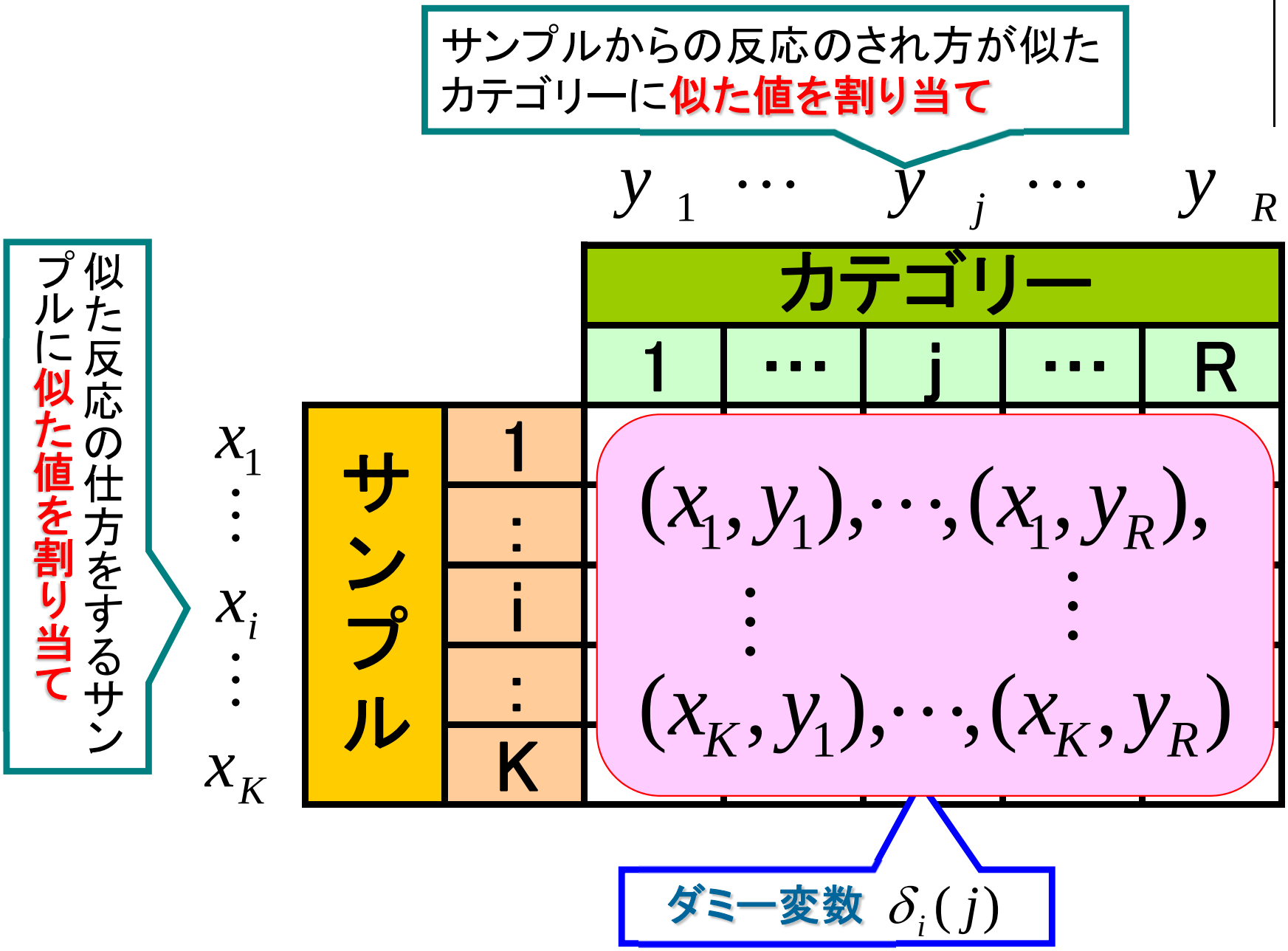
	カテゴリー		
サンプル	T	S	W
1	○	○	
2		○	○
3	○		
4		○	○
5		○	
6	○	○	○
7			○
8		○	○
9	○		
10	○		○



	カテゴリー		
サンプル	T	S	W
3	○		
9	○		
1	○	○	
10	○		○
6	○	○	○
5		○	
2		○	○
4		○	○
8		○	○
7			○

データの**特徴を知りたい**ので，**似たものどうしが近くなる**ように並べ直す！

数量化Ⅲ類とは？


サンプルからの反応のされ方が似たカテゴリーに似た値を割り当て
似た反応の仕方をするサンプルに似た値を割り当て
$y_1 \cdots y_j \cdots y_R$
カテゴリー
1 … j … R
サンプル
1
i
K
x_1
x_i
x_K
$(x_1, y_1), \cdots, (x_1, y_R),$
$(x_K, y_1), \cdots, (x_K, y_R)$
ダミー変数 $\delta_i(j)$

数量化Ⅲ類とは？

$$\begin{matrix}(x_1, y_1), \cdots, (x_1, y_R), \\ \vdots \qquad\qquad \vdots \\ (x_K, y_1), \cdots, (x_K, y_R)\end{matrix}$$

$K \times R$組の変数(x_i, y_j)を，その**相関係数 r が最大**になるように決定する．

相関係数

$$r = \frac{\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{K} f_i (x_i - \bar{x})^2 \cdot \frac{1}{n}\sum_{j=1}^{R} g_j (y_j - \bar{y})^2}}$$

$$\begin{cases} \delta_i(j) & : \text{サンプル } i \text{ がカテゴリー } j \text{ に反応するとき } 1,\ \text{それ以外 } 0 \\ f_i = \sum_{j=1}^{R}\delta_i(j) & : \text{サンプル } i \text{ が反応するカテゴリー数} \\ g_j = \sum_{i=1}^{K}\delta_i(j) & : \text{カテゴリー } j \text{ が反応されるサンプル数} \\ n = \sum_{i=1}^{K} f_i = \sum_{j=1}^{R} g_j & : \text{全反応数} \end{cases}$$

数量化Ⅲ類とは？

相関係数

$$r = \frac{\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)(x_i-\bar{x})(y_j-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{K}f_i(x_i-\bar{x})^2 \cdot \frac{1}{n}\sum_{j=1}^{R}g_j(y_j-\bar{y})^2}}$$

		カテゴリー					
		1	…	j	…	R	計
サンプル	1	1	…	0	…	1	15
	:	:		:		:	
	i	1	…	1	…	0	8
	:	:		:		:	
	K	0	…	1	…	0	21
	計	19		17		7	53

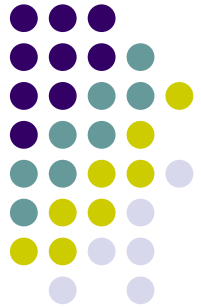
$\delta_i(j)$

f_i

g_j

n

相関係数最大化



相関係数

$$r = \frac{\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)(x_i-\bar{x})(y_j-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{K}f_i(x_i-\bar{x})^2 \cdot \frac{1}{n}\sum_{j=1}^{R}g_j(y_j-\bar{y})^2}}$$

平均 $\bar{x}=\bar{y}=0$，分散1としてよい

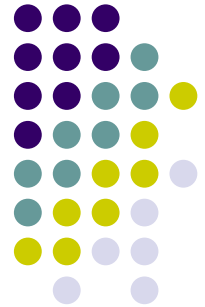
相関係数は**原点の位置**，及び**分散**に**依存しない**から

$$\max . \ \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)x_i y_j$$

$$s.t. \ \frac{1}{n}\sum_{i=1}^{K}f_i x_i^2 = 1, \ \frac{1}{n}\sum_{j=1}^{R}g_j y_j^2 = 1$$

各分散の値が1のもとで r の**分子を最大化**すればよい！

相関係数最大化



- ラグランジュの未定乗数法

$$H = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)x_i y_j - \frac{\lambda}{2}\left(\frac{1}{n}\sum_{i=1}^{K} f_i x_i^2 - 1\right) - \frac{\mu}{2}\left(\frac{1}{n}\sum_{j=1}^{R} g_j y_j^2 - 1\right)$$

- 各変数で偏微分して…

$$\frac{\partial H}{\partial x_i} = 0 (i = 1, \ldots, K), \frac{\partial H}{\partial y_j} = 0 (j = 1, \ldots, R), \frac{\partial H}{\partial \lambda} = 0, \frac{\partial H}{\partial \mu} = 0$$

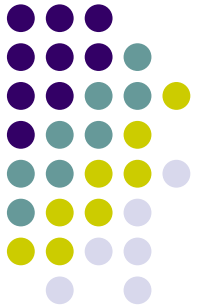
$$\Leftrightarrow \begin{cases} \sum_{j=1}^{R} \frac{\delta_i(j)}{f_i} y_j - \lambda x_i = 0 \ (i = 1, \ldots, K), \\ \sum_{i=1}^{K} \frac{\delta_i(j)}{g_j} x_i - \lambda y_j = 0 \ (j = 1, \ldots, R) \end{cases}$$

ただし，以下を仮定

$f_i > 0 \ (i = 1, \ldots, K)$,
$g_j > 0 \ (j = 1, \ldots, R)$

また，　$\lambda = \mu = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{R}\delta_i(j)x_i y_j$　が成り立つ ……

相関係数最大化



- 一般化固有値問題へ

$$\Rightarrow \begin{cases} \sum_{v=1}^{R}\sum_{i=1}^{K} \dfrac{\delta_i(v)\delta_i(j)}{f_i} y_v - \lambda g_j y_j = 0 \ (j = 1, \ldots, R), \\ \sum_{u=1}^{K}\sum_{j=1}^{R} \dfrac{\delta_u(j)\delta_i(j)}{g_j} x_u - \lambda f_i x_i = 0 \ (i = 1, \ldots, K) \end{cases}$$

$$\Leftrightarrow \begin{cases} (D^T F^{-1} D - \lambda^2 G) y = 0, \\ (D G^{-1} D^T - \lambda^2 F) x = 0 \end{cases}$$

一般化固有値問題

ただし，

$$\begin{cases} D = [\delta_i(j)] \in \Re^{K \times R}, \\ F = diag(f_i) \in \Re^{K^2}, \\ G = diag(g_j) \in \Re^{R^2} \end{cases}$$

相関係数最大化

- 固有値問題へ

$$\Leftrightarrow \begin{cases} \sum_{v=1}^{R}\sum_{i=1}^{K} \dfrac{\delta_i(v)\delta_i(j)}{f_i\sqrt{g_j}\sqrt{g_v}}\sqrt{g_v}\,y_v - \lambda\sqrt{g_j}\,y_j = 0 \ (j=1,\ldots,R), \\ \sum_{u=1}^{K}\sum_{j=1}^{R} \dfrac{\delta_u(j)\delta_i(j)}{g_j\sqrt{f_i}\sqrt{f_u}}\sqrt{f_u}\,x_u - \lambda\sqrt{f_i}\,x_i = 0 \ (i=1,\ldots,K) \end{cases}$$

$$\Leftrightarrow \begin{cases} (G^{-1/2}D^T F^{-1} D G^{-1/2} - \lambda^2 I)\,p = 0, \\ (F^{-1/2}DG^{-1}D^T F^{-1/2} - \lambda^2 I)\,q = 0 \end{cases}$$

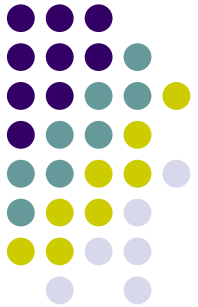
固有値問題

実対称正定値行列

$$\text{ただし,} \begin{cases} p = G^{1/2}y,\ q = F^{1/2}x, \\ F^{1/2} = diag(\sqrt{f_i}) \in \Re^{K^2}, \\ G^{1/2} = diag(\sqrt{g_j}) \in \Re^{R^2} \end{cases}$$

★より，求める固有値 λ は最大にすべき相関係数の分子に等しいので，最大固有値とその固有ベクトルを求めればよい

固有値問題



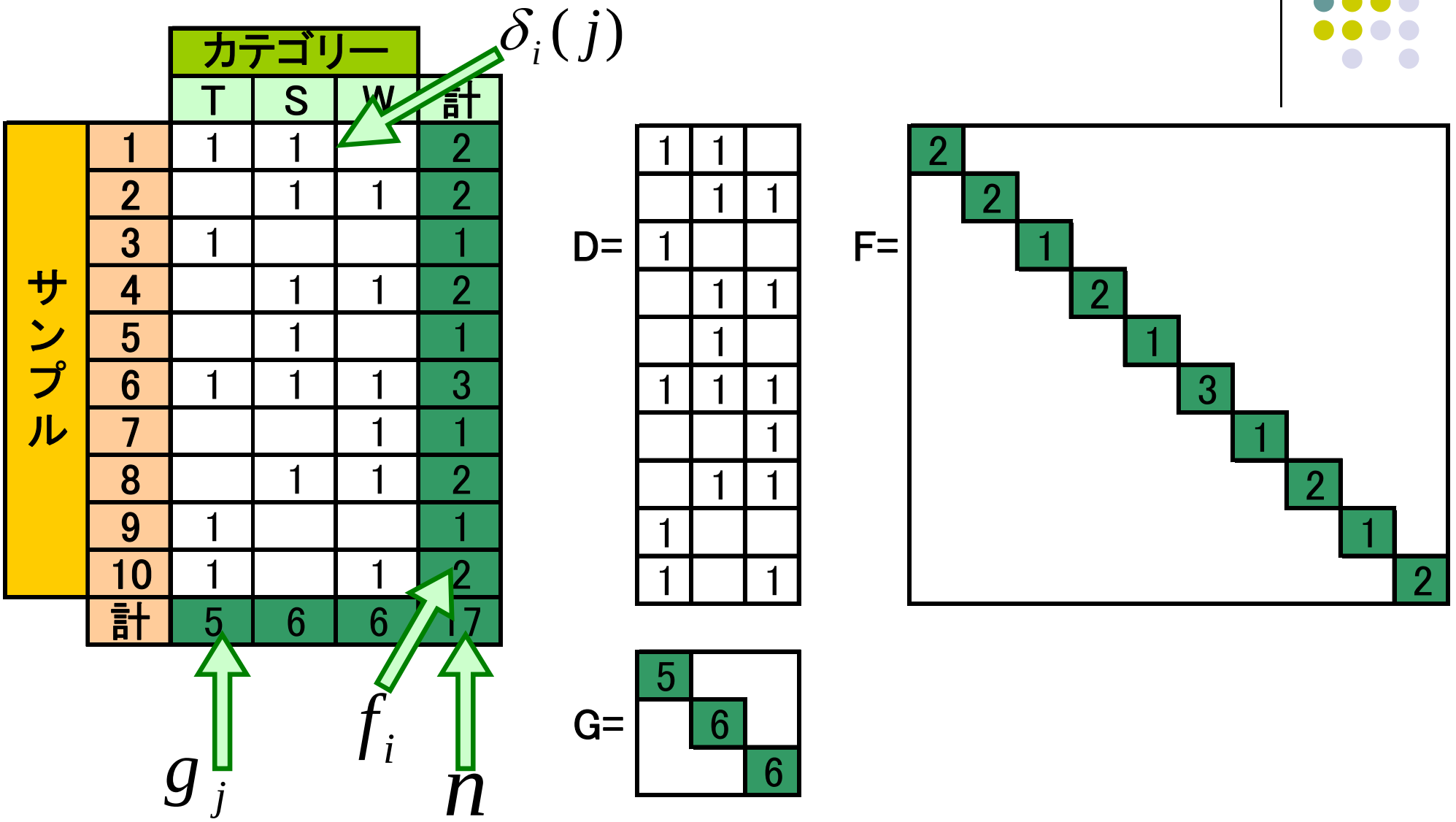
- 最大固有値に関する注意

$$\begin{cases} \lambda = 1, y_1, \cdots, y_R = c, \\ \lambda = 1, x_1, \cdots, x_K = c \end{cases}$$

は，明らかに各々の固有方程式を満たす．そのため，最大固有値は常に1であるが，これは平均が0を満たさない．
　大きさが2番目以降の固有値は満たす．

1を除いた中での最大固有値に対する固有ベクトルを各々求め，それぞれをサンプル，カテゴリーに与える数量として採用する．

例題：個人の嗜好


$\delta_i(j)$
カテゴリー
T S W 計
サンプル
1 1 1 2
2 1 1 2
3 1 1
4 1 1 2
5 1 1
6 1 1 1 3
7 1 1
8 1 1 2
9 1 1
10 1 1 2
計 5 6 6 17
g_j
f_i
n
D=
F=
G=

例題:個人の嗜好

- 結果

		カテゴリー			
		T	S	W	計
サンプル	1	1	1		2
	2		1	1	2
	3	1			1
	4		1	1	2
	5		1		1
	6	1	1	1	3
	7			1	1
	8		1	1	2
	9	1			1
	10	1		1	2
	計	5	6	6	17

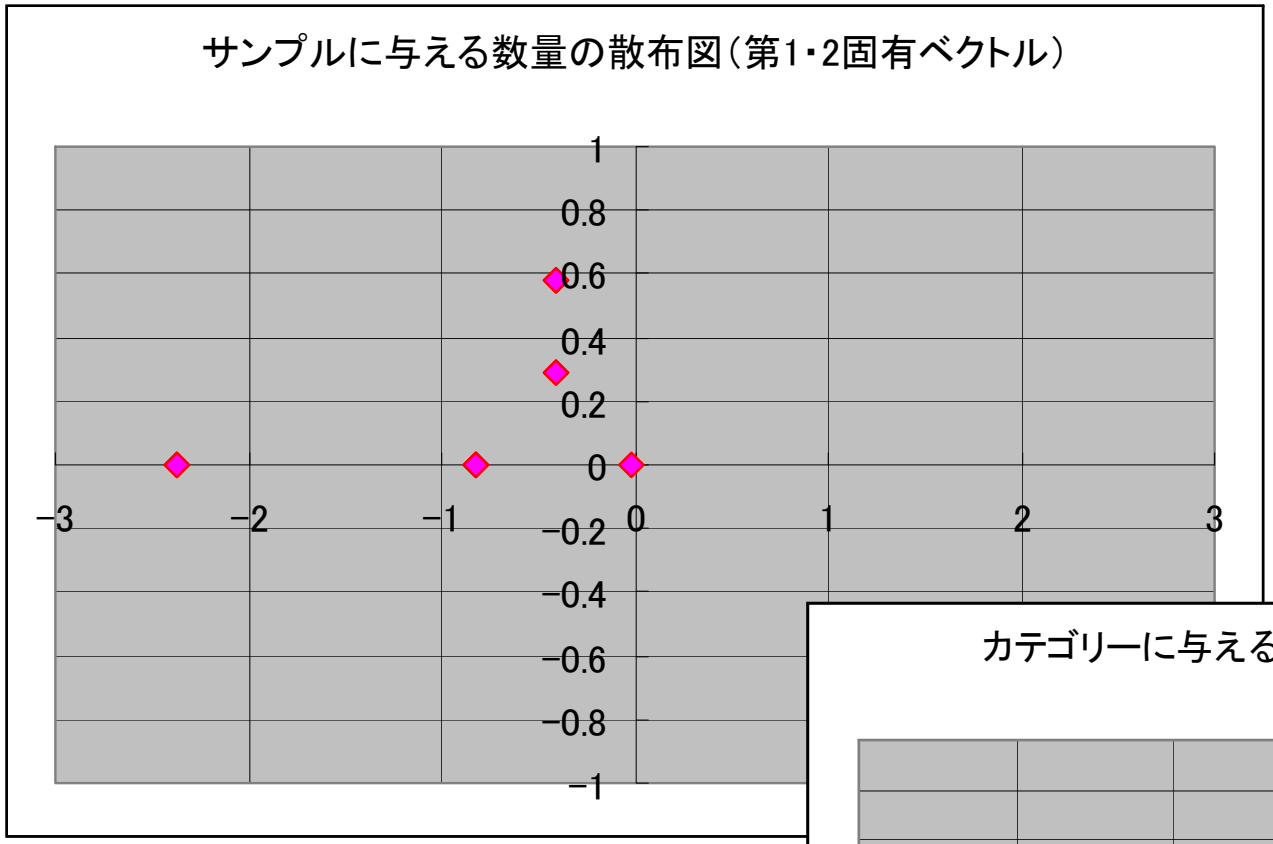
$$x = \begin{bmatrix} 0.7071 \\ -1.0102 \\ 2.4244 \\ -1.0102 \\ -1.0102 \\ 0.1347 \\ -1.0102 \\ -1.0102 \\ 2.4244 \\ 0.7071 \end{bmatrix} \qquad y = \begin{bmatrix} -0.9798 \\ 0.4082 \\ 0.4082 \end{bmatrix}$$

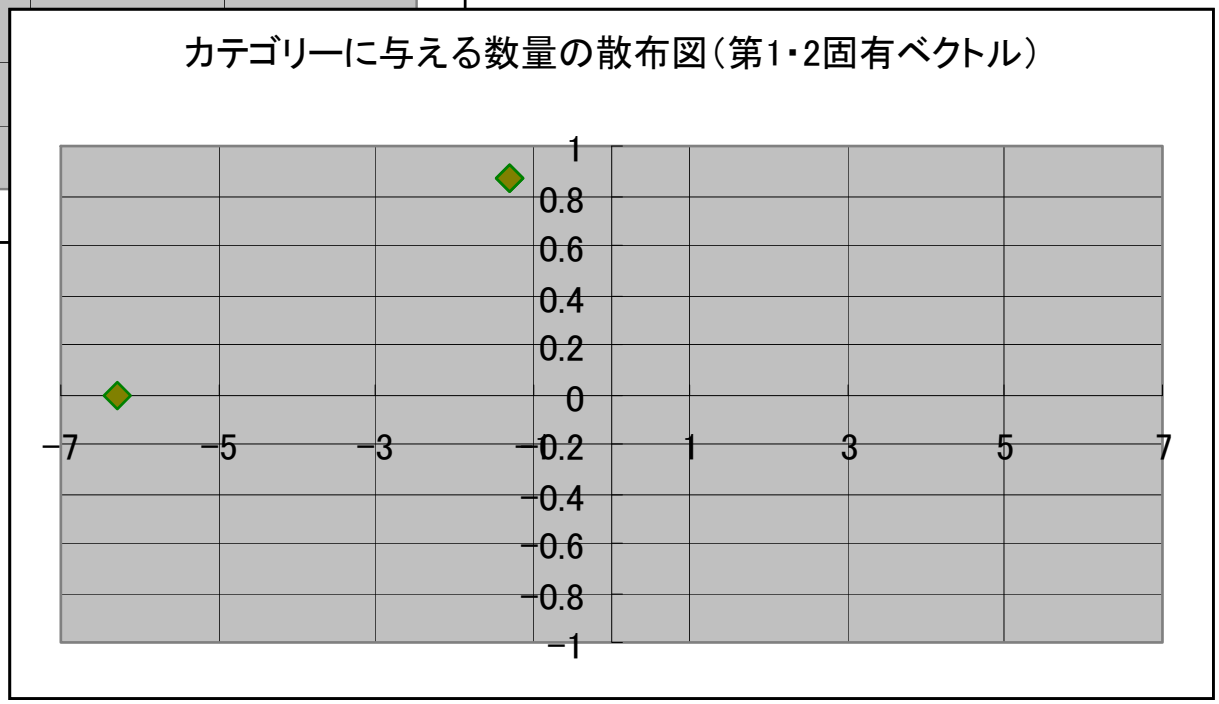
例題：個人の嗜好

			カテゴリー			
		Lambda=	T	S	W	
		0.52778	−0.9798	0.40825	0.40825	サンプルスコア
サンプル	1	0.70711	−0.6928	0.28868	0	−0.404
	2	−1.0102	0	−0.4124	−0.4124	−0.825
	3	2.42437	−2.3754	0	0	−2.375
	4	−1.0102	0	−0.4124	−0.4124	−0.825
	5	−1.0102	0	−0.4124	0	−0.412
	6	0.13469	−0.132	0.05499	0.05499	−0.022
	7	−1.0102	0	0	−0.4124	−0.412
	8	−1.0102	0	−0.4124	−0.4124	−0.825
	9	2.42437	−2.3754	0	0	−2.375
	10	0.70711	−0.6928	0	0.28868	−0.404
		カテゴリースコア	−6.2684	−1.3059	−1.3059	

3	−2.375	○		
9	−2.375	○		
2	−0.825		○	○
4	−0.825		○	○
8	−0.825		○	○
5	−0.412		○	
7	−0.412			○
1	−0.404	○	○	
10	−0.404	○		○
6	−0.022	○	○	○

例題：個人の嗜好


サンプルに与える数量の散布図（第1・2固有ベクトル）
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−3
−2
−1
0
1
2
3


カテゴリーに与える数量の散布図（第1・2固有ベクトル）
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−7
−5
−3
−1
1
3
5
7

参考文献

- 田中豊/脇本和昌「多変量統計解析法」現代数学社
- 圓川隆夫「多変量のデータ解析」朝倉書店
- 杉山高一「多変量データ解析入門」朝倉書店
- 浅利英吉他「パソコンによるデータマイニング」日刊工業新聞社
- 伊里正夫「線形代数Ｉ」岩波書店
- 荒木勉監修「Excelで学ぶデータ解析」実教出版