多変量解析

第2回　多変量解析とは

萩原•篠田
情報理工学部

測定尺度（scale of measurement）の水準

- 分類（名義）尺度　categorical (nominal) scale
  男女、職業など、順序関係のない分類
- 順序尺度　ordinal scale
  1位-2位-3位、軽症-中等度-重傷など
  大小・順序が定義される、差は定義できない
- 間隔尺度　interval scale
  温度など
  順序間の差や距離が定義される
- 比例尺度　ratio scale
  絶対0（ゼロ）が定義できる
  比を論ずることができる

質的変数

量的変数

授業スケジュール・評価

授業回	テーマ	BCPレベル1-2	BCPレベル3-4
第01回(04/11)	測定尺度の水準	ライブ配信	ライブ配信
第02回(04/18)	多変量解析とは	対面	ライブ配信
第03回(04/25)	データの集約	オンデマンド	オンデマンド
第04回(05/02)	有意差検定	対面	ライブ配信
第05回(05/09)	相関	ライブ配信	ライブ配信
第06回(05/16)	単回帰分析	対面	ライブ配信
第07回(05/23)	重回帰分析	ライブ配信	ライブ配信
第08回(05/30)	数量化1類	対面	ライブ配信
第09回(06/06)	判別分析	オンデマンド	オンデマンド
第10回(06/13)	数量化2類	対面	ライブ配信
第11回(06/20)	主成分分析	ライブ配信	ライブ配信
第12回(06/27)	数量化3類	対面	ライブ配信
第13回(07/04)	クラスター分析	オンデマンド	オンデマンド
第14回(07/11)	因子分析	対面	ライブ配信
第15回(07/18)	授業内試験	対面	ライブ配信

相関

- 3回生GPAと入試得点の関連性は？
- 3回生GPAと2回生GPAの関連性は？

keywords

相関係数、共分散・偏差積和（分散・偏差平方和）、内積

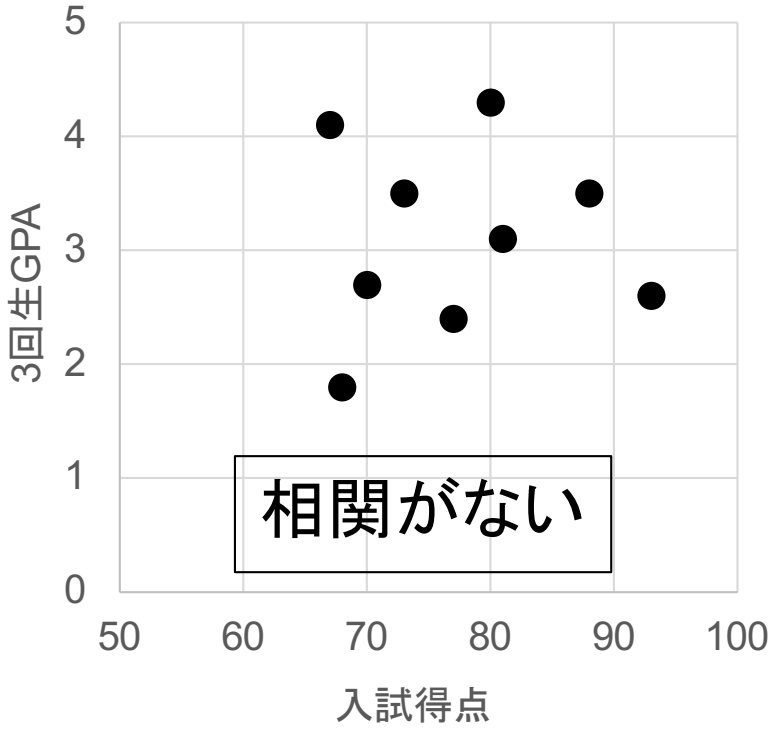
ID	3回生GPA y	入試得点 x_1	2回生GPA x_2	性別 x_3	出身高校 x_4
1	3.5	80	3.7	F	A高校
2	2.4	61	2.3	M	B高校
3	4.1	82	4.0	M	C高校
4	3.1	78	3.4	F	D高校
5	1.8	62	2.2	M	D高校
6	2.7	73	2.0	F	B高校
7	2.6	62	2.1	M	C高校
8	3.5	60	3.2	M	A高校
9	4.3	100	4.4	F	B高校

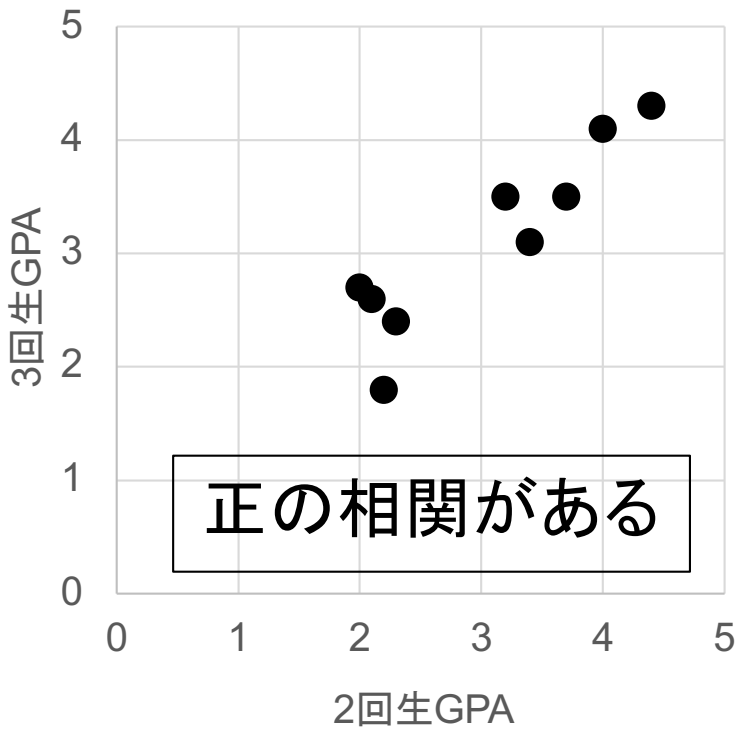
相関

- 3回生GPAと入試得点の関連性は？
- 3回生GPAと2回生GPAの関連性は？

keywords

相関係数、共分散・偏差積和（分散・偏差平方和）、内積





回帰分析

マンション価格は広さと築年数から予測可能か？

量的変数　　量的変数

回帰式：$\hat{y} = 1.02 + 0.067x_1 - 0.081x_2$

サンプル No.	広さ（m^2） x_1	築年数（年） x_2	価格（千万円） y
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

keywords

目的変数、説明変数、
線形回帰、残差、
最小二乗法、
決定係数（寄与率）、
分散共分散行列

数量化1類

量的変数 外的基準　　質的変数 アイテム

卒業時の総合成績は線形代数の成績とサークル所属の有無から予測可能か？

サンプル No.	線形代数 x_1	サークル x_2	総合成績 y
1	優	所属	96
2	優	所属	88
3	優	無所属	77
4	優	無所属	89
5	良	所属	80
6	良	無所属	71
7	良	無所属	77
8	可	所属	78
9	可	所属	70
10	可	無所属	62

keywords
質的変数、重回帰分析、ダミー変数、共線性、予測式、外的基準、カテゴリ数量、基準化

数量化1類

量的変数　外的基準　　質的変数　アイテム

卒業時の総合成績は線形代数の成績とサークル所属の有無から予測可能か？

ダミー変数導入で量的変数に変換→ 重回帰分析

回帰式：$\hat{y} = 83.0 - 10.0x_{11} - 19.0x_{12} + 9.0x_{21}$

サンプル No.	線形代数 x_{11}	線形代数 x_{12}	サークル x_{21}	総合成績 y
1	0	0	1	96
2	0	0	1	88
3	0	0	0	77
4	0	0	0	89
5	1	0	1	80
6	1	0	0	71
7	1	0	0	77
8	0	1	1	78
9	0	1	1	70
10	0	1	0	62

keywords

質的変数、重回帰分析、ダミー変数、共線性、予測式、外的基準、カテゴリ数量、基準化

↓

予測式（回帰式）の定数や係数

判別分析

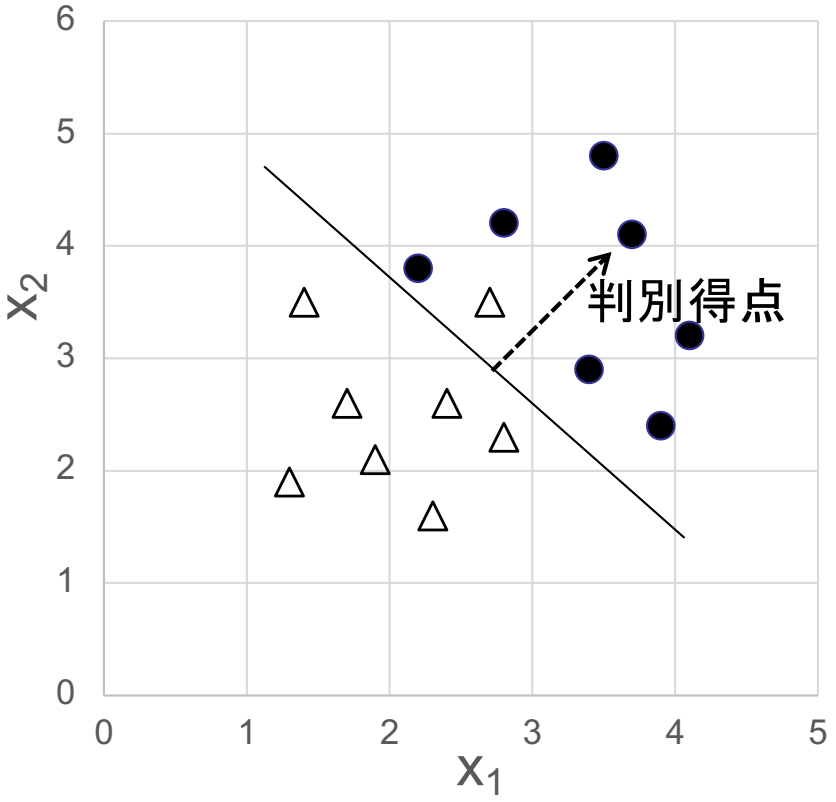
前立腺疾患を腫瘍マーカー1とマーカー2から予測可能か？

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



数量化2類

質的変数

健常者かどうか吐き気と頭痛の有無から予測可能か？

質的変数

サンプル No.	健常者/患者 y	吐き気 x_1	頭痛 x_2
1	健常者	無	少
2	健常者	少	無
3	健常者	無	無
4	健常者	無	無
5	健常者	無	無
6	患者	少	多
7	患者	多	無
8	患者	少	少
9	患者	少	多
10	患者	多	少

keywords

質的変数、判別分析、ダミー変数、予測式、相関比、外的基準、カテゴリ数量、基準化

数量化2類

質的変数

健常者かどうか吐き気と頭痛の有無から予測可能か？

質的変数

ダミー変数導入で量的変数に変換→ 判別分析

判別式：$\hat{y} = 12.8 - 9.6x_{11} - 20.8x_{12} - 6.4x_{21} - 14.4x_{22}$

サンプル No.	健常者/患者 y	吐き気		頭痛	
		x_{11}	x_{12}	x_{21}	x_{22}
1	健常者	0	0	1	0
2	健常者	1	0	0	0
3	健常者	0	0	0	0
4	健常者	0	0	0	0
5	健常者	0	0	0	0
6	患者	1	0	0	1
7	患者	0	1	0	0
8	患者	1	0	1	0
9	患者	1	0	0	1
10	患者	0	1	1	0

$\hat{y} \geq 0$ 健常者

$\hat{y} < 0$ 患者

keywords

質的変数、判別分析、ダミー変数、予測式、相関比、外的基準、カテゴリ数量、基準化

主成分分析

学力の特徴(分布)を少ない変数(主成分)で表現できないか？

第1主成分 $z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$ 総合的学力

第2主成分 $z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$ 文系･理系志向

u_1, u_2, u_3, u_4 は x_1, x_2, x_3, x_4 を標準化した変数

生徒No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

寄与率　第1主成分: 0.680

第2主成分: 0.306

累積: 0.986

↓

第2主成分までで4次元データの98.6%までが表現できる

keywords

説明変数、総合的指標、主成分、主成分得点、寄与率、情報損失量、固有値、固有ベクトル

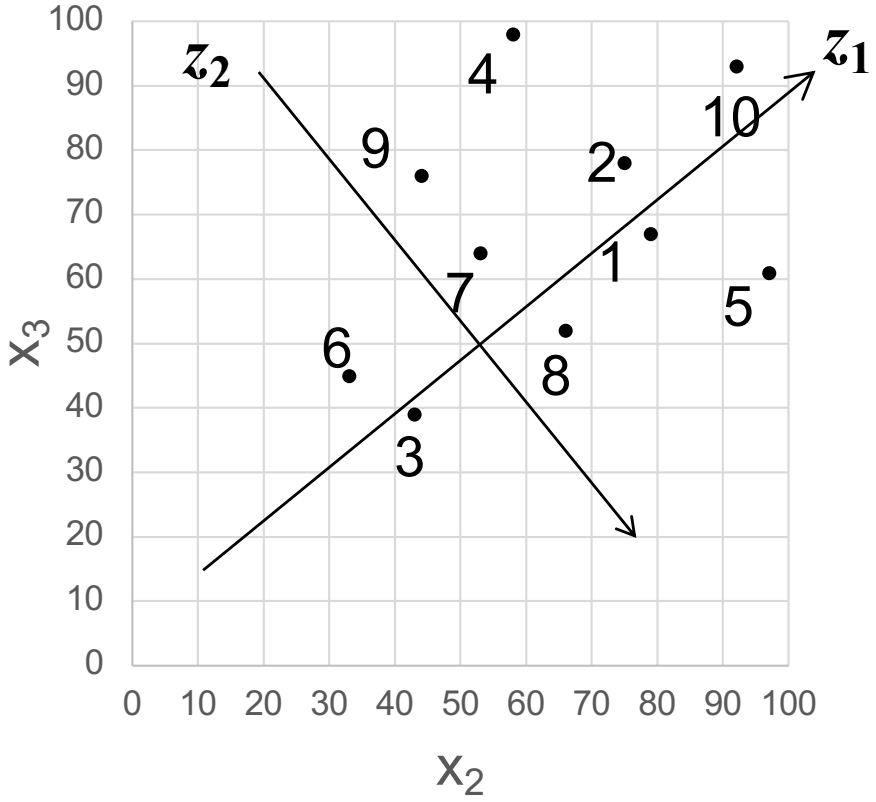
主成分分析

学力の特徴(分布)を少ない変数(主成分)で表現できないか？

第1主成分 $z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$ 総合的学力

第2主成分 $z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$ 文系･理系志向

u_1, u_2, u_3, u_4 は x_1, x_2, x_3, x_4 を標準化した変数



寄与率　第1主成分: 0.680

第2主成分: 0.306

累積: 0.986

↓

第2主成分までで4次元データの98.6%までが表現できる

keywords

説明変数、総合的指標、主成分、主成分得点、寄与率、情報損失量、固有値、固有ベクトル

数量化3類

学生と酒類の特徴づけや分類ができないか？

主成分分析（量的変数） ⟷ 数量化3類（質的変数）

学生のお酒の好み（〇印）

学生	チューハイ	日本酒	ビール
1		〇	〇
2	〇		〇
3	〇		

↓

相関係数が最大となるように割り当てた数量（a_i, b_j）を求める

学生		チューハイ b_1	日本酒 b_2	ビール b_3
1	a_1		(a_1, b_2)	(a_1, b_3)
2	a_2	(a_2, b_1)		(a_2, b_3)
3	a_3	(a_3, b_1)		

固有値・固有ベクトル

→ 成分を軸にもつ平面に酒類（学生）をプロット

↓

類似した酒類（学生）は近くに分布（＝分類可能）

直感的には並び替え！

学生	チューハイ	ビール	日本酒
1		〇	〇
2	〇	〇	
3	〇		

a_i

b_j

keywords

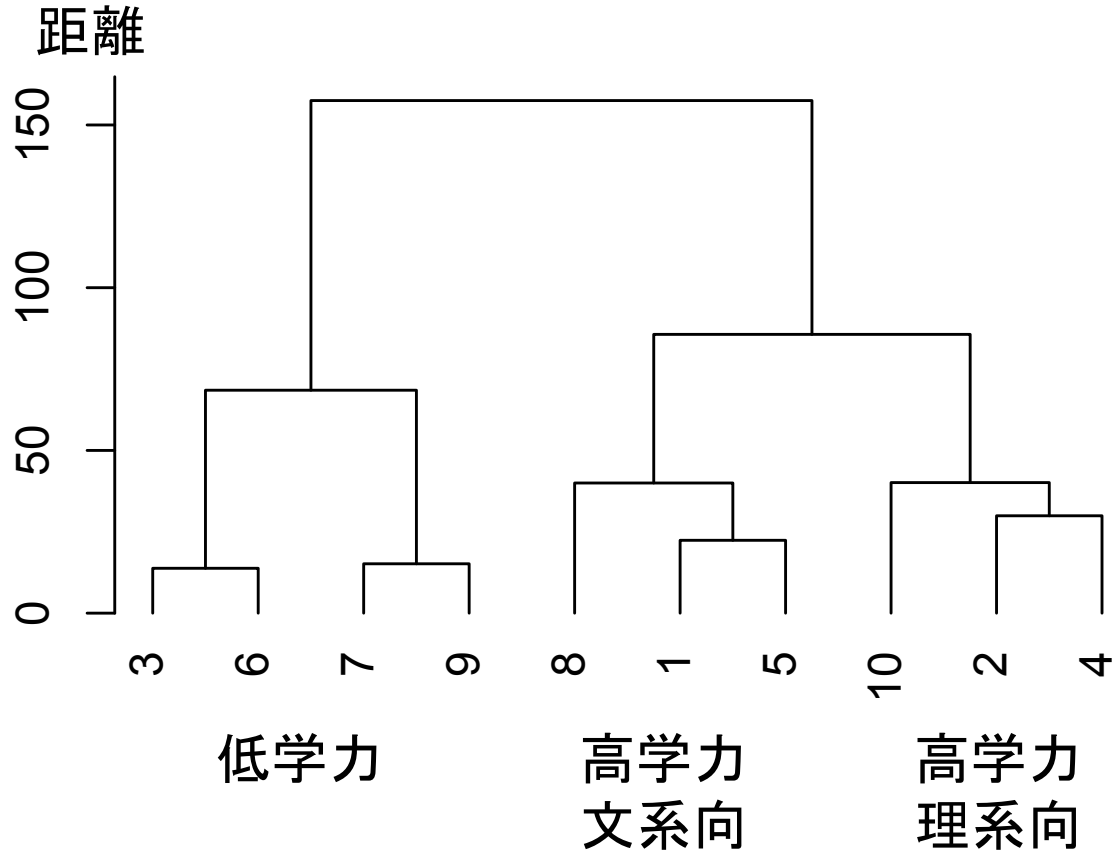
質的変数、主成分分析、相関係数、サンプルスコア、カテゴリ数量（変数スコア）、固有値、固有ベクトル

クラスター分析

類似の能力をもつ生徒をグループ化できるか？
それぞれのグループの特徴は何か？

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

クラスターを樹形図（デンドログラム）で表示

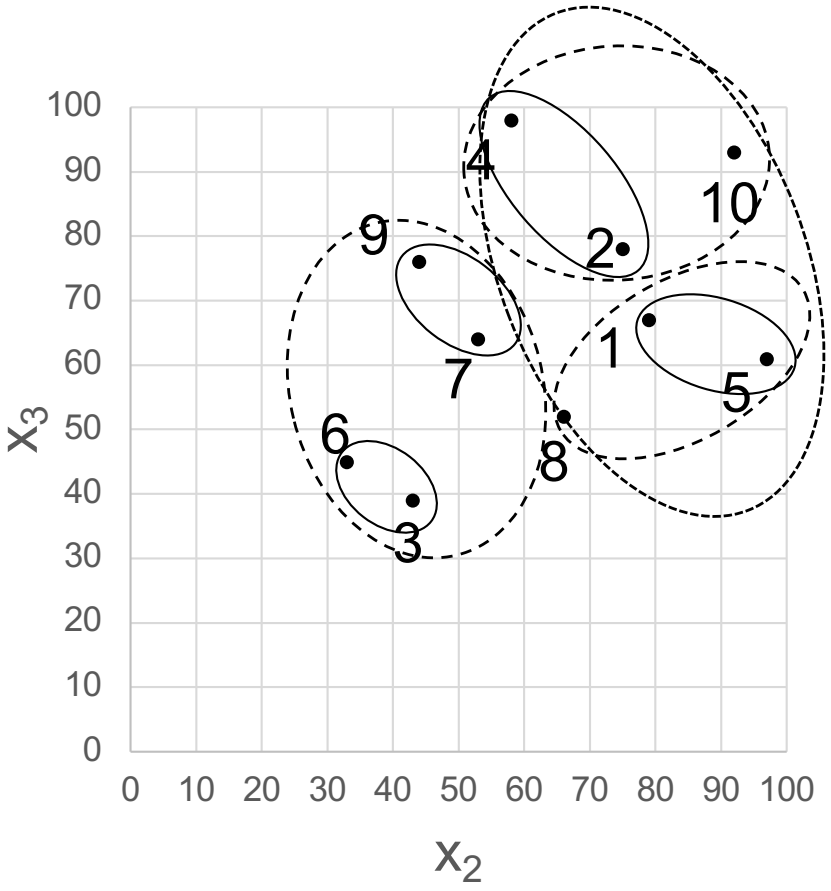


keywords

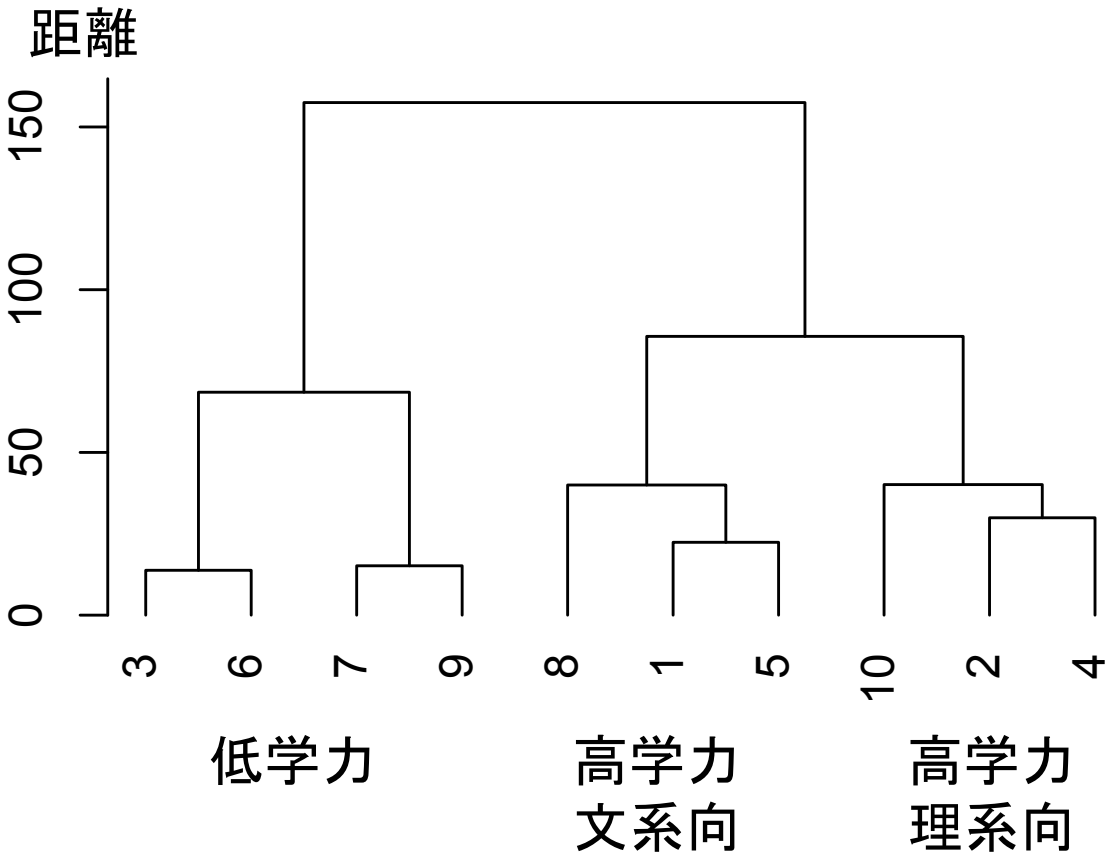
クラスター、距離（ユークリッド、マハラノビス）、デンドログラム

クラスター分析

類似の能力をもつ生徒をグループ化できるか？
それぞれのグループの特徴は何か？



クラスターを樹形図（デンドログラム）で表示



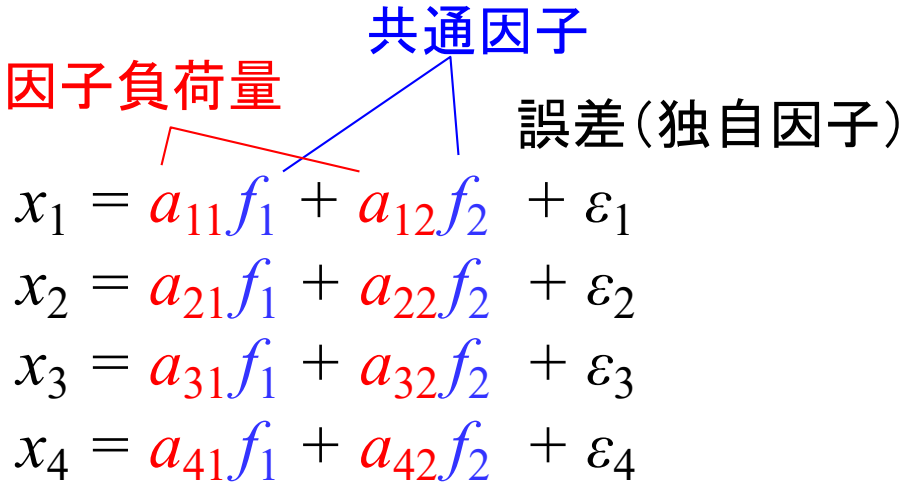
keywords

クラスター、距離（ユークリッド、マハラノビス）、デンドログラム

因子分析

多数の変数間の相関を少ない潜在因子で説明する
→ 共通因子の抽出

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91



f_1: 文系的能力に関係する因子
f_2: 理系的能力に関係する因子

keywords

要因、観測変数、潜在変数、共通因子、因子負荷量、因子得点