多元分析方法使用报告

刘士坤　郑辉杨　薛兆浩　余涛

天津商业大学　统计系

2020 年 10 月 21 日

本次报告所使用的分析方法

- 聚类分析（Cluster analysis）
  - 系统聚类（Hierarchical clustering）
  - K-means 聚类（k-means clustering）
- 主成分分析（Principal component analysis）
- 线性判别分析（Linear discriminant analysis）

本次报告所使用的数据

某中学火箭班、实验班、重点班、普通班 80 名同学的期中考试成绩

姓名	语文	数学	英语	物理	化学	生物	历史	地理	政治	分班名
杜冰燕	93.00	134.00	122.50	88.00	82.00	85.00	68.00	74.00	62.00	火箭班
刘宁涛	94.00	133.00	106.50	74.00	79.00	88.00	75.00	93.00	63.00	火箭班
刘天乐	99.00	126.00	116.50	69.00	85.00	70.00	73.00	80.50	69.00	火箭班
白煜东	88.00	138.00	125.00	68.00	71.00	82.00	64.00	75.00	74.00	火箭班
刘广胜	100.00	131.00	111.50	66.00	83.00	66.00	76.00	87.00	62.00	火箭班
曹美池	72.00	111.00	116.50	68.00	85.00	82.00	86.00	74.50	78.00	火箭班
杨晨朋	95.00	140.00	93.00	68.00	81.00	66.00	79.00	92.00	48.00	火箭班
赵敏	105.00	135.00	102.50	70.00	76.00	64.00	70.00	71.00	63.00	火箭班
邢媛媛	98.00	126.00	120.00	59.00	72.00	71.00	75.00	63.00	71.00	火箭班
张星宇	84.00	132.00	113.00	61.00	72.00	64.00	66.00	82.50	77.00	火箭班
邵振东	95.00	109.00	107.00	71.00	68.00	86.00	76.00	79.00	55.00	火箭班
郝亚欣	89.00	120.00	108.00	56.00	59.00	86.00	67.00	84.50	76.00	火箭班
张傲然	106.00	99.00	97.50	70.00	83.00	83.00	66.00	62.00	63.00	火箭班
姚伟勋	89.00	102.00	82.50	70.00	81.00	82.00	78.00	89.50	63.00	火箭班
李柯	92.00	131.00	102.00	64.00	76.00	70.00	65.00	73.00	62.00	火箭班
李天源	76.00	122.00	105.00	78.00	86.00	81.00	52.00	89.50	45.00	火箭班
赵世颖	96.00	105.00	118.50	45.00	69.00	82.00	68.00	81.50	68.00	火箭班
姜新	99.00	101.00	96.50	57.00	48.00	70.00	68.00	74.00	75.00	火箭班
张茜	105.00	61.00	115.00	44.00	61.00	61.00	75.00	63.00	59.00	火箭班
郑海鹏	78.00	114.00	88.50	70.00	66.00	81.00	76.00	91.00	67.00	火箭班
马裕祖	100.00	140.00	92.50	76.00	73.00	84.00	71.00	95.00	56.00	火箭班
葛赛楠	86.00	77.00	77.00	45.00	40.00	52.00	53.00	73.00	49.00	实验班
沈沐菲	64.00	72.00	103.00	27.00	39.00	56.00	49.00	91.50	49.00	实验班
张梓晴	93.00	64.00	106.00	35.00	37.00	45.00	54.00	70.00	46.00	实验班
王文澳	60.00	96.00	87.00	49.00	39.00	53.00	60.00	59.00	47.00	实验班
石孟矾	92.00	65.00	99.50	33.00	43.00	53.00	61.00	50.00	53.00	实验班
李宾宾	88.00	83.00	61.50	44.00	61.00	43.00	57.00	73.50	38.00	实验班
杨航凯	68.00	71.00	104.00	42.00	47.00	40.00	70.00	46.50	60.00	实验班
刘一帆	100.00	37.00	93.50	37.00	56.00	53.00	50.00	84.00	38.00	实验班
张士野	98.00	73.00	71.50	57.00	51.00	47.00	58.00	67.00	26.00	实验班
任昊楠	96.00	81.00	72.00	50.00	42.00	46.00	61.00	60.00	40.00	实验班
陈子波	84.00	63.00	45.00	49.00	54.00	66.00	84.00	57.50	45.00	实验班
王琳	77.00	61.00	123.00	30.00	39.00	49.00	48.00	65.50	55.00	实验班
白梦雨	86.00	82.00	84.50	11.00	40.00	53.00	68.00	58.00	65.00	实验班
李露露	104.00	49.00	96.00	39.00	46.00	56.00	64.00	44.50	49.00	实验班
焦树敏	89.00	43.00	82.50	28.00	39.00	51.00	71.00	89.00	55.00	实验班
李佳恒	81.00	48.00	83.00	55.00	57.00	55.00	52.00	72.00	44.00	实验班
秦雷超	80.00	68.00	75.00	47.00	42.00	51.00	59.00	71.00	54.00	实验班
苗轩旗	74.00	61.00	103.00	39.00	30.00	48.00	64.00	77.00	51.00	实验班
陈聪聪	83.00	61.00	108.00	29.00	35.00	57.00	52.00	88.00	33.00	实验班

姓名	语文	数学	英语	物理	化学	生物	历史	地理	政治	分班名
程辉	93.00	51.00	78.00	30.00	31.00	30.00	16.00	39.50	30.00	重点班
李洁玉	101.00	85.00	120.50	48.00	64.00	67.00	65.00	51.00	54.00	重点班
郑晓伟	88.00	42.00	93.50	21.00	20.00	23.00	36.00	42.00	33.00	重点班
王浩楠	84.00	29.00	54.00	13.00	30.00	38.00	48.00	68.50	34.00	重点班
闫雅雯	65.00	22.00	74.00	22.00	27.00	40.00	42.00	61.50	45.00	重点班
宋艳晶	97.00	37.00	99.50	26.00	21.00	15.00	37.00	31.50	34.00	重点班
郝可寒	57.00	30.00	76.50	20.00	43.00	41.00	36.00	53.00	41.00	重点班
孙悦	88.00	26.00	48.00	30.00	30.00	36.00	42.00	55.50	42.00	重点班
韩创	81.00	45.00	32.50	34.00	36.00	31.00	45.00	40.00	53.00	重点班
程娜娜	84.00	12.00	60.00	32.00	29.00	41.00	36.00	63.50	40.00	重点班
冯龙超	92.00	89.00	102.00	47.00	47.00	64.00	79.00	92.00	38.00	重点班
李瑞星	78.00	91.00	117.50	33.00	51.00	51.00	66.00	67.00	62.00	重点班
曹永鑫	73.00	99.00	107.50	63.00	52.00	50.00	68.00	73.50	30.00	重点班
李东娜	91.00	34.00	65.00	20.00	31.00	30.00	49.00	39.50	37.00	重点班
田欣园	79.00	27.00	95.50	27.00	36.00	28.00	39.00	40.00	25.00	重点班
谢亚培	80.00	52.00	55.00	29.00	34.00	28.00	42.00	40.50	36.00	重点班
侯亚菲	82.00	25.00	87.50	24.00	23.00	39.00	43.00	36.50	36.00	重点班
佘嘉琪	92.00	27.00	72.00	17.00	24.00	30.00	41.00	51.00	42.00	重点班
孙鹏航	82.00	58.00	57.00	15.00	28.00	27.00	45.00	54.00	30.00	重点班
李举举	72.00	32.00	20.00	30.00	38.00	41.00	54.00	60.00	49.00	重点班
李峥	45.00	7.00	27.00	20.00	20.00	24.00	20.00	22.00	33.00	普通班
刘赛恒	92.00	93.00	67.50	40.00	61.00	71.00	73.00	82.00	55.00	普通班
梁智斌	11.00	30.00	25.50	20.00	23.00	20.00	24.00	33.00	27.00	普通班
苗晓帅	77.00	133.00	86.50	51.00	72.00	79.00	69.00	90.50	44.00	普通班
任宁雪	93.00	84.00	80.50	55.00	58.00	75.00	83.00	87.00	60.00	普通班
王嘉欣	52.00	15.00	38.50	14.00	18.00	11.00	15.00	20.50	10.00	普通班
马新丽	55.00	40.00	20.50	15.00	22.00	24.00	21.00	14.00	23.00	普通班
陈涛	13.00	10.00	64.50	14.00	12.00	25.00	22.00	45.50	28.00	普通班
谷伟康	58.00	20.00	21.50	11.00	17.00	30.00	31.00	18.50	26.00	普通班
段瑞杰	56.00	15.00	19.00	21.00	21.00	19.00	23.00	30.50	25.00	普通班
孔伟潇	54.00	17.00	20.50	14.00	11.00	9.00	34.00	37.50	27.00	普通班
张贵珍	76.00	15.00	34.50	18.00	19.00	12.00	22.00	18.00	25.00	普通班
赵辰铮	57.00	20.00	30.00	21.00	26.00	22.00	24.00	28.50	11.00	普通班
姚冀远	26.00	37.00	35.50	23.00	18.00	13.00	28.00	39.50	19.00	普通班
李涵凝	44.00	20.00	67.50	20.00	17.00	14.00	23.00	26.00	16.00	普通班
沈洋	61.00	11.00	30.50	23.00	27.00	19.00	29.00	25.00	21.00	普通班
李龙琰	75.00	10.00	21.50	16.00	23.00	9.00	47.00	28.50	21.00	普通班
汲孟超	54.00	15.00	28.50	17.00	32.00	27.00	31.00	23.50	22.00	普通班
郭子豪	41.00	28.00	50.00	19.00	27.00	11.00	31.00	17.50	24.00	普通班
刘祥	60.00	20.00	44.50	11.00	27.00	13.00	27.00	25.00	21.00	普通班

第一部分

聚类分析（Cluster analysis）

内容

描述性统计

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
语文	80	11.00	106.00	78.9375	20.19233	407.730
数学	80	7.00	140.00	66.4750	41.51772	1723.721
英语	80	19.00	125.00	78.0562	31.64908	1001.665
物理	80	11.00	88.00	39.5250	20.53646	421.746
化学	80	11.00	86.00	45.4875	21.60813	466.911
生物	80	9.00	88.00	48.1875	23.37861	546.559
历史	80	15.00	86.00	52.9375	19.22113	369.452
地理	80	14.00	95.00	59.3000	23.21155	538.776
政治	80	10.00	78.00	44.7750	17.15379	294.253
Valid N (listwise)	80					

内容

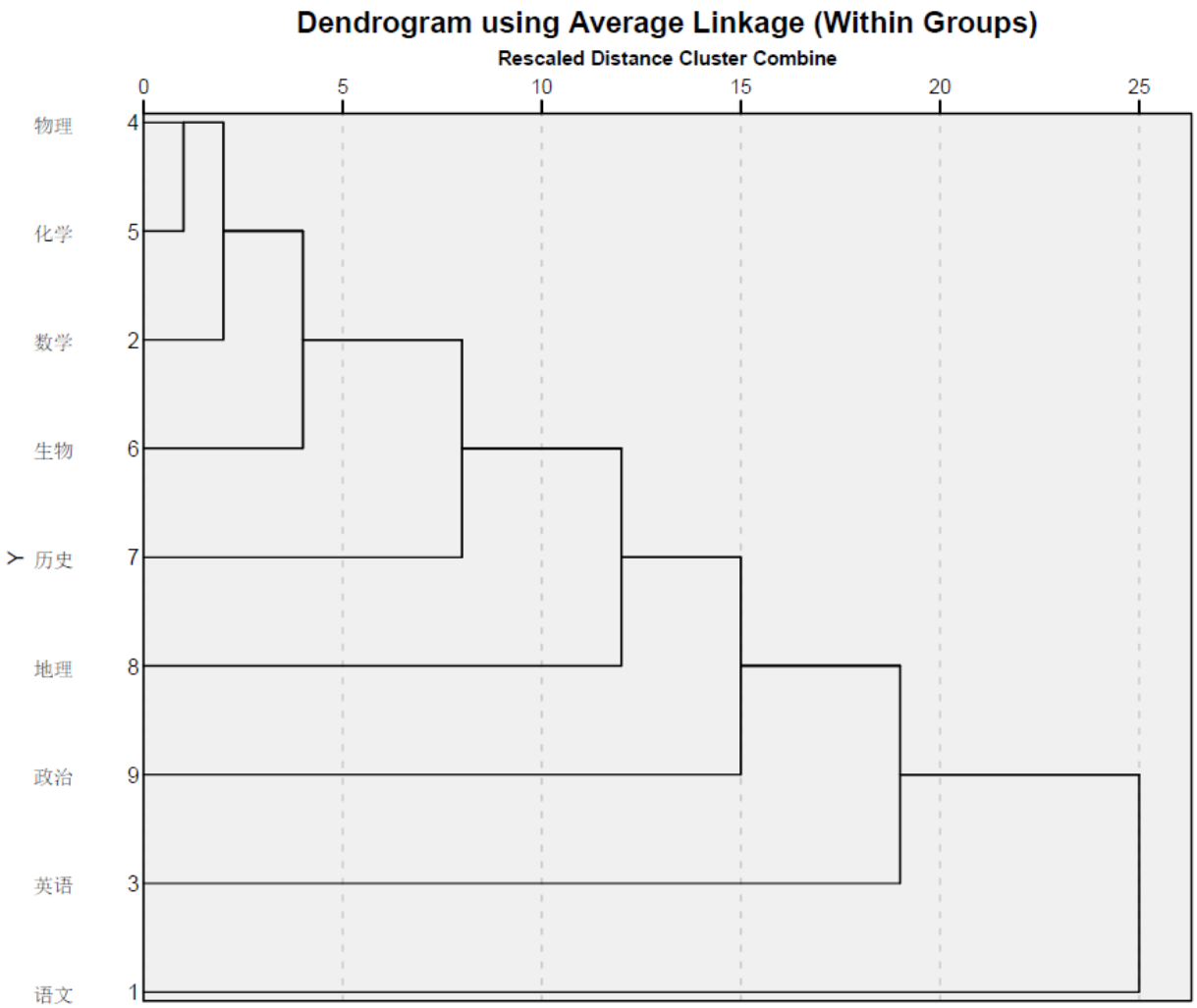
选择平方欧氏距离时的距离矩阵

Proximity Matrix

Matrix File Input

Case	语文	数学	英语	物理	化学	生物	历史	地理
语文	.000	71.490	64.116	70.850	63.851	62.179	53.230	71.046
数学	71.490	.000	43.990	17.732	16.105	20.277	31.017	37.018
英语	64.116	43.990	.000	55.247	51.283	42.976	50.987	50.317
物理	70.850	17.732	55.247	.000	13.837	21.357	36.507	42.777
化学	63.851	16.105	51.283	13.837	.000	16.033	29.734	39.775
生物	62.179	20.277	42.976	21.357	16.033	.000	22.275	24.069
历史	53.230	31.017	50.987	36.507	29.734	22.275	.000	31.705
地理	71.046	37.018	50.317	42.777	39.775	24.069	31.705	.000
政治	66.805	36.033	45.501	47.352	38.215	26.078	30.321	47.559

组内联结法的谱系聚类图



选择聚成 4 类的结果

Cluster Membership

Case	4 Clusters
语文	1
数学	2
英语	3
物理	2
化学	2
生物	2
历史	2
地理	2
政治	4

内容

4 个初始中心

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore: 语文	1.04309	-3.36452	-.34357	-.19500
Zscore: 数学	-.70994	-.87854	1.07243	-1.36026
Zscore: 英语	.48797	-1.66059	1.21469	-1.78698
Zscore: 物理	-.12295	-.95075	1.38656	-1.14552
Zscore: 化学	.48651	-1.04070	1.82859	-1.04070
Zscore: 生物	.20585	-1.20570	1.44630	-1.67621
Zscore: 历史	-.15283	-1.50550	1.72011	-.30890
Zscore: 地理	1.06413	-1.13306	.65485	-1.32693
Zscore: 政治	-.39496	-1.03621	1.93689	-1.38599

聚类结果

Cluster Membership

Case Number	姓名	Cluster	Distance
1	杜冰燕	3	1.403
2	刘宁涛	3	.900
3	刘天乐	3	.882
4	白煜东	3	1.133
5	刘广胜	3	.900
6	曹美池	3	1.686
7	杨晨朋	3	1.380
8	赵敏	3	1.054
9	邢媛媛	3	1.181
10	张星宇	3	1.167
11	邵振东	3	.847
12	郝亚欣	3	1.216
13	张傲然	3	1.371
14	姚伟勋	3	1.064
15	李柯	3	.651
16	李天源	3	1.902
17	赵世颖	3	1.283
18	姜新	3	1.647
19	张茜	1	1.719
20	郑海鹏	3	1.096
21	马裕祖	3	1.174
22	葛赛楠	1	.822
23	沈沐菲	1	1.819
24	张梓晴	1	.988
25	王文澳	1	1.586
26	石孟矾	1	1.057
27	李宾宾	1	1.460
28	杨航凯	1	1.689
29	刘一帆	1	1.594
30	张士野	1	1.791
31	任昊楠	1	1.200
32	陈子波	1	2.086
33	王琳	1	1.551
34	白梦雨	1	1.913

Cluster Membership

Case Number	姓名	Cluster	Distance
35	李露露	1	1.503
36	焦树敏	1	1.491
37	李佳恒	1	1.226
38	秦雷超	1	.840
39	苗轩旗	1	1.162
40	陈聪聪	1	1.641
41	程辉	4	1.636
42	李洁玉	1	1.820
43	郑晓伟	4	1.207
44	王浩楠	4	1.220
45	闫雅雯	4	1.221
46	宋艳晶	4	1.725
47	郝可寒	4	1.526
48	孙悦	4	.851
49	韩创	4	1.580
50	程娜娜	4	1.053
51	冯龙超	1	1.649
52	李瑞星	1	1.393
53	曹永鑫	1	1.879
54	李东娜	4	.770
55	田欣园	4	1.311
56	谢亚培	4	.766
57	侯亚菲	4	.982
58	佘嘉琪	4	.762
59	孙鹏航	4	1.005
60	李举举	4	1.981
61	李峥	2	.906
62	刘赛恒	1	1.662
63	梁智斌	2	1.984
64	苗晓帅	3	1.723
65	任宁雪	3	1.635
66	王嘉欣	2	1.063
67	马新丽	2	.990
68	陈涛	2	2.310

Cluster Membership

Case Number	姓名	Cluster	Distance
69	谷伟康	2	.989
70	段瑞杰	2	.684
71	孔伟潇	2	1.037
72	张贵珍	2	1.433
73	赵辰铮	2	.856
74	姚冀远	2	1.413
75	李涵凝	2	1.197
76	沈洋	2	.751
77	李龙琰	2	1.767
78	汲孟超	2	.759
79	郭子豪	2	.919
80	刘祥	2	.784

各类中的样品数

Number of Cases in each Cluster

Cluster	1	25.000
	2	17.000
	3	22.000
	4	16.000
Valid		80.000
Missing		.000

方差分析表

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore: 语文	16.244	3	.398	76	40.789	.000
Zscore: 数学	23.310	3	.119	76	195.342	.000
Zscore: 英语	18.278	3	.318	76	57.487	.000
Zscore: 物理	21.744	3	.181	76	120.014	.000
Zscore: 化学	22.574	3	.148	76	152.121	.000
Zscore: 生物	23.714	3	.103	76	229.308	.000
Zscore: 历史	21.625	3	.186	76	116.340	.000
Zscore: 地理	20.422	3	.233	76	87.520	.000
Zscore: 政治	19.748	3	.260	76	75.969	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

第二部分

主成分分析
（Principal component analysis）

样本相关阵

Correlation Matrix

		语文	数学	英语	物理	化学	生物	历史	地理	政治
Correlation	语文	1.000	.548	.594	.552	.596	.606	.663	.550	.577
	数学	.548	1.000	.722	.888	.898	.872	.804	.766	.772
	英语	.594	.722	1.000	.650	.675	.728	.677	.682	.712
	物理	.552	.888	.650	1.000	.912	.865	.769	.729	.700
	化学	.596	.898	.675	.912	1.000	.899	.812	.748	.758
	生物	.606	.872	.728	.865	.899	1.000	.859	.848	.835
	历史	.663	.804	.677	.769	.812	.859	1.000	.799	.808
	地理	.550	.766	.682	.729	.748	.848	.799	1.000	.699
	政治	.577	.772	.712	.700	.758	.835	.808	.699	1.000

公因子方差表

Communalities

	Initial	Extraction
语文	1.000	.994
数学	1.000	.921
英语	1.000	.993
物理	1.000	.951
化学	1.000	.945
生物	1.000	.932
历史	1.000	.904
地理	1.000	.982
政治	1.000	.976

Extraction Method: Principal Component Analysis.

各个主成分解释原始变量总方差表

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.948	77.199	77.199	6.948	77.199	77.199
2	.606	6.736	83.936	.606	6.736	83.936
3	.408	4.535	88.471	.408	4.535	88.471
4	.337	3.747	92.218	.337	3.747	92.218
5	.298	3.314	95.532	.298	3.314	95.532
6	.144	1.603	97.135			
7	.109	1.213	98.348			
8	.084	.935	99.284			
9	.064	.716	100.000			

Extraction Method: Principal Component Analysis.

成分矩阵

Component Matrix[a]

	Component				
	1	2	3	4	5
语文	.707	.640	.288	.034	.028
数学	.926	-.205	.042	.138	-.011
英语	.812	.209	-.421	.316	.112
物理	.901	-.241	.220	.175	.040
化学	.930	-.179	.186	.114	-.032
生物	.956	-.107	-.004	-.082	-.003
历史	.913	.063	.031	-.246	-.071
地理	.866	-.051	-.128	-.300	.350
政治	.871	.046	-.213	-.125	-.393

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

$$y_1 = 0.268x_1 + 0.351x_2 + 0.308x_3 + 0.342x_4 + 0.353x_5 + 0.363x_6 + 0.346x_7 + 0.329x_8 + 0.330x_9$$

$$y_2 = 0.822x_1 - 0.263x_2 + 0.268x_3 - 0.310x_4 - 0.230x_5 - 0.137x_6 + 0.081x_7 - 0.066x_8 + 0.059x_9$$

$$y_3 = 0.451x_1 + 0.066x_2 - 0.659x_3 + 0.344x_4 + 0.291x_5 - 0.006x_6 + 0.049x_7 - 0.200x_8 - 0.333x_9$$

$$y_4 = 0.059x_1 + 0.238x_2 + 0.544x_3 + 0.301x_4 + 0.196x_5 - 0.141x_6 - 0.424x_7 - 0.518x_8 - 0.215x_9$$

$$y_5 = 0.051x_1 - 0.020x_2 + 0.205x_3 + 0.073x_4 - 0.059x_5 - 0.005x_6 - 0.130x_7 + 0.641x_8 - 0.720x_9$$

第三部分

线性判别分析
（Linear discriminant analysis）

对主成分均值的检验

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
第一主成分	.261	71.627	3	76	.000
第二主成分	.634	14.628	3	76	.000
第三主成分	.867	3.877	3	76	.012
第四主成分	.661	12.965	3	76	.000
第五主成分	.827	5.306	3	76	.002

对主成分协方差矩阵的 Box's M 检验

Box's Test of Equality of Covariance Matrices

Log Determinants

分班名	Rank	Log Determinant
火箭班	5	23.596
实验班	5	27.018
重点班	5	27.267
普通班	5	26.792
Pooled within-groups	5	27.770

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		121.711
F	Approx.	2.380
	df1	45
	df2	14297.290
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Fisher 线性判别

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.218[a]	78.1	78.1	.873
2	.834[a]	20.3	98.4	.674
3	.066[a]	1.6	100.0	.248

a. First 3 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.121	157.158	15	.000
2 through 3	.512	49.929	8	.000
3	.938	4.743	3	.192

标准化/非标准化 Fisher 线性判别函数

Standardized Canonical Discriminant Function Coefficients

	Function		
	1	2	3
第一主成分	1.038	.100	-.137
第二主成分	.138	.910	.424
第三主成分	-.059	-.265	.360
第四主成分	.198	-.333	.757
第五主成分	-.260	.454	-.405

Canonical Discriminant Function Coefficients

	Function		
	1	2	3
第一主成分	.031	.003	-.004
第二主成分	.009	.061	.029
第三主成分	-.004	-.016	.022
第四主成分	.014	-.023	.052
第五主成分	-.023	.039	-.035
(Constant)	-5.416	-3.608	-.432

Unstandardized coefficients

分类函数

Classification Function Coefficients

	分班名			
	火箭班	实验班	重点班	普通班
第一主成分	.269	.208	.174	.120
第二主成分	.292	.355	.375	.229
第三主成分	.001	-.027	-.010	.020
第四主成分	.023	-.066	-.048	-.044
第五主成分	-.110	.017	.022	-.005
(Constant)	-38.401	-29.001	-24.587	-10.907

Fisher's linear discriminant functions

预测的分类结果

Classification Results[a,c]

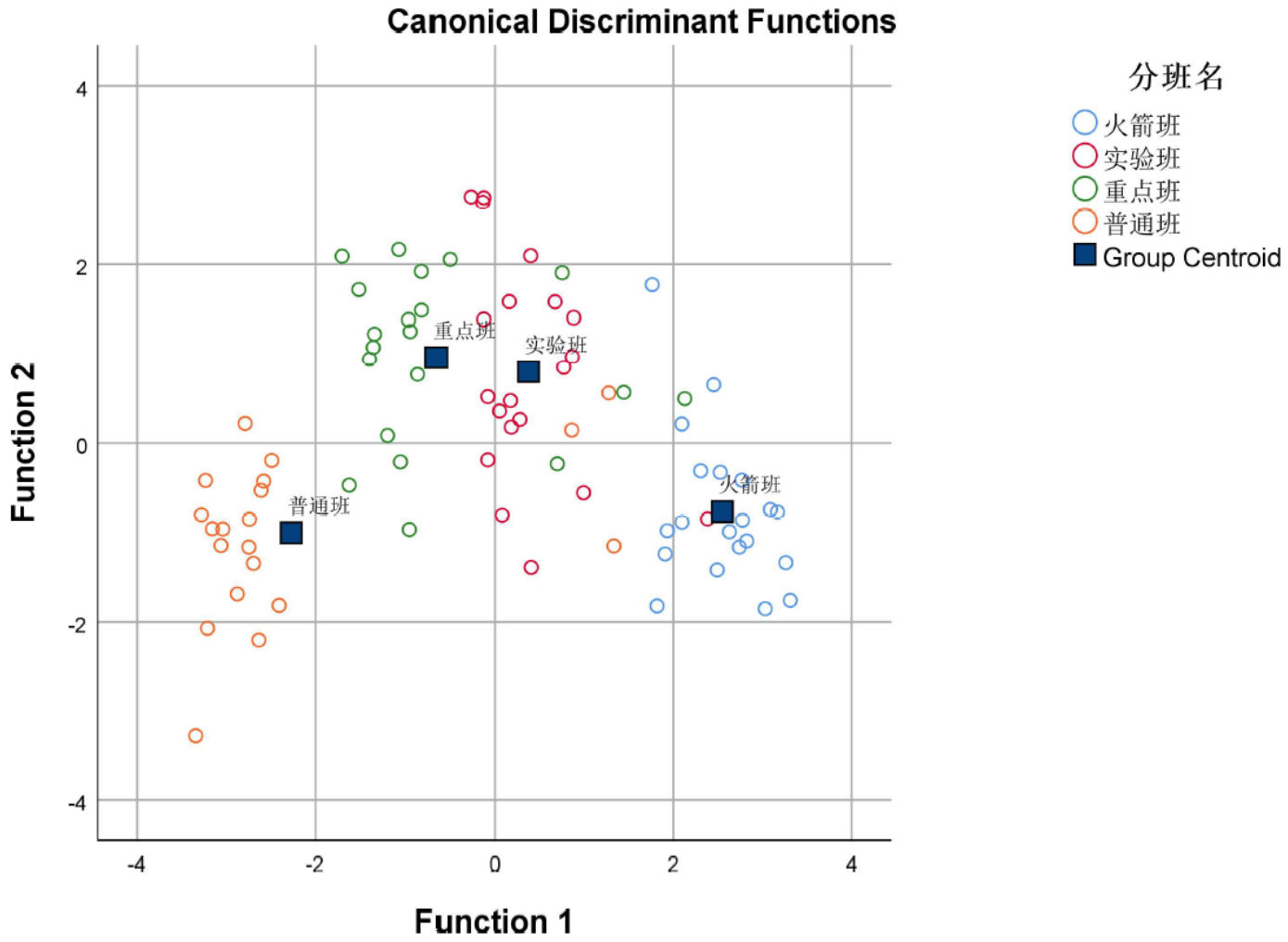
			Predicted Group Membership				
		分班名	火箭班	实验班	重点班	普通班	Total
Original	Count	火箭班	19	1	0	0	20
		实验班	2	16	2	0	20
		重点班	1	3	14	2	20
		普通班	1	2	0	17	20
	%	火箭班	95.0	5.0	.0	.0	100.0
		实验班	10.0	80.0	10.0	.0	100.0
		重点班	5.0	15.0	70.0	10.0	100.0
		普通班	5.0	10.0	.0	85.0	100.0
Cross-validated[b]	Count	火箭班	19	1	0	0	20
		实验班	3	14	3	0	20
		重点班	1	3	14	2	20
		普通班	1	2	0	17	20
	%	火箭班	95.0	5.0	.0	.0	100.0
		实验班	15.0	70.0	15.0	.0	100.0
		重点班	5.0	15.0	70.0	10.0	100.0
		普通班	5.0	10.0	.0	85.0	100.0

a. 82.5% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 80.0% of cross-validated grouped cases correctly classified.

预测的分类结果图


Canonical Discriminant Functions
Function 2
Function 1
4
2
0
-2
-4
-4
-2
0
2
4
重点班
实验班
普通班
火箭班
分班名
火箭班
实验班
重点班
普通班
Group Centroid

参考文献

何晓群.

多元统计分析 [M]. 2019.
北京: 中国人民大学出版社

白志东.

大维统计分析 [M]. 2012.
北京：高等教育出版社

VINCENT SPRUYT.

A geometric interpretation of the covariance matrix [EB/OL]. 2014.
Computer vision for dummies
https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix

多元分析方法使用报告

刘士坤　郑辉杨　薛兆浩　余涛

天津商业大学　统计系

2020 年 10 月 21 日