Statistics and Application 统计学与应用, 2018, 7(2), 263-270
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72031

Hans汉斯

The Application of Principal Component Analysis in the Analysis of Teaching Quality

Rong Hu[1,2], Zeyu Li[3], Weiyan Mu[1,2]

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Beijing Key Laboratory of Functional Materials for Building Structure and Environment Remediation, Beijing University of Civil Engineering and Architecture, Beijing
[3]Canvard College, Beijing Technology and Business University, Beijing
Email: hurong_hr@163.com

Received: Apr. 5th, 2018; accepted: Apr. 23rd, 2018; published: Apr. 30th, 2018

Abstract

The quality of the teaching of a school is often closely related to the students' academic performance. In order to study the teaching quality of several primary schools of Shaanxi, we made a principal component analysis of the average scores of 12 subjects in order to know the achievement of each school student [1]. The research results show that the quality of a school's teaching can be analyzed by student's academic record. In fact, only a few linear combinations consisting of the average scores of each subject can be considered, which can simplify the problem and improve the efficiency of analysis.

Keywords

Principal Component Analysis, Score of the Principal Component, Quality of Teaching

主成分分析在教学质量分析中的应用

胡　蓉[1,2]，李泽妤[3]，牟唯嫣[1,2]

[1]北京建筑大学理学院，北京
[2]北京建筑大学，建筑结构与环境修复功能材料北京市重点实验室，北京
[3]北京工商大学嘉华学院，北京
Email: hurong_hr@163.com

收稿日期：2018年4月5日；录用日期：2018年4月23日；发布日期：2018年4月30日

摘　要

一所学校教学质量的好坏，往往与学生的学习成绩息息相关。为了研究陕西某地区几所小学教学质量的

文章引用: 胡蓉, 李泽妤, 牟唯嫣. 主成分分析在教学质量分析中的应用[J]. 统计学与应用, 2018, 7(2): 263-270.
DOI: 10.12677/sa.2018.72031

好坏，我们对学生的12科平均成绩进行主成分分析，以便了解各个学校学生的学习成绩[1]。研究结果表明：由学生的学习成绩来分析一所学校的教学质量的好坏，实际上可以只考虑某几个由各科成绩的平均分组成的线性组合，可以简化问题，提高分析效率。

关键词

主成分分析，主成分得分，教学质量

1. 引言

教育是国之根本，社会的进步与经济发展都离不开教育。随着我国经济的迅速发展，国家对教育事业扩大投资，教育经费逐年上涨。近几年，我国将一些偏远山区的九年义务教育调整到十二年义务教育，并给农村的学生提供免费的餐宿，这足以显示我国对教育事业的重视。

国家对教育事业的投资如此之大，是希望能培育出将来报效祖国的人才，这和学校的老师就密切相关，与学校的教学质量密不可分。一般情况下，我们认为一所学校教学质量的好坏，往往根据升学率和学生的考试成绩来评价，这就需要研究学生的学习成绩，如果每一学科都考虑就会使得问题复杂化。而且，农村地区硬件设施较落后，也有可能影响学校的教学[2]，从而影响学生的学习成绩。本文通过多元统计中的主成分分析，以学生的 12 科平均考试成绩作为指标，对 9 所学校学生的学习成绩进行分析，其分析结果希望能对以后学校教学改革起到一定作用[3]。

2. 数据收集

由于农村学校一年级到二年级开设科目较少，所以，研究对象为陕西山区某镇的 9 所小学三到六年级的学生。在山区，有些学校偏远，学生人数少，为方便教学，将高年级的学生统一到一所学校，因此，有些学校没有五年级和六年级。

我们对该镇 2016~2017 学年度第二学期期末质量调研测试学生成绩进行分析，以 X_1——语文、X_2——数学、X_3——英语、X_4——思品、X_5——社科、X_6——美术、X_7——音乐、X_8——体育、X_9——劳动、X_{10}——舞蹈、X_{11}——法制安全、X_{12}——综合，这 12 科的测试平均成绩为指标，即变量数。以保安镇中心小学三年级(1)、四年级(2)、五年级(3)、六年级(4)，文峪小学三年级(5)、四年级(6)、五年级(7)、六年级(8)，瓦子坪小学三年级(9)、四年级(10)、五年级(11)、六年级(12)，乱石坪小学三年级(13)、四年级(14)，蒿坪小学三年级(15)、四年级(16)、五年级(17)，八道河小学三年级(18)、四年级(19)，西坝小学三年级(20)、四年级(21)、五年级(22)，涧底小学三年级(23)、四年级(24)、五年级(25)，眉底小学三年级(26)、四年级(27)、五年级(28)的平均成绩作为样本，即样品数。由于对原始数据进行标准化处理后，使得每个指标的作用在主成分的构成中相等，因此，本文所用的数据都是经过标准化处理后的数据。

3. 主成分分析

3.1. 主成分分析的基本思想

研究某一事物的整体情况时，我们需要对该事物的各个指标都进行考虑，这就使得问题变得复杂化。

而主成分分析正是解决这种复杂问题比较好的途径之一，该方法是研究如何通过原始变量的少数几个线性组合来解释原来变量绝大部分信息的一种多元统计方法，从而达到降维的目的[4]。本文采用 SPSS 软件和 R 语言对某镇 2016~2017 学年度第二学期期末质量调研测试学生成绩进行主成分分析。

3.2. 主成分分析的数学模型

对于一个样本资料，观测 p 个变量 $X_1, X_2, \cdots, X_p$，n 个样品的数据资料阵为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \left[X_1, X_2, \cdots, X_p \right], \; x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}, j = 1, 2, \cdots, p$$

主成分分析就是将 p 个观测变量综合成为 p 个新的变量(综合变量)，即

$$\begin{cases} F_1 = \gamma_{11} X_1 + \gamma_{12} X_2 + \cdots + \gamma_{1p} X_p \\ F_2 = \gamma_{21} X_1 + \gamma_{22} X_2 + \cdots + \gamma_{2p} X_p \\ \vdots \\ F_p = \gamma_{p1} X_1 + \gamma_{p2} X_2 + \cdots + \gamma_{pp} X_p \end{cases}$$

简写为：

$$F_j = \gamma_{j1} X_1 + \gamma_{j2} X_2 + \cdots + \gamma_{jp} X_p$$

且

$$Var\left(F_j\right) = \lambda_j$$

$$Cov\left(F_i, F_j\right) = 0$$

其中，λ_j 是样本资料阵 X 的协方差矩阵的特征根，由大到小排列，即 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$，$\gamma_j = \left(\gamma_{1j}, \gamma_{2j}, \cdots, \gamma_{pj}\right)'$ 为 λ_j 对应的标准正交特征向量。于是，称 F_1 为第一主成分，F_2 为第二主成分，依次类推，一共有 p 个主成分，主成分又叫主分量。

上述模型可用矩阵表示为

$$F = \gamma X$$

其中，

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{bmatrix}, \gamma = \left(\gamma_1, \gamma_2, \cdots, \gamma_p\right)$$

γ 为主成分系数矩阵。

主成分分析可以得到 p 个主成分，主成分的贡献率越大，说明该主成分所包含的原始变量的信息越多。但是，由于各个主成分的方差是递减的，包含的信息量也是递减的，所以实际分析时，一般不是选取 p 个主成分，而是根据累积贡献率选取前 k 个主成分，使得累积贡献率达到 85%以上，才能保证综合变量能包括原始变量的绝大多数信息，即

$$\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \geq 85\%$$

这里累积贡献率就是指前 k 个主成分的方差和占全部方差总和的比重[5]。

3.3. 主成分分析模型的应用

以为 $X_1, X_2, \cdots, X_{12}$ 这 12 个指标为变量，以上述每个学校中各个年级的平均成绩为样本，一共 28 个样本，用 SPSS 软件的因子分析模块做主成分分析[6]，输出结果如下。

表 1 显示了各主成分解释原始变量总方差的情况，前 6 个主成分的解释总方差为 85.477%，即这 6 个主成分集中了 12 个原始变量信息的 85.477%，因此保留前 6 个主成分。由图 1 也可以看出第六个和第七个特征根变化的趋势已经开始平缓，取前 6 个或前 7 个主成分都可以，但是为了更好地达到降维的目的，所以这里取前 6 个主成分。

表 2 的成分矩阵表示的是因子载荷矩阵而不是主成分系数矩阵，因此要对成分矩阵中第 j 列的每个元素分别除以第 j 个特征根的平方根 $\sqrt{\lambda_j}$ 就可以得到主成分分析的系数矩阵，即 $\gamma_{ji}=\frac{a_{ij}}{\sqrt{\lambda_j}}$，$a_{ij}$ 为成分矩阵中的元素，称为因子载荷，所得结果如表 3 所示，则主成分就可写成由原始变量表示的线性表达式：

$$\begin{aligned}F_1 &= 0.411417x_1 + 0.412534x_2 + 0.160213x_3 + 0.434863x_4 + 0.309819x_5 - 0.16245x_6\\ &\quad - 0.30982x_7 - 0.18868x_8 + 0.334381x_9 - 0.24227x_{10} - 0.12337x_{11} + 0.080385x_{12}\end{aligned}$$

$$\begin{aligned}F_2 &= 0.235393x_1 + 0.313423x_2 + 0.05007x_3 + 0.25425x_4 + 0.103391x_5 - 0.45648x_6\\ &\quad + 0.446076x_7 + 0.269206x_8 - 0.14761x_9 + 0.122248x_{10} + 0.439573x_{11} + 0.242546x_{12}\end{aligned}$$

$$\begin{aligned}F_3 &= -0.07099x_1 + 0.140462x_2 + 0.415279x_3 + 0.000763x_4 - 0.31833x_5 - 0.247335x_6\\ &\quad - 0.08779x_7 + 0.558795x_8 + 0.176341x_9 - 0.30459x_{10} - 0.17787x_{11} - 0.40536x_{12}\end{aligned}$$

$$\begin{aligned}F_4 &= 0.111978x_1 + 0.177146x_2 + 0.610374x_3 - 0.12483x_4 - 0.12391x_5 - 0.16246x_6\\ &\quad + 0.238643x_7 - 0.15695x_8 - 0.31758x_9 + 0.301975x_{10} - 0.44791x_{11} + 0.228546x_{12}\end{aligned}$$

$$\begin{aligned}F_5 &= -0.21108x_1 - 0.04242x_2 + 0.220397x_3 + 0.191425x_4 + 0.548406x_5 - 0.18004x_6\\ &\quad + 0.169696x_7 + 0.023799x_8 - 0.03208x_9 + 0.339391x_{10} + 0.112785x_{11} - 0.61359x_{12}\end{aligned}$$

$$\begin{aligned}F_6 &= 0.139163x_1 + 0.143512x_2 - 0.10655x_3 - 0.09459x_4 - 0.28159x_5 + 0.278327x_6\\ &\quad - 0.06849x_7 - 0.19026x_8 + 0.541432x_9 + 0.646892x_{10} - 0.05219x_{11} - 0.16743x_{12}\end{aligned}$$

各主成分的意义是由各线性组合中权数较大的几个指标的综合意义来确定，例如，由上述6个表达式可以看出第一主成分 F_1 中，X_1, X_2, X_4 这3个指标的系数较大，所以第一主成分主要就是 X_1, X_2, X_4 这3个指标的综合反映，第二主成分 F_2 主要是 X_7 和 X_{11} 的综合反映，第三主成分 F_3 主要是 X_8 这一个指标的反映。

我们将标准化后的原始数据代入上述 6 个表达式就可以计算出各样品的主成分得分，如表 4 所示，只取前两个主成分就能把样品的主成分得分在直角坐标系中描出来，然后可进行样品分类，由于计算量较大，主成分得分使用 R 软件来计算[7]，输出结果如图 2 所示。

得分的正负仅表示该样品与平均水平的关系，由于原始数据已经过标准化处理，所以这里的平均水平为 0。

Table 1. Interpretation of the total variance
表 1. 总方差解释

组件	初始特征值			提取载荷平方和		
	总计	方差百分比	累积%	总计	方差百分比	累积%
1	3.209	26.739	26.739	3.209	26.739	26.739
2	2.365	19.709	46.449	2.365	19.709	46.449
3	1.716	14.297	60.745	1.716	14.297	60.745
4	1.187	9.893	70.638	1.187	9.893	70.638
5	0.934	7.787	78.425	0.934	7.787	78.425
6	0.846	7.052	85.477	0.846	7.052	85.477
7	0.519	4.323	89.800	0.519	4.323	89.800
8	0.479	3.995	93.795	0.479	3.995	93.795
9	0.278	2.316	96.111	0.278	2.316	96.111
10	0.224	1.868	97.979	0.224	1.868	97.979
11	0.157	1.306	99.286	0.157	1.306	99.286
12	0.086	0.714	100.000	0.086	0.714	100.000

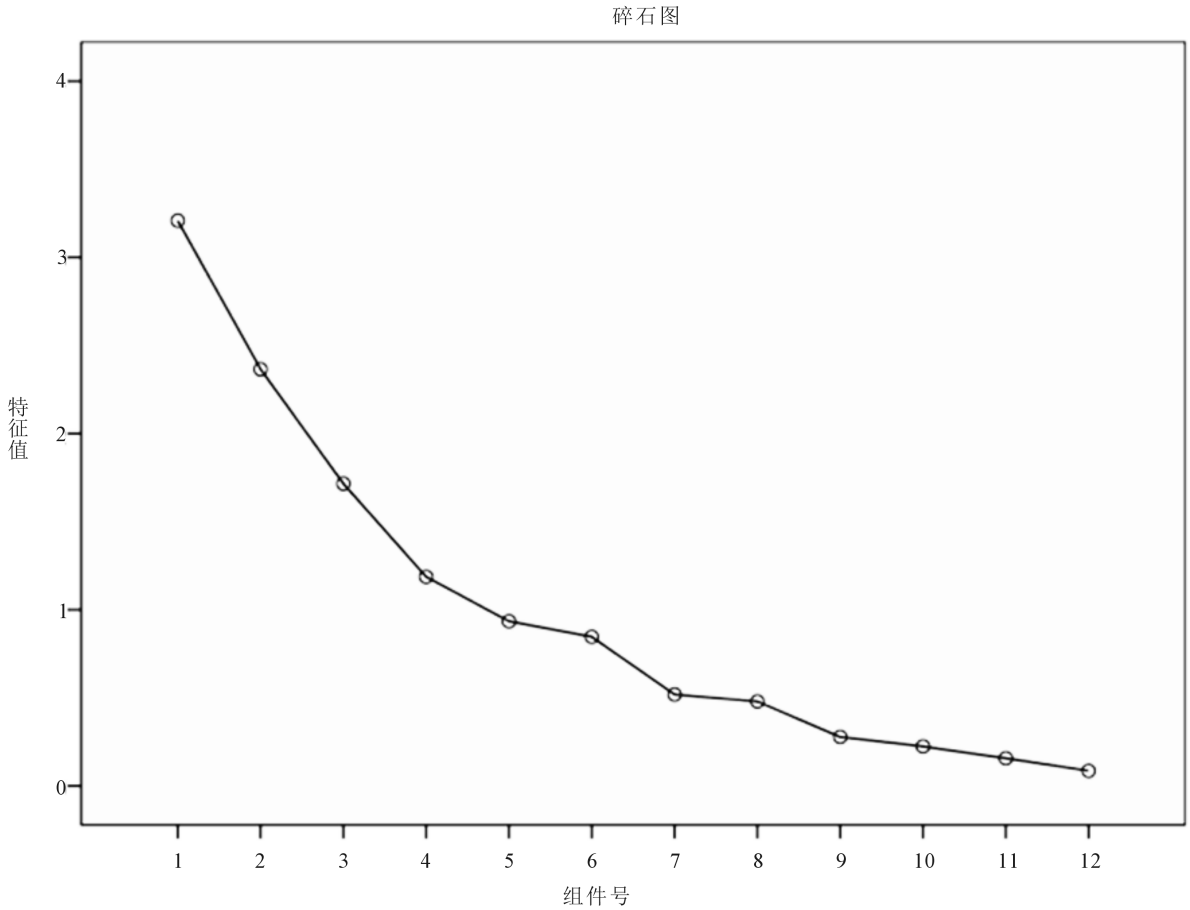

Figure 1. Scree plot
图 1. 碎石图

Table 2. Matrix of components
表 2. 成分矩阵

	组件											
	1	2	3	4	5	6	7	8	9	10	11	12
x_1	0.737	0.362	−0.093	0.122	−0.204	0.128	0.283	−0.340	−0.050	0.181	0.045	0.106
x_2	0.739	0.482	0.184	0.193	−0.041	0.132	−0.176	0.027	0.001	−0.160	−0.272	0.032
x_3	0.287	0.077	0.544	0.665	0.213	−0.098	0.223	0.165	−0.099	−0.116	0.135	−0.017
x_4	0.779	0.391	0.001	−0.136	0.185	−0.087	−0.262	−0.218	0.142	−0.037	0.154	−0.139
x_5	0.555	0.159	−0.417	−0.135	0.530	−0.259	−0.059	0.300	−0.056	0.146	−0.002	0.091
x_6	−0.291	0.702	0.324	−0.177	−0.174	0.256	−0.271	0.136	−0.284	0.086	0.099	0.011
x_7	−0.555	0.686	−0.115	0.260	0.164	−0.063	0.169	−0.015	0.024	0.198	−0.133	−0.150
x_8	−0.338	0.414	0.732	−0.171	0.023	−0.175	−0.002	0.044	0.316	0.071	0.013	0.106
x_9	0.599	−0.227	0.231	−0.346	−0.031	0.498	0.243	0.294	0.099	0.080	−0.008	−0.084
x_{10}	−0.434	0.188	−0.399	0.329	0.328	0.595	−0.094	−0.024	0.161	−0.053	0.070	0.068
x_{11}	−0.221	0.676	−0.233	−0.488	0.109	−0.048	0.341	−0.015	−0.049	−0.262	0.026	0.012
x_{12}	0.144	0.373	−0.531	0.249	−0.593	−0.154	−0.015	0.301	0.153	−0.024	0.077	−0.004

Table 3. The coefficient matrix of the principal component
表 3. 主成分系数矩阵

	F_1	F_2	F_3	F_4	F_5	F_6
X_1	0.411417	0.235393	−0.07099	0.111978	−0.21108	0.139163
X_2	0.412534	0.313423	0.140462	0.177146	−0.04242	0.143512
X_3	0.160213	0.05007	0.415279	0.610374	0.220397	−0.10655
X_4	0.434863	0.25425	0.000763	−0.12483	0.191425	−0.09459
X_5	0.309819	0.103391	−0.31833	−0.12391	0.548406	−0.28159
X_6	−0.16245	0.45648	0.247335	−0.16246	−0.18004	0.278327
X_7	−0.30982	0.446076	−0.08779	0.238643	0.169696	−0.06849
X_8	−0.18868	0.269206	0.558795	−0.15695	0.023799	−0.19026
X_9	0.334381	−0.14761	0.176341	−0.31758	−0.03208	0.541432
X_{10}	−0.24227	0.122248	−0.30459	0.301975	0.339391	0.646892
X_{11}	−0.12337	0.439573	−0.17787	−0.44791	0.112785	−0.05219
X_{12}	0.080385	0.242546	−0.40536	0.228546	−0.61359	−0.16743

4. 结果分析

本文通过主成分分析对 9 所学校中各个年级学生的学习成绩进行分析，主成分的意义是由线性组合中权重较大的几个指标的综合意义决定的，结果用前 6 个主成分进行研究。

由这 6 个主成分表达式可以看出第一主成分主要是语文、数学、思品这 3 个指标的综合反映，它反映了学生的基础学科和思想素质，第二主成分是音乐和法制安全这 2 个指标的综合反映，它代表了学生的艺术成绩和安全意识，这两个主成分反映了学生所受的主要教育，这与学校的教学质量息息相关[8]。

由图 2 可以看出分布在第一象限的是文峪小学五年级和六年级，瓦子坪小学四年级和六年级，乱石

Table 4. The score of the principal component
表 4. 主成分得分

	F_1	F_2	F_3	F_4	F_5	F_6
1	−1.35	2.031	0.526	−0.301	−0.365	0.951
2	−0.997	1.223	0.048	−0.096	−1.612	0.348
3	0.077	−0.096	−2.014	0.059	−1.19	1.082
4	−0.16	−0.246	−1.7	0.073	−0.125	−0.034
5	0.037	1.188	1.397	−0.687	0.336	1.996
6	0.726	0.036	−0.247	−0.133	−0.18	1.185
7	−0.183	−0.3	−1.845	0.67	−0.885	0.002
8	−0.011	−0.882	−0.371	0.143	2.512	0.725
9	−0.509	−1.486	0.419	−1.042	−1.444	−1.122
10	−2.451	1.426	0.414	0.935	0.171	−1.09
11	1.279	−0.006	−0.445	−1.135	−0.505	1.189
12	0.477	1.197	−0.417	−0.417	1.544	−0.393
13	1.28	−1.199	1.507	0.076	−0.877	−0.667
14	2.04	0.611	0.092	−0.476	0.39	−2.359
15	−0.145	0.961	−0.343	−0.599	1.325	−2.199
16	1.214	0.552	−0.729	1.5	1.221	0.469
17	−0.556	−0.739	0.209	−1.946	0.01	0.391
18	0.638	0.036	2.183	−0.209	−0.942	−0.515
19	−0.647	−1.093	0.715	0.833	0.367	0.093
20	0.188	0.682	1.543	0.723	0.562	0.367
21	−1.156	−1.833	0.283	−2.25	1.607	0.5
22	0.713	0.557	−0.416	−0.127	−0.708	−0.563
23	0.728	0.018	0.333	0.809	0.194	1.234
24	1.478	−0.465	−0.13	0.965	−0.334	0.216
25	−0.438	0.321	−1.371	−1.37	−0.764	−0.725
26	−0.203	−0.122	0.601	0.928	−0.936	−0.265
27	−1.208	−1.991	0.215	2.034	−0.158	−0.112
28	−0.862	−0.378	−0.455	1.04	0.786	−0.707

坪小学四年级，蒿坪小学四年级，八道河小学三年级和四年级，涧底小学四年级和五年级这 10 个年级，这说明这几所学校中的这些年级的教学质量相对较好，其中蒿坪小学四年级教学质量最好。分布在第四象限的是文峪小学四年级、蒿坪小学五年级、西坝小学五年级、眉底小学四年级共四个年级，由于第四象限的主要特征是第一主成分，而第一主成分占信息总量比重最大，所以，这四个年级的教学质量也相对较好。而分布在第二象限和第三象限的年级可以划分为同一类，教学质量相对较差。因此，总体来说，文峪小学高年级的教学质量相对最好，保安镇中心小学教学质量相对最差。

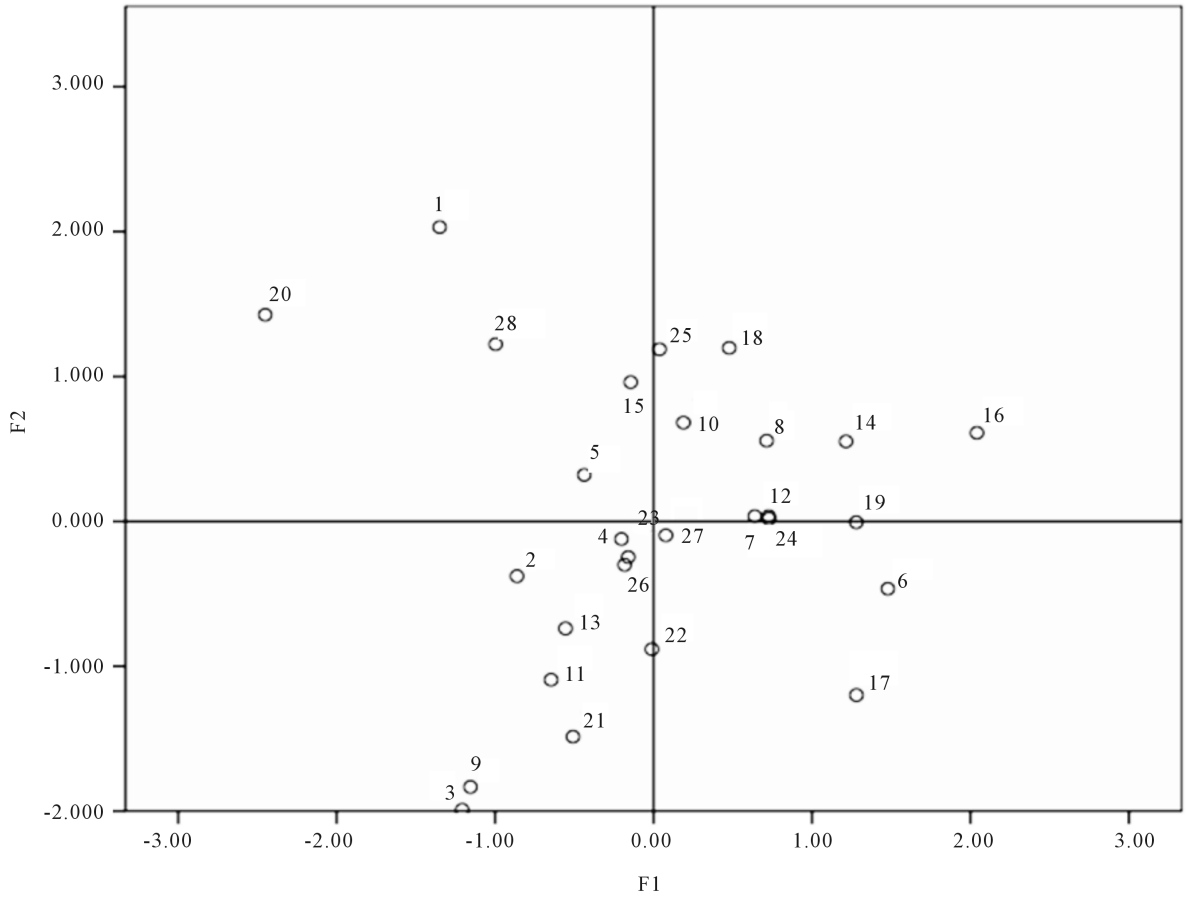

Figure 2. A scatter plot of the principal component score
图 2. 主成分得分散点图

5. 结论与建议

经过对上述几所学校中各个年级学生的成绩进行主成分分析得出，该地区文峪小学和瓦子坪小学高年级教学质量相对较好，但文峪小学低年级的教学质量也相对较差，保安镇中心小学的整体教学质量相对于其他学校差，希望学校及政府能适当改革教学方法，从基础抓起，提高学生的学习兴趣，为学生营造良好的学习氛围，以提高教学成绩。

参考文献

[1] 冷泳林, 王悦. 主成分分析法在学生评教中的应用研究[J]. 信息技术, 2012(12): 11-14.
[2] 韩际中, 缪青. 影响农村学校教学质量的客观因素[J]. 黑河教育, 2009(4): 20.
[3] 张晓. 主成分分析法在教学评价中的应用[J]. 伊犁师范学院学报, 2009(4): 45-47.
[4] 何晓群. 多元统计分析[M]. 第 4 版. 北京: 中国人民大学出版社, 2015: 113-141.
[5] 张尧庭, 方开泰. 多元分析引论[M]. 北京: 科学出版社, 1982.
[6] 陈胜可. SPSS 统计分析从入门到精通[M]. 第二版. 北京: 清华大学出版社, 2013: 300-315.
[7] Kabacoff, R.I. (2013) R in Action. Manning Publications, 300-315.
[8] 何陈. 略谈中小学教育质量评价[J]. 新课程学习: 上, 2015(13): 7.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org