BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXI No. 2/2009	33 - 38	Seria Matematică - Informatică - Fizică

Grades-based Characterization of Freshmen Using PCA

Cristian Marinoiu

Petroleum-Gas University, Ploieşti, Bd. Bucureşti, 39, Informatics Department
e-mail: cmarinoiu@upg-ploiesti.ro

Abstract

Informatics is one of the numerous specializations functioning within Petroleum-Gas University of Ploieşti that were founded during the last twenty years. Taking into account the prognosis of a continuous decrease of highschool graduates in the years to come, we consider that a better promotion of this specialisation is extremely important. This paper aims to draw a picture of the students in the first year attending Informatics, considering their marks at the disciplines included in the curriculum and the type of highschools they graduated from.

Key words: *PCA, principal components, factors, eigenvevector, eigenvalue*

Introduction

The Informatics specialization was founded in 2005, as a result of dividing the former double specialization Mathematics-Informatics functioning within Petroleum-Gas University of Ploiesti. Since then, there have been constantly registered two first year classes. The persons responsible for this specialization are interested in promoting it in order to maintain the already existing number of classes, and to increase it in the years to come by means of attracting a higher percentage of high school graduates that have proven skills in the field. A starting point in reaching these goals would be the identification and description of the structure obtained by analyzing the grades from the disciplines included in the first year curriculum and the type of high school these students graduated from.

The students that have passed all the exams included in the first year Informatics curriculum graduated mainly from the following four types of high schools: theoretical(1), technical(2), services(3) and humanistic(4).

The curriculum of Informatics specialization includes twelve disciplines, namely:

- Computer science disciplines - Procedural Programming1 (PP1), Procedural Programming2 (PP2), Computing System Architecture (CSA), Office Automation (OA), Operating Systems (OS), Algorithms and Data Structures (ADS);
- Mathematical disciplines - Computational and Mathematical Logic (CML), Mathematical Analysis (MA), Algebra (Alg), Probabilities and Statistics(PS);
- Philological disciplines - Foreign Language1 (FL1), Foreign Language2 (FL2).

In the following pages, by means of PCA (Principal Components Analysis)[1,2,3], we will characterize the students' results using only three artificial disciplines that will concentrate more than 70% of the information provided by the original disciplines.

Principal Components Analysis

We consider n available observations on p variables that are organized in a **X** (n x p) matrix. Each observation (line) may be geometrically interpreted as a point in the R^p space, and the matrix as a cloud of points in this space. When $p \geq 4$, the geometrical interpretation of the cloud is not possible. Therefore, we are interested in reducing p to values that are lower or equal to 3, still without affecting the structure of the points cloud and its variance. From a practical point of view, this implies the characterization of each observation by means of no more than 3 artificial variables, constructed from the initial p variables.

The PCA method allows finding p artificial variables that are called principal components or factors and have the following characteristics:

- each principal component is obtained as a linear combination of **X** matrix columns, in which the weigths are elements of an eigenvector to the data covariance matrix or to the correlation matrix, provided the data are centered and standardized.
- the first principal component is obtained by using as weigths the elements of the first eigenvector and it has the highest variance (equal to its first eigenvalue). Similarly, the second principal component is obtained by using as weights the elements of the second eigenvector and it has a variance equal to its second eigenvalue etc. The p th principal component has the lowest variance, that is equal to its p th eigenvalue.
- theprincipal components are orthogonal and uncorrerelated.

Lets consider $Y_1, Y_2, ..., Y_p$ the first p principal components and $Var(Y_i) = \lambda_i, i = 1,2,..., p$ - their corresponding variances, where $\lambda_i, i = 1,2,..., p$ are the data covariance matrix or the correlation matrix' eigenvalues, if the data are centered and standardized. The relation

$$r = (\sum_{i=1}^{q} \lambda_i / \sum_{i=1}^{p} \lambda_i) \times 100$$

represents the percent variation explained by the first q principal components. This may be considered as a measure of the quality of the approximation from the p initial variables, by means of the q retained components. Once the q number of retained principal components is decided, each component's interpretation is made in accordance with the degree of correlation to the initial variables.

The Representation and Analysis of Initial Variables Correlations

We used as initial data the grades of the 39 first year Informatics students (within PGU of Ploiesti) that passed all the exams, data that were organized in a table (see Table 1) in SAS[], from which we present only the first four rows. The table also contains a column specifying the student's current number (*crtno*) and another column indicating the types of the high school they graduated from (*his*). Further data may be downloaded from the site of PGU Bulletin, http://bulletin-mif.unde.ro/.

Table 1. Grades obtained by Informatics freshmen and the type of high school they graduated from

crtno	*his*	PP1	CSA	CML	MA	OA	FL1	PP2	OS	ADS	Alg	PS	PP2
1	1	8	10	5	5	10	8	8	7	7	2	4	9
2	1	8	9	5	5	10	8	8	8	5	2	4	8
3	1	10	10	5	5	9	8	10	10	10	5	6	8
4	1	6	7	5	6	8	8	6	9	7	3	6	8

In this case, **X** (*n* x *p*) matrix, previously considered, is formed only of the columns presenting grades, therefore $n = 39$ and $p = 12$. Each discipline (column respectively) represents one point in the 39-dimension space, whereas each student (row respectively) represents a point in a 12-dimension space. Our objective is to represent a student's results in a space having no more than 3 dimensions.

Table 2. The correlations of 1st year curriculum disciplines (with two rounded decimals)

	PP1	PP2	CSA	OA	OS	ADS	CML	MA	Alg	PS	FL1	FL2
PP1	1											
PP2	0.87	1										
CSA	0.56	0.67	1									
OA	0.48	0.54	0.45	1								
OS	0.66	0.77	0.63	0.65	1							
ADS	0.76	0.77	0.51	0.29	0.62	1						
CML	0.51	0.53	0.37	0.40	0.51	0.48	1					
MA	0.16	0.14	0.24	0.13	0.23	0.28	0.57	1				
Alg	0.57	0.66	0.60	0.26	0.63	0.71	0.73	0.43	1			
PS	0.56	0.62	0.47	0.34	0.63	0.66	0.76	0.45	0.84	1		
FL1	0.44	0.46	0.19	0.35	0.46	0.28	0.43	-0.16	0.40	0.47	1	
FL2	0.54	0.56	0.47	0.35	0.53	0.34	0.36	0.01	0.52	0.49	0.61	1

Table 2 contains the correlations of curriculum disciplines with two rounded decimals. If one considers the following evaluation scale: 0.00-0.20 (extremely weak correlation), 0.21-0.40 (low correlation), 0.41-0.60 (average correlation), 0.61-0.80 (strong correlation), 0.81-0.90 (very strong correlation) and 0.91-1.00 (an excellent correlation), the degree of correlation between the initial variables may be, generally, appreciated as highly satisfactory.

Extraction of the Principal Components

For the extraction of the principal components, we used the ***Factor*** procedure and the *Prin* method from SAS [2]. In order to comply with the terminology used in the above-mentioned procedure, we will use the term "factor" with the meaning of "principal component". Each of the 12 initial variables is automatically centered and standardized within the procedure, thus contributing with one unit to the total variance of the points cloud, that has, as a result, the value 12. When establishing the number of the retained factors, there were taken into account two criteria:

- the cumulated percent in variation explained by the retained factors should be higher than 70%;
- the variance (its own corresponding eigenvalue) of each retained factor should be higher than 1.

Table 3 presents the variances of the first four factors. One may notice that in order to fulfill these two criteria we should retain only the first three factors (the cumulated percentage is higher than 70%, each eigenvaleue is higher than 1.0).

Table 3. Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
Factor1	6.56583818	5.01626417	0.5472	0.5472
Factor2	1.54957402	0.52739137	0.1291	0.6763
Factor3	1.02218264	0.17264960	0.0852	0.7615
Factor4	0.84953305	0.25759628	0.0708	0.8323

In Table 4 there were inserted for each factor the so-called "factor loadings", 100 times multiplied and rounded to the nearest integer. They represent the coefficients with which the initial variables are loaded in each factor. As the changes made in this case are orthogonal, the factors loadings coincide here with the correlations between the retained factors and the initial variables. The values marked with * are higher than 50, a threshold value above which the corresponding factors loading are high and, therefore, significantly influential on the factor's value.

Table 4. Factor Pattern

	PP1	PP2	CSA	OA	OS	ADS	CML	MA	Alg	PS	FL1	FL2
Factor1	83*	89*	71*	60*	85*	79*	75*	36	85*	84*	57*	67*
Factor2	-17	-18	-2	-30	-14	13	38	81*	28	28	-48	-41
Factor3	-21	-24	-42	-19	-17	-24	34	0	16	28	58*	26

For a more accurate interpretation of factors, we operated an axis rotation (by using the *Varimax* option from the ***Factor*** procedure) that generated the values presented in table 5.

Table 5. Rotated Factor Pattern

	PP1	PP2	CSA	OA	OS	ADS	CML	MA	Alg	PS	FL1	FL2
Factor1	79*	85*	79*	63*	77*	70*	25	7	46	38	18	44
Factor2	24	25	23	3	29	44	81*	83*	72*	75*	11	12
Factor3	29	30	0	28	31	8	33	-32	30	38	92*	70*

One may notice the existence of some strong correlations between ***Factor1*** and Computer Science-related disciplines, between ***Factor2*** and mathematical disciplines, between ***Factor3*** and philological disciplines. As a consequence, factors 1, 2 and 3 reflect students' proficiency degree in the field of Informatics, mathematical or philological disciplines, respectively.

The Representation of Students on the Factorial Axes

As the field of knowledge of the students analyzed in this paper is Informatics, we are mainly interested in their relation to the first two factorial axes.

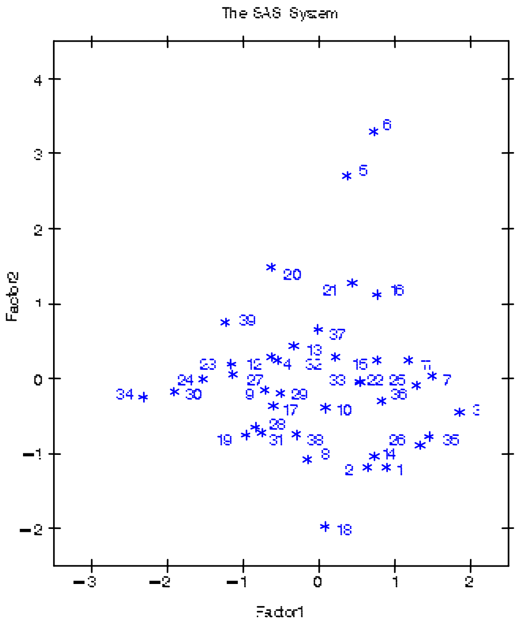

Fig. 1. The representation of students, related to the first two factorial axes

Figure 1 presents the representation of students' results related to the first two factorial axes. Considering positive correlations that were obtained in the relation with the two axes, a high score on the factorial axis number 1 or 2 represents a positive outcome for Informatics and mathematical disciplines, respectively.

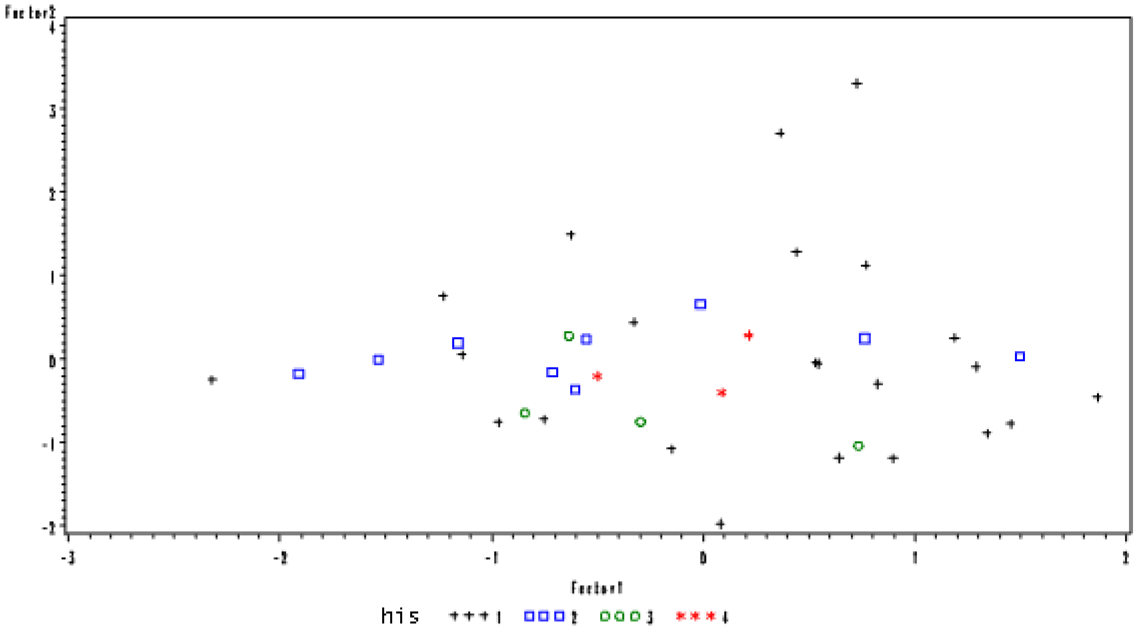

Fig. 2. The representation of students, in the relation to the first two factorial axes
(marks represent the high schools they graduated from)

In figure 2, students marked with a plus graduated from theoretical high schools, those represented by squares, from technical high schools, students marked with circles graduated from services high schools, whereas humanistics are marked by red stars.

Results and Conclusions

When analyzing figure 1, one may notice the following:

- most of the students obtained average grades at mathematical disciplines,
- excellent students at these disciplines (students number 5 and 6) have results above the average at Informatics disciplines, whereas poor students (student number 18) obtained average grades at Informatics.

When analyzing figure 2, one may ascertain the following:

- the students that graduated from theoretical high schools obtained the entire range of results: from unsatisfactory grades got at Informatics/ Mathematics-related disciplines to excellent results at the same disciplines,
- the students that graduated from technical high schools generally got average marks at Mathematics-related disciplines and the entire range of results at Informatics disciplines,
- the students that graduated from services and humanistic high schools generally got average and under-average marks, both at Informatics and Mathematics-related disciplines.

As a conclusion, Informatics specialization is to be further promoted both in theoretical high schools (that generally educate the tip-top students in the field of Informatics and Mathematics) and in other types of high schools (that not necessarily form the students with the poorest results).

References

1. Härdle, W., Simar, L. - *Applied Multivariate Statistical Analysis*, Springer-Verlag Berlin Heidelberg New York, 2003
2. Hastie, T., Tibshirani R., Friedman, J. - *The elements of Statistical learning: Data mining,,Inference and Prediction*, Springer-Verlag New York, 2001
3. Timm, N. - *Applied Mutivariate Analysis*, Springer-Verlag NewYork, 2004
4. * * * - *SAS 9.1.3 Documentation*, SAS Institute Inc, USA

Caracterizare bazată pe note a studenților din anul I folosind PCA

Rezumat

Specializarea Informatică este una din numeroasele specializări ale Universității Petrol –Gaze din Ploieşti înființate în ultimii 20 de ani. În perspectiva prognozatei scăderi a numărului de absolvenți de licee, va deveni tot mai necesară o mai bună promovare a acestei secții. Prezenta lucrare are ca scop caracterizarea studenților integralişti din anul I Informatică în raport cu performanțele obținute la disciplinele studiate şi cu profilul liceelor absolvite.