Performance of Universities in the United States of America

Heyang Dai

Abstract—**As for many students, it is indispensable for them to choose universities. However, analyzing performances of universities is never an easy work because performance of universities depends on many factors and some of them are highly correlated. In this paper, we first collected high dimensional data on various factors that can influence the performance of universities and then principal component analysis (PCA) was used to obtain uncorrelated variables and decrease dimensions of data. Finally, we provided full rankings of 30 famous universities in USA based on the score we obtained.**

Index Terms—**Principal component analysis, performance of universities, dimension reduction.**

I. Introduction

In the present age, attending universities becomes more and more crucial for most of the high school students all over the world, considering that they are capable of obtaining more professional courses, making numerous friends from different backgrounds and expanding their views about the whole nation as well as even the whole world in the college. In the meantime, as for a high school student, I face the same situation with other high school students and I consider entering an appropriate university can create a better atmosphere to do some preparations for my career. Absolutely, the very first step is to choose universities. However, choosing an appropriate university is always confusing, complicated and time-consuming for most students as well as their parents because different people have their own perspectives about the different universities. For instance, some people may prefer public universities because of lower tuition such as UC series, while others may consider expensive and qualified universities which can bring benefits than state or public universities, so people may prefer Harvard than UC series. Of course, not only can these personal opinions influence their decisions, but also many factors will affect their judges to choose universities as well such as locations, the ratio of students and professors, and the quality of dormitories. Take the ratio of students and professors in the class as an example. As for many universities, if the classes are composed of 17 or 20 members including students and one professor, it would be suitable and equal for each student to join the discussion in the lecture and express themselves. In addition, as for a professor, he can take care for every student to arrange the progress of teaching. If not, a series of problems will appear like the attendance of students, the balance of the whole class related to the GPA of students that not every student has the chance to answer questions, and so on. So this kind of factor will influence the judges of students to choose universities as well. What is more, it would be time-occupying for people to gather data from different universities. And even if we have all data, it is still arduous to analyze high dimensional correlated data. For example, the number of students leads to the arrangement of professors who teach students. Another example is that the location of the university will decide the daily cost of students, considering that different places with different levels of cost. Take students in California and NYC examples. Admittedly, when people consider these universities in these two places, their family needs to prepare some deposits for some students due to different and high price of commodities.

The good news is that there are some reputable rankings among universities from different organizations such as QS World University Rankings, U.S. News and TIMES. Here is an example from U.S. News; factors include graduation and first-year student retention rates, assessment by administrators at peer institutions, faculty resources, admissions selectivity, financial resources, alumni giving, graduation rate performance and, for National Universities and National Liberal Arts Colleges only, high school counselor ratings of colleges [1]. As U.S. News would like to choose these factors such as faculty resources and admissions selectivity, they may help people to clear how professional and sophisticated faculties are and make people realize through admissions selectivity to whether they should apply for the university. Another instance is Academic Ranking of World University, including the total number of the alumni of an institution winning Nobel Prizes and Fields Medals, the total number of the staff of an institution winning Nobel Prizes in Physics, Chemistry, Medicine and Economics and Fields Medal in Mathematics, the number of Highly Cited Researchers selected by Clarivate Analytics, the number of papers published in Nature and Science, total number of papers indexed in Science Citation Index-Expanded and Social Science Citation Index, the weighted scores of the above five indicators divided by the number of full-time equivalent academic staff [2]. These organizations not only have different factors and corresponding weights, but also use different methodologies, too. U.S. News is more inclined to focus on quality of assessment, student selectivity, faculty resources and research activity, while TIMES is more focused on the satisfaction of attended students, quality of research, standards of enrollment, the ratio of students and professors, cost of life for students, equipment and facilities, the ratio of excellent degrees students gained, the future of attended students and the ratio of graduation.

These rankings did help people to make some decisions; nevertheless, the bad news is different methodologies and

Manuscript received January 12, 2019; revised April 12, 2019.

Heyang Dai is with Chongqing Foreign Language School, Chongqing, China (e-mail: daidean0923@icloud.com).

doi: 10.18178/ijiet.2019.9.6.1235

different data sources lead to different results which make people even more confused. For instance, there exists a big gap about rankings about the same university, in USNEWS-GLOBAL in 2018, University of California Berkeley is the number four, while UCB in the U.S. News [3] is the number 21 in 2018. So there exist some differences from different organizations, which will make people feel confused about the same college with different rankings as well.

Motivated by the fact that there are multiple correlated variables influenced the performance of universities, we would like to utilize PCA to solve this problem. It is beneficial to transform correlated factors into uncorrelated factors to form new set of uncorrelated variables with less dimensions. In our project, we choose some factors such as Number of faculty, Number of students, ratio of students and professors, rejection rate , ratio of international student, R&D expenditures, Assets Under Management, salary of graduate students, students won Nobel, academic staff Nobel, Highly Cited Researchers, Nature and Science and salary of professors, which can represent precious aspects that people usually will consider, quoting from both from U.S. News and QS rankings. Then, we calculate the principal component scores based on the principal components and finally rankings the universities by their scores. In the next section, we provide an introduction to notations used in the paper and the procedures of PCA. After that, we bring the data we collected into the pragmatic process to analyze the performances of universities. Finally, we obtain a ranking of thirty famous universities in USA by their total principal component score.

II. Principal Component Analysis

A. Notations

The notations we used are defined as follows:

N: the number of observations

m: the number of random variables

X_i: random variable ($i \in \{1,2, \dots m\}$

$\boldsymbol{X} = (X_1, X_2, \dots, X_{\boldsymbol{m}})$: Random vector

S: the sample covariance matrix of $\boldsymbol{X}$ ($S \in M_m$)

λ_i : the i th largest eigenvalue of covariance matrix ($i \in \{1,2, \dots, m\}$

$\boldsymbol{a}_i$: the eigenvectors of covariance matrix corresponding to λ_i

Y_i: principal components which is linear combination of X_j ($j \in \{1,2, \dots m\}$)

k: the number of principal components

B. PCA Procedures

Principal component analysis (PCA) is a statistical method on multi-dimensional data invented by Karl Pearson in 1901, and it was independently developed and named by Harold Hotelling in the 1930s [4]. By change of basis to an orthonormal set of vectors, it can obtain linearly uncorrelated variables called principal components; each is a linear combination of previous correlated variables. Dimension reduction can also be achieved in this process by making the size of the orthonormal set k less than m. However, there is no free lunch in the world. The price of dimension reduction is information loss. The main task of PCA is to obtain lower-dimensional data with most information retained, in which case, we can make arduous problems become more easily analyzed and comprehensive for the public.

The procedures of PCA are as follows:

1) Normalize the data set

Normalized data Matrix Z can be obtained from raw data matrix R by

$$Z_{ij} = \frac{R_{ij} - \mu_j}{\sigma_j},$$

where z_{ij} is the ijth element of Z, x_{ij} is the jth term of ith observation from raw data, μ_j is the sample mean of $X_j$and σ_j is the sample standard deviation of X_j.

In PCA, the importance of one random variable depends only on the sample variance and PCA tends to provide more emphasis on variables with higher sample variance. Note that sample variance is a quantity that controlled by the unit of measurement defined on this variable. Therefore, in order to have equal weight on different variables in our analysis, it is necessary to use same measurement scale on different variables, which can be achieved by standardizing the raw data matrix or calculating sample correlation matrix instead of the sample covariance matrix in next step.

2) Calculate the sample covariance matrix from normalized data matrix Z

Sample covariance matrix S is a matrix where ijth element is defined as the sample covariance between $X_i andX_j$.

$$S_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(Z_{ki} - \mu_i)(Z_{kj} - \mu_j)$$

It is easy to show that sample covariance matrix is a symmetric matrix ($S_{ij} = S_{ji}$ for $i \neq j$), which is unitary diagonalizable and therefore there exists an orthonormal set of eigenvectors with size m.

3) Compute m eigenvalues and orthonormal set of eigenvectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m\}$ of sample covariance matrix S, where $\boldsymbol{a}_i$ is eigenvector corresponding to λ_i, eigenvalue

Foundation theorem of algebra makes sure that if $S \in M_m$, S must have m eigenvalues [5]. In addition, one of the properties of symmetric matrix is that they only have real eigenvalues. So we can have $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Furthermore, we can show that for any nonzero vector $\boldsymbol{x}$, $\boldsymbol{x}^T S \boldsymbol{x} \geq 0$, and therefore eigenvalues of S are always nonnegative.

4) Use $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_k$ as coefficient vector of k principal components $Y_1, Y_2, \dots, Y_k$ respectively, i.e.

$$Y_i = \boldsymbol{a}_i^T \boldsymbol{X}, i \in \{1,2, \dots k\},$$

where k is the smallest integer such that

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_m} \geq \alpha, 1 \leq k \leq m \qquad (2.1)$$

We reduce dimension from m to k by choosing $k \leq m$ satisfying inequality (2.1). It means that these k principal components can represent m variables with information loss less than or equal to $(1-\alpha)$%. Sometimes, we would like to make 85% to 90% as for our standard for α to make judgments about the performances of some objects. Note that we can let $k=m$ so that there is no information loss. But in this case, there is no dimension reduction because the number of principal components equals the number of variables. So this kind of step is no function to keep on the following steps.

It is easy to show that for any two random variables U and V satisfying $W=\boldsymbol{b}^T\boldsymbol{X}$ and $V=\boldsymbol{c}^T\boldsymbol{X}$, where $\boldsymbol{b}$ and $\boldsymbol{c}$ are coefficient vectors, we have

$$Var(U)=\boldsymbol{b}^T S\boldsymbol{b},$$

$$Var(V)=\boldsymbol{c}^T S\boldsymbol{c},$$

$$Cov(U,V)=\boldsymbol{b}^T S\boldsymbol{c}$$

From above results, we can see that utilizing orthonormal set of eigenvectors as coefficient vectors of principal components can make sure that any pair of principal components has covariance 0 (linearly uncorrelated). In addition, for any $1\le i\le k$,

λ_i is the variance of $Y_i=\boldsymbol{a}_i^T\boldsymbol{X}, i\in\{1,2,\dots k\}$. PCA uses variance to measure the amount of information of one random variable. The reason for choosing eigenvectors corresponding to first k largest eigenvalues as coefficient vectors of principal components is that

$$\max_{||\boldsymbol{a}||=1,\boldsymbol{a}\perp\boldsymbol{a}_1,\dots,\boldsymbol{a}_{k-1}} Var(\boldsymbol{a}^T\boldsymbol{X})=\boldsymbol{a}^T S\boldsymbol{a}=\lambda_k$$

5) Obtain new data matrix D where for any $i\in\{1,2,\dots N\}$ and $j\in\{1,2,\dots k\}$

$$D_{ij}=\sum_{l=1}^{m} z_{il}\,a_{jl}$$

The new data set has N observations and k variables and D_{ij} is called j th prinpical component score of i th observations.

6) Compute total score for each observation $i\in\{1,2,\dots,N\}$

$$F_i=\sum_{l=1}^{k}\frac{\lambda_l}{\lambda_1+\lambda_2+\cdots+\lambda_k}D_{il}$$

F_i is the weighted sum of k principal components scores for observation i. The weight for principal component score is percentage of eigenvalue as well percentage of variance.

III. Analysis of Performance of Universities Utilizing PCA

A. Factors Description

Concerning about problems on choosing universities in the United States of America, for as many people, they are still confused about the rankings, factors and methodologies used for universities from different organizations. As a result, we choose thirteen factors to measure the performance of universities.

- Number of faculty

Academic staffs are in universities including professors, lecturers and researchers. The number of faculty can show the academic capacities about the university.

- Number of students

A student is primarily a person enrolled in a school who attends classes in a course to attain the appropriate level of mastery of a subject under the guidance of an instructor. The number of students can illustrate the scale, financial funds for the daily activity in school of the university

- Ratio of students and professors

The number of students compares with that of professors in the university. This factor can help applicators decide whether they are willing to have “enlarged class” or “small class”.

- Rejection rate

In manufacturing, the rejection rate is the percentage of applicators who are refused. The rate of it can denote as %, percentage, to show the difficulty of applying the university for applicators.

- Ratio of international student

International students can be regarded as students who have crosses a national or territorial border for the purpose of education and now enrolled outside their country of origin. Ratio of it can show the tolerance of university, which means if the university comprised of different nations’ students, the university is more like a smelter that can tolerate different cultures and beliefs.

- R&D expenditures

R&D expenditures is Research and Development expenditures, which refers to the money provided to the innovative activities undertaken by corporations or governments in developing new services or products in university. This factor can attract more academic applicators to join in.

- Assets Under Management

It measures the total market value of all the financial assets which a financial institution.

- Salary of graduate students

The average money gained by graduate students in the work can illustrate capacity of graduate students in different fields, which can lure some undergraduate students enter university.

- Students won Nobel

Novel Prize is a set of six annual international awards bestowed in several categories by Swedish and Norwegian institutions in recognition of academic, cultural, or scientific advances. So this factor can demonstrate students have special ideas, capacity and so on to gain this kind of valuable award.

- Academic staff Nobel

That faculty can gain Nobel can illustrate the capacity of academic staff, leading to the precious fame comes to the university.

- Highly Cited Researchers

Researchers are mentioned by famous magazines or some distinguished organizations. This factor can illustrate the fame and consideration of researchers in the university.

- Nature and Science

The people quoted from there the organizations can show the academic abilities among the whole university no matter who are students or professors.

- Salary of professors

The money gained by professors can represent the skills, teaching experiences, teaching capacities of professors. This factor can help students according to the interests to choose sophisticated professors.

The normalization process won't change the size of our data set. So the size of normalized data set is still 13 variables with 30 observations, and therefore, the size of sample covariance matrix is 13 × 13.

TABLE I: EIGENVALUES OF COVARIANCE MATRIX

	λ_i	$\frac{\lambda_i}{\sum_{j=1}^{13} \lambda_j}$	$\frac{\sum_{k=1}^{i} \lambda_k}{\sum_{j=1}^{13} \lambda_j}$
$i = 1$	6.4775	49.83%	49.83%
$i = 2$	2.3782	18.29%	68.12%
$i = 3$	1.1878	9.14%	77.26%
$i = 4$	0.9228	7.10%	84.36%
$i = 5$	0.4902	3.77%	88.13%
$i = 6$	0.4757	3.66%	91.79%
$i = 7$	0.4043	3.11%	94.90%
$i = 8$	0.2200	1.69%	96.59%
$i = 9$	0.1669	1.28%	97.87%
$i = 10$	0.1200	0.92%	98.79%
$i = 11$	0.0833	0.64%	99.43%
$i = 12$	0.0553	0.43%	99.86%
$i = 13$	0.0179	0.14%	100%

The sum of all eigenvalues of our covariance matrix, which also equals the trace of covariance matrix, is 13. From Table I, the largest eigenvalue is nearly half of trace. It means that the first principal component, which utilize eigenvector corresponding to largest eigenvalue as coefficient vector, can represent about 50% of total information. Moreover, we can reduce dimension from 13 to 3 while still have about 80% of total information. If we would like to have more than 90% ($\alpha = 90\%$) of information, we need to utilize at least 6 principal components. The scree plot (Fig. 1) also illustrates that the first four principal components can explain most information (variance). Combining Table I and Fig. 1, $k = 4$ is a reasonable choice.

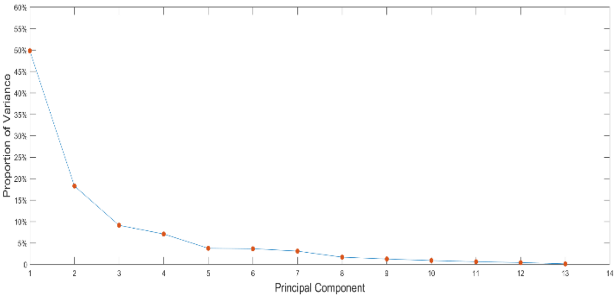

Fig. 1. Scree Plot.

Each principal component is a linear combination of previous variables. Table II and Table III list coefficient of each factors for first two principal components. (Restricted by the size of this paper, we didn't provide coefficients of all four principal components). From Table II, Rejection Rate, Average Salary of Graduates, Nobel Winners from Students, Nobel Winners from Faculties, Highly Cited Researchers, Nature and Science Index, and Average Salary of Professors have large positive coefficient. It shows that this principal component emphasis on the achievement of students and quality of faculties. Table III shows that the second principal component assigns largest weights to Number of Students and Ratio between Professors and Students. It would prefer larger universities and universities with better professor-to-student ratio.

TABLE II: COEFFICIENT OF 1ST PC

Factors	Coefficient
Faculty Members	0.2588
Students	0.0415
Ratio	0.2619
Rejection Rate	0.3084
Rate of International Students	−0.0070
R&D Expenditures	0.1665
Assets Under Management	0.2485
Average Salary of Graduates	0.3206
Nobel Winners from Students	0.3210
Nobel Winners from Faculties	0.3398
Highly Cited Researchers	0.3291
Nature and Science Index	0.3595
Average Salary of Professors	0.3454

When coefficient vector of each principal component is available, principal component score can be obtained by vector multiplication. Table IV lists the score of first principal component and second principal component of some universities. From Table IV, as well score plot (Fig. 2), Harvard University has largest 1st principal component score. This is because of its high rejection rate (94%), most Nobel winners, and most highly cited researchers among our 30 universities. Note that Harvard University has negative score on the second principal component, which is due to smaller teacher-to-student ratio compared to those of top universities, and most Nobel winners (2nd PC has negative coefficient on number of Nobel winners). Contributed by largest R&D expenditures among all universities, the Johns Hopkins University has both scores in top level, in particular 2nd PC score. And as for Stanford University, from Table II and Table IV, it also has both scores in top level as well and gains number two in 1st PC score since it has the highest rejection rate (95%) and the highest ratio between students and professors. However, it is the number four in the 2nd PC score, considering that it has fewer students who won Nobel and from Table III, the proportion to Nobel Winners from Students is comparative lower than other principal component.

TABLE III: Coefficient of 2nd PC

Factors	Coefficient
Faculty Members	0.2412
Students	0.3272
Ratio	0.3344
Rejection Rate	0.2689
Rate of International Students	−0.6364
R&D Expenditures	0.1858
Assets Under Management	−0.1020
Average Salary of Graduates	0.0563
Nobel Winners from Students	−0.3429
Nobel Winners from Faculties	−0.1766
Highly Cited Researchers	−0.2239
Nature and Science Index	−0.1166
Average Salary of Professors	0.1064

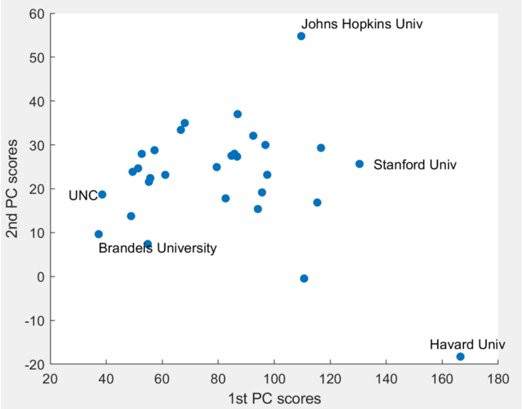

Fig. 2. Score plot.

TABLE IV: Principal Component Score

University	1st PC Score	2nd PC Score
Princeton Univ	82.6334	17.7993
Harvard Univ	166.5071	-18.2808
Univ of Chicago	94.1595	15.3798
Yale Univ	97.5010	23.1874
Columbia Univ	115.3783	16.8613
Stanford Univ	130.4639	25.6626
MIT	116.7005	29.3193
Duke Univ	86.9150	37.0086
Univ of Penn	92.4907	32.0800
Johns Hopkins Univ	109.6428	54.7987

TABLE V: Total Score

University	Total Score
Harvard University	93.8479
Stanford University	81.7622
MIT	81.4484
Princeton University	75.3528
Yale University	73.1882
Johns Hopkins University	72.2267
University of Chicago	64.8095
Columbia University	64.5431
Rice University	62.8818
Dartmouth College	62.7265

The total score can be obtained by weighted sum of each principal component scores and the weights are determined by corresponded eigenvalues. Table V shows top 10 universities based on total score. The ranking of universities in the USA is different from the conventional degrees showed in the U.S. News or other organizations. It is obvious to see that Harvard University is the number one among the ten universities and gets over ninety percent scores, which there exists a quite big gap between the number two, Stanford University, due to different scores gained in the principal component. From our rankings about universities, some outstanding and ambitious students can try them best to grab the opportunity to enter Harvard University. And then two universities, Stanford University as well as MIT exceed than eighty percent, which are precious and valuable to refer as well, having a better comprehensive and strong academic aspects for students to select. In addition, the score of three universities is more than seventy percent in our rankings, which are close to each other so that qualified students can make some rational decisions corresponding with interests. Finally, the number of University of Chicago, Columbia University, Rice University and Dartmouth College is more than sixty percent, comparing the previous universities maybe showing a little bit low but it can be options for some students who want to get high education resources regarding top 5 universities as goals.

IV. Conclusions

The article provides people with the new rankings about the performance of university in the USA, applying PCA mentioned in the introduction to solve the ranking issues. We explain each of the principal component with different purposes, furthering people to accept our rankings. Of course, our rankings are not the perfect version since we only collect 13 factors which are precious for us to consider and if following people want to get more precise results, they can find more decent, valuable and necessary factors to be incorporated. In addition, there exist other methodologies to solve the ranking problems, and people can still use and try other methodologies to resolve the same problems. So people can have more options to refer to the universities in the USA.

And as for a high school student, during the project, I lack of many academic mathematical knowledge, which promotes and pushes to me to control more knowledge. At first, I always confuses the concepts about different mathematical terms, analyzing data and writing paper; however, after leading by teacher Meng, I enhances a lot from coding the calculation in using PCA to analyze different principal component through knowledge, which is great development for me. Although this is my first try to write paper, it is meaningful and memorable.

Acknowledgments

I am grateful to U.S. News and QS world universities rankings provided various factors for us to what refer to. And thanks to the ranking given by TIMES, we employ it to compare the weights with other rankings such as U.S. News and QS world universities rankings on universities. I would like to say thank you to MATLAB for its free student version, which provides me great help in data virtualization and data analysis.

References

[1] M. Robert, B. Eric, and M. Matt. (2017). *How U.S. News Calculated the 2018 Best Colleges Rankings*. [Online]. Available: https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings

[2] *ARWU Methodology 2017*. [Online]. Available: http://www.shanghairanking.com/ARWU-Methodology-2017.html

[3] *U.S. News National University Rankings 2017*. [Online]. Available: https://www.usnews.com/best-colleges/rankings/national-universities

[4] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[5] B. Fine and G. Rosenberger, *The Fundamental Theorem of Algebra*, Springer Science & Business Media, 2012.

Heyang Dai is studying in Chongqing Foreign Language School in Chongqing, China. He was born on September 23, 2000, and he is going to go abroad to seek for knowledge in the USA and major in mathematics and computer Science.

He has participated in various activities such as Model United Nations in Philadelphia, acting different characters in the stage performances, and national photography competition.

Mr. Dai is still working to develop himself to enter a professional field.