College Desirability: A Multivariate Statistical Analysis

Andrea M Austin
St. Michael's College
Colchester, VT 05439

Terrell A Felder
North Carolina A & T St U
Greensboro, NC 27401

Lindsay M Moomaw
Baldwin-Wallace College
Berea, OH 44017

Abstract

The colleges and universities across the United States are all unique. To quantify how institutions of all sizes measure up, multivariate techniques of Principal Component Analysis, Factor Analysis, and Discriminant Analysis are used fittingly and effectively, producing a valid, unbiased evaluation of each school, and also a model to gauge any chosen seminary. The method of Principal Components reduces the number of variables, focusing on those with efficacy while Factor Analysis provides a data reduction to explain the variability of the college or university statistics. Finally, a Discriminant Analysis of the data classifies the schools and establishes a method of accurate prediction.

Introduction

As students prepare for graduation from high school, they are faced with uncertainties about their future. Of high school graduates in the United States, 65.8% answer one uncertainty by continuing their education at a college or university [BLS07]. A common quandary for these students is deciding which schools to choose. Every year, analyses of how colleges and universities rank are produced by the *Princeton Review* and by the *US News and World Report*. These are just two of many publications available to students to aid them in their decision-making process. When assessing these rankings, a question of objectivity can be raised. Many are slightly skewed, often reflecting the opinions of those involved, rather than just the facts. According to an article in TIME magazine,

> "*U.S. News* has been grading colleges and universities since 1983, and while the magazine mostly uses hard data, the largest single component of the rankings — 25% of a school's overall score — comes from a survey that asks presidents, provosts and admissions directors to assess peer institutions [JR07]."

Through statistical analysis, we use objective classifications of colleges and universities (as seen in Table I of the appendix) to show how multivariate techniques can be applied to analyze data. The variables we consider range greatly; including financial, social, athletics, and academics, shown in Table 1. To begin, we assess the variables for normality, using appropriate techniques to normalize where required. Later, by employing Principal Component Analysis, we are able to simplify the process of interpreting and summarizing the numerous variables, ultimately reducing their dimensionality. Using Factor Analysis, we further reduce dimensionality by examining the underlying commonality between variables. Finally, Discriminant Analysis allows for the separation and classification of schools into categories of desirability.

Normality Assessment

I. Theoretical Background

In order to analyze the data, it is required that each of the p variables come from a multivariate normal distribution. However, some variables may deviate from this desired distribution, so to assess for normality, a *Quantile-Quantile* (*Q-Q*) plot is constructed. A *Q-Q* plot measures the sample data versus the expected standardized observations based on a normal distribution. To construct the *Q-Q* plot, we order the n observations $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ for each variable X_i, for i=1, 2,…, p. The corresponding standardized quantiles z_j are then defined such that $P(Z < z_j) = \frac{j}{n}$, where Z~N(0,1) and j=1, 2,…, n. We plot the ordered data and corresponding normal quantiles as ordered pairs $(z_j, x_{(j)})$, thus producing the *Q-Q* plot. If the data is normally distributed, the plotted pairs will display positively correlated linear behavior based on a visual analysis of the plot. If the *Q-Q* plot does not exhibit linearity, hence normality, the sample data must be transformed to achieve normality. To substantiate our claim statistically, the linearity is also measured by calculating the sample correlation coefficient r, given by

$$r = \frac{\sum_{j=1}^{n}\left(x_{(j)} - \bar{x}\right)\left(z_j - \bar{z}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{(j)} - \bar{x}\right)^2}\sqrt{\sum_{j=1}^{n}\left(z_j - \bar{z}\right)^2}} = \frac{\hat{Cov}(z,x)}{\sqrt{\hat{Var}(z)\hat{Var}(x)}}.$$

A strong linear relationship is indicated by a value of r close to 1. In hypothesis testing, the null hypothesis (H_o) is ρ=1 and the alternative hypothesis (H_a) is $0<\rho<1$, where ρ is the population correlation coefficient. The sample size and desired α-level of significance determines the critical value c. The null hypothesis is accepted if $r \geq c$, which implies normality.

II. Analysis

For the data set under consideration, we use a sample size of n=50 colleges, with p=23 variables, and a significance level of $\alpha = 0.05$. These values result in a critical value of c=0.9768, found in Table 4.2 in [JW07]. Of the twenty-three variables, twenty are rejected under H_0 because their r value is less than 0.9768. To achieve normality for these variables, transformations are used. The non-transformed r values can be seen in Table 1. The four variables with a normal distribution are student faculty ratios, the number of intercollegiate sports, SAT scores, and average graduation debt.

The remaining nineteen variables are transformed; seven variables are transformed by taking their square root, six by raising them to a positive integer power, three by taking their log, and the remaining three by using a combination of square roots and logarithms. These changes are also shown in Table 1.

Table 1: Transformed *r* Values

Variable	Original R-Value	Transformation	Transformed R-Value
Miles From Major City (250,000+)	0.934	Square Root	0.992
% of Classes with <20 Students	0.965	x^2	0.969
Student-Faculty Ratio	0.988		0.988
% Greek Members	0.946	Square Root	0.993
Number of Intercollegiate Sports	0.993		0.993
Number of Intramural Sports	0.975	Square Root	0.985
Bowl Appearances	0.923	Square Root	0.988
Final Four Appearances	0.863	Square Root(Log(x + 1))	0.994
Overall NCAA Championships	0.788	Log(Square Root + 1)	0.984
Average SAT Score Accepted	0.984		0.984
National Merit Scholars	0.901	Log(Square Root + 1)	0.979
% of Freshmen in Top 10 % of Class	0.968	x^2	0.973
Acceptance Rate	0.969	Square Root	0.984
Faculty Awards	0.943	Square Root	0.992
Faculty in National Academy	0.818	Log(x+ 1)	0.989
Total Cost (Room, Board, Tuition)	0.907	x^3	0.932
Financial Aid	0.944	x^3	0.954
Graduation Debt	0.987		0.987
Retention Rate (Returning Sophomores)	0.952	x^7	0.981
Graduation Rate	0.927	x^4	0.963
Yield (% Attending of total Accepted)	0.956	Log	0.983
Endowment per Student	0.860	Square Root	0.962
Number of Companies Recruiting	0.642	Log(x+1)	0.942

As shown in Table 1, the variables with an *r* value less than 0.9768 are transformed in order to try to produce normal distributions. Notice that seven of our transformed variables, for example the number of companies recruiting on campus, do not meet the criteria of the null hypothesis. While the *r* values are still below the critical value, they are not significantly below, thus, a slightly smaller alpha value would have allowed for their acceptance. Despite the fact that these seven transformed variables are not quite normal, we continue to use the variables, as they have significance in the overall analysis of the colleges and universities. The variable Faculty in the National Academy is eliminated for the remainder of the analysis because it is highly correlated with Faculty Awards.

To demonstrate the linearity of the plots of the transformed variables, consider Figures 1 and 2. Figure 1 is the *Q-Q* plot of the original data for the number of companies recruiting on campus (this variable is representative of the other variables), while Figure 2 is the *Q-Q* plot of the normalized data for this same variable. Clearly Figure 2 displays greater linearity than Figure 1.

Figure 1

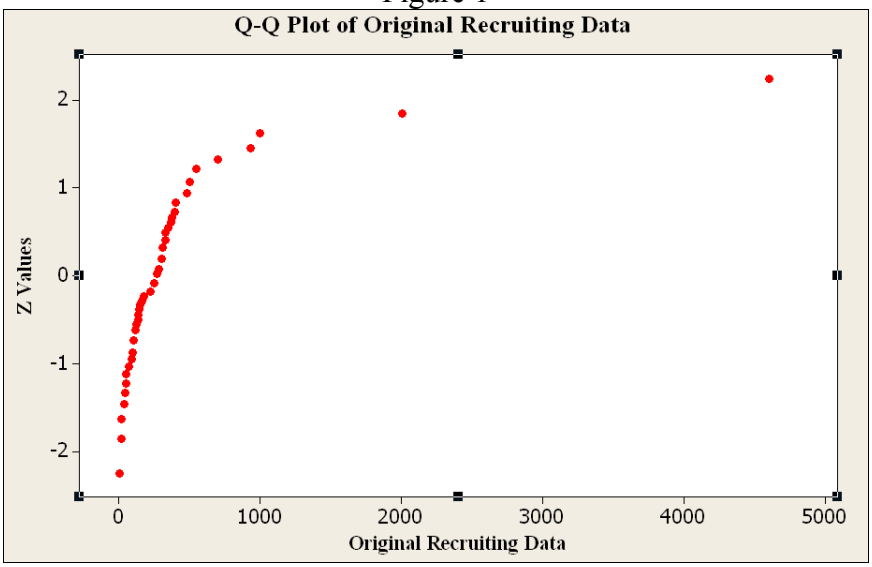
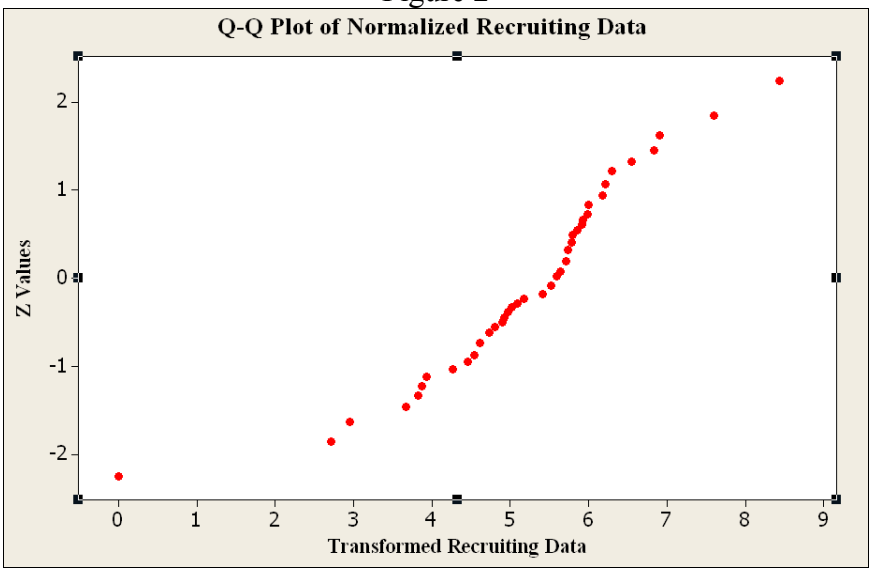


Figure 2



In the next section, we continue the statistical analysis with the transformed and normalized data, using Principal Component Analysis to reduce variable dimensionality.

Principal Components Analysis

I. Theoretical Background

Principal Component Analysis (PCA) is a method used to reduce the dimensionality of variables. The application of PCA transforms the p variables into *principal components* which are linear combinations of $X_1, X_2, \ldots, X_p$. These principal components represent a new coordinate system obtained by rotating the original variable axes. By grouping together similar variables, it is possible to reduce the number of variables analyzed from our original selection, to just a few principal components.

Ultimately, the goal is to find as few principal components as possible, to account for the largest portion of the total sample variance.

Principal components rely upon the covariance matrix **Σ.** As we have normalized $X_1, X_2, \ldots, X_p$, inferences can be made using the principal components. Let **X** be a multivariate normal random vector with mean μ and covariance matrix **Σ**. We want linear combinations of the p random variables that represent the rotated coordinate system in the direction that maximizes the variability of the data set. Let

$$Y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = \mathbf{a}_1^T\mathbf{X}$$
$$Y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p = \mathbf{a}_2^T\mathbf{X}$$
$$\vdots$$
$$Y_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p = \mathbf{a}_p^T\mathbf{X}.$$

We want to choose a vector **a** of coefficients such that it accounts for the maximum proportion of variance possible. The principal components are the uncorrelated linear combinations $Y_1, Y_2, \ldots Y_p$. In practice we extract m principal components ($m<p$). The proportion of variance contributed by the m principal components is given by

$$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j},$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ are the characteristic roots, or eigenvalues, of the covariance matrix Σ. The number of principal components needed to maximize the variance can be obtained through a visual analysis of the scree plot. The scree plot is constructed, to a desired degree, by plotting the index j=1,2,…,p against the corresponding λ_j's. Each point on the graph represents the amount of the total sample variance contributed by each eigenvalue. The number of components is determined by where the elbow occurs on the scree plot.

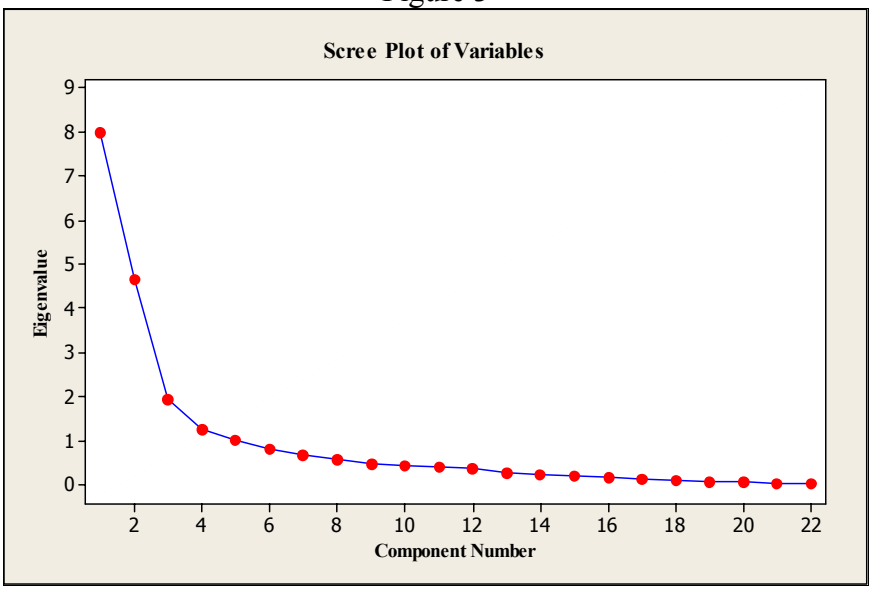
II. Analysis

Using PCA on MINITAB, we are able to analyze each component individually, in terms of the linear combinations provided, as well as the variance that each component contributes. Most importantly, Minitab is able to present the data analysis necessary without any computational error.

There are two ways to compute principal components: one, using the covariance of the variables, and the other, using the correlation among them. Principal components are calculated using either the covariance or correlation matrix. In general, when using the covariance matrix, all observations must be in equivalent units. If this is not the case, the differing units, and the resulting change in the magnitude of the data values can greatly skew the results, emphasizing only the few with large variances in relation to the others. However, scaling results or converting to equivalent units may not be practical. Therefore, when this occurs, it is conventional to use the correlation matrix instead, as is the case for our data.

Running PCA on all twenty-two of the variables considered, a scree plot (Figure 3) can be obtained to show at which component the eigenvalues begin to display little change.

Figure 3



While the elbow appears to be at 4, we want to account for a greater amount of total variance, therefore we chose to keep **seven** of the twenty-two components. These seven principal components contribute 83.4% of the variance, which is a high proportion, while our dimensionality is still greatly reduced. For the eigenvalues, the individual variances contributed, and the cumulative variance at each principal component, see Table III in the appendix. These values for our chosen seven principal components are shown in Table 2.

Table 2: Eigenanalysis with 7 PCs

Eigenvalue	7.9992	4.6497	1.9397	1.2592	1.0351	0.8021	0.6661
Proportions	0.364	0.211	0.088	0.057	0.047	0.036	0.030
Cumulative	0.364	0.575	0.663	0.720	0.767	0.804	0.834

Understanding the principal components selected is important in our analysis. Identifying the variables that are strongly represented in each component allows us to create a general title describing the meaning of each particular component. Our first principal component (PC1) is dominated by the variables: average SAT score accepted, % of freshmen in top 10 % of Class, Acceptance Rate, and Graduation Rate. So we label PC1 as *Academic Performance.* Using this method to further label the remaining principal components, we choose PC2 as *Publicity*, PC3 as *Social Opportunities*, PC4 as *Faculty Reputation*, PC5 as *Monetary Incentive*, PC6 as *Athletic Recognition*, and PC7 as *Campus Value*. Each component and the corresponding variable coefficients are displayed in Table II in the appendix.

Principal Component Analysis reduces dimensionality and represents our total data set using only seven principal components. While PCA is a very functional method for describing our data, we will now use Factor Analysis to further establish relationships between variables.

Factor Analysis

I. Theoretical Background

The purpose of Factor Analysis (FA) is also to reduce the dimensionality of the variables by identifying underlying and unobservable relationships among two or more variables. In essence, we are grouping together variables that are highly correlated under a single factor. We want to develop an *m*-factor model, with the *m* less than *p*, the number of original variables.

We account for the variation of the variables by examining the underlying *common factors* $F_1, F_2, ..., F_m$, as well as the unique factors (or errors) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ specific to each variable. Each component of the random vector $\boldsymbol{X}$ is expressed as a linear combination of the common factors, using coefficients ℓ_{ij}, or the loading of the *i*th variable on the *j*th factor, plus the specific factor. In matrix form, we obtain $\boldsymbol{X} - \mu = L\boldsymbol{F} + \boldsymbol{\varepsilon}$, where L is the loading coefficient matrix of ℓ_{ij}'s. We assume $F_1, F_2, ..., F_m$ and $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ are independent with multivariate normal distribution. Also, we assume $E(F)=0$, $Cov(F)=E(FF^T)=I$, $E(\varepsilon)=0$, and $Cov(\varepsilon) = E(\varepsilon\varepsilon^T) = \Psi = Diag(\Psi_1, \Psi_2, ..., \Psi_p)$. Then we have $X \sim MN(\mu, \Sigma)$, with mean μ and covariance matrix Σ. We can show that given the *m*-factor model, $\Sigma = LL^T + \Psi$.

To assess the adequacy of our *m*-factor model, we test the null hypothesis $H_0 : \Sigma = LL^T + \Psi$ with a given *m*, versus $H_a : \Sigma$ is any positive definite matrix. Starting with m=1, we test the goodness of fit using the χ^2 (chi-squared) test statistic. If we reject the null hypothesis where $\chi^2 > \chi^2_{\alpha,\upsilon}$, we increase *m* until H_0 can be accepted. We use the test statistic

$$\chi^2 = \left[n - 1 - \frac{2p+5}{6} - \frac{2}{3}m \right] \ln \frac{\left| \hat{\Psi} + \hat{L}\hat{L}^T \right|}{|R|}$$

where $\hat{\Psi}$ and $\hat{L}$ are the Maximum Likelihood Estimates for Ψ and *L,* and *R* is the sample correlation matrix. We then compare χ^2 to $\chi^2_{\upsilon,\alpha}$ at the α - level with

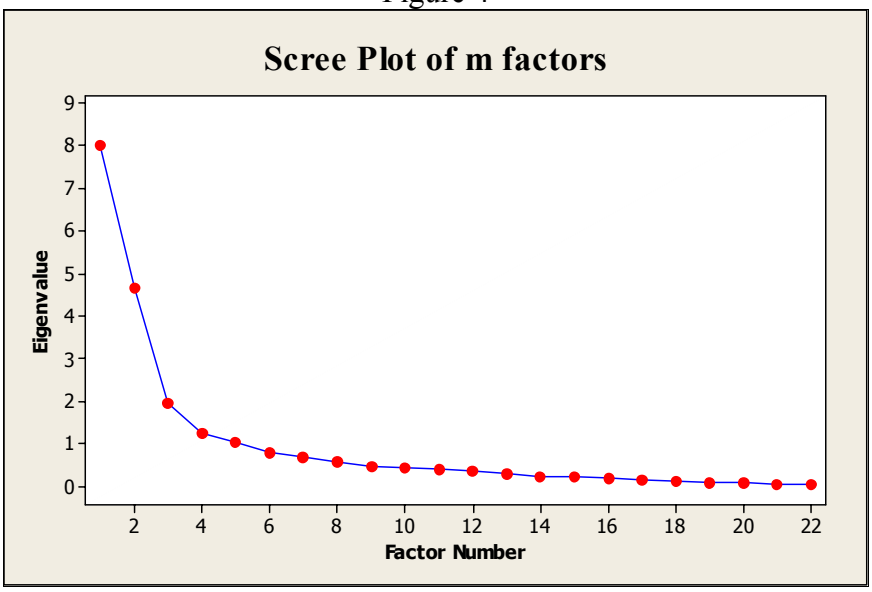
$$\upsilon = \frac{(p-m)^2 - p - m}{2}$$

degrees of freedom to determine if *m* factors present an adequate model for analysis.

II. Analysis

The goal of factor analysis is to create *m* factors that model the variation in all of our 22 variables. Observing the scree plot (Figure 4) of FA computations using MINITAB we are able to deduce a reasonable place to begin *m*-factor extraction. The factors we extract need to pass a chi-squared test; in addition, the p-value must be greater than 0.05.

Figure 4



Evaluating the FA beginning at the elbow of the scree plot (m=4), we compare the test statistic against χ^2 with $\alpha = .05$ and degrees of freedom υ=149. While the test statistic is 169.977 as seen in Table 3, the p-value is 0.115. This results in our lowest adequate p-value, but to account for a greater proportion of the total variance, 65.9%, without excessive factors, m=5 is chosen. Below, Table 3 shows the results for a range of m-factors and the corresponding p-values.

Table 3: m-Factor Test of Adequacy

***m*-Factor**	**Test Statistic**	**Degrees of Freedom**	**P-Value**
1	475.6915	209	0
2	281.3772	188	0.000012
3	252.1336	168	0.000028
4	169.977	149	0.115
5	136.4961	131	0.3535
6	108.052	114	0.6393
7	81.2978	98	0.8888
8	60.6573	83	0.969

We characterize each factor by considering the loading coefficients with the greatest influence on particular variables. Because the first factor places emphasis on Total Cost and Financial Aid, we label it *Monetary Significance*. In the same manner, we name factors two through five, *Student Academic Performance*, *Focus on Individual*, *Learning Environment*, and *Athletics/Scholarship*, respectively. The factor loading values of each factor are found in Table IV in the appendix.

Discriminant Analysis

I. Theory

Discriminant Analysis (DA) uses a linear combination of variables to classify a college into one of two groups. In choosing the variables to distinguish between the two groups, all others are discarded and not considered for this step in the analysis. Members in a population lacking the necessary criteria are discriminated against and placed into the inferior group accordingly. Equally, those possessing the required conditions are placed into the other. In essence, by separating distinct sets of observations, we can allocate new observations to these previously defined groups.

One goal in DA is to maximize the distance between the two groups, while simultaneously minimizing the variance within each. This allows for the construction of a model that will predict a grouping for new data sets, with the greatest amount of accuracy.

We define the two multivariate normal subgroups as π_1 and π_2, where $\pi_1 \sim MN$ ($\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$) and $\pi_2 \sim MN$ ($\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_2$). By convention, π_1 is usually the superior group and π_2 the inferior group. At the beginning of the analysis, it is typically assumed that $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$; thus, *Linear Discriminant Analysis* is initially applied. This equality must be verified through a hypothesis test.

Suppose we have training samples of sizes n_1 and n_2 from π_1 and π_2 respectively, with sample covariance matrices S_1 and S_2. The test of the null hypothesis $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ is carried out by first calculating the pooled unbiased estimate of the common covariance matrix Σ, given by S_p, where

$$S_p = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{2}(n_i - 1)S_i\right).$$

The test statistic is then given by $\frac{M}{c}$, where,

$$M = \left\{\sum_{i=1}^{2}(n_i - 1)\right\}\ln\left(\det\left(S_p\right)\right) - \left\{\sum_{i=1}^{2}\left[(n_i - 1)\ln\left(\det\left(S_i\right)\right)\right]\right\} \text{ and}$$

$$\frac{1}{c} = 1 - \frac{2p^2 - 3p - 1}{6(p-1)}\left(\frac{1}{n_1} + \frac{1}{n_2} - \frac{1}{n_1 + n_2 - 2}\right).$$

The test statistic has a χ^2 distribution with $\upsilon = \frac{1}{2}p(p+1)$ degrees of freedom. If the test statistic is less than $\chi^2_{\alpha,\upsilon}$, the null hypothesis is accepted. Otherwise, the null hypothesis is rejected, and *Quadratic Classification*, when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, must be applied in a similar fashion.

In order to analyze the accuracy of the discriminant function, we calculate the Apparent Error Rate (APER), the ratio of misclassifications in the training samples. Also, the Total Probability of Misclassification (TPM) is calculated to further analyze the accuracy of the method. To obtain the TPM, and classification for the remaining sample data, the Mahalanobis distance (M-distance) between the two populations is needed.

The M-distance squared is,

$$\hat{\Delta}_p^2 = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^T \mathbf{S}_p^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right).$$

Using this, we calculate,

$$TPM = \alpha = 2\Phi\left(-\frac{1}{2}\hat{\Delta}_p\right).$$

II. Analysis

To begin DA, we first need to produce two training samples from our population of 50. The first step is to classify each college into either π_1 or π_2, highly desirable or desirable, respectively, according to some criterion. The original groups are constructed by finding the median of the data for Student/Faculty Ratio, Graduation Debt, and Retention Rate (the variables we determined as most significant based on the data). Each college is now given 0 or 1 points, based upon whether their statistics are worse or better than each mean respectively (in some cases a smaller number is better, and in others, worse). The sum for each of the colleges dictates whether they are placed into π_1 (with 2 or 3 points) or π_2 (with 0 or 1 point). In total, there are fifteen samples in group one, and fifteen in group two. In order to run linear DA, a positive definite covariance matrix is needed; it follows that the size of the training sample must be greater than the number of variables. Thus, we decrease the number of variables and make our classification using: Miles from Major City, Student/Faculty Ratio, Number of Intercollegiate Sports, Total cost, Financial Aid, Graduation Debt, Retention Rate, Yield, Endowment Per Student, and Number of Companies Recruiting on Campus. Next, we must carry out a hypothesis test to determine if $\Sigma_1 = \Sigma_2$.

With the null hypothesis: H_0: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ against H_1: $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, we find the test statistic, $\frac{M}{c} = 67.86040916$. With degrees of freedom, $\upsilon = 55$, the corresponding chi-squared value is 73.31. Our test statistic is less than this chi-squared value, with a p-value of 0.114233, so we are able to accept our hypothesis that $\Sigma_1 = \Sigma_2$ at $\alpha = 0.05$, and can proceed with our application of linear DA.

Running linear DA on the training sample data (Table V in the appendix), we find that all fifteen of each group are correctly placed, giving us an APER of 0%. This is an exceptional error rate so we proceed to calculate the TPM. For this, we need the M-Distance, which is found to be 16.3722, as well as a standard normal probability table of *z*-values [JW07]. The Total Probability of Misclassification is found to be 0.0434. A summary of the linear DA results for the training sample is shown in Table 5.

Table 5: Training Sample Results

	Classified:1	Classified: 2	Total
True: 1	15	0	15
True: 2	0	15	15
Correct	15	15	

N=30 N Correct=30 Proportion Correct=1.00

Finally, we run linear DA for the twenty schools in the test sample, where 12 are classified as highly desirable and 8 as desirable. The full results of our analysis, including classification and rankings based on the squared distance from the group, can be found in Table 6.

Table 6: Discriminant Analysis Test Sample Ranks

Corresponding School	Predicted Group	Squared Distance	Probability	Rank
Tufts University	1	6.219	0.997	1
Boston College	1	6.943	0.998	2
Middlebury College	1	7.223	1.000	3
Mount Holyoke College	1	7.877	0.908	4
Washington University in St. Louis	1	10.237	0.850	5
University of North Carolina, CH	1	11.283	1.000	6
Yale University	1	11.963	1.000	7
Vanderbilt University	1	12.847	0.512	8
Georgetown University	1	14.475	0.999	9
Emory University	1	16.521	0.868	10
Princeton University	1	21.357	1.000	11
Brigham Young University	1	31.362	1.000	12
New York University	2	32.255	1.000	13
Gonzaga University	2	18.692	0.989	14
Iowa State University	2	14.367	1.000	15
Davidson College	2	13.713	0.670	16
Trinity University	2	11.006	0.983	17
Penn State University	2	8.569	0.927	18
Rutgers University	2	7.932	0.989	19
Reed College	2	7.315	0.999	20

In conclusion, we constructed a discriminant function that accurately classifies schools based on the chosen ten variables. Using this model, we can justly apply it to future observations to categorize schools as desirable or highly desirable.

Conclusion

We began our study of colleges with twenty-three variables. One variable, Faculty in the National Academy, was highly correlated to the others, and was therefore dropped from the rest of our analysis. We then proceeded to assess the normality of our remaining twenty-two variables. After evaluating for normality, and making necessary transformations, we moved on to Principal Component Analysis.

PCA allowed us to reduce the dimensionality to seven principal components. These seven components accounted for 83.4% of the total variance. Naming these seven components allowed us to better comprehend the variables that were having the greatest impact on our data.

Next, a second method to reduce dimensionality, Factor Analysis, was applied. We used five factors, and once gain, we named these factors. Using the five factors, we accepted the hypothesis of adequacy.

Finally, we used Discriminant Analysis to classify schools as either desirable or highly desirable. We were able to accomplish this by using linear DA by reducing our number of variables to 10. Both the training sample and the test sample were classified accurately.

We rank the schools in the test sample based upon their M-distance from the respective groups. In the case of π_1 (highly desirable), the smaller the M-distance, the better the rank, while the opposite is true for π_2 (desirable), where the greater the M-distance, the better the rank. The desirable rankings followed the highly desirable rankings in the overall order; these results can be seen in Table 6.

Our rankings, when compared to the *US News and World Report's Rankings* (a major source of our data), exhibit some interesting deviations. For example, the differences are seen in considering the *US News and World Report's* rankings of top national universities. The top ranked school according to their analysis is Princeton University. However, Princeton ranked 11th on our list. While, the top school according to our classification is Tufts University. Again, this differed from their results, ranking Tufts 27th. Boston College followed the same pattern, ranked 2nd in our analysis and 34th by the *US News and World Report*.

While there are some major differences, such as those mentioned above, there are also similarities. For instance, Yale University is ranked 3rd nationally [US07], and is also highly ranked 7th in our analysis. Likewise, Middlebury College is ranked 3rd in our DA, and is ranked 5th among the top Liberal Arts Colleges in the nation [US07].

It is apparent to these authors that the incorporation of opinion, in the ranking of schools, greatly distorts the final results. Our analysis is based on a statistical scientific procedure, excluding extraneous information that is usually overemphasized, thus constructing unbiased rankings.

Acknowledgements

We would like to thank Dr. Vasant Waikar, director of the statistics seminar and co-director of SUMSRI, for his tutelage throughout the research process. We would also like to thank Dr. Tom Farmer for his continued assistance during the writing and revision phases of our report, and Dr. Dennis Davenport, co-director of SUMSRI. Sweet Action Score thanks Team Amazing for continued support. We also acknowledge our graduate assistant, Kevin Tolliver, and all of the SUMSRI participants and graduate assistants for an unforgettable experience. Finally, we would like to express great appreciation to the National Security Agency, the National Science Foundation, and Miami University for giving us this opportunity.

Bibliography

[BLS07] Bureau of Labor and Statistics, *College Enrollment and Work Activity of 2006 High School Graduates*: Retrieved 20 June 2007 from http://www.bls.gov/news.release/hsgec.nr0.htm.

[JR07] Julie Rawe, TIME Magazine (2007), *The College Rankings Revolt*: Retrieved 5 July 2007 from http://www.time.com/time/nation/article/0,8599,1601485,00.html.

[JW07] Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, sixth ed., Prentice Hall, 2007.

[OL07] Ordo Ludus: College Rankings, *College and University Rankings*: Retrieved 20 June 2007 from http://www.ordoludus.com.

[US07] U.S. News and World Report, *America's Best Colleges 2007*, Retrieved 27 June 2007 from http://colleges.usnews.rankingsandreviews.com/usnews/edu/college/rankings/rankindex_brief.php

Appendix

Table I: Colleges

College	
Amherst College	Northwestern University
Arizona State University	New York University
Bates College	Oberlin College
Boston College	Ohio State University
Brigham Young University	Penn State University
Brown University	Princeton University
Bryn Mawr College	Purdue University
Boston University	Reed College
California Institute of Technology	Rice University
Columbia University	Rutgers University
Davidson College	Stanford University
DePaul University	Temple University
Duke University	New School University
Emory College	Trinity University
Georgetown University	Tufts University
Gonzaga University	Tulane University
Grinnell College	University of Kansas
Harvard University	University of California, Los Angeles
Iowa State University	University of North Carolina, Chapel Hill
Johns Hopkins University	Vanderbilt University
Michigan State University	Vassar College
Middlebury College	Villanova University
Massachusetts Institute of Technology	Washington University in St. Louis
Mount Holyoke College	Williams College
North Dakota State University	Yale University

Table II: 22 Component PC Model

Variable	PC1	PC2	PC3	PC4
Miles From Major City (250,000+ Population)	-0.072	0.006	-0.492	0.369
% of Classes with <20 Students	0.242	-0.199	-0.079	-0.307
Student-Faculty Ratio	-0.187	0.138	-0.26	0.378
% Greek Members	0.007	0.17	0.43	0.106
Number of Intercollegiate Sports	0.178	0.238	-0.137	0.156
Number of Intramural Sports	0.003	0.359	0.104	0.048
Bowl Appearances	-0.14	0.334	-0.037	-0.105
Final Four Appearances	-0.134	0.283	0.071	-0.104

Overall NCAA Championships	-0.019	0.352	-0.129	0.029
Average SAT Score Accepted	0.318	0.007	0.024	0.17
National Merit Scholars	0.145	0.329	-0.001	-0.228
% of Freshmen in Top 10 % of Class	0.315	0.118	0.017	-0.019
Acceptance Rate	-0.335	-0.044	0.019	-0.056
Faculty Awards	0.168	0.302	0.14	-0.332
Total Cost (Room, Board, Tuition)	0.294	-0.125	0.209	0.082
Financial Aid	0.3	-0.074	0.058	0.173
Graduation Debt	-0.151	-0.038	0.392	0.166
Retention Rate (Returning Sophomores)	0.292	0.105	-0.057	0.158
Graduation Rate	0.306	-0.022	0.014	0.286
Yield (% Attending of total Accepted)	0.056	0.179	-0.402	-0.307
Endowment per Student	0.297	-0.016	-0.18	-0.079
Number of Companies Recruiting	-0.021	0.36	0.167	0.315

PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
0.299	-0.115	0.369	0.047	0.179	0.136	0.186	0.399	0.026	0.276
0.299	-0.027	-0.069	0.207	0.035	-0.024	0.022	-0.153	87	0.384
0.088	-0.099	-0.502	0.27	-0.009	-0.141	0.105	-0.32	-0.111	-0.11
0.428	-0.383	-0.216	-0.387	0.141	0.136	-0.154	-0.042	0.236	0.259
-0.25	0.266	0.098	-0.455	-0.453	-0.118	0.007	-0.159	-0.011	0.449
-0.18	-0.283	0.46	0.136	0.026	-0.236	-0.004	-0.165	0.455	-0.275
-0.199	0.053	0.019	0.17	0.349	0.301	0.233	-0.365	0.027	0.358
0.102	0.557	-0.193	0.029	0.299	-0.358	0.184	0.255	0.298	0.05
0.119	0.212	0.027	0.289	-0.084	0.251	-0.781	0.087	0.005	0.024
-0.072	-0.108	-0.025	-0.006	0.21	-0.197	0.043	0.208	-0.085	0.017
-0.013	-0.36	0.106	0.26	-0.097	-0.22	0.096	-0.042	-0.331	0.219
0.068	-0.007	-0.061	0.012	0.187	-0.136	-0.089	-0.013	-0.414	-0.12
-0.031	-0.186	0.05	0.029	-0.106	0.103	0.038	0.125	0.109	-0.003
0.048	0.054	-0.076	0.046	-0.065	0.264	0.207	0.331	-0.063	-0.107
0.01	0.151	0.106	0.227	-0.026	0.285	0.168	-0.012	0.056	0.003
-0.02	0.074	-0.065	0.292	-0.293	0.296	0.19	-0.056	0.28	-0.046
0.442	0.262	0.411	0.178	-0.156	-0.253	0.059	-0.149	-0.259	0.039
-0.069	0.085	0.175	-0.267	0.398	0.18	-0.032	-0.094	-0.148	-0.237
0.138	0.091	-0.069	0.049	0.103	-0.092	-0.045	-0.247	0.227	-0.038
0.487	0.041	0.6	-0.261	-0.193	0.075	0.183	-0.25	0.072	-0.371
0.005	-0.149	-0.123	0.109	-0.127	-0.339	-0.104	0.235	0.276	0.063
0.037	-0.068	-0.196	-0.008	-0.292	0.106	0.231	0.267	-0.14	-0.11

PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
0.122	-0.001	-0.089	0.12	0.086	-0.092	0.021	-0.012
-0.098	0.275	0.019	0.234	-0.55	0	-0.195	0.049
0.371	-0.116	0.016	-0.021	-0.248	0.073	0.106	0.073
0.151	-0.038	-0.11	-0.045	0.095	0.112	0.044	-0.015
0.182	-0.004	-0.021	-0.024	-0.103	-0.029	-0.009	0.165
0.106	0.132	-0.186	0.141	-0.233	-0.087	0.04	0.052
-0.369	-0.272	-0.126	-0.135	0.052	-0.036	0.07	0.043
0.109	0.203	0.013	0.007	0.131	0.188	-0.114	0.029
-0.028	0.078	0.031	-0.125	-0.042	0.029	0.014	-0.001
-0.115	0.147	-0.042	-0.768	-0.273	-0.123	0.019	-0.029
0.162	0.189	0.341	-0.039	0.363	0.195	-0.083	-0.169
-0.007	0.011	-0.526	0.201	0.201	-0.114	-0.173	0.479
-0.005	-0.036	0.283	-0.218	-0.028	0.09	-0.411	0.697
0.358	-0.358	0.118	0.036	-0.235	-0.41	0.021	-0.018
0.15	0.208	0.102	-0.018	0.071	0.266	0.594	0.379
0.114	-0.047	-0.294	-0.15	0.201	0.168	-0.498	-0.205
-0.056	-0.336	0	-0.112	-0.119	0.041	-0.046	-0.031
0.058	-0.159	0.321	0.141	-0.231	0.452	-0.264	-0.063
-0.148	-0.011	0.483	0.095	0.286	-0.538	-0.093	0.1
-0.149	0.087	-0.032	-0.252	0.088	0.06	0.102	0.039
-0.261	-0.587	0.02	0.073	0.031	0.292	0.176	0.106
-0.548	0.222	0.06	0.247	-0.168	0.051	-0.002	-0.023

Table III: Complete Eigenanalysis

Component	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Eigenvalue	7.9992	4.6497	1.9397	1.2592	1.0351	0.8021	0.6661	0.5778
Proportion	0.364	0.211	0.088	0.057	0.047	0.036	0.03	0.026
Cumulative	0.364	0.575	0.663	0.72	0.767	0.804	0.834	0.86
Component	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Eigenvalue	0.4796	0.4378	0.4103	0.366	0.2756	0.2372	0.2119	0.1742
Proportion	0.022	0.02	0.019	0.017	0.013	0.011	0.01	0.008
Cumulative	0.882	0.902	0.921	0.937	0.95	0.961	97	0.978
Component	**17**	**18**	**19**	**20**	**21**	**22**		
Eigenvalue	0.1359	0.1117	0.0802	0.0747	0.044	0.032		
Proportion	0.006	0.005	0.004	0.003	0.002	0.001		
Cumulative	0.984	0.99	0.993	0.997	0.999	1		

Table IV: Factor Loadings

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Ψ
Miles From Major City (250,000+ Population)	-0.287	0.04	0.168	-0.053	-0.281	0.806
% of Classes with <20 Students	0.632	0.211	0.462	0.487	0.115	0.091
Student-Faculty Ratio	-0.598	0.02	-0.088	-0.196	-0.152	0.573
% Greek Members	0.008	0.015	-0.397	0.014	0.081	0.835
Number of Intercollegiate Sports	0.183	0.621	-0.33	-0.152	-0.125	0.434
Number of Intramural Sports	-0.17	0.224	-0.742	0.055	-0.135	0.35
Bowl Appearances	-0.496	0.084	-0.58	-0.001	0.17	0.382
Final Four Appearances	-0.429	0.138	-0.422	-0.145	0.429	0.413
Overall NCAA Championships	-0.267	0.345	-0.514	0.017	0.108	0.533
Average SAT Score Accepted	0.702	0.488	-0.044	0.058	-0.318	0.162
National Merit Scholars	0.093	0.452	-0.63	0.493	-0.08	0.141
% of Freshmen in Top 10 % of Class	0.598	0.691	-0.164	0.154	-0.046	0.112
Acceptance Rate	-0.718	-0.696	0	0	0	0
Faculty Awards	0.267	0.387	-0.618	0.321	0.21	0.25
Total Cost (Room, Board, Tuition)	0.999	0.053	0	0	0	0
Financial Aid	0.868	0.213	0.012	0.015	-0.174	0.171
Graduation Debt	-0.088	-0.434	-0.1	-0.229	0.306	0.648
Retention Rate (Returning Sophomores)	0.588	0.585	-0.173	-0.08	-0.161	0.25
Graduation Rate	0.738	0.464	0.061	-0.068	-0.193	0.194
Yield (% Attending of total Accepted)	-0.182	0.467	-0.019	0.377	0.142	0.586
Endowment per Student	0.523	0.553	0.154	0.335	-0.293	0.198
Number of Companies Recruiting	-0.171	0.182	-0.73	-0.175	-0.08	0.369
% Variance	0.265	0.161	0.148	0.048	0.038	
Cumulative Variance	0.265	0.426	0.574	0.622	0.659	

Table V: Training Sample

Highly Desirable	**Desirable**
Amherst College	Arizona State University
Bates College	Brown University
Brown University	DePaul University
Bryn Mawr College	Duke University
California Institute of Technology	Grinnell College
Columbia University	North Dakota State University
Harvard University	Oberlin College
Johns Hopkins University	Ohio State University
Massachusetts Institute of Technology	Purdue University
Northwestern University	Temple College
Rice University	New School University
Stanford University	Tulane University
University of California, LA	University of Kansas
Vassar College	University of Illinois
Williams College	Villanova University