A Multivariate Statistical Analysis of Crime Rate in US Cities

Kendall Williams
Howard University
Washington DC
k_r_williams@howard.edu

July 2004

Ralph Gedeon
University of Florida
Gainesville, FL
ralphael@ufl.edu

Abstract

We classify a city as safe or unsafe by using multivariate methods of Principal Components, Factor Analysis, and Discriminant Analysis. In addition, we discover which variables have salience in the identification of a city being safe or dangerous. The aforementioned analytical techniques can assist governments in finding out what variables they need to change to improve their city and make it a better place to live.

Introduction

We are living in a time when moving is second nature to us. We move to go to school. We move to find better career opportunities. We move because living in one place for the rest of our lives is not a sentence we wish to serve. Do we move to a city where safety is not an issue or to a city where safety is a major concern?

What are some of the factors that may be related to safety? On the average, approximately 506.1 violent crimes per 100,000 people in the United States are committed each year. Are there some variables you could observe to predict if a city is safe or dangerous without the use of the actual number of crimes committed in that city? In addition, the population in most of the cities in the United States is increasing rapidly. Is a more populous city safer or more dangerous than a less populous city?

We will analyze some of the variables that are used to profile U.S. cities. We will use multivariate statistical methods to classify cities as having high or low crime rates. The true classification is based upon the national average of violent crimes per 100,000 people. We will determine which variables are suitable for analysis by using Principal Components Analysis and Factor Analysis. Then, using Discriminant Analysis, we will categorize the cities into two groups: high crime rate or low crime rate.

Data

We analyze data from the year 2000 because all the information that we need is readily available for that year. Our data consists of 14 variables that fall under the categories of general characteristics, social characteristics, economic characteristics, housing characteristics, weather characteristics and political characteristics. Our data comes from the World Gazetteer, Census 2000 web search data, Yahoo Yellow Pages Search, and moving websites such as Moving.com and bestplaces.net. Our data includes profiling information from 100 US cities. Half of those cities are the most populous US cities, and the other half are much smaller US cities. The variables are: total population, percent of the population between the ages of 5 and 65, ethnicity (% Caucasian), percent of single parent homes, percent of population 25 years or older with a bachelor's degree, unemployment rate, average number of people in household, per capita income, number of churches in the city, voter turnout, percentage of Republican state officials, percent of population institutionalized, per capita alcohol consumption, and days of sunshine per year.

Normality of the data

Before we use any of the multivariate methods, we must digress in order to provide some information on the preliminary steps of this technique. Mainly, we must test for normality. Principal Components, Factor Analysis, and Discriminant Analysis cannot be applied unless the variables $x_1, x_2, \ldots, x_p$ have a multivariate normal distribution. If each x_i is normally distributed, then it is reasonable to assume that $x_1, x_2, \ldots, x_p$ has a multivariate normal distribution. If not, transformations can be applied, such as the log transformation, a square root transformation, or an exponential transformation, in order to obtain an overall multivariate normality.

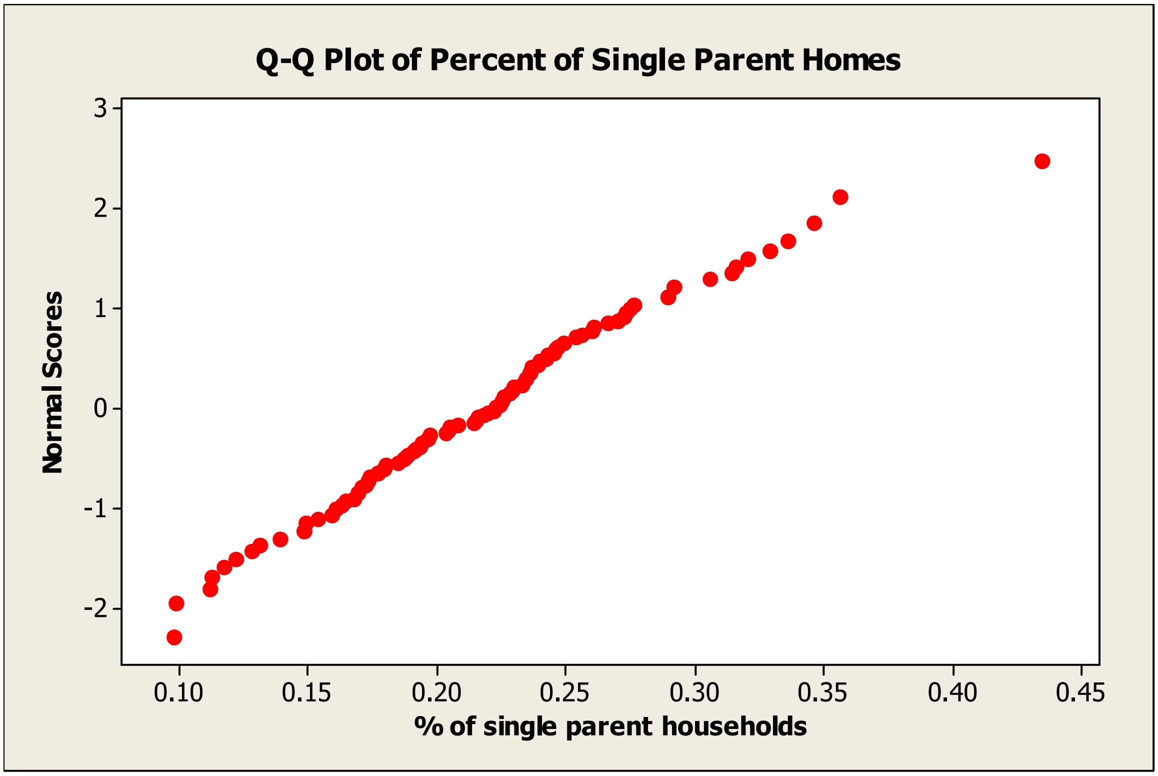
Normality of data is assessed by the linearity of quantile-quantile plots. In Q-Q plots, we compare the real values of the variables against the standardized values. For example, suppose that $f(x, \theta)$ is the density distribution of the population, where θ is the unknown parameter. To test for normality of this population we use the Q-Q plots of the order statistic, $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n)}$, where $x_{(i)}$ is the i/n^th quantile of the sample. If x is normally distributed (μ, σ^2), where μ is the population mean and σ^2 is the variance of the population, then $z = (x- \mu)/ \sigma$ is normally distributed $(0, 1^2)$. The i/n^th quantile of the z distribution is $z_{(i)}$, i.e. $P(z \leq z_{(i)}) = i/n$, which implies that $x_{(i)} = \mu + z_{(i)} \sigma$ is a straight line. Thus if the plot of $(x_{(i)}, z_{(i)})$ or the Q-Q plot is linear, then the variable is assessed as normal.

In fact, we test linearity of the Q-Q plots by using a sample correlation coefficient, r_Q. Let ρ be the correlation between $x_{(i)}$ and $z_{(i)}$. We reject H_0: $\rho=1$, i.e. linearity in favor of H_1: $\rho<1$, if the sample correlation coefficient is less than our critical value. The correlation coefficient is given by:

$$r_Q = \frac{\left(Cov(x_{(i)}, z_{(i)}\right)}{\left[\sqrt{Var(x_{(i)})Var(z_{(i)})}\right]}.$$

This coefficient can take on any value between zero and one. Values of r_Q must be relatively close to 1 to be considered as high and to cause us to accept the Q-Q plot as linear. Therefore, the values of r_Q that are less than our critical value at α=.01 are deemed low in correlation and that particular variable, which could be any x_i, must be discarded or transformed in order to obtain normality.

In the present case, since our sample size is 100 and we want α=.01, our correlation coefficient r_Q should be at least .9822. Four out of our 14 variables test normal without any transformation: percent of population from 5 to 65, percent of single parent homes, percent of population 25 or older with bachelor's degree, and percent of Republican state officials. Three out of the remaining ten test normal after some form of transformation: number in household, voter turnout, and days of sunshine. The seven remaining variables do not test normal but, with the exception of the city population variable, their r_Q values are all extremely close to .9822. It is a rare occasion that a variable such as population in a sample size of 100 will test normal. Transformed, the r_Q value for population was .9460. The Q-Q plots of the data that are not normal give the notion that there are no transformations which could achieve normality. Refer to Table 1 in the appendix for a complete list of r-values. The graph below is the Q-Q plot of the percent of single parent homes. It plots the normal scores of the variable against the original values of the variable. The linearity of the graph implies normality.



Principal Component Analysis Theory

Having a large number of variables in a study makes it difficult to decipher patterns of association. Variables sometimes tend to repeat themselves. Repetition is a sign of multicolinearity of variables, meaning that the variables may be presenting some of the same information. Principal Components Analysis simplifies multivariate data in that it reduces the dimensionality of the data. It does so by using mainly the primary variables to explain the majority of the information provided by the data set. Analysis of a smaller number of variables always makes for a simpler process.

Simply stated, in principal components analysis we take linear combinations of all of the original variables so that we may reduce the number of variables from p to m, where the number m of principal components is less than p. Further, the method allows us to take the principal components and use them to gain information about the entire data set via the correlation between the principal components and the original variables. Matrices of correlations or loadings matrices show which principal component each variable is most highly associated with. The first principal component is determined by the linear combination that has the highest variance. Variance measures the diffusion of the data. After the first principal component is obtained, we must determine whether or not it provides a sufficient amount of or all of the information displayed by the data set. If it does not provide adequate information, then the linear combination that displays the highest variance accounted for after the first principal component's variation is removed is designated as the second principal component. This process goes on until an ample amount of information/variance is accounted for. Each principal component accounts for a dimension and the process continues only on the remaining dimensions. Designating a dimension as a principal component often reveals information about correlations between remaining variables which at first was not readily available.

The main objective of Principal Components Analysis is to locate linear combinations $y_i = \underline{\ell}_i^T\underline{x} = \ell_{1i}x_1+\ell_{2i}x_2+ \dots +\ell_{pi}x_p$, with the greatest variance. We want

$$\mathrm{Var}(y_i) = \mathrm{Var}\ (\underline{\ell}_i^T\underline{x}) = \underline{\ell}_i^T\sum\underline{\ell}_i$$

, where Σ is the covariance matrix, to be the maximum among all the normalized coefficient vectors $\underline{\ell}_i$. This result is achieved by way of Lagrange Multipliers. Taking the partial derivative with respect to $\underline{\ell}_i$ of the $\mathrm{Var}(y_i) - \lambda(\underline{\ell}_i^T\underline{\ell}_i - 1)$, where λ is the Lagrange Multiplier results in the equation

$$(\Sigma - \lambda \mathrm{I})\underline{\ell}_i = \underline{0}\,,$$

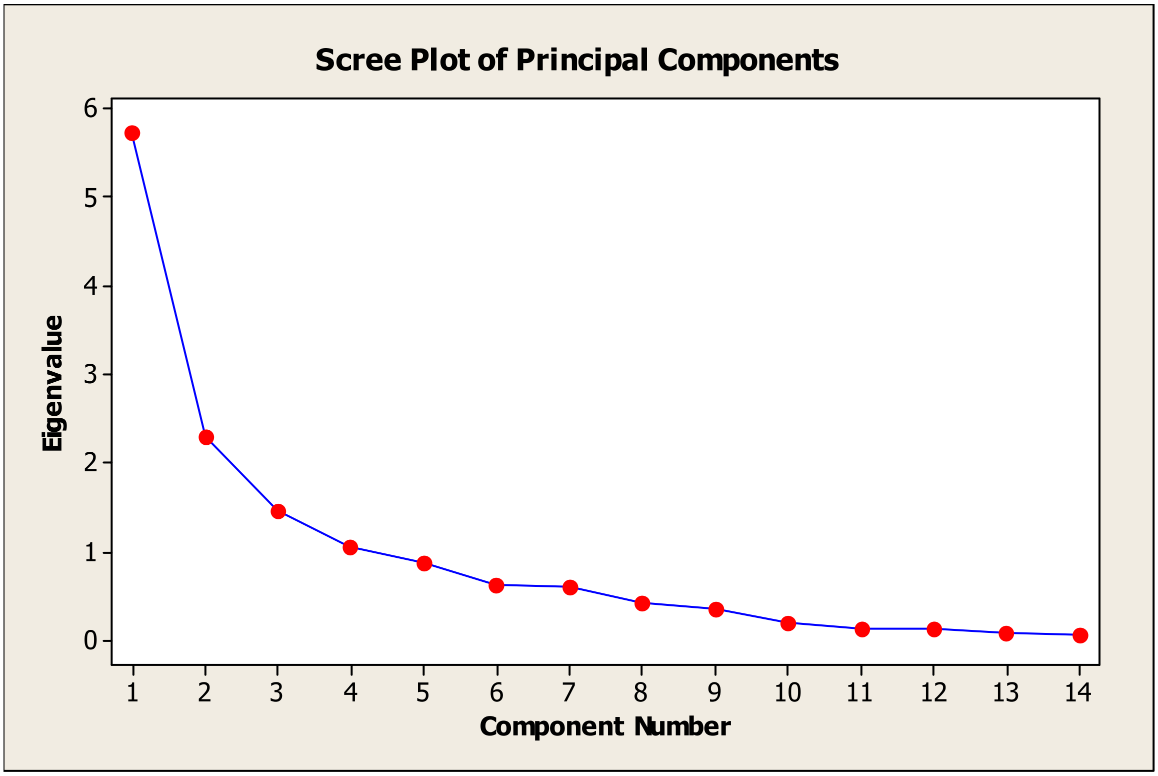
where $\underline{\ell}_i$ is not equal to the zero vector. From the above equations it can be easily verified that λ is a characteristic root of $\sum$ and λ_i is equal to the variance of y_i where $\lambda_1 > \lambda_2 > \dots > \lambda_p$ are the characteristic roots. Note that they are positive. The characteristic vector corresponding to λ_1, the root that accounts for the maximum variance, is $\underline{\ell}_1$. The percentage of variance that any particular principal component accounts for can be calculated by dividing the variance of that component by the sum of all the variances, i.e.

$$\frac{\lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}$$

We use the high correlations between the principal components and the original variables to define which components we will utilize and which ones we will discard. One device that assists us in this decision process is a scree plot. Scree plots are graphs of the variance (eigenvalue) of each principal component in descending order. A point called an "elbow" is designated. Below this point is where the graph becomes somewhat horizontal. Any principal components whose variances lie above this point are kept and the others are discarded. The original variables that are highly correlated with each principal component that is kept determine what the label of that particular component will be.

Principal Component Analysis

We now use Minitab Version 14 to run Principal Components Analysis on our fourteen variables. In cases where such large matrices are being utilized, tools such as Minitab must be used. Minitab can efficiently and accurately analyze the data without the possibility of human error. First, we take our 100 by 14 raw data matrix and run principal components for all fourteen variables. We then use our discretion along with the scree plot of the data to determine how many principal components should be kept. Our first intuition would be to aim for a high percentage of the information to be covered by the principal components no matter how many we choose. But this outlook can defeat the purpose of running principal components analysis: to sufficiently reduce the dimensionality of the data. For example, we might choose eleven of the fourteen variables as principal components in order to account for 98.1% of the variance, but the reduction of dimensionality would be insufficient. Instead, we prefer to at least cut the number of variables in half. As shown below, the "elbow" in our scree plot appears at the seventh principal component, therefore we keep the first six principal components – a valuable reduction in dimensionality.



The first six principal components account for 86% of the variance. Upon running the test, Minitab provides the eigenvalues, proportions of the variance, and the cumulative percentage of the variance covered by each principal component. See Table 3 in appendix for all these values for all 14 principal components. The table below shows these values for our first six principal components.

Eigenvalue	5.7212	2.3003	1.4565	1.0586	.8693	.6320
Proportions	.409	.164	.104	.076	.062	.045
Cumulative	.409	.573	.667	.753	.815	.860

The focal point of this table is the fact that we account for 86% of our variance while substantially reducing our dimensionality from 14 to six.

By looking at the correlations between our original variables and our principal components, we are able to name some principal components based upon which variables each one is highly correlated with. Other principal components are difficult to label due to the fact that the highly correlated variables are difficult to categorize. Principal Component One is highly correlated with the number of churches, unemployment, population, and per capita income, therefore we name it “Size”. We name PC2 “Household Characteristics”, PC3 “Political Affiliation”, PC5 “Alcohol Consumption”, PC6 “Percent Institutionalized”, and PC4 “State of Mind”. Since we have reduced the dimensionality, we perform further analysis, namely Discriminant Analysis, on these principal components. Below is a table showing the correlation between our original variables and our six principal components.

Principal Components Coefficients

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
(Log) Population	-0.377	0.010	0.079	-0.113	-0.218	-0.316
(Squared)Ethnicity	0.346	0.057	-0.275	0.219	0.180	-0.140
(Log)Unemployment	-0.380	-0.054	0.060	-0.052	-0.105	-0.101
(5sqrts) Number of Churches	-0.378	0.048	-0.018	-0.121	-0.196	-0.260
(Sqrt(log)Number in Household	0.163	-0.490	-0.016	-0.292	0.032	0.187
(Log(log(log)))Voter Turnout	0.053	0.486	-0.208	-0.314	0.285	-0.021
(Sqrt(log))Institutionalized+1	-0.221	0.281	-0.135	0.286	-0.200	0.643
(Log)Days of Sunshine	-0.018	-0.453	-0.121	0.499	-0.265	0.051
(Log)Alcohol Consumption	-0.089	-0.073	0.500	0.416	0.579	-0.219
(Log)Per Capita Income	0.383	0.017	0.174	-0.072	-0.047	0.169
Percent Population f/ 5 to 65	0.212	-0.248	0.221	-0.355	-0.227	-0.023
% of single parent households	-0.258	-0.290	0.004	-0.318	0.391	0.279
% w/ bachelor degree	0.325	0.181	0.214	0.029	-0.344	-0.291
% of Rep. St. Officials	0.021	-0.215	-0.680	0.033	0.131	-0.333

Factor Analysis Theory

Factor Analysis, like Principal Components Analysis, is a method that can be used to reduce the dimensionality of multivariate data. However, Factor Analysis is concerned with identifying underlying sources of variance common to two or more variables. These common factors describe, if possible, the covariance relationships among these variables. Factor Analysis has a common factor model that observes the variation in each variable. This model is attributable to the underlying common factors and to an additional source of variation called specific factors (or sometimes errors).

The objective of Factor Analysis is to identify these variables or common factors and explain their relationships to the observed data in terms of a few underlying, but unobservable variables. In other words, our goal is to infer factor structure from the patterns of correlation in the data by letting the observed patterns of association in the data determine the factor solution.

We apply Factor Analysis to our original data after completing principal components. In order for our m-Factors model to be adequate the number m of factors should be less than p, which is the total number of variables. We will start with m = 1 and then apply the adequacy test. In the χ^2 adequacy test, we test the null hypothesis H_0: $\Sigma = LL^T + \Psi$ against H_1: Σ, where Σ is a positive definite p x p matrix. The equation we use to do our test statistic is:

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

where n is the sample size and $\hat{L}\hat{L}^T + \hat{\Psi}$ is the maximum likelihood estimator of Σ. We accept H_0 if χ^2 is less than the critical value, $\chi^2_{\alpha,\nu}$, where α is the chance of type one error and ν is the degrees of freedom for the χ^2 test. We calculate the degrees of freedom by $\nu = ½ [(p - m)^2 - p - m]$. Remember that we start with m = 1 and apply the χ^2 adequacy test. If we reject H_0 when m = 1, then we go to the next step, where m = 2. We continue this process until we accept H_0 or we have exhausted all possible m values.

In Factor Analysis, we pay more attention to the use of rotation for facilitating the interpretation of a factor analytic solution. This is because interpreting factor solutions could be very difficult or even impossible if we cannot observe any pattern. Thus, it makes sense to rotate the factor solution in order to find a solution that displays a simpler structure, in turn making the interpretation of the factors easier.

The factor analysis model can be formed by using the common factors, the specific factors, and the data for the experimental units. Specific factors are the unique factors from the correlation of the underlying common factors. Let $\underline{X} = [x_1\ x_2\ x_3 \cdots x_p]^T$ be a p by 1 observable random vector of response variables. Suppose $\underline{X} \sim MN(\underline{\mu}, \Sigma)$, which means that $\underline{X}$ has a multivariate normal distribution with p variables, $\underline{\mu}$ is the mean of $\underline{X}$ and Σ is the covariance matrix of $\underline{X}$. The factor analysis model can be expressed in matrix notation as: $\underline{X} - \underline{\mu} = L\,\underline{F} + \underline{\varepsilon}$, where $\underline{X} - \underline{\mu}$ is a p x 1 vector, L is the p x m matrix of factor loadings , $\underline{F}$ is the m x 1 vector of unobservable common factors, and $\underline{\varepsilon}$ is the p x 1 vector of unobservable unique factors, which are unique to each x_i.

With so many unobserved quantities, a direct verification of the factor model from observations on $x_1, x_2, x_3, \cdots, x_p$ is unachievable. Thus, more assumptions about the random vectors F and ε must be made. In particular, the common factors $F_1, F_2, \cdots, F_m$ and the unique factors $\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_P$ are assumed to be independent and multivariate normal. We assume $E(\text{F}) = \underline{0}$ and $\text{Cov}(\text{F}) = E\ (\text{FF}^{\text{T}}) = \text{I}$. We also assume $E(\varepsilon) = \underline{0}$ and $\text{Cov}\ (\varepsilon) = E\ (\varepsilon\varepsilon^{\text{T}}) = \Psi$, where Ψ is a p x p diagonal matrix. The covariance of $\underline{\text{X}}$ from the factor model is:

$$\begin{aligned}\text{Cov}(\underline{\text{X}}) &= \text{E}[(\text{LF} + \varepsilon)(\text{X} - \mu)^{\text{T}}] = \text{E}[(\text{LF} + \varepsilon)\ (\text{LF} + \varepsilon)^{\text{T}}] = \text{E}\{(\text{LF} + \varepsilon)[(\text{LF})^{\text{T}} + \varepsilon^{\text{T}}]\}\\ &= \text{E}[\text{LF}(\text{LF})^{\text{T}} + \varepsilon(\text{LF})^{\text{T}} + \text{LF}\ \varepsilon^{\text{T}} + \varepsilon\ \varepsilon^{\text{T}}]\end{aligned}$$

As a result,

$$\begin{aligned}\Sigma = \text{Cov}(\underline{\text{X}}) &= \text{E}(\text{X} - \mu)\ (\text{X} - \mu)^{\text{T}} = \text{E}(\text{LF} + \varepsilon)\ (\text{LF} + \varepsilon)^{\text{T}}\\ &= \text{L}[\text{E}(\text{FF}^{\text{T}})]\text{L}^{\text{T}} + \text{E}(\varepsilon\text{F}^{\text{T}})\text{L}^{\text{T}} + \text{L}[\text{E}(\text{F}\varepsilon^{\text{T}})] + \text{E}(\varepsilon\varepsilon^{\text{T}})\end{aligned}$$

Since ε and F are independent and $\text{E}(\varepsilon) = 0$ and $\text{E}(\text{F}) = 0$, then $\text{E}(\varepsilon\text{F}^{\text{T}})\text{L}^{\text{T}} = 0$ and $\text{L}[\text{E}(\text{F}\varepsilon^{\text{T}})] = 0$. Thus $\Sigma = \text{L}\ \text{L}^{\text{T}} + \Psi$. The solution of the equation $\Sigma\ = \text{L}\ \text{L}^{\text{T}} + \Psi$ is not unique (unless the number of factors equals 1), which means that the factors are harder to interpret. Rotating any solution to obtain a new factor structure creates a solution that may show a simpler structure, having factors that are more clearly marked by high loadings for some variables and low loadings for other variables. The simpler structure makes the factors easier to interpret. Since Σ is unknown, we use a sample $x_1, x_2, ..., x_n$ to obtain the sample m-factor solution $\hat{L}$ and $\hat{\Psi}$ which are maximum likelihood estimates.

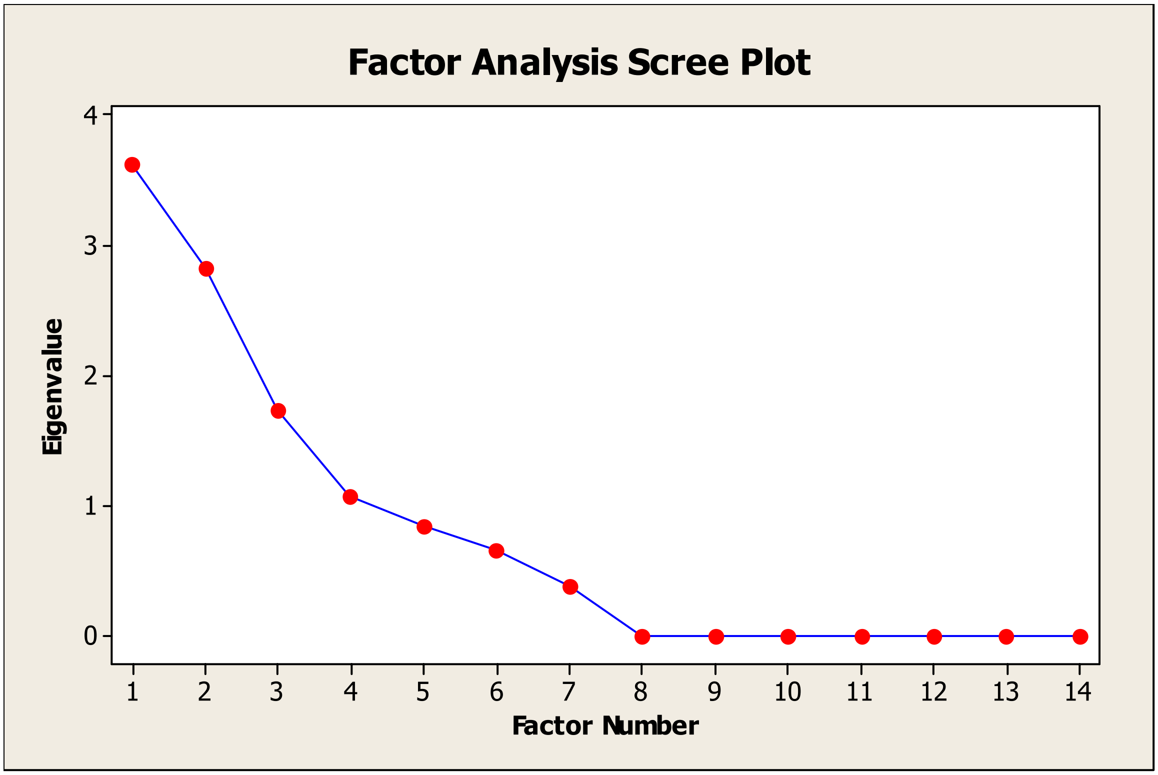
Factor Analysis Application

Our goal here is to create an m-Factors model that observes the variation in all the 14 variables. Since our data have 14 variables then our m must be less than $\frac{1}{2}(2\text{p} + 1 - \sqrt{8p+1}) \approx 9$. We let $\alpha = 0.05$ in our factor model. The first step in applying factor analysis in our study is to find the 14 x m matrix of factor loadings L for one factor via Minitab. Loadings and their communalities are recorded in our Minitab worksheet. These are maximum likelihood estimates. Being that we are looking for one factor in this first step, m is equal to one. Next we find the matrix Ψ for this factor by subtracting all of the communalities from one and diagonalizing these values. Next, we find Σ by multiplying L by its transpose and adding Ψ. It is now time for us to find the correlation matrix of our variables. We do so using Minitab. After finding this matrix, we then find the eigenvalues of it and record them in our worksheet for later use. The following step is to use Excel to calculate the determinant of our correlation matrix using its eigenvalues and the determinant of our factor loadings matrix using its eigenvalues. We utilize these determinants to find our test statistic. Next we calculate our degrees of freedom in order to determine the critical value. We compare our values for the test statistic and the critical value for each particular number of factors. Our goal is to get a critical value greater than our test statistic for some number of factors m in order to accept H_0. This determines when we have a sufficient number of factors to account for all of our variables. We carry out these steps a total of seven times, with the exception of finding the correlation matrix and its eigenvalues because these values remain the same throughout the process. This is called the χ^2 test of adequacy of the m-factor model.

m-Factor Model: Test for Adequacy

m-Factor	Test Statistic	Degrees of Freedom	Critical Values	P-Values
1 Factor Model	401.3673383	77	98.49	0
2 Factor Model	289.4783045	64	83.68	0
3 Factor Model	177.9488547	52	69.83	0
4 Factor Model	109.5053502	41	56.94	0
5 Factor Model	74.07789622	31	44.99	2.2E-05
6 Factor Model	40.97929629	22	33.92	0.008288
7 Factor Model	22.42065941	14	23.69	0.070374

Since the goal is to achieve a critical value greater than the test statistic, we finally succeed at seven factors with a test statistic of 22.42 and a critical value of 23.69. The above table shows the adequacy test results from m = 1 through m = 7. The following graph confirms the results of our m-factor model. The elbow of the scree plot is approximately at eight factors, therefore we use a total of seven factors.



Our next step is to label our seven factors that we obtain. We do this by observing the correlation between each individual factor and our original variables. If there is a high correlation between a variable and a factor, then that variable helps in the determination of what the name of that particular factor should be. In our case, rotation is unnecessary due to the fact that the seven factors are all pretty easily labeled. We name Factor One "Size" , Factor Two "Living Quarters", Factor Three "Voter Turnout", Factor Four

"Political Affiliation", Factor Five "Population Characteristics", Factor Six "Family Structure", Factor "Seven Drink or be a Republican". These seven factors account for 79.7% of the variance. See Table 2 in appendix for Factor Loadings matrix of all seven factors.

Discriminant Analysis Theory

What discriminant analysis does is to assign objects to previously defined groups. The process of classification, which is what we utilize, defines guidelines such that when followed one can determine which group an object belongs to. In our analysis, we are only concerned with the case of two groups, π_1 and π_2, where $\pi_1 \equiv MN(\underline{\mu}_1, \Sigma_1)$ and $\pi_2 \equiv MN(\underline{\mu}_2, \Sigma_2)$ are two multivariate normal populations.

We distinguish between our two groups based upon the values of our random variables $\mathbf{X}^{\mathbf{T}} = [X_1, X_2,\ldots, X_p]$, where each group's values for each variable differ to some degree. Each group has a population consisting of the values of its variables defined by a probability density function $f_1(x)$ or $f_2(x)$. The above mentioned guidelines are developed via a training sample. Two regions are formed, R_1 and R_2. The training sample splits the majority of the original sample into two known or correctly classified (by characteristics) regions and then each region R_1 and R_2 is associated with the group, π_1 and π_2 respectively. The remaining sample, n minus the size of training sample, is called the test sample. This is used to test the validity of the classification rule formed by the training sample.

Our first step in performing Discriminant Analysis is to check to see whether or not our covariance matrices, $\sum_1$ and $\sum_2$, from our two group model are equal. We check the equality of our covariance matrices in order to know if we could apply linear Discriminant Analysis or Quadratic Discriminant Analysis. We check the equality of the covariance matrices by testing the null hypothesis H_0: $\sum_1 = \sum_2$ against H_1: $\sum_1 \neq \sum_2$. To test the null hypothesis, we evaluate the pooled unbiased estimate of the common covariance matrix under H_0, which is given by

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

where N_i is the sample size of group i and S_i is the i^th^ sample covariance matrix. After evaluating S_p, we calculate the test statistic for the equality of the covariance matrices, which has a chi-square (χ^2) distribution and is equal to M/c.

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

$$1/c = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left[\left(\sum_{i=1}^{2}\frac{1}{n_i - 1}\right) - \frac{1}{n_1 + n_2 - 2}\right]$$

Incorrect classification sometimes does occur in discriminant analysis due to the fact that the characteristics or variables of the two populations may not always be readily distinguishable. Some contributing factors to misclassification are incomplete knowledge of future performance, exhaustion of the object required for faultless information, and the event of information not being readily accessible. The guidelines followed for classification should minimize the frequency of a misclassification occurring. When determining guidelines one must look at factors such as prior probabilities and the cost of misclassification. To minimize the expected cost of misclassification (ECM) one would want the following to hold for each region:

$$R_1: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right),$$

where c is the cost that an object is misclassified and p_1 and p_2 are the prior probabilities for π_1 and π_2. The left side of the inequalities is known as the density ratio. Under a multivariate normal population, the rule for assigning an object to either group becomes: allocate $\mathbf{x}_0$ to π_1 if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right].$$

Allocate $\mathbf{x}_0$ to π_2 otherwise. Another method used to attain optimal classification would be to minimize the total probability of misclassification (TPM):

$$TPM = \alpha = 2\Phi\left(-\frac{1}{2}\Delta p\right),$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal and

$$\Delta p = \sqrt{(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)}\ .$$

Discriminant Analysis Application

We classify each city in our data as high or low in crime rates according to the national average, which is 506.1 violent crimes per year per 100,000 people in US cities. The cities considered high are above the national average and the cities considered low are below the national average. Thus, our two group model of Discriminant Analysis follows this classification. According to the U.S. violent crimes national average, we classify 54 cities low and 46 cities high.

In our original data our first step in performing Discriminant Analysis is to check to see whether or not our covariance matrices, $\sum_1$ and $\sum_2$, are equal. Our test statistic M/c is equal to 346.6257. We compare our test statistic to the chi-square value we get from the chi-square table to know if you could accept or reject H_0. We want α, the chance of type one error, to be 0.01 and ν, the degrees of freedom, to be equal to ((P(P+1))/2 = 105 for our χ^2. So, we have $\chi^2_{105,\,0.01} = 138.49$ and a p-value of approximately zero. Since our test statistic is greater than the critical value, $\chi^2_{105,\,0.01}$, we reject H_0, which implies that the covariances matrices are not equal. In our case, they are not. This determines if we are able to use linear discriminant analysis versus quadratic discriminant analysis. Due to the fact of the covariance matrices of our two groups are not equal, we run quadratic discriminant analysis on our data. We use our findings from both principal components and factor analysis to perform quadratic discriminant analysis. Although our covariance matrices are not equal, we decide to run linear discriminant analysis as well. We calculate the TPM for the linear discriminant analysis to have an idea of the misclassification in our training and test samples. The TPM shows how much the two groups overlap. A high TPM tells us that the two groups overlap a lot and gives a greater probability of misclassification. For our data, the TPM is 36.1% which implies a high error rate when using linear discriminant analysis.

For principal components, we utilize the 100 x 6 data matrix which shows the correlation between each one of our samples and each particular principal component. We randomly chose 25 cities from our original sample of 100. We then separate these cities along with their information provided by principal components from the remaining 75 cities. The 75 remaining cities, better known as the training sample, are already correctly classified as having high or low crime rates. Minitab uses this information along with the differences between variables in the two groups to classify the cities in the test sample as having high or low crime rates. We determine how well our rules predict our variables by calculating the apparent error rate:

$$AER = \frac{n_{1M} + n_{2M}}{n_1 + n_2},$$

where n_1 is the sample size for π_1, n_2 is the sample size for π_2, n_{1M} is the number of π_1 items misclassified as π_2 items, and n_{2M} is the number of π_2 items misclassified as π_1 items. For our training sample Minitab misclassifies only one out of 75 cities yielding an apparent error rate of 1.3%, hence the percent correct is 98.7%. In the test sample, Minitab misclassifies four out of 25 cities yielding an AER of 16%. It misclassified Austin, TX; Sugarland, TX; Colorado Springs, CO; and Virginia Beach, VA. Refer to tables 4 through 8 in the appendix for complete PCA discriminant analysis results.

Using factor analysis information, the steps for performing discriminant analysis are almost identical to those for principal components. The only exception is the fact that we use the 100 x 7 data matrix, which is our raw data multiplied by the factor loadings for seven factors. The rest of the steps for the process remain the same. Astoundingly, the results for our factor analysis training sample are identical to those we obtain from principal components. In the test sample only three out of 25 cities were misclassified yielding an apparent error rate of 12%. Austin, TX; Colorado Springs, CO; and Virginia Beach, VA were misclassified. See tables 9 through 13 in the appendix for complete results.

Although our covariance matrices are not equal, we decide to run linear discriminant analysis anyhow. Surprisingly enough, we obtain amicable results identical to those we get using quadratic discriminant analysis.

Results

With the help of the multivariate methods of Principal Components Analysis and Factor Analysis, we reduce the fourteen distinct variables that can affect the crime rate of a city to 6 and 7 important variables that show a high correlation with all the 14 variables. Although we use both Principal Components Analysis and Factor Analysis to accomplish the reduction of variables, there are some advantages and disadvantages to each method. Principle Components Analysis did decrease the number of variables to 6 and accounted for 86% of the total variance, while Factor Analysis decreased the number to 7 and accounted for 79.7% of the total variance. Using the Factor Analysis results, we obtained the better results in Discriminant Analysis prediction of a city's crime rates. Even though both Principal Component Analysis and Factor Analysis showed a 1.3% Apparent Error rate in the training sample for Discriminant Analysis, Factor Analysis gave a 12% Apparent Error rate in the test sample while Principal Component gave a 16% Apparent Error rate. These rates indicate that our rules do a good job in classifying cities as dangerous or safe.

Conclusion

Even though we obtained exceptional results, there is always room for improvement in research. We could analyze additional variables such as: percent of households that own a computer, percent of households that own at least two vehicles, cost of living, percent of population involved within their community, available funding for extracurricular activities, as well as others. We could extend this research to the state level to find new results. With these improvements our research could eventually give results in classifying a territory as high or low in crimes and determine variables which could help in this classification process.

Safety is the most important issue for everyone in the U.S., let alone everyone in the world. We now see that we, as Americans, are willing to give up some of our freedom to make sure that our children grow up in a safe environment. With the help of the multivariate statistical methods of Principal Component Analysis, Factor Analysis, and Discriminant Analysis we have created a method that helps to classify a city as safe or unsafe without using the actual crime rate statistics of the city. We have shown certain variables such as population, household characteristic, voter turnout, political affiliation, unemployment rates, among others are suitable for analysis to classify a city as safe or unsafe and to help improve the safety of the city. City Councils will be able to decrease the crime rates of their cities by increasing or decreasing certain values of our variables.

Acknowledgements

We would like to take this time to thank all those who helped us in completing this research. We would like to thank our sponsors, NSA and NSF, for funding SUMSRI to make this research possible. Without their support we would not be here or even have the opportunity to do this research. We would like to thank Dr. Vasant Waikar for teaching us what we needed to know to perform this research and for his guidance in our methods of conducting research. We would also like to thank our Graduate Assistant James Lawrence and our fellow Statistics researchers, Sara Blight and Nick Imholte, for their support and dedication throughout this program. In addition, we want to thank Dr. Tom Farmer for revising our paper and providing suggestions to help us achieve the mathematical research paper level. Finally, we would to thank our professors, graduate assistants, other SUMSRI participants and faculty, and all SUMSRI supporters for their help in making this experience an enjoyable and unforgettable part of our lives and careers.

Bibliography

[1] Barbara Ryan, Brian Joiner and Jonathon Cryer, *Minitab Handbook Updated for Release 14 (Fifth Edition)*, Brooks/Cole, a division of Thomson Learning, Inc. (2005).

[2] Bestplaces.net, *Crime rates between two cities*. Retrieved June 30, 2004 from http://www.bestplaces.net/crime/crcompare.aspx

[3] Census.gov, *2000 United States Cities Group Quarters Population by Age and Sex: Institutionalized Population*. Retrieved June 26, 2004 from http://factfinder.census.gov/servlet/SAFFPeople?_sse=on

[4] Census.gov, *2000 United States Cities Household Population and Housing Characteristics*. Retrieved June 26, 2004 from http://factfinder.census.gov/servlet/SAFFHousing?_sse=on

[5] Census.gov, *2000 United States Cities Profile Reports*. Retrieved June 25, 2004 from http://factfinder.census.gov/servlet/SAFFFacts?_sse=on

[6] David Rapp, *Politics in America 2002 the 107th Congress,* Congressional Quarterly Inc. (2001).

[7] Donald F. Morrison, *Multivariate Statistical Methods (Fourth Edition)*, Brooks/Cole, a division of Thomson Learning, Inc. (2005).

[8] Infoplease.com, *Profiles of the 50 largest cities of the United States.* Retrieved June 22, 2004 from http://www.infoplease.com/ipa/A0108477.html

[9] James M. Lattin, J. Douglas Carroll and Paul E. Green, *Analyzing Multivariate Data,* Brooks/Cole, a division of Thomson Learning, Inc. (2003).

[10] Moving.com, *Days of Sunshine*. Retrieved June 28, 2005 from http://moving.monstermoving.monster.com/Find_a_Place/Compare2Cities/index.asp

[11] Ncsl.gov, *2000 Voter Turnout*. Retrieved June 29, 2004 from http://www.ncsl.org/programs/legman/elect/00voterturn.htm

[12] Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis (Fifth Edition),* Prentice Hall (2002).

[13] Web-lexis-nexis.com, *Apparent Alcohol Consumption for States, census regions and United States, 1998.* Retrieved June 30, 2004 from http://web.lexis-nexis.com/statuniv/doclist?_m=0303cfc4327e6bdf4a0a31a3ee728632&wchp=dGLbVlz-zSkVV&_md5=3ded29a3084545bdcbf96a4a28049148

[14] World Gazetteer, The, *Cities Population*. Retrieved June 24, 2004 from http://www.gazetteer.de/fr/fr_us.htm

[15] Yahoo.com, *Yahoo! Yellow Pages: Religion and Spirituality*. Retrieved June, 2004 from http://yp.yahoo.com/py/yploc.py?clr=ypBrowse&ycat=8104735&desc=Religion+and+Spirituality&tab=B2C&country_in=us

Appendix

Table 1: Normality Tests of Scaled and Transformed Data

Variables	r-values
(log)population	0.946044
% of Pop. between 5 and 65	0.994485
(squared)Ethnicity	0.978264
5sqrt's(# of churches)	0.978775
(log)Unemployment	0.981835
(log)income	0.973653
Sqrt(ln)# in household	0.982853
% of single parent homes	0.990454
%25 or older w/ bach. Degree	0.987421
log(log(log))Voter turnout	0.982344
sqrt(log(%of pop. Institutionalized +1))	0.952365
log(days of sunshine)	0.992472
log(alcohol cons.)	0.936483
% of Rep. off.	0.991968

Table 2: Factor Loadings

Variable	FL 1	FL 2	FL 3	FL 4	FL 5	FL 6	FL 7
% of Rep. St. Officials	0.053	-0.140	0.011	0.761	-0.106	0.344	-0.289
% w/ bachelor degree	-0.660	-0.242	0.176	-0.375	0.358	0.200	-0.160
% of single parent households	0.642	-0.029	-0.247	0.224	-0.202	-0.533	-0.184
Percent Population f/ 5 to 65	-0.257	-0.489	0.020	-0.191	-0.106	-0.017	-0.212
(Log)Per Capita Income	-0.725	-0.470	0.278	-0.246	0.077	-0.096	-0.165
(Log)Alcohol Consumption	0.057	0.070	-0.316	-0.059	0.154	-0.280	0.292
(Log)Days of Sunshine	-0.032	-0.225	-0.448	0.235	-0.248	0.114	0.133

			High	8.914	74.221