MATH 437/598 Applied Multivariate Statistics
Final Project Poster Session

Tuesday, May 8th 1-3 pm
GRL Conference Room
(across from Geology Museum)

Titles, abstracts, and authors of each project are listed in this booklet, sorted by student classification (undergraduate vs. graduate). Each student’s major field of study is given in parentheses next to his or her name.

The project number assigned to each project corresponds to the number on the poster score sheet and also to the order in which posters will be displayed around the room.

Contents

1  Undergraduate Student Projects  

2  Graduate Student Projects  

2

3
1 Undergraduate Student Projects

PROJECT 1
Author(s): Oscar Aguilar (STAT), Guo Deng (STAT), and Andrew Zerwick (STAT)

Title: Clustering Microarray Data via Gene Shaving

Abstract: The human body has between 20,000 to 25,000 genes. Each gene can be overproducing, underproducing or be at its normal level of production of cells in the human body. Under or over production can lead to the incursion of certain diseases or cancers. While monitoring each individual gene is generally not useful, genes can be grouped together on platforms called microarrays, and with these groupings certain diseases and cancers can be analyzed. With these microarrays, scientists are able to diagnose diseases at an earlier stage or find drugs to help better cure the disease. In this project, we analyzed microarray data from the National Cancer Institute. There are 5,244 genes and 61 samples in this data set; the samples are derived from tumors with different sites of origin: 7 breast, 5 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non-small-cell lung carcinoma (NSCLC), 6 ovarian, and 9 renal. We determined the number of clusters using three different approaches. The first using the k-means algorithm, second the L-method, and lastly the GAP statistic. The gene shaving method is a technique used for identifying distinct sets of genes with similar expression patterns. After clustering, we found in the heat map of cluster seven that the three types of cancers with the highest average intensity of the genes in the cluster were from three separate sites of origin: lungs, central nervous system and colon, which indicates a possible hidden relationship between these three cancers.

PROJECT 2
Author(s): Yongda Chen (STAT) and Zhen Li (STAT)

Title: Determining the Effectiveness of Classifying Diabetes with Principal Components

Abstract: Data analysis is no strange task in the field of Applied Statistics, and while it is usually the norm for statisticians to work with difficult data, a complete dataset will always be appreciated, especially when it can be explained by a fewer number of variables. In this project, an incomplete set of diabetes data will be used to illustrate our attempt at estimating the missing values. With the “complete” dataset in hand, we will then calculate its principal components (PC’s). Taking only the bare minimum number $k$ of PC’s that capture the majority ($90\%$) of the system’s total variability, the estimated expected actual error rate $\hat{E}(AER)$ can be calculated using Lachenbruch’s holdout method for both the $k$ PC’s and the $p-1$ variables for determining diabetes. If the error rates differ very little, then we will be able to conclude the validity of classifying diabetes using $k$ PC’s of the initial variables, where $k < p - 1$. 

2
2 Graduate Student Projects

PROJECT 3

Author(s): Emma Nicoletti (STAT) and Ashley Valdez (STAT)

Title: US Colleges and Universities: Who, Where, and How Big

Abstract: Using two data sets from 1995 (American Association of University Professors, and US News and World Report), we analyzed 1161 schools using factor and cluster analysis, looking for the primary ideas that lead to school groupings. In factor analysis, the five underlying concepts we identified are size, exclusivity, knowledge accessibility, region, and test scores from incoming students. We confirmed those as important grouping by performing cluster analysis on the variables of the data. Cluster analysis of the schools split them into 13 clusters, most of which were described by some combination of region, cost, size, and test scores. Overall, it became clear that school size, region, and the incoming test scores were the most influential variables, but the two analysis methods provide different types of information regarding schools and should be used with a specific purpose in mind.

PROJECT 4

Author(s): Ashley Bell (STAT)

Title: Bivariate Quality Control of Historical Radiosonde Winds

Abstract: In the atmospheric sciences, robust and accurate observational datasets are often needed to validate the climate models. A dataset that is often overlooked for this purpose is the Upper Air Network Archive, which contains historic radiosonde observations. The archive contains many important variables that with proper quality control can be of great use to many scientists. This study aims to identify techniques that will accurately and automatically identify outliers contained in the radiosonde wind variable (U, V). This multivariate variable is more difficult to quality control (QC) as the U and V components should be considered jointly, and such QC methods have never been applied to the radiosonde winds. To determine the best method for outlier identification, an extensive simulation study will be developed based on a bivariate skew-t model of winds. Data will be simulated based on parameters and autocorrelation that are estimated from observed wind data and then contaminated with known outliers. Results of the exploratory analysis of the wind variables show significant autocorrelation between pressure levels of the U and V components, and a radiosonde launch station to base the described simulation on has been identified.
PROJECT 5

Author(s): Kelly Lundstrom (STAT)

Title: Factor Analysis with Binary Indicators: An Overview and Bootstrap Study

Abstract: Factor analysis is a statistical method that is most commonly used with quantitative data, but this method can be applied to binary data with a modified approach that uses the tetrachoric correlation matrix. The Stata and product moment approximations of the tetrachoric correlation matrix are analyzed in this paper, and factor analysis is completed using the principal component and weighted least squares methods. Currently, Mplus is recognized as the superior statistical program to perform factor analysis with binary data because of the unique weighted least squares method it employs, which performs well with varying sample sizes and degrees of non-normality. In this study, ten bootstrap samples are taken from an attitudes survey with “yes” and “no” responses, and average bootstrap factor loadings for three different methods are compared to the average bootstrap Mplus factor loadings. In addition, the bootstrapped Mplus tetrachoric correlation matrices are compared to the bootstrapped Stata and product moment tetrachoric correlation matrices. The study revealed that the Stata method is superior to the product moment approximation of the tetrachoric correlation matrix; additionally, applying the principal component method of factor analysis to the Stata approximation resulted in the closest loadings to the Mplus factor solution, although the Mplus solution is still preferred because it produces better loadings when compared to the other methods.

PROJECT 6

Author(s): Sarah Valovcin (STAT)

Title: The Impact of Residential Household Characteristics on Electricity Consumption Using PCA and MCA Techniques

Abstract: Using the U.S. EIA’s Residential Energy Consumption Survey from 2005, we analyze the impact of household characteristics on annual electricity consumption. First, we reduce the dimensionality of the problem by performing a principal component analysis on the quantitative variables and a multiple correspondence analysis on the categorical variables. Then a multiple linear regression is done, where the principal components and transformed categorical variables are the predictors, and annual electricity consumption (in Btu) is the response.
PROJECT 7
Author(s): Xiaopeng Lin (STAT)

Title: Cluster Analysis of Distributions of Silver Nanoparticles and Dissolved Silver Ions in Solution

Abstract: Engineered silver nanoparticles are rapidly becoming a part of our daily life in the form of cosmetics, food packaging, drug delivery systems, therapeutics and biosensors, etc. These silver nanoparticles end up in the environment from waste disposal which introduces a potential threat to the ecological system. Due to limitations in the detection of silver nanoparticles from dissolved forms of silver, it is difficult to evaluate the environmental impacts caused by them. Different statistical methods have been studied and applied to distinguish between dissolved silver and silver nanoparticles. In this project, a model-based cluster analysis by Mixtools package in R was applied to four datasets, seeking to provide a method to separate multivariate distributions for different forms of silver. Data from samples with different sized silver nanoparticles were clustered into different distributions successively and were assigned probabilities of belonging to each cluster. Regression analysis shows good consistency between the specified fitted model and corresponding data.

PROJECT 8
Author(s): Karen Kazor (STAT)

Title: Analysis of Radius Intersection Multidimensional Scaling (RIMDS) Methods with Application to Identifying Crime Series

Abstract: This analysis evaluates the Radius Intersection Multidimensional Scaling Methods (RIMDS1 and RIMDS2) proposed by Dr. Amanda S. Hering and Mr. Sean Bair in their report “Bayesian Space-Time Point Process Models For Individually Identifiable Criminal Patterns.” The main objective of this study is to compare the efficiency of RIMDS1 and RIMDS2 to classical multidimensional scaling (MDS) based on measures of computational intensity and accuracy. We apply each method to a set of 48 robberies consisting of 6 known crime series and make comparisons based on the CPU run time for each method as well as the Stress, Scatter, and Wilk’s Lambda of the resulting configurations. The RIMDS2 method preforms reasonably well in replicating the dissimilarities structure relative to the classical MDS result. It fails however to distinctly group like crimes as measured by Wilk’s Lambda and thus frequently produces undesirable configurations. While RIMDS1 produces configurations with high Stress, these configurations have considerably lower values of Wilk’s Lambda than either the RIMDS2 or classical MDS results in the case of sparse data. Overall, classical MDS is more efficient in terms of system CPU run time when computing an initial configuration of points; however, as crimes are added to the initial dataset, the RIMSD1 and RIMSD2 algorithms surpass the efficiencies of classical MDS.
PROJECT 9
Author(s): Megan Yoder (STAT)

Title: Exploratory Analysis of Spatial Characteristics of Borehole Data for Tunnel Design and Construction

Abstract: Borehole investigations are important in geotechnical engineering for providing information on subsurface soil conditions that would otherwise be unknown. Incorporating a statistical analysis of borehole data can increase knowledge of subsurface soil parameters and lead to better decision making by project planners and engineers. The data analyzed in this project is from the Toronto-York Spadina Subway Extension (TYSSE) project, which is an ongoing tunneling project. A description of the data available is given, and attributes of interest are discussed. Empirical semivariograms are explored for one of the subway station locations, and questions are raised as to how to proceed in further research.

PROJECT 10
Author(s): Ady Van-Dunem (Mining & Earth Systems Engineering)

Title: Beyond PCA: The Min/Max Autocorrelation Factors Approach to Ore Resource Estimation

Abstract: Most ore reserve estimation studies presented in the literature will attempt to document and elaborate on the benefits of a particular suite of tools applied to the goal of accurate characterization of a given orebody. This characterization typically considers the spatial distribution of grades of a single element of interest in the deposit. The truth is that the great majority of mineral deposits are multi-element deposits which, given the nature of geological processes, will tend to have some correlation among them. In this project there will be two central multivariate statistical methods documented: (1) principal components analysis (PCA) and (2) min/max autocorrelation factors (MAF). Specifically, the application of the min/max autocorrelation factors approach to a real data set will be documented and contrasted to principal component analysis in a nickel laterite ore deposit.

PROJECT 11
Author(s): Loren John (STAT)

Title: Clustering Technique to Determine Similarities Among Telecommunications Sites for a Particular Wireless Carrier

Abstract: The purpose of the study performed was to analyze grouping and clustering techniques and apply them to telecommunications sites given specific variables. It is shown that through each site’s performance and the linkage amongst them, we can provide future modification predictions for network advancement.
PROJECT 12

Author(s): Brent Putman (Geophysics)

Title: Determining Correlation between Contaminated Vapor Intrusion and Local Weather in a Residential Setting

Abstract: Vapor intrusion in a residential setting can be dangerous to a building’s inhabitants. The intrusive vapor can be toxic if the surrounding groundwater has been contaminated with volatile chemicals. This study attempts to establish a correlation between the level of vapor intrusion and external weather variables. I show that weather variables can affect the pressure differential across a building’s external boundary, thereby changing the magnitude and direction of vapor flow. I also show the time lag difference (phase shift) between changes in weather variables and the vapor intrusion signal. This result leads to the conclusion that there are several processes that affect vapor intrusion, each of which acts on different phase and frequency scales.