Project 1: Probability Distributions and Bayesian Networks

CSE474/574: Introduction to Machine Learning (Fall 2015)

ADHIP VIHAN

Person Number: 50134774

PROBLEM INTRODUCTION

The project mainly is concerned with finding the probability distributions of the given data. It involves calculating the mean, variance, standard deviation of the univariate distribution for given 4 variables namely CS Score, Research Overhead, Admin Base Pay, Tuition, using the simple MATLAB functions. Using these statistics constructing compact representations of joint probability distributions called Bayesian Networks.

Given Dataset

The dataset that was given to us comprised of four variables.

- **CS-ranking-score:**

Each value corresponds to the score of a public university according to a survey. All being subset of the top 100 Computer Science graduate programs in US according to US News and World Report.

- **Research-Overhead (percentage):**

These values were the portion of research grants retained as infrastructure/administrative
costs by the university.

- **Administrator Base Salary :**

Base salary of administrators (Dollars).

- **Tuition (Out-of-State) :**

Out Of state tuition fee for the given 49 colleges.

- **No. of CS Graduate Students in Fall 2015**

Features and the Calculations

1. Initially the concentration of the project was on calculating the mean, covariance and standard deviation of the data. We used MATLAB software functions to calculate the values .The variables used for storing the mean, variance and Standard deviation of the data were
   mu1, mu2, mu3, mu4
   var1, var2, var3, var4 And sigma1, sigma2, sigma3, sigma4

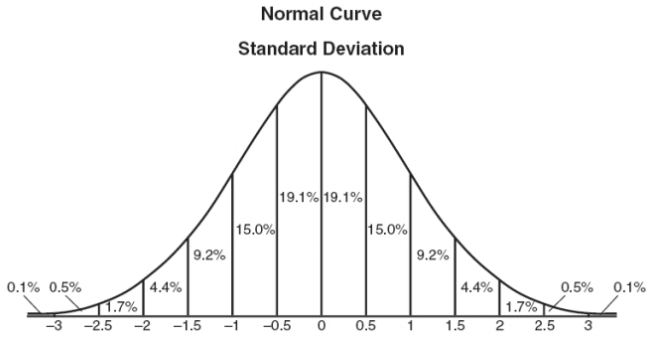
2. Subsequent part of the project involved Computing for each pair of variables their covariance and correlation and representing the result in Covariance and correlation matrix. The value were stored in
   covarianceMat , correlationMat.

3. Nextly the log likelihood of the data was calculated. Since each variable normally distributed and they are independent of each other. Mean and variance calculated previously were used to calculate this value and stored in
   logLikelihood.

4. Correlation calculated in the previous steps are used to form a Bayesian network looking at the variables that are most correlated to each other with final aim of getting a log likelihood which is greater than the log-likelihood calculated in the 3rd step.

Models and Techniques Used

Firstly our data was normally distributed. Normally distributed means arrangement of a data set in which most values cluster in the middle of the range and the remaining values are distributed symmetrically on the either side of the mean. Normally distributed variables follow simple Bell Curve which is given below.



- **Mean**
  Calculation of mean was done with the help of ***mean()*** function of MATLAB. Mean is nothing but the average of the numbers. Calculated using the formula

  $$\overline{X} = \frac{\Sigma X}{n}$$

- **Variance**
  Calculation of variance of the data was calculated using the ***var()*** function of MATLAB. ***var()*** takes as input the data set which is X,Y,ZZZ in this case. Variance was calculated using the formula.

  $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$ Where μ stands for the mean of the data.

- **Standard Deviation**
  Calculation of standard deviation was done by using the std() function of matlab. Standard deviation is calculated by the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- **Correlation:**
  Correlation means the degree to which two or more attributes or measurements on the same group of elements show a tendency to vary together. It is calculated by the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

  It was calculated by the function ***corrcoef()***.
  Firstly the data given was saved in the form of a 1-by-4 vector and the vector was passed into the ***corrcoef()*** function to calculate the 4*4 Correlation matrix.

	1	2	3	4
1	1	0.4655	0.0482	0.2794
2	0.4655	1	0.1575	0.1496
3	0.0482	0.1575	1	-0.2453
4	0.2794	0.1496	-0.2453	1

- **Covariance**

  Covariance is a measure of how much two random variables change together. It is calculated by the formula

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

The data given was saved in the form of 1-by-4 vector and then passed into the ***covar()*** function of MATLAB to calculate the 4*4 Covariance matrix of the data.

	1	2	3	4
1	0.4575	1.1184	3.8798e+03	1.0585e+03
2	1.1184	12.6161	6.6652e+04	2.9758e+03
3	3.8798e+03	6.6652e+04	1.4190e+10	-1.6369e+08
4	1.0585e+03	2.9758e+03	-1.6369e+08	3.1368e+07

- **LogLikelihood of the data:**

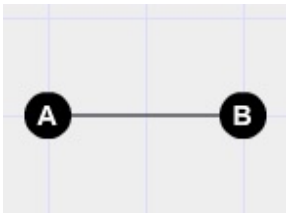
  What is likelihood?
  Likelihood according to mathworld.com website is the hypothetical probability that an event that has already occurred would yield a specific outcome. Likelihood is calculated by taking the probability distribution of the data and taking the log of the distribution. After that the sum is taken and stored in a variable.

  It was calculated by this function of MATLAB:

  sum(log(normpdf(DATA,MEAN,Covariance)))

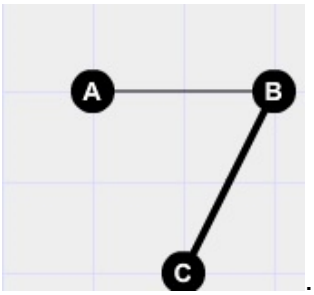
- **Constructing a Bayseian Network from the given values**
-

Taking a look at the correlation matrix we see that Variable A is most correlated with variable B since the correlation value is the greatest.



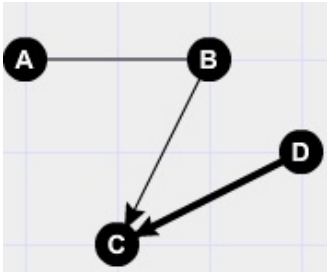
Now for the next connection we have is B is more correlated to C (considering positive

correlation). That gives us the graph

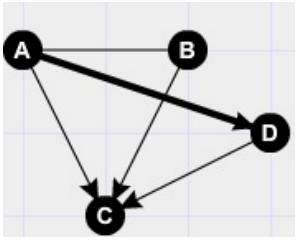


.

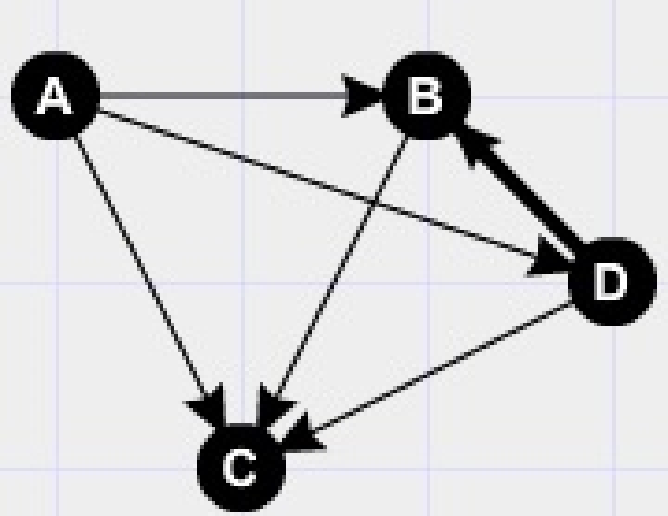
We look next in the series and see that C is most correlated to the D so making a connection from D to C we have.



A is next most correlated to the variable D and then to variable C. And the Result is going to be a Fully-connected Graph. So Constructing graph for these two we get



Now coming to Node D. Using the Correlation matrix we observe D is Most Correlated to B after C. hence, making the connection we have:

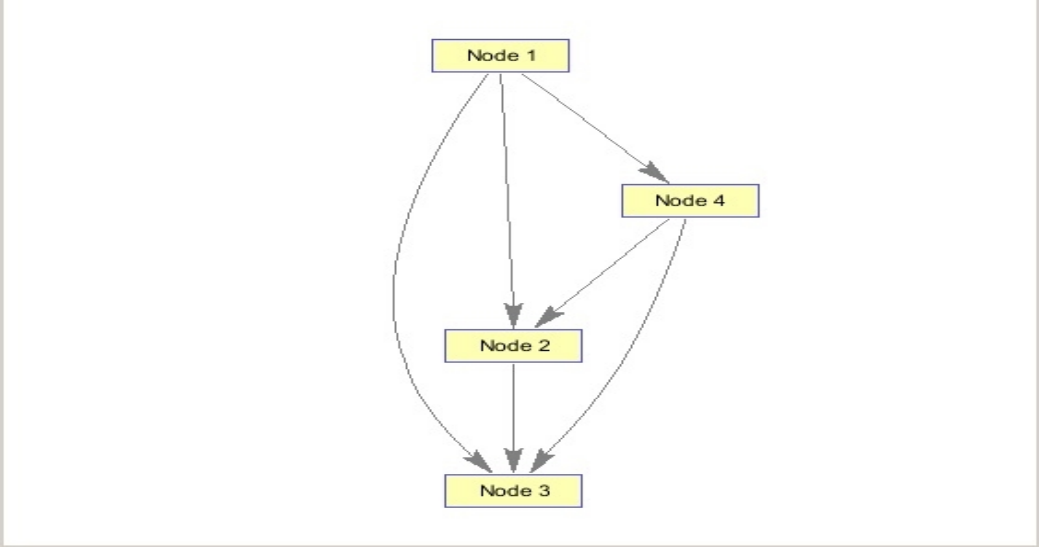


For the given Bayesian network we now construct a 4-by-4 matrix representing the acyclic directed graph showing the connections of the Bayesian network. Each entry of the matrix takes value 0 or 1.

So constructing a matrix:

	A	B	C	D
A	0	1	1	1
B	0	0	1	0
C	0	0	0	0
D	0	1	1	0

The Bayesian Network that comes out to be represented by the function ***view(Biograph()) is:***



Now to create the log likelihood of the given data we have to calculate the probability distributions. Using the formula:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Here we calculate the multivariate probability by using the ***mvnpdf()*** function of MATLAB and the univariate probability by using the ***normpdf()*** function of MATLAB.

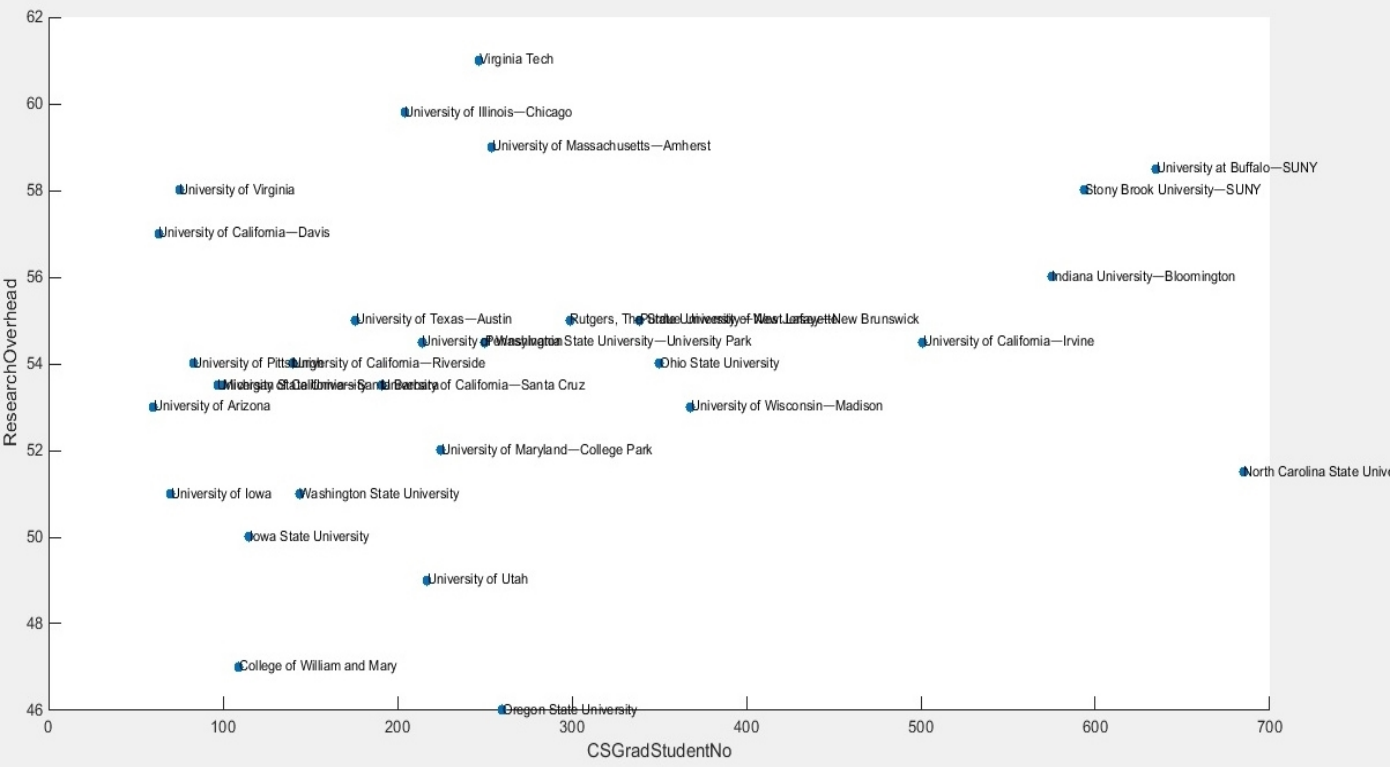
For the given Matrix We have the equation in the form of:

P(G)=P(A)*P(B|A,D)*P(C|A,B,D)*P(D|A)

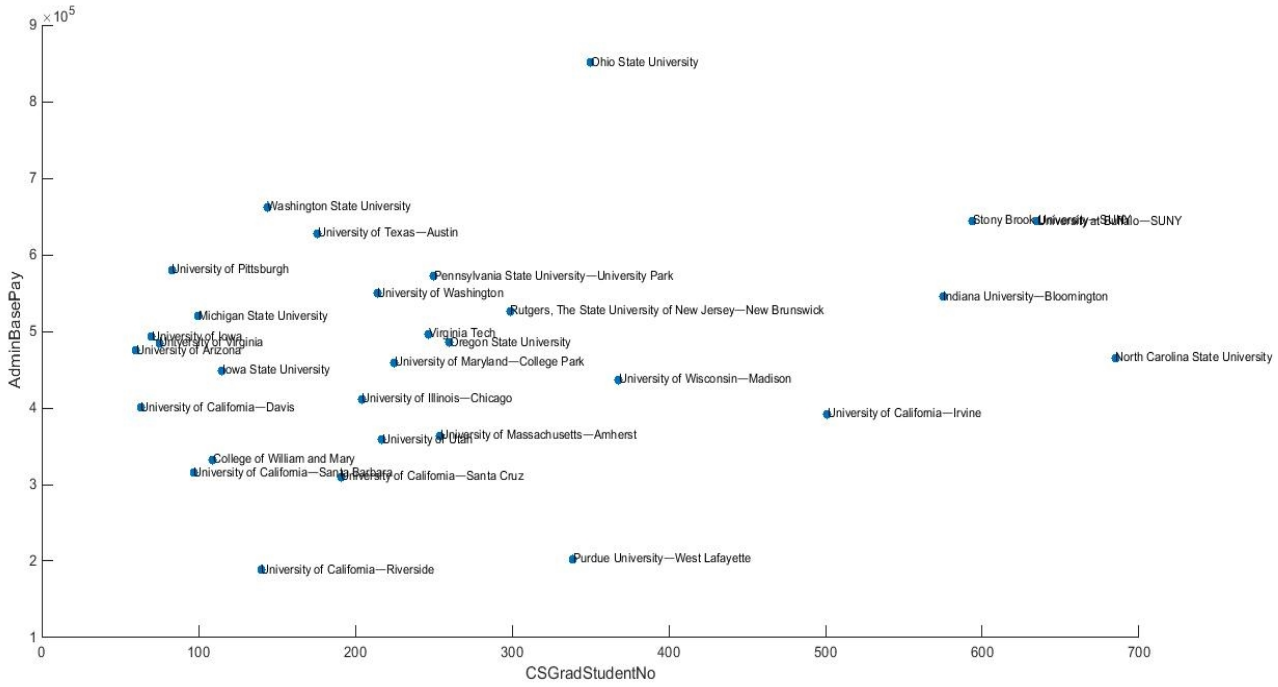
Which comes out to be: **-1.304092369999620e+03**

Results:

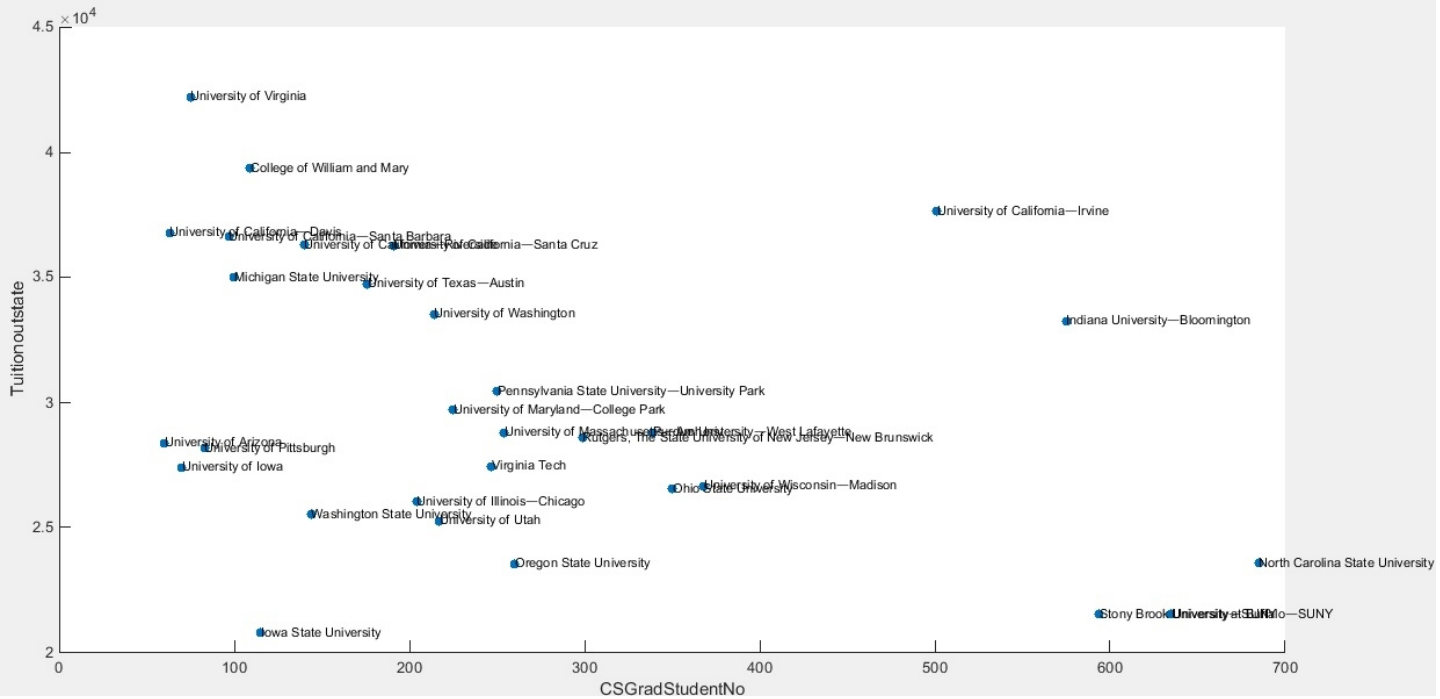
1. **CSGradStudentNo vs ResearchOverhead Plot**
   Function used: *scatter()*



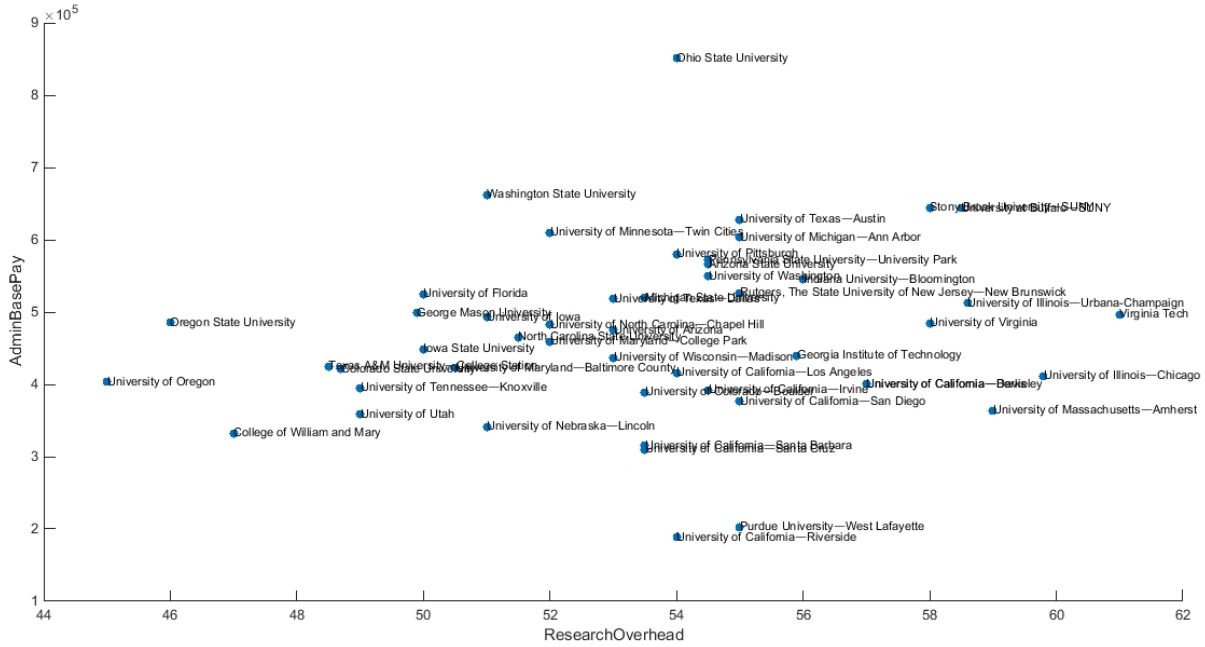
2. CSGradStudentNo Vs AdminBasePay Plot



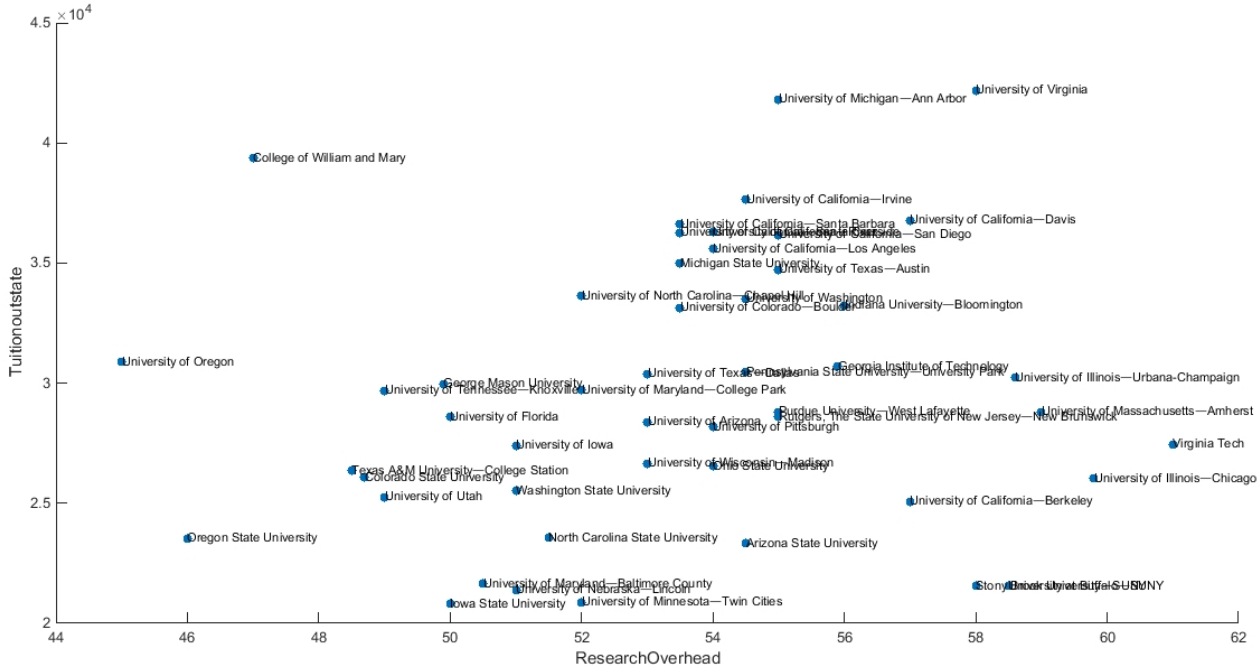
3. CSGradStudentNo vs Tuitionoutstate plot



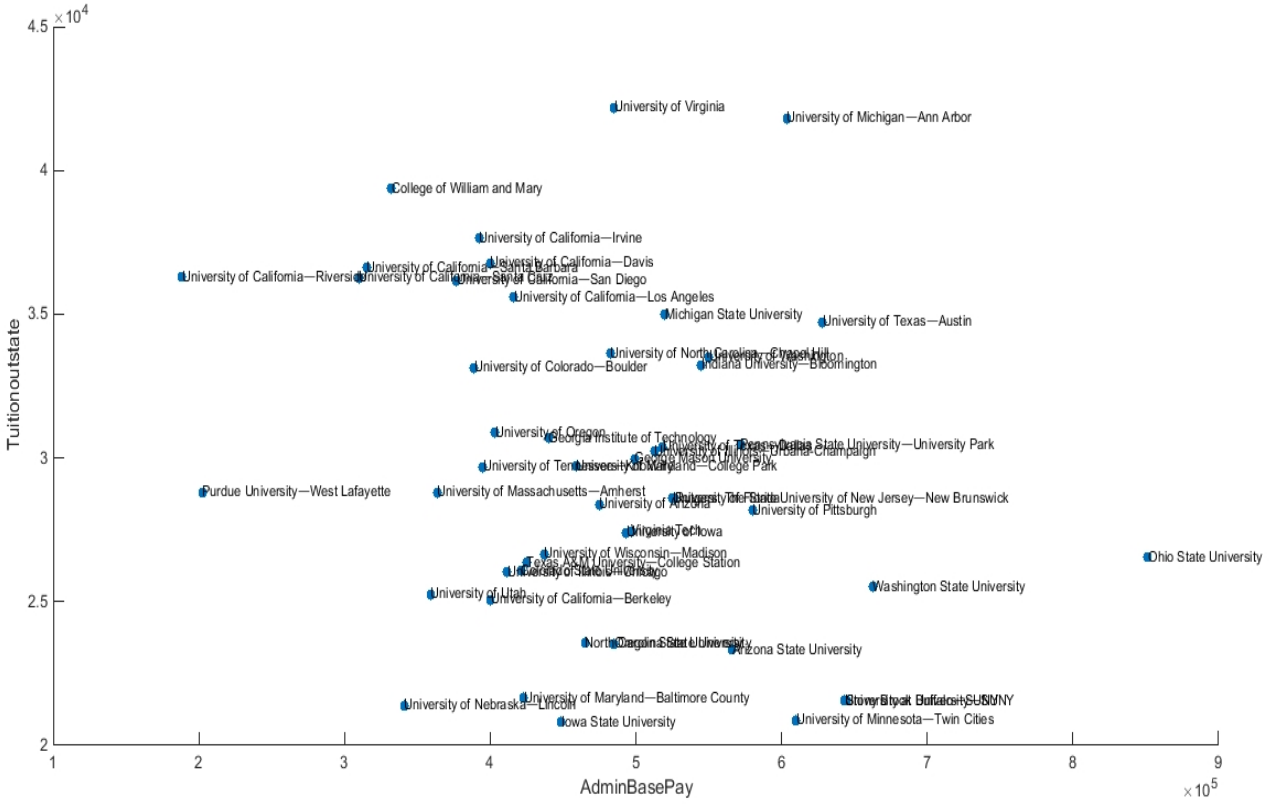
4. ResearchOverhead vs AdminBasePay Scatter Plot



5. ResearchOverhead vs Tuitionoutstate Plot



6. AdminBasePay vs Tuitionoutstate Plot



The **loglikelihood** comes out to be: **-1.314668550434506e+03**

Loglikelihood of Graph comes out to be: **-1.304092369999620e+03**

Conclusion

The project involved calculating probability distributions of several variables. MATLAB built in functions were used to calculate the mean and variance of univariate distributions and covariance and correlation coefficient of pairs of variables. The dataset given initially consist of independent variables, we observed that the log-likelihood of the data can be maximized by observing the dependency of the nodes on each other and determining the conditional probabilities. That alludes that given the optimal Network the log likelihood of the data can be optimized.

References

[1] Wikipedia: https://en.wikipedia.org/wiki/Mean,
https://en.wikipedia.org/wiki/Variance,
https://en.wikipedia.org/wiki/Standard_deviation
[2] The University Of British Columbia:
http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
[2]Mathworld: http://mathworld.wolfram.com/Likelihood.html