CSE 574: Project 1

Project Report

Submission Date: Sep 28 2015

Suramrit Singh/5016 9918/suramrit@buffalo.edu

1. Overview

Given a data set in an excel spreadsheet (UniversityData.xls), consisting of several variables, the project required the application of MATLAB to evaluate statistics, namely: mean and variance of univariate distributions and covariance and correlation coefficient of pairs of variables.

The project also involved application of the statistics to construct compact representations of Bayesian Networks and checking the goodness of these representations using the concept of likelihood

1.1 Background

Terms and Concepts:

Mean: In probability, mean gives us a central tendency of a probability distribution or a random variable.

It is represented as

$$\boldsymbol{\mu = \sum xP(x)}$$

where *x* is the value of a random variable and *P(x)* is the probability

For a data set it is defined as the sum of the values divided by the total no of values.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance: For a data set sample variance is computed as:

$$\boldsymbol{\sigma^2 = \left(\frac{1}{N-1}\right) * \sum[x(i) - \mu]^2}$$

Variance gives us a measure of dispersion of the data from the mean. A higher value of variance implies that the data set is scattered away from the mean.

Covariance: For a pair of variables X1, X2 with samples x1(i), x2(i) the covariance is given by

$$\boldsymbol{\sigma 12} = \frac{\mathbf{1}}{\boldsymbol{N-1}} * \sum[\boldsymbol{x1(i)} - \boldsymbol{\mu 1}][\boldsymbol{x2(i)} - \boldsymbol{\mu 2}]$$

Covariance gives a measure of how 2 variables change together. Since there is no square term in covariance, it can also be a negative value.

For a set of n variables **X** = {*X1,X2,….Xn*} , a covariance matrix is given as:

$$\sum = \begin{bmatrix} \boldsymbol{\sigma 1.\sigma 1} & \cdots & . & \boldsymbol{\sigma 1 d} \\ \vdots & \ddots & \vdots & . \\ \boldsymbol{\sigma 1 d - 1} & . & . & . \\ \boldsymbol{\sigma 1 d} & \cdots & . & \boldsymbol{\sigma d.\sigma d} \end{bmatrix}$$

Likelihood and Log likelihood:

The likelihood of a set of parameter values or events given certain outcomes/samples, is equal to the probability of those observed outcomes given those events.

Generally natural log of the likelihood is taken for easier estimation and analysis.
For n independent samples of a probability density, the log likelihood is given by

$$\boldsymbol{L}(\mathbf{x1}, \mathbf{x2}, .. \mathbf{xn}) = \mathbf{log} \sum \boldsymbol{p}(\mathbf{xi})$$

Data set:

The given data set for this project is multivariate with 4 variables, CS Score, Research Overhead, Base Pay, and Tuition. They have been represented by X1,X2,X3,X4 throughout the document.

2. Tools Used

Matlab and related libraries were used for the computation and analysis of the statistical data. The data used in the project was imported from an excel spreadsheet (UniversityData.xls).Wolfram alpha was also intermittently used for computations and plotting.

3. Tasks

3.1 Mean, variance and standard deviation calculation

The values for each variable were imported from the excel sheet and a resultant 49x4 matrix was made with each column containing the values of the random variables.

The mean of this matrix was calculated using: $\boldsymbol{mu} = \boldsymbol{mean}(\boldsymbol{rvMatrix})$;

Here ***rvMatrix*** is a **49x4** matrix containing the variable values.

The resultant matrix, mu, was a vector of dimension 1x4 with each column containing the mean for each variable.

After the calculation of mean, variance was calculated using: $\boldsymbol{variance} = \boldsymbol{var}(\boldsymbol{rvMat})$

The resultant matrix, variance, was a vector of dimension 1x4 with each column containing the variance for each variable.
The standard deviation was calculated taking the square root of each column of vector variance.

3.1.2 Results

The results from section 3.1.1 are tabulated below. Mu1 corresponds to mean of variable X1 and so on.
Mean:

Mu1	3.214
Mu2	53.3650612
Mu3	469178.816326531
Mu4	29711.9591836735

Tab 1

Vriance:

Var1	0.4575
Var2	12.6160629
Var3	14183920820.9031
Var4	31367695.789966

Tab 2

Standard deviation:

Sigma1	0.676387462923434
Sigma2	3.5519097574643
Sigma3	119120.614592534
Sigma4	5600.68708195396

Tab 3

3.2 Computing covariance and correlation (Matlab Implementation)

In order to calculate the covariance and correlation matrices, the following implementations were used in matlab:

For covariance , $\boldsymbol{covarianceMat} = \boldsymbol{cov}(\boldsymbol{rvMat})$;

For correlation, $\boldsymbol{correlationMat} = \boldsymbol{corrcoef}(\boldsymbol{rvMat})$;

The covariance and correlation matrices obtained are:

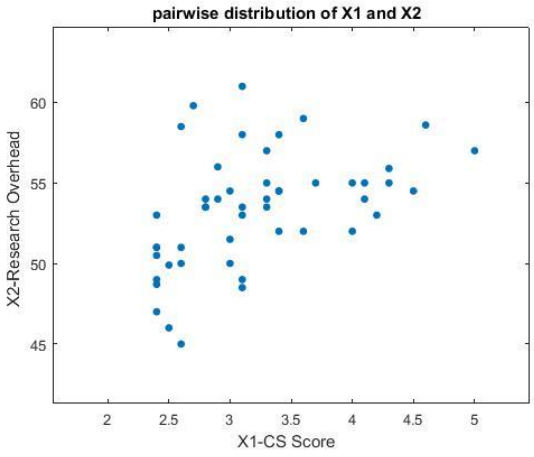
	X1	X2	X3	X4
X1	0.4575	1.11842261904762	3879.78184523809	1058.47976190476
X2	1.11842261904762	12.6160629251701	66651.6643282313	2975.82980442177
X3	3879.78184523809	66651.6643282313	14189720820.9031	-163685641.257653
X4	1058.47976190476	2975.82980442177	-163685641.257653	31367695.7899660

	X1	X2	X3	X4
X1	1	0.465530853186398	0.0481531643615803	0.279412416164465
X2	0.465530853186398	1	0.157529592573249	0.149590790502660
X3	0.0481531643615803	0.157529592573249	1	-0.245347902755565
X4	0.279412416164465	0.149590790502660	-0.245347902755565	1

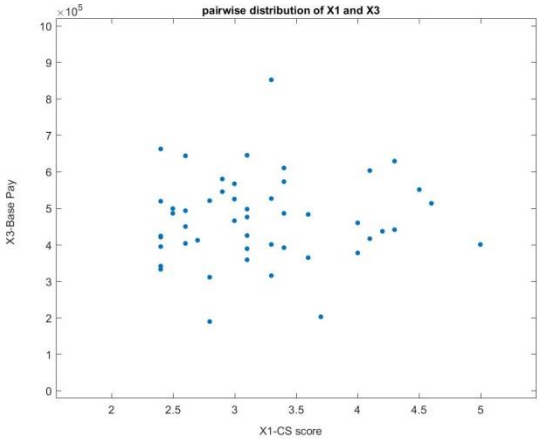
In the above tables, $(i, j)^{th}$ cell show for each variable pair X_i and X_j s the covariance and correlation values for variables respectively.

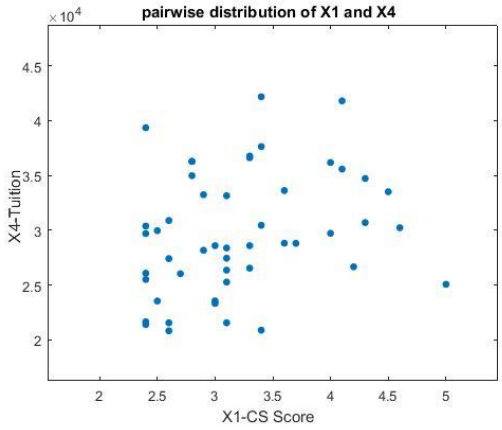
3.2.1 Pairwise data plot

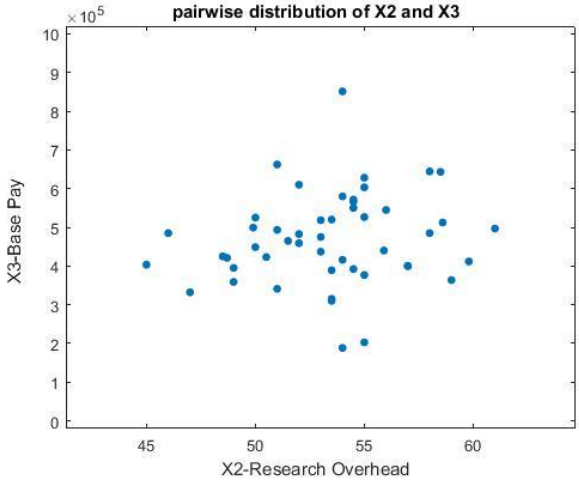
In order to understand the correlation between the variables better, pair wise data was plotted and the following plots were observed.



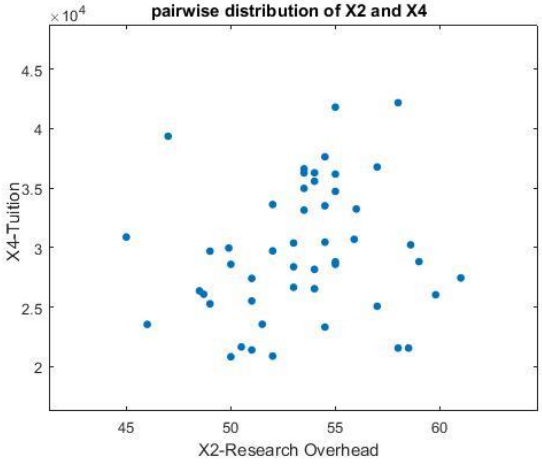
.

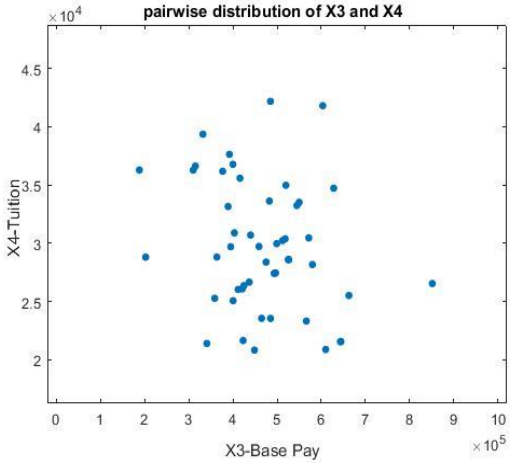

pairwise distribution of X1 and X3
×10^5
X3-Base Pay
X1-CS score


pairwise distribution of X1 and X4
×10^4
X4-Tuition
X1-CS Score


pairwise distribution of X2 and X3
×10^5
X3-Base Pay
X2-Research Overhead

.


pairwise distribution of X2 and X4
×10^4
X4-Tuition
X2-Research Overhead


pairwise distribution of X3 and X4
×10^4
X4-Tuition
X3-Base Pay
×10^5

3.2.1.1 Inference

The scatter plot for X1 and X2 can be observed to be more closely packed than other variable pairs. Also there is a positive correlation between the variables.

For other variables it cannot be directly inferred as to what their correlation is since the scatter plot is more dispersed.

3.3 Computing Log Likelihood when variables are independent

The combined probability of the variables will be represented as: $\boldsymbol{P(X1, X2, X3, X4)}$
Assuming that the variables are independent of each other, the combined probability can be represented as:

$$\boldsymbol{P(X1, X2, X3, X4) = P(X1) * P(X2) * P(X3) * P(X4)} \quad (1)$$

Therefore the Log Likelihood for the data will be:

$$\boldsymbol{Log\ Likelihood =}$$
$$\boldsymbol{Loglikelihood(X1) + Loglikelihood(X2) + Loglikelihood(X3) + Loglikelihood(X4)} \quad (2)$$

The Log Likelihood of a variable is calculated by first finding the normal probability density for the variables, in matlab this was implemented by fusing the function $(normpdf\,(rvMat(:, i), mu(i), sigma(i)))$ [4]

This function computes the pdf at each of the values in a column vector for a random variable , using the normal distribution with mean *mu* and standard deviation *sigma*

Then we take the log of the values obtained and sum them to get a scalar value. The summed values of each variable are then summed to obtain the log likelihood for the distribution.

This procedure can be implemented in matlab as:

$$\boldsymbol{sum(sum\left(log\left(\left(normpdf\left(rvMat(:, i), mu(i), sigma(i)\right)\right)\right), 1\right), 1)} \quad \text{for i = [1,4]}$$

3.3.1 Result

The value of Log likelihood is computed to be ≈ -13146

3.4 Finding a Bayesian Network and its Log likelihood

3.4.1 Finding valid Bayesian Network

Since there are 4 variables, the Bayesian network for the distribution will be represented by a 4x4 binary matrix. From a total of 2^16 possible combinations of 4X4 binary matrices, we need to find the matrices which are representations of a directed acyclic graphs (with no self loops).
In order to list all the possible 4X4 binary matrices, we take all the integers from 1 to 2^16, convert them to a binary string of 16 digits and then reshape the string to a 4X4 matrix. This can be stored in a 3d matrix array where each index of the array stores a possible combination for a 4X4 binary matrix.
The matlab implementation for the procedure will be:

$$\boldsymbol{matrices} = \boldsymbol{reshape}(\boldsymbol{dec2bin}((\mathbf{0}\colon \mathbf{2\text{^}16} - \mathbf{1}), \mathbf{16}).' - '\mathbf{0}', \mathbf{4}, \mathbf{4}, [\mathbf{1}])$$

dec2bin() converts a decimal number to binary, *reshape()* converts the 16 digit binary number into a 4x4 matrix. The result stored in '*matrices'* will be a 3-d array of matrices of size 4X4 [1][2].

Once all the combinations for the 4X4 binary matrices have been obtained we can check the matrices that satisfy the condition for representing directed acyclic networks. This can be done in matlab by

$$\boldsymbol{graphisdag}(\boldsymbol{sparse}(\boldsymbol{matrices}(:,:,\boldsymbol{i}))) \quad : \textit{for ith matrix in 'matrices'} \qquad [3]$$

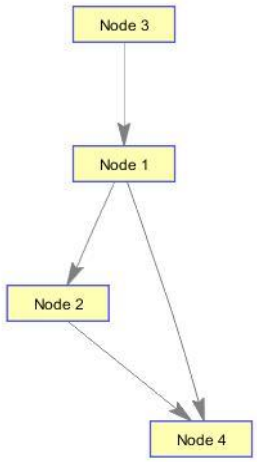
Once a Bayesian network has been obtained, we then try and find the loglikelihood for that network and compare it to the earlier found value in sec 3.3.

3.4.2 Sample Bayesian Network and Log likelihood calculation

A Bayesian network combination found by the method described in sec 3.4.1 was

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The Bayesian network represented by this matrix is



Possible Bayesian Network for given variables

For this particular matrix, the combined probability, determined by the dependencies of the network, will be given by:

$$\boldsymbol{P(n1,n2,n3,n4) = P(n3) * P(n1|n3) * P(n2|n1) * P(n4|n1,n2,n3)} \quad (3)$$

Using Bayes theorem,

$$= \boldsymbol{P(n3)} * \left(\frac{\boldsymbol{P(n1,n3)}}{\boldsymbol{P(n3)}}\right) * \left(\frac{\boldsymbol{P(n2,n1)}}{\boldsymbol{P(n1)}}\right) * \left(\frac{\boldsymbol{P(n1,n2,n3,n4)}}{\boldsymbol{P(n1,n2,n3)}}\right) \quad (4)$$

Here *P(n3)* is the univariate probability distribution for the variable X3 and rest of the terms are combined probability distributions for the variables.

Thus taking the log of each multiplicative term in (4), and taking the sum of the values, we obtain the Log Likelihood for the given Bayesian network (using (2)).

3.4.3 Matlab implementation

Once the dependencies between the nodes are determined, we can calculate $P(X1, X2, X3 \ldots . Xn)$ for an n variable distribution in matlab by using the *mvnpdf* function.
the matlab implementation for this is:

$$\boldsymbol{mvnpdf(data\ values\ of\ dependent\ variables, mu, covariance)}) \quad (5)$$

mvnpdf computes returns the density of the multivariate normal distribution with mean *mu* for the variables and covariance *sigma*[5]

We then calculate the log of density multivariate distribution and then take sum of the values to obtain the log likelihood for the Bayesian network.

Thus the entire procedure can be implemented in matlab as:

$$\boldsymbol{sum}(\log(\boldsymbol{mvnpdf}(\boldsymbol{data\ values\ of\ dependent\ variables}, \boldsymbol{mu}, \boldsymbol{covariance}))) \quad (6)$$

Once 'sum' for each term in (4) was calculated using (6) we then add the values obtained for each term to obtain the Log likelihood

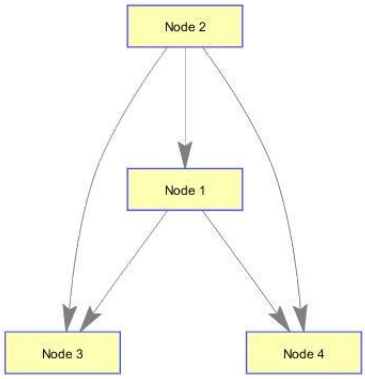
3.4.4 Results and Inference

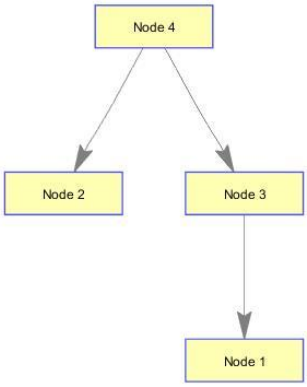
The log likelihood obtained for the Bayesian network of section 3.4.2 is -1306.6227 which is better than the obtained in section 3.3.

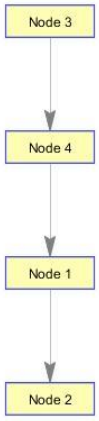
Acc to definition of Likelihood, thus it can be inferred that our variables have a higher likelihood of having dependencies as described by the Bayesian network than being independent of each other.

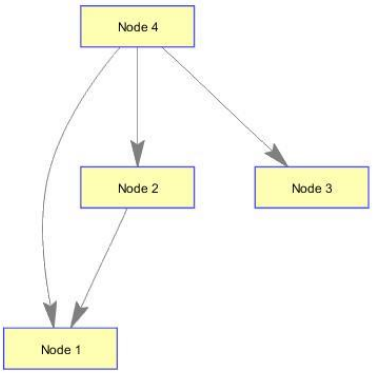
3.5.5 Further Optimization

Some samples of other possible Bayesian networks are









It was found that there are more than 500 possible matrices(out of the exhaustive 2^16 combinations) which can describe a directed and acyclic Bayesian network. Thus we can exhaustively search the log likelihood of all such possible matrices, compare the values and evaluate the Bayesian network that results in the maximum log likelihood for our set of variables.

It is possible to calculate log likelihood for all such valid Bayesian networks and then find the optimum value.

4. Conclusion

Using matlab we were able to get different statistical properties (mean, variance, covariance etc) for a multivariate data set of 4 variables. We found the covariance and correlation matrices for the variables and plotted their pair wise distribution to infer about their correlation.

We also calculated the log likelihood value for the distributions by first assuming that the variables were independent. We then found various possible Bayesian networks for the variables, and for a particular Bayesian network calculated the log likelihood value. The log likelihood value for this network was found to be greater than when the variables were considered to be independent thus implying that the there was a better likelihood of the variables being dependent as described by the given Bayesian Network than being independent.

This can also be further optimized by further analyzing the log-likelihood value for all possible Bayesian networks for the given 4 variables and finding the maximum value, which would give us the optimal Bayesian network.

5.References

1] Reshape, http://www.mathworks.com/help/matlab/ref/dec2bin.htmlReshape

2] Dag test, http://www.mathworks.com/help/bioinfo/ref/graphisdag.html

3] Mvnpdf. http://www.mathworks.com/help/stats/mvnpdf.

4] Normpdf http://www.mathworks.com/help/stats/normpdf.html

5] Aaron Goebel goaaron, Mihir Mongia mmongia Non-linear Reconstruction of Genetic Networks Implicated in AML Pathology.