**STA 546 – Home Work #4**

**Suruchi JaiKumar Ahuja**

**Question 1 –**

Consider the “cad1” data set in the package gRbase. This dataset has 236 observations on fourteen variables from the Danish Heart Clinic.

a.

This network was constructed in R, and the Conditional Probability Tables using the cad1 data were inferred. D-separation is defined as "d"-separation of two nodes, where d stands for directional. The sex smoker and inherit are d- connected. The sex and SuffHeartF are d-separated as there is no direct pathway. Also the nodes inherit and hyperhol are d- separated.

About the network constructed –

A graphNEL graph with directed edges

Number of Nodes = 6

Number of Edges = 6

Random/Generated Bayesian network

model:

[Sex][SuffHeartF][Smoker|Sex][Inherit|Smoker][Hyperchol|Smoker:SuffHeartF][CAD|Inherit:Hyperchol]

nodes: 6

arcs: 6

undirected arcs: 0

directed arcs: 6

average markov blanket size: 2.67

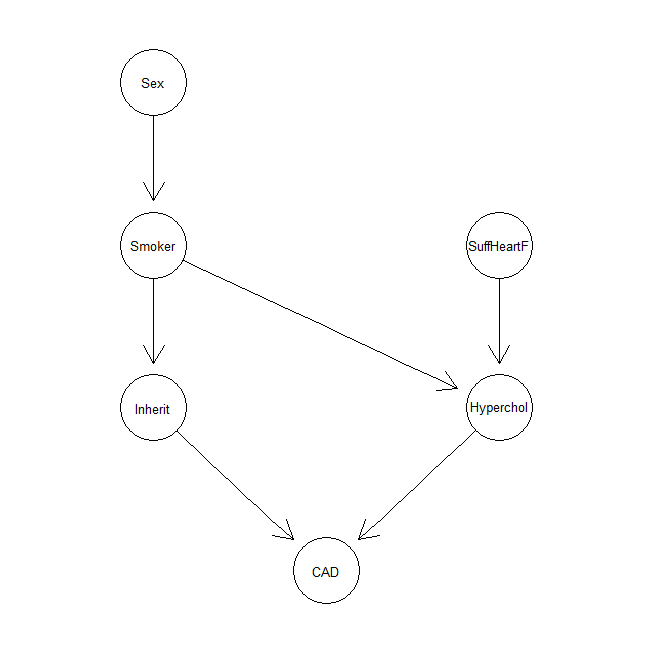
average neighbourhood size: 2.00

average branching factor: 1.00

generation algorithm: Empty

Independence network: Compiled: FALSE Propagated: FALSE

Nodes: chr [1:6] "Sex" "Smoker" "Inherit" "CAD" "Hyperchol" "SuffHeartF"



b.

Consider a female with Hypercholesterolemia (high cholesterol). Now this evidence is absorbed into the graph, and the new probabilities were observed. The change in the probability of heart-failure and coronary artery disease (CAD) after this information is taken into account, can be found using the setevidence function in the query format. So the pEvidence value is 0.1050938 and then the yes value increases to 0.60757 and the no value increases to 0.3924. looking at these values it can be stated that the probability of heart failure and cad increases after this information is taken into account.

Independence network: Compiled: TRUE Propagated: TRUE

Nodes: chr [1:6] "Sex" "Smoker" "Inherit" "CAD" "Hyperchol" "SuffHeartF"

Evidence:

nodes is.hard.evidence hard.state

1 Sex TRUE Female

2 Hyperchol TRUE Yes

pEvidence: 0.105038

> query2

$CAD

CAD

No Yes

0.3924294 0.6075706

$SuffHeartF

SuffHeartF

No Yes

0.6162534 0.3837466

The probabilities of CAD and heart failure change:

> query3

$CAD

CAD

No Yes

0.5401298 0.4598702

$SuffHeartF

SuffHeartF

No Yes

0.7076271 0.2923729

c.

A new data set is simulated with new observations conditional upon this new evidence. Using the new data set, the joint distribution of “Smoker” and “CAD” given this evidence is observed. It has changed when compared to the original observations on the previous dataset.

> cpt\_graph2

Independence network: Compiled: FALSE Propagated: FALSE

Nodes: chr [1:6] "Sex" "Smoker" "Inherit" "CAD" "Hyperchol" "SuffHeartF"

> query3

CAD

Smoker No Yes

No 0.1352618 0.1659319

Yes 0.2780331 0.4207733

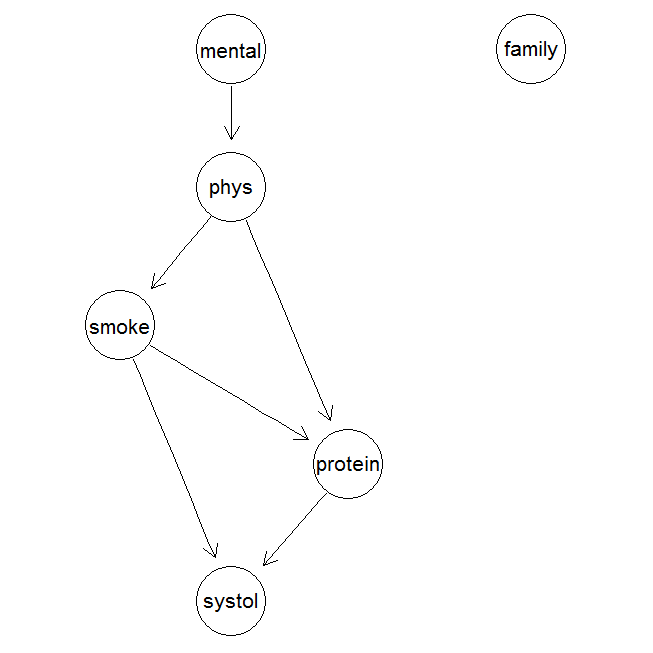
**Question 2 –**

The data “heart.txt”, was generated by the “reinis” data in the gRbase library. Data was collected at the beginning of a 15 year follow-up study of probable risk factors (yes implies risk) for coronary thrombosis. Data are from all men employed in a car factory and has 1843 observations of 6 variables. The variables are: smoking, strenuous mental work, strenuous physical work, systolic blood pressure, ratio of lipoproteins, Family anamnesis of coronary heart disease. The variables systolic blood pressure, and ratio of lipoproteins are clinical risk measurements (markers) for heart disease.

a.

The structure of the network was formulated using the hc function to learn the Bayesian Network using the hill climbing algorithm. It is a score based algorithm where each candidate is assigned a score and then tries to maximize it with some heuristic search algorithm. The score function is usually score-equivalent, so that networks that define the same probability distribution are assigned the same score.

The Family node is independent in the network.



b.

Based on the model obtained and extracting the CPT, it is seen that the person with strenuous mental work is likely to develop high systolic blood pressure. A person with strenuous physical activity and a person without strenuous physical activity will have equal probabilities of developing high systolic blood pressure. Thus a person with high systolic blood pressure is not influenced by strenuous physical activity, and both the cases are equally likely.

OUTPUT :

> data.model <- bn.fit(data.fit, data)

> data.model

Bayesian network parameters

Parameters of node smoke (multinomial distribution)

Conditional probability table:

, , phys = n

mental

smoke n y

n 0.6218487 0.5270440

y 0.3781513 0.4729560

, , phys = y

mental

smoke n y

n 0.4021244 0.4592593

y 0.5978756 0.5407407

Parameters of node mental (multinomial distribution)

Conditional probability table:

n y

0.4221378 0.5778622

Parameters of node phys (multinomial distribution)

Conditional probability table:

n y

0.4959305 0.5040695

Parameters of node systol (multinomial distribution)

Conditional probability table:

, , protein = n, family = n

smoke

systol n y

n 0.4626866 0.4482759

y 0.5373134 0.5517241

, , protein = y, family = n

smoke

systol n y

n 0.4477612 0.4857143

y 0.5522388 0.5142857

, , protein = n, family = y

smoke

systol n y

n 0.4573864 0.5081967

y 0.5426136 0.4918033

, , protein = y, family = y

smoke

systol n y

n 0.3055556 0.4375000

y 0.6944444 0.5625000

Parameters of node protein (multinomial distribution)

Conditional probability table:

, , phys = n

smoke

protein n y

n 0.4239351 0.3206651

y 0.5760649 0.6793349

, , phys = y

smoke

protein n y

n 0.5398458 0.4222222

y 0.4601542 0.5777778

Parameters of node family (multinomial distribution)

Conditional probability table:

n y

0.1421595 0.8578405

> tables <- as.grain(data.model)

> tables

Independence network: Compiled: FALSE Propagated: FALSE

N

**Question 3 –**

a.

PageRank is a [link analysis](https://en.wikipedia.org/wiki/Network_theory#Link_analysis) algorithm and it assigns a numerical [weighting](https://en.wikipedia.org/wiki/Weighting) to each element of a [hyperlinked](https://en.wikipedia.org/wiki/Hyperlink) [set](https://en.wikipedia.org/wiki/Set_(computer_science)) of documents, with the purpose of "measuring" its relative importance within the set. A PageRank results from a mathematical algorithm based on the [webgraph](https://en.wikipedia.org/wiki/Webgraph).

The page rank vector and the overall ranking of importance changes significantly when the damping factor is modified. The “black” line corresponds to a damping factor of 0.05 while the “turquoise” line to the damping factor of 0.95. Page “D” has the highest probability to be reached in the cases with d=0.95, 0.75, 0.50 but in the other cases C has the maximum probability of being reached.

The relative ranking does support the intuition made previously. Page rank of D, C and A are higher than the rest, as there is higher probability to reach these nodes when started from a node. Consider starting from Page B then there is a higher chance of reaching D and A. Also consider starting at Page F then there is a higher probability of reaching C and A (dependent on the damping factors).

Damping factor – 0.05

> pg1$vector

A B C D E F

0.1683271 0.1639395 0.1718214 0.1681380 0.1680380 0.1597361

Damping factor – 0.25

> pg2$vector

A B C D E F

0.1786588 0.1544288 0.1848587 0.1758772 0.1737324 0.1324441

Damping factor – 0.50

> pg3$vector

A B C D E F

0.19227231 0.14719411 0.18583257 0.19135235 0.18399264 0.09935603

Damping factor – 0.75

> pg4$vector

A B C D E F

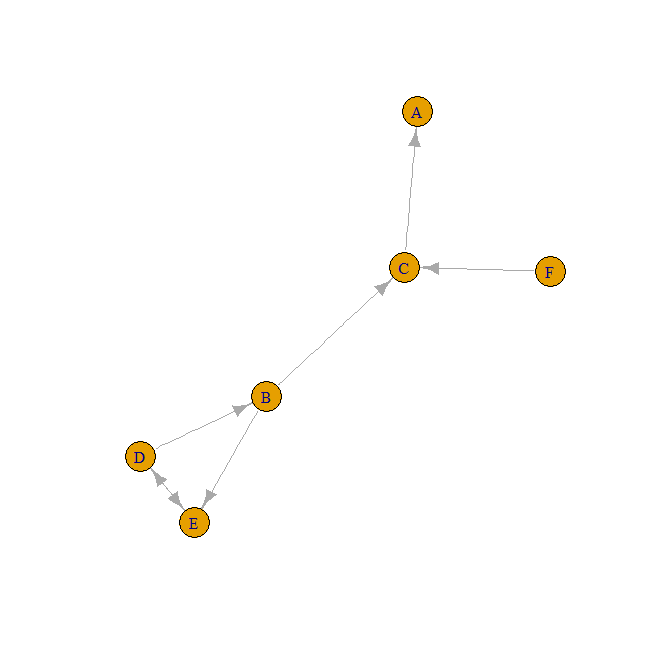
0.19399617 0.14778661 0.17077331 0.21832113 0.20320659 0.06591619

Damping factor – 0.95

> pg5$vector

A B C D E F

0.17305017 0.15761096 0.14454445 0.25658531 0.23247617 0.03573294



b)

When the damping factor value =0.15

|  |
| --- |
| pg$vector  A B C D E F G H  0.1541610 0.1418827 0.1582538 0.1091405 0.1091405 0.1091405 0.1091405 0.1091405 |
|  |
| |  | | --- | |  | |

Looking at the results values obtained above, C has the highest value, followed by A and then B.

As Page C has the maximum number of incoming nodes, it is found to have a better PageRank.

If the damping factor was increased to a higher value, Page A has the highest rank.

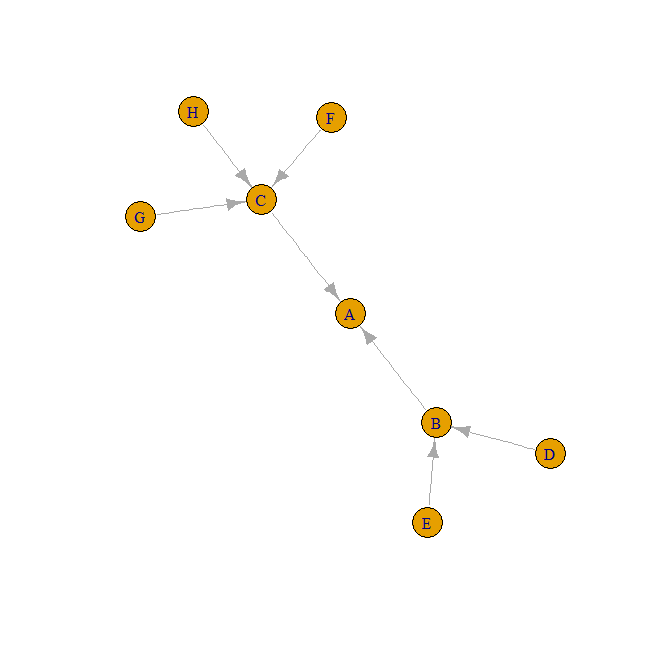
Here an intuition can be made where as the number of incoming links increases, the PageRank gets better.

This is just an assumption and cannot be generalized and used on all cases. This is possible only in cases where damping factor is taken to be low.

pg$vector

A B C D E F G H

0.37329700 0.15258856 0.20163488 0.05449591 0.05449591 0.05449591 0.05449591 0.05449591



**Question 4 –**

The sinking of the Titanic is a famous event in history. The titanic data was collected by the British Board of Trade to investigate the sinking. Many well-known facts—from the proportions of first-class passengers to the ‘women and children first’ policy, and the fact that that policy was not entirely successful in saving the women and children in the third class— are reflected in the survival rates for various classes of passenger.

a.

The arules package was used. It provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules).

|  |
| --- |
| > rules <- apriori(newdata, parameter = list(support = 0.01, confidence = 0.09))  Apriori  Parameter specification:  confidence minval smax arem aval originalSupport support minlen maxlen target ext  0.09 0.1 1 none FALSE TRUE 0.01 1 10 rules FALSE  Algorithmic control:  filter tree heap memopt load sort verbose  0.1 TRUE TRUE FALSE TRUE 2 TRUE  Absolute minimum support count: 22  set item appearances ...[0 item(s)] done [0.00s].  set transactions ...[10 item(s), 2201 transaction(s)] done [0.00s].  sorting and recoding items ... [10 item(s)] done [0.00s].  creating transaction tree ... done [0.00s].  checking subsets of size 1 2 3 4 done [0.00s].  writing ... [226 rule(s)] done [0.00s].  creating S4 object ... done [0.00s].  > inspect(rules)  lhs rhs support confidence lift  1 {} => {Class=2nd} 0.12948660 0.12948660 1.0000000  2 {} => {Class=1st} 0.14766015 0.14766015 1.0000000  3 {} => {Sex=Female} 0.21353930 0.21353930 1.0000000  4 {} => {Class=3rd} 0.32076329 0.32076329 1.0000000  5 {} => {Survived=Yes} 0.32303498 0.32303498 1.0000000  6 {} => {Class=Crew} 0.40208996 0.40208996 1.0000000  7 {} => {Survived=No} 0.67696502 0.67696502 1.0000000  8 {} => {Sex=Male} 0.78646070 0.78646070 1.0000000  9 {} => {Age=Adult} 0.95047706 0.95047706 1.0000000  10 {Age=Child} => {Class=2nd} 0.01090413 0.22018349 1.7004346  11 {Age=Child} => {Sex=Female} 0.02044525 0.41284404 1.9333398  12 {Sex=Female} => {Age=Child} 0.02044525 0.09574468 1.9333398  13 {Age=Child} => {Class=3rd} 0.03589278 0.72477064 2.2595187  14 {Class=3rd} => {Age=Child} 0.03589278 0.11189802 2.2595187  15 {Age=Child} => {Survived=Yes} 0.02589732 0.52293578 1.6188209  .  .  .  .  .  .  .  220 {Class=Crew,Age=Adult,Survived=Yes} => {Sex=Male} 0.08723308 0.90566038 1.1515647  221 {Sex=Male,Age=Adult,Survived=Yes} => {Class=Crew} 0.08723308 0.56804734 1.4127369  222 {Class=Crew,Sex=Male,Age=Adult} => {Survived=Yes} 0.08723308 0.22273782 0.6895161  223 {Class=Crew,Sex=Male,Survived=No} => {Age=Adult} 0.30440709 1.00000000 1.0521033  224 {Class=Crew,Age=Adult,Survived=No} => {Sex=Male} 0.30440709 0.99554235 1.2658514  225 {Class=Crew,Sex=Male,Age=Adult} => {Survived=No} 0.30440709 0.77726218 1.1481571  226 {Sex=Male,Age=Adult,Survived=No} => {Class=Crew} 0.30440709 0.50413845 1.2537952  Then the rhs=c("Survived=No", "Survived=Yes") is set in appearance to make sure that only  "Survived=No" and "Survived=Yes" will appear in the rhs of rules. Here it is seen that there are  Three classes – class 1, class 2, and class 3 and the crew. For all the ones who survived it is seen  that the sex is female and all the ones who perished are male. This can be used as an evidence that  the women and children passengers were evacuated before the male passengers and hence they have  the higher percentage of survival. These results alsp support the movie “Titanic”, where it was seen  in the movie that the women and children were given first preference when it came to the evacuation.      lhs rhs support confidence lift  1 {Class=2nd,Age=Child} => {Survived=Yes} 0.010904134 1.0000000 3.095640  7 {Class=2nd,Sex=Female,Age=Child} => {Survived=Yes} 0.005906406 1.0000000 3.095640  4 {Class=1st,Sex=Female} => {Survived=Yes} 0.064061790 0.9724138 3.010243  10 {Class=1st,Sex=Female,Age=Adult} => {Survived=Yes} 0.063607451 0.9722222 3.009650  2 {Class=2nd,Sex=Female} => {Survived=Yes} 0.042253521 0.8773585 2.715986  5 {Class=Crew,Sex=Female} => {Survived=Yes} 0.009086779 0.8695652 2.691861  11 {Class=Crew,Sex=Female,Age=Adult} => {Survived=Yes} 0.009086779 0.8695652 2.691861  8 {Class=2nd,Sex=Female,Age=Adult} => {Survived=Yes} 0.036347115 0.8602151 2.662916  9 {Class=2nd,Sex=Male,Age=Adult} => {Survived=No} 0.069968196 0.9166667 1.354083  3 {Class=2nd,Sex=Male} => {Survived=No} 0.069968196 0.8603352 1.270871  12 {Class=3rd,Sex=Male,Age=Adult} => {Survived=No} 0.175829169 0.8376623 1.237379  6 {Class=3rd,Sex=Male} => {Survived=No} 0.191731031 0.8274510 1.222295        > inspect(rules.pruned)  lhs rhs support confidence lift  1 {Class=2nd,Age=Child} => {Survived=Yes} 0.010904134 1.0000000 3.095640  4 {Class=1st,Sex=Female} => {Survived=Yes} 0.064061790 0.9724138 3.010243  2 {Class=2nd,Sex=Female} => {Survived=Yes} 0.042253521 0.8773585 2.715986  5 {Class=Crew,Sex=Female} => {Survived=Yes} 0.009086779 0.8695652 2.691861  9 {Class=2nd,Sex=Male,Age=Adult} => {Survived=No} 0.069968196 0.9166667 1.354083  3 {Class=2nd,Sex=Male} => {Survived=No} 0.069968196 0.8603352 1.270871  12 {Class=3rd,Sex=Male,Age=Adult} => {Survived=No} 0.175829169 0.8376623 1.237379  6 {Class=3rd,Sex=Male} => {Survived=No} 0.191731031 0.8274510 1.222295  A Naïve Bayes model can be then fit into the data and then plotted. Naïve Bayes.  Naive Bayes classifiers are a family of simple [probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classifier) based on  applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naive) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features.      A query model is fit into the data so check for the characteristics that survived and perished.  Here it is seen for each class the probability for the male and female.  > query  Independence network: Compiled: TRUE Propagated: TRUE  Nodes: chr [1:4] "Survived" "Class" "Sex" "Age"  Evidence:  nodes is.hard.evidence hard.state  1 Survived TRUE No  pEvidence: 0.676965  > query1  Sex  Class Female Male  1st 0.006924012 0.07495518  2nd 0.009477951 0.10260259  3rd 0.029966218 0.32439620  Crew 0.038195577 0.41348228 |
|  |
| |  | | --- | |  | |

**Question 5 –**

The state data released from the US department of Commerce, Bureau of the Census was loaded in the R environment. The state data set is related to the 50 states of the United States of America.

Principal Component Analysis was performed on the data - is a statistical procedure that uses an [orthogonal transformation](https://en.wikipedia.org/wiki/Orthogonal_transformation) to convert a set of observations of possibly correlated variables into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables called principal components. The number of principal components is less than or equal to the number of original variables.

The summary of the PCA is given below. Also a boxplot of the variables was plotted and shown below which a standardized way of displaying the distribution of data is based on the five number summary: minimum, first quartile, median, third quartile, and maximum. Also a biplot, i.e an enhanced scatterplot that uses both points and vectors to represent structure is also plotted. Also the first principal component was plotted against the second principal component.

> summary(fit.pca)

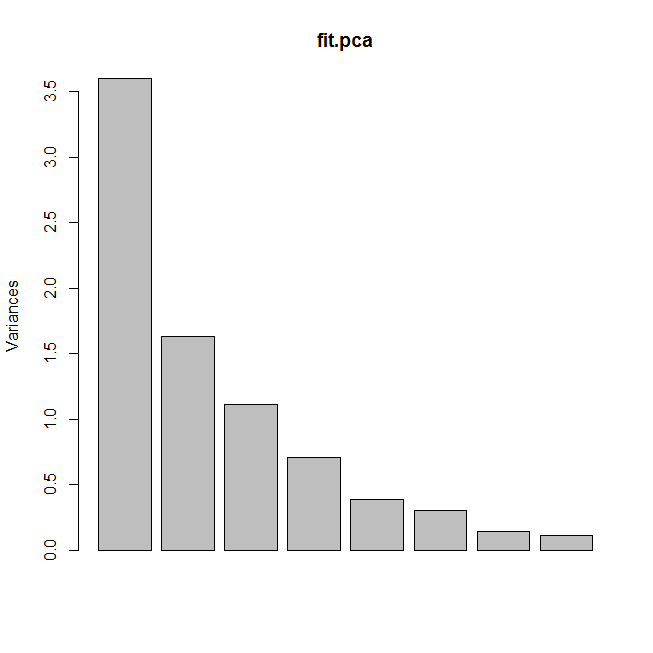
Importance of components:

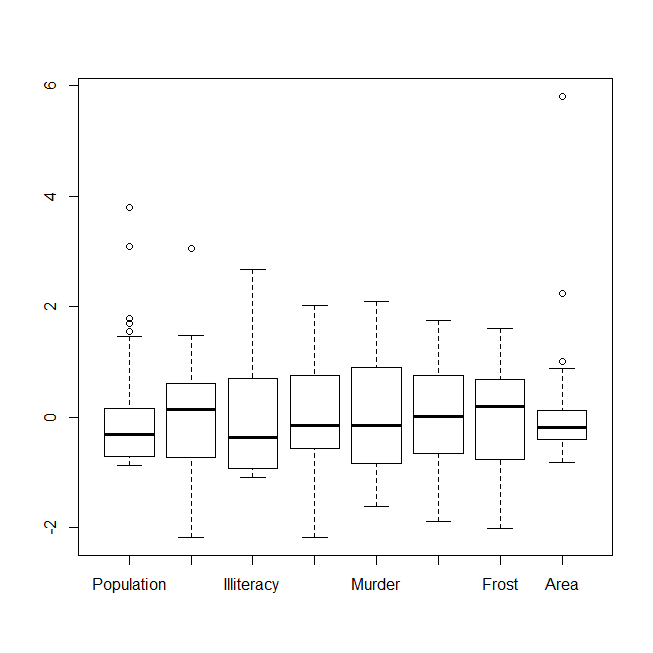
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8

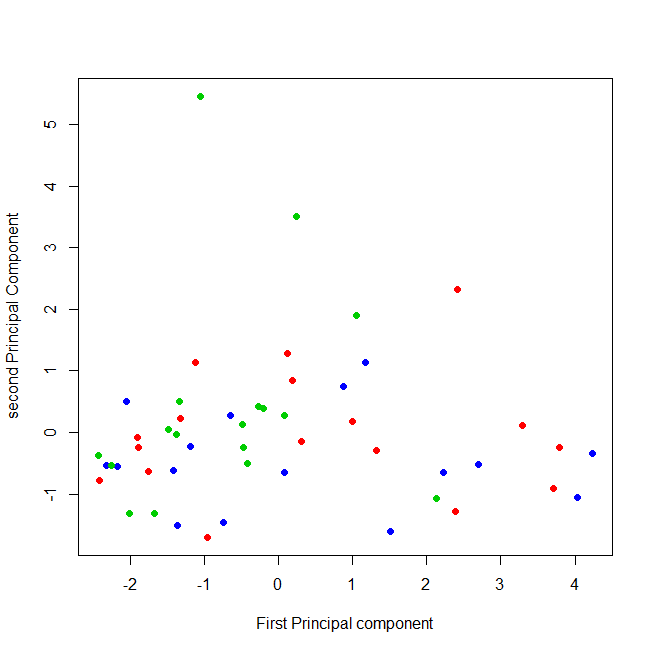
Standard deviation 1.8971 1.2775 1.0545 0.84113 0.62019 0.55449 0.38006 0.33643

Proportion of Variance 0.4499 0.2040 0.1390 0.08844 0.04808 0.03843 0.01806 0.01415

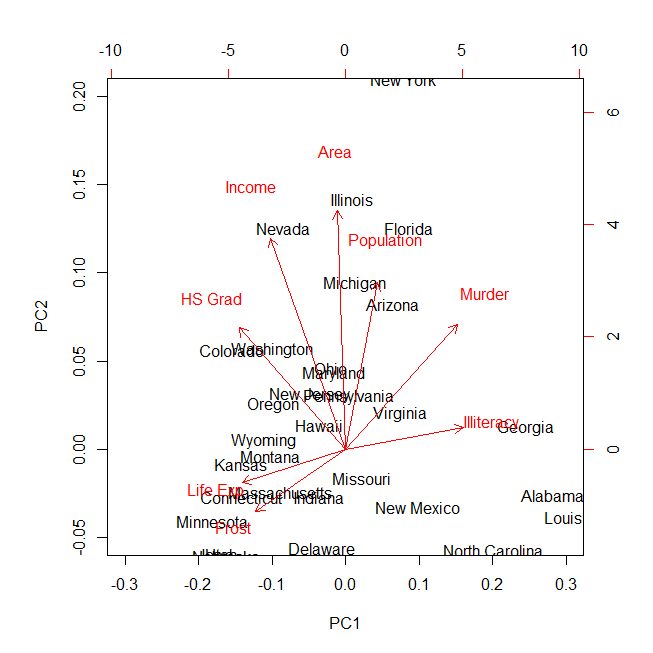
Cumulative Proportion 0.4499 0.6539 0.7928 0.88128 0.92936 0.96780 0.98585 1.00000







**BIPLOT**



Hierarchical Clustering was performed on the data – it is a method of [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) which seeks to build a [hierarchy](https://en.wikipedia.org/wiki/Hierarchy) of clusters. A cluster dendogram was plotted to show the distribution of the variables. The complete distance metric was used for the clustering.

> summary(fit.hc)

Length Class Mode

merge 98 -none- numeric

height 49 -none- numeric

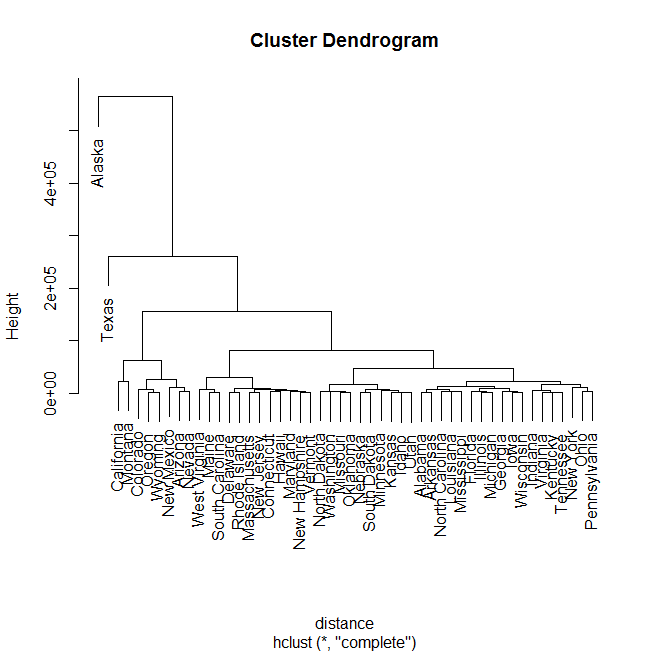
order 50 -none- numeric

labels 50 -none- character

method 1 -none- character

call 2 -none- call

dist.method 1 -none- character



b.

A Gaussian Graphical Model using the Graphical Lasso for the 8 predictors in a. was then built. The penalties used are rho = 0, 2,5,8,12,15. With the number of edges decreasing, the penalty value increases, this can be seen from the graphs attached below the question. States with high population are in the same cluster. Population is not the only factor that determines the states constituting a cluster but is the most the important factor.

The Gaussian model is very different from the methods used above, here the missing edges and penalty values are taken into account whereas in the above models it uses a clustering technique. Hierarchical clustering compares all pairs of data points and merge the one with the closest distance. The glasso function estimates a sparse inverse covariance matrix using a lasso (L1) penalty. The algorithm can also be used to estimate a graph with missing edges, by specifying which edges to omit in the zero argument, and setting rho=0.

> fit.gaussian

$w

[,1] [,2] [,3] [,4] [,5] [,6] [,7]

[1,] 19931683.7588 571229.7795 292.8678395 -4.078428e+02 5663.523730 -3551.509658 -77081.97273

[2,] 571229.7795 377573.3061 -163.7020380 2.806632e+02 -521.894277 3076.768973 7227.60408

[3,] 292.8678 -163.7020 0.3715306 -4.815123e-01 1.581776 -3.235469 -21.29000

[4,] -407.8428 280.6632 -0.4815123 1.802020e+00 -3.869480 6.312685 18.28678

[5,] 5663.5237 -521.8943 1.5817755 -3.869480e+00 13.627465 -14.549616 -103.40600

[6,] -3551.5097 3076.7690 -3.2354694 6.312685e+00 -14.549616 65.237894 153.99216

[7,] -77081.9727 7227.6041 -21.2900000 1.828678e+01 -103.406000 153.992163 2702.00857

[8,] 8587916.9976 19049013.7514 4018.3371517 -1.229410e+04 71940.429996 229873.192817 262703.89306

[,8]

[1,] 8.587917e+06

[2,] 1.904901e+07

[3,] 4.018337e+03

[4,] -1.229410e+04

[5,] 7.194043e+04

[6,] 2.298732e+05

[7,] 2.627039e+05

[8,] 7.280748e+09

$wi

[,1] [,2] [,3] [,4] [,5] [,6] [,7]

[1,] 8.089645e-08 -2.110702e-07 2.560617e-04 -1.089496e-04 -7.196392e-05 1.605118e-05 1.932607e-06

[2,] -2.110702e-07 5.278597e-06 7.583599e-04 4.585950e-05 6.093054e-05 -1.803782e-04 -1.011185e-06

[3,] 2.560614e-04 7.583606e-04 1.185379e+01 -7.113125e-02 -5.255076e-01 3.789584e-01 5.918493e-02

[4,] -1.089498e-04 4.586004e-05 -7.113652e-02 2.103262e+00 6.333410e-01 -1.029115e-01 1.206222e-02

[5,] -7.196394e-05 6.093069e-05 -5.255086e-01 6.333406e-01 3.827135e-01 -1.237630e-02 4.930909e-03

[6,] 1.605118e-05 -1.803783e-04 3.789587e-01 -1.029115e-01 -1.237630e-02 5.308857e-02 1.272295e-03

[7,] 1.932606e-06 -1.011180e-06 5.918491e-02 1.206222e-02 4.930908e-03 1.272295e-03 9.419373e-04

[8,] 2.660777e-10 -8.773352e-09 -1.785632e-05 1.552892e-07 -2.283774e-06 -1.529695e-06 -1.348096e-07

[,8]

[1,] 2.660775e-10

[2,] -8.773351e-09

[3,] -1.785631e-05

[4,] 1.552876e-07

[5,] -2.283774e-06

[6,] -1.529695e-06

[7,] -1.348096e-07

[8,] 2.458328e-10

$loglik

[1] -279.7224

$errflag

[1] 0

$approx

[1] FALSE

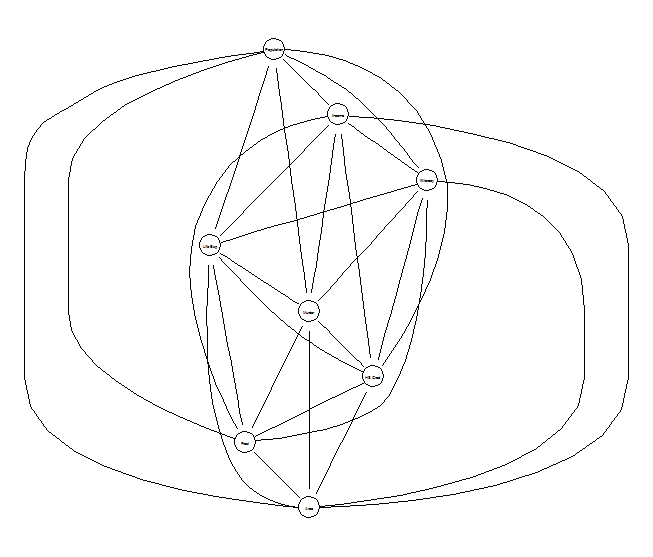
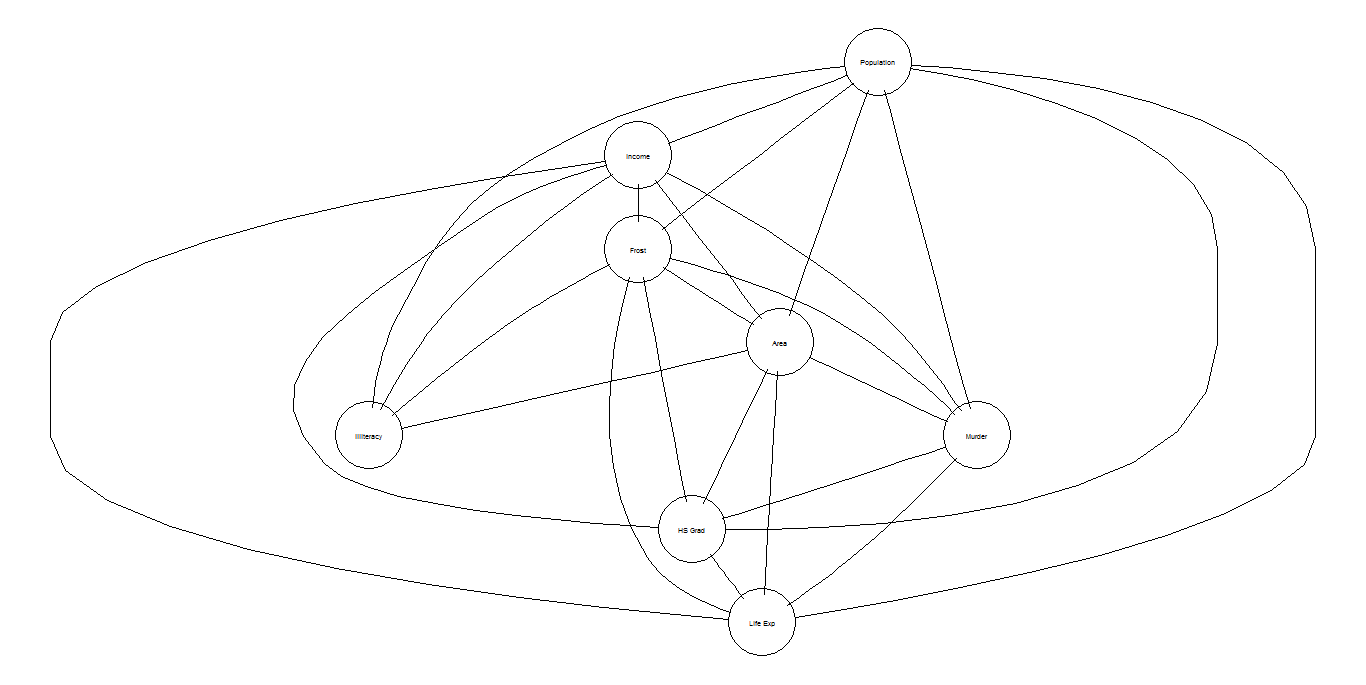
$del

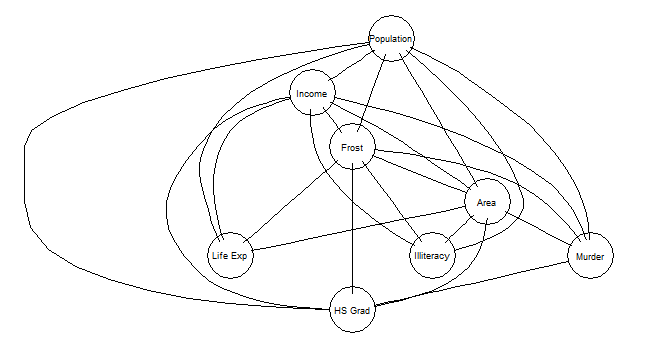
[1] 0.00696571

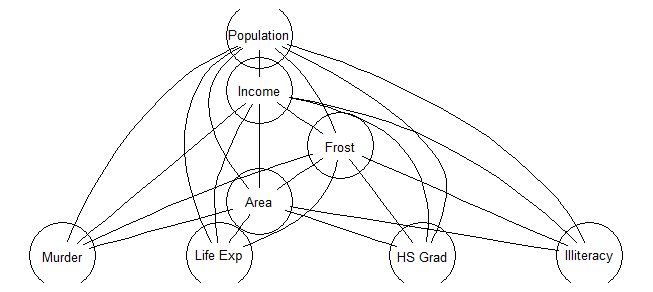
$niter

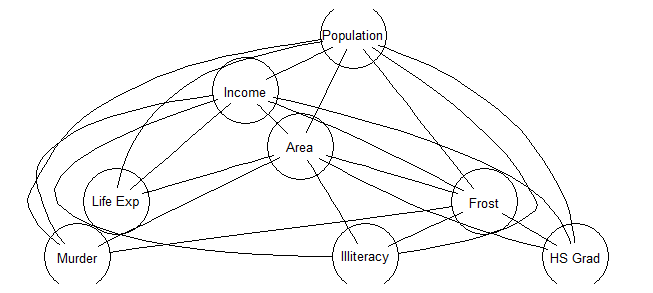
[1] 1

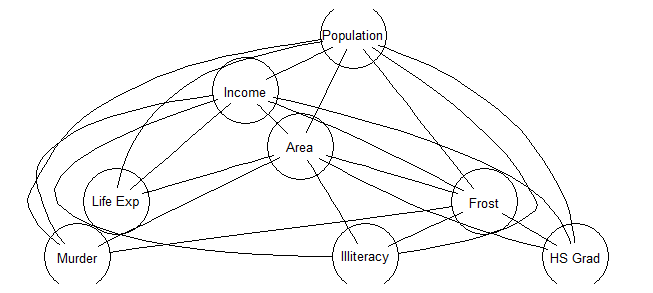
Rho=2









Extra credit –

The conditional independence relations are :

X3 indep X1, X2, X6, X5 | X4

X4 indep X2, X6, X5 | X1

X5 indep X2, X1, X3, X4 | X6

Maximal cliques are [X1, X2, X6], [X3, X4], [X4, X6], [X1, X4]

---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------The End-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------