Statistical Data Mining II - Homework 4

Ezgi Karaesmen

May 9, 2016

Problem 1: Coronary artery disease `cad1` data set and directed acyclic graph

The `cad1` data set in the package `gRbase` was loaded into the environment. `cad1` data set consists of 236 observations on fourteen variables from the Danish Heart Clinic. An optimal network for the data as identified previously by a structural learning algorithm was given. For simplicity, not all variables in the network were represented.

(a) Construct the optimal network in R, and infer the Conditional Probability Tables using the cad1 data. Identify any d-separations in the graph.

The given optimal network was a directed acyclic graph (DAG), hence it was constructed via `dag()` function. The constructed optimal network is shown in Figure 1. The network was then applied to the `cad1` data, and possible probability tables were extracted via `extractCPT()` and `compileCPT()` functions. The conditional probability tables were identified as follows (Tables 1-6):

```
## CPTspec with probabilities:
##  P( CAD | Inherit Hyperchol )
##  P( Inherit | Smoker )
##  P( Hyperchol | SuffHeartF Smoker )
##  P( SuffHeartF )
##  P( Smoker | Sex )
##  P( Sex )
```

Table 1: Sex conditional probability table

Sex	
Female	Male
0.199	0.801

Table 2: Smoker conditional probability table

	Sex	
Smoker	Female	Male
No	0.362	0.180
Yes	0.638	0.820

Table 3: Inheritance conditional probability table

	Smoker	
Inheritance	No	Yes
No	0.824	0.649
Yes	0.176	0.351

Table 4: Heart Failure conditional probability table

Heart Failure	
No	Yes
0.708	0.292

Table 5: Hypercholesterolemia conditional probability table

	Smoker = No		Smoker = Yes	
	Heart Failure		Heart Failure	
Hypercholesterolemia	No	Yes	No	Yes
No	0.675	0.273	0.465	0.328
Yes	0.325	0.727	0.535	0.672

Table 6: Coronary artery disease (CAD) conditional probability table

	Hypercholesterolemia = No		Hypercholesterolemia = Yes	
	Inheritance		Inheritance	
CAD	No	Yes	No	Yes
No	0.821	0.5	0.449	0.26
Yes	0.179	0.5	0.551	0.74

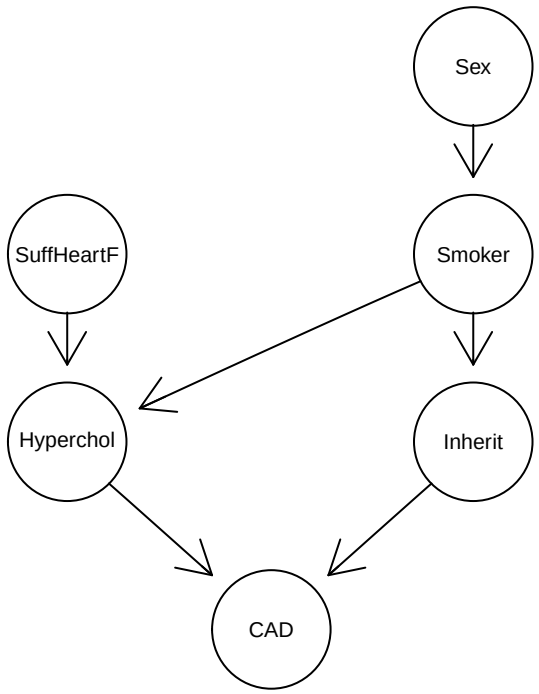

Figure 1: Optimal network for the ‘cad1‘ data set as given in problem 1. For simplicity, not all variables in the network were represented.

Thanks to simplicity of the network, d-separations can be detected easily. Then the `dSep()` function from the `ggm` package can be used to validate the detected d-separations (please see Appendix for the code):

1. When there is evidence for the `Smoker`, we would expect `Sex` and `Inherit` to be d-separated, as the information can be directly obtained from the `Smoker` for `Inherit`. As expected, `dSep()` would validate this fact.
2. When both `Hyperchol` and `Inherit` are known, we would expect `Smoker` and `CAD` to be d-separated. Similarly `SuffHeartF` and `CAD` would be d-separated as well.
3. We can also see that `SuffHeartF` and `Smoker` are d-separated independent from the evidence.
4. Similarly `Hyperchol` and `Inherit` are also d-separated independent from the evidence.

(b) Suppose it is known that a new observation is female with Hypercholesterolemia (high cholesterol). The new evidence was absorbed into the graph, and the probabilities were revised. The changes in the probability of heart-failure and coronary artery disease (CAD) were identified.

The new information was absorbed into the existing network via `setEvidence()` function. As shown in Table 7, upon addition of the new evidence, probabilities of getting both Heart Failure and CAD were increased. This is plausible as according to our new evidence all females suffer from Hypercholesterolemia, resulting in increased probability for having Hypercholesterolemia. On the other hand Hypercholesterolemia depends on the probability of having Heart Failure, and since the probability of getting Hypercholesterolemia has to increase with the new evidence, probability of getting Heart Failure has to increase as well. Similarly, increased probability of having Hypercholesterolemia will increase CAD, as CAD depends on Hypercholesterolemia.

Table 7: Probabilities of Heart Failure and CAD, before and after absorbing the new evidence.

	Heart Failure		CAD	
	No	Yes	No	Yes
Before New Evidence	0.708	0.292	0.54	0.46
After New Evidence	0.616	0.384	0.392	0.608

(c) A new data set with new observations conditional upon the new evidence was simulated. Using the new data set the joint distribution of "Smoker" and "CAD" was determined.

New data set upon the new evidence was simulated with `simulate.grain()` function. Simulated data set can be found in the attached file `newCAD.txt`. Joint distribution of `Smoker` and `CAD` before and after the new evidence was presented in Table 8. As explained in part (b), probability of having CAD was increased overall with the new evidence. Furthermore, probability of having CAD for non-smokers was also increased. This is due to the fact that less females smoke compared to males, but still get Hypercholesterolemia, hence contribute to the risk of CAD overall although they are not smokers.

Table 8: Joint distribution of 'Smoker' and 'CAD' before and after the new evidence

	Before New Evidence		After New Evidence	
	Smoker		Smoker	
CAD	No	Yes	No	Yes
No	0.132	0.408	0.126	0.266
Yes	0.084	0.376	0.177	0.430

Problem 2: `heart` data set and high systolic blood pressure risk

(a) Structure of the generated network

`heart` data set consists of 6 discrete variables and 1843 observations. The variables are `smoke`: smoking, `mental`: strenuous mental work, `phys`: strenuous physical work, `systol`: systolic blood pressure, `protein`: ratio of lipoproteins, `family`: Family anamnesis of coronary heart disease. A plausible directed acyclic graph network was generated according to the variable information, and is shown in Figure 2.

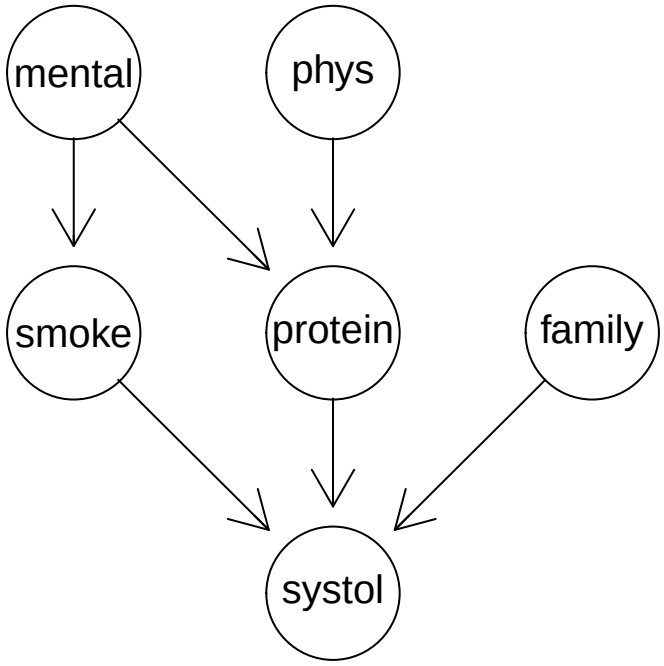

Figure 2: Optimal network for the 'cad1' data set as given in problem 1. For simplicity, not all variables in the network were represented.

(b) Risk of systolic blood pressure risk with respect to mental and physical stress

Based on the generated model, risk of having high systolic blood pressure (Systolic BP = Yes) for individuals with strenuous mental or physical work was compared. The results are shown in Table 9. Physical stress did not effect the systolic blood pressure risk with respect to mental stress and mental stress increased the risk of systolic blood pressure. Hence, it can be argued that an individual with mental stress is more likely to develop high systolic blood pressure.

Table 9: Systolic blood pressure risk (Systolic BP) with respect to mental and physical stress

	Physical Stress = No		*Physical Stress = Yes*	
	Mental Stress		Mental Stress	
Systolic BP	No	Yes	No	Yes
No	0.091	0.120	0.093	0.124
Yes	0.118	0.167	0.120	0.167

Problem 3: Webgraphs

PageRank is an algorithm to determine the most relevant website for a certain keyword search. Beyond for looking the occurrence of a certain keyword, PageRank scores the websites according to their referencing status (i.e. how often is a certain website referenced by others or how many websites a certain website references). This inter-referencing (or "link popularity") relationship between websites results in a directed network (webgraphs). Webgraph A and B was constructed via `graph.data.frame()` function, and shown in Figures 3 and 4.

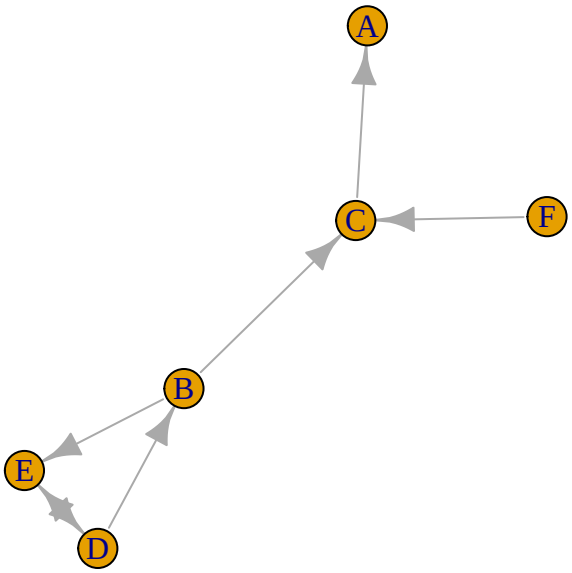

Figure 3: Webgraph A

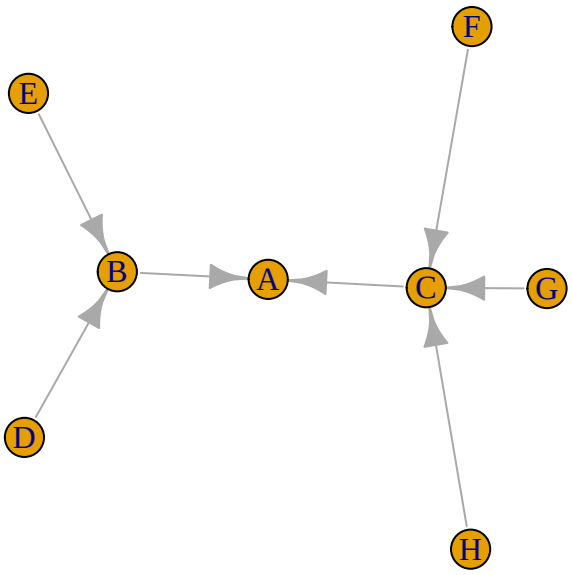

Figure 4: Webgraph B

(a) Compute the PageRank vector of Webgraph A for damping constants p = 0.05, 0.25, 0.50, 0.75, and 0.95.

Above mentioned webgraph A network was used to generate the PageRank vector, which determines the ranking of the given websites. The PageRank algorithm requires a webgraph network and a damping factor `p` as inputs. Damping factor `p` reflects the probability that the user quits the current page and "teleports" to a new one. This solves the issues with "disconnected" networks (networks that has no links with each other) within the global network, or node with no outgoing edges (dangling nodes). In terms of computing the PageRank vector, `p` determines the weights of the adjacency matrix A, which is generated according to the given network and matrix B, where n stands for the total number of pages. B simply represents the probability of "teleporting" to any web page, and since the user can teleport any page, each page has $\frac{1}{n}$ probability to be chosen. Hence `p` determines how much random "teleporting" will be allowed in the model.

The PageRank vector of a webgraph is defined as:

$$M = (1-p)\cdot A + p\cdot B \qquad \text{where} \quad B = \frac{1}{n}\cdot \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

The PageRank vector of Webgraph A (Figure 3) for damping constants `p` = 0.05, 0.25, 0.50, 0.75, and 0.95 was computed. Results are presented in Tables 10-11, where ranking value or the actual ranking of each node (page) was obtained from the PageRank vector that was computed with a certain `p`. It can be seen that for lower values of `p`, ranking is plausible, yet it becomes less intuitive for values larger than `0.25` This is expected as with higher values of `p` more randomness is introduced in the model. Since `0.15` is a widely accepted value for `p`, it is safe to assume that for values of `p` that are larger than `0.25` should not be preferred.

Table 10: Ranking values of each page for different damping constants `p`.

	A	B	C	D	E	F
p=0.05	0.168	0.164	0.172	0.168	0.168	0.160
p=0.25	0.179	0.154	0.185	0.176	0.174	0.132
p=0.5	0.192	0.147	0.186	0.191	0.184	0.099
p=0.75	0.194	0.148	0.171	0.218	0.203	0.066
p=0.95	0.173	0.158	0.145	0.257	0.232	0.036

Table 11: Ranking of each page for different damping constants `p`.

	1st	2nd	3rd	4th	5th	6th
p=0.05	C	A	D	E	B	F
p=0.25	C	A	D	E	B	F
p=0.5	A	D	C	E	B	F
p=0.75	D	E	A	C	B	F
p=0.95	D	E	A	B	C	F

(b) Compute the PageRank vector of Webgraph B for damping constant p = 0.15.

The PageRank vector of Webgraph B (Figure 4) was computed with damping constant `p = 0.15`. Results were presented on Table 12. It can be seen that ranking highly depends on the incoming number of links to a certain node. `C` has the highest number of incoming links and ranked first. Then the outgoing number of links are taken into account. Although `A` and `B` have the same number of incoming links, `A` is ranked second as it has no outgoing links, whereas `B` was ranked third as it has one outgoing link. Similarly `D, E, F G` and `H` are all ranked 4th, since they all only have one outgoing link, and no incoming link, resulting in the exact same ranking value.

Table 12: Ranking values of each Webgraph B page for p = 0.15.

	Ranking Value	Ranking
A	0.154	2nd
B	0.142	3rd
C	0.158	1st
D	0.109	4th
E	0.109	4th
F	0.109	4th
G	0.109	4th
H	0.109	4th

Problem 4: Titanic data

Titanic data consists of 4 variables that specifies the class, sex, age (adult or child) and survival status of 2201 passengers that experienced the historical Titanic accident. The data was analyzed with association rules to answer the certain questions concerning the accident. Association rules identifies which items tend to occur together in a transaction, and helps determining certain rules from these occurrences. For the case of the Titanic data set, it helps to determine which passenger characteristics are tend to occur together more frequently, and allows us to determine rules such as if a passenger is a class 1 female adult, do they survive? The algorithm produces exhaustive number of rules, and certain constraints are required to determine interesting and important rules. These 3 important constraints are:

- **Support (**$supp(X)$**) :** The proportion of transactions (passengers) in the data set which contain the itemset (passenger characteristics).
- **Confidence (**$conf(X => Y) = supp(X \cup Y)/supp(X)$**) :** Probability of right hand side itemset of a rule to occur with the left hand side item.
- **Lift (**$lift(X => Y) = supp(X \cup Y)/(supp(X)supp(Y))$ **) :** Greater lift values indicate stronger associations.

One problem with this approach was that since the frequency of children passengers is very low, support threshold had to be set low, to obtain rules with children. As shown in the item frequency plot in Figure 5, children passengers are less then 10% of the whole passengers. Constraints were set to $support = 0.01$ & $confidence = 0.2$ to generate rules, which resulted in 200 rules.

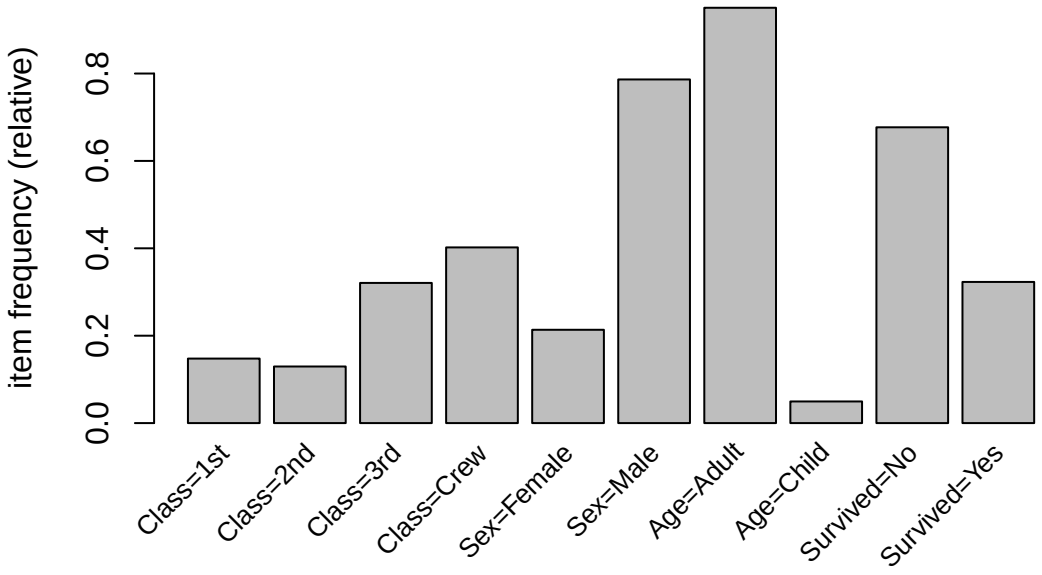

Figure 5: Item frequecy plot of the Titanic data set.

1. Is there evidence that "women and children" were the first evacuated? What characteristics/ demographics are more likely in surviving passengers or in passengers that perished?

To answer this question, 200 rules were filtered for either `Survived=Yes` or for `Survived=No` on the right hand side (RHS), and sorted for lift in decreasing order. First 10 rules for either two of the filtering conditions were presented in Tables 13 and 14. As shown in Table 13, majority of the rules for survivors include either women from different classes or children, and men does not appear on the list. On the other hand, rules resulting in death are mostly comprised of adult men in 2nd, 3rd classes or the crew. Furthermore, no women passengers appear on the top of the list where RHS was `Survived=No`. Overall, these findings suggest that women and children were the first evacuated, but there were differences between classes.

Table 13: Association rules for passengers that survived, sorted by lift.

	lhs		rhs	support	confidence	lift
60	{Class=2nd,Age=Child}	=>	{Survived=Yes}	0.0109041	1.0000000	3.095640
92	{Class=1st,Sex=Female}	=>	{Survived=Yes}	0.0640618	0.9724138	3.010243
170	{Class=1st,Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.0636075	0.9722222	3.009650
77	{Class=2nd,Sex=Female}	=>	{Survived=Yes}	0.0422535	0.8773585	2.715986
163	{Class=2nd,Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.0363471	0.8602151	2.662916
119	{Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.1435711	0.7435294	2.301699
29	{Sex=Female}	=>	{Survived=Yes}	0.1562926	0.7319149	2.265745
22	{Class=1st}	=>	{Survived=Yes}	0.0922308	0.6246154	1.933584
65	{Sex=Female,Age=Child}	=>	{Survived=Yes}	0.0127215	0.6222222	1.926176
101	{Class=1st,Age=Adult}	=>	{Survived=Yes}	0.0895048	0.6175549	1.911728

Table 14: Association rules for passengers that did not survive, sorted by lift.

	lhs		rhs	support	confidence	lift
168	{Class=2nd,Sex=Male,Age=Adult}	=>	{Survived=No}	0.0699682	0.9166667	1.354083
87	{Class=2nd,Sex=Male}	=>	{Survived=No}	0.0699682	0.8603352	1.270871
191	{Class=3rd,Sex=Male,Age=Adult}	=>	{Survived=No}	0.1758292	0.8376623	1.237379
130	{Class=3rd,Sex=Male}	=>	{Survived=No}	0.1917310	0.8274510	1.222295
158	{Sex=Male,Age=Adult}	=>	{Survived=No}	0.6038164	0.7972406	1.177669
55	{Sex=Male}	=>	{Survived=No}	0.6197183	0.7879838	1.163995
148	{Class=Crew,Sex=Male}	=>	{Survived=No}	0.3044071	0.7772622	1.148157
199	{Class=Crew,Sex=Male,Age=Adult}	=>	{Survived=No}	0.3044071	0.7772622	1.148157
48	{Class=Crew}	=>	{Survived=No}	0.3057701	0.7604520	1.123325
151	{Class=Crew,Age=Adult}	=>	{Survived=No}	0.3057701	0.7604520	1.123325

To further investigate the differences between classes for woman and children, rules obtained from the analysis were filtered for women, children and survival. Results for women were presented in Tables 15 nd 16. The support for these rules are low, since there were much fewer female passengers compared to males, as shown in the item frequency plot (Figure 5). However, confidence and lift values were very suggestive. Overall, confidence and lift of the surviving women decreases as their classes decreases. The rules suggest with a high confidence that almost all women in the first class and majority of women in the second class survived the accident. However, nearly half of the women in third class could not be saved, and died in the accident. These facts are supported by the contingency table generated for women in Table 17.

Table 15: Association rules for females that survived, sorted by lift.

	lhs		rhs	support	confidence	lift
92	{Class=1st,Sex=Female}	=>	{Survived=Yes}	0.0640618	0.9724138	3.010243
170	{Class=1st,Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.0636075	0.9722222	3.009650
77	{Class=2nd,Sex=Female}	=>	{Survived=Yes}	0.0422535	0.8773585	2.715986
163	{Class=2nd,Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.0363471	0.8602151	2.662916
119	{Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.1435711	0.7435294	2.301699
29	{Sex=Female}	=>	{Survived=Yes}	0.1562926	0.7319149	2.265745
65	{Sex=Female,Age=Child}	=>	{Survived=Yes}	0.0127215	0.6222222	1.926176
180	{Class=3rd,Sex=Female,Age=Adult}	=>	{Survived=Yes}	0.0345298	0.4606061	1.425871
109	{Class=3rd,Sex=Female}	=>	{Survived=Yes}	0.0408905	0.4591837	1.421467

Table 16: Association rules for females that did not survive, sorted by lift.

	lhs		rhs	support	confidence	lift
112	{Class=3rd,Sex=Female}	=>	{Survived=No}	0.0481599	0.5408163	0.7988837
184	{Class=3rd,Sex=Female,Age=Adult}	=>	{Survived=No}	0.0404362	0.5393939	0.7967826
31	{Sex=Female}	=>	{Survived=No}	0.0572467	0.2680851	0.3960103
123	{Sex=Female,Age=Adult}	=>	{Survived=No}	0.0495229	0.2564706	0.3788535

Table 17: Survival of female passengers according to class.

	Died	Survived
1st	4	141
2nd	13	93
3rd	106	90
Crew	3	20

Results for children were presented in Tables 18 and 19. Again, the support for these rules are very low, since the frequency of children is very low, as shown in the item frequency plot (Figure 5). However, confidence and lift values strongly suggest that all children in 2nd class survived, yet many children in 3rd class died and majority of them were males. These facts are supported by the contingency table generated for children as shown in Table 20. There were only 6 children in the 1st class, due to this low frequency, no rule appeared for these children. Nevertheless, all children in 2nd class were saved, yet more than half of the children in 3rd class died as expected by the rules.

Table 18: Association rules for children that survived, sorted by lift.

	lhs		rhs	support	confidence	lift
60	{Class=2nd,Age=Child}	=>	{Survived=Yes}	0.0109041	1.0000000	3.095640
65	{Sex=Female,Age=Child}	=>	{Survived=Yes}	0.0127215	0.6222222	1.926176
11	{Age=Child}	=>	{Survived=Yes}	0.0258973	0.5229358	1.618821
74	{Sex=Male,Age=Child}	=>	{Survived=Yes}	0.0131758	0.4531250	1.402712
67	{Class=3rd,Age=Child}	=>	{Survived=Yes}	0.0122672	0.3417722	1.058004

Table 19: Association rules for children that did not survive, sorted by lift.

	lhs		rhs	support	confidence	lift
160	{Class=3rd,Sex=Male,Age=Child}	=>	{Survived=No}	0.0159019	0.7291667	1.0771113
69	{Class=3rd,Age=Child}	=>	{Survived=No}	0.0236256	0.6582278	0.9723218
76	{Sex=Male,Age=Child}	=>	{Survived=No}	0.0159019	0.5468750	0.8078335
12	{Age=Child}	=>	{Survived=No}	0.0236256	0.4770642	0.7047103

Table 20: Survival of children according to class.

	Died	Survived
1st	0	6
2nd	0	24
3rd	52	27
Crew	0	0

Overall most of the women and children in the first and second classes were saved, however this was not the case for women or child passengers in the third class. Furthermore, most of the men that died in the accident were either in second and third classes, or were the crew died, whereas many men in the first class most likely survived.

2. What is the probability that Rose (1st class adult and female) and Jack (3rd class adult and male) would not survive?

As shown in Table 14, with 0.84 probability Jack would die, and as shown in Table 15, Rose would survive with 0.97 probability. This points out that the plot of the movie Titanic was accurate.

Problem 5: US State data

State data consists of several vectors specifying the states' region, division, geographic center, abbreviation of the state names etc. and a data matrix with 8 variables (Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area) for 50 states.

(a) Exploring the clustering patterns of states

To explore the clustering patterns of states according to these 8 variables, first a PCA was conducted. Variables were scaled and centered for the analysis. A screeplot showing the percent variability captured by principal components, and a cumulative proportion barplot showing the proportion of variance captured cumulatively by each principal component is presented in Figure 6. About 40% of the variability in the data was captured by the first principal component (PC1) and 20% by the second principal component (PC2). Furthermore, PC1 and PC2 explain about 60% of the variability cumulatively.

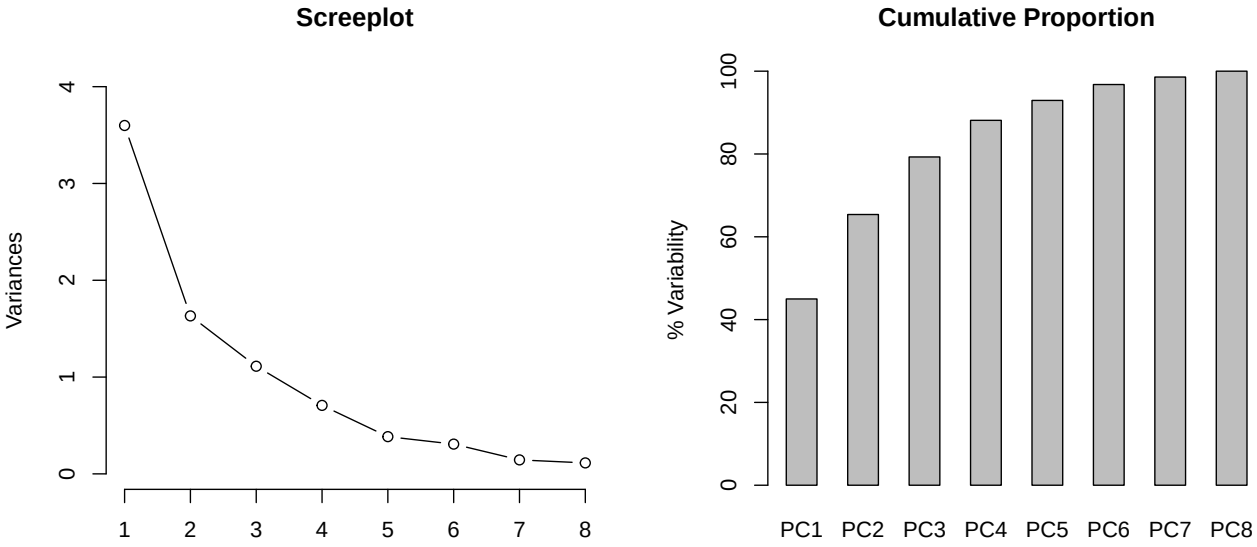

Figure 6: Screeplot and cumulative proportion plot of the PCA of state data.

PC1 and PC2 were used to produce biplots of the analysis, to show the clustering of the states and loadings of the variables. Variable loadings are visualized with red arrows, indicating which component explains what proportion of variance of that certain variable. Additionally variables with smaller angles, that are closer to the same principal component and pointing out the same direction are likely to be correlated or have the similar variances. Consequently, arrows pointing opposite directions are likely to have a negative correlation.

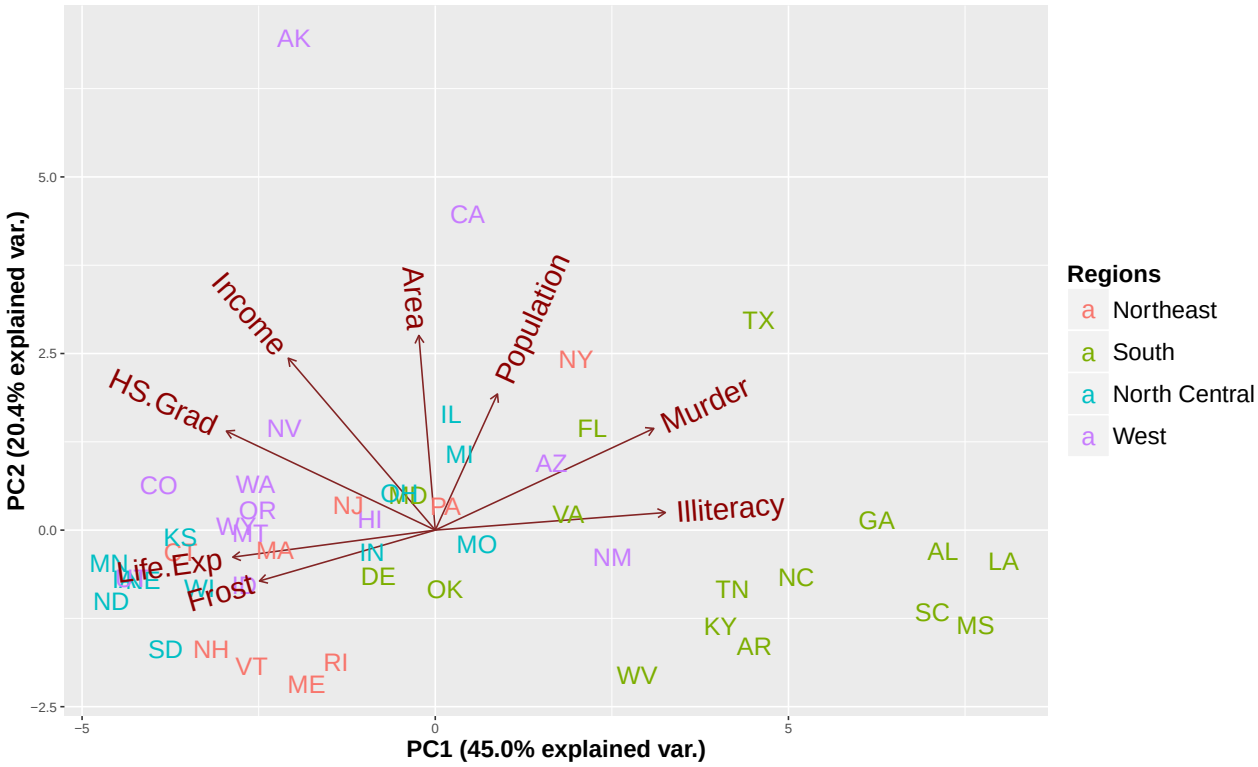

Figure 7: Biplot of the state data, labeled by regions.

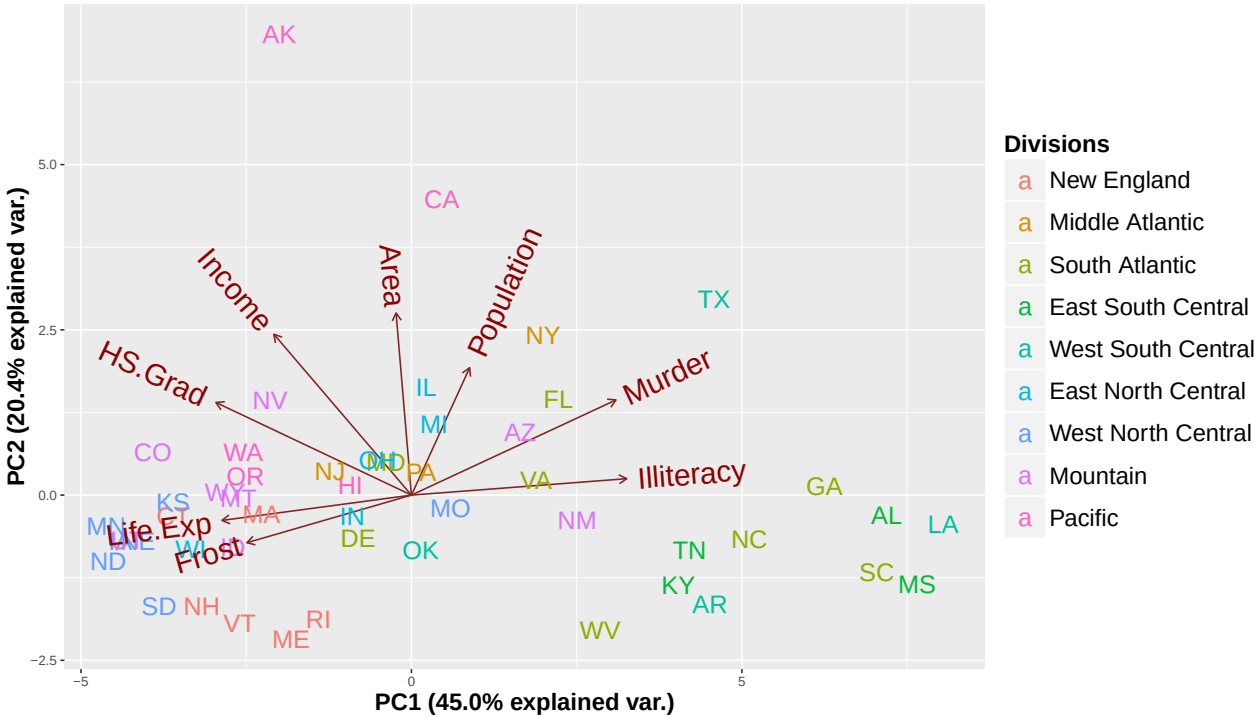

Figure 8: Biplot of the state data, labeled by divisions

Biplots are shown in Figures 7 and 8, where the states were labeled with their abbreviations, and color labeled according to regions or divisions they belong to. Variables `Life expectancy` and `Frost` are very close to each other, and pointing at the same direction, indicating that these variables have very similar variance profiles across states. This also indicates that states that have higher `Frost`, which stands for "mean number of days with minimum temperature below freezing (1931–1960) in capital or large city", also has higher life expectancy. On the other hand, `Murder` and `Illiteracy` variables are closer to each other and pointing at the opposite direction than `Frost` and `Life Expectancy`, indicating that these variables are negatively correlated. Therefore, it can be argued that `Murder` is increased as `Illiteracy` increases and these variables are much higher in states that experience less `Frost` and have a lower `Life Expectancy`. Furthermore, states that have a increased `Murder` and `Illiteracy` rates tend to have larger `Population` size; yet the states with higher `Life Expectancy` and `Frost` rates tend to have higher `HS.Grad` (percent high-school graduates) and `Income` rates.

These characteristics are also seem to be specific to regions and divisions, although regions seem to be explaining the data in much simpler and straight forward way. As shown in Figure 7, `South` states tend to have much higher `Murder`, `Illiteracy` rates and relatively higher population sizes, yet lower `Frost`, `Life Expectancy`, `HS Grad` rates and lower `Income` (except Texas and to some extend Florida). Northeast and North Central regions tend to have higher `Life Expectancy` and `Frost` rates, yet lower `HS Grad` and `Income` compared to West. Furthermore, Alaska is an outlier of the West region, and Arizona and New Mexico have more similar characteristics to South region states.

Biplots suggested that clustering states in 4 clusters and seeing whether the clusters would be enriched for certain regions would be interesting approach. Therefore, k-means clustering was applied to the scaled and centered data for K=4. Enrichment of the regions in the clusters were visualized with mosaic plot in Figure 9. The width of each column (clusters determined by Kmeans) represents the number of states it contains. Wider clusters contain more states compared to others. Colors represent different regions, and the height of each color represents the proportion of the states represented in that certain cluster. It can be seen that clusters 2 and 4 are enriched for South and West states respectively. However, clusters 1 and 3 were not specifically enriched for any of the states, and show a very similar enrichment pattern for Northeast and North Central states. Poor clustering for Northeast and North Central is likely due to the similar variance profiles of these states, as shown in biplot labeled for regions (Figure 7).

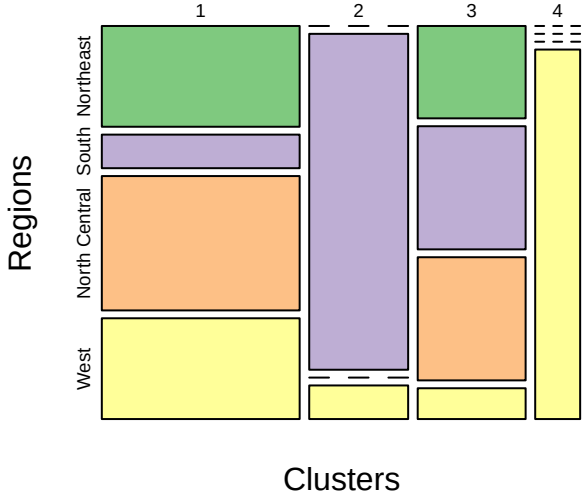

Figure 9: Mosaic plot of the kmeans clustered state data, labeled by regions

To determine which states were mis-clustered, biplots were labeled according to clusters assigned by kmeans algorithm. As shown in Figure 10, kmeans performance was acceptable, and clustered the states well according to 8 variables. Most of the "traditional" South states were clustered together, as well as "northern" states such as Northeast and North Central. Cluster 4 was enriched for "outlier states" such as Alaska, Hawaii and Arizona but also for Northwest states such as Washington and Oregon, which are in the Pacific division with Alaska and Hawaii. On the other hand, larger states such as New York, Texas, California and Illinois with larger population size were enriched in cluster 3. Although not indicated in the data, large metropolitan cities are found within these states that play an important role as economic hubs. These large cities likely affect the statistics on these states, although the variables are distributed unevenly throughout the other cities within these states. Overall the analysis indicated that regions were important indicators for these 8 variables especially for southern and northern region states; however for larger states with larger populations, region was not a good indicator.

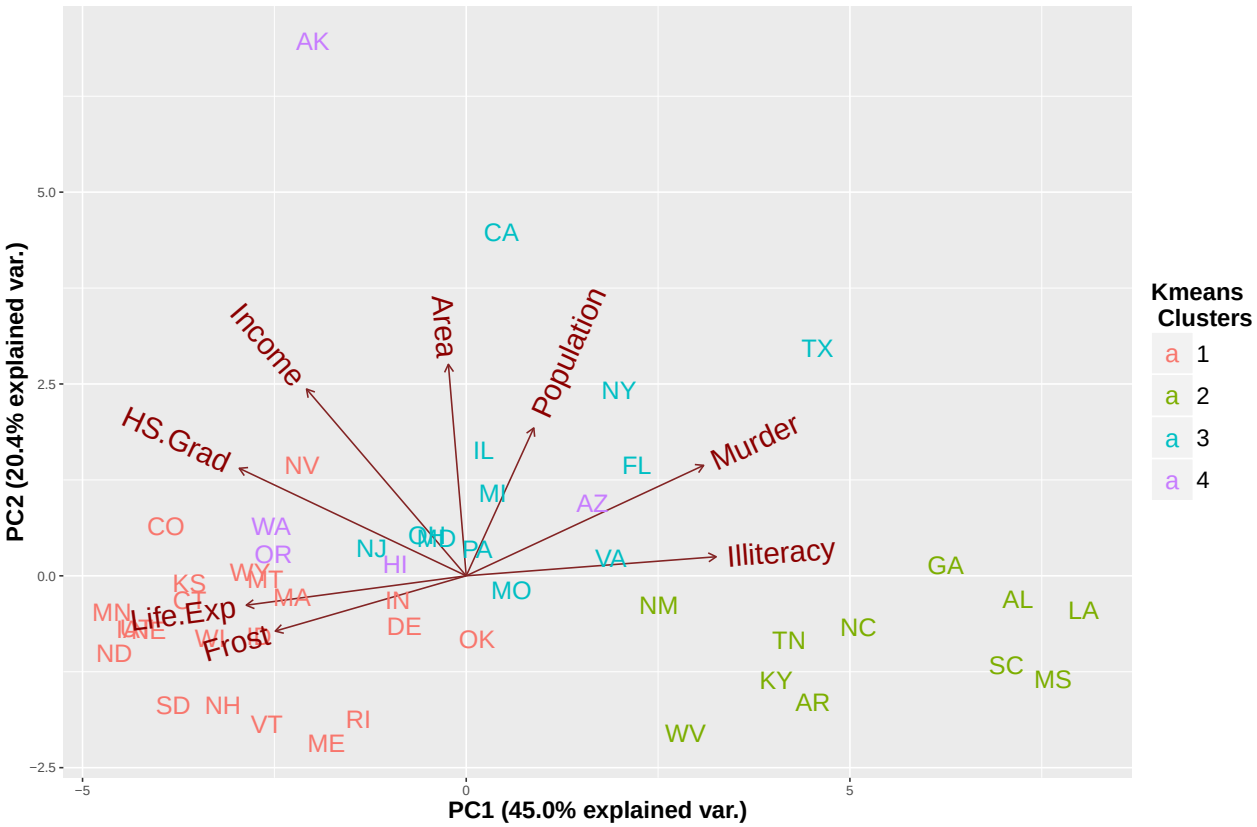

Figure 10: Biplot of the state data, labeled by Kmeans clusters

(b) Build a Gaussian Graphical Model using the Graphical Lasso for the 8 predictors mentioned in Part (a).

Gaussian Graphical Models (GGMs) are continuous graphical models with Gaussian assumptions. These graphs are useful in terms of exploring the relationships between variables, but do not provide definitive evidence of relationships. GGMs may require filtering, as "saturated" graphs can be too complex or not valuable as every node is connected to one another. Lasso algorithm is used to solve this problem. Lasso "shrinks the map" by excluding edges, and even nodes by introducing penalties to the model. The shrinkage can be controlled with the parameter ρ, which determines the degree of penalties. As ρ increases, complexity of the model decreases, which forces "less important" variables to leave the "nest". Lasso requires a covariance matrix of the data variables as the input, and outputs an inverse covariance matrix. Values that are equal to zero in this output matrix indicates the nodes that should be penalized, and excluded from the model. A heatmap of the correlation matrix of the variables is presented in Figure 11. The heatmap includes hierarchical clustering of the variables, which shows an expected clustering and suggest similar relationships between variables to the report described in part (a). Looking at this heatmap, murder, illiteracy, frost, life expectancy are expected to have some links that would connect them.

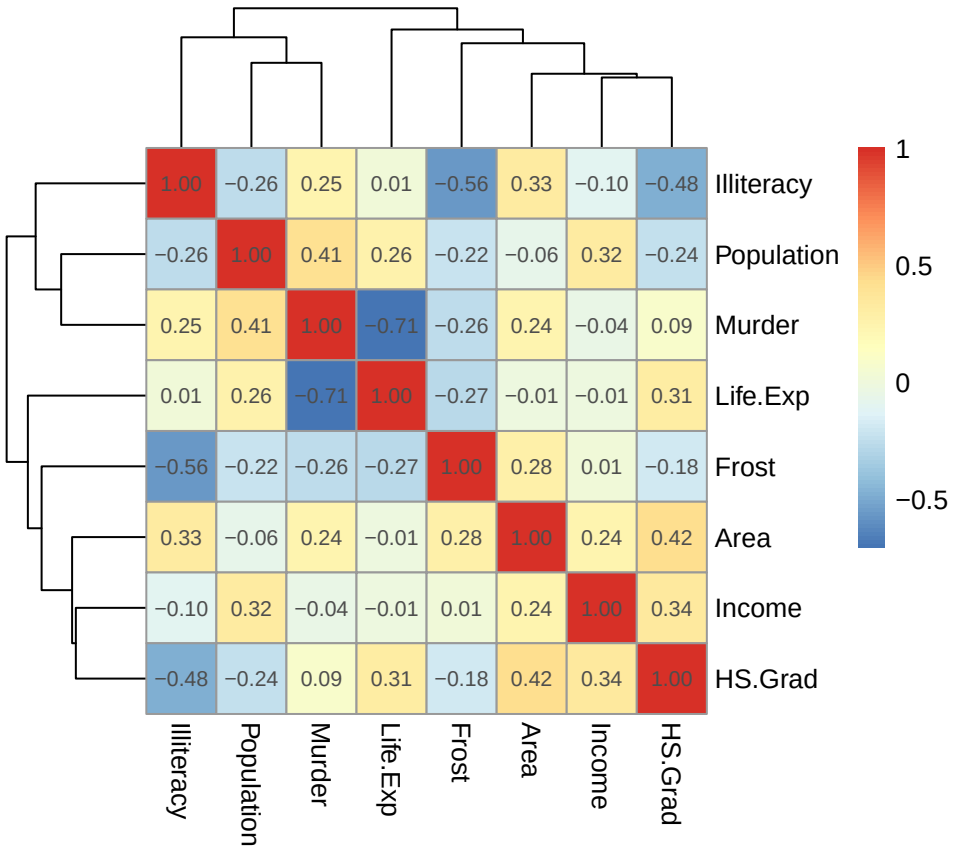

Figure 11: Correlation matrix of the state data set variables

Before generating the GGMs, data was scaled and centered. Unscaled data resulted in very different GGMs that were very insensitive to large ρ values and remained complex (data not shown). GGMs were shrunk for six different ρ values ($\rho = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55$). Results are shown in Figure 12. As expected, for increasing values of ρ complexity was reduced, and starting at $\rho = 0.35$ nodes started leaving the nest. Already for $\rho = 0.25$ model was simplified, and showed expected relationships where illiteracy was linked to murder, frost and life expectancy. These relationships remained even for the highest ρ values indicating its strength. Furthermore, population and area left the nest for higher values of ρ, indicating that these variables were less important in terms of explaining the relationships. Again, this was expected, as both population and are were mostly explained by the second principal component, which explains the data variability less than the first principal component. Overall, lasso performed well in reducing the complexity while keeping the important relationships determined in part (a).

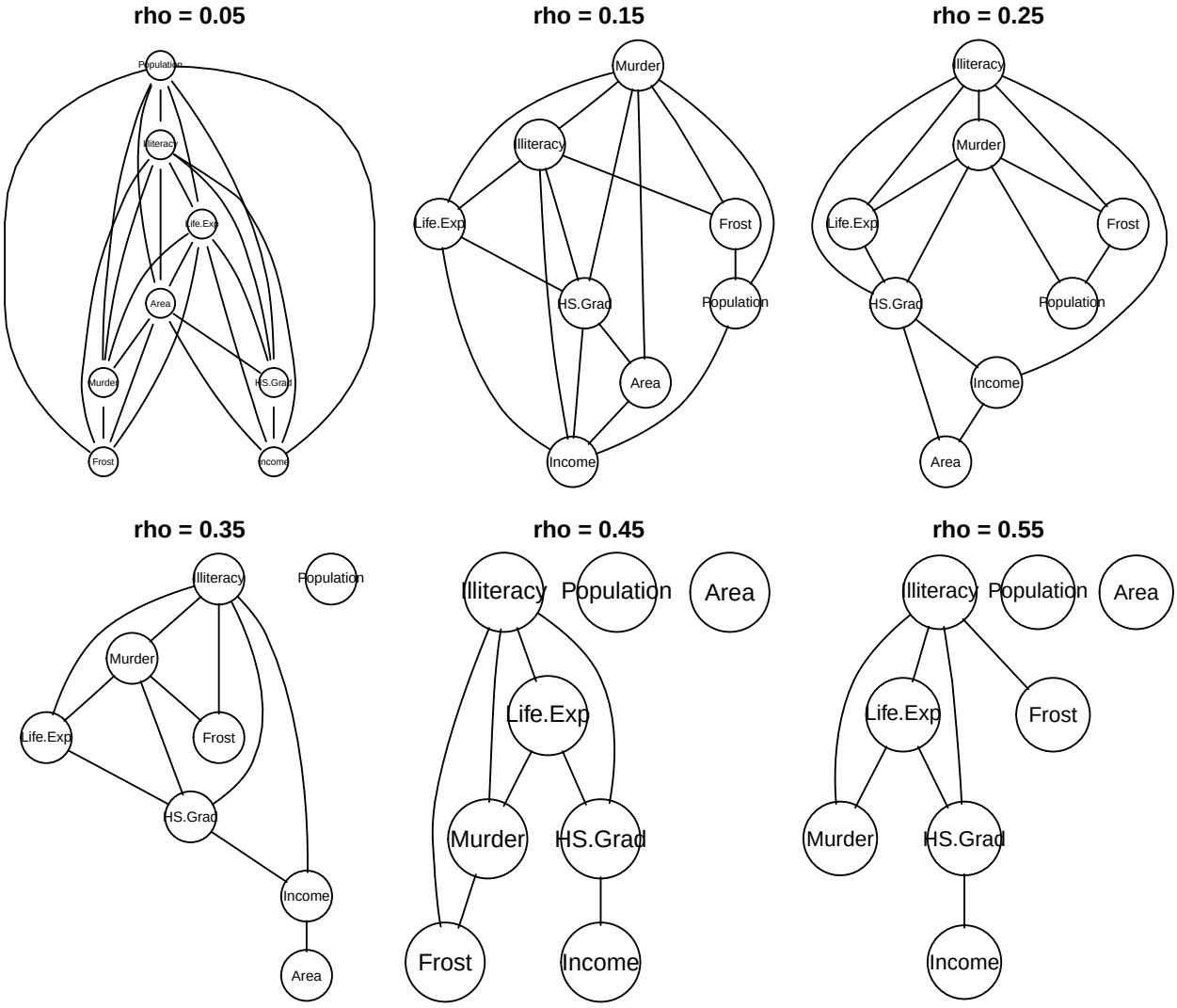

Figure 12: GGM netforks for different values of rho.

Appendix

Problem 1

```
library(gRain)
library(gRbase)
library(ggm)

## Load CAD Data
data(cad1)

###### Part (a) ######
DAG.opt <- dag(~CAD:Inherit:Hyperchol + Hyperchol:SuffHeartF +
                 Inherit:Smoker + Hyperchol:Smoker + Smoker:Sex)
CPT <- extractCPT(cad1, DAG.opt)
CPTlist <- compileCPT(CPT)
cad.DAG <- grain(CPTlist)

# plot network
plot(cad.DAG)

# conditional probability tables
CPTlist$Sex
CPTlist$Smoker
CPTlist$Inherit
CPTlist$SuffHeartF
CPTlist$Hyperchol
CPTlist$CAD

# d-separations

CAD.netw <- list(~Sex, ~Smoker|Sex, ~Inherit|Smoker, ~SuffHeartF,
                 ~Hyperchol|Suffheartf:Smoker,
                 ~CAD|Hyperchol:Inherit)
## Generating the adjacency matrix of the network
CAD.nw <- dagList(CAD.netw, "matrix")

# 1. When there is evidence for the `Smoker`,
#   we would expect `Sex` and `Inherit` to be d-separated,
#    as the infomation can be directly obtained from the `Smoker` for `Inherit`.
#    As expected, `dSep()` would validate this fact.
dSep(CAD.nw, "Inherit", "Sex", "Smoker")

# 2. When both `Hyperchol` and `Inherit` are known,
#    we would expect `Smoker` and `CAD` to be d-separated.
#    Similarly `SuffHeartF` and `CAD` would be d-separated as well.
dSep(CAD.nw, "CAD", "Smoker", c("Inherit", "Hyperchol"))
dSep(CAD.nw, "CAD", "SuffHeartF", c("Inherit", "Hyperchol"))

# 3. We can also see that `SuffHeartF` and `Smoker` are d-separated
#     independent from the evidence.
dSep(CAD.nw, "SuffHeartF", "Smoker", "Hyperchol") # evidence for Hyperchol
```

```r
dSep(CAD.nw, "SuffHeartF", "Smoker", "CAD") # evidence for CAD
dSep(CAD.nw, "SuffHeartF", "Smoker", "Sex") # evidence for Sex
dSep(CAD.nw, "SuffHeartF", "Smoker", NULL) # no evidence

# 4. Similarly `Hyperchol` and `Inherit` are also d-separated independent from the evidence.
dSep(CAD.nw, "Hyperchol", "Inherit", "Smoker") # evidence for Smoker
dSep(CAD.nw, "Hyperchol", "Inherit", "CAD") # evidence for CAD
dSep(CAD.nw, "Hyperchol", "Inherit", "Sex") # evidence for Sex
dSep(CAD.nw, "Hyperchol", "Inherit", NULL) # no evidence


###### Part (b) ######
cad.newEvid <- setEvidence(cad.DAG, evidence=list(Sex="Female", Hyperchol="Yes"))

# before new evidence
querygrain(cad.DAG)$CAD
querygrain(cad.DAG)$SuffHeartF

# after new evidence
querygrain(cad.newEvid)$CAD
querygrain(cad.newEvid)$SuffHeartF

###### Part (c) ######

new.cad <- simulate.grain(cad.newEvid, nsim = 5000)
head(new.cad)
#joint distribution of "Smoker" and "CAD" given the new evidence
new <- querygrain(cad.newEvid, nodes=c("Smoker", "CAD"), type="joint", result = "data.frame")
new$Freq <- round(new$Freq, 3)
new

old <- querygrain(cad.DAG, nodes=c("Smoker", "CAD"), type="joint", result = "data.frame")
old$Freq <- round(old$Freq, 3)
old
```

Problem 2

```r
heart <- read.table("./hw4/heart-2.txt", header=T)
heartDAG <-  dag(~systol:smoke:protein:family + smoke:mental + protein:phys:mental)
heartCPT <- extractCPT(heart, heartDAG)
plist <- compileCPT(heartCPT)
heart.N <- grain(plist)

#plot the network
plot(heart.N)

# Based on your model, who is more likely to develop
# high systolic blood pressure (risk = yes),
# a person with strenuous mental work, or one with strenuous physical work, or both?
sys <- querygrain(heart.N, nodes=c("systol", "mental", "phys"), type="joint")
round(sys, 3)
```

Problem 3

```
####### part (a) ######

nodesA <- data.frame(names=c("A","B","C","D","E","F"))
relationsA <- data.frame(
  from = c("E", "D", "F", "D", "B", "B", "C"),
  to   = c("D", "E", "C", "B", "E", "C", "A")
)
gA <- graph.data.frame(relationsA, directed = TRUE, vertices = nodesA)
plot(gA) #, main="Webgraph A")

damp <- c(0.05,0.25,0.50,0.75,0.95) # a typical damping factor is 0.15
pgA <- list()
rankedA <- list()
for(i in 1:length(damp)){
  name <- paste('p=',damp[i],sep='')
  temp <- page.rank(gA, damping=damp[i])$vector
  pgA[[name]] <- temp
  rankedA[[name]] <- names(sort(temp, decreasing = T))
}

pgA.table <- do.call(rbind, pgA)
rankedA.table <- do.call(rbind, rankedA)
colnames(rankedA.table) <- c("1st", "2nd", "3rd", "4th", "5th", "6th")
kable(round(pgA.table, 3), caption = "Ranking values of each page for different damping constants `p`.")
kable(rankedA.table, caption = "Ranking of each page for different damping constants `p`.")


####### part (b) ######

nodesB <- data.frame(names=c("A","B","C","D","E","F","G","H"))
relationsB <- data.frame(
  from = c("B","C","D","E","F","G","H"),
  to = c("A","A","B","B","C","C","C")
)
gB <- graph.data.frame(relationsB, directed = TRUE, vertices = nodesB)
plot(gB) #, main="Webgraph B")

damp <- 0.15 # a typical damping factor is 0.15
pgB <- page.rank(gB, damping=damp)$vector
ranking <- c("1st", "2nd", "3rd", rep("4th", 5)) #, "5th", "6th", "7th", "8th")
rankedB <- c()
rankedB[order(pgB, decreasing = T)] <- ranking
pgB.table <- data.frame(round(pgB, 3), rankedB)
colnames(pgB.table) <- c("Ranking Value", "Ranking")
kable(pgB.table, caption = "Ranking values of each Webgraph B page for p = 0.15.")
```

Problem 4

```
load("./hw4/titanic.raw-2.rdata")
kable(head(titanic.raw), caption = "Head of the Titanic data set")

## Convert to a binary incidence matrix
library(arules)
titanic <- as(titanic.raw, "transactions")

## itemFrequencyPlot
itemFrequencyPlot(titanic, support = 0.01, cex.names = 0.8)

## Apply the apriori algorithm
rules  <- apriori(titanic, parameter = list(support = 0.01, confidence = 0.2))

## Filter for Survived Yes and No
survived <- subset(rules, subset = rhs %in% "Survived=Yes")
died <- subset(rules, subset = rhs %in% "Survived=No")
surv.table <- inspect(sort(survived, by = "lift"))
died.table <- inspect(sort(died, by = "lift"))

kable(surv.table[1:10,], caption = "Association rules for passengers that survived, sorted by lift.")
kable(died.table[1:10,], caption = "Association rules for passengers that did not survive, sorted by li

## Women sruvivors
survivors.fem <- subset(rules, subset = rhs %in% "Survived=Yes" & lhs %in% "Sex=Female")
died.fem <- subset(rules, subset = rhs %in% "Survived=No" & lhs %in% "Sex=Female")
surv.fem.table <- inspect(sort(survivors.fem, by = "lift"))
died.fem.table <- inspect(sort(died.fem, by = "lift"))
kable(surv.fem.table, caption = "Association rules for females that survived, sorted by lift.")
kable(died.fem.table, caption = "Association rules for females that did not survive, sorted by lift.")

## Child survivors
survivors.child <- subset(rules, subset = rhs %in% "Survived=Yes" & lhs %in% "Age=Child")
died.child <- subset(rules, subset = rhs %in% "Survived=No" & lhs %in% "Age=Child")
surv.child.table <- inspect(sort(survivors.child, by = "lift"))
died.child.table <- inspect(sort(died.child, by = "lift"))
kable(surv.child.table, caption = "Association rules for children that survived, sorted by lift.")
kable(died.child.table, caption = "Association rules for children that did not survive, sorted by lift.

## Contingency table for women
fem <- titanic.raw[which(titanic.raw$Sex == "Female"), ]
levels(fem$Survived) <- c("Died", "Survived")
kable(table(fem$Class, fem3$Survived), caption = "Survival of female passengers according to class.")

## Contingency table for children
child <- titanic.raw[which(titanic.raw$Age == "Child"), ]
levels(child$Survived) <- c("Died", "Survived")
kable(table(child$Class, child$Survived), caption = "Survival of children according to class.")
```

Problem 5

```
####### Part (a) ########

## Data released from the US Department of Commerece, Beureau of the Census is available in R
data(state)
?state
## First need to pre-process data
states <- data.frame(state.x77)
pc <- prcomp(states, center=T, scale=T)
#summary(pc)
screeplot(pc, type="lines")

# Other info concerning states
region <- state.region
state.abb <- state.abb
division <- state.division
st.names <- state.name


# Biplots
library(ggbiplot)
mybiplot <- function(groups=region, group.title="Legend"){
g <- ggbiplot(pc, obs.scale = 2, var.scale = 1, groups=groups, labels =state.abb,
              labels.size = 6, varname.size = 7, varname.adjust = 1.2)
g <- g + theme(legend.text = element_text(size = 16), legend.key.size=unit(1, "cm"),
               legend.title= element_text(size = 16, face="bold" ),
               axis.title=element_text(size=16,face="bold"))
g <- g +  labs(colour = group.title)

g
}

mybiplot(region, "Regions")
mybiplot(division, "Divisions")

# Kmeans clustering
sc.states <- scale(states)
km.states <- kmeans(sc.states, 4)
clus.reg <- data.frame(Clusters = km.states$cluster, Regions = region)
tab.clus.reg1 <- table(clus.reg$Clusters, clus.reg$Regions)
mosaicplot(tab.clus.reg1, color=c('#7fc97f','#beaed4','#fdc086','#ffff99'),
           cex.axis=1, main=NULL, xlab="Clusters", ylab="Regions")

####### Part (b) ########

library(gRbase)
library(gRim)
library(gRain)
library(glasso)
library(pheatmap)
library(ggm)
library(igraph)
```

```r
# Generate Weighted Covariance Matrices
S.body <- cov.wt(sc.states, method = "ML") #scaled state data

# Partial correlation matrix
PC.body <- cov2pcor(S.body$cov)

# correlation matrix heatmap
pheatmap(PC.body, cex=1.5, display_numbers=T)

# assign covariate values
S <- S.body$cov

# Estimate over a range of rho's
rhos <- seq(0.05, 0.6, 0.10) # complexity parameter

m0.lasso <- glassopath(S, rho = rhos)
par(mfrow=c(2,3))
for (i in 1:length(rhos)){
  my.edges <- m0.lasso$wi[, , i] != 0 #stack of matrices
  diag(my.edges) <- FALSE #adjacency matrix
  g.lasso <- as(my.edges, "graphNEL") # convert for plotting
  nodes(g.lasso) <- names(state.x77)
  glasso.net <- cmod(g.lasso, data = state.x77)
  plot(glasso.net)
  title(main=paste("rho = ", rhos[i], sep=""))
}
```