Report Homework -5

Name: Akshay Adlakha
UBIT: akshayad
Person#: 50317479

Problem 2: In this problem, I had to specify the structure of bayesian network containing 4 nodes W, X, Y and Z given a set of independencies. I implemented this in R code.

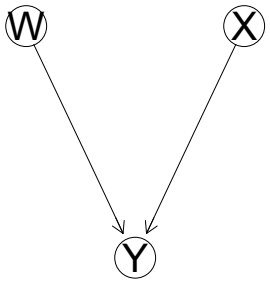
The independence of given conditions checked if there is any dependency between the nodes. If the nodes are d-separated that means they are independent. I checked the d-separation of all the conditions given in the question.

To find the d-separation of nodes, all the paths should be inactive between the nodes. And, if there is any path found to be active then they are not d-separated or there is a dependency between the nodes.

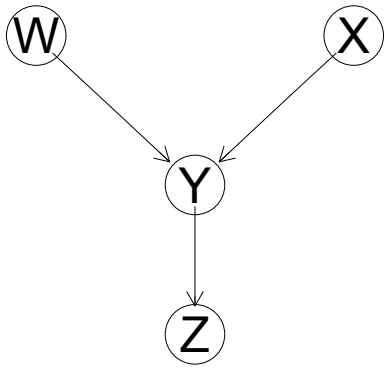
Following conditions were given in the question:

$$
\begin{aligned}
W &\perp X \\
W &\not\perp Z \mid X \\
Z &\perp W \mid Y \\
W &\not\perp Y \\
X &\not\perp Y \\
W &\not\perp X \mid Z \\
X &\perp Z \mid W, Y
\end{aligned}
$$

Based on the conditions: W is independent of X, W and X are not independent of Y. We get the following structure:



And, it given that W is not independent of Z given X and Z is independent of W given Y. And, we know W and X are independent. X and Z are independent given W, Y. Based on this, we get this.



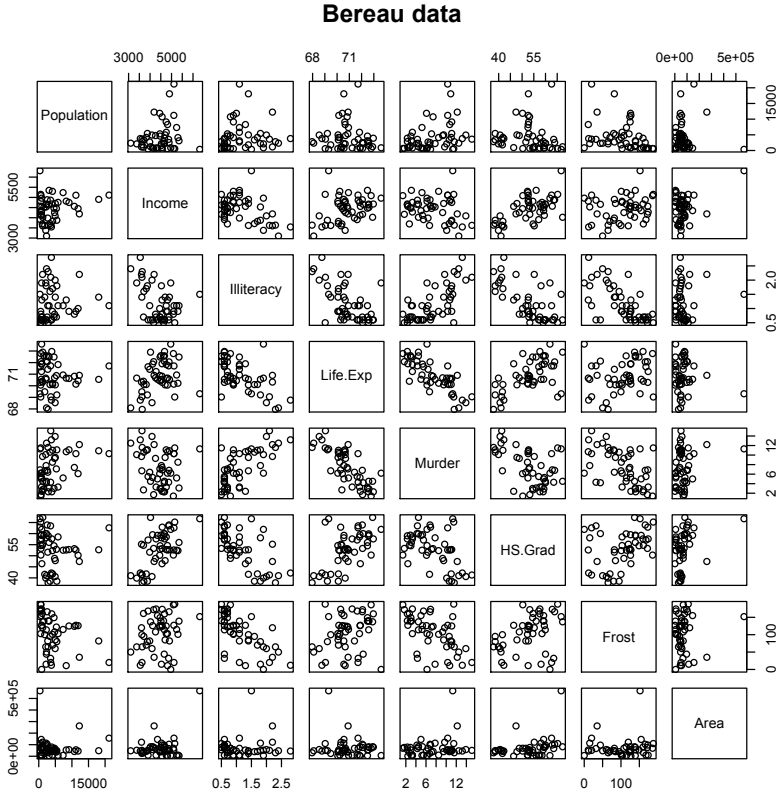
I checked all these condition using dsep, Here is the result:

```
> # W and X independent given no condition
> dSep(as(graphdag, "matrix"),"W","X",cond = NULL)
[1] TRUE
>
> # W and Z independent given X
> dSep(as(graphdag, "matrix"),"W","Z",c("X"))
[1] FALSE
>
> #  W and Z given Y
> dSep(as(graphdag, "matrix"),"W","Z",c("Y"))
[1] TRUE
>
> # W and Y given no condition
> dSep(as(graphdag, "matrix"),"W","Y",cond = NULL)
[1] FALSE
>
> # Y and X given no condition
> dSep(as(graphdag, "matrix"),"Y","X",cond = NULL)
[1] FALSE
>
> # W and X given Z
> dSep(as(graphdag, "matrix"),"W","X",c("Z"))
[1] FALSE
>
> # X and Z given W and Y
> dSep(as(graphdag, "matrix"),"X","Z",c("W","Y"))
[1] TRUE
```

As we can see that the graph is satisfying all the conditions given in the question.

Problem 5: In this problem, I was given a Bureau of Census data from the US Department. I had to perform the Hierarchical clustering, SOM and Gaussian graphical model using graphical lasso.

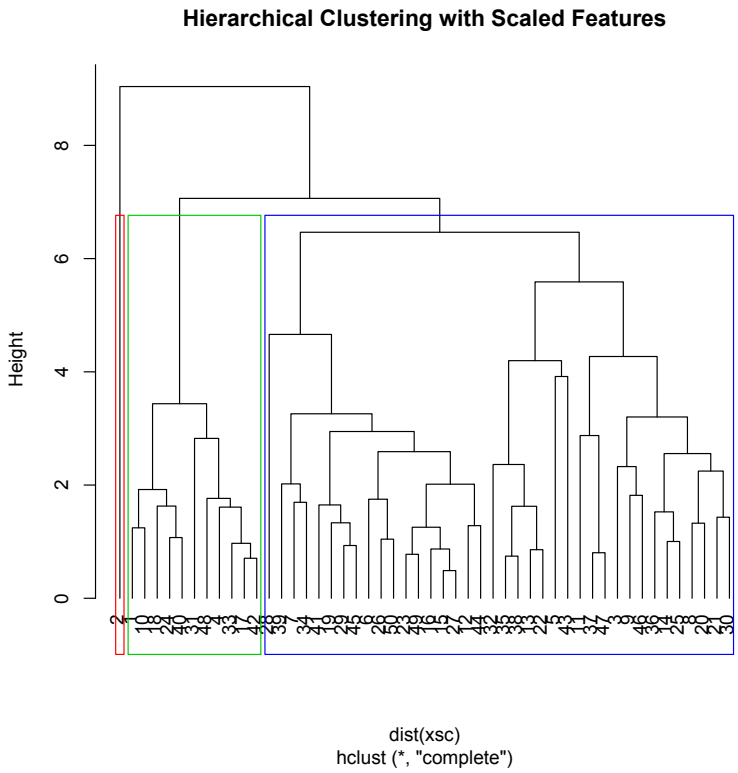
After loading data, did a sanity check by checking the NA values. Plotted this correlation plot to see the relation between data.



It was given in the question not to use the state names in the modeling. I removed the state name from the data and kept it other variable.

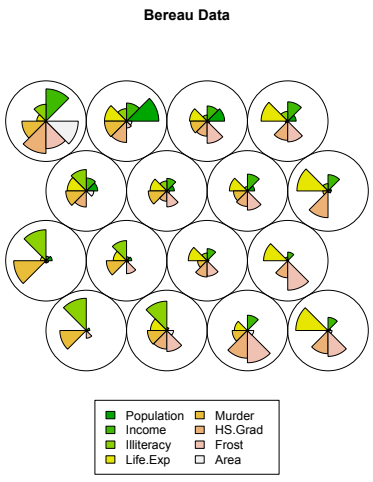
```
 [1] "Alabama"       "Alaska"        "Arizona"       "Arkansas"      "California"    "Colorado"      "Connecticut"   "Delaware"      "Florida"
[10] "Georgia"       "Hawaii"        "Idaho"         "Illinois"      "Indiana"       "Iowa"          "Kansas"        "Kentucky"      "Louisiana"
[19] "Maine"         "Maryland"      "Massachusetts" "Michigan"      "Minnesota"     "Mississippi"   "Missouri"      "Montana"       "Nebraska"
[28] "Nevada"        "New Hampshire" "New Jersey"    "New Mexico"    "New York"      "North Carolina" "North Dakota"  "Ohio"          "Oklahoma"
[37] "Oregon"        "Pennsylvania"  "Rhode Island"  "South Carolina" "South Dakota"  "Tennessee"     "Texas"         "Utah"          "Vermont"
[46] "Virginia"      "Washington"    "West Virginia" "Wisconsin"     "Wyoming"
```

Then, scaled the data by using the scale method. And, did the hierarchical clustering with complete linkage and euclidean distance to the scaled data. Here is the result:

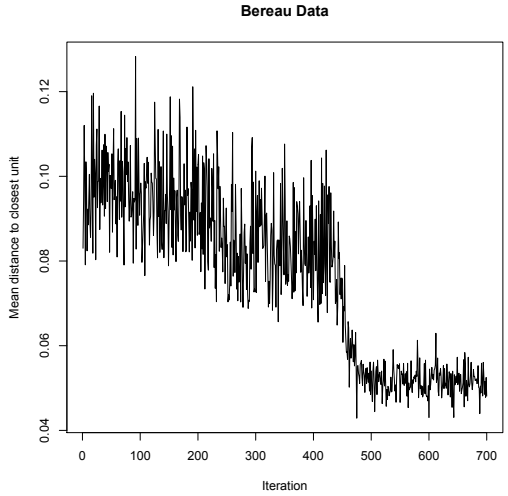


Cut the tree with three clusters horizontally. As here we can that 2 which is Alaska seems to different structure. It is making the cluster alone. And, most of states or region are into one cluster because of scaling. It makes a significant impact on the clustering. So, it is important to scale the variables first if the feature range is much greater than others, then the distance would be bigger and impact on the overall distance is high. Each variable has given an equal importance in the hierarchical clustering as variables scaling is done before the inter-observation dissimilarities are computed.

Then, did the SOM to the data. SOM finds grid structure to see the similar input vector to classify. Here is the result.

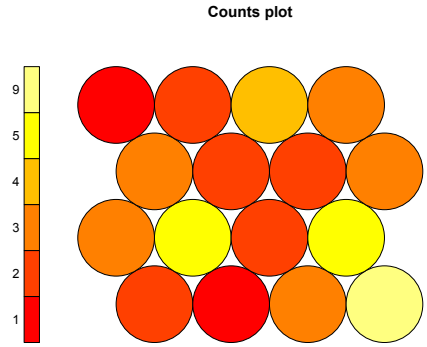


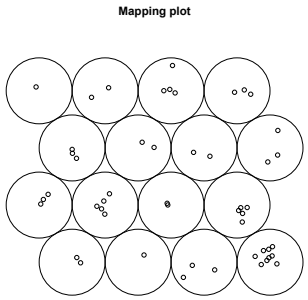
We can see that the population and income is more at the top side of the plot. HS grad and Frost on the bottom of the plot. And, the other plot.



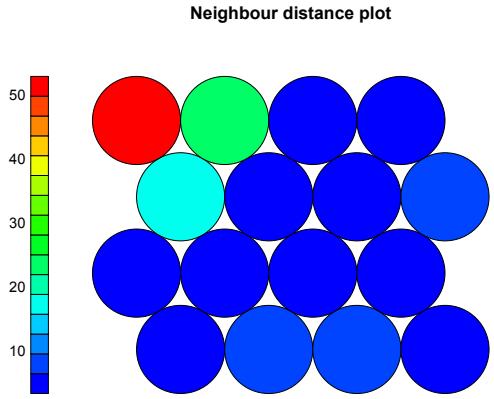
This training iterations progress plot shows the progress over time. Here, the distance is reaching to minimum level. More iterations are required if it keeps on decreasing.

This is the count plot. This to see the number of samples mapped to each node. Red nodes are denoting the less number of distribution together. And, yellow denoting the high number of associated data.



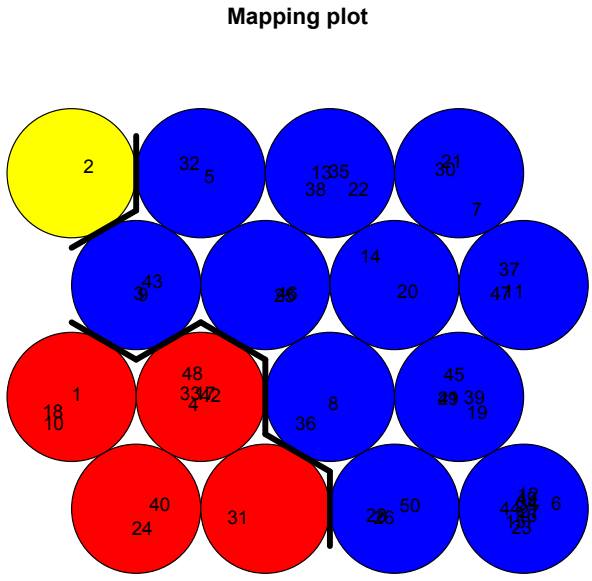


The Neighbor distance Plot:



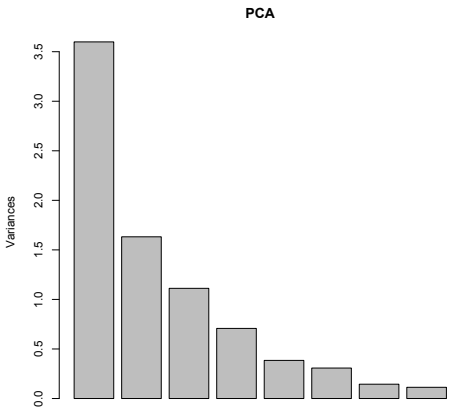
This plot shows the distance between the nodes. The lower value is on the right side and bottom of the plot and the group of similar nodes and value is high on the top left side and has a group of dissimilar nodes. Here, that red node might be having that Alaska which we saw in the result of hierarchical clustering. We can plot the SOM with the found cluster to confirm this.

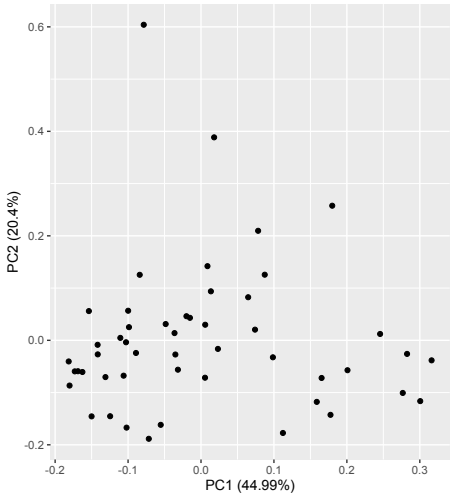
Then, plotted the SOM with the found cluster. Here is the result:

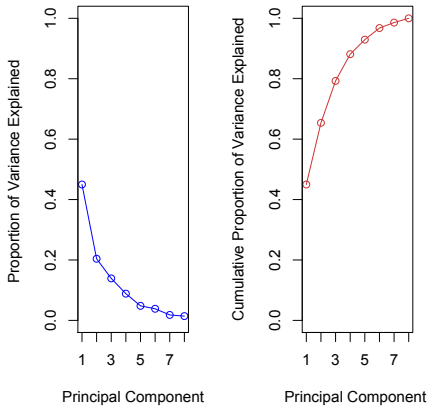


It is expected to have this result as dataset is small. And, We can see that 2 which is Alaska making a single cluster here also. It is supporting the result of hierarchical clustering in part A.The number of states in each cluster from the both method are same. Generally, hierarchical clustering and SOM algorithm give the better results when used with small dataset.

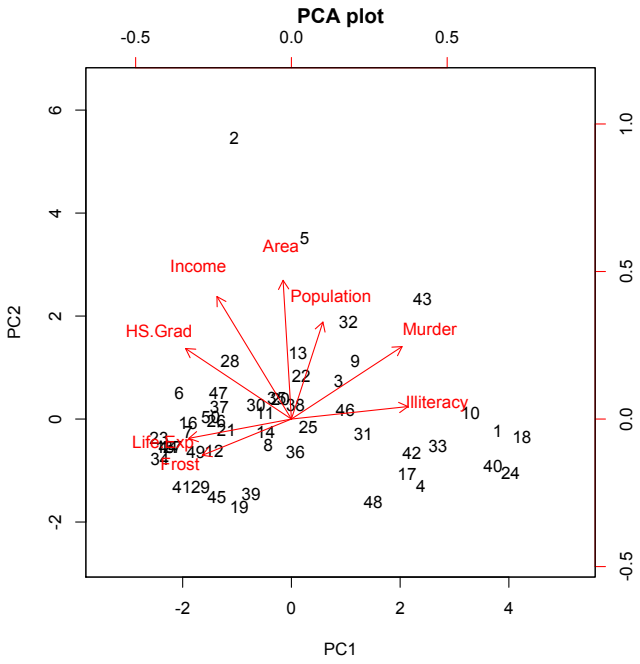
Then, went for the Gaussian Graphical model using the graphical lasso. To start with graphical lasso, performed PCA on the data. Here is the result of PCA.







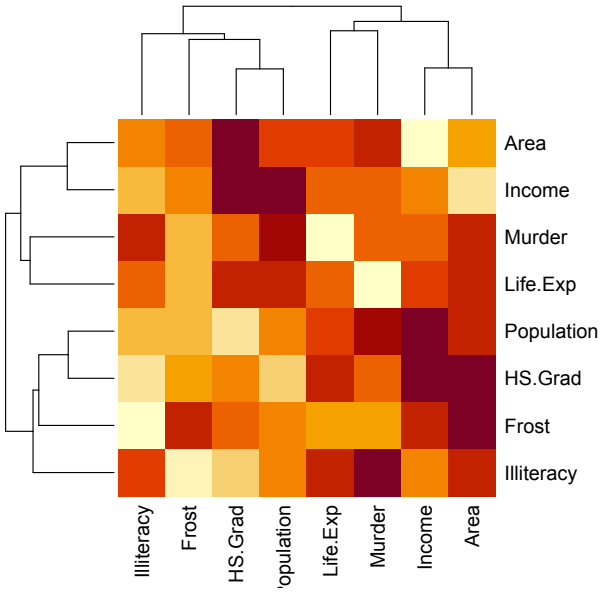
From the above result, we can see that first three components are contributing the 80% of the variance in the data. And, last are not contributing negligible of the variance.



From the above result, we can see that Frost and Murder/Population/Illiteracy are in the opposite direction. They are negatively correlated with each other. Area,Population are in positive correlation.

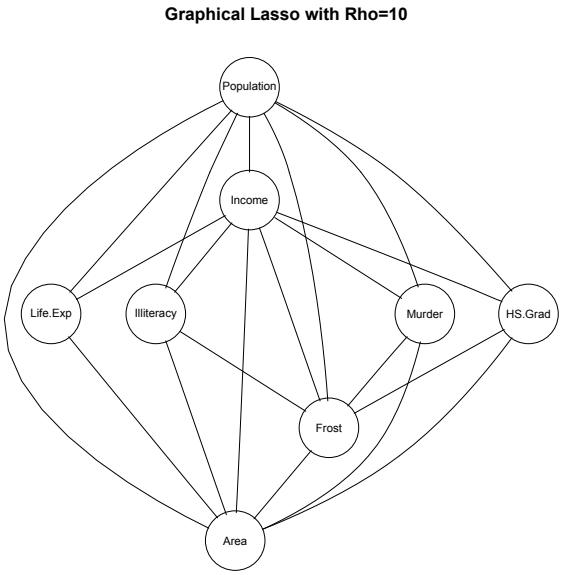
Alaska 2 is in the direction of Area. It has a large area value. It is an outlier in the data. This we also observed in the hierarchical clustering and SOM. We can remove this value.

Here is the result of partial correlation.



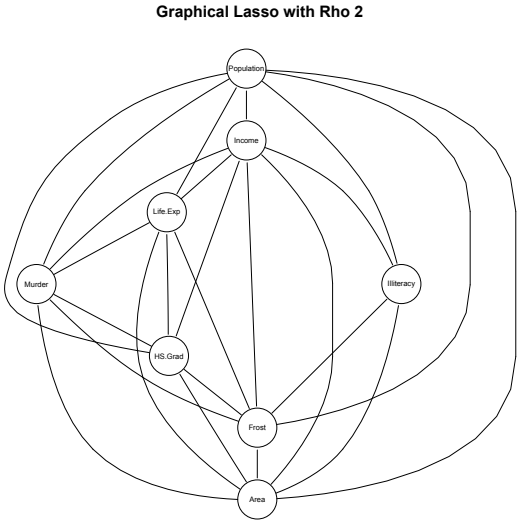
Then implemented the graphical lasso to understand the distribution of the data. Glasso method tries to estimate sparse matrix using the regularization penalty. This is based on the L1 regularization. It introduces sparsity by squishing some of the features to zero.

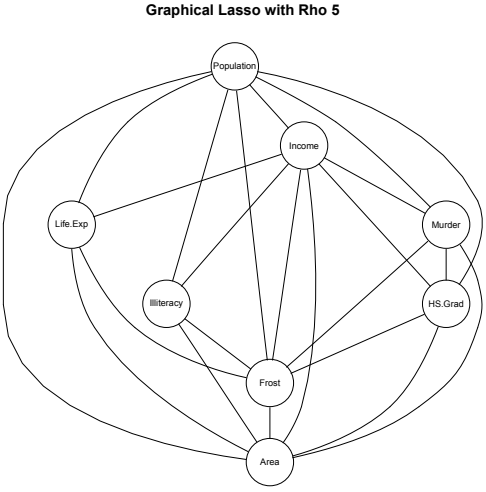
The graphical lasso result:



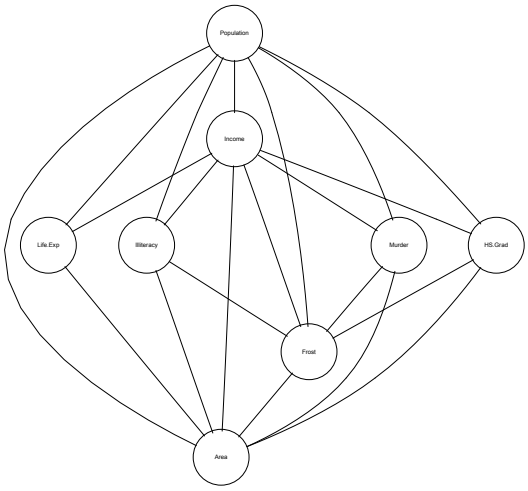
Here the graph is more dense. The features are related to each other.

Then, Tried with different value of rho eg 2,5,15,30,50,100.

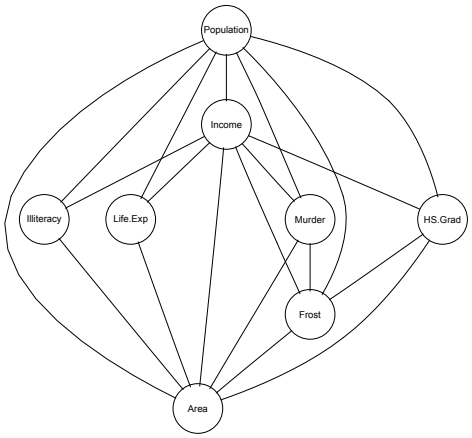




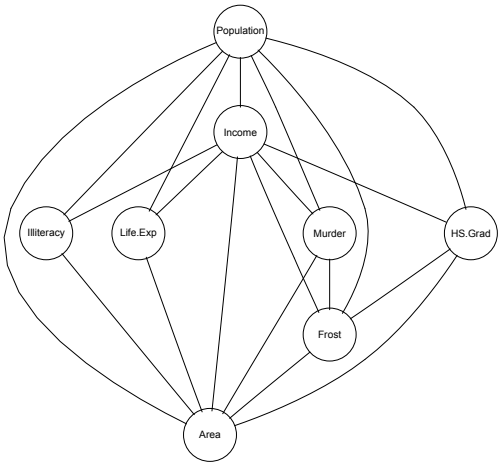
Graphical Lasso with Rho 15



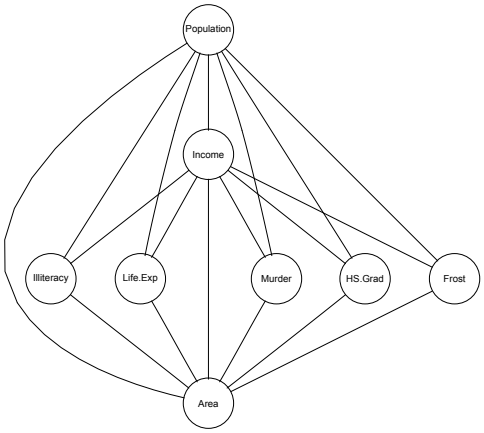
Graphical Lasso with Rho 30



Graphical Lasso with Rho 50



Graphical Lasso with Rho 100



For the first few penalty the graph comes out to be dense. As we increase the rho i.e. regularization penalty, the graph starts getting sparse. The edges with less weight are penalized and their values get squished to zero and making the graph sparse. And, we can see that for rho 30 and 50, most of features are connected with each other.

We see that the Frost, Population, Area, Murder, HS Grad are more contributing features in the data. This we also saw in the SOM.

Advantages of Clustering to Graphical model.

The graphical lasso models maximizes the gaussian log likelihood, It adds the l1 penalty. The graphical lasso performs two steps. First, it does single linkage hierarchical clustering on the features to find out the connected components. And then it adds a l1 penalty and log likelihood is maximized on the subset variables within each connected components. Single linkage usually doesn't perform well. Cluster graphical lasso is being used clustering the features and performing the lasso on the subset features within each cluster.

Clustering finds out the behavior of the each data point in the data and clusters out them based on their similar or dissimilar properties. Clustering model tells us about the feature space. We can look out the performance of each of the feature in the data.

Graphical model tells us about the feature space and it is not able to find the unknown pattern of the data. Thus unable to find the differences in each of the data point.

So, clustering comes out to be better as compared to graphical model as it looks the each individual datapoint.