SAINEY MANGA
MASTER’S IN DATA SCIENCE AND ECONOMICS, UNIVERSITY OF MILAN

HAPPINESS AND ECONOMIC INDICATORS

Author Note

This is a final term paper for Probabilistic Modeling course of academic year 2020/21

Course Coordinator: **Prof. FEDERICA NICOLUSSI**

1. Dataset Description

The dataset "World happiness report 2019" is downloaded from Kaggle https://www.kaggle.com/unsdsn/world-happiness?select=2019.csv. It is a landmark survey of the state of global happiness released at the United Nations at an event celebrating International Day of Happiness on March 20th. The attributes following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher. Basically, the dataset consists of happiness score of a country and other economic variables. The following columns: GDP per Capita, Life Expectancy, Freedom, Generosity, Trust Government Corruption describe the extent to which these factors contribute to evaluating the happiness in each country.

Table 1. first six rows of the data

Score	GDP.per.capita	Social.support	Life.exp.	freedom	generosity	Perc.of.corrupt.
7.769	1.340	1.587	0.986	0.596	0.153	0.393
7.600	1.383	1.573	0.996	0.592	0.252	0.410
7.554	1.488	1.582	1.028	0.603	0.271	0.341
7.494	1.380	1.624	1.026	0.591	0.354	0.118
7.488	1.396	1.522	0.999	0.557	0.322	0.298
7.480	1.452	1.526	1.052	0.572	0.263	0.343

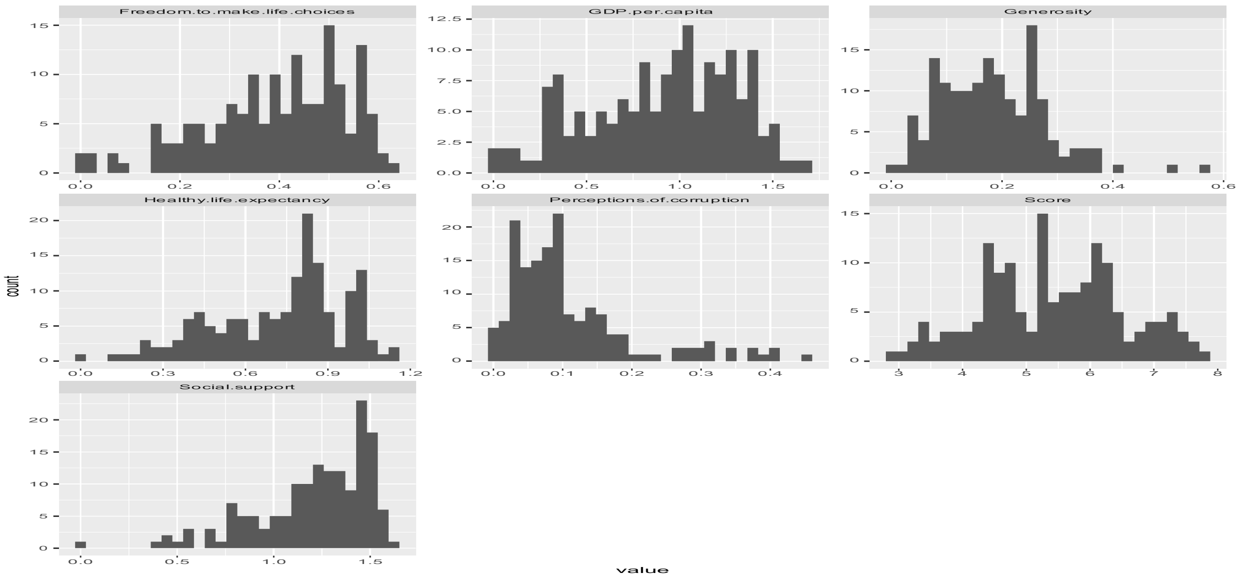


figure 1: histograms of respective variables

2. Methodology:

In this report, we will use Gaussian graphical models (for undirected graphs) to investigate the mutual associations amongst variables. The dataset applied is a recent one and to the knowledge of the author, there has not been a similar methodology in the literature that investigated relations amongst these specific indicators. However, John F. Helliwell (university of British Columbia) etal.; Durand, M & Exton, c(2019)- adopting a well-being approach in central government policy, both found significant correlation of all variables with happiness except corruption and generosity. They both applied regression techniques.

2.1 TOPIC OVERVIEW:

Gaussian graphical models provide a framework for modelling how variables are mutually related. Consider a random vector $y = (y_1,. . . , y_d)$ following a multivariate normal $N_d (\mu, \Sigma)$ distribution. The key quantity in Gaussian graphical models is the inverse of the covariance matrix $K = \Sigma^{-1}$ known as the concentration matrix. The partial correlation between y_u and y_v given all other variables can be simply derived from K.

In GGMs, the primary objective is to characterize conditional dependencies between pairs of variables. This is typically accomplished by determining which off-diagonal elements in the inverse of the covariance matrix (i.e., the precision matrix) are non-zero. This is referred to as covariance selection (Dempster, 1972; Peng et al., 2009). When these elements are standardized and the sign reversed, this results in partial correlations (ρ) that are pairwise relations controlling for all other variables in the model (Baba and Sibuya,2005; Baba et al., 2004). The non-zero partial correlations constitute the underlying network structure of conditionally dependent effects. In a graph, they are represented visually as “connections” that link variables (Jones et al., 2018).

2.2 Hypothesis Testing and Regression/Estimation

We use the likelihood ratio test by applying the ciTest_mvn() function to investigate the conditional independence of key variables on happiness score (which is the deviance). Alternatively, we applied the F-statistics to compare our results with the deviance methods. We also applied the asymptotic normality of Fisher’s z transform of the partial correlation by using the gaussCItest() in pcalg package.

On the estimation part, we applied a multiple linear regression using the concentration matrix since there is a close connection between the two.

2.3 Model Selection

2.3.1. Stepwise Methods: This method can be implemented using the AIC or BIC criteria to perform a search in the model for the deletion or addition of edges in the graph base on optimality. We applied both the different methods to see possible variants in our results. The gRim package in R will make the implementation of the method easier.

2.3.2. Convex Optimization: This gives a fast technique to find the Gaussian graphical model that maximizes a log-likelihood for K which is penalized by the L1-norm |K|; this is implemented in the package glasso. The smaller the value of ρ, the denser the graph that results. No penalization occurs for values of ρ close to zero.

2.3.3. Thresholding: A simple and apparently naive method for selecting a UGGM is to set a specific threshold for the partial correlations, so edges are removed for all partial correlations less than a given value. Basically, we set a threshold where any entry that is less than the threshold is round down to zero and otherwise round up to 1.

2.3.4. Simultaneous P-values: It follows that if we construct a graph G(α) by including precisely those edges with puv< α, then the probability of incorrectly including one or more edges is less than or equal to α. Or to put this another way, the probability that G(α) is not a subgraph of the true model is less or equal to α. We use two α thresholds as suggested in the literature so as to partition the simultaneous p- values into three sets: a significant set S, an intermediate set I and a non-significant set N (hence the spicy acronym, SIN).

2.4. Final Model Selection

Finally, we implemented a model that took all the common edges that were included in the previous methods and plot our final graph to provide a decisive conclusion.

3. RESULTS:

Table 2: The concentration matrix

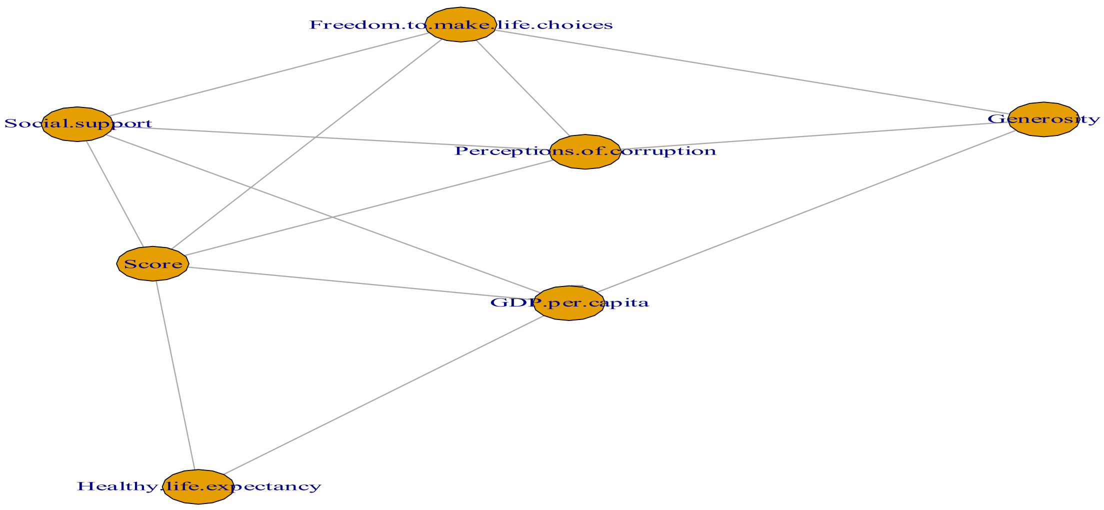
	Score	GDP.per.capita	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Score	368	-285	-414	-397	-535	-180	-358
GDP.per.capita	-285	2831	-815	-2159	526	969	838
Social.support	-414	-815	3541	-544	-830	373	1706
Healthy.life.expectancy	-397	-2159	-544	6561	292	31	-333
Freedom.to.make.life.	-535	526	-830	292	8499	-1933	-2993
Generosity	-180	969	373	31	-1933	13666	3852
Perceptions.of.corruption	-358	-838	1706	-333	-2993	-3852	16469

Table 3: Partial Covariance Matrix

	Score	GDP.per.capita	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Score	100	28	36	26	30	8	15
GDP.per.capita	28	100	26	50	-11	-16	12
Social.support	36	26	100	11	15	-5	-22
Healthy.life.expectan	26	50	11	100	-4	0	3
Freedom.to.make.life.	30	-11	15	-4	100	18	25
Generosity	8	-16	-5	0	18	100	26
Perceptions.of.corrupti	15	12	-22	3	25	26	100

From table 1, we have seen that most of our variables were measured in a different scale. Our target variable happiness score was measured on a scale out of ten, GDP per capita is a ratio of a countries GDP and population, and most other variables measured in terms of percentages. Therefore, it is prudent we use the partial covariance instead of the concentration in order to link the connections amongst the indicators. At the initial stage, we connect nodes that have an entry less than 12 in the partial covariance matrix. This is also supported by our a priori intuition about the relationship between the various indicators.

Figure 2: ***Graph built base on the partial covariance***



The graph in figure 2 was built base on the partial covariance matrix. Variables represented in the graph as nodes are not connected if their entry in the matrix is significantly low, indicating that corresponding variables are conditionally independent. Alternatively, we say an edge does not exist between two variables if their partial covariance is significantly low. From the graph, all other variables have a connection with the happiness score except Generosity. All other variables also have a connection with GDP per capita except Freedom to make life choices.

Hypothesis Testing and Regression/Estimation:

```
Testing Score _|_ GDP.per.capita | Social.support Healthy.life.expectancy Freedom.to.make.life.choices Generosity Perceptions.of.corruption
Statistic (DEV):  12.687 df: 1 p-value: 0.0004 method: CHISQ

Testing Score _|_ GDP.per.capita | Social.support Healthy.life.expectancy Freedom.to.make.life.choices Generosity Perceptions.of.corruption
Statistic (F):  12.624 df: 1 p-value: 0.0005 method: F

Testing Score _|_ Generosity | Healthy.life.expectancy Social.support GDP.per.capita Freedom.to.make.life.choices Perceptions.of.corruption
Statistic (DEV):  1.010 df: 1 p-value: 0.3148 method: CHISQ

Testing Score _|_ Perceptions.of.corruption | Generosity Healthy.life.expectancy Social.support GDP.per.capita Freedom.to.make.life.choices
Statistic (DEV):  3.329 df: 1 p-value: 0.0681 method: CHISQ
```

From the above conditional tests, both the P-values of Score on GDP per capita (including the F-statistics) are not plausible. We could deduce from the test of Score on Generosity is significantly plausible, also that on Perception of corruption is arguably plausible.

Applying the **Asymptotic normality of Fisher's z transform of the partial correlation** using the gaussCItest() function gives us an estimate of **0.000135792** which is in agreement with our first two cases.

The matrix below shows the regression estimate using the concentration matrix.

```
GDP.per.capita      Social.support    Healthy.life.expectancy Freedom.to.make.life.choices Generosity Perceptions.of.corruption
     0.7753716          1.1241916            1.0781427               1.4548324                 0.4897834        0.9722802
```

From our estimation above, it is only Generosity that has a lower prediction of Happiness Score. Healthy life expectancy, Social support and Freedom to make life choices all have very high prediction for Happiness Score. GDP per capita has a slightly lower prediction than the ones mentioned. This might be attributed to the distribution of GDP in a particular country, for example. When a larger part of GDP per capita is own by upper percentile of income earners.

One of the fears of multiple regression is spuriousness as it is important to note that edges in GGM may also result from latent variables. Let us see what happens in the models to follow.

Model Selection

Figure 3: ***Stepwise based on AIC***

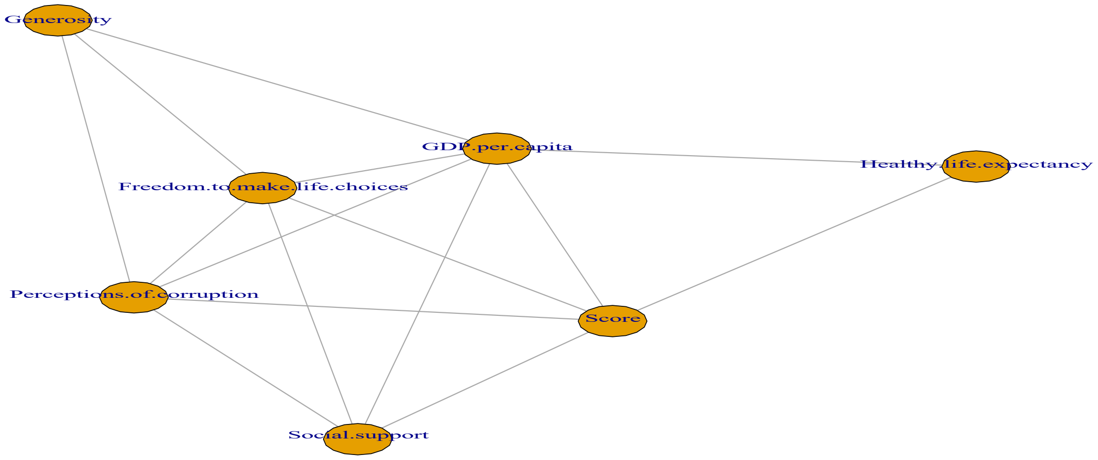
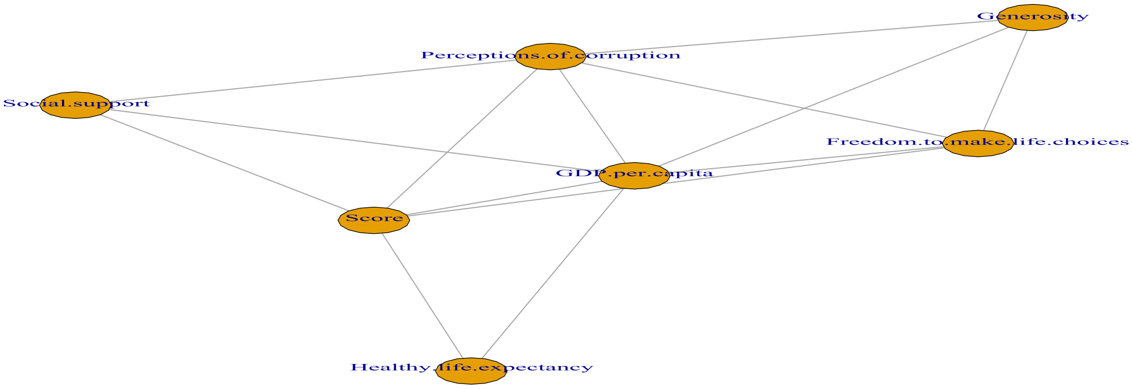


Figure 4: ***Stepwise based on BIC***



From the two methods, we notice that there is one main difference. The edge between Social support and Freedom to make life choices from AIC have been deleted by the BIC. Intuitively this suggests (in BIC) that the two cannot be used as a prediction for each other.

Figure 5: **stepwise selection using significance tests**

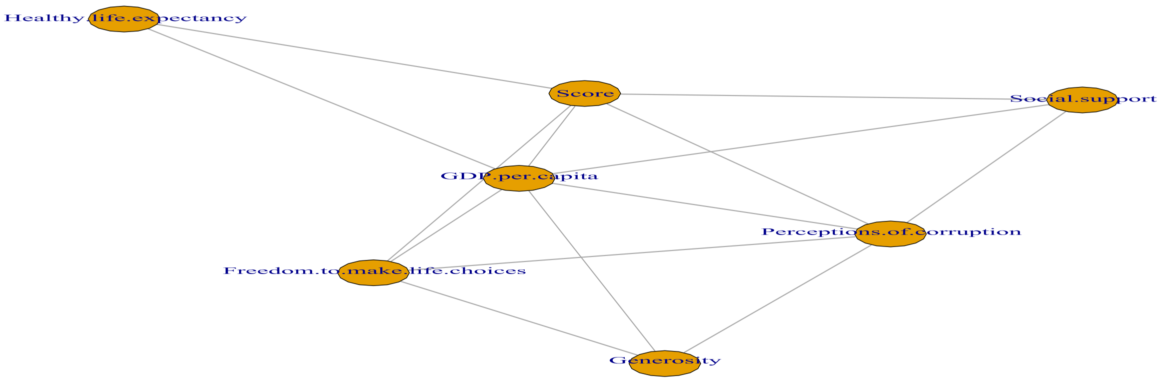
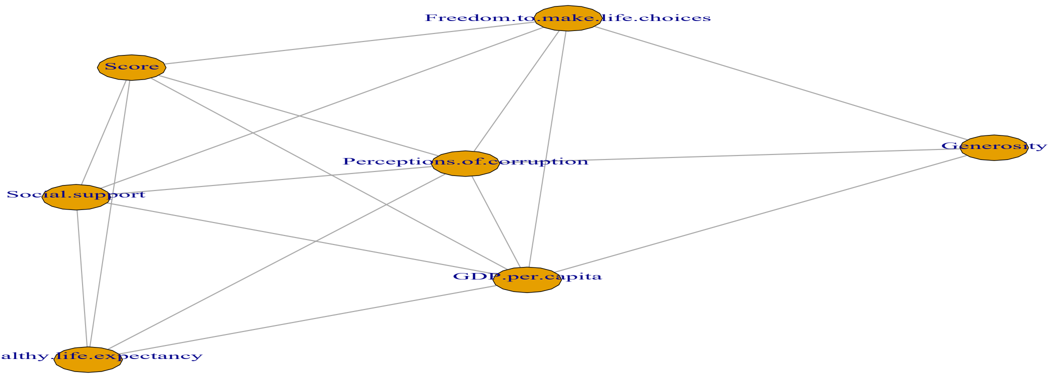


Figure 6: **headlong stepwise selection using the AIC criterion**



In figure 6, the headlong stepwise selection added two edges to Healthy life expectancy in contrast to figure 5. One of the edges that somehow does not conform to our intuition is the connection between Healthy life expectancy and perception of corruption, which in previous methods never connect.

Figure 7: ***Convex Optimization Method***

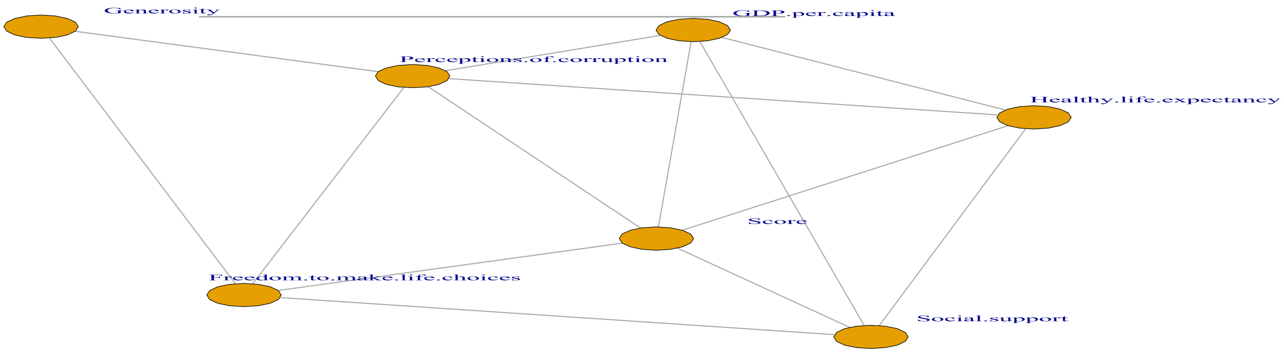
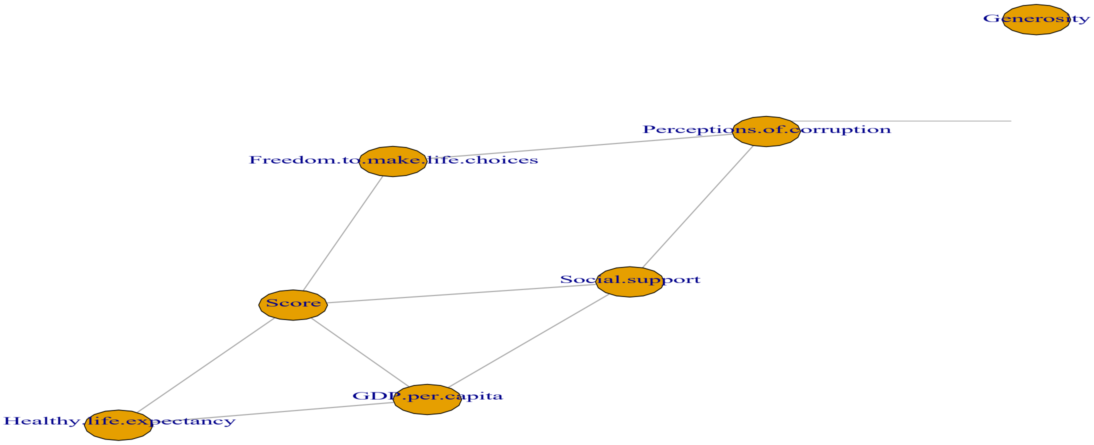


Figure 8: ***Threshold Method***



The Convex optimization is more or less the same as the previous models seen before. Interestingly, the thresholding of correlation actually has resonated economic intuitions after setting a threshold of 0.2, any entry less than or equal to threshold is rounded to zero, otherwise to 1. Only one edge is connected to Generosity.

Figure 9: ***Simultaneous P-values***

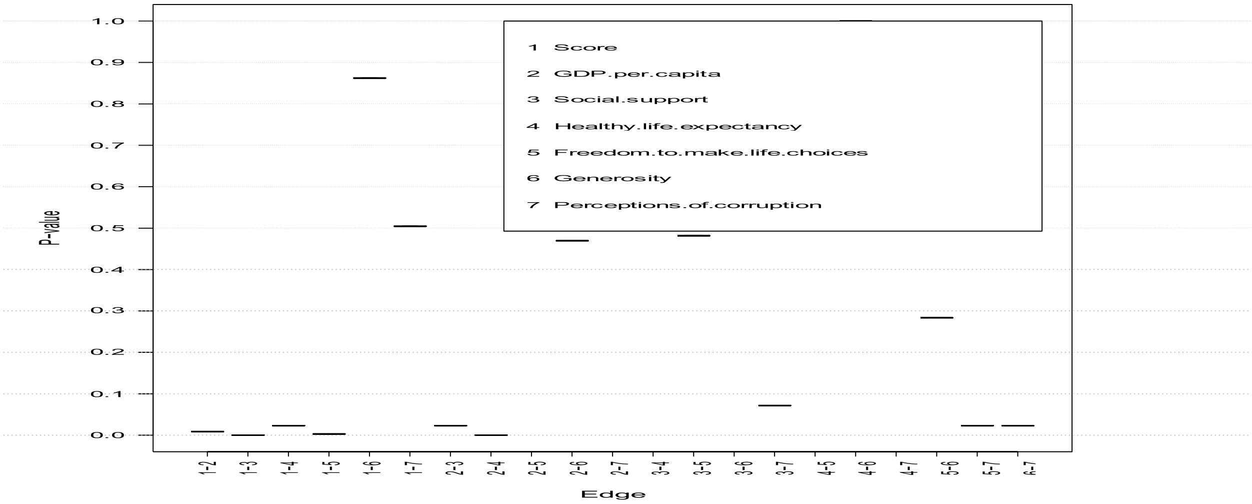
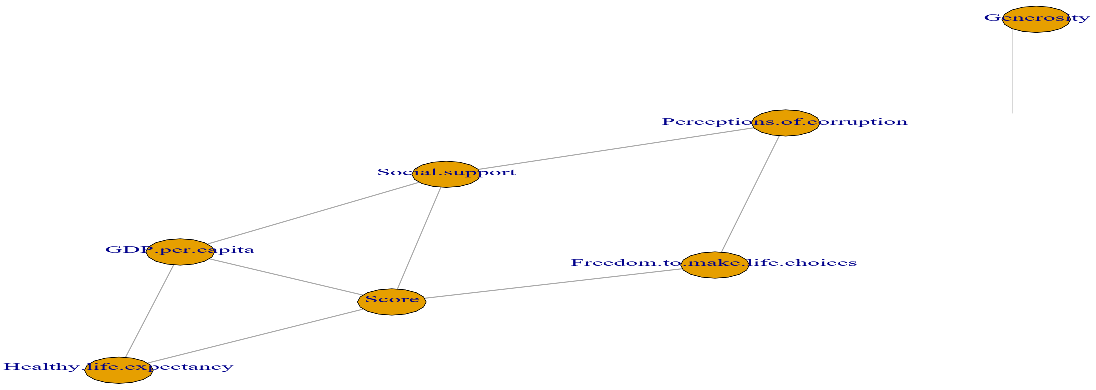


*Figure 10****: Thresholding the simultaneous P-values***



It so happens that both the Thresholding and the Simultaneous p-values method produce the same results. This is so because we set a threshold of 0.2 on the correlation and of 0.1 on the p-values. This is literally the same as there are no entry that are exactly 0.2, so setting 0.2 brings values from 0.1and below.

Figure 11: ***Final Model Selection from common edges of all the previous Models***.

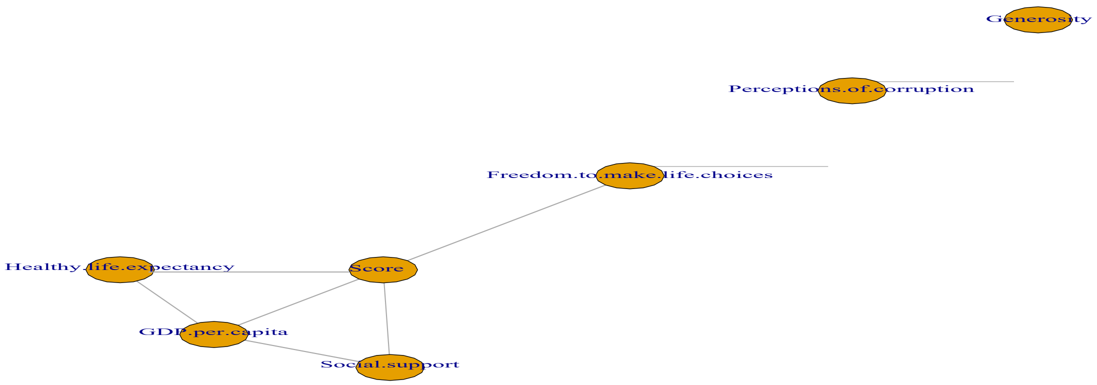


Figure 11 summarizes all common edges from our previous models into one graph. This leads us to a conclusion that Happiness Score is only affected by GDP per capita, Healthy life expectancy, Social Support and Freedom to make life choices. Therefore, based on the economic variable our data provides, we can say that what makes people of a country happy is the above mention variables, thus answering our main research questions.

CLASSIFICATION ERROR:

Table 4:

	Variable	**RMSE**	**R2**
1	**Score**	**0.469**	**0.779**
2	**GDP.per.capita**	**0.472**	**0.776**
3	**Social.support**	**0.562**	**0.682**
4	**Healthy.life.expectancy**	**0.511**	**0.737**
5	**Freedom.to.make.life.choices**	**0.769**	**0.405**
6	**Generosity**	**0.904**	**0.177**
7	**Perceptions.of.corruption**	**0.824**	**0.316**

The validation test through the mgm package, with an interaction of k=2 and a cross validation produced the results above. The variables Score, GDP per capita, Social support and Healthy life expectancy are associated with a relatively low RMSE and a high R^2. This indeed indicates a significant predictability and a strong explanation of variance amongst these variables. The remaining variables have much higher values in RMSE estimates and lower values in R^2. These results somewhat confirm our previous findings in the previous models about the relations amongst variables and how they connect in our graphical models.

4. Conclusion

Although we have applied a different methodology in this report, but our findings are in agreement with the findings of the quoted authors from the literature. Basically, Corruption and Generosity do not have much to explain on Happiness.

Reference:

1. Lecture Notes, Prof. FEDERICA NICOLUSSI, University of Milan
2. Hojsgaard, David Edwards, and Steffen Lauritzen, Graphical models with R, Springer Science & Business Media, 2012
3. Koller, D., Friedman, N., and Bach, F. (2009) Probabilistic graphical models: principles and techniques
4. Donald R. Williamsa, Joris Mulder, Bayesian Hypothesis Testing for Gaussian

   Graphical Models: Conditional Independence and Order Constraints(2020)

APPENDICES: R-CODE

```
>library('dplyr')
>library('gRbase')
>library('igraph')
>library('gRim')
>library('RBGL')
>library('pcalg')
>library('Rgraphviz')
>library('glasso')
>library('SIN')
>library('mvtnorm') # multivariate Gaussian distribution library('matrixcalc') # operations with matrices
>happiness<- read.csv("/Users/saineymanga/Desktop/2019.csv")
>any(is.na(happiness))
>attach(happiness)
>head(happiness)
>clean_happiness<-select(happiness, -c(Overall.rank, Country.or.region))#####droping off the country of origin and the overall rank.
>head(clean_happiness)
####Data visualization)
>library('purrr')
>library('tidyr')
>library('ggplot2')

>clean_happiness %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

####Covariance and concentration/partial covariance
>S.happiness <- cov.wt(clean_happiness,method="ML")$cov
>K.happiness <- solve(S.happiness) # inverse
>round(K.happiness*100)
>PC.happiness <- cov2pcor(S.happiness)
>round(PC.happiness*100)
```

```
    ####ploting the graphical representation of the model
    >gen.happiness <- cmod(~Score*GDP.per.capita*Social.support*Perceptions.of.corruption+
                Score*GDP.per.capita*Healthy.life.expectancy+
                Score*Social.support*Freedom.to.make.life.choices+
                Generosity*GDP.per.capita*Perceptions.of.corruption+
                Generosity*Freedom.to.make.life.choices*Perceptions.of.corruption,data=clean_happiness)
    >gen.happiness
    >plot(as(gen.happiness,"igraph"))

    >gen.happiness.graph<- ug(~Score*GDP.per.capita*Social.support*Perceptions.of.corruption+
                  Score*GDP.per.capita*Healthy.life.expectancy+
                  Score*Social.support*Freedom.to.make.life.choices+
                  Generosity*GDP.per.capita*Perceptions.of.corruption+
                  Generosity*Freedom.to.make.life.choices*Perceptions.of.corruption)
    >plot(as(gen.happines.graph,'igraph'))
    ##In order to make faster the algorithm it is advisable to pass the cliques
    >cl.happiness <- getCliques(gen.happiness.graph)
    >gen.happiness.cl <- cmod(cl.happiness,data=clean_happiness)
    >gen.happiness.cl
    ####Hypothese testing
    >ciTest_mvn(list(cov=S.happiness,n.obs=nrow(clean_happiness)),#### the independence of happiness
given with GDp given other variable

set=~Score+GDP.per.capita+Social.support+Healthy.life.expectancy+Freedom.to.make.life.choices+Generosity+
         Perceptions.of.corruption)
    >ciTest_mvn(list(cov=S.happiness,n.obs=nrow(clean_happiness)),#### the independence of happiness
given with GDp given other variable

set=~Score+GDP.per.capita+Social.support+Healthy.life.expectancy+Freedom.to.make.life.choices+Generosity+
         Perceptions.of.corruption, statistic = 'F')

    >ciTest_mvn(list(cov=S.happiness,n.obs=nrow(clean_happiness)),

set=~Score+Healthy.life.expectancy+Social.support+GDP.per.capita+Freedom.to.make.life.choices+
         Perceptions.of.corruption+Generosity)
    >ciTest_mvn(list(cov=S.happiness,n.obs=nrow(clean_happiness)),
set=~Score+Generosity+Healthy.life.expectancy+Social.support+GDP.per.capita+Freedom.to.make.life.choices+
        Perceptions.of.corruption)
    >ciTest_mvn(list(cov=S.happiness,n.obs=nrow(clean_happiness)),
       set=~Score+
Perceptions.of.corruption+Generosity+Healthy.life.expectancy+Social.support+GDP.per.capita+
        Freedom.to.make.life.choices)

    ###Asymptotic normality of Fisher's z transform of the partial correlation
    >cS<-cov2cor(S.happiness)
    >gaussCItest(1,2,3:6,list(C=cS,n=nrow(clean_happiness)))
    >gaussCItest(7,2,c(1,3,4,5,6),list(C=cS,n=nrow(clean_happiness)))

    #####CONCENTRATION AND REGRESSION
    >-K.happiness[1,-1]/K.happiness[1,1]
    >1/K.happiness[1,1] ## residual variance of score
    ##### MODEL SELECTION#####
    >sat.happiness <- cmod(~.^., data=clean_happiness)
    >aic.happiness<- stepwise(sat.happiness)##base on AIC, note
    >plot(as(aic.happiness,"graphNEL"),"fdp")
    >plot(as(aic.happiness, 'igraph'))
    >bic.happiness<-stepwise(sat.happiness,k=log(nrow(clean_happiness))) ###base on BIC
    >bic.happiness
    >plot(as(bic.happiness,"graphNEL"),"fdp")
    >plot(as(bic.happiness, 'igraph'))
    ##stepwise method##
    >test.happiness <- stepwise(sat.happiness, details=1,"test")###Stepwise using the default significance
    >plot(test.happiness,"neato")
    >plot(as(test.happiness, "igraph"))###using "igraph" to improve visibility
    >ind.happiness<-cmod(~.^1,data=clean_happiness)#####another stepwise method using gRim package
    >set.seed(123)
    >forw.happiness<-stepwise(ind.happiness,search="headlong",
           direction="forward",k=log(nrow(clean_happiness)),details=0)
    >forw.happiness
    >plot(forw.happiness,"neato")
    >plot(as(forw.happiness, 'igraph'))
    ##convexx optimization method##
```

```
>res.lasso<-glasso(cS,rho=0.1)
>AM <- res.lasso$wi != 0
>diag(AM) <- F
>g.lasso <- as(AM, "graphNEL")
>nodes(g.lasso)<-names(clean_happiness)
>glasso.happiness <- cmod(edgeList(g.lasso),data=clean_happiness)
>plot(as(glasso.happiness,"igraph"))
##thesholding method
>threshold <- .2
>Z <- abs(PC.happiness)
>Z[Z<threshold] <- 0
>diag(Z)<-0
>Z[Z>0] <- 1
>g.thresh<-as(Z, "graphNEL")
>thresh.happiness <- cmod(edgeList(g.thresh), data=clean_happiness)
>thresh.happiness
>plot(thresh.happiness, "neato")
>plot(as(thresh.happiness, "igraph"))### using 'igraph' to improve the label visibility
##Simultaneous p-values
>psin.happiness<-sinUG(S.happiness,n=nrow(clean_happiness))
>plotUGpvalues(psin.happiness)
>gsin.happiness <- as(getgraph(psin.happiness, 0.1), "graphNEL")##thresholding the simultaneous p-
values
>plot(gsin.happiness, "neato")
>plot(as(gsin.happiness, "igraph"))### improving the label visibility
####COMMON EDGES###
>commonedges.happiness<-intersection(as(aic.happiness,"graphNEL"), as(bic.happiness,"graphNEL"))
>othermodels<-list(test.happiness,forw.happiness, thresh.happiness,gsin.happiness,glasso.happiness)
>othermodels<-lapply(othermodels, as, "graphNEL")
>for(ii in 1:length(othermodels)){
    commonedges.happiness<-intersection(commonedges.happiness,othermodels[[ii]])
}
>plot(commonedges.happiness,"fdp")
>plot(as(commonedges.happiness, "igraph"))###improving visibility
```