Assignment 5

Network Analysis 2017

Part 1: Conceptual

Often, when networks are formed on symptom data, the data is ordinal and highly skewed. For example, an item "do you frequently have suidical thoughts" might be rated on a three point scale: 0 (not at all), 1 (sometimes) and 2 (often). Especially when the sample is based on the general public, we often see that the majority of people respond with 0 and only few people respond with a 2. This presents a problem for network estimation, as such data is obviously not normally distributed.

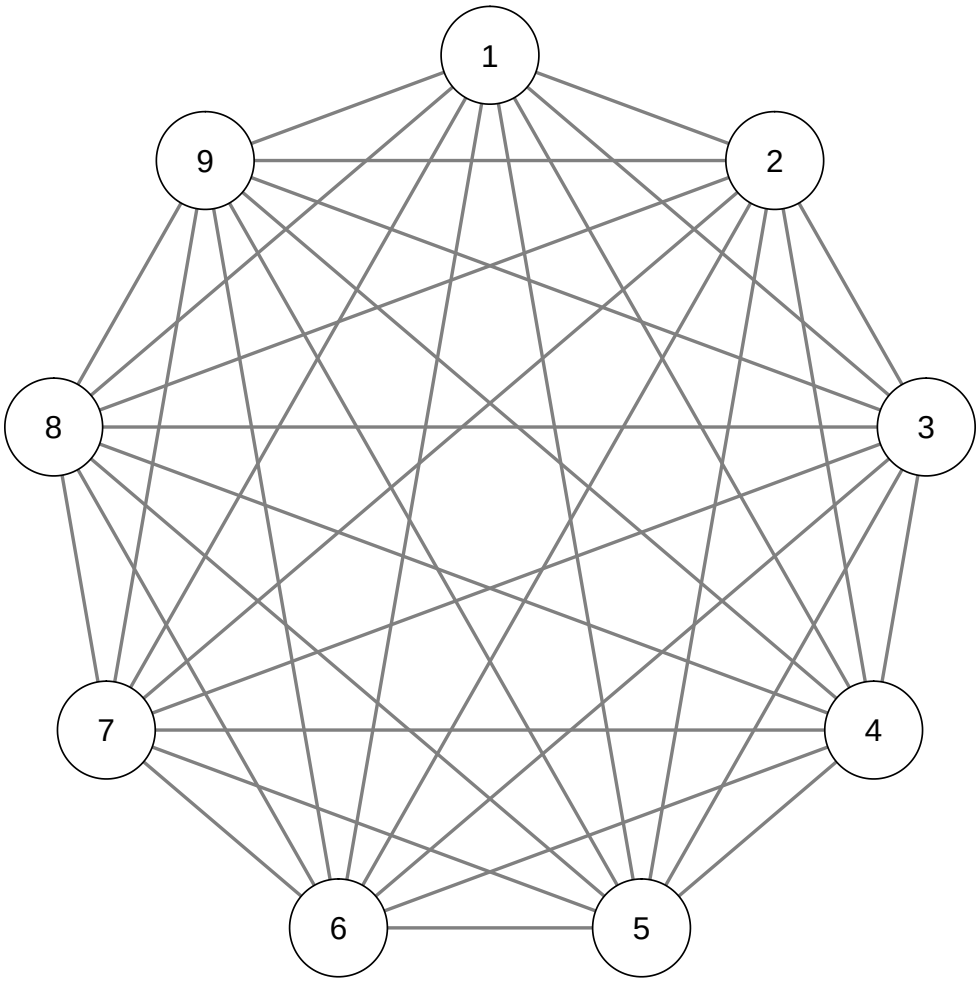
Exercise 1 (2 points)
Which of the following methods would you prefer to analyze such highly skewed ordinal data?

- First binarize the data (e.g., dichotomize everything above 0 to be a 1), and then estimate an Ising model.
- Estimate a Gaussian graphical model based on a polychoric correlation matrix.
- Transform the data using the non-paranormal transformation, and then estimate a Gaussian graphical model.
- Treat the data as categorical rather than ordinal, and estimate a mixed model.

Motivate your answer why you would or would not choose for each method. Note that this question does not have a clear correct answer.

■

Given the following GGM:



Exercise 2 (2 points)
Suppose we have limited data generated by the above shown model. Do you expect LASSO estimation with EBIC model selection to be able to retrieve the true network structure? Why (not)? Describe in your answer the assumption(s) made when using LASSO estimation and what you know about the performance of LASSO estimation.

■

Part 2: Data-analysis

In R, run the following code:

```
install.packages("psych")
```

```
library("psych")
data("bfi")
bfiData <- bfi[,1:25]
```

The data frame `bfiData` contains the questions of the `bfi` (Big Five Inventory) data contained in the *psych* package. More information on this dataset can be obtained by running:

```
?bfi
```

The questions are designed to measure five central personality traits: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Opennness. The following table gives the item descriptions:

Item label	Item description	Trait
A1	Am indifferent to the feelings of others	Agreeableness
A2	Inquire about others' well-being	Agreeableness
A3	Know how to comfort others	Agreeableness
A4	Love children	Agreeableness
A5	Make people feel at ease	Agreeableness
C1	Am exacting in my work	Conscientiousness
C2	Continue until everything is perfect	Conscientiousness
C3	Do things according to a plan	Conscientiousness
C4	Do things in a half-way manner	Conscientiousness
C5	Waste my time	Conscientiousness
E1	Don't talk a lot	Extraversion
E2	Find it difficult to approach others	Extraversion
E3	Know how to captivate people	Extraversion
E4	Make friends easily	Extraversion
E5	Take charge	Extraversion
N1	Get angry easily	Neuroticism
N2	Get irritated easily	Neuroticism
N3	Have frequent mood swings	Neuroticism
N4	Often feel blue	Neuroticism
N5	Panic easily	Neuroticism
O1	Am full of ideas	Opennness
O2	Avoid difficult reading material	Opennness
O3	Carry the conversation to a higher level	Opennness
O4	Spend time reflecting on things	Opennness
O5	Will not probe deeply into a subject	Opennness

We can compute a polychoric correlation matrix based on this data as follows:

```
library("qgraph")
corMat <- cor_auto(bfiData)
```

Next we can use *qgraph* to compute a partial correlation network:

```
qgraph(corMat, graph = "pcor", layout = "spring", cut = 0)
```

We can use the *bootnet* function `estimateNetwork` to automate this process:

```
library("bootnet")
Result_pcor <- estimateNetwork(bfiData, default = "pcor")
plot(Result_pcor, layout = "spring", cut = 0)
```

Exercise 3 (1 point) Obtain the weights matrices from qgraph and bootnet by applying the `getWmat` function to output of both. Confirm that the results are identical (tip: the operator == tests if values in R are equal).

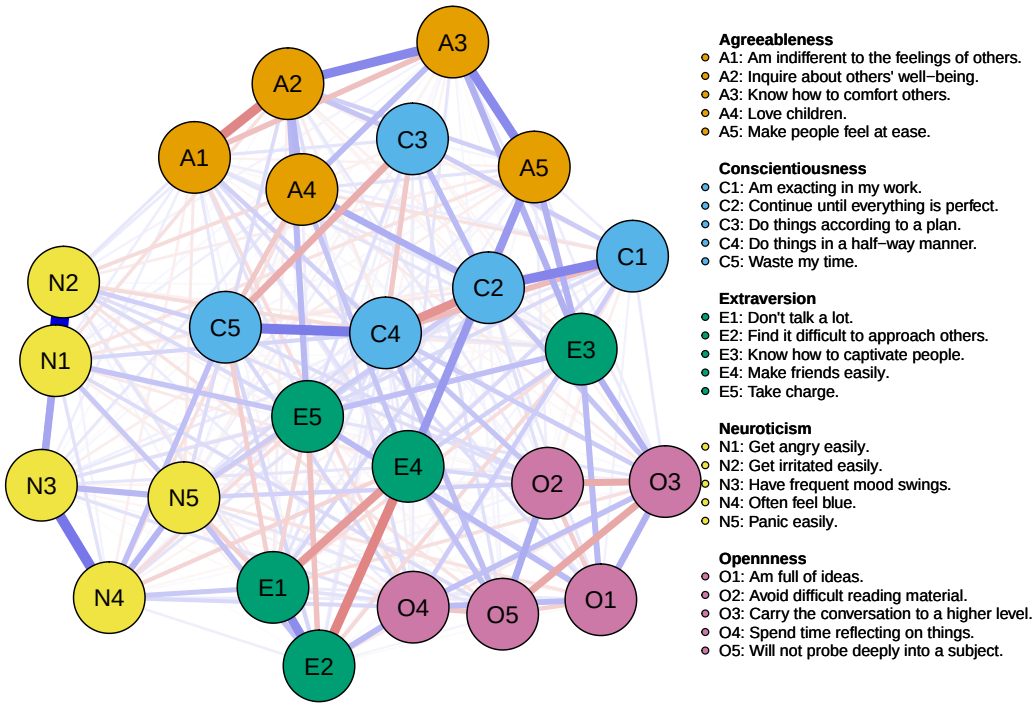
■

We can load for each node the item description and factor the item is aimed to measure as follows:

```
Names <- scan("http://sachaepskamp.com/files/BFIitems.txt",
              what = "character", sep = "\n")
Traits <- rep(c(
  'Agreeableness',
  'Conscientiousness',
  'Extraversion',
  'Neuroticism',
  'Openness'
),each=5)
```

These can be used to plot a legend next to the graph. In combination, we can make the graph friendly to colorblind viewers using the `theme` option:

```
plot(Result_pcor,
     layout = "spring",
     cut = 0,
     theme = "colorblind",
     groups = Traits,
     nodeNames = Names,
     legend.cex = 0.4)
```



Exercise 4 (1 point) What do the arguments `groups` and `nodeNames` do?

■

In `estimateNetwork`, the `fun` argument can be specified a custum function estimating the network from data. To aid the user, several default functions have been built in. For example, `default = "pcor"` specified a function that estimates a partial correlation networks (in the help file this function is called `bootnet_pcor`).

Exercise 5 (1 point) Use the `default` argument in `estimateNetwork` to estimate a partial correlation network using glasso and EBIC model selection.

■

Exercise 6 (1 point) Set the hypertuningparameter γ to 0. Did the network change?

■

Read the literature on Blackboard on how to perform accuracy and stability checks.

Exercise 7 (2 points) Perform a *non-parametric* bootstrap on the EBICglasso network that uses $\gamma = 0.5$, and plot the confidence intervals of the edge-weights.

■

Exercise 8 (2 points) Perform a *case-drop* bootstrap on the EBICglasso network that uses $\gamma = 0.5$, and plot the stability of centrality indices.

■

Exercise 9 (1 point) Give the *CS*-coefficient of the three centrality indices, and explain how this measure can be interpreted.

■

Part 3: SEM re-analysis

Doosje, Loseman, and Bos (2013) analyzed radicalization of Islamic youth in the Netherlands using a large-scale structural equation model (SEM), which can be drawn as a directed causal network:

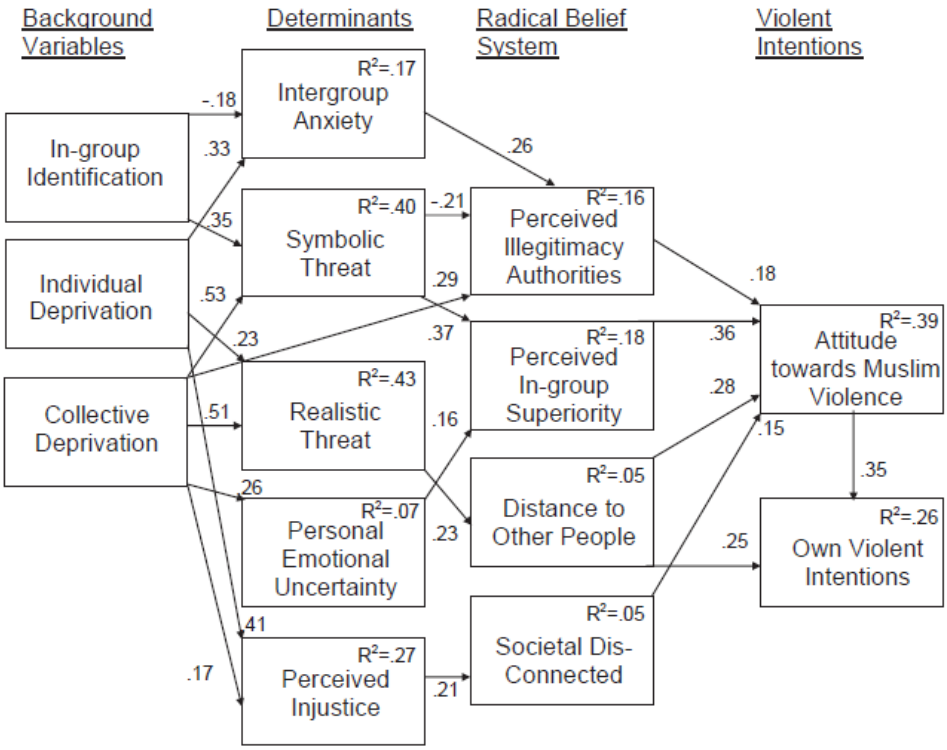


Fig. 1. Final structural equation model. All paths are significant. R^2 = % variance explained.

As with many SEM papers, Doosje et al. (2013) reported the correlation matrix and sample size ($N = 131$) to reproduce their analyses. We can load the correlation matrix in R as follows:

```
source("http://sachaepskamp.com/files/DoosjeData.R")
View(corMat)
```

The `corMat` object now contains the correlation matrix. A SEM analysis as shown above can be used to test a confirmatory model, as is done by Doosje et al. (2013). Suppose however we had no theory and want to exploratory find a good fitting model. SEM is less useful for exploratory model search, as there are many equivalent models possible that fit just as well. For this reason, undirected networks offer a powerfull tool in gaining exploratory insight in which variables might interact.

Exercise 10 (3 points)
Estimate a Gaussian graphical model using LASSO regularization and EBIC model selection on the data from Doosje et al. (2013). Note that you do not have the raw data, so you can not use `estimateNetwork` and need to use the underlying estimation function from the *qgraph* package (`EBICglasso`). Set the EBIC tuning parameter γ to zero. Compare your estimated network to the SEM model reported. Are there edges in your network that are not included in the model shown by Doosje et al. (2013)?

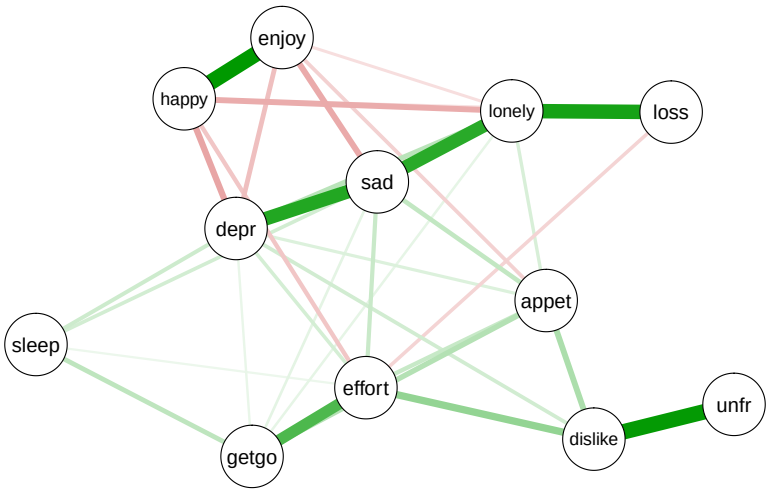
■

Challenge Question

In a recent publication, Fried et al. (2015) analyzed depressive symptoms in elderly people who did or did not lose a spouse. The authors estimated an Ising model, using LASSO regularization with EBIC model selection as explained in the lecture. We have received the network parameters and the thresholds (which encode the difficulty of each item) from the authors:

```
trueNetwork <- read.csv('http://sachaepskamp.com/files/weiadj.csv')[,-1]
trueNetwork <- as.matrix(trueNetwork)
Symptoms <- rownames(trueNetwork) <- colnames(trueNetwork)
Thresholds <- read.csv('http://sachaepskamp.com/files/thr.csv')[,-1]

library("qgraph")
qgraph(trueNetwork, labels = Symptoms, layout='spring')
```



We will use this network to investigate the performance of LASSO estimation. The *IsingSampler* package can be used to simulate data given an Ising model. For example, the following codes simulate a dataset of 5,000 observations:

```
library("IsingSampler")
sampleSize <- 5000
set.seed(1)
newData <- IsingSampler(sampleSize, graph = trueNetwork, thresholds = Thresholds)
```

We can use the IsingFit package to estimate a regularized Ising network (alternatively use bootnet with `default = "IsingFit"`):

```
library("IsingFit")
Res <- IsingFit(newData,progressbar = FALSE,plot = FALSE)
estNetwork <- Res$weiadj
```

Suppose the weights matrix of your estimated network is called `estNetwork`. We can then compute the number of *true positives*—edge weights you estimated to be non-zero that were also non-zero in the original network—as follows:

```
sum(trueNetwork != 0 & estNetwork != 0)

## [1] 64
```

Similarly, the '==' operator can be used instead of '!=' to test if edges are equal to zero, allowing you to obtain:

- The number of *false positives*: edge weights you estimated to be non-zero that were zero in the original network.
- The number of *true negatives*: edge weights you estimated to be zero that were also zero in the original network.

- The number of *false negatives*: edge weights you estimated to be zero that were non-zero in the original network.

Note that in this terminology a "positive" indicates a non-zero edge, not necessarily a positive edge weight. An edge with a negative edge-weight is also non-zero and therefore also a "positive".
Sensitivity, also termed the true positive rate, gives the ratio of the number of true edges that were detected in the estimation versus the total number of edges in the true model:

$$\text{sensitivity} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ of false negatives}}$$

Specificity, also termed the true negative rate, gives the ratio of true missing edges detected in the estimation versus the total number of absent edges in the true model:

$$\text{specificity} = \frac{\#\text{ true negatives}}{\#\text{ true negatives} + \#\text{ false positives}}$$

Exercise 11 (1 bonus point)
Repeat the above simulation procedure and simulate 100 datasets with sample sizes of 50, 250, 1,000, and 2,500 (you should simulate 400 datasets in total). For each network, compute the correlation between edge weights, the sensitivity and the specificity. Report your findings in a table. What do you conclude about the performance of LASSO regularization?

■

References

Doosje, B., Loseman, A., & Bos, K. (2013). Determinants of radicalization of islamic youth in the netherlands: Personal uncertainty, perceived injustice, and perceived group threat. *Journal of Social Issues*, *69*(3), 586–604.

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, O. J., … Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of abnormal psychology*, *124*(2), 256–265.