Workshop Network Modeling for Psychological (and Attitudinal) Data

Exercises Day 2-2: Advanced methods in R

Oldenburg, 12/01/2018

Part 1. Network Comparison Test

We use the same Big Five Inventroy (personality) data as yesterday.

```
install.packages(“psych”)
library(”psych”)
data(“bfi”)
bfiData = bfi[,1:25]
```

Exercise 1. Split the data into two groups. You can decide to just cut the dataset in half, or split it according to gender (column 26), education (column 27), or age (column 28). For now, let’s start by splitting the data according to gender. Make sure that your two datasets do not contain any missing data (use for example `complete.cases`).

Solution:

```
sam1 = bfiData[bfi[,26]==1,]
sam2 = bfiData[bfi[,26]==2,]
sam1 = sam1[complete.cases(sam1),]
sam2 = sam2[complete.cases(sam2),]
```

```
install.packages(“NetworkComparisonTest”)
library(NetworkComparisonTest)
```

Exercise 2. Use the NetworkComparisonTest (`NCT`) to test for global strength invariance, network structure invariance, and edge strength invariance. Are there any differences?

Solution:

```
compare1 = NCT(sam1, sam2, it = 100, test.edges = TRUE, edges =
"all")
compare1$glstrinv.pval #p-value = 0.56; no difference in global strength
compare1$glstrinv.sep #global strength values for the two samples
compare1$nwinv.pval #p-value = 0.7; no difference in network structure
compare1$einv.pvals #p-values for each edge
which(compare1$einv.pvals<.05) #the edge numbers that are significantly different
```

N.b.: these numbers may slightly differ for you. If you would like your results to stay the same you can use `set.seed(1)` before running your code.

Exercise 3. The two samples are of different size: sam1 has *n = 805* while sam2 has *n = 1631*. Yesterday we've learned that networks based on smaller samples are sparser. Why is this not a problem when performing a NCT?

Solution: When we perform a NCT, we keep the original sample sizes. The effect that any difference in sample sizes might have will therefore not only be present in the difference between the two original networks, but also in all differences between the permutation networks.

Part 2. Mixed Graphical Models

In the following, we use the R package mgm to estimate a Mixed Graphical Model on a data set consisting of questionnaire responses of individuals diagnosed with Autism Spectrum Disorder. This dataset includes variables of different domains, such as age (continuous), type of housing (categorical) and number of treatments (count).

The dataset consists of responses of 3521 individuals diagnosed with Autism Spectrum Disorder (ASD) to a questionnaire including 28 variables of domains continuous, count and categorical and is automatically loaded with the mgm package.

```
> dim(autism_data_large$data)
[1] 3521   28

> autism_data_large$data[1:4, 1:5]
  Gender IQ Age diagnosis Openness about Diagnosis Success selfrating
1      1  6    -0.9605781                        1               2.21
2      2  6    -0.5156103                        1               6.11
3      1  5    -0.7063108                        2               5.62
4      1  6    -0.4520435                        1               8.00
```

We use our knowledge about the variables to specify the domain (type) of each variable and the number of levels for categorical variables (for non-categorical variables we choose 1 by convention). "c", "g", "p" stands for categorical, Gaussian and Poisson (count), respectively. In this case, these are already encoded in the data within the package, so there is no need to create new object. We can extract them as follows:

```
> autism_data_large$type
 [1] "c" "g" "g" "c" "g" "c" "c" "p" "p" "p" "p" "p" "p"
[14] "c" "p" "c" "g" "p" "p" "p" "p" "g" "g" "g" "g" "g"
[27] "c" "g"

> autism_data_large$level
 [1] 2 1 1 2 1 5 3 1 1 1 1 1 1 2 1 4 1 1 1 1 1 1 1 1 1 1 3
[28] 1
```

mgm allows to estimate k-order MGMs (for more details see here). Here we are interested in fitting a pairwise MGM, and we therefore choose k = 2. In order to get a sparse graph, we use L1-penalized regression, which minimizes the negative log likelihood together with the L1 norm of the parameter vector. This penalty is weighted by a parameter λ, which can be selected either using cross validation (lambdaSel = "CV") or an information criterion, such as the Extended Bayesian

Information Criterion (EBIC) (lambdaSel = "EBIC"). Here, we choose to use the EBIC with a hyper parameter of γ=0.25.

Exercise 4. Run this analysis.

Solution:

```
fit_ADS <- mgm(data = as.matrix(autism_data_large$data),
               type = autism_data_large$type,
               level = autism_data_large$level,
               k = 2,
               lambdaSel = 'EBIC',
               lambdaGam = 0.25)
```

Exercise 5. Next, use the qgraph package to visualize the weighted adjacency matrix. What do you see? What is the relationship between age and age of diagnosis? Does this relationship make sense?

Solution:

```
library(qgraph)

qgraph(fit_ADS$pairwise$wadj,
       layout = 'spring', repulsion = 1.3,
       edge.color = fit_ADS$pairwise$edgecolor,
       nodeNames = autism_data_large$colnames,
       color = autism_data_large$groups_color,
       groups = autism_data_large$groups_list,
       legend.mode="style2", legend.cex=.4,
       vsize = 3.5, esize = 15)
```

There is a strong positive relationship between age and age of diagnosis, which makes sense because the two variables are logically connected (one cannot be diagnosed before being born).

While the interaction between continuous variables can be interpreted as a conditional covariance similar to the well-known multivariate Gaussian case, the interpretation of edge-weights involving categorical variables is more intricate as they are comprised of several parameters. In the following section we show how to retrieve necessary parameters from the fit_ADS object (i.e., your *mgm* object if you named it differently) in order to interpret interactions between continuous and categorical variables.

We first consider the edge weight between the continuous Gaussian variable ‘Working hours’ and the categorical variable ‘Type of Work’, which has the categories (1) No work, (2) Supervised work, (3) Unpaid work and (4) Paid work.

In order to get the necessary parameter, we look up in which row this pairwise interaction is listed in `fit_ADS$rawfactor$indicator[[1]]`. We look at the first list entry here, because we are looking for a pairwise interaction. If we estimated an MGM involving 3-way interactions, the 3-

way interactions would be listed in the second list entry, etc. Here, however, we look for the pairwise interaction between ‘Type of Work’ (16) and ‘Working hours’ (17), and find it in row 86:

```
fit_ADS$rawfactor$indicator[[1]][86, ]
[1] 16 17
```

Using the row number, we can now look up all estimated parameters in `fit_ADS$rawfactor$weights`:

```
fit_ADS$rawfactor$weights[[1]][[86]]

[[1]]
[1] -14.6460488  -0.7576681   0.7576681   1.4885513

[[2]]
           [,1]
V16.2 0.5150313
V16.3 1.3871043
V16.4 1.7926628
```

Exercise 6. What do these coefficients mean?

Solution: The first entry corresponds to the regression on ‘Type of Work’ (16). Since we model the probability of every level of a categorical variable, we get for each of the four levels of ‘Type of Work’ a parameter for ‘Working hours’. We see that there is a huge negative parameter for the first category of ‘Type of Work’, which is ‘No work’. This makes sense, since in the data all individuals with no work logically also work 0 hours. The differences between the remaining categories are less strong. However, we see that the more hours one works, the more one is likely to be in category (4) ‘Paid work’.

The second entry corresponds to the regression on ‘Working hours’ (17). Now the categorical variable is a predictor variable, which means that the first category is coded as a dummy category which is absorbed in the intercept. Note that we could also model all categories explicitly by using the overparameterized parameterization by setting overparameterize = TRUE in mgm(). Here we see that being in category (3) ‘Unpaid work’ predicts a larger amount of working hours than being in category (2) ‘Supervised work’, and that being in category (4) ‘Paid work’ predicts a larger amount of working hours than being in category (3) ‘Unpaid work’, which makes sense.