Assignment № 3

released: 15.11.2021 at 19:00 **due:** 01.12.2021 at 12:00

Task 1: Simulations from SAOM **10 points**

The file `simSAOMs.R` contains the code to simulate the network evolution between two observations from a SAOM with an evaluation function specified by the outdegree and the reciprocity statistics. It also includes the code to produce violin plots for the cumulative indegree and outdegree distribution.

(1) Implement the missing code so that the function `simulation` can be used to simulate the network evolution. Document the code. The algorithm is described as an appendix of the slides of Practical 3. Unconditional simulation is used.

Hint: a useful function for the implementation is `sample`

(2) Consider the two adjacency matrices in the files `net1.csv` and `net2.csv`. They are observations of two networks collected on a set of 22 actors at time t_1 and t_2, respectively. Estimate the parameters of the SAOM with outdegree and reciprocity statistics using the function `siena07`.

(3) Conditioning on the first observation, generate 1,000 simulations of the network evolution using the function `simulation` developed in (1). Compute the indegree and outdegree distribution for each simulated network. Save the results in two `R` objects[1], named *indegDist* and *outdegDist* respectively, in which rows are the simulations and columns are the number of nodes with a given in(out)degree.

Hint: The helper function `degreeDistribution` is included in the code. You can use it to compute the in(out)degree distribution

(4) Use the simulated values of the in(out)degree distributions to evaluate the model's goodness of fit. The second part of the code was written to this aim, complete the missing pieces of code to produce the violin plots. Additionally, write the code to compute the Mahalanobis distance and the p-value used in `RSiena` to assess the fit of the model with respect to the in(out)degree auxiliary statistics. The code should compute the following quantities:

[1] You can use objects of class matrix/arrays or data frame

i. the variance-covariance matrix of the simulated network statistics $\hat{\Sigma}$ and its inverse.
Hint: useful functions are `cov` and `solve`

ii. center the simulated values of the cumulative degree distribution census obtained in (3).
(The centered values are computed as (x-$\hat{\mu}$) with $\hat{\mu}$ the average)

iii. center the observed values of the cumulative degree distribution census in the second observation with $\hat{\mu}$ as in ii.

iv. compute the Mahalanobis distance for each simulated and observed network using the centered values (compute on ii. and iii.).
(The Mahalanobis distance is computed as $(x-\hat{\mu})^T\hat{\Sigma}^{-1}(x-\hat{\mu})$)

v. compute the percentage of simulated networks with Mahalanobis distance equal or greater than the observed network Mahalanobis distance.

Run the complete code to obtain the violin plots and the test on the Mahalanobis distance. Would you think that the model has a good fit based on the cumulative degree distribution and the p-value compute in (4)? Justify your answer.

Hint: a useful function to apply the same function to the rows or columns of a data frame(array) is the function `apply`

(Please do not modify existing code even though more efficient solutions can be implemented)

Task 2: Estimation and interpretation of SAOMs **10 points**

The folder `Glasgow.zip` contains data collected by Michell and West (1996) under the "Teenage Friends and Lifestyle Study" [2]. The dataset was collected on a cohort of 160 students followed over two years starting in February 1995, when the pupils were aged 13, and ending in January 1997. The friendship network of the pupils was observed at three-time points. Pupils were asked to name up to six friends and provide information on their socio-demographic characteristics along with the use of substances, such as tobacco and alcohol consumption. In the following, we analyse the data of the 129 pupils who were present at all three-time points.

The folder contains the following files

- `f1, f2, f3.csv`: adjacency matrices of the friendship networks
- `demographic.csv`: data frame containing information on gender (1 boy, 2 girl) and age
- `logdistance.csv`: logarithm of the distance (in kilometers) between the houses of the pupils
- `alcohol.csv`: alcohol consumption coded as 1 (non), 2 (once or twice a year), 3 (once a month), 4 (once a week) and 5 (more than once a week);

(1) Let us start by considering the friendship network as the only dependent variable

(1.1) Compute the Jaccard index to evaluate if the data contains enough information to investigate the evolution of the friendship network. Comment on the results.

(1.2) Specify a reasonable model to test the following hypotheses on the friendship evolution:

i. Students tend to be friends with popular pupils
ii. Students tend to be friends with pupils with similar alcohol consumption to their own
iii. Students tend to be friends with students that live in the same neighborhood (living nearby)

Do not forget to control for the basic endogenous and exogenous variables.
Hint: take a look at the practicals on SAOM and on the introductory paper on SAOMs for inspiration

[2] Data description and download from `http://www.stats.ox.ac.uk/~snijders/siena/siena.html`

(1.3) Estimate the model, check its convergence and fit, and comment on its parameters.

(1.4) Are the hypotheses i.-iii. supported by the data? Argue for your answer.

(2) We now investigate the co-evolution of friendship (network dependent variable) and alcohol consumption (behavioral dependent variable).

(2.1) Use the model specification developed in (1) for the selection part of the model. Specify a reasonable model for the influence part to test the following hypotheses:

iv. Popular students tend to increase or maintain their level of alcohol consumption

v. Students tend to adjust their alcohol consumption to that of their friends

(2.2) Estimate the model, check its convergence and fit, and comment on its parameters.

(2.3) Are the hypothesis iv.-v. supported by the data? How the conclusions about the test on the hypothesis i.-iii. differ with respect to the findings in (1). Argue for your answers.

(2.4) Given the model estimated in point (2.2.), do we have evidence for selection processes only, influece processes only, or both selection and influence processes? Argue for your answer.

You are encouraged to work in groups of 3 or 4 people.
Please submit your solution in a PDF using moodle. Also, remember to put your names on the R scripts that you submit to moodle. Only one member of the group should submit the solution. Do not forget to report the names of all the group members in the documents you submit.