SEM with networks - background

Network data can be integrated into the SEM framework in different ways. We focus on two main approaches here. The first approach extracts the information from a network based on each participant and then use that information as variable(s) in a SEM model. In this method, each participant (node) in the network is the basic unit for analysis. The second approach extracts information from a network based on each relationship present. In this method, each pair of participants or nodes are used as the basic unit for analysis.

In our software, we propose and implement four types of models.

Network nodes as analysis units

In this method, each participant is treated as the basic unit of analysis. Therefore, the sample size is equal the sample size n. We use two approaches here: (1) we extract information as network statistics from a network, and (2) we extract information through a latent space model.

Use network statistics

We denote a network through a square adjacency matrix $\mathbf{M}=[m_{ij}]$ with each m_{ij} denoting the connection between subject i and subject j. Based on the adjacency matrix, many node-based network statistics can be defined. For example, the statistic *degree* is a centrality measure that simply counts how many subjects a subject connects to in the network. The statistic *betweenness* measures the extent to which a subject lies on the paths between other subjects. Subjects with high betweenness influence how the information flows in the network. Both degree and betweenness quantify the importance of a subject in a network. For example, for our friendship network, if a student has a larger degree, he or she is more popular in the network. From a network, we can derive a vector of network statistics for each subject i as $\mathbf{t}_{i}(\mathbf{M})$.

Because the network statistics are node based, the dimension of the resulting network statistics data will match the non-network data, and they can be combined to be used in SEM as any regular SEM analysis.

Use latent space model

In this approach, each subject assumes a position in a Euclidean space. The distance of two subjects in the latent space is assumed to be related to how likely they are connected in the network. The idea of latent space modeling is similar to that of factor analysis with a latent factor space and factor scores. Let $\mathbf{z}_{i}$ be a vector of latent positions of subject i in the latent space. For subjects i and j, the Euclidean distance between them is:

$$
d_{ij}({\bf z}_{i},{\bf z}_{j})=\sqrt{({\bf z}_{i}-{\bf z}_{j})^{t}({\bf z}_{i}-{\bf z}_{j})}=\sqrt{\sum_{d=1}^{D}({z}_{i,d}-{z}_{j,d})^{2}}
\label{eq:distance}
$$

where $(\cdot)^{t}$ is the transpose of a matrix or vector, D is the dimension of the Euclidean latent space, $\mathbf{z}_{i}=(z_{i,1},z_{i,2},\cdots,z_{i,D})^{t}$ and $\mathbf{z}_{j}=(z_{j,1},z_{j,2},\cdots,z_{j,D})^{t}$ are the latent positions of subjects i and j, respectively. With the distance, the latent space model can be written as

$$
\begin{cases}
m_{ij} & \sim\text{Bernoulli}(p_{ij})\\
\text{logit}[p(m_{ij})] & =\alpha+\boldsymbol{\beta}'{\bf h}_{ij}-\kappa\times d_{ij}({\bf z}_{i},{\bf z}_{j})
\end{cases}\label{eq:LSM}
$$

where α is an intercept, ${\bf h}_{ij}$ is a vector of covariates and $\boldsymbol{\beta}$ contains the coefficients of the covariates. Note that the network is assumed to be unweighted here. In our software, following the tradition in network analysis, the coefficient κ for d_{ij} is fixed as 1 because κ can be rescaled together with the distance (Hoff et al., 2002). Therefore, the closer of two subjects are in the latent space, the higher the probability is for them to be connected after controlling the covariates in the model.

Here, we adapt and extend the latent space model to have the form shown below:

$$
\begin{cases}
E(m_{ij}) & =\mu_{ij}\\
g(\mu_{ij}) & =\alpha-d_{ij}({\bf z}_{i},{\bf z}_{j})
\end{cases}\label{eq:SEM-LSM}
$$

where g is a link function. First, we assume the connection between two subjects is solely explained by the latent space. Second, we relax the requirement of the Bernoulli distribution to use any exponential family of distributions. Using this model, we can extract information from a network. The idea is similar to principal component analysis. In our model, the latent positions will be used along with non-network variables in the SEM framework.

Network edges as analysis units

Another approach we take is to use edges as the unit of interest. In this case, non-network data are reformatted for analysis to be based on pairs of individuals. In this case, given a non-network covariate c, we define $c_{ij} = f(c_i, c_j)$, where c_i and c_j are the covariate values for individual i and individual j. The function f can be chosen according to the purpose of the analysis. For example, c_{ij} can be the average of c_i and c_j, or it can be the difference. Then, these pairwise non-network variables can be used as either endogenous or exogenous variables.

Use network statistics

Similar as in the node-based framework, in the edge-based framework, network statistics that can be obtained free from assuming underlying models to the social network can be used in SEM. The network statistics are constructed based on each pairs of subjects. For example, the shortest path length between each pair of nodes can be used as the edge-based network statistics.

Use latent space model

The latent space modeling approach can also be used when using a pair of subjects as the unit of analysis. In this case, the latent distance between two subjects $d_{ij}(z_i, z_j)$ can be used in SEM instead of the latent positions z_i and z_j.

Revision #8
Created 22 October 2024 20:05:23 by Admin
Updated 24 October 2024 00:18:52 by Admin