Measure Social Inference Using Latent Space Model

Yi Chen

Teachers College Columbia University

Final Report

Multidimensional Scaling, Clustering, and Network Models (HUDM 5124)

Measure Social Inference Using Latent Space Model

Social network analysis is a ubiquitous area of study, which is widely discussed in many distinct fields (Vivar & Banks 2012). In particular, dynamic network is getting more and more attention since it represents the structure and evolution of the relationships between entities (Sewell & Chen,2015). In this report, we introduce a statistical framework of latent space model (LSM; Hoff, Raftery, & Handcock, 2002), which embeds longitudinal network data as trajectories in a (usually low dimensional) latent Euclidean space. For each discrete time interval, we incorporate the latent distance to accommodate network dependence.

Social inference (sometimes referred to as spillover, contagion, or diffusion) represents the propensity that individuals become more similar (perform the similar behavior) to some reference group (Manski, 1993), such as one's social contacts. An example of the social influence could be: adolescent's delinquency behavior may be inferenced by the delinquency behavior of friends and also the delinquency behavior of the popular student in his/she network (even they are not friend directly). Meanwhile, adolescent's social-economics background as an exogenous factor may also explain his/her delinquency behaviors.

In dynamic network analysis, it is difficult to measure the independent social influence effects since there may have many other influential processes operate at the same time. In other words, some exogenous individual behavior, common social-environmental factors, or psychological states may co-determine the probabilities of change in network and/or behavior (Steglich et al, 2010). Meanwhile, individuals can be influenced by those with whom they do not directly interact in the network. Consequently, using latent space model for social influence has the following four benefits: (1) identifying the underlying mechanism that generate the observed

pattern in social network, (2) specifying the latent structure, which is able to identify the dependence/inference when there is no direct tie in the observed social network, and (3) the latent structure is also able to represent the unobserved latent co-determining variables.

Model Framework

Social network at a discrete time range among n individual or nodes can be coded by adjacent or tie matrix $A_{N\times N}$. For the undirected network, A_{ij} is the value of the edge from actor i to actor j, which is usually binary value to represent the ties absent or present. We can also assign different value on the edge as weight to describe the connection more accurately. Consequently, $A_{N\times N}$ is a symmetric matrix (like dissimilarity or distance matrix that is used in multidimensional scaling). For the directed network, we can also code the observed network into a matrix using the positive or negative sign to indicate the direction of connection. In this report, we focus on the undirected network.

There are two main assumptions in LSM, which model the connection coded into the adjacent matrix. The first assumption is *conditional independence assumption*, that is the ties in the network are independent given the latent space variable.

$$P(\boldsymbol{A}|\alpha_0, \boldsymbol{Z}) = \prod_{i\neq j} p\left(A_{ij}\middle|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0\right)$$

, where α_0 is the intercept term, $\boldsymbol{Z_i}$ and $\boldsymbol{Z_j}$ are two latent variable that indicate the location of node i and node j in a lower dimension latent space. Thus, LSMs are sometimes called latent variable network models (Rastelli et al., 2016).

The second assumption is *monotonous assumption*, that is closer two nodes in the latent space, more likely the tie between the two nodes will be 1 (present).

$$p\left(A_{ij}\middle|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0\right) = logit\ \left(P\left(A_{ij}=1\right)\right) = \alpha_0 - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$

, where α_0 is the intercept term and represent the baseline probability of a tie for node in the network (overall density of the network), $\left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$ is the distance between node i and j in the latent space, which can be calculate based on Euclidean distance or any other distance measurements.

Under these two assumptions, latent space models have been used to estimate covariate or selection effects, the association between actor attributes and the presence of a tie. After incorporating meaningful covariates, the latent space positions function more as a way to account for unobserved network tie dependencies than as an interpretable parameter (Sweet & Adhikari, 2020). We can further formulate a general causal influential model framework for social influence.

$$Y_i^t = \beta_0 + \beta_1 Y_i^{t-1} + \beta_2 \sum_{g \in G_i} (w_g Y_g^{t-1}) + \in_i$$

$$w_g = \frac{||Z_g - Z_i||^{-1}}{\sum_{g \in G_i} \left|\left|Z_g - Z_i\right|\right|^{-1}}$$

, where Y is the collection of nodal outcomes measured at two different times t and $t-1$, W_g is a weight matrix of neighbor g of the node i (closer the neighbor is, bigger the weight it will have), β_0 is the intercept coefficient, β_1 is the effect of node i outcome at the previous time on the outcome at the current time, and β_2 is the influence of the network on the outcome Y (i.e., the social influence effect). Thus, for any outcome variable of interest, we can model the social influence effect in the network by estimating the lower dimensional latent space location and using the corresponding distance in latent space to calculate the weight of influence from other nodes in the network.

In the Bayesian statistics framework, the latent variables can be estimated using MCMC. The prior and conditional distribution can be assumed as following:

$$A_{ij} \sim Bernoulli(p_{ij})$$

$$logit(p_{ij}) \sim \alpha_0 - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$

$$\boldsymbol{Z_i} \sim_{iid} MVM_d(\mathbf{0}, \tau I)$$

$$Y_i^t \sim N(\beta_0 + \beta_1 Y_i^{t-1} + \beta_2 \boldsymbol{N}(\boldsymbol{Z}) \boldsymbol{Y_{j \neq i}^{t-1}}, \sigma)$$

$$\beta_i \sim N(\mu_0, \sigma_0^2), i = 0,1,2$$

$$\sigma \sim Inv - Gamma(a, b)$$

$$\tau_z \sim Inv - Gamma(c, d)$$

, where $\boldsymbol{N}(\boldsymbol{Z})$ is any format of weighting vector depend on the latent variable $\boldsymbol{Z}$. The joint posterior distribution is

$$P(A, Z, \beta_0, \beta_1, \beta_2, Y^t, Y^{t-1}) = P(A|Z, \tau)P(Y^t|\beta_0, \beta_1, \beta_2, Y^t, \sigma)P(Z|\tau)P(\beta_0, \beta_1, \beta_2|\sigma_0)P(\sigma_0)P(\tau)$$

Empirical Example

In this section, we will show an empirical example of LSM using R. The data set we used comes from the teenage friends and lifestyle study (Michell 2000, Pearson and West 2003), which is available online through https://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm. The data set provide the friendship network and substance use were recorded for a cohort of 50 female pupils in a school in the West of Scotland. The panel data were recorded over a three-year period starting in 1995, when the pupils were aged 13, and ending in 1997. The friendship networks were formed by allowing the pupils to name up to twelve best friends. Pupils were also asked about their attributes on smoking (s), drug use (d), sport (sp), and alcohol use (a). The observed social network is shown in the *Figure 1*.

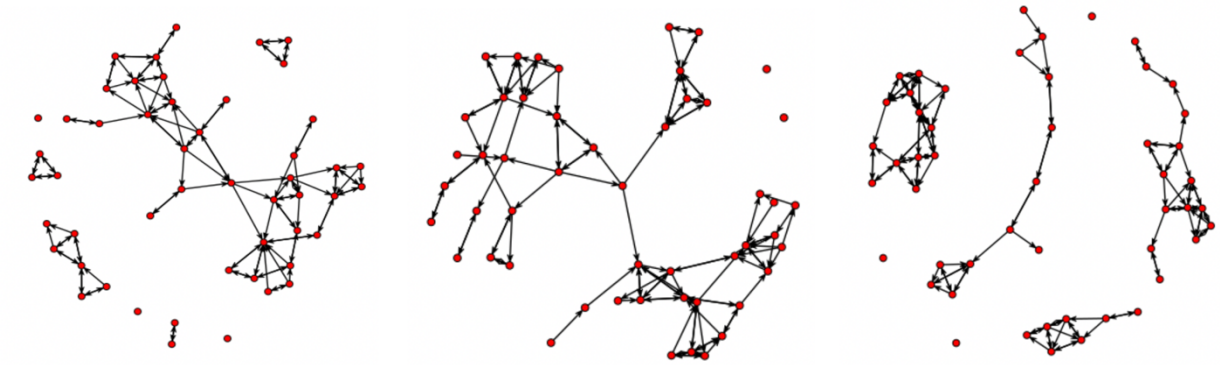

Figure 1. Girls' friendship network from 1995 to 1997

For the first step, we estimate the latent space variables based on the adjacent matrix.

$$p(A_{ij}|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0) = logit\left(P(A_{ij} = 1)\right) = \alpha_0 + |sp| - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$

, where $|sp|$ is the absolute difference between node i and node j. We assume the friendship connection depend on their attitude towards sport. Thus, latent variable $\boldsymbol{Z}$ capture the combined effects from other factors, which may have an effect on the friendship. For simplicity, we take assume the latent space in one dimension.

For the social inference model, the outcome variable we care about is the grills' attitude towards alcohol use (a). We want to explore, to what extent, does girl's attitude towards alcohol use is affected by the social influence of their friends' attitudes towards alcohol use. The model framework is shown as following:

$$Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \beta_2 \sum w_g Y_{gt-1} + \beta_3 d_{it} + \beta_4 s_{it} + e_{it}$$

$$w_g = \frac{||Z_g - Z_i||^{-1}}{\sum_{g \in G} \left|\left|Z_g - Z_i\right|\right|^{-1}}$$

, where the attitude of the individual i about alcohol use at time t is Y_{it}, β_0 is the intercept term, β_1 is the lag effect of previous one-year's attitude, β_2 is the social influence effect, β_3 and β_4 are the coefficient for attitude of on smoking (s) and drug use (d). After estimating the latent

variable, we can calculate the Euclidean distance and weight correspondingly. *Table 1* shows the result of regression. According to the result, the effect of social influence is significant. While the biggest effect comes from the last years' perception. Controlling the social inference and last years' perception, the perception on smoke and drug are not significant any more. The detailed information about R code are provided in the appendix.

Table 1 model result of regression

Coefficients	Estimate	Std. Error	T value	P-value
Intercept (β_0)	0.606	0.321	1.890	0.062$^{\cdot}$
Lag alcohol (β_1)	0.545	0.080	6.779	0.000***
Social influence (β_1)	0.264	0.121	2.181	0.031**
Smoke (β_2)	-0.054	0.115	-0.473	0.637
Drug (β_3)	0.168	0.115	1.452	0.149

Discussion

Latent space model provides an opportunity of measuring the social influence in the observed network, which could apply the conventional causal inference technique. The latent space model itself follows the logic of latent structure model. The framework we proposed in this report can be easily generalized for different research purpose. We hope that the present article will encourage researchers to use latent space mode. The result will be a more in-depth and more informative analysis of social network analysis.

References

Sweet, T. and Adhikari S. (2020). A Latent Space Network Model for Social Influence. *Psychometrika*, https://doi.org/10.1007/s11336-020-09700-x.

Xu, R. (2019) Estimating Social Influence Using Latent Space Adjusted Approach in R. *arXiv preprint*, https://arxiv.org/abs/1903.05999.

Daniel K. Sewell & Yuguo Chen (2015) Latent Space Models for Dynamic Networks, *Journal of the American Statistical Association*, 110:512, 1646-1657, DOI: 10.1080/01621459.2014.988214

Vivar, J. C., and Banks, D. (2012), "Models for Networks: A Cross-Disciplinary Science," *Wiley Interdisciplinary Reviews: Computational Statistics*, 4, 13–27.

Steglich, C., Snijders, T. A., & Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, *40*(1), 329-393.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098

Rastelli, R., Friel, N., & Raftery, A. E. (2016). Properties of latent variable network models. *Network Science*, *4*, 407–432.

Michell, M. P. L. (2000). Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs: education, prevention and policy*, *7*(1), 21-37

Pearson, M., & West, P. (2003). Drifting smoke rings. *Connections*, *25*(2), 59-76.

Appendix

```
library(RSiena)
library(latentnet)
s50s<-read.table("s50-smoke.dat",header=FALSE)
s50d<-read.table("s50-drugs.dat",header=FALSE)
s50sp<-read.table("s50-sport.dat",header=FALSE)
s50a<-read.table("s50-alcohol.dat", header=FALSE)
g1<-network(s501,directed=TRUE)
g1%v%"a" <- s50a[,1]
g1%v%"s" <- s50s[,1]
g1%v%"sp" <- s50sp[,1]
g1%v%"d" <- s50d[,1]
g2<-network(s502,directed=TRUE)
g2%v%"a" <- s50a[,2]
g2%v%"s" <- s50s[,2]
g2%v%"sp" <- s50sp[,2]
g2%v%"d" <- s50d[,2]
g3<-network(s503,directed=TRUE)
g3%v%"a" <- s50a[,3]
g3%v%"s" <- s50s[,3]
g3%v%"sp" <- s50sp[,3]
g3%v%"d" <- s50d[,3]
plot(g1)
plot(g2)
plot(g3)
```

```
m1<-ergmm(g1 ~ euclidean(d =

1)+absdiff("sp"),control=ergmm.control(sample.size=5000,burnin=20000,interval=10,Z.delta=5)

)

m2<-ergmm(g2 ~ euclidean(d =

1)+absdiff("sp"),control=ergmm.control(sample.size=5000,burnin=20000,interval=10,Z.delta=5)

)

latent_pos_1 <- m1$mkl$Z
```

```
latent_pos_2 <- m2$mkl$Z
w <- c()
for (i in 1:50){
  current_position <- latent_pos_2[i]
  neighbor_alcho <- s50a[,2][-i]
  neighbor <- latent_pos_2[-i]
  distances <- c()
  for (j in neighbor){ distances <- c(distances,1/abs(current_position-j))  }
  cur_w <- 0
  for (k in 1:length(distances)){ cur_w <- cur_w + (distances[k] / sum(distances)) *
neighbor_alcho[k] }
  w <- c(w,cur_w)
}
for (i in 1:50){
  current_position <- latent_pos_1[i]
  neighbor_alcho <- s50a[,1][-i]
  neighbor <- latent_pos_1[-i]
  distances <- c()
  for (j in neighbor){distances <- c(distances,1/abs(current_position-j)) }
  cur_w <- 0
  for (k in 1:length(distances)){cur_w <- cur_w + (distances[k] / sum(distances)) *
neighbor_alcho[k]  }
  w <- c(w,cur_w)
```

```
}
alcohol<-c(s50a[,3],s50a[,2])
lag_alc<-c(s50a[,2],s50a[,1])
drug<-c(s50d[,3],s50d[,2])
smoke<-c(s50s[,3],s50s[,2])
infl<-data.frame(cbind(alcohol,lag_alc,w,drug,smoke,rep(c(1:50),2),rep(c(1:2),each=50)))
summary(lm(alcohol~lag_alc+w+smoke+drug,data=infl))
```