Statistical analysis of network data lecture 9

Sofia Olhede

November 13, 2024

1. Latent Space Ct'd
2. β–model
3. Variants of SBMs
4. Network Sampling
5. Barabasi-Albert
6. Statistical Temporal Models

Latent Space VI

- Hoff introduced the **latent space** models in 2002. Young and Scheinerman in 2007 introduced the **random dot product graphs (RDPG)**.
- Let F be a d-dimensional inner product distribution with $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, collected in the rows of the matrix

$$\mathbf{X} = [X_1, X_2, \ldots, X_n]^T \in \mathbb{R}^{n \times d}.$$

  In this model $X_j \in \mathbb{R}^d$.
- Conditional on $\mathbf{X}$ we have that

$$\Pr\{\mathbf{A} \mid \mathbf{X}\} = \prod_{i<j} \left(\rho_n X_i^T X_j\right)^{a_{ij}} \left(1 - \rho_n X_i^T X_j\right)^{1-a_{ij}},$$

  if we assume $0 \leq \rho_n X_i^T X_j \leq 1$. ρ_n is not identifiable, but there to scale the whole set of probabilities in one go.

Latent Space VII

- Given a graph distributed as an RDPG, you may seek to estimate X_i.
- Note that if $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a unitary matrix then

$$\begin{aligned}(\mathbf{XW})(\mathbf{XW})^T &= \mathbf{XWW}^T\mathbf{X}^T \\ &= \mathbf{XX}^T\end{aligned}$$

- As this is true, latent positions $\mathbf{X}$ and $\mathbf{XW}$ give rise to the same distribution over graphs in Equation (4).
- We note that the (positive semi-definite) stochastic blockmodel can be reformulated as a RDPG.
- We say an RDPG with latent positions $\mathbf{X}$ is an SBM with K blocks if the number of distinct rows in $\mathbf{X}$ is K, denoted $\mathbf{X}(1), \ldots, \mathbf{X}(K)$. In this case, we define the block membership function $\tau : [n] \mapsto [K]$ to be a function such that $\tau(i) = \tau(j)$ if and only if $X_i = X_j$.

Latent Space VIII

- In addition, we also consider the case of a stochastic block model in which the block memberships of each vertex is randomly assigned. More precisely, let $\pi \in (0,1)^K$ with $\sum_k \pi_k = 1$ and suppose that $\tau(1), \ldots, \tau(n)$ are now i.i.d. random variables multinomial π that is, $\Pr(\tau(i) = k) = \pi_k$ for all k.
- We begin with describing the notations for the spectral decomposition of the rank d positive semidefinite matrix $\mathbf{P} = \mathbf{X}\mathbf{X}^T$.
- As $\mathbf{P}$ is symmetric and positive semidefinite it has a spectral decomposition of $\mathbf{P} = \mathbf{U}_P \mathbf{S}_P \mathbf{U}_P^T$, with $\mathbf{U}_P$ having orthogonal columns, and $\mathbf{S}_P$ a diagonal matrix with nonincreasing entries.

Latent Space IX

- One can take the adjacency spectral embedding (ASE) of $\mathbf{A}$ into $\mathbb{R}^d$, by first calculating $|\mathbf{A}| = \left(\mathbf{A}^T\mathbf{A}\right)^{1/2}$, and then determine its spectral embedding of $\mathbf{U}_A$ and $\mathbf{S}_A$.
- Our goal is to estimate the latent position matrix $\mathbf{X}$. Now, if the matrix $\mathbf{P}$ were actually observable, then the spectral embedding of $\mathbf{P}$, given by $\mathbf{U}_P\mathbf{S}_P^{1/2}$, is simply some orthogonal transformation $\mathbf{P}$ of $\mathbf{X}$.
- As long as we only observe $\mathbf{A}$ and not $\mathbf{P}$, we have to use that $\mathbf{A}$ can be viewed as a small perturbation of $\mathbf{P} = \mathbf{X}\mathbf{X}^T$. More precisely

$$\left\|\mathbf{A} - \mathbf{X}\mathbf{X}^T\right\| = o\left(\left\|\mathbf{X}\mathbf{X}^T\right\|\right).$$

- By the Davis-Kahan theorem (Davis and Kahan, 1970), the subspace spanned by the top d eigenvectors of $\mathbf{X}\mathbf{X}^T$ is well-approximated by the subspace spanned by the top d eigenvectors of $\mathbf{A}$. d is assumed known.

Latent Space X

- Non-parametric summaries calculable from $\mathbf{X}\mathbf{X}^T$ can therefore be calculated from $\mathbf{A}$.

β–model

- We have already looked at the Chung–Lu or configuration model which sets
$$\mathbb{E}\, A_{ij} = p_{ij} = \min(\pi_i \pi_j, 1),$$
but even if this has a closed form estimate from d it does not naturally stay in the permitted range.
- The natural way out of this problem is to find a simpler way to parameterise p_{ij}. Instead of defining n parmeters $\pi_i \in [0, 1]$ we define $\beta_i \in \mathbb{R}$ and set
$$p_{ij} = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}.$$
In terms of of the log-odds ratio we then have
$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \beta_j, \quad 1 \leq i < j \leq n.$$
- The magnitude and sign of β_i quantifies the propensity of node i to have ties.

β–model

- The degree of node i is expected to be large (small) if β_i is positive (negative).
- The maximum likelihood estimate (MLE) $\widehat{\beta}_i$ satisfies

$$d_i = \sum_{j \neq i} \frac{e^{\widehat{\beta}_i + \widehat{\beta}_j}}{1 + e^{\widehat{\beta}_i + \widehat{\beta}_j}}, \quad i = 1, \ldots, n.$$

- Questions may arise about the existence, uniqueness and accuracy of the MLE.
- Let $d_1, \ldots, d_n$ be the degree sequence of G generated from the β–model. Let $L = \max |\beta_i|$. Then there is a constant $C(L)$ such that with probability at least $1 - C(L)n^{-2}$ there exists a unique solution of the maximum likelihood equations , that satisfies

$$\max_{1 \leq i \leq n} |\widehat{\beta}_i - \beta_i| \leq C(L) \sqrt{\frac{\log(n)}{n}}.$$

Variants of Block models

- We have observed the general utility of the stochastic block model. However many human interactions are more complex than a stochastic block model would give.
- In Lecture 1, we were introduced to the mixed membership stochastic blockmodel. Generate latent variable $\boldsymbol{\xi}_i$ for each node i from the Dirichlet distribution of dimension k with parameters $\boldsymbol{\alpha}$. Define $\boldsymbol{\Theta} = (\theta_{pq})$ and draw

$$A_{ij} \,|\, \boldsymbol{\xi}_i \, \boldsymbol{\xi}_j, \sim \mathrm{Ber}(\boldsymbol{\xi}_i^T \boldsymbol{\Theta} \boldsymbol{\xi}_j). \tag{1}$$

- This allows individuals to not behave only like one group.
- There are many other generalizations of the stochastic block model, namely the degree-corrected stochastic block model, the overlapping stochastic block model, the hierarchical stochastic block model and the geometric block model (combining GRG and block model).

Variants of Block models

- Let the expectation of the adjacency matrix be written as P.
- Assume that P admits the representation of

$$P = \rho_n \mathbf{Q} \mathbf{\Xi}^T \mathbf{\Lambda} \mathbf{\Xi} \mathbf{Q}, \tag{2}$$

  where $\mathbf{\Xi}$ is a $k \times n$ matrix where each row denotes a position in a latent space, $\mathbf{\Lambda}$ denotes a diagonal matrix, and $\mathbf{Q}$ is an $n \times n$ matrix.
- ρ_n and all entries of the diagonal matrix $\mathbf{\Lambda}$ are non-negative. In this representation, $\boldsymbol{\xi}_i$ denotes the community membership of node i.
- If this is the case then we say that the network is generated according to the overlapping stochastic blockmodel.

Network sampling

- What ways can we sample a network?
- **Relational or edge sampling.** Relational sampling corresponds to sampling exactly relations or edges. This could be sampling phone calls.
- In many networks applications the relations are the **primitive** objects and the vertices are **derivative** from these.
- **Hyperedge sampling.** Sampling academic articles from a research repository involves more than actor in every relationship. Then every article represents an hyperedge.
- **Path sampling.** In the early days of network science it was thought that one could ascertain network topology by analyzing the paths traversed when sending information from one part of the Internet to another. **Traceroute** is an example of such a method.
- The high level realization is that **we rarely see the entire network**.

Network sampling

- Snowball sampling. Here we see a network by initialsing the sampling mechanism at $N(s,0) = \{s\}$ by sampling the node s (could be at random).
- We then sample the set of students $N(s,1) = \{s' : A_{ss'} = 1\}$.
- Then we sample

$$N(s,2) = \cup_{s' \in N(s,1)} \{s'' : A_{s's''} = 1\} \backslash (N(s,0) \cup N(s,1)).$$

- We can keep going, and then take

$$N(s,k) = \cup_{s' \in N(s,k-1)} \{s'' : A_{s's''} = 1\} \backslash (\cup_{j=0}^{k-1} N(s,j)).$$

- Then the snowball sampling is the set of neighbourhoods $\{N(s,k)\}_{0 \leq k \leq r}$.

Network sampling

- It used to be standard to think if you observed one network then you had one observation.
- How can we bring the notion of large samples to this idea?
- Instead we must ask if we have many observational units.
- The sample size is the number of observational units. In theory we could view a network in terms of the number of observational units (edges) $\binom{n}{2}$. Yet this does not reflect reality when the network is sparse-(expected edges) $\rho\binom{n}{2}$ is then more realistic.

Barabasi-Albert

- Finally we shall introduce one more model into the mix, partially because it is not statistical.
- The Barabasi-Albert preferential attachment model was introduced to explain a number of empirical characteristics observed in real networks.
- It was actually introduced by Yule in 1925 not Barabasi and Albert in 1999, which is the usual reference.
- To be clear, I have modified the description to that given in chapter 8 of van der Hofstad. The description of Durrett assumed continuous time.
- A network evolves in this model by adding a vertex at a time, with each new vertex attaching to existing vertices according to their degree. We assume there are t vertices at time t. We fix $m \in \mathbb{N}$.
- We write the degree of vertex $v_i^{(m)}$ as $d_i(t)$.

Barabasi-Albert

- To map the network at time point t to that at time point $t+1$, we add a new vertex to the network. This has m edges.
- We now need to decide how these connect to the existing vertices.
- We take $\delta > -m$ and at each time step a new vertex appears and attaches at random to m existing vertices, with a probability proportional to the degree of that vertex shifted to be offset by δ.
- When $m = 1$ the probability there is an edge between nodes i and $t+1$ is (self-loops admitted)

$$\Pr\{v_i^{(1)} \text{ connects to } v_{t+1}^{(1)}\} = \begin{cases} \frac{d_i(t)+\delta}{t(2+\delta)+1+\delta} & \text{for} \quad i = t+1 \\ \frac{1+\delta}{t(2+\delta)+1+\delta} & \text{for} \quad i \in \{1,\ldots,t\} \end{cases}.$$

- For $m > 1$ the network is constructed from a PA with $m = 1$ and $\delta' = \delta/m$.
- You start with some existing graph $\mathbf{G}_0$ and keep going, getting $\mathbf{G}_1$, $\mathbf{G}_2$ etc. Assume $\mathbf{G}_0$ has k_0 edges.
- Normally networks are sparse and having a few very large degrees.

Barabasi Albert

- We define the collection of edge variables at time step n to be $a_{ij}^{(n)}$.
- A network is sparse if

$$\lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i<j} a_{ij}^{(n)} = \lim_{n\to\infty} \widehat{\rho} = 0.$$

- By the generating dynamics of the BA model there are m new edges at each step so $k_n = mn + k_0$.
- Thus

$$\widehat{\rho} = \frac{2(mn + k_0)}{n(n-1)} \to 0.$$

- The degree distribution counts the relative proportion of vertices of any integer degree

$$p_{A^{(n)}}(k) = \sum_{i=1}^{n} \mathsf{I}(d_i^{(n)} = k),$$

and $p_{A^{(n)}}(k) \sim k^{-\gamma}$. This corresponds to a power law distribution.

Barabasi Albert

- We define the collection of edge variables at time step t to be $a_{ij}^{(t)}$.
- A network is sparse if

$$\lim_{t\to\infty} \frac{2}{t(t-1)} \sum_{i<j} a_{ij}^{(t)} = \lim_{n\to\infty} \hat{\rho} = 0.$$

- By the generating dynamics of the BA model there are m new edges at each step so $k_n = mn + k_0$.
- Thus

$$\widehat{\rho} = \frac{2(mn + k_0)}{n(n-1)} \to 0.$$

Barabasi Albert

- The degree distribution counts the relative proportion of vertices of any integer degree

$$p_{A^{(n)}}(k) = \sum_{i=1}^{n} \mathrm{I}\left(d_i^{(n)} = k\right),$$

and $p_{A^{(n)}}(k) \sim k^{-\gamma}$. This corresponds to a power law distribution.

- The generating mechanism of the preferential attachment model is very mechanistic.
- What if we want to generate a set of edges (or contacts) over time?
- We could simply start from the graph limit model

$$\mathbb{E}\{A_{ij} \mid \xi\} = \rho_n f(\xi_i, \xi_j).$$

- We shall assume we observe multiple graphs across time and so for a given edge ij we study edge-variable $A_{ij}(t)$.
- The simplest model takes the form:

$$\Pr\{A_{ij}(t) = 1 \,|\, A_{ij}(t-1), \dots\} = h(t,\, A_{ij}(t-1),\, A_{ij}(t-2),\, A_{ij}(t-3), \dots),$$

for some appropriately chosen function $h()$.

- We shall now look to how the edges can be modelled across times (see Süveges and Olhede (2023)) with label vector z:

$$A_{ij}(t) \,|\, A_{ij}(t-1), \ldots, A_{ij}(t-K) \sim \text{Bern}\left(\text{logit}^{-1}\left(\sum_{k=1}^{K} b_{z_i z_j k} A_{ij}(t-k) + c_{z_i z_j}(t)\right)\right). \quad (3)$$

- This model defines a correlated process across time, where the correlation is specified by b. Uses the trick of using a modelling framework that naturally limits the success probabilities between zero and one.
- Introduces flexible forms of serial correlation.
- The generating mechanism can still be estimated, and the properties of the network determined.
- Non-stationary models can we included by replacing $c_{ij,g}$ by a time-varying alternatives as we have above.
- Parameters naturally stay within the range of allowed values.

- The simplest version of this model takes $K = 1$.
- This introduces series correlation of length one.
- It is different from just letting the parameters of the stochastic block model change over time.
- Let us discuss how we would expect the realized edges to change?