Day-3 Practical Session, 27 May 2021

Part 2: T-step Targeted Random Walk Sampling (TTRWS)

Li-Chun Zhang[1,2,3] and *Melike Oguz-Alper*[2]

[1]*University of Southampton (L.Zhang@soton.ac.uk)*, [2]*Statistics Norway*, [3]*University of Oslo*

In this illustration, we will work with the basis of inference for the 1st and the 3rd order graph parameters under T-step targeted random walk sampling (TTRWS).

Description of the population and sampling strategies

- Population graph: $G = (U, A)$; $|U| = N$
- Let $X_t = i$ be the node or *state* at step t form a Markov chain
- The observation procedure (OP) of TRWS applied to the seed sample $s = \{X_0, X_1, \dots X_T\}$
- A node can be observed more than once in the seed sample in contrast to the T-wave SBS
- Under T-step TRWS, $\{i\} \times \alpha_i$ and $\alpha_i \times \{i\}$ observed at each $X_t = i$, such that the referense set given by $s_{ref} = s \times U \cup U \times s$
- *Generating states* of T-step walk with seed sample s: $\mathcal{C}_s = \{M : M \subseteq s\}$
- *Actual sampling sequence of states* (AS3) of motif κ: $s_\kappa = (X_t, \dots, X_{t+q})$
- *Equivalent sampling sequence of states* (ES3) of s_κ: $R_\kappa = \{\tilde{s}_\kappa : \tilde{s}_\kappa \sim s_\kappa\}$, contains any possible sequence of states with $|\tilde{s}_\kappa| = |s_\kappa|$, such that the motif κ would be observed given $(X_t, X_{t+1}, \dots, X_{t+q}) = \tilde{s}_\kappa$

Formula sheet

- The 1st order parameter of interest

  $\mu = \theta/N = \sum_{i \in U} y_i / N$

- Let $r_i = r/(d_i + r)$ be the probability of taking a *random jump*. The transition probability from $X_t = i$ to $X_{t+1} = j$ under TRWS

  $p_{ij} = \frac{1}{d_i+r}\left(1 + \frac{r}{N}\right)$ if $a_{ij} = 1$, and $p_{ij} = \frac{r}{d_i+r}\frac{1}{N}$ if $a_{ij} = 0$ or $i = j$

- *Stationary probabilities* in undirected graphs under TRWS

  $\pi_i \propto d_i + r$

- Let $\mathfrak{s}_n = \{X_{t_1}, X_{t_2}, \dots, X_{t_n}\}$ be the n states after the walk becomes *stationary draw-by-draw at equilibrium*, such that π same for each draw. The *generalised ratio estimator* for the 1st order parameter given by

  $\hat{\mu} = \left(\frac{1}{n}\sum_{i \in \mathfrak{s}_n} \frac{y_i}{c_i}\right)\left(\frac{1}{n}\sum_{i \in \mathfrak{s}_n} \frac{1}{c_i}\right)^{-1}$, where c_i known or unknown values, such that the stationary probabilities $\pi_i \propto c_i$. Under *uniform walk*, $c_i \equiv 1$ and $\pi_i \equiv 1/N$; under *random walk* in undirected graphs, $c_i = d_i$, where c_i unknown for unvisited nodes; under *targeted random walk* in undirected graphs, $c_i = d_i + r$.

- The naive variance estimator of $\hat{\mu}$ under T-step TRWS given by

  $\hat{\mathrm{V}}_{naive}(\hat{\mu}) = \frac{\sum_{t=0}^{T}(y_{X_t} - \bar{y})^2}{T+1}$, where $\bar{y} = \sum_{t=0}^{T} y_{X_t}/(T+1)$

- The analytical variance of $\hat{\mu}$ under SRS with replacement, an example of *independent and identically distributed* (IID) random sample, given by

  $\mathrm{V}_{IID}(\hat{\mu}) = \mu(1-\mu)/(T+1)$

- The 3rd order graph parameter

  $\mu = \frac{\theta}{\theta'}$, where $\theta = \sum_{\kappa \in \Omega} y_\kappa$ is the total number of triangles among cases with all the three nodes being cases, i.e., $y_i = 1$ for all $i \in M$, and $\theta' = \sum_{\kappa \in \Omega'} y'_\kappa$ is that of the other triangles with at least one noncase node.

- An estimator of θ using IWE given by

  $\hat{\theta}(X_t, \dots, X_{t+q}) = \sum_{\kappa \in \Omega} \sum_M \frac{\delta_M}{\pi_M} I_\kappa(M) w_{M\kappa} y_\kappa$, where $w_{M\kappa}$ is the incidence weight, such that $\sum_{M \in R_\kappa} w_{M\kappa} = 1$, $I_\kappa(M) = 1$ if $M \in R_\kappa$, and 0 otherwise, and $\delta_M = 1$ if M is realised and 0 otherwise

  - The *multiplicity* weight and the *proportional-to-probability weight* (PPW) given, respectively, by

    $w_{M\kappa} \equiv \frac{1}{|R_\kappa|}, \quad w_{M\kappa} = \frac{\pi_M}{\pi_\kappa}$, with $\pi_\kappa = \sum_{M \in R_\kappa} \pi_M$, where π_M is the *stationary successive sampling probability* (S3P) defined by

    $\pi_M = \Pr(X_{t_1}, X_{t_2}, \dots, X_{t_q}) = \pi_{X_{t_1}} \prod_{i=1}^{q-1} p(X_{t_i} X_{t_{i+1}})$, where $p(X_{t_i} X_{t_{i+1}})$ is the transition probability from X_{t_i} to $X_{t_{i+1}}$

- Let $\mathbb{I}_t = 1$ if $M = \{X_t, \dots, X_{t+q}\}$ yields $\sum_{\kappa \in \Omega} I_\kappa(M) > 0$, and 0 otherwise. Provided TRWS stationary sequence-by-sequence, an estimator of θ combining all $\hat{\theta}_t$ given by

  $\hat{\theta} = \left(\sum_{t=1}^{n-q} \mathbb{I}_t \hat{\theta}_t\right) / \left(\sum_{t=1}^{n-q} \mathbb{I}_t\right)$

- A generalised ratio estimator of the population ratio $\mu = \frac{\theta}{\theta'}$ given by

  $\hat{\mu} = \frac{\hat{\theta}}{\hat{\theta}'}$, which is invariant towards the unknown proportionality constant in S3P

- Given $(X_t, X_{t+1}) = (i, j)$ between two adjacent nodes under TRWS, all the triangles including i and j as two of the nodes observed. Given $(X_t; X_{t+1}) = (i, j)$ as the AS3, the ES3s of any sampled triangle are the six possible adjacent moves along the triangle. Under TRWS, the S3P given by

  $\pi_i p_{ij} = 1 + \frac{r}{N}$, which leads to that both the multiplicity and the PPW reduce to $w_{M\kappa} \equiv 1/6$

- The expectation $E(\hat{\mu})$ is estimated by $\bar{\hat{\mu}} = \sum_{b=1}^{B} \hat{\mu}_b / B$, where $\hat{\mu}_b$ is the estimate for the bth replication and B is the total number of replications

- The variance of $\hat{\mu}$ is estimated, over all replications, by

  $\mathrm{V}(\hat{\mu}) = \frac{\sum_{b=1}^{B}(\hat{\mu}_b - \bar{\hat{\mu}})^2}{B-1}$

- The simulation error of $\bar{\hat{\mu}}$ is estimated by $\sqrt{\mathrm{V}(\bar{\hat{\mu}})/B}$

NB. R-package **igraph** has to be installed

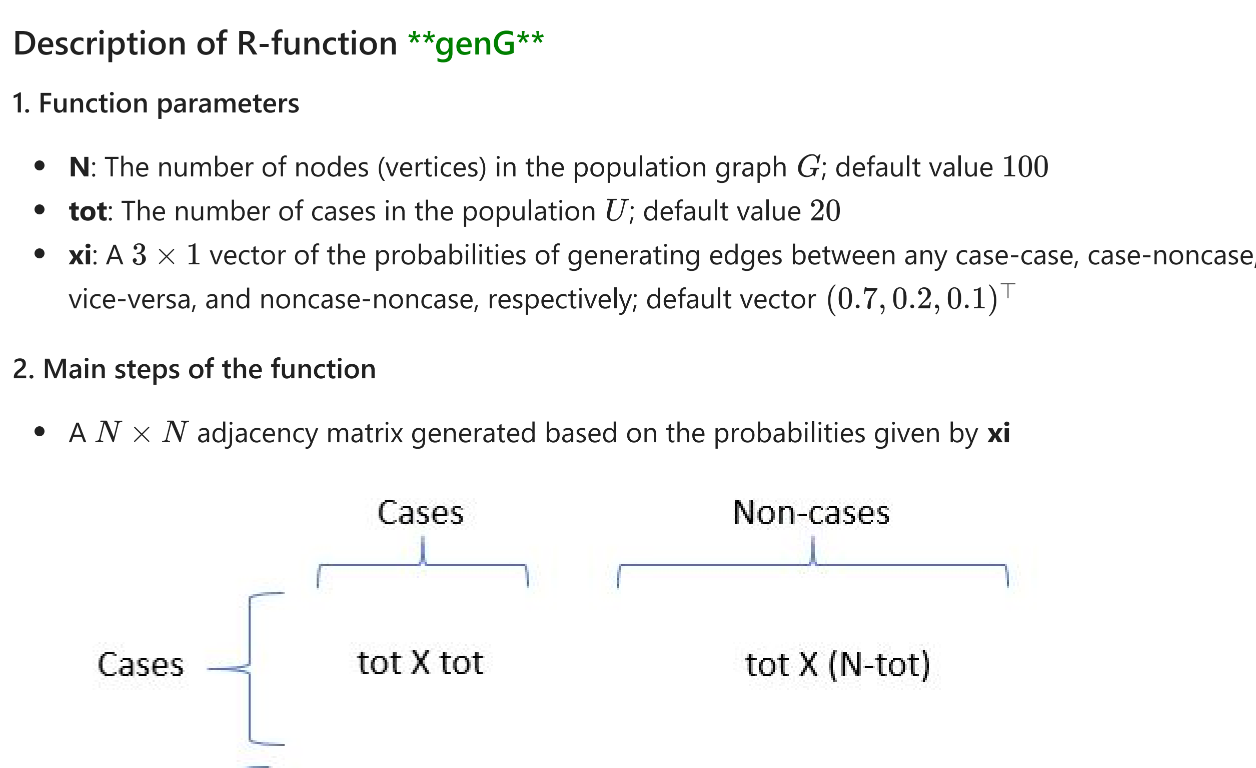
Description of R-function **genG**

1. Function parameters

- **N**: The number of nodes (vertices) in the population graph G; default value 100
- **tot**: The number of cases in the population U; default value 20
- **xi**: A 3×1 vector of the probabilities of generating edges between any case-case, case-noncase, or vice-versa, and noncase-noncase, respectively; default vector $(0.7, 0.2, 0.1)^\top$

2. Main steps of the function

- A $N \times N$ adjacency matrix generated based on the probabilities given by **xi**



3. Main outputs of the function

- The adjacency matrix returned

Description of R-function **triangle**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**

2. Main steps of the function

- Counts the triangles in the population graph based on the adjacency matrix

3. Main outputs of the function

- The number of triangles in the population graph returned

Description of R-function **trw**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**
- **y**: An $N \times 1$ vector of y-values which are equal to 1 for cases, and 0 for non-cases
- **init**: How to choose the initial state: if 0, $\Pr(X_0 = i) = \pi_i$; if < 0, $\Pr(X_0 = i) \equiv 1/N$; if > 0, $\Pr(X_0 = i) = 1$; default value 0
- **K**: The number of steps after the initial state; default value 100
- **r**: The probabilities of random jumps reduced given small value; default value 0.1

2. Main steps of the function

- After specification of the initial state X_0 based on **init**, sample nodes are selected for each step with the transition probabilities given above

3. Main outputs of the function

- A list of two elements returned: the first element is a vector states $X_0, X_1, \dots, X_K$ and the second element is an $N \times 1$ vector of *TRUE/FALSE* indicating which nodes are visited in the targeted random walk

Description of R-function **cnv.y**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**
- **y**: An $N \times 1$ vector of y-values which are equal to 1 for cases, and 0 for non-cases
- **init**: How to choose the initial state: if 0, $\Pr(X_0 = i) = \pi_i$; if < 0, $\Pr(X_0 = i) \equiv 1/N$; if > 0, $\Pr(X_0 = i) = 1$; default value 0
- **K**: The number of steps after the initial state; default value 2
- **r**: The probabilities of random jumps reduced given small value; default value 1
- **B**: The number of replications; default value 100

2. Main steps of the function

- The expectation $E(Y_\infty) = \sum_{i \in U} \pi_i y_i$, where $y_i = 1$ if case, and $y_i = 0$ otherwise, at equilibrium calculated
- The TRW sampling applied independently for each replication
- The expectation $E(Y_t) = \sum_{i \in U} p_{t,i} y_i$ estimated by calculating the mean of the y values of the sample states at step $t = K + 1$ over all replications and the square-root of the Monte-Carlo variance (MC-SD) is calculated

3. Main outputs of the function

- The expectation at equilibrium
- Estimate for the expectation $E(Y_t)$ and its MC-SD

Description of R-function **est.y**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**
- **y**: An $N \times 1$ vector of y-values which are equal to 1 for cases, and 0 for non-cases
- **init**: How to choose the initial state: if 0, $\Pr(X_0 = i) = \pi_i$; if < 0, $\Pr(X_0 = i) \equiv 1/N$; if > 0, $\Pr(X_0 = i) = 1$; default value 0
- **K**: The number of steps after the initial state; default value 100
- **r**: The probabilities of random jumps reduced given small value; default value 0.1
- **B**: The number of replications; default value 100

2. Main steps of the function

- For each replication, the TRW sampling applied for chosen parameters
- The case prevalence is estimated with the generalised ratio estimator for each Markov-Chain
- The SD of the y-values corresponding to the sample states calculated for each Markov-Chain

3. Main outputs of the function

- The expectation of the mean over B replications and its SD returned
- The square-root of the variance under an IID sample (SD-IID) (e.g. SRS with replacement) returned

Description of R-function **muFun**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**
- **y**: An $N \times 1$ vector of y-values which are equal to 1 for cases, and 0 for non-cases

2. Main steps of the function

- The population graph is generated from the adjacency matrix as a graph object. The R-package **igraph** used here
- All the triangles are idenfified in the population graph
- The total number of case-triangles and the total number of triangles with at least one non-case node calculated

3. Main outputs of the function

- The population ratio between the total number of case-triangles and the total number of triangles with at least one non-case node returned

Description of R-function **est.tri**

1. Function parameters

- **amat**: The adjacency matrix for the population graph. Use the output of the function **genG**
- **y**: An $N \times 1$ vector of y-values which are equal to 1 for cases, and 0 for non-cases
- **init**: How to choose the initial state: if 0, $\Pr(X_0 = i) = \pi_i$; if < 0, $\Pr(X_0 = i) \equiv 1/N$; if > 0, $\Pr(X_0 = i) = 1$; default value 0
- **K**: The number of steps after the initial state; default value 100
- **r**: The probabilities of random jumps reduced given small value; default value 0.1
- **B**: The number of replications; default value 100

2. Main steps of the function

- An $N \times N$ matrix of transition probabilities constructed, where the transition probabilities given in the formula sheet above
- For each replication, the TRW sampling applied for chosen parameters
- For each walk, the S3P calculated for any given $(X_t, X_{t+1}) = (i, j)$, where $t = 1, \dots, T-1$
- For each step t of a TRW, it is determined if any triangle observed
- For each observed triangle, the PPW calculated
- For each step of a TRW, the weighted sum of the number of case triangles and the weighted sum of the number of triangles with at least one non-case node calculated. The PPW used.
- For each replicate, the ratio of θ/θ' estimated if the walk is valid, which is defined as a walk where at least one triangle with at least one non-case node observed
- The mean and the variance of the estimates obtained over all B replications calculated

3. Main outputs of the function

- The number of nodes visited and the estimates of the ratio for replicates with valid walks returned
- The mean and the variance of the estimates over all B replications returned
- Simulation error of the mean of the estimates returned
