Data Mining in Big Dynamic Networks

Tutorial 2: inferring dynamic networks

Ernst C. Wit

May 7, 2024

1 Preliminary information

1.1 Theoretical Overview

Inferring Relational Event Models A relational event is a behavioural action from a sender $s \in \mathscr{V}_1$ towards a receiver $r \in \mathscr{V}_2$ at a certain time $t \in [0, \tau]$ (Butts, 2008).

$$e = (s, r, t)$$

Relational events may be seen as time-stamped edges of a dynamic graph. We will focus on the underlying intensity rate of this process, $\lambda_{sr}(t)$. This rate is a function of the baseline hazard function and a number of covariates that are allowed to vary in time.

$$\lambda_{sr}(t) = \lambda_0(t) \cdot \exp[\beta^T \boldsymbol{x}_{sr}(t)]$$

Given relational event data,

$$\{(t_1, s_1, r_1), \ldots, (t_n, s_n, r_n)\}$$

the parameters can be estimated by means of maximizing the (partial) likelihood. The computational bottleneck in the partial likelihood is the sum over the risk set in its denominator. By sampling for each event (t_k, s_k, r_k) one non-event (t_k, s_k^*, r_k^*), the likelihood reduces to the simplified form, i.e.,

$$\begin{aligned} L_{PS}(\beta) &= \prod_{k=1}^{n} \frac{\exp\{\beta^\top x_{s_k r_k}(t_k)\}}{\exp\{\beta^\top x_{s_k r_k}(t_k)\} + \exp\{\beta^\top x_{s_k^* r_k^*}(t_k)\}} \\ &= \prod_{k=1}^{n} \frac{\exp\left\{\beta^\top \left(x_{s_k r_k}(t_k) - x_{s_k^* r_k^*}(t_k)\right)\right\}}{1 + \exp\left\{\beta^\top \left(x_{s_k r_k}(t_k) - x_{s_k^* r_k^*}(t_k)\right)\right\}}, \end{aligned}$$

where the last expression is the likelihood of a logistic regression with

- response: $(1, 1, \ldots, 1)$
- covariates: $\left(x_{s_k r_k}(t_k) - x_{s_k^* r_k^*}(t_k)\right)$ for $k = 1, \ldots, n$.

1.2 Packages and Functions

You will need the library mgcv in R.

2 Inference of Relational Event Models

1. **Recovering simulated parameters.** In the previous tutorial we simulated a dataset of email exchanges sent between colleagues that depended on repetition, reciprocity and weekday/weekend.

   (a) We will use the sampled partial likelihood to infer the underlying parameters. Unfortunately, it is impossible to recover the parameters λ_{0WD} and λ_{0WE} using this method. Why is this?

   (b) However, we can infer the parameters β associated with reciprocity and repetition.

   (c) First, we need to prepare the dataset in such a fashion that it can be analysed as an event history model. Perform the following five steps:

      i. Determine the potential sender V_1 and receiver sets V_2. Are the sets the same (a socalled one-mode network) or different (a socalled two-mode or bipartite network)? Are all possible events $V_1 \times V_2$ always possible or does the risk set change?

      ii. Sample for each event (s_k, r_k) among those events at risk at that time point $\mathscr{R}(t_k)$ a single non-event (s_k^*, r_k^*).

      iii. For your dataset of choice, create for every event and for every non-event the variable `repetition, reciprocity`. Each variable records by means of a 1 or 0 whether or not for the current (s,r) (non)-event,

         - `repetition`: the event (s,r) occurred before less than two time units ago;
         - `reciprocity`: the event (r,s) occurred before less than two time units ago;

      iv. Put together the following vectors in a data matrix:

         - `Time`, which record for each event when it happened;
         - `Event`, which records that the event happened (all 1s!);
         - `Sender`, which records the sender of the events;
         - `Receiver`, which records the receiver of the events;
         - `NonSender`, which records the sender of the non-events;
         - `NonReceiver`, which records the receiver of the non-events;
         - `Repetition`: repetition indicator of the events;
         - `Reciprocity`: reciprocity indicator of the events;
         - `NonRepetition`: repetition indicator of the non-events;
         - `NonReciprocity`: reciprocity indicator of the non-events;

   (d) Use logistic regression modelling `gam(... , family= binomial)` to analyze the effect of the repetition and reciprocity on the dynamic interactions. Compare the results to the values that you selected for your simulation.

2. **Importance of repetition, reciprocity and triadic closure in real networks.** In this example we consider one of two datasets. You can choose which one you prefer. The data is an RData file and can be found on `http://ci.inf.usi.ch/pakdd24`.

   - **Class interactions.** This describes the 692 interactions between 18 pupils and 2 teachers in a class room over 51 minutes of teaching.
      - `Class`: a matrix with 3 columns consisting of time, sender and receiver, resp.
      - `ClassIsFemale`: vector of length 20 indicating whether person is female.
      - `ClassIsTeacher`: vector of length 20 indicating whether person is a teacher.

- **World trade center.** During the 9/11 attacks in New York, there was communication ongoing between emergency services. The matrix `WTCPoliceCalls` consists of 481 communications between the police in a sequential fashion.

(a) First, we need to prepare the dataset in such a fashion that it can be analysed as an event history model. Perform the following five steps:

i. Determine the potential sender V_1 and receiver sets V_2. Are the sets the same (a socalled one-mode network) or different (a socalled two-mode or bipartite network)? Are all possible events $V_1 \times V_2$ always possible or does the risk set change?

ii. Sample for each event (s_k, r_k) among those events at risk at that time point $\mathscr{R}(t_k)$ a single non-event (s_k^*, r_k^*).

iii. For your dataset of choice, create for every event and for every non-event the variable `repetition, reciprocity, triadic closure`. Each variable records by means of a 1 or 0 whether or not for the current (s, r) (non)-event,

- `repetition`: the event (s, r) occurred before
- `reciprocity`: the event (r, s) occurred before;
- `triadic closure`: the events (r, v) and (v, s) occurred before, for some vertex v.

It is possible to include a time-window, if you prefer.

iv. Put together the following vectors in a data matrix:

- `Time`, which record for each event when it happened;
- `Event`, which records that the event happened (all 1s!);
- `Sender`, which records the sender of the events;
- `Receiver`, which records the receiver of the events;
- `NonSender`, which records the sender of the non-events;
- `NonReceiver`, which records the receiver of the non-events;
- `Repetition`: repetition indicator of the events;
- `Reciprocity`: reciprocity indicator of the events;
- `TriadicClosure`: triadic closure indicator of the events;
- `NonRepetition`: repetition indicator of the non-events;
- `NonReciprocity`: reciprocity indicator of the non-events;
- `NonTriadicClosure`: triadic closure indicator of the non-events;

v. Put together other variables of interest in your study:

- **WTC**: The vector `WTCPoliceICR` indicates which of the 37 actors in this communication network had an institutionalized coordinator role. It might be hypothesized that their sending and receiving is different from those that do not have such a role.
- **Class**: Use `ClassIsFemale, ClassIsTeacher` to create possibly interesting monadic and dyadic covariates that may drive the interactions.

(b) Use logistic regression modelling `gam(... , family= binomial)` to analyze the effect of the covariates on the dynamic interactions. Use all relevant variables (such as the monadic/dyadic variables you created in the question above) to your disposal to correct for any confounding.