Data Mining in Big Dynamic Networks

Tutorial 3: fitting non-linear effects in dynamic networks

Ernst C. Wit

May 7, 2024

1 Preliminary information

1.1 Theoretical Overview

Additive mixed effect REMs In the models we considered so far, we assumed that (i) the effect of covariates is linear and (ii) does not change over time, and (iii) the relational behaviour of actors is identical, except for different covariates. All of these assumptions are restrictive. we can write the cumulative hazard as

$$\Lambda_{sr}(t) = \int_0^t \lambda_{sr}(\tau)\, d\tau,$$

where λ_{sr} is the hazard function of the relational event (s,r). The general REM is defined as

$$\lambda_{sr}(t) = 1_{\{(s,r)\in\mathscr{R}(t)\}}\, \lambda_0(t)\, \exp\left\{\beta(t)^\top x_{sr}(t) + f(w_{sr}(t)) + \gamma^\top z_{sr}(t)\right\},$$

where $\lambda_{sr}(t)$ is only non-zero if the event (s,r) is contained in the risk set $\mathscr{R}(t)$ of possible events at time t, $\lambda_0(t)$ is the baseline hazard function unrelated to (s,r), $x_{sr}(t)$, $w_{sr}(t)$ and $z_{sr}(t)$ are the $\mathscr{H}_t$ measurable set of endogenous and exogenous (possibly) time-varying variables, and $\beta(t)$ are the time-varying effect sizes, f is a non-linear effect, and γ captures the inherent heterogeneity in the system.

Table 1: Notation in Relational Event Model

Notation	Meaning
(t,s,r)	relational event: sender s interacts with receiver r at time t
$\lambda_{sr}(t)$	Rate/hazard at which sender s contacts receiver r at time t
$\mathscr{H}_t$	History of process up until time t
L, L_P	Likelihood and partial likelihood
$\mathscr{R}(t)$	Risk set at time t
$\widetilde{\mathscr{R}}(t)$	Sampled risk set at time t
$x_{sr}(t)$	Dyadic covariate(s) with corresponding effect(s) $\beta(t)$
$w_{sr}(t)$	Dyadic covariate(s) with non-linear effect f
$z_{sr}(t)$	Dyadic covariate(s) with corresponding random effect(s) γ
a	Alter, i.e., an individual different from sender or receiver

Time-varying effects. The simplest extension of the linear model βx_{sr} is that the coefficient β varies over time. To include such an effect in `mgcv`, we create a matrix of weights with two columns consisting of the event covariate and minus the non-event covariate:

$$W = [x_{sr}; -x_{s^*r^*}]$$

The time-varying effect can be included in R, via `gam(y ~ ...+ s(Time, by=W), family = binomial)`, where `Time` represents the event times.

Non-linear effects. It could also be that the linear model βx_{sr} is not sufficient, in the sense that a non-linear function $f(x_{sr})$ would be more suitable. To include such an effect in `mgcv`, we create a matrix of weights with two columns consisting of the 1 and -1 in the columns:

$$W = [\text{rep}(1,n); \text{rep}(-1,n)]$$

and a matrix with two columns with the values of the event covariate and the non-event covariate, respectively.

$$C = [x_{sr}; x_{s^*r^*}]$$

The time-varying effect can be included in R, via `gam(y ~ ...+ s(C, by=W), family = binomial)`.

Random effect. In order to capture heteregeneity, it may be useful to include random effects in the model. In particularly, sender and receiver effects. To include e.g. sender heterogeneity in the model, we first define the matrices,

$$W = [\text{rep}(1,n); \text{rep}(-1,n)]$$

$$C = [\texttt{Sender; NonSender}]$$

Then this can be done via `gam(y ~ ...+ s(C, by=W, bs="re"), family = binomial)` in R. It is important that the columns of C are `factors`.

1.2 Packages and Functions

You will need the library mgcv in R.

2 Inference of non-linear Relational Event Models

1. **Importance of non-linear effects.** 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`.

    - **Patent citations.** This describes the 50K patent citations between 1976 and 2022. The data set contains
        - `rec_pub_year`: year of patent that is cited.
        - `lag`: time (in days) between the publication year of the citing and cited patent.
        - `rec_outd`: cited patent number of citations.
        - `jac_sim`: Jaccard similarity between citing and cited patents.
        - `cumu_cit_rec`: cumulative citations received by the cited patent.
        - `tfe`: time (in days) since the last citation of the cited patent.
        - `sim`: similarity measure between the abstract of the citing and cited patents.

- **Manufacturing company email.** 57,791 emails between 176 employees of a mid-sized manufacturing company over a nine month period starting in January 2010. The dataset contains:
  - `sender`: id of sending colleague;
  - `receiver`: id of receiving colleague;
  - `time`: time (in days) since 1 January 2010 at which moment email was sent.
  - `r4a`: reciprocity (0=non reciprocal; 1=reciprocal)
  - `t9`: transitive (0=not transitive; 1=transitive)
  - `c9`: cyclic closure (0=non cyclic 1=cyclic)
  - `rb9`: receiver balance (0=not balanced; 1=balanced)
  - `sb9`: sender balance (0=not balanced; 1=balanced)

(a) Download the data of your choice in R. Explore the data with a small exploratory analysis.

(b) Create the matrices W and C described above for the non-linear effects that you would like to fit.

(c) Use logistic regression modelling `gam(... , family= binomial)` to analyze the effect of the covariates on the dynamic interactions.