SAND2019-9785R

A Model: Time-Clustered Events

Alex Foss

August 19, 2019

1 Model Statement

We have m sources and n sinks. Let Θ_{ij} denote a collection of parameters describing the relationship between source i and sink j, with

$$\Theta_{ij} = (\alpha_{ij}, \beta_{ij}, \kappa_{ij}, \gamma_{ij}, \mu_{ij}).$$

Each pair of source and sink generate clusters of events over time under the following model. Cluster inter-arrival times are Weibull-distributed. That is, cluster k for source i and sink j begins at time

$$S_{ijk} = S_{ij(k-1)} + W_{ijk}$$

$$W_{ijk} \sim Weibull(\alpha_{ij}, \beta_{ij}).$$

Event start-times within cluster k are denoted $\vec{Z}_{ijk}$, with

$$|\vec{Z}_{ijk}| \sim Pois(\mu_{ij}) \tag{1}$$

$$Z_{ijk\ell} \sim Unif(0, E_{ijk}) \tag{2}$$

$$E_{ijk} \sim Gamma(\kappa_{ij}, \gamma_{ij}). \tag{3}$$

The vector of event start-times for cluster ijk is given by $S_{ijk} + \vec{Z}_{ijk}$. In other words, cluster ijk consists of $|\vec{Z}_{ijk}|$ events which occur between time S_{ijk} and time $S_{ijk} + E_{ijk}$.

2 The data

Assume we have a large (say, about 1 million) set of event timestamps, each associated with a source and sink. Clusters are unknown and must be inferred from the data. All sink IDs are known, but some source IDs are missing. Assume that there exist sources that have never been observed, that is, we can't assume that a missing source belongs to one of the observed sources.

3 Notes

Some notes:

1. Primary goal: Forecast event timestamps. Ideally, forecast timestamps for the source and sink pairs.
2. Many sources and sinks have no mutual events. Handling these pairs with no events is important. Do these sources truly have no relationship, or have we simply not observed enough events yet? For now, we can assume no relationship.
3. Overlapping clusters are likely; however, overlapping uniform clusters are not necessarily identifiable: for example, a sample in equal proportions from $Unif(0, 3)$ and $Unif(1, 2)$ is indistinguishable from a sample in equal proportions from $Unif(0, 2)$ and $Unif(1, 3)$. This identifiability problem will have to be addressed, perhaps by changing the distribution specified in equation 2.

Some complications and future extensions:

1. Equation 1 may need to be generalized; something more skewed like the negative binomial is probably more appropriate.
2. The clusters may be of a few distinct flavors. This flavor would control the parameters and distributions in equations 2 and 3. Some sources can only produce certain flavors.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

3. It may be more appropriate to use some time-dependent intensity function instead of Weibull inter-arrival times. This intensity function may be related to time-of-day and seasonality, as well as external events that are unknown. Different sources probably require their own intensity functions.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy.s National Nuclear Security Administration under contract DE-NA0003525.