The Dynamic Latent Block Model for Sparse and Evolving Count Matrices

Giulia MARCHELLO

Università degli Studi di Perugia
Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai research team
giulia.marchello@inria.fr

Joint work with M. Corneli & C. Bouveyron

The problem

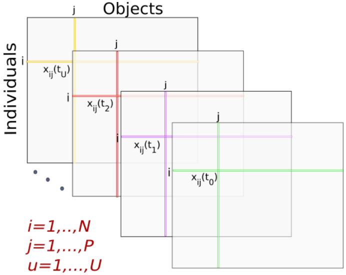

Figure: The data

- $X_{ij}(t)$ indicates the number of interactions occurring between the individual i and the item j in the time period $t \in [0, T]$.
- When no interaction between individuals and objects occurs, we have missing values and the number of interactions is 0 for time t.
- Segmentation of the continuous time period $[0, T]$ in U subintervals, with $I_u = [t_{u-1} - t_u[$.

$$0 = t_0 < t_1 < \cdots < t_U = T.$$

The goal is to cluster each individual, i, each product, j, and each time partition, u, to homogeneous hidden clusters respectively identified by K (row clusters), L (column clusters) and C (time clusters).

The Dynamic Latent Block Model

- The number of interactions between individuals and objects follows a non-homogeneous Poisson process (NHPP) where the intensity function $\lambda(t)$ only depends on the clusters they belong to.

$$p(X_{iju}|z_{ik}w_{jl}s_{uc} = 1, \lambda_{k\ell c}) = \left(\frac{(\lambda_{klc})^{X_{iju}}}{X_{iju}!} \exp\left(-\lambda_{klc}\right) \right) \tag{1}$$

- X_{iju} represents the number of interactions between i and j in the considered time partition I_u, it is the generic element of the tensor X with dimensionality $N \times P \times U$.
  The latent structure of the model is identified by:
  - z $= (z_{ik}; i = 1, ..., n; k = 1, ..., K)$: it represents the clustering of rows into K groups.
    - The z_i are i.i.d with $\mathrm{z}_i \sim \mathscr{M}(1; \gamma)$.
  - w $= (w_{j\ell}; j = 1, ..., p; \ell = 1, ..., L)$: it represents the clustering of columns into L groups.
    - The w_j are i.i.d with $\mathrm{w}_j \sim \mathscr{M}(1; \rho)$.
  - s $= (s_{uc}; u = 1, ..., U; c = 1, ..., C)$: it represents the clustering of time intervals into C time clusters.
    - The s_u are i.i.d with $\mathrm{s}_u \sim \mathscr{M}(1, \delta)$.
- As inference algorithm a stochastic version of the EM algorithm is used: the SEM-Gibbs while for the model selection we used the ICL criterion.

Real data application: Amazon Fine Food dataset

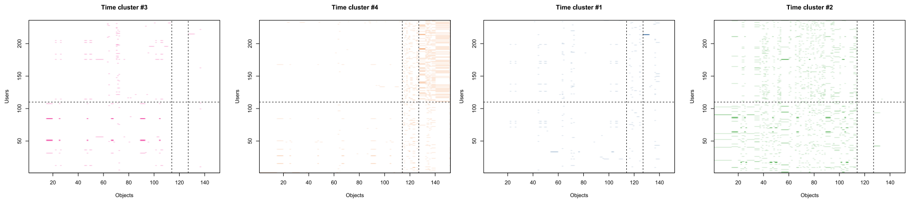

Figure: Reorganized incidence matrix for each time cluster. Rows and columns clusters are delimited by the dashed lines while the colored dots marks an interaction (i.e. review) between a user and a product.

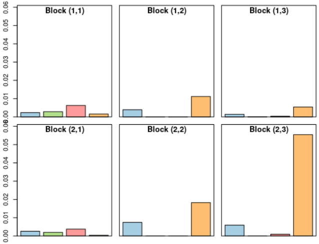

Figure: Estimated parameters of the probability distribution of interactions between users and products, according to time clusters