A Dynamic Latent Block Model for Co-clustering of Count Data Streams

Working group on Model-Based clustering

Giulia MARCHELLO

Joint work with
C. Bouveyron & M. Corneli.

Introduction

The goal:

- Perform co-clustering of a stream of evolving count data,
- identify atypical behavior in a real-time,
- detect abrupt changes of clusters memberships.

Further Application:

- automatic safety signal detection from Pharmacovigilance data,
- find unexpected patterns in the spontaneous reports received by the RCPV (Regional Center of Pharmacovigilance),
- assist medical experts thanks to the highly intuitive interpretation of the results.

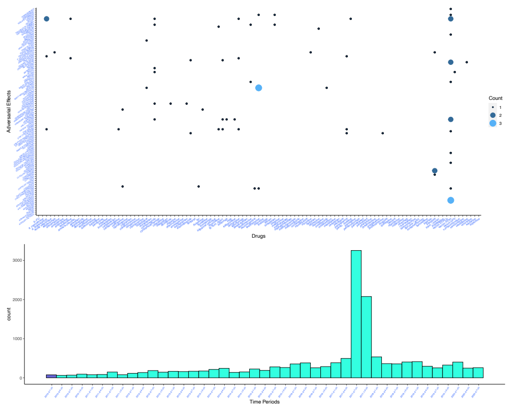

Figure: Evolution of spontaneous reports received by the RCPV from 2010 to 2020, a small sample is considered.

Stream-dLBM

- Modeling cluster memberships:
  - Time-dependent multinomial random variables:
    - $Z_i(t) \sim \mathcal{M}(1, \alpha(t) := (\alpha_1(t), \ldots, \alpha_Q(t)))$,
    - $\alpha_q(t) = \text{softmax}(a_q(t))$,
    - $a(t) = a(t-1) + f_Z(a(t-1))$.
    - $W_j(t) \sim \mathcal{M}(1, \beta(t) := (\beta_1(t), \ldots, \beta_L(t)))$,
    - $\beta_\ell(t) = \text{softmax}(b_\ell(t))$,
    - $b(t) = b(t-1) + f_W(b(t-1)) \forall t \in \{1, \ldots, T\}$,
- Modeling the interaction process:
  - $X_{ij}(t)|Z_i(t), W_j(t) = 1 \sim \mathcal{P}\left(\Lambda_{Z_i(t)W_j(t)}\right)$,

Algorithm 1 VEM-SGD Algorithm

Require: X, Q, L, max_iter, H, K, (where H and K are the hidden layer sizes.)
Initialization of τ and η with kmeans.
Initialization of $\alpha_q(t) = 1/Q$ and $\beta_\ell(t) = 1/L$
M-Step: $\hat{\Lambda}_{q\ell} = \dfrac{\sum_{i,j,t} \tau_{iq}(t)\eta_{j\ell}(t)X_{ij}(t)}{\sum_{i,j,t} \tau_{iq}(t)\eta_{j\ell}(t)}$.
for epoch = 1 to Epochs **do**
  Update $\hat{\alpha}(t)$:
  Loss Evaluation: $\ell = -\sum_{T,Q,M} \tau_{iq}(t)\text{softmax}(\alpha_q(t))$;
  Algorithm backpropagation;
  Numerical optimization of $\alpha(t)$ with SGD.
  Update $\hat{\beta}(t)$:
    Similar steps as for $\alpha(t)$.
end for
E-Step: Updatings of $\tau(t)$ and $\eta(t)$ by maximizing the lower bound $\mathcal{L}(q, \theta)$.

Figure: Pseudocode of the VEM-SGD inference algorithm.

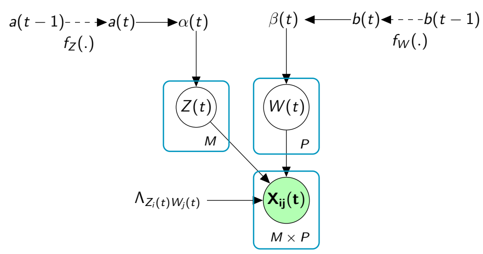

Figure: Graphical representation of stream-dLBM.

Experiment on simulated data

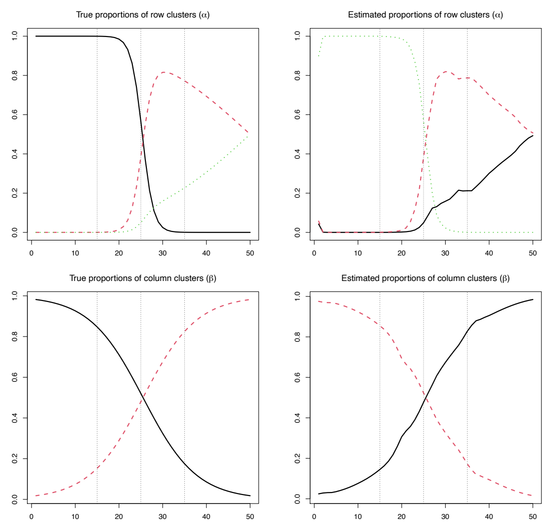

Figure: Comparison between the evolution of the true mixing parameters (left side) and the estimated ones (right side).

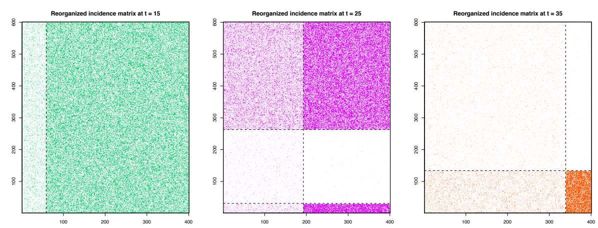

Figure: Reorganized incidence matrices in 3 different time instants according to the estimations of row and column partitions.The blocks are also delimited by black dashed lines.