Supplementary material for the article "A Deep Dynamic Latent Block Model for the Co-clustering of Zero-Inflated Data Matrices"

Giulia Marchello[1], Marco Corneli[1,2], and Charles Bouveyron[1]

[1] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.
[2] Université Côte d'Azur, Laboratoire CEPAM, Nice, France.

1 Proofs

1.1 Likelihood of complete data

The model described so far can be adapted to any zero-inflated distribution. The first formulation as well as the most well-known concerns the Zero-Inflated Poisson, from an article by Lambert (1992). However, other distributions such as Zero-Inflated Negative Binomial (Ridout et al., 2001), Zero-Inflated Beta (Ospina and Ferrari, 2012), Zero-Inflated log-normal (Li et al., 2011) could be coupled with the present modeling.
In the following to ease the readability of the inference procedure we make use of the Zero-Inflated Poisson (ZIP) formulation to illustrate our approach. Hence, we can write $X_{ij}(t)|Z_i(t), W_j(t) \sim ZIP(\Lambda_{Z_i(t),W_j(t)}, \pi(t))$ and develop Eq. 2 as follows:

$$\begin{cases} X_{ij}(t)|Z_i(t), W_j(t) = 0 & \text{with probability } \pi(t) \\ X_{ij}(t)|Z_i(t), W_j(t) \sim \mathcal{P}(\Lambda_{Z_i(t),W_j(t)}) & \text{with probability } 1-\pi(t) \end{cases} \tag{1}$$

where Λ is a $Q \times L$ matrix, denoting the block-dependent Poisson intensity function and $\pi(t)$ represents the sparsity at any given time period, with $t = 0, \dots, T$. Then, we rewrite Eq. (1) by introducing the hidden random matrix, $A \in \{0,1\}^{N \times M}$, where for all i and j:

$$A_{ij}(t) \sim \mathcal{B}(\pi(t)),$$

with $\mathcal{B}(\pi)$ denoting the Bernoulli probability mass function of parameter π and such that

$$\begin{aligned} &A_{ij}(t) = 1 \Rightarrow X_{ij}(t)|Z_i(t), W_j(t) = 0 \\ &A_{ij}(t) = 0 \Rightarrow X_{ij}(t)|Z_i(t), W_j(t) \sim \mathcal{P}(\Lambda_{Z_i(t),W_j(t)}). \end{aligned} \tag{2}$$

As expressed in Section 2.1 of the main paper, the likelihood of the complete data can be written as

$$p(X, A, Z, W|\theta) = p(X|A, Z, W\Lambda, \pi)p(A \mid \pi)p(Z|\alpha)p(W|\beta), \tag{3}$$

the components can be further developed as

$$p(X|A, Z, W, \Lambda, \pi) = \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{t=1}^{T} \mathbf{1}_{\{X_{ij}(t)=0\}}^{A_{ij}(t)} \left\{ \left(\frac{\Lambda_{Z_i(t)W_j(t)}^{X_{ij}(t)}}{X_{ij}(t)!} \exp(-\Lambda_{Z_i(t)W_j(t)}) \right)^{(1-A_{ij}(t))} \right\}, \tag{4}$$

$$p(A|\pi) = \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{t=1}^{T} \pi(t)^{A_{ij}(t)} \Big(1-\pi(t)\Big)^{(1-A_{ij}(t))}, \tag{5}$$

$$p(Z|\alpha) = \prod_{i=1}^{N}\prod_{q=1}^{Q}\prod_{t=1}^{T}\alpha_q(t)^{Z_{iq}(t)}, \tag{6}$$

$$p(W|\beta) = \prod_{j=1}^{M}\prod_{\ell=1}^{L}\prod_{t=1}^{T}\beta_\ell(t)^{W_{j\ell}(t)}. \tag{7}$$

1.2 Optimization of the factor q(A)

Let us consider the derivation of the update equation for the factor $q(A)$. The sequential update for the factor $q(A)$ can be computed through the log of the optimized factor, where all the terms that do not depend on A are absorbed in the constant term. Starting from Eq. 13, we use the decomposition in Eq. 10 of the main paper, then we substitute the conditional distributions on the right-hand side.

$$\begin{aligned}
\log(q^*(A)) &= E_{W,Z}[\log(p(X,Z,W,A|\theta)] + const;\\
&= E_{W,Z}[\log(p(X|Z,W,A,\Lambda,\pi)] + E_{W,Z}[\log(A|\pi)] + const\\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\Bigg\{A_{ij}(t)[\log \mathbf{1}_{\{X_{ij}(t)=0\}}] + (1-A_{ij}(t))\Bigg[\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big[E[Z_{iq}(t)]E[W_{j\ell}(t)]X_{ij}(t)\log\Lambda_{q\ell}\\
&- E[Z_{iq}(t)]E[W_{j\ell}(t)]\Lambda_{q\ell}\Big] - \log X_{ij}(t)!\Bigg] + A_{ij}(t)\log\pi(t) + (1-A_{ij}(t))\log(1-\pi(t))\Bigg\} + const\\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\Bigg\{A_{ij}(t)\log\pi\mathbf{1}_{\{X_{ij}(t)=0\}} + (1-A_{ij}(t))\Bigg[\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big[E[Z_{iq}(t)]E[W_{j\ell}(t)]X_{ij}(t)\log\Lambda_{q\ell}\\
&- E[Z_{iq}(t)]E[W_{j\ell}(t)]\Lambda_{q\ell}\Big] - \log X_{ij}(t)! + \log(1-\pi(t))\Bigg]\Bigg\} + const\\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\Bigg\{A_{ij}(t)\log\pi\mathbf{1}_{\{X_{ij}(t)=0\}} + A_{ij}(t)\Bigg[\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big[-E[Z_{iq}(t)]E[W_{j\ell}(t)]X_{ij}(t)\log\Lambda_{q\ell}\\
&+ E[Z_{iq}(t)]E[W_{j\ell}(t)]\Lambda_{q\ell}\Big] + \log X_{ij}(t)! - \log(1-\pi(t))\Bigg]\Bigg\} + const\\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\Bigg\{A_{ij}(t)\Bigg[\log\pi\mathbf{1}_{\{X_{ij}(t)=0\}} + \sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big[-E[Z_{iq}(t)]E[W_{j\ell}(t)]X_{ij}(t)\log\Lambda_{q\ell}\\
&+ E[Z_{iq}(t)]E[W_{j\ell}(t)]\Lambda_{q\ell}\Big] + \log X_{ij}(t)! - \log(1-\pi(t))\Bigg]\Bigg\} + const.
\end{aligned} \tag{8}$$

We can then recognize the functional form of the Bernoulli distribution by indicating:

$$\begin{aligned}
\log q^*(A) &\propto \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} A_{ij}(t)\log\delta_{ij}(t) + (1-A_{ij}(t))\log(1-\delta_{ij}(t)),\\
&\propto \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} A_{ij}(t)\frac{\log\delta_{ij}(t)}{1-\log\delta_{ij}(t)}
\end{aligned} \tag{9}$$

where $\delta_{ij}(t)$ is defined as:

$$\delta_{ij}(t) = \frac{\exp(R_{ij}(t))}{1+\exp(R_{ij}(t))}, \tag{10}$$

with $R_{ij}(t)$ defined as:

$$\begin{aligned} R_{ij}(t) = \log(\pi(t)\mathbf{1}_{\{X_{ij}(t)=0\}}) + \sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big[- E[Z_{iq}(t)]E[W_{j\ell}(t)]X_{ij}(t)\log \Lambda_{q\ell} + \\ + E[Z_{iq}(t)]E[W_{j\ell}(t)]\Lambda_{q\ell}\Big] + \log X_{ij}(t)! - \log(1-\pi(t)). \end{aligned} \tag{11}$$

1.3 Optimization of the factor q(Z)

Let us now consider the derivation of the update equation for the factor $q(Z)$. The sequential update for the factor $q(Z)$ can be computed through the logarithm of the optimized factor, where all the terms that do not depend on Z are absorbed in the constant term. Starting from Eq. 14, we use the decomposition in Eq. 10 of the main paper, then we substitute the conditional distributions on the right-hand side.

$$\begin{aligned} \log q^*(Z|\theta) &= E_{W,A}[\log p(X, A, Z, W \mid \theta)] \\ &= E_{W,A}[\log(p(X \mid A, Z, W, \Lambda, \pi) + \log p(Z \mid \alpha)] \\ &= E_{W,A}\Bigg[\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\Bigg\{(1-A_{ij}(t))\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big\{Z_{iq}(t)W_{j\ell}(t)X_{ij}(t)\log(\Lambda_{q\ell}) - Z_{iq}(t)W_{j\ell}(t)\Lambda_{q\ell}\Big\} \\ &- (1-A_{ij}(t))\log(X_{ij}(t)!)\Bigg\}\Bigg] + \sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}(t)\log(\alpha_q(t)) + \text{const}, \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}(1-E[A_{ij}(t)])\Bigg[\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big\{Z_{iq}(t)E[W_{j\ell}(t)]X_{ij}(t)\log(\Lambda_{q\ell}) + \\ &- Z_{iq}(t)E[W_{j\ell}(t)]\Lambda_{q\ell}\Big\}\Bigg] + \sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}(t)\log(\alpha_q(t)) + \text{const}, \\ &= \sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}(t)\Bigg[\sum_{j=1}^{M}\sum_{\ell=1}^{L}\Big\{(1-E[A_{ij}(t)])\Big[E[W_{j\ell}(t)]X_{ij}(t)\log(\Lambda_{q\ell}) + \\ &- E[W_{j\ell}(t)]\Lambda_{q\ell}\Big]\Big\} + \log(\alpha_q(t))\Bigg] + \text{const}. \end{aligned} \tag{12}$$

We can then recognize the functional form of the multinomial distribution. Thus, we can write:

$$\log q^*(Z|\theta) = \sum_i \sum_t \sum_q Z_{iq}(t)\log r_{iq(t)} + const. \tag{13}$$

Taking the exponential on the two sides, we obtain:

$$q(Z_i) = \prod_{t=1}^{T}\prod_{q=1}^{Q} r_{iq}(t)^{Z_{iq}(t)}, \tag{14}$$

where $r_{iq}(t)$ is denoted by:

$$r_{iq}(t) \propto \exp\left(\sum_{j=1}^{M}\sum_{\ell=1}^{L}\left\{(1-E[A_{ij}(t)])\Big[E[W_{j\ell}(t)]X_{ij}(t)\log(\Lambda_{q\ell}) - E[W_{j\ell}(t)]\Lambda_{q\ell}\Big]\right\} + \log(\alpha_q(t))\right). \quad (15)$$

However, this distribution needs to be normalized because the matrix $Z(t)$ is a binary matrix and the elements sum to 1 over the values of Q. We can then obtain:

$$q(Z_i) = \prod_{t=1}^{T}\prod_{q=1}^{Q}\tau_{iq}(t)^{Z_{iq}(t)}, \quad (16)$$

where

$$\tau_{iq}(t) = \frac{r_{iq}(t)}{\sum_{q_0=1}^{Q} r_{iq_0}(t)}. \quad (17)$$

1.4 Optimization of the factor q(W)

Let us now consider the derivation of the update equation for the factor $q(W)$. The sequential update for the factor $q(W)$ can be computed through the log of the optimized factor, where all the terms that do not depend on W are absorbed in the constant term.

Proposition 1. *Denoting by $\eta_{j\ell}(t) := q(W_j\ell(t) = 1)$ the variational probability of success of $W_{j\ell}(t)$, the optimal update of is:*

$$\eta_{j\ell}(t) = \frac{s_{j\ell}(t)}{\sum_{\ell_o=1}^{L} s_{j\ell_o}(t)}, \quad (18)$$

where :

$$s_{j\ell}(t) \propto \exp\left(\sum_{i=1}^{N}\sum_{q=1}^{Q}\left\{(1-E[A_{ij}(t)])\Big[E[Z_{iq}(t)]X_{ij}(t)\log(\Lambda_{q\ell}) - E[Z_{iq}(t)]\Lambda_{q\ell}\Big]\right\} + \log(\beta_\ell(t))\right). \quad (19)$$

The proof is symmetric to the one developed for $\tau_{iq}(t)$ in Proposition 2.

1.5 Derivation of the lower bound

In order to obtain the updating of the parameter set θ, the objective of the M-Step is the maximization of the lower bound $\mathcal{L}(q,\theta)$ with respect to $\theta = (\Lambda, \alpha(t), \beta(t), \pi(t))$, while holding the variational distribution $q(\cdot)$ fixed.

Proposition 2. *By developing the Eq. 11 of the main paper, the variational lower bound $\mathcal{L}(q,\theta)$ can be written as*

$$\begin{aligned}\mathcal{L}(q,\theta) = &\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\Bigg\{\delta_{ij}(t)\log(\pi(t)\mathbf{1}_{\{X_{ij}(t)=0\}}) + (1-\delta_{ij}(t))\Bigg[\log(1-\pi(t)) + \\ &+\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big\{\tau_{iq}(t)\eta_{j\ell}(t)X_{ij}(t)\log\Lambda_{q\ell} - \tau_{iq}(t)\eta_{j\ell}(t)\Lambda_{q\ell}\Big\}\Bigg] + \\ &-(1-\delta_{ij}(t))\log(X_{ij}(t)!)\Bigg\} + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}(t)\log(\alpha_q(t)) + \\ &+\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{\ell=1}^{L}\eta_{j\ell}(t)\log(\beta_\ell(t)) - \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}(t)\log\tau_{iq}(t) + \\ &-\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{\ell=1}^{L}\eta_{j\ell}(t)\log(\eta_{j\ell}(t)) - \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\Big(\delta_{ij}(t)\log(\delta_{ij}(t)) + (1-\delta_{ij}(t))\log(1-\delta_{ij}(t))\Big).\end{aligned} \tag{20}$$

Proof. Starting from Eq. 11 we obtain the final expression of the variational lower bound $\mathcal{L}(q,\theta)$ by developing the expression

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_{A,Z,W} q(A,Z,W)\log\frac{p(X|A,Z,W,\Lambda)p(A\mid\pi)p(Z|\alpha)p(W|\beta)}{q(A,Z,W)} \\ &= E_{A,Z,W}\left[\log\frac{p(X|A,Z,W,\Lambda)p(A\mid\pi)p(Z|\alpha)p(W|\beta)}{\prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{t=1}^{T}q(A_{ij}(t))\prod_{i=1}^{N}\prod_{t=1}^{T}q(Z_i(t))\prod_{j=1}^{M}\prod_{t=1}^{T}q(W_j(t))}\right] \\ &= E_{A,Z,W}[\log p(X|A,Z,W,\Lambda)] + E_A[\log p(A\mid\pi)] + E_Z[\log p(Z|\alpha)] + \\ &+ E_W[\log p(W|\beta))] - E_Z[\log\prod_i q(Z_i)] + \\ &- E_W[\log\prod_j q(W_j)] - E_A[\log\prod_i\prod_j q(A_{ij})].\end{aligned} \tag{21}$$

Then we substitute the results obtained in the VE-Step, denoting $E[A_{ij}(t)] = \delta_{ij}(t)$, $E[Z_{iq}(t)] = \tau_{iq}(t)$ and $E[W_{j\ell}(t)] = \eta_{j\ell}(t)$, in order to obtain the final expression of the lower bound that can be written as follows:

$$\begin{aligned}\mathcal{L}(q,\theta) = &\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\Bigg\{\delta_{ij}(t)\log(\pi(t)\mathbf{1}_{\{X_{ij}(t)=0\}}) + (1-\delta_{ij}(t))\Bigg[\log(1-\pi(t)) + \\ &+\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\Big\{\tau_{iq}(t)\eta_{j\ell}(t)X_{ij}(t)\log\Lambda_{q\ell} - \tau_{iq}(t)\eta_{j\ell}(t)\Lambda_{q\ell}\Big\}\Bigg] + \\ &-(1-\delta_{ij}(t))\log(X_{ij}(t)!)\Bigg\} + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}(t)\log(\alpha_q(t)) + \\ &+\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{\ell=1}^{L}\eta_{j\ell}(t)\log(\beta_\ell(t)) - \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}(t)\log\tau_{iq}(t) + \\ &-\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{\ell=1}^{L}\eta_{j\ell}(t)\log(\eta_{j\ell}(t)) - \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\Big(\delta_{ij}(t)\log(\delta_{ij}(t)) + (1-\delta_{ij}(t))\log(1-\delta_{ij}(t))\Big).\end{aligned} \tag{22}$$

1.6 Update of Λ

Here our goal is to derive the update of the Zero-inflated Poisson intensity parameter, Λ. The variational distribution $q(A, Z, W)$ is kept fixed, while the lower bound is maximized with respect to Λ, to obtain its update, $\hat{\Lambda}$. To find the optimal update expression of Λ we compute the derivative of the lower bound $\mathcal{L}(q, \theta)$ in Eq. 20 of this Appendix, with respect to Λ

$$\frac{\partial \log \mathcal{L}(q,\theta)}{\partial \Lambda_{q\ell}} = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}(1-\delta_{ij}(t))\Big[\frac{\tau_{iq}(t)\eta_{j\ell}(t)X_{ij}(t)}{\Lambda_{q\ell}} - \tau_{iq}(t)\eta_{j\ell}(t)\Big], \tag{23}$$

Equaling $\frac{\partial \log \mathcal{L}(q,\theta)}{\partial \Lambda_{q\ell}}$ to zero leads to:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}(1-\delta_{ij}(t))\Big[\tau_{iq}(t)\eta_{j\ell}(t)X_{ij}(t) - \tau_{iq}(t)\eta_{j\ell}(t)\Lambda_{q\ell}\Big] = 0, \tag{24}$$

and then

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}(1-\delta_{ij}(t))\tau_{iq}(t)\eta_{j\ell}(t)\Lambda_{q\ell} = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\tau_{iq}(t)\eta_{j\ell}(t)\Big[X_{ij}(t) - X_{ij}(t)\delta_{ij}(t)\Big]. \tag{25}$$

We finally get:

$$\hat{\Lambda}_{q\ell} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\tau_{iq}(t)\eta_{j\ell}(t)\Big(X_{ij}(t) - \delta_{ij}(t)X_{ij}(t)\Big)}{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\tau_{iq}(t)\eta_{j\ell}(t)\Big(1-\delta_{ij}(t)\Big)}. \tag{26}$$

2 Numerical experiments

The main purpose of this section is to highlight the most important features of our zero-inflated dLBM algorithm over simulated data sets in the Poisson scenario. We aim at demonstrating the validity of the inference algorithm and model selection criterion presented in the previous sections. The first experiment consists in applying $ZI_{\mathcal{P}}$-dLBM to a specific data set with evolving block pattern and sparsity to show that it recovers the data structure. The second experiment shows that $ZI_{\mathcal{P}}$-dLBM is able to uncover clusters being initially empty, filling up over time, then emptying again. The third experiment shows the robustness of $ZI_{\mathcal{P}}$-dLBM when the initial number of clusters is not the actual one, thus testing the performance of the model in case of poor initialization. The fourth experiment demonstrates the model selection procedure on 50 simulated date sets. In the fifth experiment, we finally compare the performances of $ZI_{\mathcal{P}}$-dLBM with the dLBM model in two simulated scenarios, with constant and evolving sparsity, respectively. All the experiments on simulated data were realized on data sets with $N = 600$ rows, $M = 400$ columns and $T = 50$ time instants.

2.1 Introductory example

As a first example, we simulate a data set with dimension $600 \times 400 \times 50$ and with $Q = 3$ groups of rows, $L = 2$ groups of columns. The level of sparsity ranges from 80% to 90% in the time period. The values of the other simulated parameters in this experiment are:

Cluster	α	β
1	0.2 to 0.8	0.1 to 0.99
2	0.18 to 0.14	0.99 to 0.1
3	0.6 to 0.06	-

$$\Lambda = \begin{bmatrix} 6 & 4 \\ 1 & 2 \\ 7 & 3 \end{bmatrix}$$

We apply $ZI_{\mathcal{P}}$-dLBM to the simulated data set with the actual values of Q and L to show the ability of the model to fully recover the model parameters.
Figure 1 shows the evolution of the the lower bound, expressed in Eq. 20, that $ZI_{\mathcal{P}}$-dLBM aims to maximize. We can notice the convergence is reached in less than 10 iterations, in this example.
Figure 2, displays the reorganized incidence matrices at time instants $t = 10$ and $t = 30$, respectively: the rows and columns of the incidence matrix are permuted according to the estimates of the latent variables $\hat{Z}$ and $\hat{W}$, in such a way that nearby rows (columns) belong to the same cluster of rows (columns). The blocks are also delimited by black dashed lines. The density of points in each block depends on the intensity function of the Poisson distribution Λ.
Figure 3 shows the evolution of the estimated mixture parameters $\hat{\alpha(t)}$, $\hat{\beta(t)}$ and $\hat{\pi(t)}$ along the time period, represented on the x-axes. These parameters are estimated through the stochastic gradient descent technique, linked with the neural networks. By looking at these figures, we see the true parameters on the left column, the output of the initialization procedure in the middle and the results the $ZI_{\mathcal{P}}$-dLBM estimates on the right. The comparison between the simulated and estimated parameter evolution shows that the model fully recovers the actual values over time, modulo the switched labels for the mixture proportions.
From these results we can clearly see that our algorithm perfectly identifies the composition of the original clusters and it recovers the evolution of the mixing proportion over time.

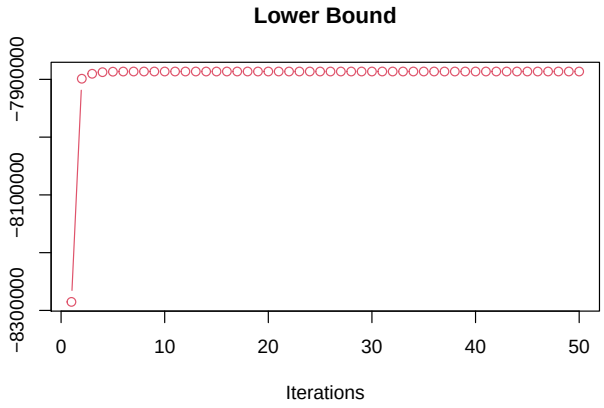


Fig. 1: Lower bound maximization throughout the iterations of the $ZI_{\mathcal{P}}$-dLBM algorithm.

2.2 Robustness of the initialization procedure

In this section we perform two experiments to test the robustness of the model to initialization. In the first experiment, we initialize parameters with wrong number of clusters, while in the second experiment the data are simulated with particularly complex dynamics. In fact, we test the model's ability to identify a cluster that is empty at the beginning of the period, which fills up and then empties again. As for the first experiment to test the robustness of our initialization strategy, $ZI_{\mathcal{P}}$-dLBM was intentionally initialized with a higher than the optimal number of clusters. In fact, although the data were simulated with $Q = 3$ and $L = 2$, both the initialization process and $ZI_{\mathcal{P}}$-dLBM were run with $Q = 5$ and $L = 4$. Figure 4 shows on the left column the evolution of the simulated mixture proportions and the sparsity parameter,

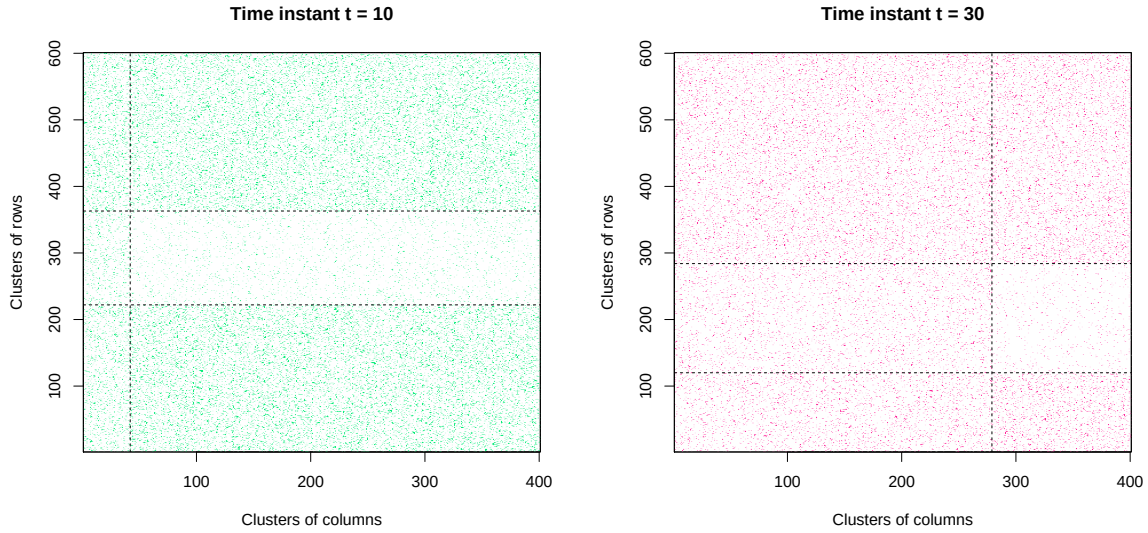


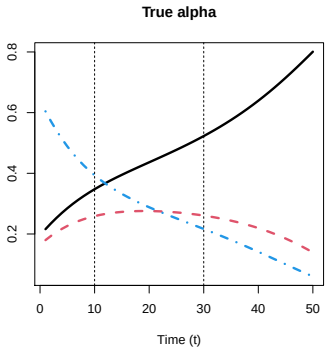
Fig. 2: Reorganized incidence matrices at time instants $t = 10$ and $t = 30$ according to the estimates of the cluster memberships. Nearby rows (columns) belong to the same cluster of rows (columns). The blocks are also delimited by black dashed lines.

in the middle column their initialization, and on the right column the results of the estimates provided by $ZI_{\mathcal{P}}$-dLBM. We can see that the initialization of $\alpha(t)$ in Figure 4b is rather poor. Nevertheless, $ZI_{\mathcal{P}}$-dLBM finds the right trend of the mixture proportions over time, effectively emptying the two superfluous clusters. Furthermore, to evaluate the quality of the clustering, we use a measure called CARI, recently introduced by **?**. This new criterion is based on the Adjusted Rand Index (**?**) and it was developed especially for being applied to co-clustering methods. The closer the index is to 1, the more both the row and column partitions are close to the actual ones, whereas the closer the value is to 0, the greater the difference between the true and estimated labels. In this experiment a CARI index value was calculated for each time instant; the obtained CARI index is 0.98.

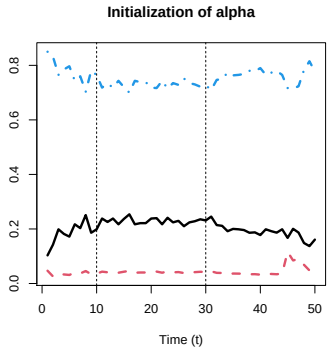
Now, as for the second experiment of testing the robustness of the model to initialization, we simulate the data in such a way that in the clusters in line there is one that is empty at the beginning of the period under consideration, then fills up towards the middle of the period, and then empties again at the end, Figure 5a shows this dynamic. Through this experiment we want to show how $ZI_{\mathcal{P}}$-dLBM is able to find the right evolution of mixture proportions despite the complex dynamics. Figure 5 depicts the simulated parameters on the left, the initial estimates in the middle, and the final estimates on the right. Looking at the middle part, in Figure 5b, we can see that the initialization process is not particularly helpful to the model because of the switched labels and a poor parameters estimation. Despite the initialization, we see in Figure 5c that $ZI_{\mathcal{P}}$-dLBM is perfectly able to recognize the initially empty cluster which then gradually fills up and empties again. In this experiment, the CARI index, obtained by averaging the indexes over time, is 0.95.

2.3 Model selection experiment

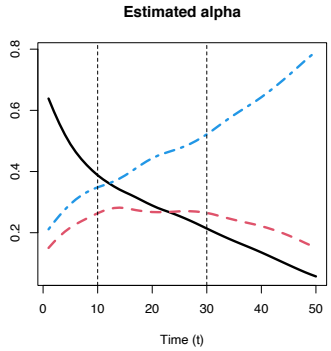
Previous experiments have allowed us to attest that the initialization strategy is globally robust and that the application of $ZI_{\mathcal{P}}$-dLBM allows us to correct for poor initializations with respect to the number of clusters in rows or columns. Therefore, in this experiment we test the global capability in choosing the



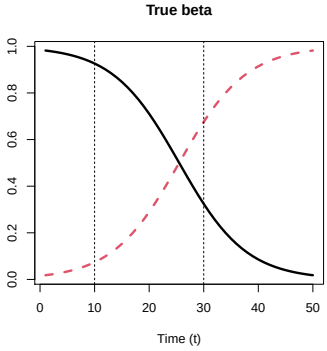
(a) True $\alpha(t)$.



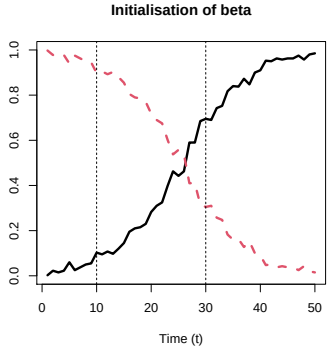
(b) Initialization of $\alpha(t)$.



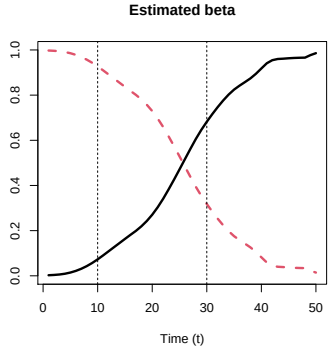
(c) Estimated $\alpha(t)$.



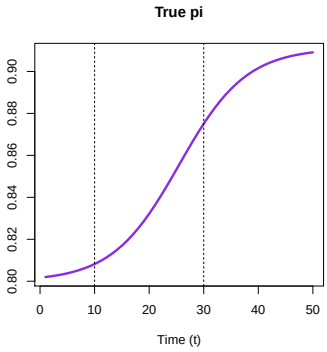
(d) True $\beta(t)$.



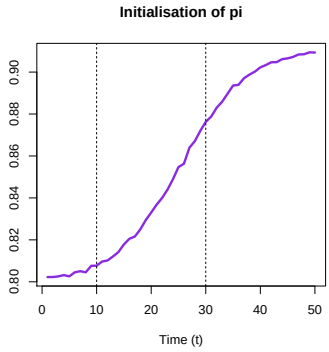
(e) Initialization of $\beta(t)$.



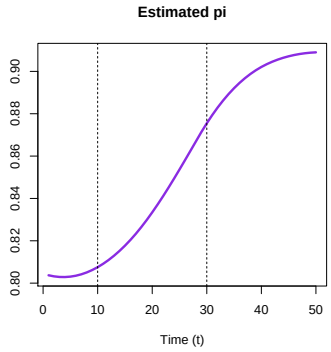
(f) Estimated $\beta(t)$.



(g) True $\pi(t)$.

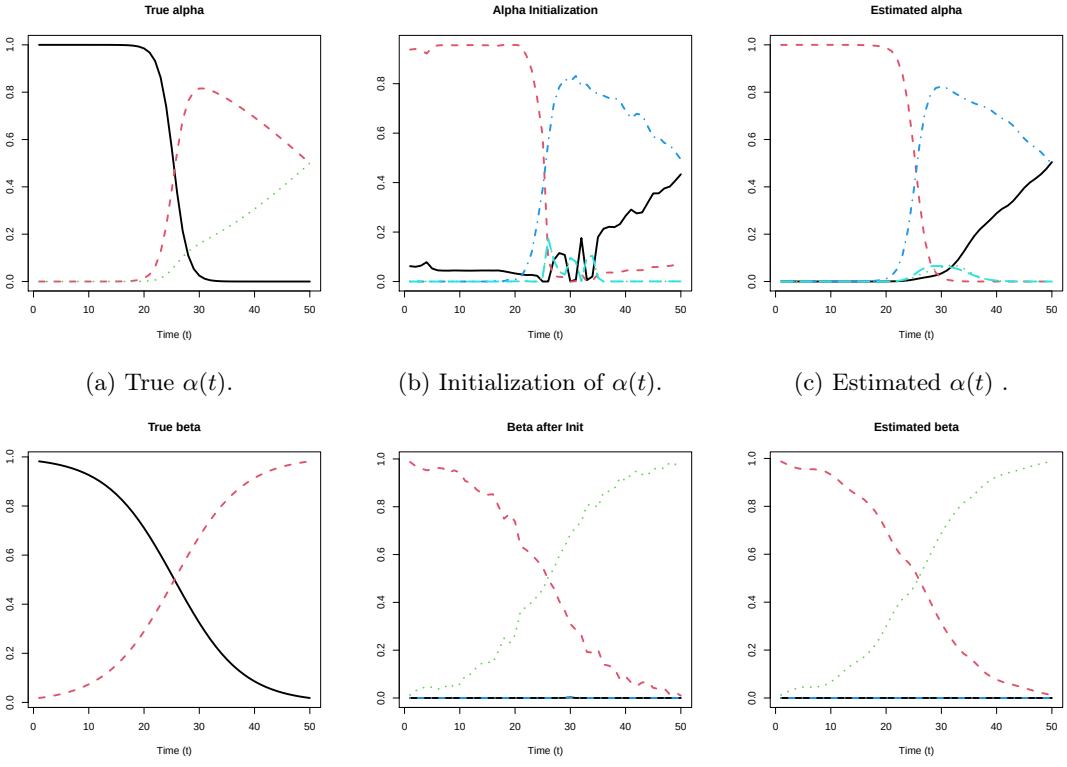


(h) Initialization of $\pi(t)$.



(i) Estimated $\pi(t)$.

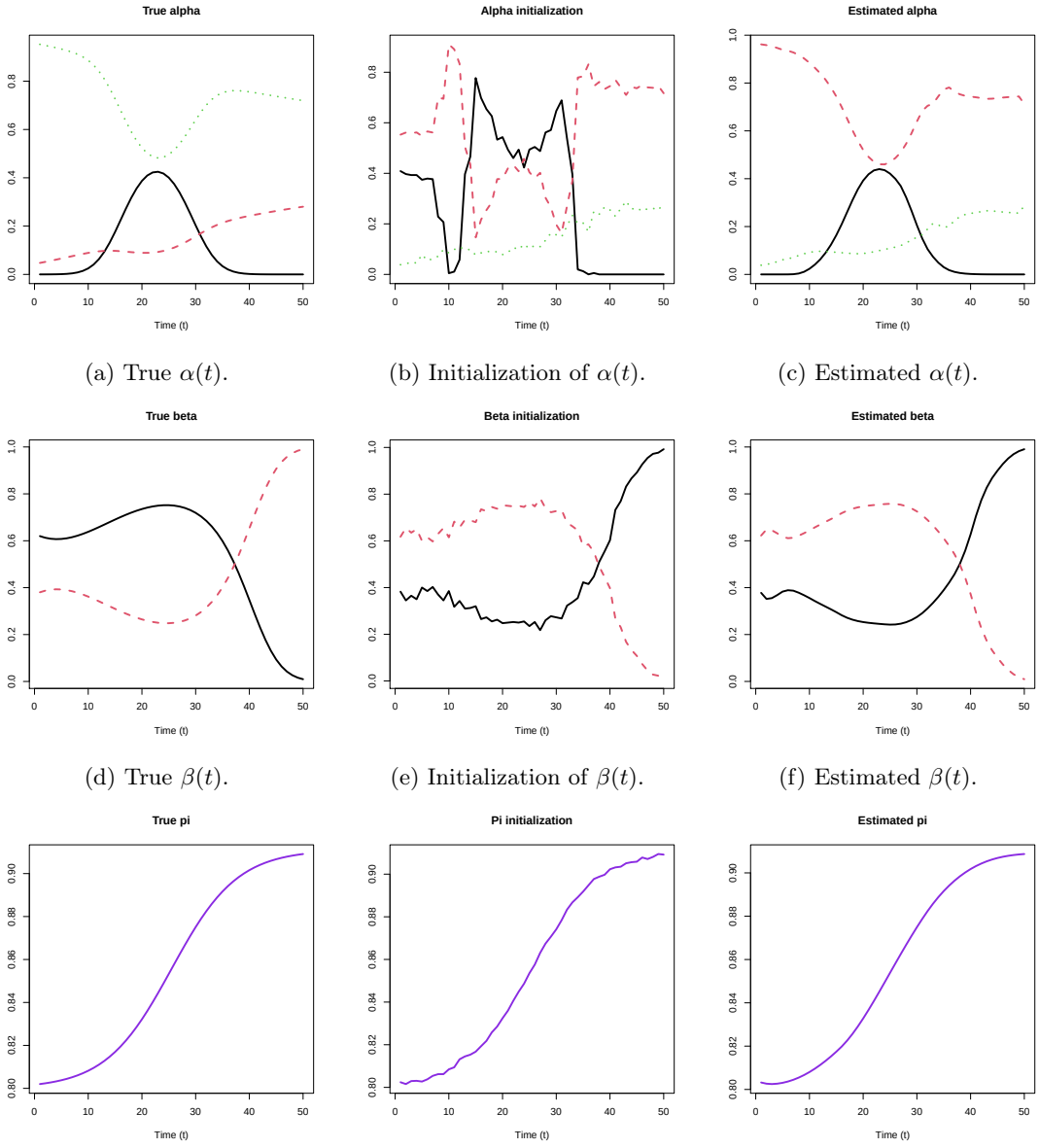
Fig. 3: Evolution of the true (left), initialized (center) and estimated (right) proportions of the parameters $\alpha(t)$, $\beta(t)$ and $\pi(t)$, respectively.



(a) True $\alpha(t)$. (b) Initialization of $\alpha(t)$. (c) Estimated $\alpha(t)$.

(d) True $\beta(t)$. (e) Initialization of $\beta(t)$. (f) Estimated $\beta(t)$.

Fig. 4: Evolution of the true (left), initialized (center) and estimated (right) proportions of the parameters $\alpha(t)$ and $\beta(t)$, respectively.



(a) True $\alpha(t)$. (b) Initialization of $\alpha(t)$. (c) Estimated $\alpha(t)$.

(d) True $\beta(t)$. (e) Initialization of $\beta(t)$. (f) Estimated $\beta(t)$.

(g) True $\pi(t)$. (h) Initialization of $\pi(t)$. (i) Estimated $\pi(t)$.

Fig. 5: Evolution of the true (left), initialized (center) and estimated (right) proportions of the parameters $\alpha(t)$, $\beta(t)$ and $\pi(t)$, respectively.

optimal number of clusters in rows and columns over a larger number of simulated datasets through the combination of the initialization procedure and the application of the $ZI_{\mathcal{P}}$-dLBM algorithm. Let us recall that, as mentioned in Section 3, the ICL criterion identifies the optimal number of clusters only at one time instant in order to initialize the parameters optimally. Subsequently, $ZI_{\mathcal{P}}$-dLBM is run with a higher number of clusters than those identified by ICL. Hence, to validate the performances on the component activation, 50 independent data sets are generated with the setup explained in Section 2.1, with $Q = 3$ row clusters and $L = 2$ column clusters, a level of sparsity varying between 80% and 90% and the other model parameters equal to:

Cluster	α	β
1	0.2 to 0.8	0.1 to 0.99
2	0.18 to 0.14	0.99 to 0.1
3	0.6 to 0.06	-

$$\Lambda = \begin{bmatrix} 6 & 4 \\ 1 & 2 \\ 7 & 3 \end{bmatrix}$$

Then, $ZI_{\mathcal{P}}$-dLBM is applied on those simulated data sets using values of Q and L equal to 10. Table 1 shows the percentage of selections. The highlighted cell corresponds to the actual value of Q and L. $ZI_{\mathcal{P}}$-dLBM succeeds 86% of the time to identify the correct model. Specifically, to evaluate the results of this experiment, we averaged the membership probability of the two estimated mixing parameters, $\alpha(t)$ and $\beta(t)$; exceeding clusters having an average membership probability of less than 1e-3 were considered to be off. Among the results of the 50 simulated data sets, we report in Figure 6, as an illustrative example, one of the component activation results. We see that not only the unnecessary clusters remained empty, but also the estimates of the $\alpha(t)$ and $\beta(t)$ are good, as $ZI_{\mathcal{P}}$-dLBM manages to identify the evolution of the two mixing parameters over time, despite the number of clusters given as input is not the optimal one.

Q/L	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	86	0	0	0	0	0	0	0	0
4	0	2	0	0	0	0	0	0	0	0
5	0	2	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	4	0	0	0	0	0	0	0	0
9	0	2	0	0	0	0	0	0	0	0
10	0	4	0	0	0	0	0	0	0	0

Table 1: Model selection. Percentage of activated components on 50 simulated data sets. The highlighted cell corresponds to the actual value of Q and L.

2.4 Benchmark study

The goal of this last experiment is to compare $ZI_{\mathcal{P}}$-dLBM with two state-of-the-art methods to recover the data structure. First, $ZI_{\mathcal{P}}$-dLBM is compared with a model based on the same assumptions but which does not take into account the sparsity modeling over time. Denoted by Zip-dLBM$_{\pi(\cdot)=0}$, the model does not take into account the excess of zeros in the data and it is obtained by setting the sparsity parameter $\pi(t)$, with t in $[0, T]$, equal to zero. The other model $ZI_{\mathcal{P}}$-dLBM is compared with is dLBM proposed by Marchello et al. (2022) where not only the sparsity is not taken into account but the cluster memberships

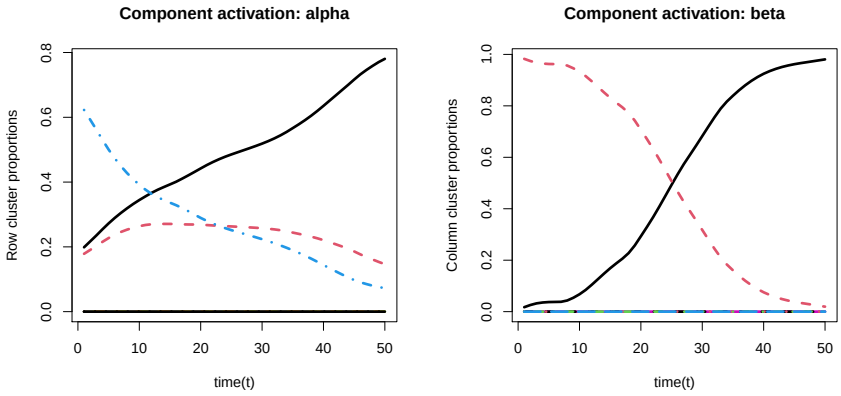


Fig. 6: View of a model selection result: the useful clusters are activated while the useless ones lie empty on the basis of the figure.

Z and W are not time-dependent, i.e. cluster switches are not allowed. However, the expected number of interactions between co-clusters (the parameter Λ) changes in time in dLBM.
We chose to evaluate the results with the CARI index. In order to compare the affectations to the clusters over time, the cluster labels in dLBM were repeated as many times as the number of time instants, and then compared to the affectations of the simulated data using the CARI index. To make this comparison more complete, we defined two simulation scenarios. In Scenario A, the data are simulated as described in Section 2.1 but with a constant sparsity level of 80%, fixed in time. In Scenario B the sparsity evolves in time from 80% to 90% . Table 2 displays the results of this comparison, in terms of average CARI values, reported with standard deviations.
In Scenario A, $ZI_{\mathcal{P}}$-dLBM performs well reaching a CARI value of 0.93, on the other hand $ZI_{\mathcal{P}}$-dLBM$_{\pi(\cdot)=0}$ suffers from the excessive number of zeros, whose treatment is not considered, probably affecting the clustering performance. Even worse for dLBM whose CARI index is 0. This is certainly due to the fact that the two latent clustering variables, Z and W, do not evolve over time.
In scenario B, $ZI_{\mathcal{P}}$-dLBM performs comparably with the previous scenario, with an average CARI index of 0.94 and a smaller standard deviation. Thus, we see that an increasing level of sparsity does not degrade the performance of the model since it is able to distinguish structural zeros from those coming from the Poisson process. This could even help in improving clustering performance. On the contrary, Zip-dLBM$_{\pi(\cdot)=0}$ performs worse than the results obtained in the scenario A probably due to the increased sparsity in the data.

	$ZI_{\mathcal{P}}$-dLBM	Zip-dLBM$_{\pi(\cdot)=0}$	dLBM
Scenario A	0.93 ± 0.13	0.27 ± 0.1	0 ± 0
Scenario B	0.94 ± 0.03	0.16 ± 0.11	0 ± 0.01

Table 2: Co-clustering results for $ZI_{\mathcal{P}}$-dLBM, Zip-dLBM$_{\pi(\cdot)=0}$ and dLBM on 50 simulated data according to the two scenarios. Average CARI values are reported with standard deviations.