The Dynamic Random Subgraph Model for the Clustering of Evolving Networks

Rawya Zreik, Pierre Latouche, Charles Bouveyron

The Dynamic Random Subgraph Model for the Clustering of Evolving Networks

Rawya ZREIK[1,2], Pierre LATOUCHE[1] and Charles BOUVEYRON[2]

[1] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne*
[2] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes & Sorbonne Paris Cité*

Abstract

In recent years, many clustering methods have been proposed to extract information from networks. The principle is to look for groups of vertices with homogenous connection profiles. Most of these techniques are suitable for static networks, that is to say, not taking into account the temporal dimension. This work is motivated by the need of analyzing evolving networks where a decomposition of the networks into subgraphs is given. Therefore, in this paper, we consider the random subgraph model (RSM) which was proposed recently to model networks through latent clusters built within known partitions. Using a state space model to characterize the cluster proportions, RSM is then extended in order to deal with dynamic networks. We call the latter the dynamic random subgraph model (dRSM). A variational expectation maximization (VEM) algorithm is proposed to perform inference. We show that the variational approximations lead to an update step which involves a new state space model from which the parameters along with the hidden states can be estimated using the standard Kalman filter and Rauch-Tung-Striebel (RTS) smoother. Simulated data sets are considered to assess the proposed methodology. Finally, dRSM along with the corresponding VEM algorithm are applied to an original maritime network built from printed Lloyd's voyage records.

Key words: Dynamic networks , subgraphs, random subgraph model, state space model, clustering, variational inference, variational expectation maximization, maritime data.

1. Introduction

Network analysis has become a mature discipline, since the original work of Moreno (1934) [1], which is no longer limited to sociology and is now applied in many areas such as biology [2, 3, 4], geography [5] or history [6]. The growing interest in network analysis is explained partly by the strong presence of this type of data in the digital world, and by recent advances in the modeling and the processing of these data. The clustering methods allow in particular clusters of vertices sharing homogeneous connection profiles to be uncovered. Most

methods look for specific structures, so called communities, which exhibit a transitivity property such that nodes of the same community are more likely to be connected [7]. A popular approach for community discovering, though asymptotically biased [8], is based on the modularity score given by Girvan et Newman [9]. Alternative methods usually rely on the latent position cluster model (LPCM) of Handcock, Raftery and Tantrum [10] which assumes that the links between the vertices depend on their positions in a social latent space.

The stochastic block model (SBM) [11, 12] is a flexible random graph model which can also characterize communities, but not only. It is based on a probabilistic generalization of the method applied by White *et al.* [13] on Sampson's famous monastery [14]. The SBM model assumes that each vertex belongs to a latent group, and that the probability of connection between a pair of vertices depends exclusively on their group. Because no assumption is made on the connection probabilities, various types of structures of vertices can be taken into account. While SBM was originally developed to analyze mainly binary networks, many extensions have been proposed since to deal for instance with valued edges [15] or to take into account prior information [16, 17]. In particular, the random subgraph model (RSM) of Jernite *et al.* [18] aims at modeling categorical edges using prior knowledge of a partition of the network into subgraphs. These known subgraphs are assumed to be made of latent clusters which have to be inferred. The vertices are then connected with a probability depending only on the subgraphs whereas the edge type is assumed to be sampled conditionally on the latent groups. This model was applied in the original paper to analyze a historical network in merovingian Gaul. Note that other extensions of SBM have focused on looking for overlapping clusters [19, 20]. The inference of SBM like models is usually done using variational expectation maximization (VEM) [21], variational Bayes EM (VBEM) [22], or Gibbs sampling [12]. Moreover, we emphasize that various strategies have been derived to estimates the number of corresponding clusters using model selection criteria [21, 22], allocation sampler [23], greedy search [24], or non parametric schemes [25].

Recently, a few attempts have been made to extend the models mentioned previously in order to deal with dynamic networks. The main idea consists in introducing temporal processes in order to characterize the temporal evolution of nodes and edges through time. Thus, Yang *et al.* [26] proposed a dynamic version of SBM allowing a node to switch its class at time $t+1$ depending on its state at time t. The switching probabilities are all characterized by a transition matrix. The alternative extension for SBM of Xu *et al.* [27] focuses on modeling the temporal changes through a state space model and relies on the Kalman filter for inference. Contrary to Yang *et al.* [26], Xu *et al.* [27] treated the edge probabilities as time varying parameters. In parallel, the mixed membership SBM (MMSBM) of Airoldi *et al.* [19], capable of characterizing overlapping clusters, was adapted to deal with dynamic networks by Xing *et al.* [28], Ho *et al.* [29] and Kim and Leskovec [30]. Moreover, Sarkar and Moore [31] derived a dynamic version of the LPCM model of Handcock, Raftery and Tantrum [10] keeping the transitivity property that nodes which are close in a social latent space should be more likely to connect. Finally, we would like to highlight the

work of Dubois, Butts and Smyth [32] and Heaukulani and Gharamani [33]. In [32] a non homogeneous Poisson process is considered. Thus, contrary to most clustering models for dynamic networks, a continuous time period is taken into account and events, *i.e.* the creation or removal of an edge, occur one at a time. While models usually focus on modeling the dynamic of networks through the evolution of their latent structures, Heaukulani and Gharamani [33] extended the dynamic latent feature model of Foulds *et al.* [34] to define how observed social interactions can affect future unobserved latent structures. In the same vein, a dynamic model inspired by SBM was proposed recently by Xu [35].

In this paper, we aim at modeling dynamic networks with binary or more generally typed edges, for which a partition of the nodes is given. As an example, we will consider an original network, built from printed Lloyd's voyage records and describing maritime flows between ports where the geographical positions of the ports play an important role. The partition was obtained by associating each port to a region according to its geographical position. Figure 1 presents the evolution of network navigations, for 23 years between October 1985 and October 2008. A (given) partition of the nodes is seen here as a decomposition of the network into known subgraphs that we propose to model using unobserved clusters that have to be inferred from the data in practice. Thus, considering a slightly different version of the original RSM model of Jernite *et al.* [18] and relying on a state space model as in [28], we propose a new random graph model for evolving networks that we call the dynamic RSM (dRSM) model. The model focuses on describing the network dynamic by characterizing the evolution of the cluster proportions within the known subgraphs. A logistic transformation is used to link the hidden states and the clusters proportions, as in [36, 37]. The inference of the model is done using a VEM algorithm.

The article is organized as follows. In Section 2, we introduce the dRSM model along with an inference procedure in Section 3. Variational techniques are considered and a model selection criterion is derived. Finally, the methodology is tested on simulated data in Section 4 and on the maritime network built from Lloyd's data in Section 5.

2. The dynamic random subgraph model

This section presents the context of the work and introduces the dRSM model along with the modeling of its dynamic. The joint distribution associated with the model is also detailed.

2.1. Context and notations

We consider a set of T networks $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$, where $\mathcal{G}^{(t)}$ is a directed graph observed at time t. Each $\mathcal{G}^{(t)}$ is represented by its $N \times N$ adjacency matrix $X^{(t)}$ where N denotes the number of nodes. The edge $X_{i,j}^{(t)}$, describing the relationship between nodes i and j, is assumed to take its values in $\{0, \dots C\}$ such that $X_{ij}^{(t)} = c$ means that nodes i and j are linked by a relationship of type c at time t and $X_{ij}^{(t)} = 0$ indicates the absence of relationship between the two

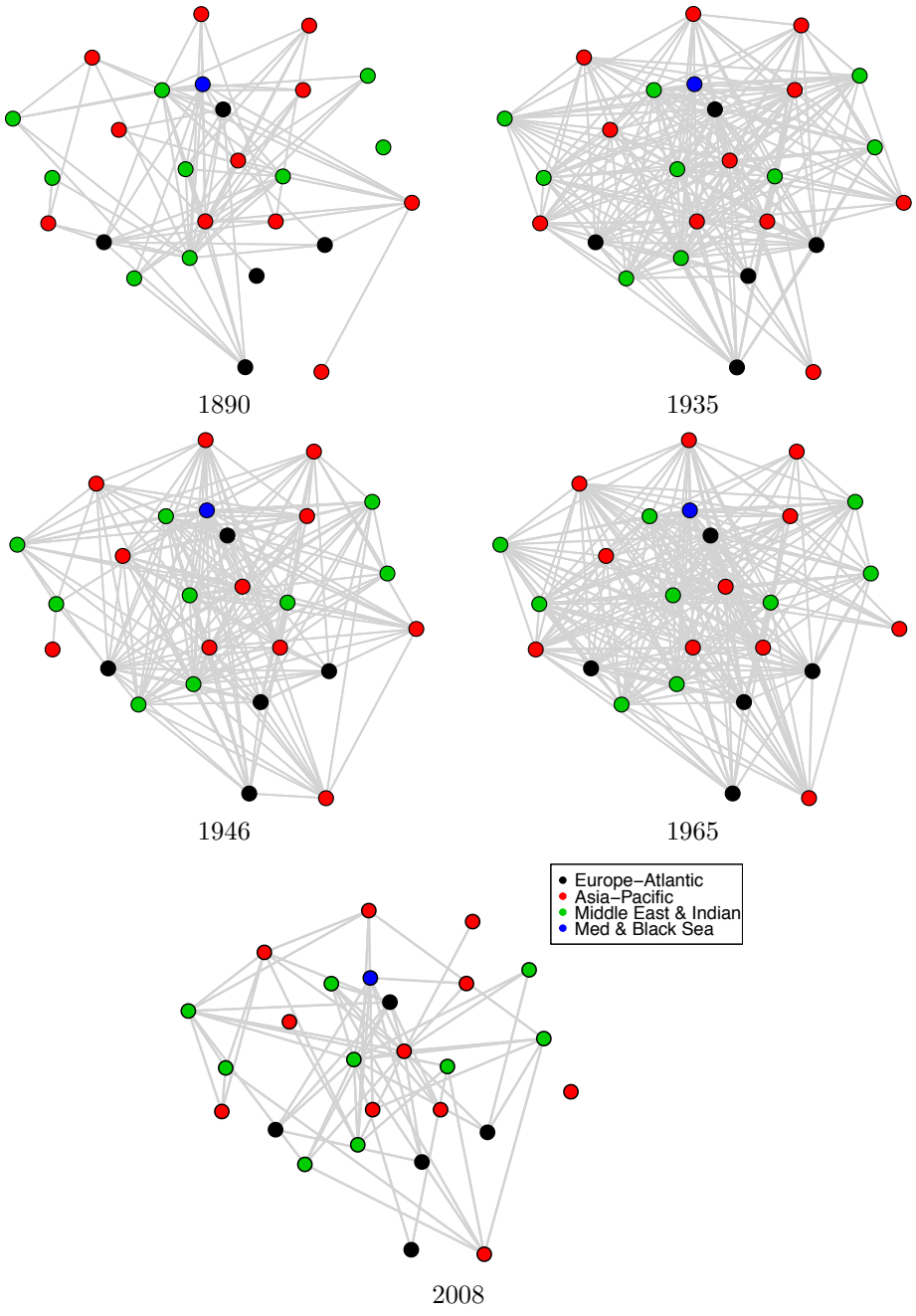

Figure 1: Connections between a subset of 26 ports (from October 1890 to October 2008). Data extracted from Lloyd's list. The known subgraphs correspond to geographical regions (continents) indicated using colors.

nodes at time t. Note that no self loops are considered, *i.e.* the connection of a node to itself, thus $X_{ii}^{(t)} = 0$, $\forall\, i, t$.

Moreover, a partition $\mathcal{P}$ of the network into S classes of vertices is assumed to be given. We emphasize that the observed partition induces a decomposition of the graph into subgraphs where each class of vertices corresponds to a specific subgraph. To describe the subgraph membership of each vertex, the variable s is introduced. The variable takes its values in $\{1, \dots S\}$ and is such that s_i indicates the subgraph of vertex i. In some cases, and in order to clarify the equations, we will also consider the indicator variables y_{is} such that $y_{is} = 1$ if node i is in subgraph s, 0 otherwise. Finally, because the vertex i can only belong to a single subgraph, we have $\sum_{s=1}^{S} y_{is} = 1$.

Our goal is to cluster at each time t the N nodes into K latent groups with homogeneous connection profiles, *i.e.* find an estimate of the set Z of latent variables $Z_{ik}^{(t)}$ such that $Z_{ik}^{(t)} = 1$ if at time t, the node i belongs to the class k, and 0 otherwise. Please note that N, C, $\mathcal{P}$, S and K are all assumed to be constant over time.

2.2. The model at each time t

As in the original RSM model, the (known) subgraphs are assumed to be built from K unobserved clusters of vertices, with varying proportions. Thus, each subgraph s has its own mixing proportion vector $\alpha_s^{(t)} = (\alpha_{s1}^{(t)}, \dots, \alpha_{sK}^{(t)})$ where $\alpha_{sk}^{(t)}$ is the proportion of cluster k in subgraph s at time t and $\sum_{k=1}^{K} \alpha_{sk}^{(t)} = 1$, $\forall\, s, t$. The network is then assumed to be generated at each time t as follows.

Each vertex i is first associated to a latent cluster k with a probability depending on its subgraph s_i. In practice, the variable $Z_i^{(t)}$ is drawn from a multinomial distribution of parameter $\alpha_{s_i}^{(t)}$:

$$Z_i^{(t)} \sim \mathcal{M}(1, \alpha_{s_i}^{(t)}),$$

and therefore $\sum_{k=1}^{K} Z_{ik}^{(t)} = 1$. Note that $Z_{ik}^{(t)} = 1$ indicates that vertex i belongs to cluster k at time t, 0 otherwise.

On the other hand, the type of link between nodes i and j is assumed to be sampled from a multinomial distribution depending on the latent vectors $Z_i^{(t)}$ and $Z_j^{(t)}$:

$$X_{ij}^{(t)} | Z_{ik}^{(t)} Z_{jl}^{(t)} = 1 \sim \mathcal{M}(1, \Pi_{kl}),$$

with $\Pi_{kl} \in [0,1]^{C+1}$ and $\sum_{c=0}^{C} \Pi_{kl}^{c} = 1, \forall k, l$.

As in the RSM model, and more generally in SBM like models, all vectors $Z_i^{(t)}$ are sampled independently, and, conditionally on these membership vectors, the edges are assumed to be independent. Thus, contrary to the original RSM model, the edges depend directly on the latent clusters exclusively, and there is no direct dependency on the subgraphs (see Figure 3). Each edge between a pair (i, j) of vertices does depend on the subgraphs s_i and s_j, but only through the fact that the edge depends on the latent clusters of the vertices, which

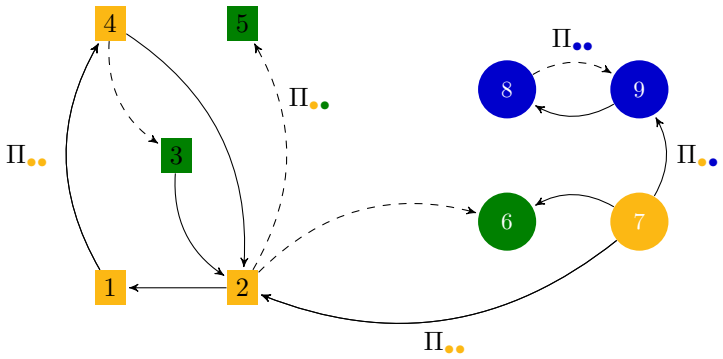

Figure 2: A dRSM network observed at time t. The network is made of 9 nodes belonging to $S = 2$ subgraphs (denoted through the form of the nodes) and split into $K = 3$ clusters (indicated by the colors). According to the dRSM model, the directed edges between the nodes can be of different types ($C = 2$ types are considered here). Given the clusters, the presence of an edge depends on the connection probabilities between clusters (Π).

themselves depend on the subgraphs. The dependency is indirect while in the original RSM model, the latent clusters along with the subgraphs are all involved in the creation of edges and have different roles. Indeed, the presence or absence of an edge between (i, j) is first drawn from a Bernoulli distribution depending on s_i and s_j. If an edge is present, the edge type is then sampled depending on the latent clusters. The separation of roles between the latent clusters and the subgraphs was originally motivated by assumptions regarding the nature of the networks analyzed. We do not make such assumptions in this paper. The latent clusters explain both the creation of an edge and its type.

Figure 2 presents an example of a dRSM network, observed at time t, made of 9 nodes belonging to 2 subgraphs (denoted through the form of nodes) and split into 3 clusters (indicated by the colors).

2.3. Modeling the evolution of random subgraphs

In order to model the evolution of the cluster proportions within the subgraphs through time, a state space model is considered as in [28]. Thus, the latent variable $\gamma_s^{(t)}$ is introduced and a logistic transformation $f(\cdot)$ is used to link the mixing vector $\alpha_s^{(t)}$ with $\gamma_s^{(t)}$:

$$\alpha_s^{(t)} = f(\gamma_s^{(t)}),$$

such that

$$\alpha_{sk}^{(t)} = f_k(\gamma_s^{(t)}) = \exp(\gamma_{sk}^{(t)} - C(\gamma_s^{(t)})), \forall s, k, t,$$

where $\gamma_{sK}^{(t)} = 0$ and $C(\gamma_s^{(t)}) = \log(\sum_{k=1}^{K} \exp(\gamma_{sk}^{(t)}))$. The choice to fix the last component of the vector $\gamma_s^{(t)}$ arbitrarily to 0 is widely used in the literature (see for instance [28, 37, 38, 39]) and is due to the bijectivity constraint of this logistic transformation which requires $\gamma_s^{(t)}$ to live in a $(K-1)$ dimensional

space since $\alpha_s^{(t)}$ has $(K-1)$ degrees of freedom. This induces that $\gamma_{sk}^{(t)} = \log(\alpha_{sk}^{(t)}/\alpha_{sK}^{(t)}), \forall s, k, t$. In addition, the $(K-1)$ first components of the vector $\gamma_s^{(t)}$ are assumed to be distributed according to a Gaussian distribution with mean $B\nu^{(t)}$ and covariance matrix Σ:

$$\gamma_{s\backslash K}^{(t)} \sim \mathcal{N}(B\nu^{(t)}, \Sigma), \tag{1}$$

where $\gamma_{s\backslash K}^{(t)}$ is the vector $\gamma_s^{(t)}$ without his last component. Both Σ and B are matrices of size $(K-1) \times (K-1)$ while $\nu^{(t)}$ is a $(K-1)$ dimensional vector. Let us notice that even though the $\gamma_s^{(t)}$ have the same mean in the state-space, they are actually independent and thus play different roles.

The rest of the model now involves a classic state space model for linear dynamic systems. It is defined as follows:

$$\begin{cases} \nu^{(t)} = A\nu^{(t-1)} + \omega \\ \nu^{(1)} = \mu_0 + u. \end{cases}$$

The noise terms ω and u are supposed to be Gaussian and independent:

$$\begin{cases} \omega \sim \mathcal{N}(0, \Phi) \\ u \sim \mathcal{N}(0, V_0). \end{cases}$$

Again, A, Φ and V_0 are matrices of size $(K-1) \times (K-1)$ while μ_0 is a $(K-1)$ dimensional vector.

Notice that the state space model for linear dynamic systems may suffer from model identifiability issues and constraints have to be introduced (see [40] for instance). In the following, we derive the inference procedure in a general context since different constraints can be considered. In practice, in all the experiments that we carried out, we fixed A, B, and V_0 to be equal to the identity matrix I_{K-1} and all components of μ_0 to zero.

The model described here has three sets of latent variables ($\nu = (\nu^{(t)})_t, \gamma = (\gamma_s^{(t)})_{st}, Z = (Z_{ik}^{(t)})_{ikt}$) and is parameterized by $\theta = (\mu_0, A, B, \Phi, V_0, \Sigma, \Pi)$. Note that all parameters in θ depend neither on time nor subgraphs. This model is called the *dynamic random subgraph model* (dRSM) in the rest of the document. Figure 3 gives the graphical model for dRSM and Table 1 summarizes the notations used in the model.

At this point, it is possible to see some links and differences between dRSM and dM3SBM [29], which is the closest model in the litterature. On the one hand, dRSM and dM3SBM share a common way to model the latent clusters and the temporal dynamic trough a state space model. On the other hand, dRSM is able to handle categorical edges, which is a useful feature when working on real-world networks, whereas dM3SBM cannot. In addition, dRSM requires the knowledge of the subgraphs whereas dM3SBM proposes to estimate them. Furthermore, dM3SBM allows the nodes to belong to different clusters. However, allowing to estimate the subgraphs and multi-group belongings may conduce dM3SB to be a too flexible model and thus to fail in recovering the network

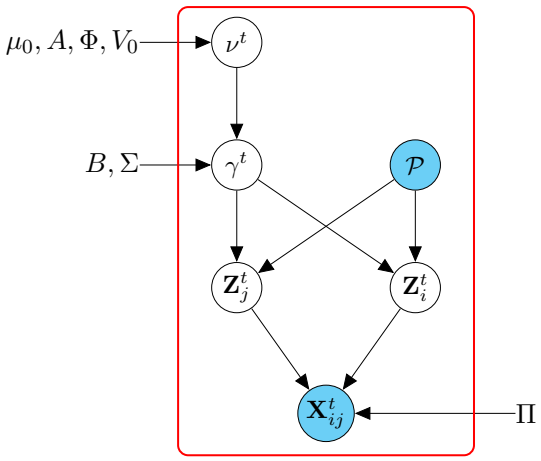

Figure 3: *Graphical representation of the dRSM model.*

structure. Indeed, providing the subgraphs to dRSM allows it to avoid looking for obvious structures such that it can focus on the search of hidden patterns. The comparisons presented in Section 4 seem to confirm this thesis.

2.4. Joint distribution of dRSM

The dRSM model proposed above is defined by the joint distribution:

$$p(X, Z, \gamma, \nu|\theta) = p(X|Z, \Pi)p(Z|\gamma)p(\gamma_{\setminus K}|B, \nu, \Sigma)p(\nu|\mu_0, A, \Phi, V_0), \tag{2}$$

where $\gamma_{\setminus K} = (\gamma^{(t)}_{s\setminus K})_{st}$. Moreover

$$p(X|Z, \Pi) = \prod_{t=1}^{T}\prod_{k,l}^{K}\prod_{c=0}^{C}(\Pi^c_{kl})^{\sum_{i\neq j}^{N}\delta(X_{ij}^{(t)}=c)Z_{ik}^{(t)}Z_{jl}^{(t)}},$$

and

$$\begin{aligned} p(Z|\gamma) &= \prod_{t=1}^{T}\prod_{i=1}^{N}\prod_{k=1}^{K} f_k(\gamma_{s_i}^{(t)})^{Z_{ik}^{(t)}} \\ &= \prod_{t=1}^{T}\prod_{k=1}^{K}\prod_{s=1}^{S} f_k(\gamma_s^{(t)})^{\sum_{i=1}^{N} y_{is}Z_{ik}^{(t)}}. \end{aligned} \tag{3}$$

Note that

$$p(\gamma_{\setminus K}|B, \nu, \Sigma) = \prod_{t=1}^{T}\prod_{s=1}^{S}\mathcal{N}(\gamma^{(t)}_{s\setminus K}; B\nu^{(t)}, \Sigma),$$

Notations	Description
X	Adjacency matrix $X_{ij}^{(t)} \in \{0, \dots, C\}$ at each t
Z	Binary matrix. $Z_{ik}^{(t)} = 1$ indicates that i belongs to cluster k at t
N	Number of vertices in the network
K	Number of latent clusters
S	Number of subgraphs
C	Number of edge types
Π	Π_{kl}^{c} is the probability of having an edge of type c between vertices of clusters k and l
α	$\alpha_{sk}^{(t)} = f_k(\gamma_s^{(t)})$ is the proportion of cluster k in the subgraph s at t

Table 1: *Summary of the notations used in the paper.*

where $\mathcal{N}(\gamma_{s\backslash K}^{(t)}; B\nu^{(t)}, \Sigma)$ denotes the multivariate Gaussian distribution, with mean vector $B\nu^{(t)}$ and covariance matrix Σ, evaluated at $\gamma_{s\backslash K}^{(t)}$. Finally

$$p(\nu|\mu_0, A, \Phi, V_0) = p(\nu^{(1)}|\mu_0, V_0) \prod_{t=2}^{T} \log p(\nu^{(t)}|\nu^{(t-1)}, A, \Phi).$$

3. Estimation

This section focuses on the inference of the model proposed above. A variational EM algorithm is considered and a model selection criterion is derived.

3.1. A variational framework

We aim at maximizing the log-likelihood $\log p(X|\theta)$ associated with the model. To achieve this maximization, a common approach consists in using an expectation maximization (EM) algorithm [41, 42]. However, such an algorithm cannot be derived here since $p(Z, \gamma, \nu|X, \theta)$ is intractable. Therefore, we propose to use a variational EM-type algorithm (VEM) [43] which locally optimizes the model parameters with respect to a lower bound of the log-likelihood. Thus, given a distribution q for the three sets of latent variables (Z, γ, ν), the log-likelihood can be written:

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + KL(q(.) \parallel p(.|X, \theta)), \tag{4}$$

where $\mathcal{L}$ is defined as follows:

$$\mathcal{L}(q, \theta) = \sum_{Z} \int_{\gamma} \int_{\nu} q(Z, \gamma, \nu) \log \frac{p(X, Z, \gamma, \nu|\theta)}{q(Z, \gamma, \nu)} d\gamma \, d\nu, \tag{5}$$

and KL denotes the Kullback-Leibler divergence between the true and approximate posterior distributions:

$$KL(q(.) \parallel p(.|X, \theta)) = -\sum_{Z} \int_{\gamma} \int_{\nu} q(Z, \gamma, \nu) \log \frac{p(Z, \gamma, \nu|X, \theta)}{q(Z, \gamma, \nu)} d\gamma \, d\nu. \tag{6}$$

Looking for the best approximation of the posterior distribution $p(Z, \gamma, \nu|X, \theta)$ in the sense of the KL divergence becomes equivalent to searching for a distribution $q(\cdot)$ that maximizes the lower bound $\mathcal{L}$ of the integrated log-likelihood. Unfortunately, because the joint distribution (2) in the lower bound involves the quantity $p(Z|\gamma)$ which depends on the normalizing constant $C(\gamma_s^{(t)})$, $\mathcal{L}$ has no analytical form and cannot be optimized with respect to $q(\cdot)$. Indeed, $C(\gamma_s^{(t)}) = \log(\sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)}))$ is based on a non linear transformation of the vector $\gamma_s^{(t)}$ which makes some expectations of the standard VEM algorithm impossible to derive.

Following the work of Blei and Lafferty [38] on correlated topic models, we propose a new bound of $\mathcal{L}(q, \theta)$ based on a variational lower bound of $p(Z|\gamma)$, as in Jordan et al. in [44].

Proposition 3.1. *(Proof in A) Given any set ξ of variational parameters $\xi_s^{(t)} \in \mathbb{R}^{*+}$, a lower bound of the first lower bound $\mathcal{L}(q, \theta)$ is given by:*

$$\log p(X|\theta) \geqslant \mathcal{L}(q, \theta) \geqslant \tilde{\mathcal{L}}(q, \theta, \xi),$$

where

$$\begin{aligned} &\tilde{\mathcal{L}}(q, \theta, \xi) \\ &= \sum_{Z} \int_{\gamma} \int_{\nu} q(Z, \gamma, \nu) \log \frac{p(X|Z, \Pi) h(Z, \gamma, \xi) p(\gamma_{\setminus K}|B, \nu, \Sigma) p(\nu|\mu_0, A, \Phi, V_0)}{q(Z, \gamma, \nu)} d\gamma d\nu \end{aligned} \tag{7}$$

with

$$\log h(Z, \gamma, \xi) = \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{s=1}^{S} y_{is} Z_{ik}^{(t)} \Big(\gamma_{sk}^{(t)} - \big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)}) - 1 + \log(\xi_s^{(t)}) \big) \Big).$$

Note that the variational parameters $\xi_s^{(t)}$ can be optimized to obtain tight bounds (see the end of Section 3.2). Moreover, we emphasize that a variational parameter $\xi_s^{(t)}$ is considered for each subgraph s and each time t for more flexibility and to improve the inference procedure. We point out that the quality of the variational approximation we propose cannot be tested analytically since $\tilde{\mathcal{L}}(q, \theta, \xi)$ and the Kullback-Leibler divergence in (6) are not tractable. Nevertheless, we rely on them for inference purposes. Note that similar approximation schemes have been used for instance by [45] and [46], in the context of model selection.

In order to maximize $\tilde{\mathcal{L}}(q, \theta, \xi)$, we further assume that $q(Z, \gamma, \nu)$ can be factorized:

$$q(Z, \gamma, \nu) = q(Z) q(\gamma) q(\nu) = \Big(\prod_{t=1}^{T} \prod_{i=1}^{N} q(Z_i^{(t)}) \Big) q(\gamma) q(\nu).$$

Finally $q(\gamma)$ is chosen within the family of Gaussian distributions of the form:

$$q(\gamma) = \prod_{t=1}^{T} \prod_{s=1}^{S} \prod_{k=1}^{K} \mathcal{N}(\gamma_{sk}^{(t)}; \hat{\gamma}_{sk}^{(t)}, \hat{\sigma}_{sk}^{2^{(t)}}),$$

to derive analytical expectations in the E step, as in [38]. Since the last component of each vector $\gamma_s^{(t)}$ has to remain equal to zero, to preserve the bijectivity constraints of the transformation $f(\cdot)$, the terms $\hat{\gamma}_{sK}^{(t)}$ and $\hat{\sigma}_{sK}^{2^{(t)}}$ are all set to zero to ensure a Dirac mass at zero. All other mean and variance terms $(\hat{\gamma}_{sk}^{(t)}, \hat{\sigma}_{sk}^{2^{(t)}}), \forall s, k \neq K, t$, are parameters to be estimated.

3.2. A VEM algorithm for the dRSM model

In this section, we first assume that the variational terms ξ, which were introduced for approximation purposes, are given. This allows the use of a VEM algorithm [44] to maximize the lower bound $\tilde{\mathcal{L}}(q, \theta, \xi)$ with respect to $q(Z, \gamma, \nu)$ and the model parameters θ. Such an optimization procedure is iterative and involves a series of successive updates. In the E step, the model parameters are fixed and the lower bound is optimized with respect to $q(Z, \gamma, \nu)$. Conversely, during the M step, the variational distribution is held fixed while $\tilde{\mathcal{L}}(q, \theta, \xi)$ is maximized with respect to θ. In standard VEM algorithms, a unique set of latent variables is usually considered. In our case, there are three sets (Z, γ, ν) of latent variables and therefore the E step itself involves iterative updates (as in [46] for instance). All distributions in $q(Z, \gamma, \nu)$ are held fixed, except one, which is optimized. This procedure is repeated for all distributions in turn.

In the following, we give the update formulae for the E and M steps. The details of the calculations along with the derivation of the lower bound are given in the appendix.

Proposition 3.2. *The VEM update step for each distribution $q(Z_i^{(t)})$ is given by:*

$$q(Z_i^{(t)}) \sim \mathcal{M}\left(Z_i^{(t)}; 1, \tau_i^{(t)} = (\tau_{i1}^{(t)}, \ldots, \tau_{iK}^{(t)})\right) \ \forall i, t,$$

where

$$\tau_{ik}^{(t)} \propto \exp\Bigg(\sum_{l=1}^{K} \sum_{c=0}^{C} \sum_{i \neq j}^{N} \delta(X_{ij}^{(t)} = c) \tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] + \sum_{s=1}^{S} y_{is} \Big(\hat{\gamma}_{sk}^{(t)} - \Big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\hat{\gamma}_{sl}^{(t)} + \frac{\hat{\sigma}_{sl}^{2^{(t)}}}{2}) - 1 + \log(\xi_s^{(t)}) \Big) \Big) \Bigg).$$

Note that $\tau_{ik}^{(t)}$ is the approximate posterior probability that node i belongs to cluster k at time t.

Proposition 3.3. *The VEM update step for the distribution $q(\nu)$ is given by:*

$$q(\nu) \propto p(\nu^{(1)} | \mu_0, V_0) \Big[\prod_{t=2}^{T} p(\nu^{(t)} | \nu^{(t-1)}, A, \Phi) \Big] \Big[\prod_{t=1}^{T} \mathcal{N}(\frac{\sum_{s=1}^{S} \hat{\gamma}_s^{(t)}}{S}; B\nu^{(t)}, \frac{\Sigma}{S}) \Big].$$

At this step, we recall that the terms $\hat{\gamma}_s^{(t)}$ are fixed and so is the variable $x^{(t)} = \sum_{s=1}^{S} \hat{\gamma}_s^{(t)}/S$. Therefore, it is remarkable to note that the functional form of $q(\nu)$ corresponds exactly to the form of the posterior distribution associated with a state space model where ν is the set of all latent state variables and $x = (x^{(t)})_t$ the set of observed outputs. Thus, each $x^{(t)}$ can be written as $x^{(t)} = B\nu^{(t)} + \tilde{v}$ where $\tilde{v} \sim \mathcal{N}(0, \Sigma/S)$ while the variables in ν are defined as previously:

$$\begin{cases} \nu^{(t)} = A\nu^{(t-1)} + \omega \\ \nu^{(1)} = \mu_0 + u, \end{cases}$$

with

$$\begin{cases} \omega \sim \mathcal{N}(0, \Phi) \\ u \sim \mathcal{N}(0, V_0). \end{cases}$$

Contrary to the original state space model introduced in Section 2, where both γ and ν were sets of unobserved variables, we obtain here a standard linear dynamic system from which the corresponding parameters, *i.e.* $\theta' = (\mu_0, A, B, \Phi, V_0, \Sigma/S)$ can be estimated using Kalman filter and Rauch-Tung-Striebel (RTS) smoother [47] (details can also be found in [48]). The expectations $\hat{\nu}^{(t)}$ and covariance matrices $\hat{V}^{(t)}$ of the random variables $\nu^{(t)}$, given all the observed data x, are determined relying on backward forward recursions.

Proposition 3.4. *After the E step of the VEM algorithm, the lower bound* $\tilde{\mathcal{L}}(q, \theta, \xi)$ *simplifies into:*

$$\begin{aligned}
\tilde{\mathcal{L}}(q,\theta,\xi) \quad = \quad & \sum_{t=1}^{T}\sum_{k,l}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)} = c)\tau_{ik}^{(t)}\tau_{jl}^{(t)} \log(\Pi_{kl}^{c}) \\
+ \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(r_s^{(t)}\hat{\gamma}_{sk}^{(t)} - N_s\xi_s^{-1(t)}\sum_{l=1}^{K}\exp(\hat{\gamma}_{sl}^{(t)} + \frac{\hat{\sigma}_{sl}^{2^{(t)}}}{2}) + N_s - N_s\log(\xi_s^{(t)})\Big) \\
+ \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(\log\mathcal{N}(\hat{\gamma}_s^{(t)}, B\hat{\nu}_s^{(t)}, \Sigma) - \frac{1}{2}tr(\Sigma^{-1}B^{\intercal}\hat{V}^{(t)}B) - \frac{1}{2}tr(\Sigma^{-1}\hat{\sigma}_s^{(t)^2})\Big) \\
- \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{k=1}^{K-1} -\log\Big((2\pi)^{\frac{1}{2}}\hat{\sigma}_{sk}^{(t)}\Big) + \frac{TKS}{2} \\
- \quad & \sum_{t=1}^{T}\Big(\log\mathcal{N}(x^{(t)}; B\hat{\nu}^{(t)}, \frac{\Sigma}{S}) + \frac{1}{2}tr(\Sigma^{-1}SB^{\intercal}\hat{V}^{(t)}B)\Big) \\
- \quad & \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log(\tau_{ik}^{(t)}) \\
+ \quad & \log p(x|\theta')
\end{aligned}$$

where $r_s^{(t)} = \sum_{i=1}^{N}\tau_{ik}^{(t)}y_{is}$, N_s *is a number of nodes in the subgraph* s, *and* $\log p(x|\theta')$ *is the log likelihood of the linear dynamic system associated with the variational distribution* $q(\nu)$ *(see Proposition 3.3)*

The maximization of this bound allows to obtain the updating formula for the tensor matrix Π:

$$\hat{\Pi}^c_{kl} = \frac{\sum_{t=1}^{T}\sum_{i\neq j}^{N}\delta(X_{ij}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}}{\sum_{t=1}^{T}\sum_{c=0}^{C}\sum_{i\neq j}^{N}\delta(X_{ij}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}}, \forall k,l,c.$$

For the parameters $\hat{\gamma}_{sk}^{(t)}$ and $\hat{\sigma}_{sk}^{2^{(t)}}$, we do not obtain analytical expressions, and therefore we rely on a quasi-Newton algorithm for the optimization task.

3.3. Optimization of ξ

So far, we have seen that a VEM algorithm could be implemented from approximations depending on the variational parameters $\xi_s^{(t)}$. However, we have not addressed yet how these parameters could be estimated from the data. We follow the work of Bishop and Svensén [49] on Bayesian hierarchical mixture of experts. Thus, the lower bound $\tilde{\mathcal{L}}(q,\theta,\xi)$ is optimized with respect to the variational terms $\xi_s^{(t)}$ to obtain the tightest bound $\tilde{\mathcal{L}}(q,\theta,\xi)$ of $\mathcal{L}(q,\theta)$. This leads to new estimates $\hat{\xi}_s^{(t)}$ of $\xi_s^{(t)}$:

$$\hat{\xi}_s^{(t)} = \sum_{l=1}^{K}\exp(\hat{\gamma}_{sl}^{(t)} + \hat{\sigma}_{sl}^{2^{(t)}}), \forall s,t.$$

This procedure gives rise to a three step optimization scheme. Given all $\xi = (\xi_s^{(t)})_{st}$, the VEM algorithm described previously is used to maximize the lower bound with respect to $q(Z,\gamma,\nu)$ and θ. These terms are then held fixed and a new estimate of ξ is computed. The three steps are repeated until convergence of the lower bound.

3.4. Model selection: choice of the number K of latent groups

Using the VEM algorithm proposed in the previous paragraphs, the estimation of the model parameters and of the group memberships is fully automatic for a given value of K. Since we consider here a model-based approach, two dRSM models with different values of K can be considered as two different models. The problem of choosing K can therefore be viewed as a model selection problem. It can be tackled in a model-based context using model selection criteria, such as the Akaike information criterion (AIC) [50] or the Bayesian information criterion (BIC) [51]. Due to its popularity and its asymptotic properties [52], we use BIC in the numerical experiments presented in the following sections. BIC relies on an asymptotic approximation of the marginal log-likelihood, also called integrated log-likelihood, and is defined in the specific context of the dRSM model $\mathcal{M}$ by:

$$BIC(\mathcal{M}) = \log p(X|\hat{\theta}) - \frac{\eta(\mathcal{M})}{2}\log(TN^2).$$

where $\eta(\mathcal{M}) = CK^2 + (K-1)(K-2) + K - 1$ is the number of free model parameters depending on K, for the identifiability constraints we consider in

Parameters	Scenario 0	Scenario 1	Scenario 2	Scenario 3	Scenario 4
N	300				
K	4				
T	10 (indep.)	10 (SSM)			
S	1	1	1	2	2
C	1	1	1	1	2
$(\Pi^0_{ll})_{l=1,..,K}$	(0.1,0.4,0.6,0.5)				
$\Pi^0_{kl,k\neq l}$	0.99		0.8	0.99	
$\Pi^{c\neq 0}_{kl}$	$(1-\Pi^0_{kl})/C$				

Table 2: Parameter values for the five types of graphs used in the experiments. In scenario 0, the networks are drawn without an explicit temporal dependance whereas, in the other scenarios, the temporal dependance is generated through a state space model (SSM).

this paper. Unfortunately, the log-likelihood $\log p(X|\hat{\theta}) = \log\left(\sum_Z p(X,Z|\hat{\theta}\right)$ is not tractable here because it involves marginalizing over all latent vectors $Z_i^{(t)}$ in Z. Therefore, we propose to replace the log-likelihood with its variational approximation $\tilde{\mathcal{L}}(q,\theta,\xi)$. Thus, the VEM algorithm is run for various values of K. For each K, the algorithm iterates until convergence of the lower bound. $\hat{K}$ is then chosen such that the (approximate) BIC criterion is maximized.

4. Numerical experiments and comparisons

This section aims at proving on synthetic data the validity of the inference algorithm presented in section 3. An introductory example is first considered to highlight the main features of the proposed approach. Model selection is then considered to validate the criterion choice. Extensive comparisons with state-of-the-art methods conclude this section.

4.1. Experimental setup

In order to validate our approach, we use in this section artificial data generated according to a common experimental setup. To simplify the characterization and facilitate the reproducibility of the experiments, we designed five different scenarios. The generation setup for each scenario is summarized in Table **??**. Data from scenario 0 are drawn using SBM at each time t and without an explicit temporal dependence. The data sets for all other scenarios (scenarios 1 to 4) are drawn according to the dRSM model. Therefore, the temporal dependence is generated through a state space model. All generated networks are made of $N = 300$ nodes, distributed into $K = 4$ latent groups and have $T = 10$ time points. Depending on the scenario, the networks have $S = 1$ or 2 subgraphs, with binary ($C = 1$) or categorical ($C = 2$) edges. When $S > 1$, the nodes are randomly assigned uniformly to the subgraphs. Notice that scenario 2 has a parameter $\Pi^0_{kl,k\neq l}$ equal to 0.8 which leads to less heterogeneous latent groups.

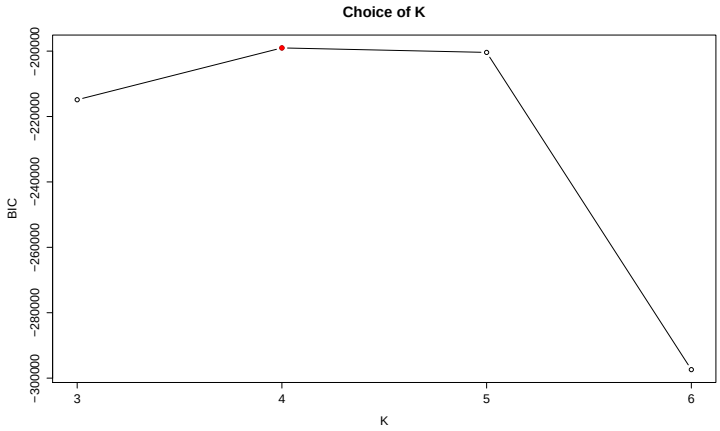

Figure 4: Choice of K by model selection with BIC for a simulated network. The actual value for K is 4.

The model parameters used for the simulation are as follows. For the simulation of γ, it is assumed that the matrices A, B and V_0 are set to I_{K-1}, and that $\Sigma = 0.1 \times K \times I_{K-1}$ and $\Phi = 0.01 \times I_{K-1}$. Finally, the tensor matrix Π, which defines the connection probabilities between clusters for the C different types, is set up such that, within the clusters, the probability $1 - \Pi^0_{ll}$ of having an edge of any type is larger than the corresponding connection probabilities between clusters $1 - \Pi^0_{kl,k\neq l}$ (see Table **??**). Notice that such a choice of parameters induces networks made of communities. Then, in case of a connection between two nodes, the edge type is sampled uniformly, *i.e.* $\Pi^{c\neq 0}_{kl} = (1 - \Pi^0_{kl})/C, \ \forall k, l$.

4.2. An introductory example

We first focus on an introductory example to illustrate the global behavior of the proposed methodology. To this end, we simulated a single network according to scenario 2 for facilitating the understanding of the results. We remind that in this setup the number K of latent groups is fixed to 4 and that $C = 1$. Therefore, the network is binary and Π^1_{kl} indicates the occurrence probability of an edge. We ran the VEM algorithm on it for a number K of groups ranging from 3 to 6. We selected afterward the most appropriate number of groups using the BIC criterion.

Figure 4 shows the BIC values associated to the results provided by our VEM algorithm for the different values of K. One can observe that the criterion picks at $K = 4$, which is the actual simulated value for K. Figure 5 presents the evolution of the bound $\tilde{\mathcal{L}}$ for this specific value of K along the 10 iterations of the VEM algorithm. A clear plateau of the bound is visible on the figure, which indicates the convergence of the algorithm.

To quickly assess the estimation quality, Table 2 allows to compare the actual (left panel) and estimated (right panel) values of the terms Π^1_{kl} in the tensor

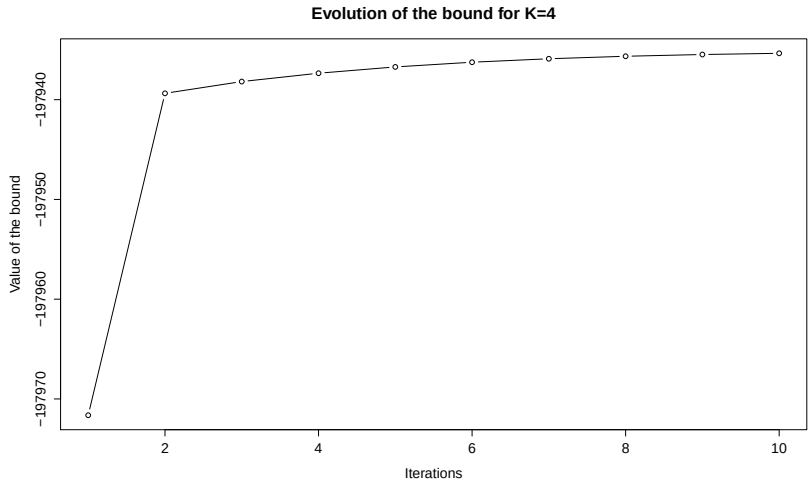

Figure 5: Evolution of the bound $\tilde{\mathcal{L}}$ for $K = 4$.

matrix Π, which define the connection probabilities between the latent clusters. On this single example, the estimated values Π^1_{kl} turn out to be extremely close to the true ones. Similarly, Figure 6 compares the actual (dashed red lines) and estimated (solid black lines) values of the group proportions α for the simulated example. Once again, the estimation of α appears to be very close to the true proportions.

4.3. Choice of K

We now focus on the evaluation of the criterion we proposed to select the number K of latent groups. Since our approach aims at searching the unobserved clustering partition of the nodes, we chose here to evaluate the combination of our VEM algorithm with the BIC criterion by comparing the resulting partition with the actual one (the simulated partition). In the clustering community, the adjusted Rand index (ARI) [53] serves as a widely accepted criterion for the difficult task of clustering evaluation. The ARI looks at all pairs of nodes and check wether they are classified in the same group or not in both partitions. As a result, an ARI value close to 1 means that the partitions are similar and, in our case, that the VEM algorithm succeeds in recovering the simulated partition.

To validate the combination of our VEM algorithm with the BIC criterion, the analysis was repeated for 50 different data sets, generated according to scenario 2, for a number K of latent groups ranging from 3 to 6. This allows us to both verify the consistency of the BIC criterion and to study the clustering ability of our approach. Figure 7 shows the repartition of the criterion values (left panel) as well as the associated ARI values (right panel). These results first confirm that BIC is a valid criterion for selecting the number of groups in

Cluster	1	2	3	4
1	0.90	0.01	0.01	0.01
2	0.01	0.60	0.01	0.01
3	0.01	0.01	0.40	0.01
4	0.01	0.01	0.01	0.50

Actual values

Cluster	1	2	3	4
1	0.89	0.01	0.01	0.01
2	0.01	0.59	0.01	0.01
3	0.01	0.01	0.39	0.01
4	0.01	0.01	0.01	0.48

Estimated values

Table 3: Actual (left) and estimated (right) values for the terms Π^1_{kl} of the tensor matrix Π. See text for details.

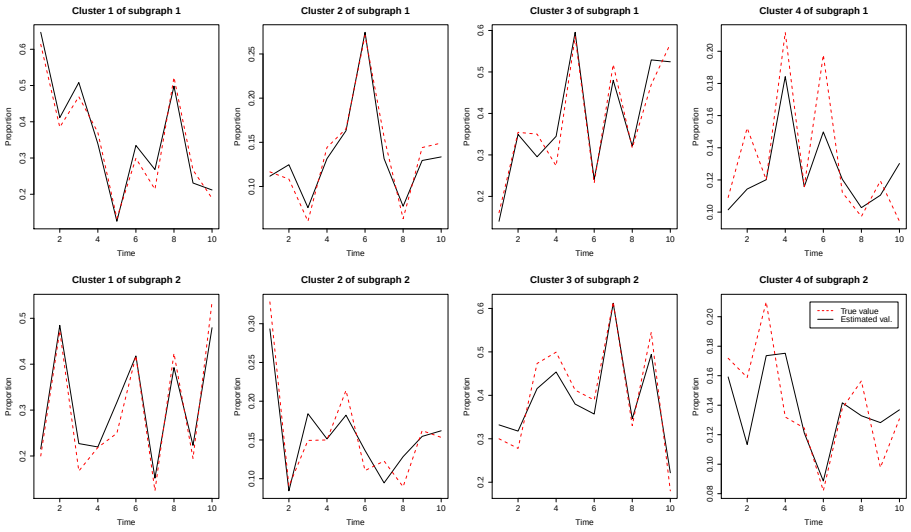

Figure 6: Actual (dashed red lines) and estimated (solid black lines) values of the group proportions for the simulated example ($K = 4$ groups and $S = 2$ subgraphs).

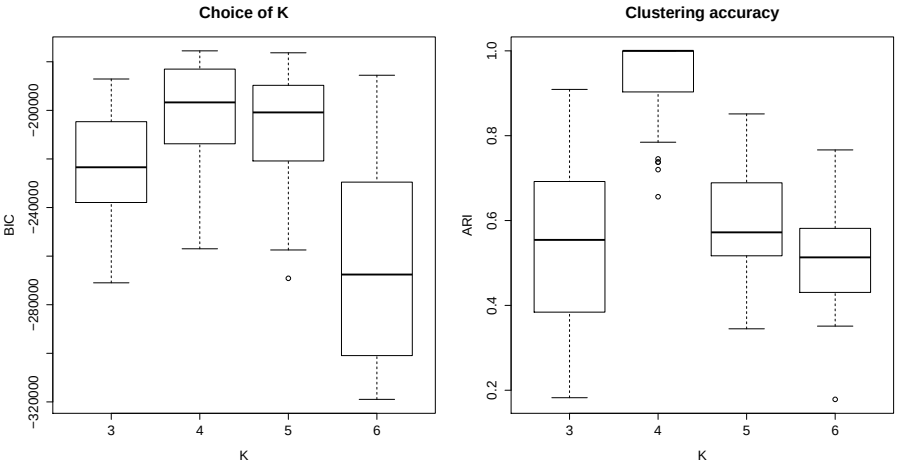

Figure 7: *Criterion and ARI values over 50 networks generated.*

this context. Indeed, the value $K = 4$ is the one which is the most frequently associated with the highest value of BIC. We remind that $K = 4$ is the actual number of latent groups. One can also observe that the partition resulting from our VEM algorithm is associated, for this value of K, to an ARI value extremely close to 1 which denotes a good matching with the actual partition of the data.

4.4. Comparison with the other stochastic models

Our third set of experiments now aims at comparing the performance of our approach to that of state-of-the-art methods. We are here interested in the comparison of dRSM with the following methods: SBM [12], RSM [18] and dM3SBM [29]. Once again, the evaluation of the results is done using the ARI criterion. In order to fit a SBM on a dynamic network, we ran the `mixer` package [54] for the R software at each time t and the ARI is then computed on the concatenation of all group labels. However, let us notice that SBM was not able to handle networks with categorical edges (scenario 3). For RSM, we used the `Rambo` package [55] for R, on an aggregated version of the whole network. Conversely to SBM, RSM is only able to deal with categorical networks and, consequently, it works only in scenario 4. Finally, we used the Matlab toolbox `dM3SBM`, kindly provided by the authors, to fit the dM3SBM on the dynamic networks. However, dM3SBM is also not able to handle networks with categorical edges (scenario 4).

In order to consider a wide type of networks, we compare here the methods over the five simulation scenarios. We remind that Table **??** summarizes the main features of each scenario. This comparison has been conducted in two different situations: with and without the knowledge of the actual number of clusters. Table **??** presents the clustering results for the four studied methods in the case where the actual number K = 4 of groups has been provided to each

Method	Scenario 0	Scenario 1	Scenario 2	Scenario 3	Scenario 4
SBM	0.10±0.04	0.12±0.05	0.18±0.07	0.14±0.09	–
RSM	–	–	–	–	0.01±0.01
dM3SBM	0.36±0.09	0.30±0.16	0.25±0.16	0.32±0.20	–
dRSM	1.00±0.00	0.98±0.04	0.90±0.20	0.97±0.07	0.75±0.24

Table 4: Clustering results for the four studied methods on networks simulated according to the five scenarios. The actual number $K = 4$ of groups has been provided to each method here. Average ARI values are reported (with standard deviations) and results are averaged on 20 networks for each scenario.

method. Conversly, Table **??** presents the clustering results when the methods have to look for the value of K. Reported values are averaged ARI values (with standard deviations) on 20 networks for each scenario. The average selected number K of latent groups is also provided for Table **??**.

First, for scenarios 0, 1 and 2, which consider dynamic networks with binary edges ($C = 1$) and with only one subgraph ($S = 1$), one can see on Tables **??** and **??** that SBM is, as expected, not able to handle the network dynamic. Indeed, SBM obtains a low ARI value in all situations, even though it correctly estimates the number of clusters (Table **??**). Conversely, the two dynamic methods (dM3SBM and dRSM) turn out to be able to recover the dynamic of the network. One can however notice that dRSM significantly outperforms dM3SBM in this situation. Notice also the accurate estimation of the number K of clusters made by dRSM (Table **??**).

In scenario 3, the simulated dynamic networks are now made of two subgraphs ($S = 2$), still with binary edges ($C = 1$). Naturally, SBM does not perform well in this situation too. The dM3SBM provides clustering results similar to the ones of previous scenarios: it globally succeeds in recovering the dynamic but fails in recognizing the clustering pattern. On the other hand, dRSM provides again accurate clustering results associated with good estimations of K, meaning that it succeeds in identifying both the dynamic and clustering patterns.

Finally, scenario 4 considers the case of dynamic networks with two subgraphs ($S = 2$) and categorical edges ($C = 2$). Only RSM and dRSM are able to deal with this kind of networks. Similarly to SBM in previous scenarios, RSM does not succeed in recovering the dynamic and provides very unsatisfactory clustering results. Conversely, dRSM gives very good clustering results regarding the difficulty of the situation. It is worth noticing the sharp estimation made by dRSM of the number K of group in this case too. This confirms the efficiency of both our inference algorithm and our model selection criterion.

5. Maritime network

This section presents an application of the proposed methodology for the analysis of a network of maritime flows in which a temporal dynamic is present. The dynamic network was provided by Dr. César Ducruet, from the Géographie-Cités laboratory, who is interested in studying the evolution of maritime flows

	Scenario 0		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	ARI	K	ARI	K	ARI	K	ARI	K	ARI	K
SBM	0.01±0.04	4.00±0.00	0.18±0.13	3.94±0.71	0.21±0.11	3.97±0.46	0.13±0.05	4.16±0.79	–	–
RSM	–	–	–	–	–	–	–	–	0.01±0.01	2.00±0.00
dM3SBM	0.01±0.01	5.55±1.39	0.35±0.21	5.95±1.15	0.30±0.21	4.35±1.63	0.32±0.19	5.15±1.17	–	–
dRSM	1.00±0.00	4.00±0.00	0.87±0.17	4.01±0.65	0.89±0.21	4.10±0.30	0.85±0.22	4.10±0.45	0.68±0.30	4.05±0.51

Table 5: Clustering results for the four studied methods on networks simulated according to the five scenarios. Average ARI values are reported (with standard deviations) as well as the selected number K of latent groups. Results are averaged on 20 networks for each scenario.

Time point	Date
t_1	October 1890
$t_2 \dots t_4$	October 1925 to October 1940, every five years
t_5	October 1946
t_6	October 1951
t_7	October 1960
$t_8 \dots t_{16}$	October 1965 to October 2000, every five years
t_{17}	October 2008

Table 6: The time points considered in the maritime network.

over time. The data was extracted from the well-known Lloyd's list which has recorded almost all ship movements worldwide since 1890.

5.1. Data and study protocol

Data was obtained from the printed Lloyds voyage record published every October between 1890 until 2008. The list details, for each merchant vessel, its successive movements from one port to another. From the raw database of vessel flows, we extracted a dynamic network with 17 time points. The first observation is October 1890 and the network ends in October 2008. Table 3 provides the correspondence between the 17 time points and the actual dates.

At each time point, the adjacency matrix between ports was constructed as follows. First, for every pair of ports, we calculated the total number of ship movements between those ports. Then, we set the associated entry in the adjacency matrix to 1 if the number of ship movements between the two ports is greater or equal to 1, and to 0 otherwise. The original network contained 4472 ports worldwide. We however had to reduce the network size to only 286 ports since most of the ports were not active throughout the whole period of the study.

We finally applied dRSM to a maritime network which describes the navigation of ships among 286 ports in the world at 17 time points. Let us highlight that the study period includes many major historical or economical events (the two world wars, the oil crisis, the economic crisis in Europe, ...), which could directly affect the navigation movements at a global scale and could also change the port behaviors.

The partition of the network into subgraphs is here provided by the port memberships to the four main maritime basins: Asia – Pacific, Europe – Atlantic, Mediterranean – Black Seas, and Middle East – Indian Ocean. Figure 8 presents this partition of the ports where the colors indicates the different subgraphs.

To summarize, the network is a undirected and binary network without self loops, *i.e.* $C = 1$ and $X_{ij}^t = 1$ if the port i and the port j exchange at least one ship during the period t, 0 otherwise, with $t \in \{1, \dots, 17\}$ and $S = 4$. Figure 9 shows the adjacency matrix, in 1890 and in 2008, between the 286 ports organized by subgraph.

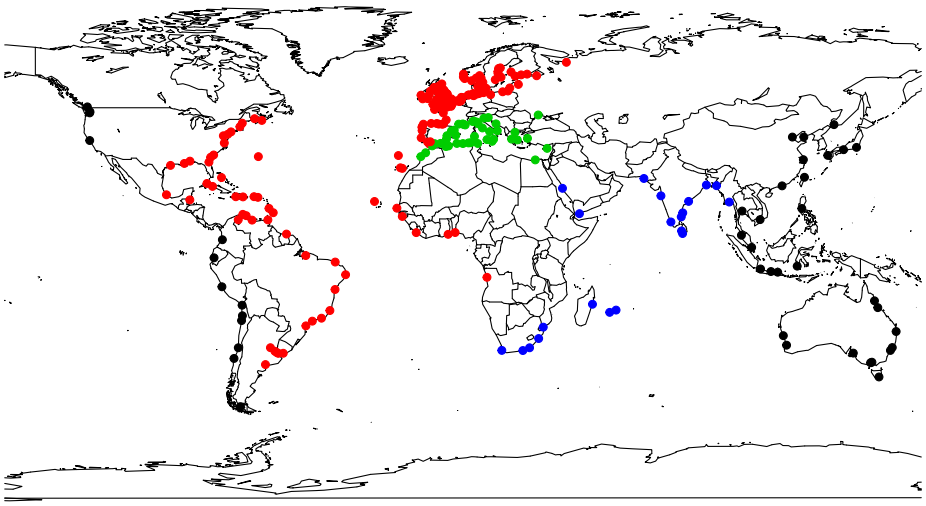

Figure 8: The given partition of the 286 nodes (ports) into 4 subgraphs.

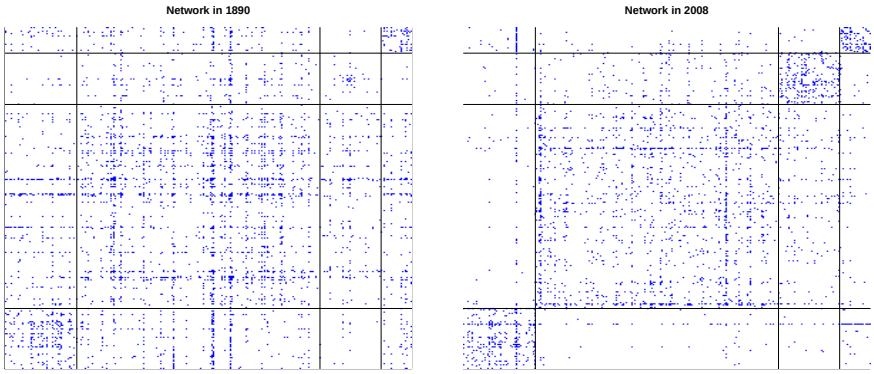

Figure 9: Adjacency matrix of the maritime network organized by subgraph (basin) in 1890 (left) and 2008 (right).

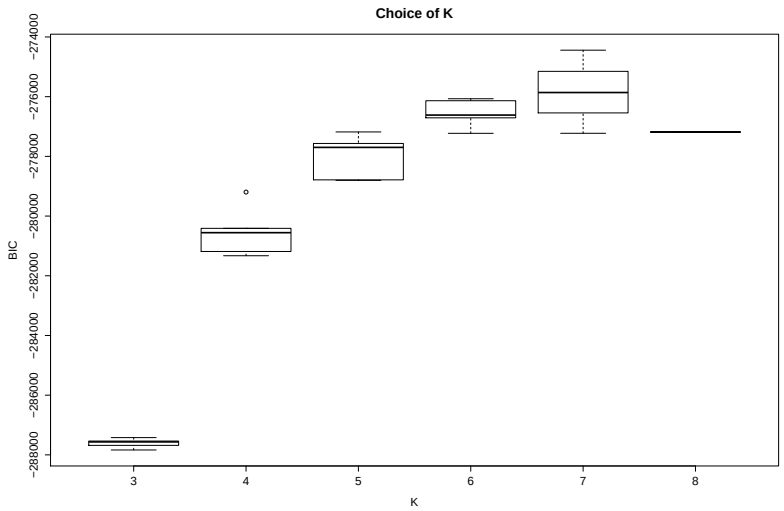

Figure 10: BIC values according to the number K of groups for the maritime network.

5.2. Results

We used the variational EM algorithm introduced in the previous section in order to find the latent groups that may be hidden in the data. The choice of the number of groups is made by applying the VEM algorithm for $K = 3, ..., 8$ and by then computing the associated BIC values. The retained value for K is the one associated with the highest BIC value. To ensure a good accuracy of the results, the VEM algorithm was run 5 times for each value of K. Figure 10 shows the evolution of the BIC criterion according to K. One can observe that BIC peaks at $K = 7$, meaning that 7 latent groups seem to organize the network. We therefore chose this specific value for K and retained the best run for $K = 7$ over the five runs as the final clustering result.

On the one hand, it is of main interest to look at the estimated tensor matrix Π in order to understand and characterize the found latent groups. Indeed, the tensor matrix Π describes the connection probabilities between the groups and allows to figure out the different connection patterns. Since the network considered here is binary, it is enough to look at the terms Π^1_{kl} since $\Pi^0_{kl} + \Pi^1_{kl} = 1$, for all k, l. Figure 11 presents those estimated values. From the figure, clusters 6 and 7 appear to be groups of hubs for which the connection probabilities are large within and between clusters.

On the other hand, the estimated group proportions over time should allow to understand the dynamic of the network. Figure 12 presents the evolution of those proportions over time for each subgraph. One can first observe that the proportion of cluster 6 is low and rather stable over time. This confirms that cluster 6 is a group of a limited number of hubs with a high connectivity and probably a high level of traffic. Cluster 6 includes ports such as Anvers,

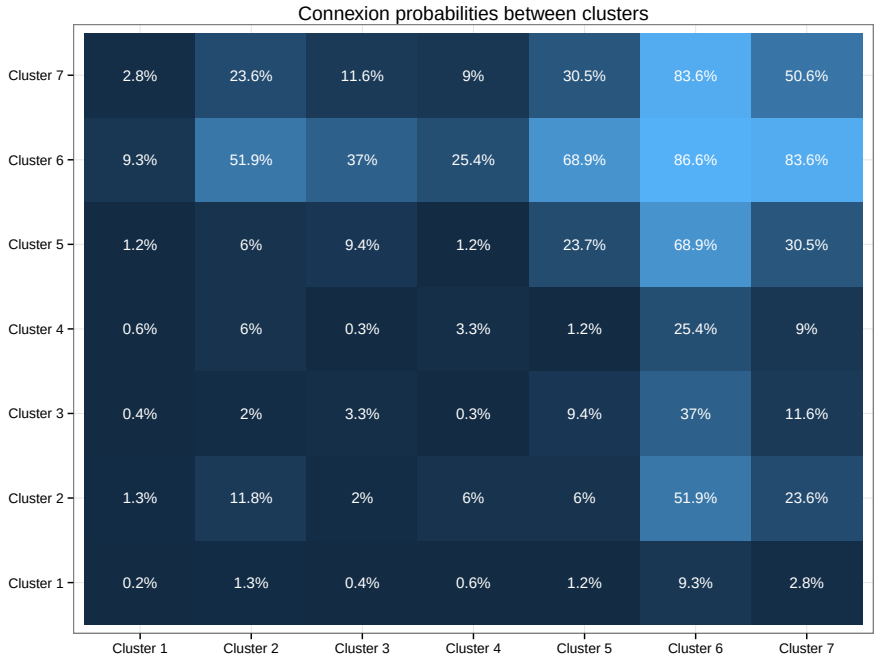

Figure 11: Terms Π^1_{kl} of the tensor matrix Π estimated using the VEM algorithm.

Rotterdam or Singapore. It is also interesting to see that, in subgraph 2 (Europe – Atlantic), the number of hubs increased until 1930, was then perturbed during the second world war and finally decreased from 1951. Conversely, in subgraph 1 (Asia – Pacific), the proportion of hubs was low until 1975 and then significantly increased. From a global point of view, one can also observe a clear and recent reorganization of the network in which hubs tend to be less numerous worldwide (and probably bigger).

Regarding cluster 7, one can see on Figure 12 that its proportions in the subgraphs are higher than those of cluster 6. The ports of cluster 7 can be qualified as hubs of second class which are subordinated to the main hubs of cluster 6. Most of them are marked by a colonial logic, such as Marseille, Calcutta or Cape Town. The evolution of this cluster until the recent period shows a persisting link North-South (*e.g.* Le Havre - Casablanca) or East-West (*e.g.* Spain - Brazil - Canaries).

The cluster 5 is mainly made of ports from the Asia – Pacific and Middle East – India basins except during major crises, such as World War II and the oil crisis. During those crises, the cluster mainly contains European ports. The rapid modification of this cluster appears clearly on Figure 12 around 1946, 1980 and 2008. This cluster can be interpreted as made of active ports from the developing world which move to cluster 2 during the crises. This may highlight the disintegration of long distance links during such crises. Conversely, cluster 2 turns out to be mostly made, except during crises, of European ports of average size, mainly on the atlantic coast. Those ports are rather a reflection of a past glory and most of them have declined over the century. This may due to a failed industrialization or a significant distance to the major trade routes.

Finally, clusters 3 and 4 are made of very small ports with low activity. Those ports are usually not connected together and communicate with the rest of the network only through ports of clusters 2 and 5. The connection with clusters 2 and 5 explains the brutal changes in the proportions of clusters 3 and 4 that one can also observe.

6. Conclusion

This work has considered the problem of analyzing dynamic networks with categorical edges and for which a subgraph partition is known. This kind of networks is frequent in a wide range of scientific fields, such as Geography in particular. For this purpose, we proposed an extension of the RSM model to the dynamic setting. The new model, called dRSM, uses a state space model to model the evolution of the latent group proportions over time. A variational expectation maximization (VEM) algorithm is proposed to perform inference. We have shown in particular that the variational approximations lead to a new state space model from which the parameters can be estimated using the standard Kalman filter and the Rauch-Tung-Striebel (RTS) smoother. Model selection is also considered through an approximate BIC criterion.

Numerical experiments have highlighted the main features of the dRSM model and have demonstrated the efficiency of both the VEM algorithm and the

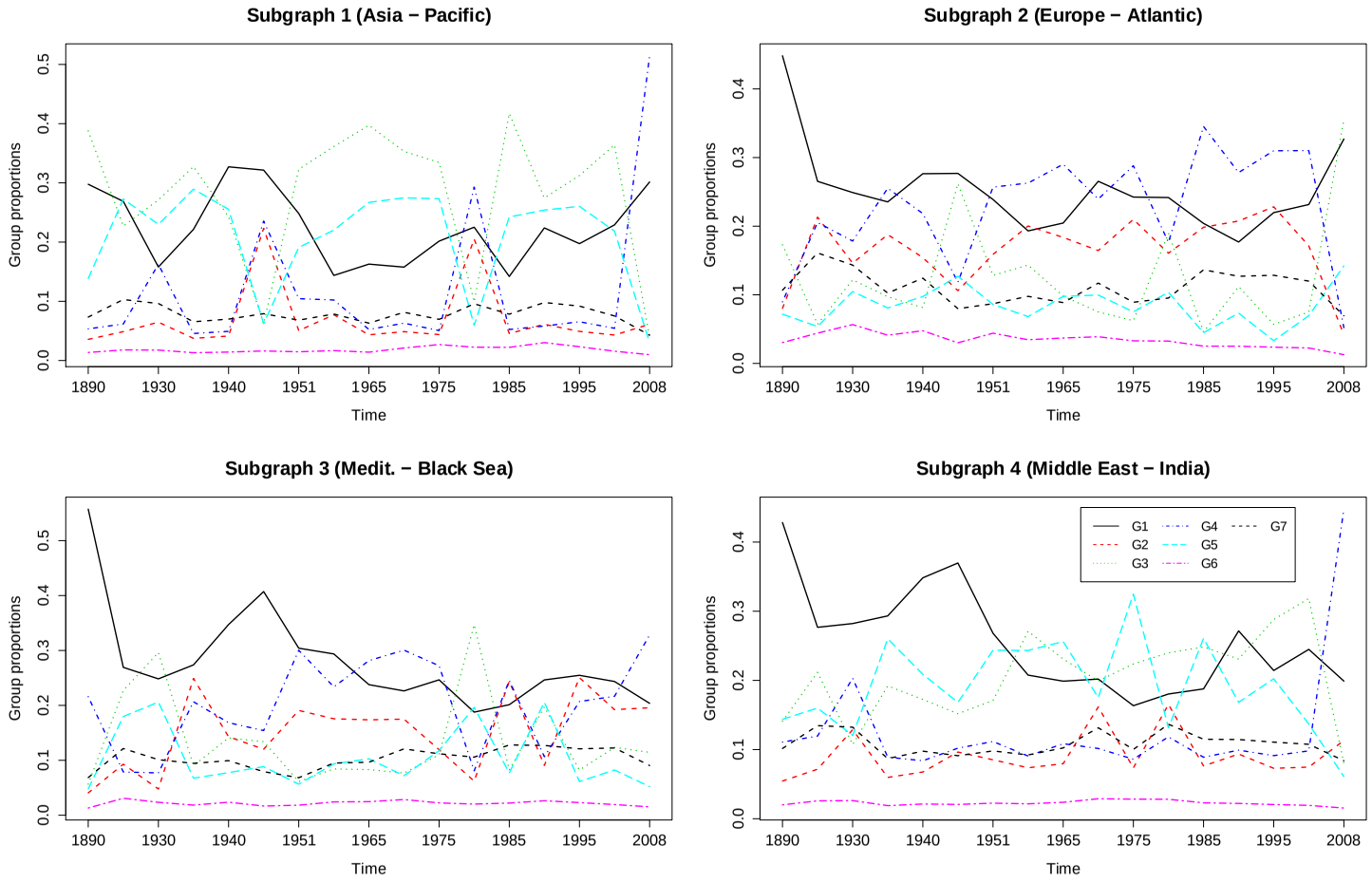

Figure 12: Evolution of the proportions of the $K = 7$ latent clusters.

model selection criterion. A numerical comparison has also shown that existing methods, dynamic or not, are less flexible and efficient than dRSM when applied to dynamic networks. Finally, dRSM has been applied to a dynamic maritime flow network, build from the famous Lloyd's list, and has allowed to characterize interesting dynamic phenomena.

Acknowledgments

The authors would like to greatly thank César Ducruet, from the Géographie-Cités laboratory, Paris, France, for providing the maritime network and for his painstaking analysis of the results. The data were collected in the context of the ERC Grant N.313847 "World Seastems" (`http://www.world-seastems.cnrs.fr`). The authors would like also to thank Catherine Matias and Stéphane Robin for their useful remarks and comments on this work.

References

[1] J. Moreno, Who shall survive?: A new approach to the problem of human interrelations., Nervous and Mental Disease Publishing Co, 1934.

[2] R. Albert, A. Barabási, Statistical mechanics of complex networks, Modern Physics 74 (2002) 47–97.

[3] A. Barabási, Z. Oltvai, Network biology: understanding the cell's functional organization, Nature Rev. Genet 5 (2004) 101–113.

[4] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.

[5] C. Ducruet, Network diversity and maritime flows, Journal of Transport Geography 30 (2013) 77–88.

[6] F. Rossi, N. Villa-Vialaneix, F. Hautefeuille, Exploration of a large database of French notarial acts with social network methods, Digital Medievalist 9 (2014) 1–20.

[7] J. Hofman, C. Wiggins, Bayesian approach to network modularity, Physical review letters 100 (25) (2008) 258701.

[8] P. Bickel, A. Chen, A nonparametric view of network models and newman–girvan and other modularities, Proceedings of the National Academy of Sciences 106 (50) (2009) 21068–21073.

[9] M. Girvan, M. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences 99 (12) (2002) 7821.

[10] M. Handcock, A. Raftery, J. Tantrum, Model-based clustering for social networks, Journal of the Royal Statistical Society: Series A (Statistics in Society) 170 (2) (2007) 301–354.

[11] Y. Wang, G. Wong, Stochastic blockmodels for directed graphs, Journal of the American Statistical Association 82 (1987) 8–19.

[12] K. Nowicki, T. Snijders, Estimation and prediction for stochastic block-structures, Journal of the American Statistical Association 96 (455) (2001) 1077–1087.

[13] H. White, S. Boorman, R. Breiger, Social structure from multiple networks. i. blockmodels of roles and positions, American Journal of Sociology (1976) 730–780.

[14] S. Fienberg, S. Wasserman, Categorical data analysis of single sociometric relations, Sociological Methodology 12 (1981) 156–192.

[15] M. Mariadassou, S. Robin, C. Vacher, Uncovering latent structure in valued graphs: a variational approach, Annals of Applied Statistics 4 (2) (2010) 715–742.

[16] H. Zanghi, S. Volant, C. Ambroise, Clustering based on random graph model embedding vertex features, Pattern Recognition Letters 31 (9) (2010) 830–836.

[17] C. Matias, S. Robin, Modeling heterogeneity in random graphs through latent space models: a selective review., Esaim Proc. and Surveys 47 (2014) 55–74.

[18] Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, S. Lamassé, The random subgraph model for the analysis of an acclesiastical network in merovingian gaul, Annals of Applied Statistics 8 (1) (2014) 55–74.

[19] E. Airoldi, D. Blei, S. Fienberg, E. Xing, Mixed membership stochastic blockmodels, The Journal of Machine Learning Research 9 (2008) 1981–2014.

[20] P. Latouche, E. Birmelé, C. Ambroise, Overlapping stochastic block models with application to the french political blogosphere, Annals of Applied Statistics 5 (1) (2011) 309–336.

[21] J.-J. Daudin, F. Picard, S. Robin, A mixture model for random graphs, Statistics and Computing 18 (2) (2008) 173–183.

[22] P. Latouche, E. Birmelé, C. Ambroise, Variational bayesian inference and complexity control for stochastic block models, Statistical Modelling 12 (1) (2012) 93–115.

[23] A. Mc Daid, T. Murphy, F. N., N. Hurley, Improved bayesian inference for the stochastic block model with application to large networks, Computational Statistics and Data Analysis 60 (2013) 12–31.

[24] E. Côme, P. Latouche, Model selection and clustering in stochastic block models with the exact integrated complete data likelihood, Statistical Modelling (2015) in press.

[25] C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, N. Ueda, Learning systems of concepts with an infinite relational model, in: Proceedings of the National Conference on Artificial Intelligence, Vol. 21, 2006, pp. 381–391.

[26] T. Yang, Y. Chi, S. Zhu, Y. Gong, R. Jin, Detecting communities and their evolutions in dynamic social networksa bayesian approach, Machine learning 82 (2) (2011) 157–189.

[27] K. S. Xu, A. O. Hero III, Dynamic stochastic blockmodels: Statistical models for time-evolving networks, in: Social Computing, Behavioral-Cultural Modeling and Prediction, Springer, 2013, pp. 201–210.

[28] E. Xing, W. Fu, L. Song, A state-space mixed membership blockmodel for dynamic network tomography, The Annals of Applied Statistics 4 (2) (2010) 535–566.

[29] Q. Ho, L. Song, E. P. Xing, Evolving cluster mixed-membership blockmodel for time-evolving networks, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 342–350.

[30] M. Kim, J. Leskovec, Nonparametric multi-group membership model for dynamic networks, in: Advances in Neural Information Processing Systems (25), 2013, pp. 1385–1393.

[31] P. Sarkar, A. W. Moore, Dynamic social network analysis using latent space models, ACM SIGKDD Explorations Newsletter 7 (2) (2005) 31–40.

[32] C. Dubois, C. Butts, P. Smyth, Stochastic blockmodelling of relational event dynamics, in: International Conference on Artificial Intelligence and Statistics, Vol. 31 of the Journal of Machine Learning Research Proceedings, 2013, pp. 238–246.

[33] C. Heaukulani, Z. Ghahramani, Dynamic probabilistic models for latent feature propagation in social networks, in: Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 275–283.

[34] J. R. Foulds, C. DuBois, A. U. Asuncion, C. T. Butts, P. Smyth, A dynamic relational infinite feature model for longitudinal social networks, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 287–295.

[35] K. S. Xu, Stochastic block transition models for dynamic networks, in: International Conference on Artificial Intelligence and Statistics, 2015, pp. 1079–1087.

[36] A. Ahmed, E. P. Xing, On tight approximate inference of logistic-normal admixture model, In Proceedings of the International Conference on Artificial Intelligence and Statistics (2007) 1–8.

[37] D. Blei, J. Lafferty, A correlated topic model of science, The Annals of Applied Statistics (2007) 17–35.

[38] J. D. Lafferty, D. M. Blei, Correlated topic models, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems 18, MIT Press, 2006, pp. 147–154.

[39] D. Blei, J. Lafferty, A correlated topic model of science, Annals of Applied Statistics 1 (1) (2007) 17–35.

[40] A. Harvey, Forecasting, structural time series models and the Kalman filter, Cambridge University Press, Cambridge, UK, 1989.

[41] T. Krishnan, G. McLachlan, The EM algorithm and extensions, John Wiley, New York, 1997.

[42] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38.

[43] R. J. Hathaway, Another interpretation of the EM algorithm for mixture distributions, Statistics & Probability Letters 4 (2) (1986) 53–56.

[44] M. Jordan, Z. Ghahramani, T. Jaakkola, L. K. Saul, An introduction to variational methods for graphical models, Machine learning 37 (2) (1999) 183–233.

[45] C. Bishop, M. Svensén, Bayesian hierarchical mixtures of experts, in: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence, U. Kjaerulff and C. Meek, 2003, pp. 57–64.

[46] P. Latouche, E. Birmelé, C. Ambroise, Model selection in overlapping stochastic block models, Electronic Journal of Statistics 8 (1) (2014) 762–794.

[47] H. Rauch, F. Tung, T. Striebel, Maximum likelihood estimates of linear dynamic systems, AIASS Journal 3 (8) (1965) 1445–1450.

[48] T. Minka, From hidden markov models to linear dynamical systems, Tech. rep., MIT (1998).

[49] M. Svensén, C. Bishop, Robust bayesian mixture modelling, Neurocomputing 64 (2004) 235–252.

[50] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.

[51] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.

[52] B. Leroux, Consistent estimation of amixing distribution, Annals of Statistics 20 (1992) 1350–1360.

[53] W. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical association (1971) 846–850.

[54] C. Ambroise, G. Grasseau, M. Hoebeke, P. Latouche, V. Miele, F. Picard, The mixer R package (version 1.8), *http://cran.r-project.org/web/packages/mixer/* (2010).

[55] C. Bouveyron, Y. Jernite, P. Latouche, L. Nouedoui, The rambo R package (version 1.1), *http://cran.r-project.org/web/packages/Rambo/* (2013).

A. Construction of a tractable lower bound

We rely on a bound introduced in [44]. Such a general bound can easily been derived by noticing that $C(\cdot)$ is a concave function of $\sum_{l=1}^{K} \exp(\gamma_{s_i l}^{(t)})$ and therefore a first order Taylor expansion of the normalizing constant, at any $\xi_s^{(t)} \in \mathbb{R}^{*+}$, will lead to the inequality:

$$\log(\sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)})) \leq \xi_s^{-1(t)} (\sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)})) - 1 + \log(\xi_s^{(t)}). \tag{8}$$

The bounds (8) on the $C(\gamma_s^{(t)})$ terms induce a lower bound on the quantity $\log p(Z|\gamma)$:

$$\begin{aligned} \log p(Z|\gamma) &= \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} Z_{ik}^{(t)} \log(f_k(\gamma_{s_i}^{(t)})) \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{s=1}^{S} y_{is} Z_{ik}^{(t)} \Big(\gamma_{sk}^{(t)} - \log(\sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)})) \Big) \\ &\geq \log h(Z, \gamma, \xi), \end{aligned}$$

where ξ denotes the set of all variational parameters $(\xi_s^{(t)})_{st}$ and the function $h(\cdot, \cdot, \cdot)$ is such that:

$$\log h(Z, \gamma, \xi) = \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{s=1}^{S} y_{is} Z_{ik}^{(t)} \Big(\gamma_{sk}^{(t)} - \big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)}) - 1 + \log(\xi_s^{(t)}) \big) \Big).$$

Replacing $\log p(Z|\gamma)$ by $\log h(Z, \gamma, \xi)$ in $\mathcal{L}(q, \theta)$, leads to a new lower bound $\tilde{\mathcal{L}}(q, \theta, \xi)$ for $\log p(X|\theta)$ which satisfies:

$$\log p(X|\theta) \geqslant \mathcal{L}(q, \theta) \geqslant \tilde{\mathcal{L}}(q, \theta, \xi),$$

where

$$\begin{aligned} &\tilde{\mathcal{L}}(q, \theta, \xi) \\ &= \sum_{Z} \int_{\gamma} \int_{\nu} q(Z, \gamma, \nu) \log \frac{p(X|Z, \Pi) h(Z, \gamma, \xi) p(\gamma|B, \nu, \Sigma) p(\nu|\mu_0, A, \Phi, V_0)}{q(Z, \gamma, \nu)} d\gamma d\nu. \end{aligned}$$

B. E-step of the VEM algorithm

Distribution $q(Z)$

The VEM update step for each of the distributions $q(Z_i)$ in $q(Z)$ is given by:

$$
\begin{aligned}
\log q(Z_i) &= E_{\gamma,\nu,Z^{\setminus i}}[\log p(X|Z,\Pi) + \log h(Z,\gamma,\xi)] + \text{const} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} Z_{ik}^{(t)} \left(\sum_{l=1}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)} = c)\tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] \right) \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{s=1}^{S} y_{is} E_{\gamma}\Big[Z_{ik}^{(t)} \log h(Z^{(t)}, \gamma^{(t)}, \xi^{(t)}) \Big] + \text{const.} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} Z_{ik}^{(t)} \left(\sum_{l=1}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)} = c)\tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] \right) \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{s=1}^{S} y_{is} E_{\gamma}\Big[\gamma_{sk}^{(t)} - (\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\gamma_{sk}^{(t)}) - 1 + \log(\xi_s^{(t)})) \Big] + \text{const.} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} Z_{ik}^{(t)} \left(\sum_{l=1}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)} = c)\tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] \right) \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{s=1}^{S} Z_{ik}^{(t)} y_{is} \left(\hat{\gamma}_{sk}^{(t)} - \Big[\xi_s^{-1(t)} \sum_{l=1}^{K} E(\exp(\gamma_{sl}^{(t)})) - 1 + \log(\xi_s^{(t)}) \Big] \right) + \text{const.} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} Z_{ik}^{(t)} \left(\sum_{l=1}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)} = c)\tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] \right. \\
&+ \left. \sum_{s=1}^{S} y_{is} \Big(\hat{\gamma}_{sk}^{(t)} - \Big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\hat{\gamma}_{sl}^{(t)} + \frac{\hat{\sigma}_{sl}^{2^{(t)}}}{2}) - 1 + \log(\xi_s^{(t)}) \Big) \Big) \right) + \text{const},
\end{aligned}
$$

where all terms that do not depend on Z_i have been put into the constant terms const. Moreover since $\gamma_{sk}^{(t)} \sim \mathcal{N}(\hat{\gamma}_{sk}^{(t)}, \hat{\sigma}_{sk}^{2^{(t)}})$ we have used:

$$\mathbb{E}[\exp(\gamma_{sk}^{(t)})] = \exp(\hat{\gamma}_{sk}^{(t)} + \frac{\hat{\sigma}_{sk}^{2^{(t)}}}{2}).$$

We then recognize the functional form of a multinomial distribution:

$$q(Z_i^{(t)}) \sim \mathcal{M}(Z_i^{(t)}; 1, \tau_i^{(t)}), \ \ \forall i, t.$$

Distribution $q(\nu)$:

The VEM update step for the distribution $q(\nu)$ is given by:

$$\begin{aligned}
\log q(\nu) &= E_{Z,\gamma}\Big(\log p(\gamma|\nu,\Sigma,B) + \log p(\nu|\mu_0,V_0,A,\Phi)\Big) + \text{const} \\
&= \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(E_\gamma\big(\log \mathcal{N}(\gamma_s^{(t)}; B\nu^{(t)},\Sigma)\big)\Big) + \log p(\nu^{(1)}|\mu_0,V_0) \\
&+ \sum_{t=2}^{T}\log p(\nu^{(t)}|\nu^{(t-1)},A,\Phi) + \text{const} \\
&= \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(E_\gamma\big(-\frac{1}{2}(\gamma_s^{(t)})^\intercal\Sigma^{-1}(\gamma_s^{(t)}) + (\gamma_s^{(t)})^\intercal\Sigma^{-1}B\nu^{(t)} - \frac{1}{2}(\nu^{(t)})^\intercal B^\intercal\Sigma^{-1}B\nu^{(t)}\big)\Big) \\
&+ \log p(\nu^{(1)}|\mu_0,V_0) + \sum_{t=2}^{T}\log p(\nu^{(t)}|\nu^{(t-1)},A,\Phi) + \text{const.} \\
&= \sum_{t=1}^{T}\Big(\sum_{s=1}^{S}\Big(\hat{\gamma}_s^{(t)}\Sigma^{-1}B\nu^{(t)}\Big) - \frac{1}{2}(\nu^{(t)})^\intercal B^\intercal(S\Sigma^{-1})B\nu^{(t)}\Big) \\
&+ \log p(\nu^{(1)}|\mu_0,V_0) + \sum_{t=2}^{T}\log p(\nu^{(t)}|\nu^{(t-1)},A,\Phi) + \text{const},
\end{aligned}$$

where all terms that do not depend on ν have been put into the constant terms const. We recognize the functional form of the posterior distribution of a linear dynamic system:

$$\begin{aligned}
\log q(\nu) &= \sum_{t=1}^{T}\Big(\log\mathcal{N}(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}; B\nu^{(t)},\frac{\Sigma}{S})\Big) \\
&+ \log p(\nu^{(1)}|\mu_0,V_0) + \sum_{t=2}^{T}\log p(\nu^{(t)}|\nu^{(t-1)},A,\Phi) + \text{const.}
\end{aligned}$$

C. Derivation of the lower bound

In the following, we denote $x^{(t)} = \frac{\sum_{s=1}^{S} \hat{\gamma}_s^{(t)}}{S}$ an observed variable. The lower bound is given by:

$$\begin{aligned}
\tilde{\mathcal{L}}(q,\theta,\xi) &= \sum_{Z} \int_{\gamma} \int_{\nu} q(Z,\gamma,\nu) \log \frac{p(X|Z,\Pi)h(Z,\gamma,\xi)p(\gamma|\nu,\Sigma)p(\nu|\mu_0,A,\Phi,V_0)}{q(Z,\gamma,\nu)} d\nu d\gamma \\
&= E_{Z,\gamma,\nu}[\log \frac{p(X|Z,\Pi)h(Z,\gamma,\xi)p(\gamma|\nu,\Sigma,B)p(\nu|\mu_0,A,\Phi,V_0)}{q(\gamma)q(\nu)\prod_{i=1}^{N} q(Z_i)}] \\
&= E_Z(\log p(X|Z,\Pi)) + E_{Z,\gamma}(\log h(Z,\gamma,\xi)) + E_{\gamma,\nu}(\log p(\gamma|\nu,\Sigma,B)) \\
&+ E_\nu(\log p(\nu|\mu_0,A,\Phi,V_0)) - E_\gamma(\log q(\gamma)) - E_\nu(\log q(\nu)) - E_Z(\log(\prod_{i=1}^{N} q(Z_i))).
\end{aligned}$$

Note that (see Proposition 3.3),

$$q(\nu) \propto p(\nu^{(1)}|\mu_0,V_0)\Big[\prod_{t=2}^{T} p(\nu^{(t)}|\nu^{(t-1)},A,\Phi)\Big]\Big[\prod_{t=1}^{T} \mathcal{N}(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}; B\nu^{(t)}, \frac{\Sigma}{S})\Big].$$

As pointed out in this proposition, this corresponds to the form of the posterior distribution associated with a state space model with parameter θ' and with observed outputs $x = (x^{(t)})_t$. If we denote $p(x|\theta')$ the likelihood associated with this model, and the joint likelihood $p(x,\nu|\theta')$, we have

$$q(\nu) = \frac{p(x,\nu|\theta')}{p(x|\theta')}.$$

Therefore

$$E_\nu(\log q(\nu)) = E_\nu(\log p(\nu|\mu_0,A,\Phi,V_0)) + E_\nu(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)}, \frac{\Sigma}{S}, B)) - \log p(x|\theta').$$

This leads to,

$$E_\nu(\log p(\nu|\mu_0,A,\Phi,V_0)) - E_\nu(\log q(\nu)) = -E_\nu(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)}, \frac{\Sigma}{S}, B)) + \log p(x|\theta'),$$

and $\tilde{\mathcal{L}}(q,\theta,\xi)$ can be written as follows:

$$\begin{aligned}
\tilde{\mathcal{L}}(q,\theta,\xi) &= \sum_{Z}\int_{\gamma}\int_{\nu} q(Z,\gamma,\nu)\log\frac{p(X|Z,\Pi)h(Z,\gamma,\xi)p(\gamma|\nu,\Sigma)p(\nu|\mu_0,A,\Phi,V_0)}{q(Z,\gamma,\nu)} \\
&= E_{Z,\gamma,\nu}[\log\frac{p(X|Z,\Pi)h(Z,\gamma,\xi)p(\gamma|\nu,\Sigma,B)p(\nu|\mu_0,A,\Phi,V_0)}{q(\gamma)q(\nu)\prod_{i=1}^{N}q(Z_i)}] \\
&= E_Z(\log p(X|Z,\Pi)) + E_{Z,\gamma}(\log h(Z,\gamma,\xi)) + E_{\gamma,\nu}(\log p(\gamma|\nu,\Sigma,B)) \\
&- E_\gamma(\log q(\gamma)) - E_\nu(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)},\frac{\Sigma}{S},B)) - E_Z(\log(\prod_{i=1}^{N}q(Z_i))) \\
&+ \log p(x|\theta^{'}).
\end{aligned}$$

We explicit below each of the terms of the bound $\tilde{\mathcal{L}}(q,\theta)$.

1. $E_Z(\log p(X|Z,\Pi))$:

$$\begin{aligned}
E_Z(\log p(X|Z,\Pi)) &= \sum_{t=1}^{T}\sum_{k,l}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} E_z(\delta(X_{ij}^{(t)}=c)Z_{ik}^{(t)}Z_{jl}^{(t)}\log(\Pi_{kl}^{c})) \\
&= \sum_{t=1}^{T}\sum_{k,l}^{K}\sum_{c=0}^{C}\sum_{i\neq j}^{N} \delta(X_{ij}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}\log(\Pi_{kl}^{c})
\end{aligned}$$

2. $E_{Z,\gamma}(\log h(Z,\gamma,\xi))$:

$$\begin{aligned}
E_{Z,\gamma}(\log h(Z,\gamma,\xi)) &= E_{Z,\gamma}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{s=1}^{S} y_{is}Z_{ik}^{(t)}\Big(\gamma_{sk}^{(t)} - \big(\xi_s^{-1(t)}\sum_{l}\exp(\gamma_{sk}^{(t)}) \\
&- 1+\log(\xi_s^{(t)})\big)\Big)\Big] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{s=1}^{S} y_{is}\Big(\tau_{ik}^{(t)}\hat{\gamma}_{sk}^{(t)} - \tau_{ik}^{(t)}\xi_s^{-1(t)}\sum_{l=1}^{K}\exp(\hat{\gamma}_{sl}^{(t)}+\frac{\hat{\sigma}_{sl}^{2^{(t)}}}{2}) \\
&+ \tau_{ik}^{(t)} - \tau_{ik}^{(t)}\log(\xi_s^{(t)})\Big) \\
&= \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(r_s^{(t)}\hat{\gamma}_{sk}^{(t)} - N_s\xi_s^{-1(t)}\sum_{l=1}^{K}\exp(\hat{\gamma}_{sl}^{(t)}+\frac{\hat{\sigma}_{sl}^{2^{(t)}}}{2}) + N_s - N_s\log(\xi_s^{(t)})\Big)
\end{aligned}$$

where denote $r_s^{(t)}$ is a quantity $\sum_{i=1}^{N}\tau_{ik}^{(t)}y_{is}$.

3. $E_{\gamma,\nu}(\log p(\gamma|\nu,\Sigma,B))$:

$$\begin{aligned} E_{\gamma,\nu}(\log p(\gamma|\nu,\Sigma,B)) &= E_{\gamma,\nu}\Big(\log\prod_{t=1}^{T}\prod_{s=1}^{S}\mathcal{N}(\gamma_s^{(t)};B\nu_s^{(t)},\Sigma)\Big) \\ &= \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(\log\mathcal{N}(\hat{\gamma}_s^{(t)},B\hat{\nu}_s^{(t)},\Sigma)-\frac{1}{2}tr(\Sigma^{-1}B^T\hat{V}^{(t)}B)-\frac{1}{2}tr(\Sigma^{-1}\hat{\sigma}_s^{(t)^2})\Big) \end{aligned}$$

4. $E_{\gamma}(\log q(\gamma))$:

$$\begin{aligned} E_{\gamma}(\log q(\gamma)) &= E_{\gamma}\Big(\prod_{t=1}^{T}\prod_{s=1}^{S}\prod_{k=1}^{K}\mathcal{N}(\gamma_{sk}^{(t)};\hat{\gamma}_{sk}^{(t)},\hat{\sigma}_{sk}^{2^{(t)}})\Big) \\ &= \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{k=1}^{K}-\log\Big((2\pi)^{\frac{1}{2}}\hat{\sigma}_{sk}^{(t)}\Big)-\frac{TKS}{2}. \end{aligned}$$

5. $E_{\nu}(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)},\frac{\Sigma}{S},B))$:

$$E_{\nu}(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)},\frac{\Sigma}{S},B)) = \sum_{t=1}^{T}\Big(\log\mathcal{N}(x^{(t)};B\hat{\nu}^{(t)},\Sigma/S)-\frac{1}{2}tr(\Sigma^{-1}SB^T\hat{V}^{(t)}B)\Big).$$

6. $E_Z(\log(\prod_{i=1}^{T}q(Z_i)))$:

$$\begin{aligned} E_Z(\log(\prod_{i=1}^{T}q(Z_i))) &= \sum_{i=1}^{N}E_Z(\log q(Z_i)) \\ &= \sum_{i=1}^{N}E_Z(\sum_{t=1}^{T}\sum_{k=1}^{K}Z_{ik}^{(t)}\log(\tau_{ik})) \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log(\tau_{ik}^{(t)}). \end{aligned}$$