UN MODÈLE DYNAMIQUE DE SOUS-GRAPHES ALÉATOIRES. ÉTUDE DU SCANDALE ENRON

R Zreik, P Latouche, Charles Bouveyron

UN MODÈLE DYNAMIQUE DE SOUS-GRAPHES ALÉATOIRES. ÉTUDE DU SCANDALE ENRON

R. Zreik, P. Latouche & C. Bouveyron

Résumé. — Ces dernières années, de nombreux modèles de graphes aléatoires ont été proposés pour extraire des informations à partir de réseaux dans des domaines variés. Le principe de ces modèles consiste à chercher des groupes de nœuds ayant des profils de connexion homogènes. La plupart de ces modèles sont adaptés pour des réseaux statiques ayant des arêtes binaires ou discrètes mais sans prendre en compte la dimension temporelle. Ce travail est motivé par la nécessité d'analyser un réseau dynamique décrivant les communications électroniques (e-mail) entre les employés de l'entreprise Enron où les positions sociales jouent un rôle important. Nous proposons dans cet article une extension au cadre dynamique du modèle de graphe aléatoire RSM qui a été récemment proposé pour modéliser à l'aide de groupes latents des réseaux statiques pour lesquels une partition en sous-graphes est connue. Notre approche est basée sur l'utilisation d'un *state-space model* pour modéliser l'évolution au cours du temps des proportions des groupes latents. Le modèle ainsi obtenu est appelé modèle de sous-graphes aléatoires dynamiques (dRSM) et un algorithme de type EM variationnel (VEM) est proposé pour en effectuer l'inférence. Nous montrons que les approximations variationnelles conduisent à un nouveau *state-space model* à partir duquel les paramètres ainsi que les états cachés peuvent être estimés en utilisant le filtre de Kalman et le Rauch-Tung-Striebel (RTS) *smoother*. La méthodologie est finalement appliquée au jeu des données d'e-mails de l'entreprise Enron et permet de mettre en évidence une réaction anticipée des cadres par rapport aux autres employés concernant le scandale à venir.

Mots clefs. — Réseau dynamique, sous-graphes, random subgraph model (RSM), *state-space model*, classification, algorithme VEM, données Enron.

Abstract. — In recent years, many random graph models have been proposed to extract information from networks. The principle is to look for communities or groups of vertices with homogenous connection profiles. Most of these models are suitable for static networks, that is to say, not taking into account the temporal dimension, but can handle different types of edges, whether binary or discrete. This work is motivated by the need of analysing an evolving network describing email communications between employees of the Enron compagny where social positions play an important role. 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 maximisation (VEM) algorithm is proposed to perform inference. We show that the variational approximations lead to a new state space model from which the parameters along with hidden states can be estimated using the standard Kalman filter and Rauch-Tung-Striebel (RTS) smoother. The methodology is finally applied to the Enron email dataset and allows to discover a early reaction of the partners and directors compared to the other employees regarding the coming scandal.

1. Introduction

Depuis les travaux précurseurs de Moreno [**16**], l'analyse des réseaux est devenue une discipline forte, qui ne se limite plus à la sociologie et qui est à présent appliquée à des domaines très variés tels que la biologie, la géographie ou l'histoire. L'intérêt croissant pour l'analyse des réseaux s'explique d'une part par la forte présence de ce type de données dans le monde numérique d'aujourd'hui et, d'autre part, par les progrès récents dans la modélisation et le traitement de ces données. En effet, informaticiens et statisticiens ont porté leurs efforts depuis plus d'une dizaine d'années sur ces données de type réseau et ont proposé de nombreuses techniques permettant leur analyse. Les méthodes de clustering permettent en particulier de recouvrir une structure en groupes cachés dans le réseau. Parmi ces méthodes, on peut citer les travaux de Hofman et Wiggins [**10**] qui cherchent une partition des sommets où les groupes présentent une propriété de transitivité. Le modèle de Handcock, Raftery et Tantrum [**9**] suppose quant à lui que les liens entre les sommets dépendent des positions des sommets dans un espace latent. Une approche très populaire également, même si celle-ci est asymptotiquement biaisée [**3**], est celle proposée par Girvan et Newman [**8**] qui repose sur la notion de modularité.

Outre l'approche de Handcock *et al.*, les méthodes statistiques récentes sont généralement basées sur le modèle à blocs stochastiques (SBM) [**6, 17, 18**], qui est une généralisation probabiliste de la méthode appliquée par [**19**] sur les

fameuses données de Sampson [**7**]. Le modèle SBM suppose que chaque sommet appartient à un groupe latent et que la probabilité de connexion entre une paire de sommets dépend exclusivement de leur groupe. Parmi les nombreuses extensions récentes du modèle SBM, on peut citer d'une part les modèles autorisant un nœud du réseau à appartenir à un ou plusieurs groupes : le mixed membership stochastic block model (MMSBM) [**2**] et le overlapping stochastic block model (OSBM) [**13**]. D'autre part, certains auteurs se sont intéressés à la modélisation de réseaux avec des arêtes valuées [**14**] et une prise en compte d'une information a priori [**15**]. En particulier, le travail de Jernite *et al.* [**11**] considère des arêtes catégorielles et la connaissance d'une partition du réseau en sous-graphes. Ce modèle, baptisé random subgraph model (RSM), suppose que chaque sous-graphe possède son propre mélange de groupes, ces derniers pouvant être présents dans tous les sous-graphes. Les sommets sont alors connectés entre eux avec une probabilité dépendante seulement des sous-graphes alors que le type d'arête est supposé être échantillonné conditionnellement aux groupes latents. Ce modèle a été appliqué avec succès à l'analyse d'un réseau historique dans la Gaule mérovingienne.

Toutefois, dans ce dernier travail, la dynamique temporelle du réseau, à l'origine présente dans les données, avait dû être écartée à cause de l'incapacité du modèle à gérer cet aspect. Il apparaît donc aujourd'hui nécessaire de pouvoir prendre en compte la dimension temporelle afin de modéliser au mieux les réseaux qui sont, pour la plupart, observés dans un cadre dynamique. La littérature statistique est malheureusement peu fournie sur le sujet. Le travail le plus représentatif est probablement le modèle dMMSBM (dynamic mixed membership stochastic block model), proposé dans [**20**], qui étend le modèle MMSBM au cadre dynamique.

Dans cet article, nous nous proposons d'étendre le modèle RSM aux réseaux dynamiques. Le modèle RSM nous apparaît comme un modèle flexible pour ce passage au temporel puisqu'il généralise le modèle SBM sur deux aspects. En effet, le modèle SBM peut être retrouvé en considérant le modèle RSM avec un unique sous-graphe et un seul type de connexion. Le modèle que nous présentons dans ce travail, que nous baptiserons dynamic random subgraph model (dRSM), permettra de modéliser l'évolution temporelle des nœuds et leurs connexions grâce à un *state-space model* (SSM) [**4**, chapitre 13]. Notre modèle relie les probabilités de connexions des nœuds à chaque instant t à des paramètres d'état dans un espace latent à travers une transformation logistique [**1, 5**]. L'inférence de ce modèle sera faite grâce à un algorithme de type EM variationnel (VEM). Le modèle dRSM sera finalement appliqué au célèbre jeu de données Enron qui présente l'évolution des communications électroniques entre les employés pendant les deux ans (2001, 2002) avant la faillite de l'entreprise en décembre 2001. La Figure 1 présente l'évolution du

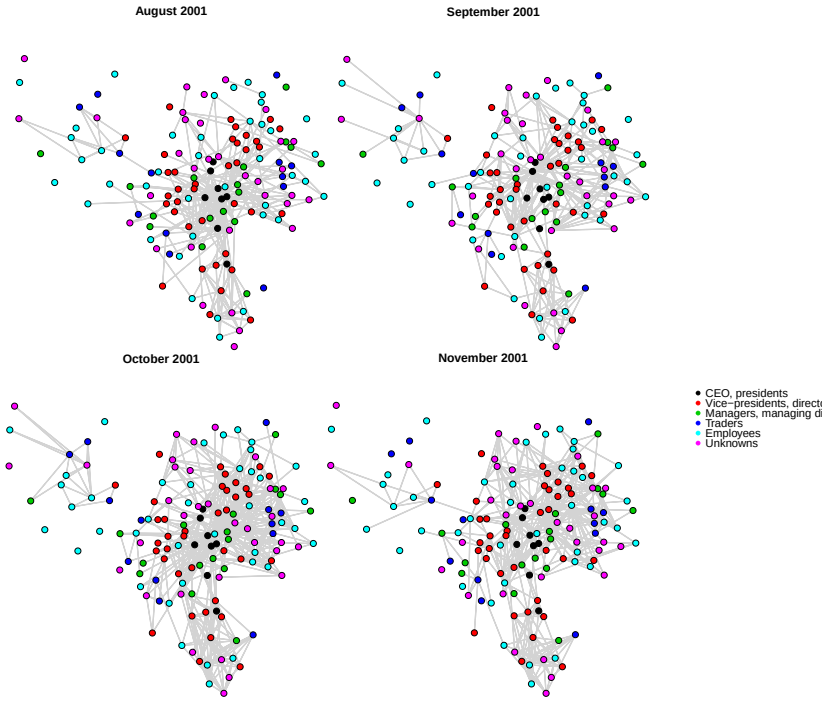

FIGURE 1. Réseau des communications électroniques entre 148 employés d'Enron durant les 4 mois (août - décembre 2001) précédant la faillite de l'entreprise.

réseau de communications e-mail dans les quatre mois clés (août – décembre 2001) de la crise Enron.

2. Le modèle de sous-graphes aléatoires dynamiques

Nous considérons un ensemble de T réseaux $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$, où $\mathcal{G}^{(t)}$ est un graphe dirigé observé au temps t et qui est représenté par une matrice d'adjacence $X^{(t)}$ de taille $N \times N$, N désignant le nombre de nœuds (supposé constant au cours du temps). Chaque arête $X_{ij}^{(t)}$, décrivant la relation entre les nœuds i et j, est supposée prendre ses valeurs dans $\{0, \dots C\}$ tel que $X_{ij}^{(t)} = c$ signifie que les nœuds i et j sont liés par une relation de type c au temps t et $X_{ij}^{(t)} = 0$ indique en particulier l'absence de relation entre les deux nœuds à cet instant. Notons que ne nous considérons pas les boucles, c'est à dire les connexions d'un noeud sur lui même, donc $X_{ii}^{(t)} = 0$, $\forall i, t$.

Nous supposons également, qu'à chaque instant de temps t, une partition $\mathcal{P}^{(t)}$ du réseau en S sous-graphes est également connue. Nous introduisons la

variable s qui prend ses valeurs dans $\{1, \dots S\}$ et qui est telle que $s_i^{(t)}$ indique à quel sous-graphe le nœud i appartient au temps t.

Notre objectif est finalement de regrouper à chaque temps t les N nœuds en K groupes latents de profils de connexions homogènes, *i.e.* trouver une estimation à chaque instant t de la matrice binaire Z telle que $Z_{ik}^{(t)} = 1$ si, à l'instant t, le nœud i appartient à la classe k et 0 sinon.

2.1. ***Le modèle à un instant de temps.*** — Le réseau, représenté par sa matrice d'adjacence $X^{(t)}$, est supposé être généré à chaque instant t comme suit. Tout d'abord, chaque nœud i est associé à une classe latente k avec une probabilité dépendante du sous-graphe auquel il appartient. Nous supposons donc que, pour un nombre K de groupes latents donné, la variable $Z_i^{(t)}$ est distribuée selon une loi multinomiale de paramètre $\alpha_{s_i}^{(t)}$:

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

où $\alpha_s^{(t)} = (\alpha_{s1}^{(t)}, \dots, \alpha_{sK}^{(t)})$ est le vecteur des probabilités a priori des K groupes latents dans le sous-graphe s à l'instant t et est tel que :

$$\forall s \ \in 1, \dots, S, \ \sum_{k=1}^{K} \alpha_{sk}^{(t)} = 1.$$

Notons que le modèle autorise chaque sous-graphe à avoir des proportions $\alpha_s^{(t)}$ des groupes latents différentes et cela pour chaque instant de temps.

Nous supposons d'autre part que le type de lien entre les nœuds i et j est, conditionnellement aux variables $Z_i^{(t)}$ et $Z_j^{(t)}$, distribué à nouveau selon une loi multinomiale :

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

avec $\Pi_{kl} \in [0,1]^{C+1}$ et $\sum_{c=0}^{C} \Pi_{klc} = 1$. Remarquons que, par souci de parcimonie, les paramètres Π_{kl} sont supposés constants au cours du temps.

2.2. ***Modélisation de l'évolution des sous-graphes aléatoires.*** — Nous ajoutons à présent un état caché, sous la forme d'un *state-space model*, pour modéliser l'évolution des sous-graphes au cours du temps. Nous introduisons donc une nouvelle variable latente, $\gamma_s^{(t)}$, permettant de faire le lien entre les $\alpha_s^{(t)}$aux différents temps et celà grâce à une transformation de type logistique :

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

La variable latente $\gamma_s^{(t)}$ est quant à elle supposée être distribuée selon une loi normale centrée en $B\nu^{(t)}$ et de matrice de covariance Σ :

$$\gamma_s^{(t)} \sim \mathcal{N}(B\nu^{(t)}, \Sigma). \tag{1}$$

Notations	Descriptions
X	Matrice d'adjacence. $X_{i,j}^{(t)} \in \{0, \dots, C\}$ à chaque instant t.
Z	Matrice binaire d'appartenance aux groupes. $Z_{i,k}^{(t)} = 1$ si $i \in$ cluster k au temps t.
N	Nombre de sommets dans le réseau.
K	Nombre des groupes latents.
S	Nombre de sous-graphes.
C	Nombre de types d'arêtes.
T	Nombre de pas de temps.
α	$\alpha_{sk}^{(t)}$ est la proportion du groupe k dans le sous-graphe s au temps t.
Π	$\Pi_{k\ell c}$ est la probabilité d'une arête de type c entre les groupes k et ℓ.
ν	$\nu^{(t)}$ est la probabilité moyenne des groupes au temps t dans le *state-space model*.

TABLE 1. *Résumé des principales notations utilisées.*

Notons que cette variable, introduite dans notre modèle pour la modélisation des processus temporels, a un degré de liberté égal à $(K-1)$ du fait de la nature de $\alpha_s^{(t)}$. De ce fait, $\alpha_s^{(t)}$ est généré en échantillonnant $(K-1)$ des composantes de $\gamma_s^{(t)}$ selon une distribution normale de moyenne $B\nu^{(t)}$ de dimension $(K-1)$ et une matrice de covariance Σ également de taille $(K-1) \times (K-1)$. La dernière composants du vecteur γ_s est arbitrairement fixée à zéro. Ainsi, les variables $\alpha_s^{(t)}$ et $\gamma_s^{(t)}$ sont liées par la relation :

$$\alpha_{sk}^{(t)} = \exp(\gamma_{sk}^{(t)} - C(\gamma_s^{(t)})) \ \forall k = 1, \dots, K-1 \tag{2}$$

où $C(\gamma_s^{(t)}) = \sum_{\ell=1}^{K-1} \exp(\gamma_{s\ell}^{(t)})$ et $\alpha_{sK}^{(t)} = 1 - \sum_{\ell=1}^{K-1} \exp(\alpha_{s\ell}^{(t)})$.

Le reste de la modélisation fait maintenant intervenir un *state- space model* classique pour des systèmes dynamiques linéaires qui évolue au cours de temps. Le modèle est comme suit :

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

où A et B sont deux matrices de transition de taille $(K-1) \times (K-1)$. Les termes de bruit ω, u et v sont d'autre part supposés gaussiens :

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

Ainsi, les proportions $\alpha_s^{(t)}$ du mélange dans chaque sous-graphe aux différents temps $t = 1, \dots, T$ sont liées entre elles au travers du *state-space model*.

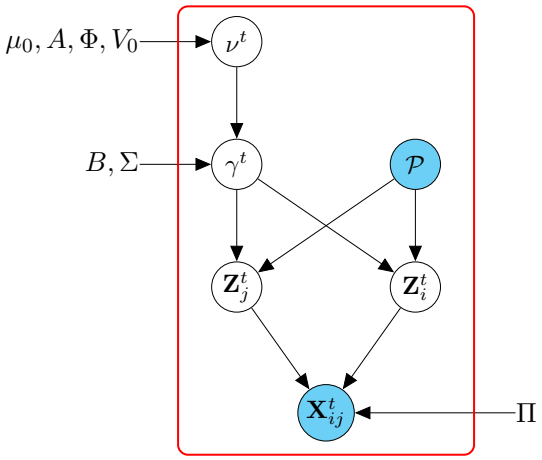

FIGURE 2. *Modèle graphique associé à dRSM.*

Le modèle ainsi décrit possède trois variables latentes (ν, γ, Z) et est paramétré par $\theta = (\mu_0, A, B, \Phi, V_0, \Sigma, \Pi)$. Ce modèle est appelé *dynamic random subgraph model* (dRSM) dans la suite du document. La Figure 2 présente le modèle graphique associé à dRSM et la Table 1 résume les principales notations utilisées.

Le modèle dRSM est donc défini par la distribution jointe suivante :

$$\begin{aligned} p(X, Z, \gamma, \nu|\theta) &= p(X, \gamma, \nu|Z, \Pi, \Sigma, \mu_0, A, \Phi, V_0)p(Z|f(\gamma)) \\ &= p(X|Z, \Pi)p(Z|f(\gamma))p(\gamma|\nu, \Sigma)p(\nu|\mu_0, A, \Phi, V_0), \end{aligned}$$

où :

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

$$\begin{aligned} p(Z|f(\gamma)) &= \prod_{t=1}^{T} \prod_{i}^{N} \prod_{k=1}^{K} f(\gamma_{s_i,k}^{(t)})^{Z_{ik}^{(t)}} \\ &= \prod_{t=1}^{T} \prod_{k=1}^{K} f(\gamma_{s_i,k}^{(t)})^{\sum_{i=1}^{N} Z_{ik}^{(t)}}, \end{aligned}$$

$$p(\gamma|\nu, \Sigma) = \prod_{t=1}^{T} \prod_{s=1}^{S} \mathcal{N}(\gamma_s^{(t)}, \Sigma),$$

et finalement

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

Les détails des calculs de cette vraisemblance complétée associée au modèle dRSM sont donnés en annexe A.1.

2.3. ***Un cadre variationnel pour l'inférence du modèle dRSM.*** — Nous considérons à présent l'inférence du modèle dRSM introduit dans les paragraphes précédents. Nous cherchons à maximiser la log-vraisemblance $\log p(X|\theta)$. Pour réaliser cette maximisation, il est d'usage de recourir à l'utilisation d'un algorithme EM. Malheureusement, l'utilisation de l'algorithme EM n'est pas possible dans notre cas. En effet, la distribution a posteriori $p(Z, \gamma, \nu|\theta)$ n'est pas calculable. Il est donc nécessaire d'utiliser un algorithme de type EM variationnel (VEM) qui optimise localement les paramètres du modèle par rapport à une borne inférieure de la log-vraisemblance. Cet algorithme permet de déterminer les distributions variationnelles a posteriori pour toutes les variables latentes.

Étant donnée une distribution variationnelle q de (Z, γ, ν), la log-vraisemblance peut s'écrire :

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

où $\mathcal{L}$ est définie comme suit :

$$\mathcal{L}(q, \theta) = \sum_z \int_\gamma \int_\nu q(Z, \gamma, \nu) \log \frac{p(X, Z, \gamma, \nu|\theta)}{q(Z, \gamma, \nu)} d\gamma \, d\alpha, \tag{3}$$

et la divergence de Kullback-Leibler est donnée par :

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

Trouver la meilleure approximation de la distribution a posteriori $p(Z, \gamma, \nu|X, \theta)$ au sens de la divergence KL est équivalent à trouver $q(\cdot)$ qui maximise la borne inférieure $\mathcal{L}$ de la log-vraisemblance. Or, $p(X, Z, \gamma, \nu|\theta)$, qui pour rappel à la forme suivante :

$$p(X, Z, \gamma, \nu|\theta) = p(X|Z, \Pi)p(Z|f(\gamma))p(\gamma|\nu, \Sigma)p(\nu|\mu_0, A, \Phi, V_0),$$

fait intervenir la quantité $p(Z|f(\gamma))$ dont le calcul de l'espérance est rendu difficile du fait de la constante de normalisation $C(\gamma_s^{(t)}) = \sum_{\ell=1}^{K} \exp(\gamma_{s\ell}^{(t)})$.

Il nous est alors nécessaire, pour maximiser la borne $\mathcal{L}(q, \theta)$, d'utiliser une approximation de Taylor **[12]** de $\log C(\gamma_s^{(t)})$ pour en pour trouver une borne

supérieure :

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

On obtient ensuite facilement une borne inférieure de $\log p(Z|f(\gamma))$:

$$\begin{aligned}
\log p(Z|f(\gamma)) &= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N} Y_{is}Z_{ik}^{(t)}\log(f(\gamma_{s_ik}^{(t)})) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N} Y_{is}Z_{ik}^{(t)}\Big(\gamma_{sk}^{(t)} - \log(\sum_{l=1}^{K}\exp(\gamma_{sl}^{(t)}))\Big) \\
&\geq \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N} Y_{is}Z_{ik}^{(t)}\Big(\gamma_{sk}^{(t)} - \big(\xi_s^{-1(t)}\sum_{l=1}^{K}\exp(\gamma_{sl}^{(t)}) - 1 + \log(\xi_s^{(t)})\big)\Big) \\
&= \log h(Z,\gamma,\xi),
\end{aligned}$$

où $\xi = (\xi_s^t)_s^t$ est un ensemble de paramètres variationnels.

En remplaçant $\log p(Z|f(\gamma))$ par $\log h(Z,\gamma,\xi)$ dans l'équation (3), on obtient une nouvelle borne inférieure $\tilde{\mathcal{L}}(q,\theta)$ pour $\log p(X|\theta)$:

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

où :

$$\tilde{\mathcal{L}}(q,\theta) = \sum_z \int_\gamma \int_\nu q(Z,\gamma,\nu)\log\frac{p(X|Z,\Pi)h(Z,\gamma,\xi)p(\gamma|\nu,\Sigma)p(\nu|\mu_0,A,\Phi,V_0)}{q(Z,\gamma,\nu)}.$$

Pour permettre la maximisation de $\tilde{\mathcal{L}}(q,\theta)$, nous supposons en outre que $q(Z,\gamma,\nu)$ a la forme variationnelle suivante :

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

et $q(\gamma)$ est un produit de lois normales de paramètres $\hat{\gamma}_{sk}^{(t)}$,$\hat{\sigma}_{sk}^{(t)}$:

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

où $\hat{\gamma}_s^{(t)}$ est un vecteur moyenne de dimension K et dont la dernière composante est égale à zéro.

Algorithm 1 Algorithme VEM pour le modèle dRSM à K groupes latents.

Initialisation de $\theta^0 = (\mu_0, A, \Phi, V_0, \Pi, B, \Sigma)$
Initialisation de la matrice τ à chaque instant t
Déterminer $\hat{\nu}, \hat{V}, \hat{\Phi}, \hat{\Sigma}$ en utilisant le filtre de Kalman et le RTS *smoother*
Calculer $\mathcal{L}(q, \theta)$
Tant que $\mid \mathcal{L}^{new} - \mathcal{L}^{old} \mid > \varepsilon$
    Étape E : mise à jour de $\tau, \hat{\nu}, \hat{V}$
    Étape M : mise à jour de Σ, Π, ξ
    Calcul de $\mathcal{L}(q, \theta)$
    Optime $\hat{\gamma}_{s_{ik}}^{(t)}$ et $\hat{\sigma}_{sl}^{(t)^2}$
fin boucle

2.4. ***Algorithme VEM pour l'inférence de dRSM.*** — Avec les approximations faites au paragraphe précédent, il est maintenant possible de mettre en œuvre un algorithme de type EM variationnel (VEM). L'algorithme VEM, dont une esquisse est donnée par l'algorithme 1, permet dans un tel cadre de maximiser itérativement la borne $\tilde{\mathcal{L}}(q, \theta)$ par rapport à q (étape E) et par rapport aux paramètres du modèle $\theta = (\mu_0, A, B, \Phi, V_0, \Sigma, \Pi)$ et variationel ξ (étape M). Nous donnons ci-après les formules de mise à jour de ces deux étapes. Les détails des calculs sont donnés en annexe.

Etape E. — L'étape de mise à jour de VEM pour la distribution $q(Z_i^{(t)})$ est donnée par :

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

où :

$$\tau_{ik}^{(t)} \propto \exp\left(\sum_{l=1}^{K} \sum_{c=1}^{C} \sum_{i \neq j}^{N} \delta(X_{i,j}^{(t)} = c) \tau_{jl}^{(t)} \Big[\log(\Pi_{kl}^{c}) + \log(\Pi_{lk}^{c}) \Big] \right.$$
$$\left. + \sum_{s=1}^{S} \hat{\gamma}_{s_{ik}}^{(t)} - \Big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\hat{\gamma}_{s_{ik}}^{(t)} + \frac{\hat{\sigma}_{sl}^{(t)^2}}{2}) - 1 + \log(\xi_s^{(t)}) \Big) \right) + \text{const.}$$

Concernant $\gamma_{s_k}^{(t)}$, rappelons que $\gamma_{s_k}^{(t)} \sim \mathcal{N}(\hat{\gamma}_{sk}^{(t)}, \hat{\sigma}_{sk}^{(t)^2})$ et, par conséquent, il est facile de montrer que :

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

Pour $q(\nu)$, il n'est en revanche pas possible de reconnaitre une loi de probabilité mais nous avons pu établir la forme suivante pour la distribution (dont le calcul

est détaillé dans l'annexe A.2) :

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

Cette formulation de $q(\nu)$ est tout à fait remarquable puisqu'il s'agit d'une distribution associée à un *state-space model*. Il est donc possible de faire correspondre à $q(\nu)$ un système linéaire de la forme :

$$\begin{cases} \nu^{(t)} = A\nu^{(t-1)} + \omega \\ \omega \sim \mathcal{N}(0, \Phi) \\ \nu^{(1)} = \mu_0 + u, \end{cases}$$

où $\nu^{(t)}$ est une variable latente et $x^{(t)} = \dfrac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}$ est considéré comme une variable observée telle que :

$$x^{(t)} = C\nu^{(t)} + \tilde{v},$$

et

$$\tilde{v} \sim \mathcal{N}(0, \frac{\Sigma}{S}).$$

En conséquence, l'approximation variationnelle de $q(\nu)$ conduit à un nouveau *state-space model* avec une variable d'observation $x^{(t)}$. L'estimation des paramètres de ce modèle peut être fait grâce à un filtre de Kalman standard et le Rauch-Tung-Striebel (RTS) *smoother*. Le paquet MARSS pour le logiciel R permet notamment de faire l'estimation d'un tel *state-space model*. Notons que x rassemble l'ensemble des variables observées $x^{(t)}$, et nous notons $\left(\hat{\nu}^{(t)}\ , \hat{V}^{(t)}\right)$ les estimations associées à l'espérance et la matrice variance-covariance de la variable latente $\nu^{(t)}$ sachant x.

Etape M. — Les résultats précédents nous permettent à présent de construire la borne $\tilde{\mathcal{L}}(q,\theta)$ qui prend alors la forme suivante :

$$\begin{aligned}
\tilde{\mathcal{L}}(q,\theta) \quad = \quad & \log p(x|\theta) + \sum_{t=1}^{T}\sum_{k,l}^{K}\sum_{c=1}^{C}\sum_{i\neq j}^{N} \delta(X_{i,j}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}\log(\Pi_{kl}^{c}) \\
+ \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(r_s^{(t)}\hat{\gamma}_{s_{ik}}^{(t)} - N_s\xi_s^{-1(t)}\sum_{l=1}^{K}\exp(\hat{\gamma}_{s_l}^{(t)} + \frac{\hat{\sigma}_{sl}^{(t)^2}}{2}) + N_s - N_s\log(\xi_s^{(t)})\Big) \\
+ \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\Big(\log\mathcal{N}(\hat{\gamma}_s^{(t)}, B\hat{\nu}_s^{(t)}, \Sigma) - \frac{1}{2}tr(\Sigma^{-1}B^{\intercal}\hat{V}^{(t)}B) - \frac{1}{2}tr(\Sigma^{-1}\hat{\sigma}_s^{(t)^2})\Big) \\
- \quad & \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{k=1}^{K-1}\Big(-\log(2\pi)^{\frac{1}{2}}\hat{\sigma}_{sk}^{(t)}\Big) + \frac{TKS}{2} \\
- \quad & \sum_{t=1}^{T}\Big(\log\mathcal{N}(x^{(t)}; B\hat{\nu}^{(t)}, \frac{\Sigma}{S}) + \frac{1}{2}tr(\Sigma^{-1}SB^{\intercal}\hat{V}^{(t)}B)\Big) \\
- \quad & \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log(\tau_{ik}^{(t)}).
\end{aligned}$$

Le détail du calcul de la borne $\tilde{\mathcal{L}}(q,\theta)$ est donné en annexe A.3. La maximisation de cette borne permet d'obtenir les formules de mise à jour pour les paramètres Π et ξ :

$$\hat{\Pi}_{kl}^{c} = \frac{\sum_{t=1}^{T}\sum_{i\neq j}^{N}\delta(X_{i,j}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}}{\sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i\neq j}^{N}\delta(X_{i,j}^{(t)}=c)\tau_{ik}^{(t)}\tau_{jl}^{(t)}},$$

et

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

Les paramètres $\hat{\gamma}_{s_{ik}}^{(t)}$ et $\hat{\sigma}_{sl}^{(t)^2}$ doivent quant à eux être obtenus par une maximisation numérique de la borne. Cela peut notamment être fait grâce à un algorithme de type quasi-Newton.

3. Application aux réseau du scandale Enron

Nous appliquons ici la méthodologie décrite dans ce papier au jeu de données Enron. L'entreprise Enron était spécialisée dans l'énergie (gaz, électricité, ...) et est devenue célèbre au début des années 2000 du fait d'un scandale financier lié à ses activités de courtage. Enron avait dévelopé une activité de spéculation autour de l'électricité et des manipulations comptables visant à couvrir les

t	périodes
t_1	du 01/01/2000 au 01/12/2000
t_2	du 01/01/2001 au 01/03/2001
t_3	du 01/04/2001 au 01/06/2001
t_4	du 01/07/2001 au 01/09/2001
t_5	du 01/10/2001 au 01/03/2002

TABLE 2. Cinq périodes de temps considérées pour l'analyse des échanges d'emails au sein de l'entreprise Enron.

pertes de l'entreprise ont mené à sa faillite décembre 2001. A la suite de ce scandale financier, l'agence de régulation de l'énergie américaine a rendu public l'ensemble des emails de l'entreprise dans le cadre de ses investigations.

3.1. Données et protocole d'étude. — Nous disposons ainsi de tous les échanges d'emails entre 148 personnes d'intérêt ayant travaillé pour l'entreprise à chaque temps t. Afin de faire apparaître les changements structurels au sein de l'entreprise, les données sont d'abord regroupées par mois, puis par période de temps, de manière à ce que deux personnes soient considérées comme connectées si elles ont échangé au moins un email pendant la période associée. Nous nous intéressons donc ici à cinq périodes de temps notées $t_1, t_2, \ldots, t_5$ (voir Table 2). Les opérations de maquillage des pertes occasionnées par des opérations spéculatives furent révélées en octobre 2001 suite à l'ouverture d'une enquête par l'agence de régulation de l'énergie. L'entreprise fit finalement faillite en décembre 2001. Ces deux évènements clés correspondent à la période t_5.

Nous disposons d'autre part d'une partition des employés en trois sous-graphes selon leur statut dans l'entreprise. La Table 3 détaille cette partition. Notons que le sous-graphe s_1 comprend tous les cadres de l'entreprise, c'est à dire le directeur général, les présidents, vice-présidents, directeurs, managers, et directeurs managers. Les traders sont également associés à ce sous-graphe. Les individus du sous-graphe s_2 sont tous employés par l'entreprise mais n'ont pas le statut de cadre. Enfin, s_3 rassemble tous les autres individus en contact avec l'entreprise ayant également été touchés lors de la crise d'octobre 2001.

En résumé, le réseau dirigé sans boucle considéré ici est binaire : $C = 1$ et $X_{ij}^t = 1$ si i et j ont échangé au moins un email durant la période t, 0 sinon, avec $t \in \{1, \ldots, T\}$ et $T = 5$. Nous considérons trois sous-graphes connus, $S = 3$, et nous utilisons l'algorithme varitionnel EM décrit précédemment afin de rechercher $K = 4$ groupes latents dans les données. Ce choix est motivé par des considérations empiriques et permet d'obtenir un modèle dRSM décrivant l'apparition et surtout la gestion de la crise suite à l'ouverture de l'enquête.

Sous-graphes	statuts
s_1	Cadres
s_2	Employés
s_3	Autres

TABLE 3. Partition connue des personnes ayant travaillé pour Enron, en fonction de leur statut dans l'entreprise.

	cluster 1	cluster 2	cluster 3	cluster 4
cluster 1	0.478	0.037	0.005	0.023
cluster 2	0.020	0.181	0.006	0.012
cluster 3	0.001	0.002	0.001	0.003
cluster 4	0.012	0.012	0.024	0.119

TABLE 4. Termes Π_{kl1} de la matrice Π estimée à l'aide de l'algorithme variationnel EM pour $K = 4$ clusters.

3.2. Résultats. — Intéressons-nous tout d'abord à la topologie des clusters obtenus. Le réseau étant binaire, le modèle dRSM utilisé correspond à un cas particulier du modèle décrit à la Section 2 où $C = 1$. Par construction la matrice Π vérifie $\Pi_{kl0} + \Pi_{kl1} = 1, \forall(k, l)$ et par conséquent seules sont données les probabilités Π_{kl1} d'apparition d'arêtes dans la Table 4. Les clusters 1, 2, et 4 sont définis par des termes diagonaux significativement plus forts que les termes extra-diagonaux. Ces trois clusters correspondent donc à des communautés où la probabilité de connexion entre deux noeuds d'une même communauté est plus forte qu'entre des noeuds de communautés différentes. Ces clusters se distinguent principalement par le fait qu'ils ont des probabilités intra-cluster différentes de connexion. Ainsi, le cluster 1 est le cluster ayant la densité Π_{kk1} la plus forte, suivi du cluster 2 et du cluster 4. Finalement, notons que le cluster 3 est construit à partir de probabilités de connexion faibles. Il rassemble en fait tous les individus participant à des échanges d'emails peu structurés dans le réseau.

Le principal avantage du modèle dRSM est qu'il permet de caractériser l'évolution des sous-graphes en fonction de clusters latents estimés par l'approche d'inférence variationnelle. Toutes les proportions estimées sont données dans la Figure 3.2 et nous nous concentrons ici sur l'interprétation de ces résultats. Comme indiqué précédemment le cluster 3 rassemble les noeuds peu structurés du réseau. Les personnes associées à ce cluster à un moment t échangent des emails avec d'autres personnes du réseau, sans profil type de connexion. Notons que l'augmentation de la proportion de ce cluster coïncide avec la diminution de la proportion du cluster 2 (densité intra-cluster moyenne), quelques soient les sous-graphes et temps t, et inversement. Ces deux proportions renseignent

donc de manière inversée sur la structuration des échanges d'emails dans le réseau.

Nous observons également une chute importante de la proportion du cluster 3, dans tous les sous-graphes, entre t_4 et t_5 c'est à dire juste avant et après l'ouverture de l'enquête par l'organisme fédéral américain. Cette structuration du réseau est ici une réaction à la crise d'octobre 2001. Les personnes échangent des emails sur le sujet et contactent des personnes de manière préférentielle. Sur cette période, la proportion du cluster 4 (densité intra-cluster plus faible), comme celle du cluster 3, augmentent. Il est fondamental de noter que la structuration du réseau commence plus tôt (à t_3) chez les cadres que chez les employés et les autres. Il y a bien une légère réaction à t_3 pour ces deux derniers sous-graphes, mais elle disparaît à t_4. Ces observations laissent penser que les cadres ont eu connaissance des conditions d'arrivée de la crise avant les employés et les autres personnes considérées qui ont eu une légère réaction à t_3, ont été rassurés en t_4, et on finalement réagi à t_5, plusieurs mois plus tard.

Finalement, intéressons-nous au cluster 1 (densité intra-cluster forte). Pour les sous-graphes 2 et 3, la proportion de ce cluster a une tendance générale à diminuer jusqu'à (t_4, t_5) où au contraire elle augmente. Cette remarque va également dans le sens d'une structuration du réseau liée à l'ouverture de l'enquête. Les cadres sont les seuls individus du réseau pour lesquels nous observons au contraire une diminution de la proportion du cluster 1 à ce moment là. En d'autres termes, le noyau dur des cadres, où l'échange d'emails se fait de manière très préférentielle, se désolidarise du reste réseau.

4. Conclusion

Nous avons considéré dans ce travail le problème de l'analyse de réseaux dynamiques avec des arêtes catégorielles et pour lesquels une partition en sous-graphe est connue. Pour ce faire, nous avons proposé une extension au cadre dynamique du modèle RSM. Le nouveau modèle, appelé dRSM, est basé sur l'utilisation d'un *state-space model* pour modéliser l'évolution au cours du temps des proportions des groupes latents. Un algorithme de type EM variationnel (VEM) a été proposé pour en effectuer l'inférence. Nous avons en particulier montré que les approximations variationnelles conduisent à un nouveau *state-space model* à partir duquel les paramètres ainsi que les états cachés peuvent être estimés en utilisant le filtre de Kalman et le Rauch-Tung-Striebel (RTS) *smoother*. La méthodologie a été finalement appliquée au jeu des données e-mail de l'entreprise Enron et a permis de mettre en évidence une réaction anticipée des cadres par rapport aux autres employés concernant le scandale à venir. Concernant les travaux futurs, nous souhaitons nous intéresser

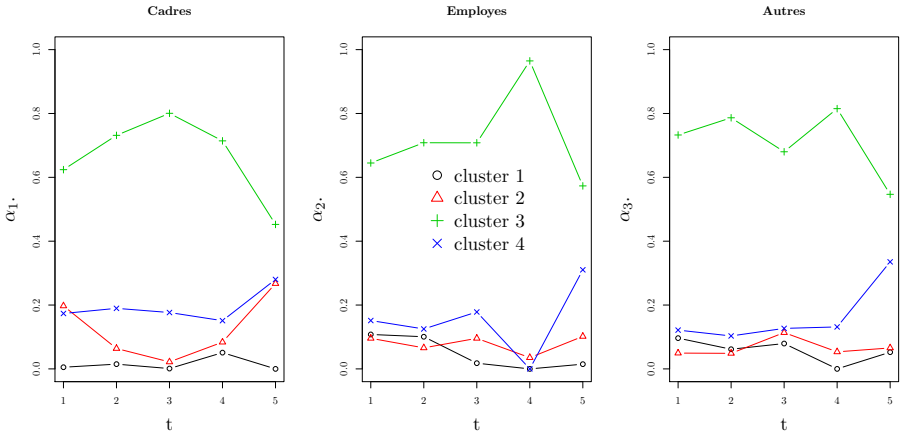

FIGURE 3. Proportions de chacun des $K = 4$ clusters, à chaque temps $t \in \{1, \ldots, 5\}$. Sous-graphe 1 (cadres), figure de gauche ; sous-graphe 2 (employés), figure du milieu ; sous-graphe 3 (autres), figure de droite.

au problème du choix du nombre K de groupes latents et cela pourrait être fait en utilisant une approche de choix de modèles.

Références

[1] A. AHMED & E. P. XING – « On tight approximate inference of logistic-normal admixture model », *In Proceedings of the Eleventh International Conference on Artifical Intelligence and Statistics. Omnipress, Madison, WI.* (2007).

[2] E. AIROLDI, D. BLEI, S. FIENBERG & E. XING – « Mixed membership stochastic blockmodels », *The Journal of Machine Learning Research* **9** (2008), p. 1981–2014.

[3] P. BICKEL & A. CHEN – « A nonparametric view of network models and newman–girvan and other modularities », *Proceedings of the National Academy of Sciences* **106** (2009), no. 50, p. 21068–21073.

[4] C. BISHOP – *Pattern recognition and machine learning*, Springer-Verlag, 2006.

[5] D. BLEI & J. LAFFERTY – « A correlated topic model of science », *The Annals of Applied Statistics* (2007), p. 17–35.

[6] J.-J. DAUDIN, F. PICARD & S. ROBIN – « A mixture model for random graphs », *Statistics and Computing* **18** (2008), no. 2, p. 173–183.

[7] S. FIENBERG & S. WASSERMAN – « Categorical data analysis of single sociometric relations », *Sociological Methodology* **12** (1981), p. 156–192.

[8] M. GIRVAN & M. NEWMAN – « Community structure in social and biological networks », *Proceedings of the National Academy of Sciences* **99** (2002), no. 12, p. 7821.

[9] M. Handcock, A. Raftery & J. Tantrum – « Model-based clustering for social networks », *Journal of the Royal Statistical Society : Series A (Statistics in Society)* **170** (2007), no. 2, p. 301–354.

[10] J. Hofman & C. Wiggins – « Bayesian approach to network modularity », *Physical review letters* **100** (2008), no. 25, p. 258701.

[11] Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou & S. Lamassé – « The random subgraph model for the abalysis of an acclesiasrical network in merovingian gaul », *Annals of Applied Statistics* (2013).

[12] M. Jordan, Z. Ghahramani, T. Jaakkola & L. K. Saul – « An introduction to variational methods for graphical models », *Machine learning* **37** (1999), no. 2, p. 183–233.

[13] P. Latouche, E. Birmelé & C. Ambroise – « Overlapping stochastic block models with application to the french political blogosphere », *Annals of Applied Statistics* **5** (2011), no. 1, p. 309–336.

[14] M. Mariadassou, S. Robin & C. Vacher – « Uncovering latent structure in valued graphs : a variational approach », *Annals of Applied Statistics* **4** (2010), no. 2, p. 715–742.

[15] C. Matias & S. Robin – « Modeling heterogeneity in random graphs through latent space models : a selective review », *HAL preprint hal-00948421v2* (2014).

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

[17] K. Nowicki & T. Snijders – « Estimation and prediction for stochastic blockstructures », *Journal of the American Statistical Association* **96** (2001), no. 455, p. 1077–1087.

[18] Y. Wang & G. Wong – « Stochastic blockmodels for directed graphs », *Journal of the American Statistical Association* **82** (1987), p. 8–19.

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

[20] E. Xing, W. Fu & L. Song – « A state-space mixed membership blockmodel for dynamic network tomography », *The Annals of Applied Statistics* **4** (2010), no. 2, p. 535–566.

Appendice A
Détails des calculs

A.1. Détails du calcul de la log-vraisemblance. — Nous détaillons ci-dessous le calcul des termes de la forme factorisée de $p(X, Z, \gamma, \nu|\theta)$. D'une part,

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

D'autre part, la variable Z_i à chaque t est telle que :

$$Z_i^{(t)}|f(\gamma_{s_{i,k}}^{(t)}) \sim \mathcal{M}(1, f(\gamma_{s_{i,k}}^{(t)})),$$

avec $f(\gamma_{s_{i,k}}^{(t)}) = \exp(\gamma_{s_{i,k}}^{(t)})/\sum_{l=1}^{K}\exp(\gamma_{s_{i,l}}^{(t)}) = \exp\{\gamma_{s_{i,k}}^{(t)} - \log(\sum_{l=1}^{K}\exp(\gamma_{s_{i,l}}^{(t)})\}$ $\in [0,1]$ et $\sum_{k=1}^{K} f(\gamma_{s_k}^{t}) = 1$. Par conséquent,

$$
\begin{aligned}
p(Z|f(\gamma)) &= \prod_{t=1}^{T} p(Z^{(t)}|f(\gamma^{(t)})) \\
&= \prod_{t=1}^{T}\prod_{i}^{N}\prod_{k=1}^{K} Y_{is} f(\gamma_{si,k}^{(t)})^{Z_{ik}^{(t)}} \\
&= \prod_{t=1}^{T}\prod_{k=1}^{K} Y_{is} f(\gamma_{si,k}^{(t)})^{\sum_{i=1}^{N} Z_{ik}^{(t)}}.
\end{aligned}
$$

De même, à chaque instant t, la variable latente $\nu^{(t)}$ est distribuée comme suit :

$$p(\nu^{(t)}|\nu^{(t-1)}, A, \Phi) = \mathcal{N}(\nu^{(t)}; A\nu^{(t-1)}, \Phi),$$

et

$$\gamma_s^{(t)} \sim \mathcal{N}(B\nu^{(t)}, \Sigma).$$

En conséquence, on a :

$$p(\gamma|\nu,\Sigma) = \prod_{t=1}^{T}\prod_{s=1}^{S}\mathcal{N}(\gamma_s^{(t)}, \Sigma).$$

Finalement,

$$
\begin{aligned}
\log p(X, Z, \gamma, \nu|\theta) &= \log p(X|Z, \Pi) + \log p(Z|f(\gamma) + \log p(\gamma|\nu, \sigma, B) + \log p(\nu|\mu_0, A, \Phi, V_0) \\
&= \sum_{t=1}^{T}\sum_{k,l}^{K}\sum_{c=1}^{T}\sum_{i\neq j}^{N} \delta(X_{i,j}^{(t)} = c) Z_{ik}^{(t)} Z_{jl}^{(t)} \log(\Pi_{kl}^{c}) \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N} Y_{is} Z_{ik}^{(t)} \log(f(\gamma_{si,k}^{(t)}) \\
&+ \sum_{t=1}^{T}\sum_{s=1}^{S} \log \mathcal{N}(\gamma_s^{(t)}; B\nu^{(t)}, \Sigma) \\
&+ \log p(\nu^{(1)}|\mu_0, V_0) + \sum_{t=2}^{T} \log p(\nu^{(t)}|\nu^{(t-1)}, A, \Phi).
\end{aligned}
$$

A.2. Détails des calculs de l'étape E du VEM. —

Distribution $q(Z)$:—

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

On reconnait alors la forme fonctionnelle d'une loi multinomiale :

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

Distribution $q(\nu)$:— On a :

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

On reconnait la forme fonctionnelle de la distribution d'un *state-space model* :

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

ou plus synthétiquement :

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

Concernant la constante de normalisation de la distribution, on peut démontrer qu'elle est dépendante du paramètre θ. Elle est en fait égale à la vraisemblance de x sachant θ. Ce résultat peut être obtenu par : $q(\nu)\propto p(x,\nu|\theta)$ pour la normalisation on a $\int q(\nu)d\nu = 1 = 1/C\int p(x,\nu|\theta)$ donc $C = p(x|\theta)$.

A.3. Détails des calculs de l'étape M du VEM. — Nous donnons ci-dessous le détail du calcul de la borne :

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

Comme,

$$E_\nu(\log q(\nu)) = E_\nu(\log p(\nu|\mu_0,A,\Phi,V_0))+E_\nu(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^t,\frac{\Sigma}{S},B)),$$

alors,

$$E_\nu(\log p(\nu|\mu_0,A,\Phi,V_0))-E_\nu(\log q(\nu)) = -E_\nu(\log p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^t,\frac{\Sigma}{S},B)),$$

et on peut réécrire $\tilde{\mathcal{L}}(q,\theta)$ sous la forme :

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

Nous explicitons ci-dessous chacun des termes de la borne $\tilde{\mathcal{L}}(q,\theta)$. Le terme de log-vraisemblance $\log p(x|\theta)$ du modèle SSM est quant à lui obtenu lors de l'inférence du *state-space model*.

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

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

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

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

où $r_s^{(t)}$ est un $\sum_{i=1}^N \tau_{ik}^{(t)} y_{i,s}$.

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

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

où $\hat{V}^{(t)}$ est la matrice de variance-covariance de la variable latente $\nu^{(t)}$ sachant toutes les variables observées.

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

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

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

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

Version 1, novembre 2014

R. Zreik, Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes
E-mail : rawyazreik@gmail.com • *Url :* http://www.sfds.webasso.fr/
P. Latouche, Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne
E-mail : pierre.latouche@univ-paris1.fr • *Url :* http://www.sfds.webasso.fr/

C. Bouveyron, Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes
E-mail : charles.bouveyron@parisdescartes.fr
Url : http://www.sfds.webasso.fr/