Intégration tardive de données multimodales par modèles à blocs stochastiques

Kylliann De Santiago [1] & Marie Szafranski [2] & Christophe Ambroise [3]

[1] *LaMME, France, kylliann.desantiago@univ-evry.fr*
[2] *ENSIIE, France, marie.szafranski@univ-evry.fr*
[3] *LaMME, France, christophe.ambroise@univ-evry.fr*

Résumé. Dans ce travail, nous présentons une méthode originale permettant d'agréger différentes sources d'information. Chaque partition est encodée par une matrice de co-appartenance des observations aux classes. Notre approche est fondée sur un mélange de modèles de blocs stochastiques multicouches pour conjointement définir des composantes de sources sur les matrices de co-appartenance similaires et partitionner les observations en différents groupes selon ces composantes. L'identifiabilité des paramètres du modèle est établie et un algorithme EM variationnel bayésien est proposé pour l'estimation de ces paramètres. Le cadre bayésien permet de sélectionner un nombre optimal de groupes et de composantes.

Mots-clés. Modèle à blocs stochastiques, apprentissage multivues, réseaux multicouches, cadre bayésien, vraisemblance complétée intégrée (ICL).

Abstract. In this work, we introduce an innovative method for aggregating multiple clusters originating from different sources of information. Each partition is represented by a co-membership matrix among observations. Our approach employs a mixture of multilayer Stochastic Block Models (SBM) to cluster co-membership matrices with similar information into components and to assign observations to distinct clusters, considering their specificities within the components. The identifiability of the model parameters is established, and a variational Bayesian EM algorithm is proposed for parameter estimation. The Bayesian framework facilitates the selection of an optimal number of clusters and components.

Keywords. Stochastic Block Model, Multiview clustering, Multilayer Network, Bayesian Framework, Integrated Classification Likelihood.

1 Introduction

La plupart des situations d'apprentissage font appel à différentes sources d'information, appelées ici *modalités* ou *vues*, telles que la vision, le toucher ou encore l'ouïe. Les objectifs de l'apprentissage machine multimodal ou multi-vues vont de l'apprentissage de nouvelles représentations, en passant par la traduction de textes ou encore la fusion d'information (cf. Zhao et al., 2017; Baltrušaitis et al., 2018; Cornuéjols et al., 2018, par exemple).

Les graphes constituent un moyen puissant et intuitif de représenter des systèmes complexes de relations entre individus. Ils fournissent une représentation efficace et informative

du système. La construction de graphes à partir de chaque vue permet d'utiliser l'apprentissage machine non supervisé dans une perspective de classification multimodale (Ektefaie et al., 2023).

Dans le cadre de la classification non supervisée, les algorithmes produisent souvent une partition ou une matrice d'appartenance $\mathbf{Z}$. Cette information, bien qu'utile, présente l'inconvénient de dépendre fortement du nombre de groupes choisi. Pour éviter ce problème, $\mathbf{Z}$ peut être transformé en une matrice d'adjacence $\mathbf{A}$ avec

$$A_{ij} = \begin{cases} 1, & \text{si les individus } i,\ j \text{ sont dans le même groupe,} \\ 0, & \text{sinon.} \end{cases}$$

Le terme de méta-classification(*meta clustering*) est utilisé pour désigner les approches permettant de combiner des partitions découvertes à l'aide de différentes méthodes. Parmi elles, la classification consensuelle (*consensus clustering*) est un algorithme de référence (Monti et al., 2003; Li et al., 2015; Liu et al., 2018). Lorsque ces approches sont fondées sur des modèles, elles permettent d'une part de construire un sous-typage final à partir des résultats déjà obtenus, mais aussi de mettre en lumière la redondance et / ou la complémentarité des sources d'information. En outre, utiliser un modèle permet aussi de disposer de critères d'évaluation de performance (log-vraisemblance, évidence, etc.) et, au moins dans le cadre bayésien, de critères de sélection de modèle (Biernacki et al., 2010).

Les modèles d'apprentissage varient en fonction de la stratégie de fusion de vues, précoce, intermédiaire ou tardive. Nous privilégierons ici la fusion tardive, où pour chaque vue, la matrice d'adjacence peut être obtenue par des algorithmes de classification efficaces et dédiés.

Contribution. À partir d'une classification indépendante de chaque vue, nous proposons d'apprendre une représentation coordonnée par le biais d'un modèle probabiliste. Notre modèle est un mélange de SBM multicouches associées à différentes sources d'information, avec une stratification des observations transversale comme illustré en Figure 1. L'algorithme associé est appelé mimi-SBM (De Santiago et al., 2024b, MIxture of Multilayer Integrator Stochastic Block Model). De plus, grâce au cadre bayésien, il est possible de développer un critère de sélection du modèle issu de la borne inférieure de l'évidence, à la fois pour le mélange de vues et le nombre de groupes. Enfin, l'identifiabilité des paramètres du modèle est établie et un algorithme EM bayésien variationnel est proposé pour estimer les paramètres.

2 Modèle de mélange de SBM multicouches

Notre modèle s'appuie sur un SBM avec deux ensembles de variables latentes correspondant respectivement à la structure des observations et à la structure des vues. Cette proposition se situe à la croisée du *Multilayer SBM* (MLSBM), qui recherche une matrice de co-appartenance traversante des observations sur toutes les couches, et du *Mixture of Multilayer SBM* (MMLSBM), qui recherche des motifs structurels au sein des couches.

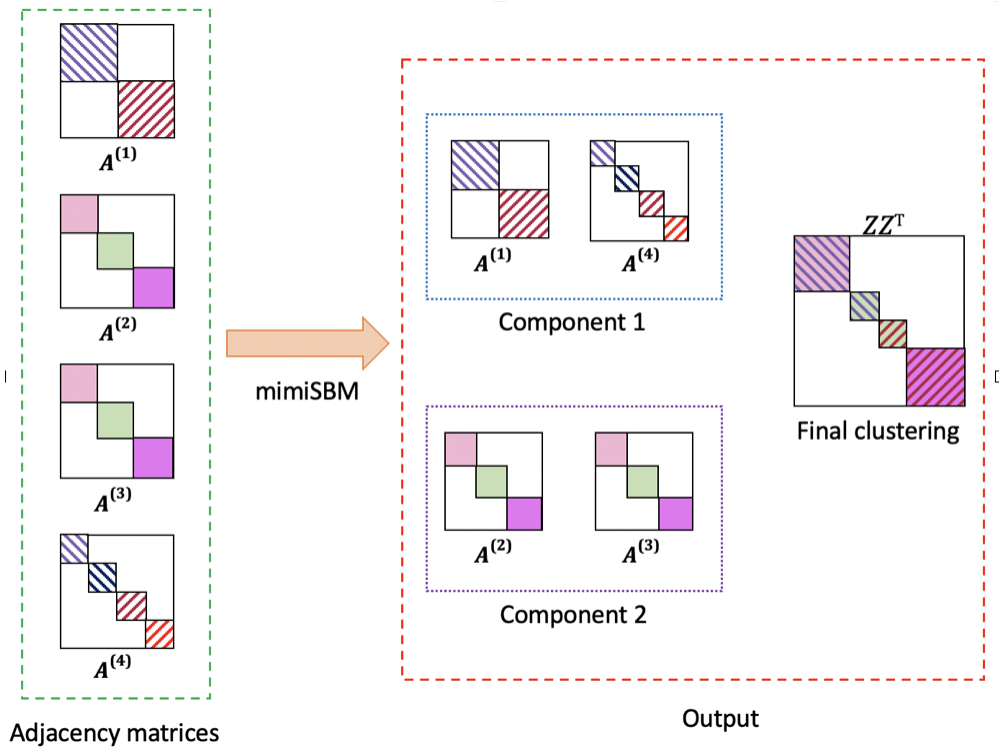

FIGURE 1 – Illustration de mimi-SBM. Gauche : 4 matrices d'adjacences $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(4)}$ provenant de quatre vues différentes organisées en deux composantes. Droite : identification des deux composantes à partir des vues (informations locales et complémentaires) et partition des observations décrites par la matrice d'appartenance $\mathbf{Z}$ (information globale et consensus).

Observations. Nous considérons, $\mathbf{A} \in \{0,1\}^{N \times N \times V}$, une représentation tensorielle des observations, où N est le nombre d'observations (nœuds) et V est le nombre de vues. Chacune des V tranches de $\mathbf{A}$ est une matrice d'adjacence correspondant à un graphe $\mathcal{G}^v$. Le tenseur est donc un empilement de matrices d'adjacence relié aux graphes multi-vues $(\mathcal{G}^1, \cdots, \mathcal{G}^V)$ et nœuds correspondants. Soit (i,j), une arête entre les observations i et j. Par définition, $A_{ijv} = \mathbb{I}_{((i,j) \in E^v)}$, où E^v est l'ensemble des arêtes du graphe $\mathcal{G}^v$.

Structures latentes. Soit $\mathbf{Z} \in \{0,1\}^{N \times K}$ la matrice d'appartenance des nœuds, où K est le nombre de groupes traversant les vues. Pour le nœud i et le groupe k, $Z_{ik} = \mathbb{I}_{(i \in k)}$. Notons aussi $\mathbf{W} \in \{0,1\}^{V \times Q}$, la matrice d'appartenance des vues, où Q est le nombre de composantes du mélange des vues. Ainsi pour la vue v et la composante s, $W_{vs} = \mathbb{I}_{(v \in s)}$.

2.1 Un mélange d'observations à travers un mélange de vues

Nous supposons que les V vues sont générées par un modèle de mélange de Q composantes, où chaque composante s est un SBM. Nous supposons également que chaque ligne de la

matrice $\mathbf{W}$ suit une distribution multinomiale, $\mathbf{W}_v \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q))$, avec

$$\mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho}) = \prod_{v=1}^{V} \prod_{s=1}^{Q} \rho_v^{W_{vs}}. \tag{1}$$

Nous supposons aussi une *structure traversante* à *toutes les vues* décrite par la variable latente $\mathbf{Z}$. En exploitant toutes les sources d'information disponibles, notre objectif est d'obtenir des composantes cohérentes sur l'ensemble des vues. On suppose ainsi que les individus proviennent d'un nombre K de sous-populations. Chaque vecteur de classe latente pour l'observation i suit une distribution multinomiale, avec $\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K))$, et

$$\mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{Z_{ik}}. \tag{2}$$

Enfin, chaque observation A_{ijv} conditionnellement à la structure latente $\mathbf{Z}$ suit une distribution de Bernoulli : $A_{ijv} \mid Z_{ik} = 1, Z_{jl} = 1 W_{vs} = 1 \sim \mathcal{B}(\alpha_{kls})$. La probabilité de toutes les observations sachant les variables latentes $\mathbf{Z}$, $\mathbf{W}$, et un vecteur de paramètres $\boldsymbol{\Theta}$, est donc

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\Theta}) = \prod_{\substack{i=1,\\ i<j}}^{N} \prod_{\substack{k=1\\ l=1}}^{K} \prod_{v=1}^{V} \prod_{s=1}^{Q} \left(\alpha_{kls}^{A_{ijv}} \left(1 - \alpha_{kls}\right)^{1 - A_{ijv}} \right)^{Z_{ik} Z_{jl} W_{vs}}. \tag{3}$$

2.2 Identifiabilité

Théorème 1 *Soient $N \geq \max(2K, 4Q)$ et $V \geq 2K$. Supposons que pour tout $k, l \in \{1, \ldots, K\}$ et chaque $s \in \{1, \ldots, Q\}$, les coordonnées de $\left(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{k\cdot\cdot} \boldsymbol{\rho}\right)$ sont toutes différentes, $(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{\cdot\cdot s} \boldsymbol{\pi})_{s=1:Q}$ sont distinctes, et chaque $(\boldsymbol{\alpha}_{kl\cdot} \boldsymbol{\rho})_{k,l=1:K}$ est différent. Alors, les paramètres du mimi-SBM $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ sont identifiables.*

Preuve 1 *La preuve de ce théorème est disponible dans (De Santiago et al., 2024a).*

2.3 Formulation bayesienne

Les modèles bayésiens offrent un cadre naturel pour incorporer des connaissances *a priori* permettant d'améliorer la précision de la structure estimée, en particulier lorsque les données disponibles sont limitées ou bruitées. Dans ce contexte, la définition des distributions conjuguées choisies à la fois pour la proportion du mélange et les proportions des blocs s'appuie sur les travaux de Latouche et al. (2012). Les lois *a priori* conjuguées conduisent à des distributions *a posteriori* explicites, où Dir(.) désigne la loi de Dirichlet :

$$\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_K^0)) \;=\; \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}^0), \tag{4}$$

$$\mathbb{P}(\boldsymbol{\rho} \mid \boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_Q^0)) \;=\; \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}^0), \tag{5}$$

$$\mathbb{P}(\boldsymbol{\alpha} \mid \boldsymbol{\eta}^0 = (\eta_{kls}^0), \boldsymbol{\xi}^0 = (\xi_{kls}^0)) \;=\; \prod_{k,k<l} \prod_{s} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}^0, \xi_{kls}^0). \tag{6}$$

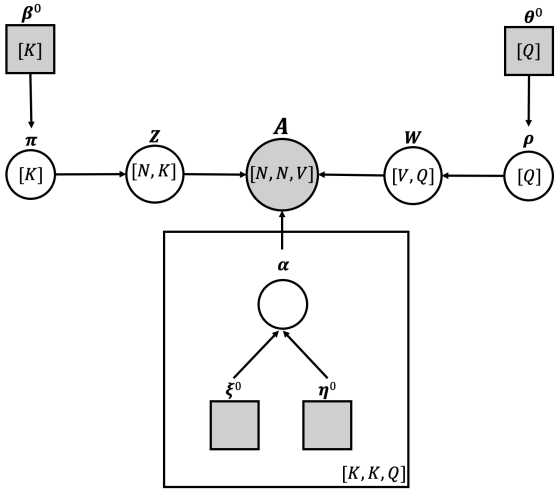

FIGURE 2 – Illustration du mimi-SBM avec les notations bayesiennes.

Les paramètres $\boldsymbol{\beta}^0, \boldsymbol{\theta}^0, \boldsymbol{\eta}^0, \boldsymbol{\xi}^0$ sont choisis selon les lois *a priori* de Jeffreys, qui sont souvent considérées comme non informatives ou faiblement informatives. Elles n'introduisent pas d'hypothèses *a priori* fortes ou de biais marqués dans l'analyse.

Pour la distribution de Dirichlet, fixer β_k^0 et θ_s^0 à $1/2$ revient à considérer une distribution *a priori* de Jeffreys. De même, pour la distribution Beta, η_{kls}^0 et ξ_{kls}^0 peuvent être choisis comme étant égaux à $1/2$ pour tous les indices k, l, et s correspondants.

3 Algorithme EM variationnel pour le mimi-SBM

La vraisemblance marginale dans les modèles de blocs stochastiques s'exprime par :

$$\mathbb{P}(\mathbf{A}) = \sum_{\mathbf{Z}} \sum_{\mathbf{W}} \int \int \int \mathbb{P}(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \; d\boldsymbol{\alpha} \; d\boldsymbol{\pi} \; d\boldsymbol{\rho}. \tag{7}$$

Le calcul des intégrales de cette vraisemblance marginale ne présente pas de solution analytique. De plus, les sommes sur $\mathbf{Z}$ et $\mathbf{W}$ deviennent très difficiles à calculer lorsque le nombre de paramètres ou d'observations est conséquent. Pour contourner cela, on utilise généralement l'approximation de distributions postérieures complexes réalisée soit par échantillonnage (Monte-Carlo par chaînes de Markov ou des approches similaires), soit par l'inférence bayésienne variationnelle introduite par Attias (1999).

3.1 Borne inférieure de l'évidence

L'inférence variationnelle est efficace du point de vue computationnel et extensible pour les grands ensembles de données, et elle fonctionne particulièrement bien pour les modèles SBM. Le problème est formulé comme une tâche d'optimisation, où l'objectif est de trouver la meilleure approximation de la véritable distribution postérieure. Ce cadre d'optimisation

permet un calcul efficace des paramètres variationnels en maximisant une borne inférieure de la log-vraisemblance, connue sous le nom de borne inférieure de l'évidence (ELBO). Des techniques d'optimisation telles que la descente de gradient stochastique (SGD) ou l'algorithme d'EM peuvent être utilisées pour trouver les paramètres variationnels optimaux.

La distribution $\mathbb{P}(\mathbf{Z}, \mathbf{W}|\mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})$ étant incalculable dans le cadre des SBM, nous allons approcher l'ensemble de cette distribution. Soit une distribution variationelle q sur $\{\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\}$, nous pouvons décomposer la log-vraisemblance marginale en deux parties : la borne inférieure de l'Évidence (ELBO) et la divergence de Kullback-Leibler **KL** entre la distribution variationnelle et la distribution postérieure :

$$\begin{aligned} \log P(\mathbf{A}) &= \mathbb{E}_q\left[\log \frac{P\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{P(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}|\mathbf{A})}\right] \quad (8) \\ &= \underbrace{\mathbb{E}_q\left[\log \frac{P\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}\right]}_{ELBO=\mathcal{L}(q(.))} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}{P(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho} \mid \mathbf{A})}\right]}_{\mathbf{KL}(q(\cdot)\|\,\mathbb{P}(\cdot|\mathbf{A}))} \quad (9) \end{aligned}$$

où $\mathbf{KL}\left(q(.) \mid \mathbb{P}(\cdot \mid \mathbf{A})\right) = -\mathbb{E}_q\left[\log \frac{p}{q}\right] \geq -\log \mathbb{E}_q\left[\frac{p}{q}\right] \geq 0$ selon l'inégalité de Jensen.

L'ELBO est donnée par

$$\mathcal{L}\left(q(\cdot)\right) = \sum_{\mathbf{Z},\mathbf{W}} \int\int\int q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \log \frac{p\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}\, d\boldsymbol{\alpha}\, d\boldsymbol{\pi}\, d\boldsymbol{\rho}. \quad (10)$$

La distribution variationnelle est généralement choisie dans une famille de distributions plus facile à manipuler, comme la famille exponentielle. Les paramètres de la distribution variationnelle sont ensuite ajustés pour réduire la divergence de Kullback-Leibler par rapport à la distribution postérieure. Si $q(\cdot)$ est exactement égale à $p(\cdot|\mathbf{A})$, le terme **KL** est égal à 0, et l'ELBO est maximisée. Par une approximation du champ moyen, on définit $q(\cdot)$ comme :

$$\begin{aligned} q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) &= \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{v=1}^{V} q(\mathbf{W}_v) \prod_{s=1}^{Q} \prod_{k,k\leq l}^{K} q(\alpha_{kls})\; q(\boldsymbol{\pi})\; q(\boldsymbol{\rho}), \\ &= \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta})\; \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}) \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\tau}_i) \prod_{v=1}^{V} \mathcal{M}(\mathbf{W}_i; 1, \boldsymbol{\nu}_v) \quad (11) \\ &\quad \prod_{s=1}^{Q} \prod_{k,k\leq l}^{K} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}, \xi_{kls}), \end{aligned}$$

où τ_{ik} (resp. ν_{vs}) sont les paramètres variationels indiquant la probabilité que l'individu i (resp. une vue v) appartienne au groupe k (resp. à la composante s).

D'après (10), avec $\Gamma(.)$ désignant la fonction gamma et étant donné la distribution $q(\cdot)$

choisie, l'ELBO est donné par

$$\begin{aligned}\mathcal{L}\left(q(\cdot)\right) = &\log\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\} \\ &+ \sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\} \\ &- \sum_{i}^{N}\sum_{k}^{K}\tau_{ik}\log\tau_{ik} \; - \sum_{v}^{V}\sum_{s}^{Q}\nu_{vs}\log\nu_{vs}.\end{aligned} \tag{12}$$

Cette expression, également appelée *Integrated Likelihood variational Bayes* (ILvb, Latouche et al., 2012)), peut être utilisée pour la sélection de modèle.

3.2 Optimisation des paramètres de la borne inférieure

Nous utilisons un algorithme EM variationnel bayésien pour estimer les paramètres (cf. Algorithme 1). L'algorithme débute par l'initialisation des paramètres du modèle, puis effectue de manière itérative deux étapes : l'étape d'Espérance Variationnelle Bayésienne (étape VBE) et l'étape de Maximisation (étape M).

Dans l'étape VBE, les distributions variationnelles $q(\mathbf{Z}_i)$ et $q(\mathbf{W}_v)$ sont optimisées sur les variables latentes $\forall i \in \{1,\ldots,N\}$ et $\forall v \in \{1,\ldots,V\}$ afin d'approximer la vraie distribution postérieure. Dans l'étape M, les paramètres du modèle $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ et $\boldsymbol{\xi}$ sont mis à jour pour maximiser une borne inférieure sur la log-vraisemblance, sachant les paramètres calculés lors de l'étape VBE.

Il existe plusieurs techniques pour initialiser l'algorithme EM. Une approche prévalente consiste à initialiser aléatoirement les paramètres selon une distribution choisie. Néanmoins, cette méthode peut manquer de fiabilité et ne pas fournir de valeurs de départ satisfaisantes pour l'algorithme. Dans Stanley et al. (2016), les paramètres (τ_{ik}) et (ν_{vs}) sont initialisés avec les résultats d'un modèle de blocs stochastiques appliqué séparément sur chaque vue. Pour initialiser mimi-SBM, nous combinons, avec l'algorithme des K-means, les résultats de SBM appliqués indépendamment sur les V vues.

4 Sélection de modèle

Dans le contexte de la classification, la sélection de modèle fait souvent référence au processus de détermination du nombre idéal de groupes pour un ensemble de données donné. Dans notre situation, la décision clé réside dans la sélection des valeurs appropriées pour K et Q afin de trouver un équilibre entre les performances et la complexité du modèle. Pour cela, plusieurs critères basés sur la log-vraisemblance pénalisée peuvent être utilisés, tels que le Critère d'Information d'Akaike (AIC, Akaike, 1998), le Critère d'Information Bayésien (BIC, Schwarz, 1978) et plus récemment le critère de Vraisemblance Complétée Intégrée (ICL,

Algorithme 1 mimi-SBM

Require: Tenseur $\mathbf{A}$, Nombre de groupes K, Nombre de composantes Q, précision *eps*.
Initialisation : $\tau_{ik}^{(old)}$ et $\nu_{vs}^{(old)}$
while $\|\mathcal{L}\left(q^{new}(.)\right) - \mathcal{L}\left(q^{old}(.)\right)\| < eps$ **do**
VBE-step
Calculer $\tau_{ik}^{(new)}$ $\forall i \in \{1, \ldots, N\}$ et $\forall k \in \{1, \ldots, K\}$
Calculer $\nu_{vs}^{(new)}$ $\forall v \in \{1, \ldots, V\}$ et $\forall s \in \{1, \ldots, Q\}$
M-step
Optimiser $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ sachant les paramètres $(\tau_{ik}^{(new)})$ et $(\nu_{vq}^{(new)})$
ELBO
Calculer $\mathcal{L}\left(q^{new}(.)\right)$
end while

Biernacki et al., 2000). Nous considérons spécifiquement le critère ICL et ses pénalisations associées bien adapté aux modèles de mélange (Biernacki et al., 2010).

L'ICL est fondé sur la log-vraisemblance intégrée des paramètres sur les données complètes. De plus, si nous supposons l'indépendance des paramètres de la probabilité de connexion des composantes-groupes $\boldsymbol{\alpha}$, des paramètres du mélange de communautés $\boldsymbol{\pi}$ et des paramètres du mélange de vues $\boldsymbol{\rho}$, alors

$$\begin{aligned} \mathrm{ICL}(\mathbf{A}, K, Q) &= \log \mathbb{P}(\mathbf{A}, \mathbf{Z}, \mathbf{W} \mid K, Q), \\ &= \log \int_{\boldsymbol{\alpha}} \mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha})\, \mathbb{P}(\boldsymbol{\alpha}) d\boldsymbol{\alpha} + \log \int_{\boldsymbol{\pi}} \mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi})\, \mathbb{P}(\boldsymbol{\pi}) d\boldsymbol{\pi}, \\ &\quad + \log \int_{\boldsymbol{\rho}} \mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho})\, \mathbb{P}(\boldsymbol{\rho}) d\boldsymbol{\rho}. \end{aligned} \tag{13}$$

Dans le cadre variationnel, $\mathbf{Z}$ et $\mathbf{W}$ doivent être estimés. Ainsi, $\hat{\mathbf{Z}}$ (resp. $\hat{\mathbf{W}}$) peut être choisi directement comme le vecteur des paramètres variationnels $\boldsymbol{\tau}$ (resp. $\boldsymbol{\nu}$) ou par un Maximum a Posteriori (MAP) :

$$\hat{\mathbf{Z}}_i = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ \tau_{ik}.$$

En utilisant des approximations telles que la formule d'approximation de Stirling sur $\mathbb{P}(\boldsymbol{\pi})$ et $\mathbb{P}(\boldsymbol{\rho})$ et l'approximation asymptotique de Laplace sur $\mathbb{P}(\boldsymbol{\alpha})$, nous pouvons définir un *ICL approximatif* :

$$\begin{aligned} \mathrm{ICL}(\mathbf{A}, K, Q) &\approx \log\left(\mathbb{P}(\mathbf{A}, \hat{\mathbf{Z}}, \hat{\mathbf{W}} \mid K, Q)\right) - \mathrm{pen}(K, Q), \\ &\approx \mathcal{L}\left(q(.)\right) - \mathrm{pen}(K, Q). \end{aligned} \tag{14}$$

où $\mathrm{pen}(K, Q) = \dfrac{1}{2}\dfrac{K(K+1)}{2} Q \log\left(V \dfrac{N(N-1)}{2}\right) + \dfrac{1}{2}(K-1)\log(N) + \dfrac{1}{2}(Q-1)\log(V)$.

Rappelons d'abord que notre modèle est fondé sur des matrices d'adjacence non orientées (symétriques) où seules les matrices triangulaires supérieures sans la diagonale sont prise

en compte. L'*ICL approximatif* (14) est composé d'une partie dépendant du nombre de paramètres de $\boldsymbol{\alpha}$ et du nombre d'arêtes des graphes associés aux matrices d'adjacence et d'une autre partie fonction du nombre de degrés de liberté dans les paramètres des mélanges ($\boldsymbol{\pi}$, $\boldsymbol{\rho}$) et du nombre de variables liées.

Dans le cadre bayésien, avec des lois *a priori* conjuguées, il est possible de définir un ICL exact (Côme and Latouche, 2015). Il peut être obtenu à partir de l'ILvb (12) lorsque l'entropie des variables latentes est nulle et que l'algorithme EM est un *Classification EM* (CEM, Celeux and Govaert, 1992). En d'autres termes, les paramètres variationnels sont égaux à 1 s'ils sont le MAP et 0 sinon. Ainsi, cet *ICL exact* peut être défini comme :

$$\begin{aligned}\mathrm{ICL}_{\mathrm{exact}}(\mathbf{A}, K, Q) = \log\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\} \\ + \sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\}. \qquad (15)\end{aligned}$$

Par ailleurs, il est possible d'utiliser directement les paramètres variationnels, à la place du CEM, et de dériver ainsi un *ICL variationnel* à partir du critère précédent.

Conclusion

Cet article décrit une nouvelle méthode de modèle de mélange de SBM multicouches et son algorithme associé *mimi-SBM*. Afin d'obtenir une borne inférieure de l'évidence calculable, une approche variationelle bayésienne a été utilisée, où chaque paramètre du modèle est estimé à l'aide d'un algorithme EM bayésien variationnel. De plus, le cadre bayésien permet de développer une stratégie de sélection de modèle. Enfin, une preuve de l'identifiabilité des paramètres est établie dans la pré-publication soumise de De Santiago et al. (2024a).

Références

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Hagai Attias. A variational baysian framework for graphical models. *Advances in neural information processing systems*, 12, 1999.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning : A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2) : 423–443, 2018.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7) :719–725, 2000.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Exact and monte carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11) :2991–3002, 2010.

Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3) :315–332, 1992.

Etienne Côme and Pierre Latouche. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6) :564–589, 2015.

Antoine Cornuéjols, Cédric Wemmert, Pierre Gançarski, and Younès Bennani. Collaborative clustering : Why, when, what and how. *Information Fusion*, 39 :81–95, 2018.

Kylliann De Santiago, Marie Szafranski, and Christophe Ambroise. Mixture of multilayer stochastic block models for multiview clustering. *arXiv preprint arXiv :2401.04682*, 2024a.

Kylliann De Santiago, Marie Szafranski, and Christophe Ambroise. mimisbm : Mixture of multilayer integrator stochastic block models. CRAN, Package R, 2024b. https ://cran.r-project.org/package=mimiSBM.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4) :340–350, 2023.

P. Latouche, É. Birmele, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1) :93–115, 2012.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Liangchen Liu, Feiping Nie, Arnold Wiliem, Zhihui Li, Teng Zhang, and Brian C Lovell. Multi-modal joint clustering with application for unsupervised attribute discovery. *IEEE Transactions on Image Processing*, 27(9) :4345–4356, 2018.

S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering : a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52 :91–118, 2003.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Natalie Stanley, Saray Shai, Dane Taylor, and Peter J Mucha. Clustering network layers with the strata multilayer stochastic block model. *IEEE transactions on network science and engineering*, 3(2) :95–105, 2016.

J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview : Recent progress and new challenges. *Information Fusion*, 38 :43–54, 2017.