université
PARIS-SACLAY

From stratification to prediction: multimodal machine learning with Latent Block Models and Mixtures of Experts

De la stratification à la prédiction : apprentissage automatique multimodal par modèles à blocs latents et mélanges d'experts

Thèse de doctorat de l'université Paris-Saclay

École doctorale n◦ 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques. Référent : Université d'Évry Val d'Essonne

Thèse préparée dans le Laboratoire de Mathématiques et Modélisation d'Evry (Université Paris-Saclay, CNRS, Univ Evry) sous la direction de Christophe AMBROISE, professeur, le co-encadrement de Marie SZAFRANSKI, maîtresse de conférences, et de Guillaume ANDEOL, médecin en chef

Thèse soutenue à Paris-Saclay, le 28 janvier 2025, par

Kylliann DE SANTIAGO

Composition du Jury

Membres du jury avec voix délibérative

Christine KERIBIN Professeure, Université Paris-Saclay	Présidente
Pierre LATOUCHE Professeur, Université Clermont-Auvergne	Rapporteur & Examinateur
Mohamed NADIF Professeur, Université Paris Cité	Rapporteur & Examinateur
Tabea REBAFKA Professeure, AgroParisTech	Examinatrice

THESE DE DOCTORAT

NNT : 2025UPASM001

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre : De la stratification à la prédiction : apprentissage automatique multimodal par modèles à blocs latents et mélanges d'experts

Mots clés : Apprentissage multimodal - Modèle à blocs stochastique - Modèle à blocs latents - Mélanges d'experts - Sélection de modèle

Résumé : Cette thèse explore l'application de méthodes d'apprentissage automatique multimodales pour l'analyse de données médicales, en mettant l'accent sur la stratification des patients et la prédiction de la récupération auditive après un traumatisme sonore aigu. L'étude repose sur des données hétérogènes (audiologiques, génomiques et protéomiques) collectées à différents moments après le traumatisme. L'objectif principal est d'extraire des caractéristiques pertinentes en combinant ces données multimodales, afin de permettre une analyse plus précise du comportement individuel des patients et des tendances globales. Dans un premier temps, les problématiques de l'apprentissage multimodal et les particularités de la fusion des données sont abordées. Ensuite, un modèle de fusion tardive basé sur les modèles à blocs stochastiques est développé. Ce modèle permet de caractériser la redondance et la complémentarité de l'information disponible : (i) en regroupant les différentes sources en composantes, (ii) en maintenant une stratification globale des individus, permettant ainsi la définition de communautés. Par ailleurs, l'utilisation de l'approche bayésienne permet de mettre en œuvre une méthode de sélection de modèle. Enfin, un modèle de fusion intermédiaire est proposé, étendant le cadre des Mixture of Experts en intégrant une modélisation par modèle à blocs latents conditionnels des entrées. L'objectif est de réduire la complexité algorithmique en résumant les variables par composantes, tout en préservant l'interprétabilité et en assurant de bonnes performances de prédiction.

Title : From stratification to prediction: multimodal machine learning with latent block models and mixtures of experts

Keywords : Multimodal machine learning - Stochastic block model - Latent block model - Mixture of experts - Model selection

Abstract: This thesis explores the application of multimodal machine learning techniques for the analysis of medical data, with a focus on patient stratification and the prediction of hearing recovery after acute sound trauma. The study relies on heterogeneous data (audiological, genomic, and proteomic) collected at various time points following the trauma. The main objective is to extract relevant features by combining these multimodal data, thus enabling a more accurate analysis of individual patient behavior and global trends. First, the challenges of multimodal learning and the specificities of data fusion are addressed. Next, a late fusion model based on stochastic block models is developed. This model allows the characterization of redundancy and complementarity of the available information by (i) grouping the different sources into components, and (ii) maintaining a global stratification of individuals, thereby defining communities. Moreover, the use of a Bayesian approach enables the implementation of a model selection method. Finally, an intermediate fusion model is proposed, extending the Mixture of Experts framework by incorporating conditional latent block modeling of the inputs. The objective is to reduce algorithmic complexity by summarizing variables into components while preserving interpretability and ensuring good predictive performance.

université
evry
val-d'essonne

université
PARIS-SACLAY

LaMME

Laboratoire de
Mathématiques
et Modélisation
d'Évry

IRBA

Sensorion

Remerciements

Il est difficile de trouver les mots justes pour exprimer toute la gratitude que j'éprouve envers celles et ceux qui ont rendu cette thèse possible. Ce travail, fruit de plusieurs années de recherche, d'échanges et de remises en question, n'aurait pu voir le jour sans le soutien indéfectible de nombreuses personnes et institutions que je tiens à remercier ici.

Avant tout, je souhaite exprimer ma profonde reconnaissance à mon directeur de thèse Christophe, qui m'a accordé sa confiance dès mon Master 1 et qui a soutenu ce projet de thèse bien avant qu'il ne prenne forme officiellement. Son engagement, sa bienveillance et son expertise ont été des repères constants tout au long de ce parcours. Sa patience et sa rigueur scientifique ont été des guides essentiels, et son exigence intellectuelle m'a permis de progresser bien au-delà de mes espérances. Nos discussions stimulantes, parfois animées, ont toujours été une source d'inspiration, et son soutien constant m'a encouragé à repousser mes limites et à affiner ma démarche de recherche. Son humilité et son humanité ont fait de ces années de collaboration une expérience précieuse dont je garderai un souvenir impérissable.

Je tiens également à exprimer ma gratitude infinie à ma co-encadrante de thèse, Marie, dont le soutien indéfectible a été une véritable boussole durant ces années. Toujours présente, d'une écoute incroyable et d'une aide si précieuse, elle a su m'accompagner avec une bienveillance rare. Sa disponibilité et ses conseils ont été d'une valeur inestimable, me permettant de surmonter les difficultés et d'avancer avec sérénité dans mon travail. Je lui dois une grande partie de l'équilibre et de la persévérance qui m'ont permis d'aller au bout de cette aventure.

J'ai également eu la chance d'être co-encadré par Guillaume, dont l'apport a été fondamental pour donner une dimension concrète à mes travaux de recherche. Son expertise médicale et son regard appliqué m'ont permis de comprendre les véritables enjeux des outils développés durant cette thèse. Grâce à lui, j'ai pu mieux appréhender les besoins du terrain et percevoir comment mes recherches pouvaient, à terme, contribuer à améliorer la prise en charge des patients. Son approche pragmatique a été un complément essentiel à l'approche théorique et algorithmique de mon travail.

Mes remerciements vont aussi aux membres du jury qui ont accepté d'évaluer ce travail. Leur expertise et leur bienveillance m'honorent, et leurs observations éclairantes m'aideront à poursuivre mes recherches avec un regard toujours plus affûté. Merci à Pierre Latouche et Mohamed Nadif pour le temps consacré à la lecture de ce manuscrit et pour leurs retours pertinents, ainsi qu'à Tabea Rebafka et Christine Keribin pour leur présence, leur engagement lors de la soutenance. Je tiens à remercier toute l'équipe de Sensorion, pour avoir rendu cette thèse CIFRE possible. Leur confiance et leur soutien ont été déterminants dans la concrétisation de ce projet. L'opportunité de mener cette recherche en collaboration avec le monde industriel m'a apporté une richesse d'expérience inestimable et m'a permis de relier plus étroitement la recherche fondamentale aux besoins du terrain. L'encadrement de Viviana a été un véritable moteur dans ce projet, et son accompagnement a joué un rôle clé dans le développement et l'aboutissement de mes travaux.

Je souhaite également exprimer ma gratitude à mes collègues et amis, avec qui j'ai partagé ces années intenses de recherche. Ces compagnons de route ont rendu cette aventure bien plus enrichissante, en mêlant sérieux scientifique et moments de partage. Leurs encouragements, nos discussions passionnées et les instants de convivialité ont été d'un soutien inestimable. Un immense merci à Claire, Ludmila, Ludivine, Arnaud, Vincent, Cyril, Carène, Franck, Margot, et Maurice pour tous ces bons moments, ainsi qu'à toute l'équipe du LaMME de manière générale. Enfin, je voudrais adresser mes plus sincères remerciements à ma famille et à mes proches. Leur soutien indéfectible, leur patience et leur compréhension ont été des piliers sur lesquels je me suis appuyé tout au long de cette aventure. À Mathilde, pour sa présence constante, son soutien et pour avoir partagé avec moi les joies et (surtout) les épreuves de cette quête intellectuelle. À Céline, Stéphane, Natasha, Simone, André et Carole pour toutes ces années à mes côtés, à m'épauler sur toutes mes décisions.

À vous tous, qui avez contribué à cette aventure, par un mot, un regard, une idée ou simplement votre présence, je vous adresse ma plus profonde gratitude. Cette thèse est autant la vôtre que la mienne.

Contributions scientifiques

Publications

- Courbariaux, M., De Santiago, K., Dalmasso, C., Danjou, F., Bekadar, S., Corvol, J.-C., Martinez, M., Szafranski, M., and Ambroise, C. (2022). A sparse mixture-of-experts model with screening of genetic associations to guide disease subtyping. Frontiers in Genetics, 13, 859462. https://doi.org/10.3389/fgene.2022.859462

- De Santiago, K., Szafranski, M., and Ambroise, C. (2023). Mixture of stochastic block models for multiview clustering. In ESANN 2023 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (pp. 151–156).

- De Santiago, K., Szafranski, M., and Ambroise, C. (2024). Mixture of multilayer stochastic block models for multiview clustering. arXiv preprint arXiv:2401.04682.

- De Santiago, K., Szafranski, M., and Ambroise, C. (2024, May). Intégration tardive de données multimodales par modèles à blocs stochastiques. In 55e Journées de Statistique de la SFdS.

Bibliothèque logicielle

- De Santiago, K., Szafranski, M., and Ambroise, C. (2024). mimiSBM: Mixture of multilayer integrator stochastic block models (R package version 0.0.1.3). https://CRAN.R-project.org/package=mimiSBM

Introduction

Motivation

L'audiologie est la science de l'audition et des troubles qui y sont associés. Les problèmes audiologiques peuvent varier selon le type de perte auditive: perte brutale d'audition, exposition au bruit, surdité de naissance, etc. Leur prise en charge nécessite souvent une approche holistique et personnalisée car les profils audiologiques sont variés, contextuels, et dépendent de l'objectif fixé : entendre "normalement", récupérer une audition dans le bruit ou récupérer un certain seuil d'audition par exemple. Dans ce cadre, l'analyse de données multimodales — issues de différentes sources d'information — est essentielle pour améliorer la précision des diagnostics et l'efficacité des traitements.

Cette thèse présente des modèles potentiellement utilisables pour mener une étude audiologique fondée sur des données multimodales telles que des mesures d'audiométrie tonale, vocale, des mesures biochimiques ou des questionnaires. L'objectif est de développer des outils d'apprentissage machine pour étudier une cohorte de patients ayant subi un traumatisme sonore suite à une exposition au bruit.

Les données

L'oreille humaine se divise en trois parties distinctes : l'oreille externe, l'oreille moyenne et l'oreille interne, illustrées à la Figure 1. L'oreille externe, composée du pavillon et du conduit auditif, capte et concentre les ondes sonores vers le tympan, une fine membrane située à la jonction avec l'oreille moyenne. Les vibrations du tympan sont transmises par une chaîne d'os minuscules (le marteau, l'enclume et l'étrier) dans l'oreille moyenne, qui amplifient ces signaux et les acheminent vers l'oreille interne. Celle-ci abrite la cochlée, une structure spiralée remplie de liquide, où les cellules ciliées convertissent les vibrations mécaniques en impulsions électriques transmises au cerveau via le nerf auditif.

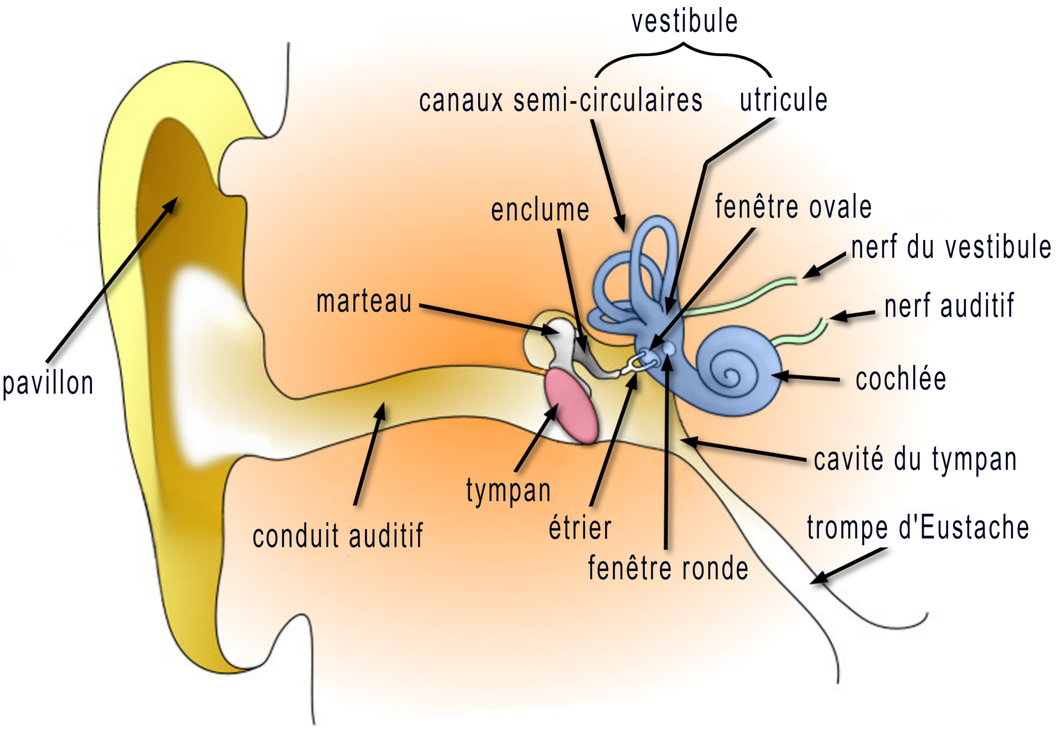


Figure 1: Schéma de l'oreille humaine, issue de Wikipédia (2024).

Le traumatisme sonore aigu (TSA) est une maladie complexe qui se développe à la suite d'une exposition soudaine et brutale à un bruit de haute intensité (Billot, 2010; Lachaux et al., 2024). Cette onde sonore puissante provoque des vibrations excessives du tympan et des osselets, mais surtout, elle perturbe profondément l'oreille interne. Les cellules ciliées, responsables de la conversion des sons en signaux électriques, sont extrêmement sensibles et peuvent être irréversiblement endommagées (Lachaux et al., 2024). En cas de lésion, comme illustré sur la Figure 2, ces cellules ne se régénèrent pas. Cela peut ainsi entraîner une perte auditive temporaire (hypoacousie) ou permanente, des acouphènes (bourdonnements ou sifflements dans les oreilles) et parfois des sensations de vertige ou de déséquilibre.

Le point de départ de cette thèse concerne l'étude d'une cohorte de patients atteints de TSA pour lesquels ont été collectés des données audiologiques et plasmatiques à différents points de temps : jour 0, jour 1, jour 3, jour 7 et

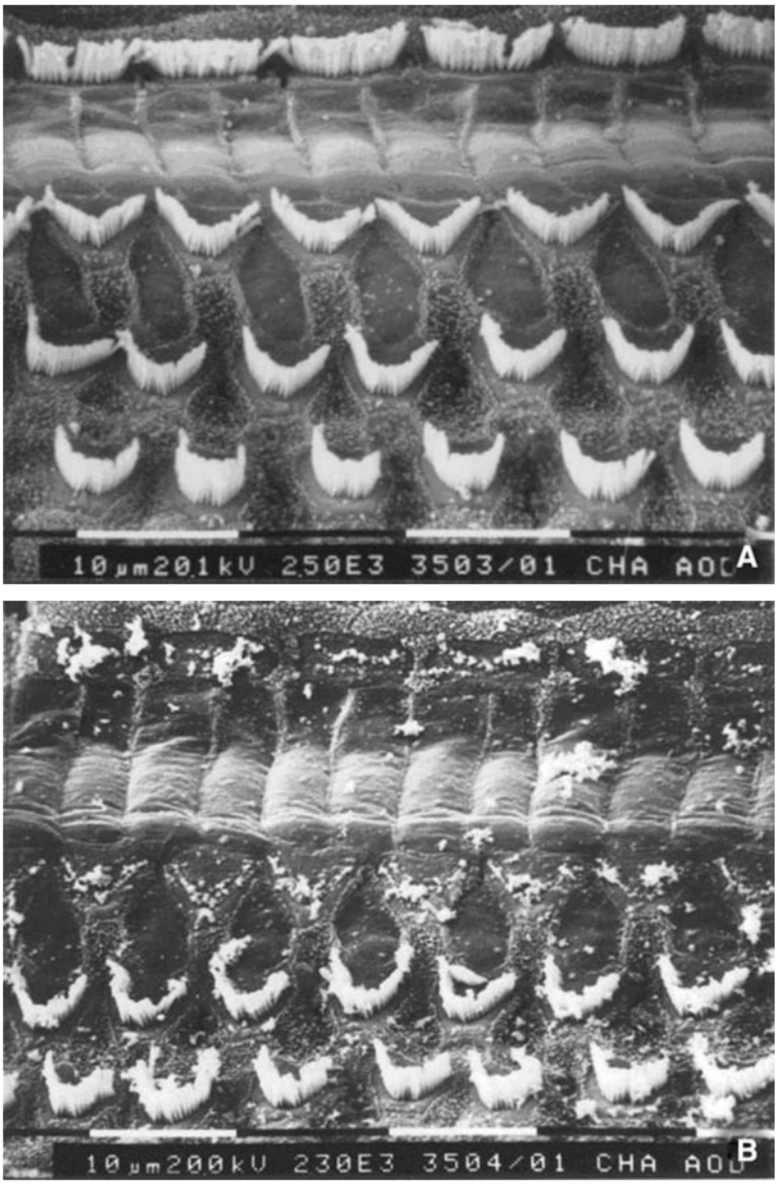

10μm201kV 250E3 3503/01 CHA AO
A
10μm200kV 230E3 3504/01 CHA AO
B

Figure 2: Traumatisme sonore aigu obtenu chez le chat au moyen d'un microscope électronique à balayage, issue de l'article écrit par Nottet et al. (2009). A. La cochlée avant le traumastime (normale). B. Cochlée après TSA : désorganisation des cellules ciliées. Le TSA a complètement détruit la structure des stéréocils.

jour 30, où le jour 0 étant défini comme le jour où le traumatisme a eu lieu. De plus, des données génomiques et protéomiques ont aussi été récoltées.

Les données audiologiques collectées incluent des mesures de la capacité auditive des participants subjectives (audiogrammes) et objectives (mesures d'otoémissions acoustiques) qui correspondent principalement à des séries temporelles. Dans cette étude, une partie des mesures d'otoémissions acoustiques correspondent aux DPOAEs (Distortion Product Otoacoustic Emissions) et aux TEOAEs (Transient Evoked Otoacoustic Emissions). Un audiogramme est un test de l'audition réalisé par un audiologiste qui permet de mesurer la capacité d'une personne à entendre des sons de différentes fréquences et intensités. Il repose sur la réponse volontaire du patient qui indique lorsqu'il perçoit un son. Cette méthode est donc subjective.

Les mesures d'otoémissions acoustiques sont des méthodes objectives, elles ne nécessitent pas de réponse active du patient. Les otoémissions acoustiques sont des sons émis par l'oreille interne en réponse à une stimulation auditive et leur mesure reflète le bon fonctionnement des cellules ciliées externes de la cochlée, comme illustré à la Figure 3. Les DPOAEs sont générées lorsqu'on envoie deux sons purs de fréquences légèrement différentes dans l'oreille. Si l'oreille interne fonctionne normalement, elle produit une troisième fréquence, appelée produit de distorsion, qui est ensuite mesurée. Ce test est particulièrement sensible à la détection des altérations cochléaires dans les fréquences médianes à hautes. Les TEOAEs, en revanche, sont déclenchées par un bref stimulus sonore (un clic) et mesurent la réponse généralisée de la cochlée sur une plus large gamme de fréquences.

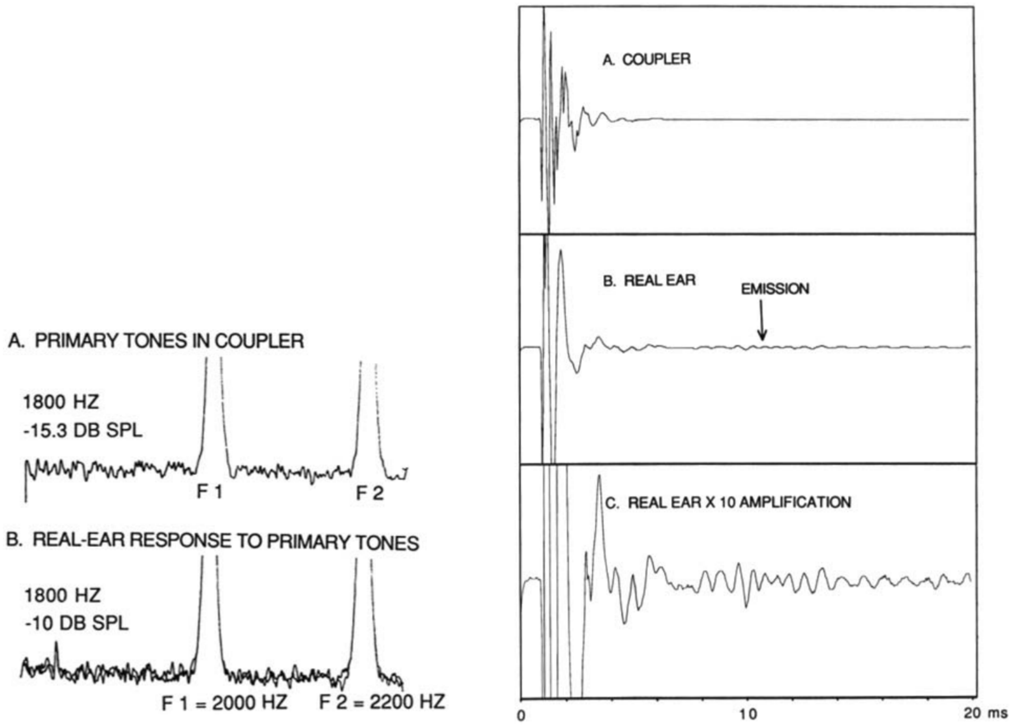
Les données génomiques englobent les séquences d'ADN pour identifier les variations génétiques pertinentes. Les données protéomiques, quant à elles, se composent de l'analyse des profils protéiques permettant d'étudier l'expression et la régulation des protéines au sein des échantillons biologiques.

Tous les échantillons ont été prélevés et analysés selon un protocole standardisé strict afin de garantir la cohérence et la comparabilité des résultats. Ce protocole permet d'assurer la robustesse des données et de faciliter l'intégration et la comparaison des résultats obtenus à différents moments.

Les données issues de ces différentes modalités permettent de caractériser chaque patient de manière unique. L'objectif est de les combiner à l'aide de modèles d'apprentissage machine afin d'identifier la complémentarité et la redondance des différentes sources d'information pour affiner et renforcer la robustesse du diagnostic.

Stratification et prédiction de récupération

Deux axes principaux sont considérés dans ce travail:



(a) DPOAE

(b) TEOAE

Figure 3: Exemples de DPOAE et TEOAE issus de Glattke and Kujawa (1991). (a) Les produits de distorsion ont été mesurés pour les sons purs $F_1 = 2$kHz et $F_2 = 2.2$kHz. Le produit de distortion le plus robuste est obtenu pour $F_{DP} = 2F_1 - F_2 = 1.8$kHz. Le coupleur, qui simule l'oreille occluse, n'a pas dépassé le seuil du bruit à F_{DP} (-15dB). En revanche, lorsque ces fréquences sont présentées à une oreille humaine saine, une DPOAE est apparue à F_{DP}, 5dB au-dessus du seuil de bruit. (b) Dans le panneau A, le stimulus transitoire a été enregistré dans un coupleur passif, un appareil qui imite l'oreille, et aucune émission n'est détectée après le stimulus. Dans le panneau B, un enregistrement effectué sur une oreille humaine montre l'apparition d'une émission environ 6 ms après le stimulus initial. Enfin, dans le panneau C, cet enregistrement a été amplifié par un facteur 10 pour rendre l'émission plus visible.

1. Stratification des patients : Diviser les patients en sous-groupes homogènes basés sur des caractéristiques spécifiques permet de mettre en évidence l'aspect multifactoriel du traumatisme sonore aigu et de confirmer l'hypothèse de sous-atteintes spécifiques. Cette tâche correspond à une mission d'apprentissage non supervisé, le clustering, qui vise à identifier des structures latentes au sein des données.

2. Prédiction de la récupération spontanée de l'audition : Les cellules ciliées internes ne se renouvellent pas et il n'existe pas de régénération cellulaire (El-Amraoui and Petit, 2010). Traiter les patients rapidement et efficacement afin de préserver leur audition, est essentiel. Les données multimodales recueillies sont combinées pour prédire la récupération — ou non — de l'audition avant que les séquelles ne deviennent irréversibles. Cette tâche d'apprentissage est dite supervisée. Elle consiste à entraîner un modèle à partir de données étiquetées (récupération *vs* non-récupération) pour effectuer des prédictions sur la récupération auditive de nouveaux patients. Ces prédictions seront réalisées à partir des données, covariables, précédemment décrites.

Les modèles développés pour répondre aux critères établis seront présentés en détails dans le chapitre 2 pour la stratification de patients et dans le chapitre 3 pour la prédiction.

En raison de la nature confidentielle des données de cette étude, les modèles inventés dans la thèse seront évalués et comparés sur d'autres ensembles de données. L'évaluation des performances des modèles est principalement réalisée sur des données simulées, ce qui permet de tester leur robustesse et leur fiabilité à travers différents schémas de simulation et niveaux de difficultés.

Organisation du manuscrit

Le premier chapitre est une mise en contexte bibliographique de l'apprentissage machine multimodal. Une attention particulière est portée à la synthèse des données ainsi qu'aux méthodes permettant de les fusionner. De plus, les modélisations statistiques nécessaires seront introduites pour établir les bases théoriques des modèles développés.

Le deuxième chapitre présente un modèle original de clustering multi-vues destiné à la stratification de patients. L'hypothèse de ce modèle repose sur la notion de communautés traversantes stables à travers les différentes vues. Ce chapitre aborde la modélisation de la redondance et de la complémentarité de l'informationen s'appuyant sur un modèle de mélange de modèles à blocs stochastiques multicouches. Il comprend également l'étude de l'identifiabilité du modèle, ses performances dans divers scénarios, sa robustesse face aux perturbations, ainsi que les critères de sélection de modèle.

Ces travaux ont été valorisées lors des conférences (De Santiago et al., 2023) et (De Santiago et al., 2024a). Un article long a été soumis (De Santiago et al., 2024c).

Le troisième chapitre introduit un mélange d'experts pour la prédiction à partir de données multimodales. Ce modèle d'apprentissage supervisé utilise un modèle à blocs latents comme première couche d'orientation. L'intégration de ce modèle à blocs latents a pour objectif de tirer parti des fortes capacités prédictives des modèles d'experts tout en conservant une grande interprétabilité.

Ces travaux incluent des bibliothèques documentées pour le logiciel R. La bibliothèque développée pour le modèle du Chapitre 2 est disponible le CRAN (De Santiago et al., 2024b), celle pour le Chapitre 3 est disponible sur mon profil GitHub (*Kdesantiago*) [1].

[1] https://github.com/Kdesantiago.

Résumé détaillé

Chapitre 1. Ce chapitre explore l'importance de l'apprentissage multimodal, qui consiste à combiner et analyser des informations de nature hétérogène pour améliorer les diagnostics et l'efficacité des traitements, en particulier dans des domaines comme l'audiologie où les profils des patients sont complexes et multifactoriels. Différents types d'intégration de données sont présentés :

- Intégration verticale : Elle utilise des données provenant de multiples procédures expérimentales pour prédire ou caractériser des patients. Par exemple, la combinaison de profils omiques pour prédire les sous-types de maladies ou la survie des patients.

- Intégration horizontale : Elle combine les résultats du même type de test appliqué à différentes populations ou ensembles de données. La méta-analyse est un exemple d'intégration horizontale.

- Intégration diagonale : Elle combine les approches verticales et horizontales, en réalisant plusieurs tests sur différentes populations et en fusionnant les résultats à différents niveaux pour une analyse intégrée.

De plus, on y souligne l'importance de distinguer la redondance de la complémentarité lors de l'intégration de données multimodales. La redondance met en évidence les informations communes et les points de convergence entre les sources, tandis que la complémentarité explore les caractéristiques spécifiques de chaque source. Parmi l'ensemble des méthodes associées à l'apprentissage multimodal, notre étude se concentre spécifiquement sur deux axes principaux : (i) L'apprentissage de représentations pour les données multimodales vise à synthétiser les données provenant de diverses sources en un format unifié et dense, en préservant l'essence de chaque modalité tout en établissant des liens cohérents entre elles. (ii) L'apprentissage par fusion, quant à lui, cherche à exploiter de manière synergique les informations provenant de multiples modalités pour résoudre un problème spécifique. Il existe trois principales approches de fusion :

- Fusion précoce : Elle combine les données brutes de différentes modalités avant tout traitement, permettant aux modèles de capturer les interactions de bas niveau.
- Fusion tardive : Elle traite chaque modalité indépendamment avant de combiner les résultats, en utilisant des techniques telles que la moyenne pondérée ou le vote majoritaire.
- Fusion intermédiaire : Elle combine les informations à différents niveaux d'abstraction, offrant une plus grande flexibilité et une meilleure synergie entre les modalités.

Enfin, la modélisation mathématique est traitée en détaillant plusieurs modèles statistiques, à savoir le modèle à blocs stochastiques, le modèle à blocs latents, et les modèles d'experts. Par la suite, les méthodes d'optimisation correspondantes sont présentées, incluant d'une part l'optimisation basée sur les chaînes de Markov cachées, et d'autre part les approches variationnelles.

Chapitre 2. Dans le chapitre 2, on propose le modèle *mimi-SBM*, une nouvelle approche pour le clustering de données multi-vues. Ce modèle s'avère particulièrement pertinent pour l'analyse de données issues de sources multiples et potentiellement hétérogènes, permettant ainsi de découvrir des structures cachées et d'améliorer la performance du clustering. L'un des principaux atouts du modèle *mimi-SBM* réside dans sa capacité à prendre en compte à la fois la redondance et la complémentarité des informations provenant des différentes vues. Le modèle repose sur deux variables latentes:

- $\mathbf{Z}$, représentant la structure en communautés des observations.
- $\mathbf{W}$, représentant la structure en composantes des vues.

Le modèle *mimi-SBM* suppose que chaque vue est issue d'un modèle de mélange de Q composantes, chaque composante étant un SBM. L'estimation des paramètres du modèle est réalisée à l'aide d'un algorithme d'Espérance-Maximisation variationnel (VEM) qui permet de contourner la complexité du calcul de la vraisemblance marginale.

De plus, une extension au cadre bayésien pour le modèle *mimi-SBM* y est développée. L'utilisation de priors conjugués simplifie l'inférence, permet d'obtenir des distributions a posteriori sous une forme analytique. De plus, cette modélisation permet d'obtenir une méthode de sélection de modèle.

Les expériences menées sur des données synthétiques et réelles démontrent l'efficacité du modèle *mimi-SBM* en matière de clustering. Les résultats montrent que le modèle surpasse les méthodes concurrentes en termes de précision, tant pour le clustering des observations que pour l'identification des composantes des vues.

Chapitre 3. Ce chapitre examine l'intégration d'une modélisation de biclustering conditionnel dans les modèles de mélange d'experts à travers le modèle *MoEBIUS* pour Mixture of Experts and BIclustering Unified Strategy. L'objectif est d'améliorer l'interprétabilité et les performances prédictives dans le contexte de l'apprentissage automatique multimodal. Le modèle cherche à identifier des relations de redondance et de complémentarité au niveau des variables au sein de chaque communauté:

- La redondance de l'information: Certaines variables peuvent fournir une information similaire au sein d'une communauté. *MoEBIUS* regroupe ces variables redondantes en composantes et utilise une variable représentative pour chaque composante. Cela permet de simplifier l'interprétation et de réduire la complexité du modèle.

- La complémentarité de l'information: Différentes composantes peuvent apporter des informations complémentaires pour expliquer la variabilité de la variable cible $\mathbf{y}$ au sein d'une communauté. Le modèle capture ces interactions en modélisant la cette variable en fonction des variables représentatives de chaque composante, avec des paramètres de régression spécifiques à chaque communauté.

Le modèle développé se nomme *MoEBIUS*, pour Mixture Of Experts and BIclustering Unified Strategy. Il utilise un réseau d'attribution (gating network) pour associer les observations à différentes communautés en fonction de leurs caractéristiques. Ensuite, pour chaque communauté, les variables sont regroupées en composantes, et une variable représentative est créée pour chacune. Enfin, une régression est effectuée sur ces variables représentatives, avec des paramètres de régression spécifiques à chaque communauté.

Des mesures de performances sur des données simulées ont confirmé l'efficacité de MoEBIUS. Il surpasse significativement un modèle de régression linéaire global et un modèle en deux étapes utilisant K-means suivi d'une régression linéaire. *MoEBIUS* a démontré son apport en termes de performance de régression et de précision de co-clustering, en particulier dans des scénarios complexes avec un nombre limité d'observations (Table 3.1), comparé aux autres approches analysées.

De plus, le développement d'un critère de sélection, le critère BIC_ICL, s'est avéré être la méthode de sélection de modèle la plus fiable pour MoEBIUS, permettant une sélection précise des hyperparamètres K (nombre de communautés) et Q (nombre de composantes) dans la plupart des scénarios de simulation (Figures 3.6 et 3.7).

Contents

1 Context 17

1.1 Multimodal Learning 17

1.1.1 Schemes of multimodal integration 19

1.1.2 Problematics of multimodal integration 22

1.2 Some model-based approaches 29

1.2.1 Mixture models 29

1.2.2 Classical block models 31

1.2.3 Mixture of experts 40

1.3 Optimization methods 43

1.3.1 Variational approach 44

1.3.2 MCMC approaches 46

2 MIxture of Multilayer Integrator Stochastic Block Model 51

2.1 Introduction 52

2.2 Background 54

2.2.1 Consensus Clustering 54

2.2.2 Block Models for Multiview Clustering 55

2.3 Mixture of Multilayer Stochastic Block Models 57

2.3.1 A Mixture of Observations across a Mixture of Views . 58

2.3.2 Identifiability of the Model 59

2.3.3 Variational Expectation-Maximization for mimi-SBM . 60

2.4 Bayesian Framework 62

2.4.1 Evidence Lower Bound 63

2.4.2 Lower Bound Optimization 65

2.5 Initialization and model selection 68

2.5.1 Initialization 68

2.5.2 Model selection 69
2.6 Experiments 71
2.6.1 Simulation study 71
2.6.2 Worldwide Food Trading Networks 81
2.7 Bilan 85

3 Mixture of Experts models × Conditional LBM 87
3.1 Introduction 89
3.2 Material and Methods 93
3.2.1 General modeling 93
3.2.2 Mixture Of Experts and BIclustering Unified Strategy model 95
3.3 Optimization via SEM-Gibbs algorithm 98
3.3.1 Gibbs-sampling Expectation step 99
3.3.2 Maximization step 99
3.4 Model Selection 101
3.5 Experiments 102
3.5.1 Simulation process 103
3.5.2 Co-clustering and regression performance 105
3.5.3 Parameter estimations 107
3.5.4 Model selection 107
3.6 Bilan 110

4 Conclusion and perspectives 113

A Identifiability of mimiSBM 119
A.1 Assumptions 119
A.2 Identifiability of $\boldsymbol{\pi}$ 120
A.3 Identifiability of $\boldsymbol{\rho}$ 122
A.4 Identifiability of $\boldsymbol{\alpha}$ 124

B Details of VEM algorithm for mimiSBM - frequentist framework 127
B.1 Variational parameters of clustering τ_{ik} 127
B.2 Variational parameters of component membership ν_{vs} 128
B.3 Optimization of $\boldsymbol{\pi}$ 130
B.4 Optimization of $\boldsymbol{\rho}$ 130

B.5 Optimization of $\boldsymbol{\alpha}$ 131

C Details of VBEM algorithm for mimiSBM - Bayesian framework 135

C.1 Variational parameters of clustering τ_{ik} 135

C.2 Variational parameters of component membership ν_{vs} 137

C.3 Optimization of $q(\boldsymbol{\pi})$ (β_k) 139

C.4 Optimization of $q(\boldsymbol{\rho})$ (θ_s) 140

C.5 Optimization of $q(\boldsymbol{\alpha})$ (η_{kls} and ξ_{kls}) 141

D Evidence Lower Bound 145

E Details of Gibbs Sampling - EM with gradient ascent algorithm for MoEBIUS model 149

E.1 Variational parameters of clustering τ_{ik} 149

E.2 Variational parameters of component membership ν_{vs} 150

E.3 Optimization of $\boldsymbol{\rho}$ 150

E.4 Optimization of $\boldsymbol{\pi}$ 151

E.5 Optimization of $\boldsymbol{\beta}$ 152

E.5.1 Optimization of $\boldsymbol{\beta}$ - Regression 152

E.5.2 Optimization of σ_k^2 153

E.5.3 Optimization of $\boldsymbol{\beta}$ - multiclass classification 154

F Co-conditional Latent Block Model 157

F.1 Introduction 157

F.2 Materials and Methods 159

F.3 Optimization by Variational EM 160

F.4 Model selection 161

F.5 Experiments 162

F.5.1 Co-clustering Performance 163

F.5.2 Co-clustering Performance 163

F.5.3 Parameter Estimations 165

F.5.4 Model Selection 165

G Details of Variation EM algorithm for Co-conditional Latent Block Model 169

G.1 Variational parameters of clustering τ_{ik} 169

G.2 Variational parameters of component membership ν_{vs} 170
G.3 Parameters of distribution μ_{ks} 171
G.4 Parameters of distribution σ^2_{ks} 171

H Details of Variation EM algorithm for Co-CoLBMoE 173
H.1 Objective function 173
H.2 Community membership $\boldsymbol{\tau}$: 174
H.3 Component membership $\boldsymbol{\nu}$: 175
H.4 Expert selection $\mathbf{g}$: 177
H.5 Community proportions $\boldsymbol{\rho}$: 178
H.6 Community membership regression parameters $\boldsymbol{\pi}$: 179
H.7 Expert selection regression parameters $\boldsymbol{\gamma}$: 180
H.8 Regression parameters : $\boldsymbol{\beta}$ 181
H.9 Noise σ^2_{ks} 183
H.10 Multinomial logistic regression by SEM-Gibbs algorithm $\mathbf{W}$ $\boldsymbol{\beta}$ 183

1 Context

1.1 Multimodal Learning 17

1.1.1 Schemes of multimodal integration 19

1.1.2 Problematics of multimodal integration 22

1.2 Some model-based approaches 29

1.2.1 Mixture models 29

1.2.2 Classical block models 31

1.2.3 Mixture of experts 40

1.3 Optimization methods 43

1.3.1 Variational approach 44

1.3.2 MCMC approaches 46

1.1 Multimodal Learning

Most learning situations involve the integration of different sources of information, such as vision, touch and hearing. An information source in a given format will be called *modality* or *view*. Multimodal or multiview machine learning aims to learn models from multiple views (e.g. text, sound, image, etc.) in order to represent, translate, align, merge or co-learn (see Zhao et al., 2017; Baltrušaitis et al., 2018; Cornuéjols et al., 2018, for instance).

The benefits of multimodal machine learning in the medical field are manifold. Firstly, combining different data modalities enables a more holistic understanding of patient health (Acosta et al., 2022). For example, by merging medical image analysis with textual medical history and real-time biometric data, doctors can obtain a more complete overview, facilitating diagnosis and treatment planning (Bertsimas and Ma, 2024; Shivahare et al., 2024).

Moreover, this approach can significantly improve the accuracy of predictions and diagnoses. By integrating various data sources, models can identify subtle patterns and correlations that are not evident with a single modality. This is particularly useful in fields such as neurology or oncology where the combination of brain imaging, behavioral data, and genetic biomarkers can lead to earlier and more accurate detection of diseases (Boehm et al., 2022; Zhou et al., 2019; Biffi et al., 2010).

Multimodal machine learning also offers opportunities for more personalized medicine. By analyzing a wide range of patient-specific data, systems can propose tailored treatments, predict drug responses, and identify potential risks more accurately. For example, a meta-analysis confirmed the accuracy of lung ultrasound scores for early prediction of bronchopulmonary dysplasia in premature newborns (Pezza et al., 2022). This study demonstrates how the use of a specific imaging modality can significantly improve predictive accuracy and enable early diagnosis.

However, it's important to note that using these models also entails risks. Incomplete or unrepresentative predictive models can lead to biased results, resulting in misdiagnosis or overdiagnosis (Phillips et al., 2022). It is, therefore, obviously crucial to develop and use these tools in conjunction with a medical team, thus ensuring the consistency of predictions and the quality and representativeness of the data used.

Organization of the Chapter. This chapter provides an overview of the key concepts in multimodal learning and introduces the mathematical foundation that will be used to define both our unsupervised and supervised models within the multimodal framework.

In Section 1.1, we provide an overview of multimodal learning and its potential benefits. We examine the vertical, horizontal, and diagonal data integration schemes. Then, we discuss the need to optimize the balance between redundancy and complementarity of information coming from the different modalities. Finally, unified representations and fusion learning are presented. Section 1.2 shifts the focus to model-based approaches for multimodal learning. It begins with mixture models, emphasizing their ability to identify hidden structures within complex data. The section also covers classical block models, especially Stochastic Block and Latent Block Models which are widely applied to network analysis and biclustering, respectively. It concludes with a discussion of Mixture of Experts models, underlining their strength in managing complex tasks by leveraging multiple specialized models.

Section 1.3 reviews common optimization techniques for latent variable models, including Markov Chain Monte Carlo (MCMC) methods like Gibbs sampling, Metropolis-Hastings or Hamiltonian Monte Carlo. It also covers the variational approach to the Expectation-Maximization algorithm, providing an efficient solution for models with many parameters or observations by optimizing a lower bound on the likelihood.

1.1.1 Schemes of multimodal integration

Data integration is crucial in today's era of growing information sources and complex problems. By combining diverse analyses, it enhances understanding of the studied phenomena and strengthens the robustness and reliability of machine learning models (Boehm et al., 2022). In multimodal machine learning, data integration plays a key role in developing strategies to merge and analyze heterogeneous information, with three primary types of integration: vertical, horizontal, and diagonal (Briere, 2022).

Vertical integration

Vertical integration aims to characterize patients by using data from various experimental procedures. This scheme corresponds to the issue presented in the introduction. It combines results from different tests or analyses on the same population or dataset to provide *complementary information* about the same observations (see Figure 1.1).

For instance, this can be done by taking into account the different omic profiles of a group of patients to perform disease subtype prediction based on multi-omic data (Saria and Goldenberg, 2015; Mihaylov et al., 2019) or to predict patient survival based on their multi-omic profiles (Mihaylov et al., 2019; Wu et al., 2021). Moreover, vertical integration can also anticipate or describe the behavior or importance of certain features or groups of features derived from experiments (Wu et al., 2021; Cantini et al., 2021).

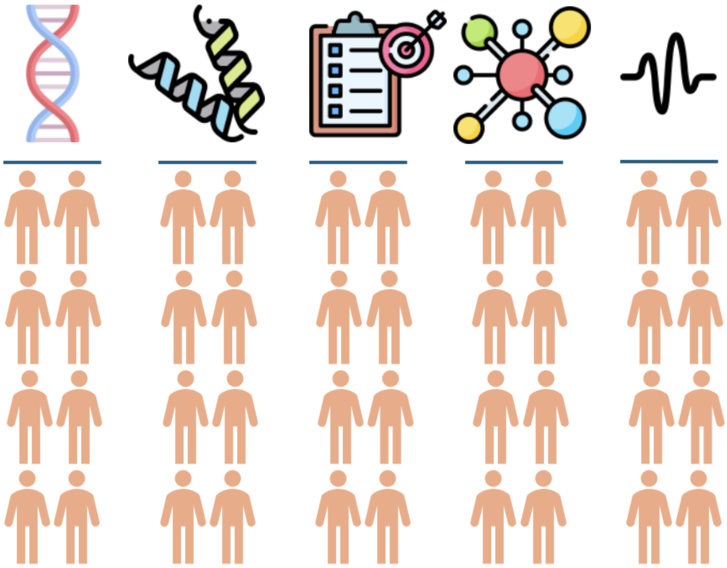


Figure 1.1: Vertical integration: The same population is characterized by a set of data derived from tests of different natures.

Horizontal integration

Horizontal integration involves combining the results of the same type of test or analysis applied to different populations or data sets, as shown in Figure 1.2. When those analyses aim to answer the same question, they may be pooled into a meta-analysis (Gurevitch et al., 2018; Borenstein et al., 2021). Meta-analyses aggregate results from independant multiple studies to determine if conclusions align, potentially enhance statistical power, and provide more reliable results than a single population (Franke et al., 2010; Tzenios et al., 2024). In cases where independent experiments are designed to answer different questions, horizontal integration can be used to define standard features between experiments, such as similar test behaviors or similar profiles (see Shen and Tseng, 2010; Subramanian et al., 2020; Tseng et al., 2012, for instance).

Diagonal integration

Diagonal integration merges vertical and horizontal integration approaches, as shown in Figure 1.3. It involves conducting several tests on different populations.

The diagonal approach to multimodal data integration presents several significant challenges in its implementation (Xu and McCord, 2022):

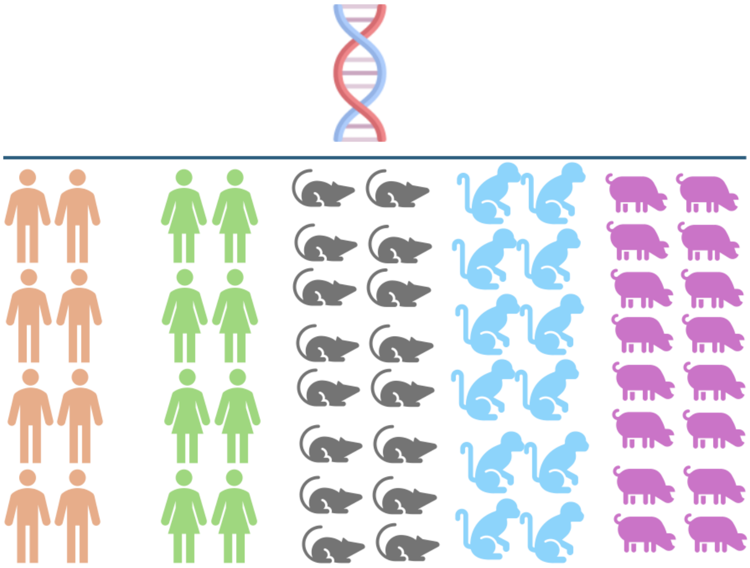


Figure 1.2: Horizontal integration. Data from genetic studies are collected on different populations.

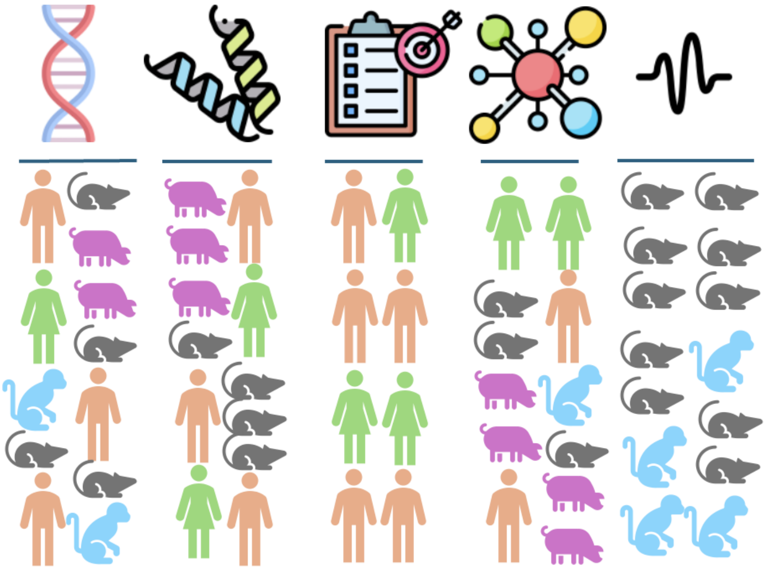


Figure 1.3: Diagonal integration. Various tests were carried out on different populations. Only mice and monkeys were tested for audiological data, while proteomic tests were carried out on all populations except monkeys.

- *Anchoring data across different modalities.* Finding commonalities or reliable correspondences across various data types can be extremely difficult (Calaon et al., 2024), especially when modalities have very different structures, scales, or domains. For example, in the field of biology, aligning spatial data with single-cell data can require sophisticated techniques to establish meaningul correspondences between modalities (Athaya et al., 2023; Chen et al., 2024).

- *The complexity of signal analysis.* Relevant information is often diffused across different modalities, making distinguishing between significant signal and noise difficult. This difficulty is particularly pronounced in diagonal integration approaches, where relationships between modalities can be subtle, indirect, and shared across different combinations of analyses (Chen et al., 2024).

For further details and possible solutions, the thesis by Briere (2022) may be helpful.

1.1.2 Problematics of multimodal integration

From representation to fusion learning

Representation learning for multimodal data aims to unify information from diverse sources into a coherent and dense format (Baltrušaitis et al., 2018), retaining the essential features of each modality while creating meaningful connections between them, enabling a thorough analysis of heterogeneous data. Unified representation encompasses a versatile range of integration strategies. It is particularly complex and crucial for several reasons (see Liang et al., 2022; Kline et al., 2022; Baltrušaitis et al., 2018):

- *Efficient compression*: The challenge is to reduce the dimensionality of the data while retaining the relevant information. This involves using advanced methods such as multimodal Principal Component Analysis (PCA) or multimodal autoencoders.

- *Preservation of specificities*: Each modality has its own characteristics and structures. It is essential to design architectures that can capture these specificities. For example, convolutional neural networks for images or transformers for text can be combined into a unified framework.

- *Inter-modal alignment*: A major challenge is correctly aligning information from different modalities. Cross-attention mechanisms or contrastive learning techniques can be employed to establish meaningful correspondences between modalities.

- *Handling missing data*: In many practical cases, some modalities may be missing. The learned representations need to be robust to these scenarios and capable of functioning with partial inputs.

- *Interpretability*: While the goal is to summarize the information, it is also crucial to maintain a certain level of interpretability. This allows researchers to understand how different modalities contribute to the final representation.

- *Generalization*: The learned representations should be sufficiently general to be useful in various downstream tasks such as classification, generation, or multimodal retrieval.

Fusion Learning

Fusion learning is a fundamental approach to harnessing information from multiple modalities to solve specific problems collaboratively. The timing of data fusion can occur at three stages: Early, Late, or Intermediate. Early fusion integrates data at the raw input stage, late fusion combines results from individual models, and intermediate fusion balances the two approaches. This timing is essential in balancing cross-modal interactions with specific signal extraction from each modality.

Early Fusion Early integration merges raw data or their initial representations (Baltrušaitis et al., 2018), mainly by concatenating them together to form a large description vector, as illustrated in Figure 1.4. The fused feature vector is then used as input for a learning model (e.g., a neural network, a support vector machine, etc.). The model learns from this combined vector, aiming to capture the interactions and correlations between the different modalities from the very beginning of the process.

This type of method has two major advantages:

- *Capturing interactions*: By combining the data at the beginning, the model can learn to identify complex interactions and dependencies between modalities.

- *Single model*: Only one model is required to process the fused data, thus simplifying the processing pipeline.

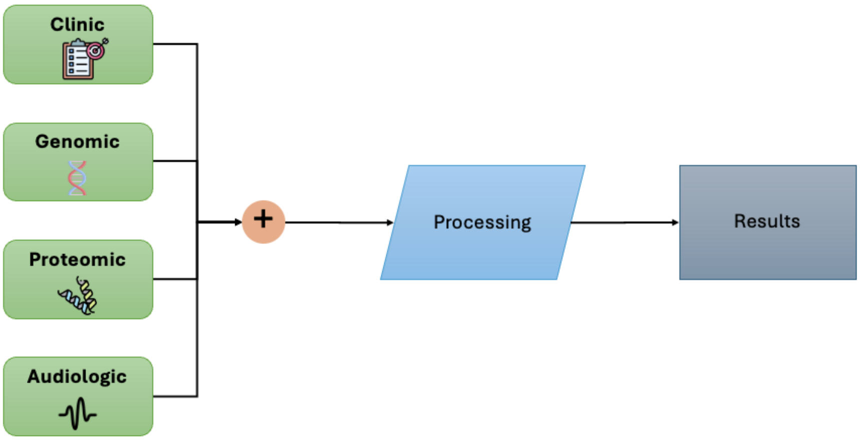


Figure 1.4: Early fusion of multimodal data. (i) Data Collection: The process begins with the collection of clinical, genomic, proteomic, and audiologic data. (ii) Data Processing: The collected data is integrated and analyzed together. (iii) Results: Following data processing, significant results are obtained. These results may include diagnoses, predictions, or recommendations for patient treatment.

However, early fusion presents several significant challenges (Briere, 2022; Liang et al., 2022):

- *Bias toward major modalities*: Without appropriate adjustments, early fusion tends to excessively favor data types with the most features. This can lead to the loss of important signals from smaller but potentially crucial modalities.

- *Amplification of dimensionality issues*: This method exacerbates challenges related to statistical power, particularly pronounced in genomic data analysis. The simple concatenation of multiple datasets significantly increases dimensionality, worsening already existing issues in high-dimensional data analysis.

Late Fusion Late fusion is an approach in multimodal learning that involves processing each modality independently before combining the results (Briere, 2022, see Figure 1.5). In this method, distinct models are applied to each type of data, and then the predictions from these models are merged using techniques such as weighted averaging, majority voting, or ensemble methods (Moshawrab et al., 2023).

For example, in the context of horizontal integration, meta-analysis is a late fusion method. Meta-analyses aim to synthesize the results of multiple independent studies on a given topic, based on the results obtained through various studies.

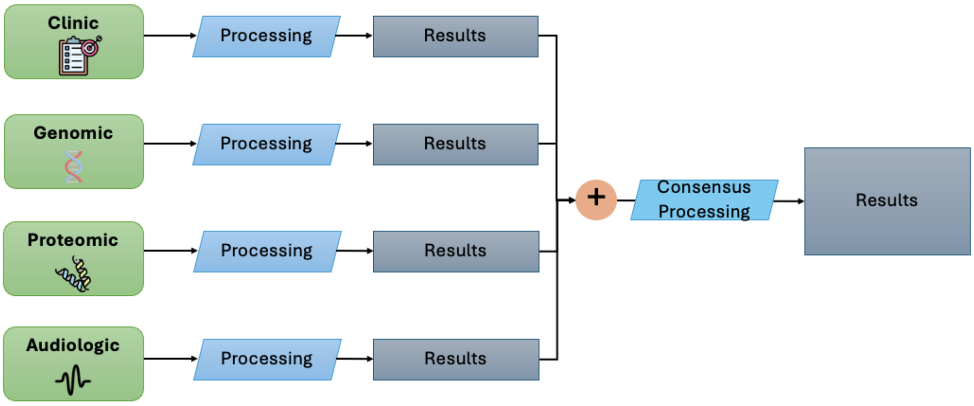


Figure 1.5: Late fusion of multimodal data. (i) Data Collection: The process begins with the collection of clinical, genomic, proteomic, and audiological data. (ii) Data Processing and Intermediate Results: The collected data is integrated and analyzed independently, allowing for specialized processing for each modality, as well as specific results that fully exploit the available signal. (iii) Consensus Processing and Results: The intermediate results are integrated and processed to obtain a final consensus result.

This approach leverages the specific strengths of each modality, which can improve the overall accuracy of the model. By processing each modality separately, late fusion is also more resilient to errors that may occur in a particular modality. For instance, in certain medical prognosis or diagnosis contexts, late fusion is used to combine data from radiographs, blood tests, and medical records, each analyzed by a specialized model (Cui et al., 2023).

However, late fusion has some limitations. It may lack synergy between modalities, as it does not account for direct interactions between multimodal features. This means that valuable information, which could emerge from early data fusion, might be lost.

In Chapter 2, a consensus processing method will be developed in the context of late fusion. This approach aims to combine the results of multiple clustering analyses extracted from each modality to obtain a final synthesis and partitioning of the observations.

Intermediate Fusion Intermediate fusion, or hybrid fusion, refers to an approach where data from different modalities are integrated at an intermediate level of the learning process (Guarrasi et al., 2024; Stahlschmidt et al., 2022). It occurs after pre-processing the data, making the format homogeneous across each modality, as illustrated in Figure 1.6.

Intermediate fusion allows for the exploitation of relationships at different levels of abstraction, offering greater flexibility and better synergy between modalities. However, it can increase the complexity of the models, which must be carefully designed to efficiently manage interactions between modalities.

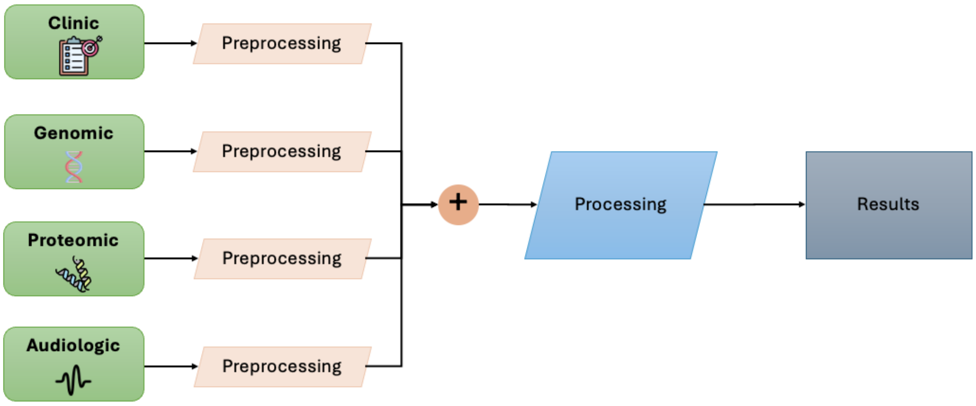


Figure 1.6: Intermediate fusion of multimodal data. (i) Data Collection: The process begins with the collection of clinical, genomic, proteomic, and audiologic data. (ii) Data Preprocessing: The collected data is preprocessed independently, allowing for a homogeneous format across each modality, while extracting as much signal as possible. (iii) Processing and Results: The results from preprocessing are integrated and processed to obtain a final outcome.

There are a multitude of methods that allow for intermediate fusion; here are a few examples:

- *Multiple Kernel Learning* (MKL): Each modality can be represented by a different kernel, and MKL learns an optimal combination of these kernels to improve model performance. This method is flexible and can capture complex nonlinear relationships between modalities (Wang et al., 2021; Lauriola et al., 2020). However, it can be computationally expensive and memory-intensive if the number of observations becomes large.

- *Latent vector concatenation* (VAE, AE, PCA): This method involves using dimensionality reduction techniques such as autoencoders (Hinton and Zemel, 1993, AE), variational autoencoders (Kingma and Welling, 2022, VAE) or Principal Component Analysis (PCA) to extract latent representations of each modality. These representations are then concatenated to form a merged feature vector, which is used as input for further data processing. Latent vector concatenation allows for capturing essential features while reducing dimensionality, which can improve computational efficiency and model performance.

- *Transformers with Cross-Attention*: Transformers with cross-attention mechanisms explicitly model the interactions between different modalities. In this context, the representations of each modality are crossed through attention layers, allowing each modality to "inform" the others and enabling interaction between the latent embeddings of each modality (Wang et al., 2024).

In Chapter 3, we introduce an intermediate fusion method, where variables are first grouped into distinct components, conditional on communities. A representative variable is then extracted from each component to reduce dimensionality while retaining relevant information. These aggregated variables are subsequently used for prediction tasks, providing a robust and interpretable approach for integrating complex data.

Redundancy versus complementarity

The kind of integration and fusion are important, but so is the information to be extracted. Depending on requirements, it may be appropriate to focus on extracting *common information*, revealing *redundancies* and points of convergence between different sources, for meta-analyses for example.

Alternatively, we might choose to extract the *specific behaviors* associated with each modality, which helps identify *complementarities* between them. This enables the exploration of differences and enhances the unique contributions of each piece of information, enriching the overall analysis.

Redundancy

Research in multimodal machine learning has led to various methods aimed at maximizing information redundancy between different modalities. In such cases, merging modalities may not significantly enhance model performance but instead help identify fundamental data characteristics across different sources. However, concentrating solely on dominant patterns can result in overlooking specific features and unique observations, diminishing the distinctiveness of each source (Baltrušaitis et al., 2018).

Multiple Kernel Learning (MKL) has played a significant role to capture complementary information from various data sources (see below). Yet, consensus kernel approaches, like those developed by Mariette and Villa-Vialaneix (2018), use kernel combination techniques to align the spaces of different modalities, optimizing informative redundancy through cosine distance between each kernel and a consensus kernel that summarizes the overall information.

The use of generative adversarial networks (GANs) has been shown to be effective in maximizing information redundancy. For instance, Zhu et al. (2017) have proposed CycleGAN, an architecture capable of learning mappings between two different domains. This technique captures features common to each modality, exploiting intra- and inter-domain variations to generate rich and informative representations.

Complementarity

Complementarity occurs when different modalities captures specific information of the studied phenomenon, enhancing overall model performance. The challenge lies in obtaining relevant modality-specific representations of the data without being affected by noise or measurement disturbances.

Traditional statistical methods incorporate multimodal information through techniques such as principal component analysis (PCA), hidden Markov models (HMM) or neural networks. For instance, The MOLI (Multi-Omics Late Integration) approach, developed by Sharifi-Noghabi et al. (2019), leverages intermediate integration of multi-omics data to predict drug response using deep learning. This model employs deep neural networks as feature extractors for each type of omics data, ensuring a modality-specific representation before merging them in a final layer. The last hidden layers from each omics source are concatenated to capture the complementarity between these distinct modalities. The *DIABLO* (Data Integration Analysis for Biomarker discovery using Latent cOmponents) model, developed by Singh et al. (2019), is a supervised multi-omics method based on a latent variable model. This model use the partial least squares approach to integrate several types of omics data. The key idea is to maximize the covariance between variables projected from each modality onto shared latent components, enabling common relationships to be captured while preserving the specificities of each data block.

Another effective method to improve model performance while identifying complementary information is the use of cross-attention in deep learning architectures, which explicitly models interactions by capturing complex dependencies, resulting in accurate representations between modalities (Khan et al., 2024; Song et al., 2023; Huang et al., 2018, for instance)

In chapters 2 and 3, the proposed stratification and prediction method, respectively, will merge redundant modalities via a clustering process, thereby enhancing complementary aspects across the various clusters.

1.2 Some model-based approaches

In the multimodal context, the model-based approach allows for modeling the relationships between the data and the studied problem. It considers the type of integration, the information to be maximized, and the link between the different modalities. Moreover, the statistical model approach is particularly useful for making predictions, estimating parameters based on hypotheses about the processes that generate the data, and integrating random variations through probability distributions. This modeling provides a robust framework for handling the inherent uncertainty in observed data and adapting to the data used (Bishop, 2013).

In the following, we define a community as a group of individuals who share similar characteristics, and a component represents a grouping of variables or descriptors that exhibit similar patterns or dependencies.

1.2.1 Mixture models

A mixture model is a statistical approach that represents the distribution of a dataset as a combination of several community, each with its own distribution (McLachlan and Peel, 2000). These models are particularly effective at uncovering hidden structures and patterns in complex and heterogeneous data (Biernacki et al., 2000).

In the context of multimodal learning, mixture models allow for grouping data of the same nature, with similar characteristics (redundancy), and estimate parameters characterizing the heterogeneity between different modalities and communities (complementarity).

Let $\mathbf{X}$ be a matrix of N observed data represented by random vectors $(\mathbf{X}_i)_{i=1:N}$, with $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)$. Let $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ be a set of parameters that define the K *communities* of the mixture model. The density function of $\mathbf{X}_i$ can be expressed as:

$$f(\mathbf{X}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{X}_i|\theta_k), \tag{1.1}$$

where π_k denotes the weigth of the k^{th} community, $f_k(.|\theta_k)$ its density, and the parameters θ_k determine the properties of its distribution. An example of a mixture of two Gaussian distributions, with changing proportions, is shown in Figure 1.7.

The key challenge of this approach is determining the optimal set of parameters $\boldsymbol{\Theta}$. Typically, this is done by maximizing the log-likelihood function, which in this case is given by

$$l(\boldsymbol{\Theta}; \mathbf{X}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{X}_i|\theta_k) \right). \tag{1.2}$$

Direct log-likelihood optimization can be complex, due to the non-trivial form of the objective function.

In latent variable models, we assume that the community membership of the observations $(\mathbf{X}_i)_{i=1:N}$ is described with latent variable Z_i following a multinomial distribution :

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)), \tag{1.3}$$

where $\boldsymbol{\pi} \in (0,1)^K$ is the mixture proportion vector, such that $\sum_{k=1}^{K} \pi_k = 1$.

In this family of models, the Expectation-Maximization (EM) algorithm is often employed to estimate the model parameters, thus enabling unsupervised classification (Dempster et al., 1977). A detailed discussion of the algorithm is provided in Section 1.3.

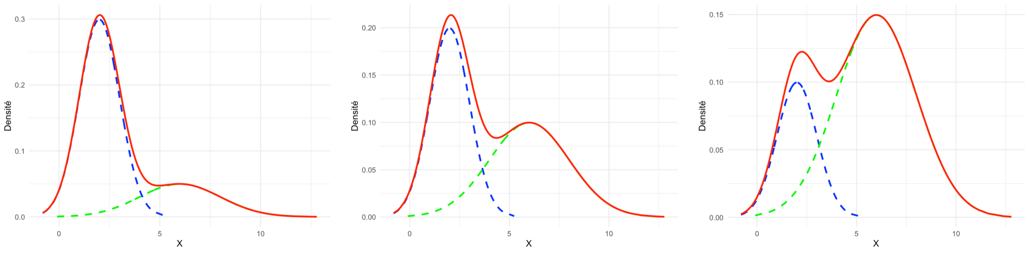
For this category of models, the following assumptions are typically made:

- The latent variables are assumed to be independent: $\mathbf{Z}_i \perp\!\!\!\perp \mathbf{Z}_j$, for all

$i \neq j$.

- The observed data are conditionally independent given the latent variables: $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{Z}$, for all $i \neq j$.

- For observations within the same community, the observed data are assumed to be independent and identically distributed (iid).



(a) $(\pi_1 = 0.75, \pi_2 = 0.25)$ (b) $(\pi_1 = 0.5, \pi_2 = 0.5)$ (c) $(\pi_1 = 0.25, \pi_2 = 0.75)$

Figure 1.7: Impact of weights on the density of mixtures of two Gaussian distributions. The Gaussian with the density in blue is $\mathcal{N}\left(\mu_1 = 2, \sigma_1^2 = 1\right)$ and the one in green is $\mathcal{N}\left(\mu_2 = 6, \sigma_2^2 = 4\right)$.

1.2.2 Classical block models

Stochastic block models

Stochastic block models (SBMs) are latent variable statistical models widely used for analyzing social networks and other types of complex graphs (Holland et al., 1983). This approach is particularly useful for community detection in various contexts, such as social networks, biological interactions, or communication networks (Miele and Matias, 2017; Leger et al., 2015), as illustrated in Figure 1.8.

These models aim to model the probability of links between nodes or individuals based on their membership in groups (Nowicki and Snijders, 2001), which are composed of nodes in a network with their own interaction methods. More specifically, the SBM uses a latent structure to partition the vertices of a network into blocks, with each block representing a group of nodes sharing common characteristics (Nowicki and Snijders, 2001).

In an intermediate fusion appoach, SBMs can be used to characterize communities based on interactions defined by proximities calculated across different modalities. From a late fusion perspective, it could be used to aggregate into a final clustering a set of independant clusterings obtained from various modalities.

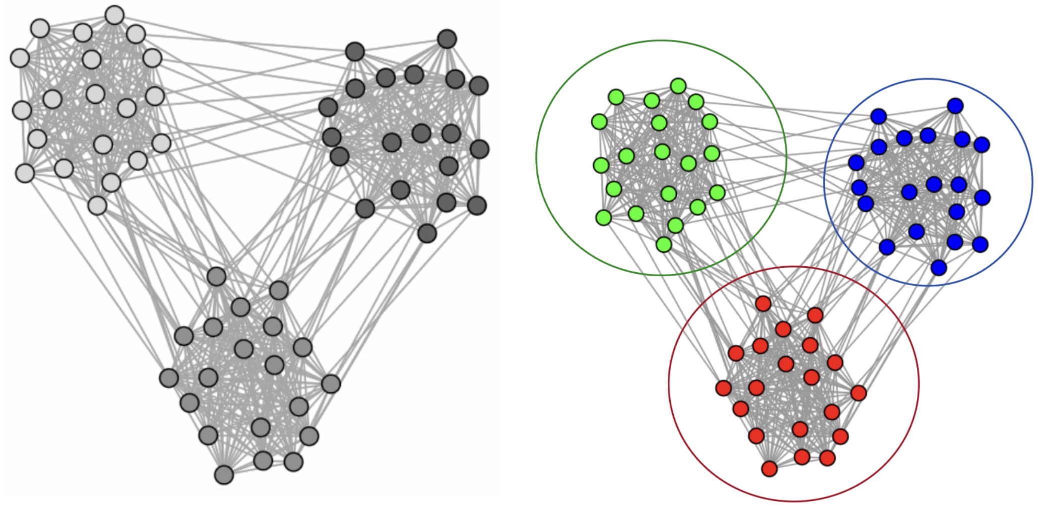


Figure 1.8: Community detection in an interaction network. On the left, a graph of nodes interconnected by numerous edges, without any group structure. On the right, the same network is divided into three distinct communities, each within a colored circle (green, blue, and red). Each group contains the nodes of the corresponding color, and there are significantly more interactions within the groups compared to those between the groups.

We focus here on SBMs for unweighted and undirected graphs. Let $\mathcal{G}$ be a random binary undirected graph, represented by a symmetric $N \times N$ matrix denoted by $\mathbf{A}$, as shown in Figure 1.9. Each entry A_{ij} is a binary random variable indicating the presence or absence of an edge between nodes i and j. More specifically,

$$A_{ij} = \begin{cases} 1, & \text{if individuals } i, j \text{ are linked in } \mathcal{G} \ , \\ 0, & \text{otherwise} \ . \end{cases} \tag{1.4}$$

In this model, each node i belongs to a group or community Z_i, and the probability of an edge existing between two nodes i and j depends solely on the groups to which these nodes belong. This probability is given by an interaction probability matrix between the groups, denoted $\boldsymbol{\alpha}$, where the element $\boldsymbol{\alpha}_{kl}$ represents the probability of an edge existing between a node from group k and a node from group l, as illustrated in Figure 1.10.

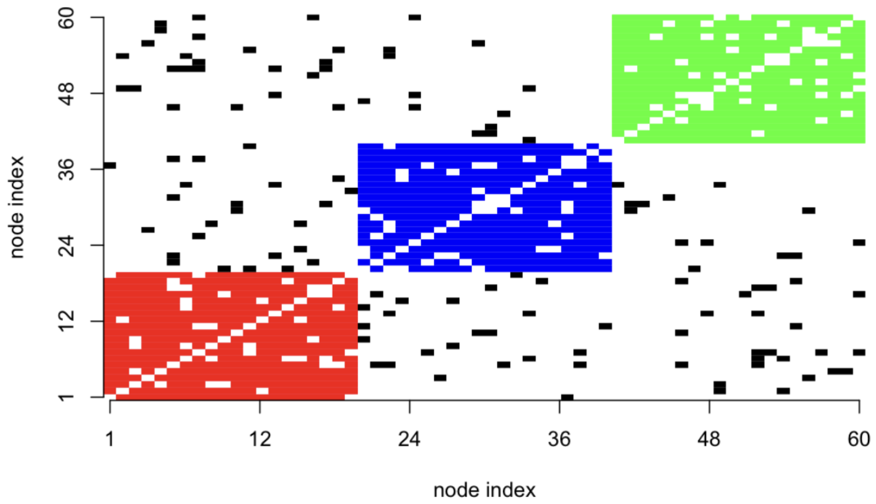


Figure 1.9: An adjacency matrix corresponding to the encoding of the graph in Figure 1.8. The presence of a red, blue, or green point represents an interaction between two individuals of the same corresponding group. Black points indicate an interaction between two individuals from different groups.

Thus, for two nodes i and j, the conditional probability of an edge is

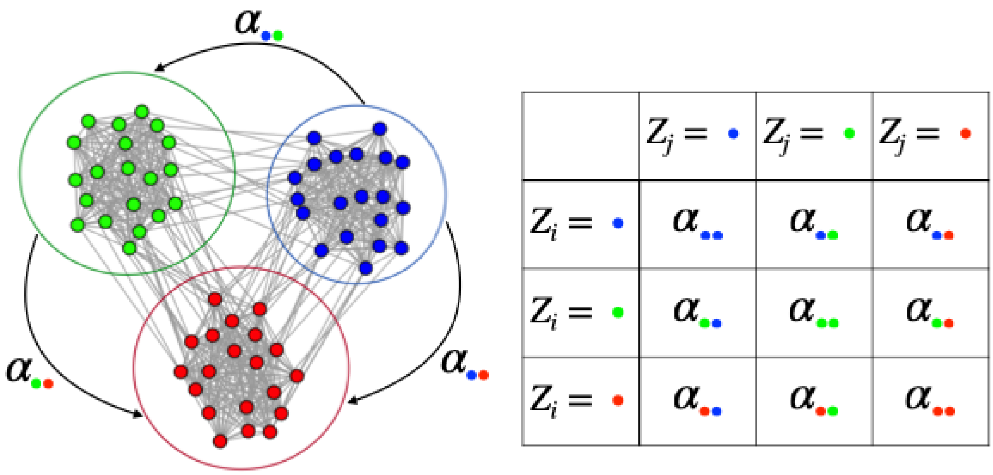
$$A_{ij}|Z_{ik} = 1, Z_{jl} = 1 \sim \mathcal{B}(\alpha_{kl}), \tag{1.5}$$

where $\mathcal{B}(.)$ denotes the Bernoulli distribution.

This stochastic block model relies on several fundamental assumptions that enable effective modeling of community structures in networks:

- *Connection probabilities*: All nodes within the same block have the same probability of connecting with nodes from other blocks.
- *Homogeneity within blocks*: Nodes within the same block are assumed to be homogeneous in terms of connection probability. This means that nodes in a block exhibit similar connection behaviors, simplifying the modeling of interactions within blocks.
- *Edge independence*: Generally, it is assumed that edges are independent of each other, given the blocks to which the connected nodes belong.

In this document, it is assumed that, conditional on the latent variables, the distribution of edges A_{ij} follows a Bernoulli distribution and that the values of $\boldsymbol{\alpha}$ are symmetric due to the undirected graph assumption. However,



	Z_j = 🔵	Z_j = 🟢	Z_j = 🔴
Z_i = 🔵	α🔵🔵	α🔵🟢	α🔵🔴
Z_i = 🟢	α🟢🔵	α🟢🟢	α🟢🔴
Z_i = 🔴	α🔴🔵	α🔴🟢	α🔴🔴

Figure 1.10: SBM Modeling of an Interaction Network. Individuals are divided into three communities (green, red, blue). They connect according to a probability defined by their group membership, as given in the table.

the SBM can be more generic, as shown in Table 1.1 which details some possible extensions.

Also, SBMs have been extended to include additional features, such as node attributes (Zanghi et al., 2010) or temporal dynamics, to better represent the complexity of real networks (Matias and Miele, 2017). For instance, the weighted stochastic block model considers the distribution of edge weights to better capture interactions within and between groups (Airoldi et al., 2008). Furthermore, inference methods, such as variational methods and spectral clustering, have been developed to efficiently estimate model parameters and block memberships in large networks (Daudin et al., 2008; Rohe et al., 2011).

At this point, several fundamental questions arise regarding:

(i) *Identifiability:* Can a unique set of model parameters be determined from the observed data distribution?

(ii) *Convergence:* Does the estimator converge to the true model parameters?

(iii) *Model selection:* How to choose the hyperparameter of this model (*ie*, the number of communities)?

Identifiability of the parameters and convergence of (variational) estimators have been demonstrated by Celisse et al. (2012). Additionally, properties of estimators have been studied. Specifically, Mariadassou and Matias

Model Name (Authors)	**Innovation**	**Latent Variable Distribution**	**Emission Distribution**
Degree-Corrected SBM (Karrer and Newman, 2011)	Models the propensity of nodes to be connected, allowing for better modeling of heterogeneous networks.	$\mathcal{M}\left(1;\boldsymbol{\pi}\right)$.	$\text{Pois}\left(\theta_i\alpha_{kl}\theta_j\right)$, where θ_i represents the propensity of individual i to be connected.
Overlapping SBM (Latouche et al., 2011)	Allows nodes to belong to multiple blocks or communities simultaneously.	$\prod_k B\left(\pi_k\right)$.	$B\left(g(a_{\mathbf{Z}_i\mathbf{Z}_j})\right)$, where g is the sigmoid function, and $a_{\mathbf{Z}_i\mathbf{Z}_j} = \mathbf{Z}_i^T\mathbf{W}\mathbf{Z}_j + \mathbf{Z}_i^T\mathbf{U} + \mathbf{V}^T\mathbf{Z}_j + W^*$.
Hierarchical SBM (Peixoto, 2014)	Introduction of a hierarchical structure to capture communities at multiple levels of granularity.	$\mathbb{P}(\mathbf{Z}_i^{(0)}) = \boldsymbol{\pi}^{(0)}$, $\mathbf{Z}^{(t+1)} = \text{Markov}\left(Z^{(t)}, J\right)$, with J as a transition function.	$B(\alpha_{\mathbf{Z}_i^{(l)}\mathbf{Z}_j^{(l)}})$.
Latent Space SBM (Hoff et al., 2002)	Models proximity between nodes beyond blocks by incorporating a latent Gaussian space.	$\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 I_K)$, with $\mu \in \mathbb{R}^K$ and I_K being the identity matrix of size $K \times K$.	$B\left(f(\mathbf{Z}_i, \mathbf{Z}_j)\right).$, where f is a similarity function.
Dynamic SBM (Matias and Miele, 2017)	Models temporal networks with state transitions to capture community changes over time.	$\mathbf{Z}^{(t+1)} = \text{Markov}\left(Z^{(t)}, \boldsymbol{\pi}\right)$, where $\boldsymbol{\pi}$ is a transition matrix.	$\phi\left(X_{ij}^{(t)}; \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}\right) =$ $\left(1-\beta_{Z_iZ_j}^{(t)}\right)\mathbb{1}_{X_{ij}^{(t)}=0}$ $+\,\beta_{Z_iZ_j}^{(t)} f(X_{ij}, \boldsymbol{\gamma}_{Z_iZ_j}^{(t)})$.

Table 1.1: Summary of extended Stochastic Block Models (not exhaustive).

(2015) proved the convergence of the posterior distribution of blocks to a Dirac mass centered on the true node memberships. Furthermore, the consistency of variational estimators was established by Celisse et al. (2012), while Bickel et al. (2013) proved their asymptotic normality under certain conditions. These results have recently been extended to more complex contexts, including networks with missing data Mariadassou and Tabouy (2020) and bipartite networks (latent block model with Bernoulli emission law) Brault et al. (2020), thereby expanding the applicability of SBMs to more realistic and heterogeneous environments.

Regarding model selection, a lot of research has been done. The review by Lee and Wilkinson (2019) covers a multitude of detailed criteria. Some approaches rely on likelihood modularity criteria (Bickel and Chen, 2009), some on Bayesian information criteria (BIC) (Fienberg et al., 2008; Xing et al., 2010), and others on integrated classification likelihood (ICL) (Biernacki et al., 2000; Côme and Latouche, 2015; Daudin et al., 2008). In Chapter 2, model selection will rely on the latter but will be adapted to the model developed.

Latent Block Models

The Latent Block Model (LBM) is a statistical framework primarily used in unsupervised learning for biclustering (Govaert and Nadif, 2010; Keribin et al., 2017). This model is designed to simultaneously partition the rows and columns of a matrix into clusters, ensuring that the resulting blocks exhibit maximum homogeneity, as shown in Figure 1.11. It can also be viewed as an early fusion and unified representation method, where various modalities are concatenated into a single matrix, and the goal is to identify blocks with common characteristics.

In addition to the observation membership matrix $\mathbf{Z}$, which represents the K *communities*, let $\mathbf{W}$ denote the indicator matrix representing the membership of variable accross Q *components*. Given the latent variables, the observations $\forall i \in \{1, \ldots, N\}$ and $\forall j \in \{1, \ldots, p\}$, follow:

$$X_{ij} \mid Z_{ik} = 1, W_{js} = 1 \sim \mathbb{P}(X_{ij} \mid \alpha_{ks}) \ , \tag{1.6}$$

where α_{ks} represents the parameter of the block distribution formed from community k and component s, as illustrated in Figure1.12.

The LBM is particularly advantageous for analyzing large datasets with

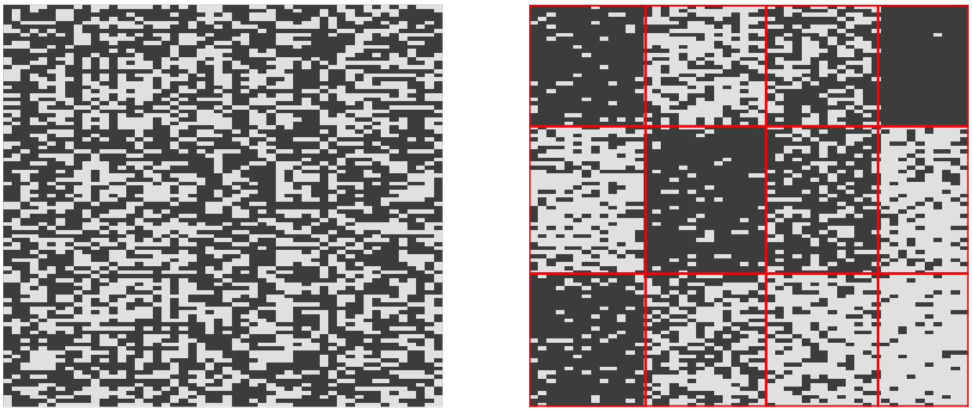


Figure 1.11: Reorganization of a random matrix $\mathbf{X}$: transitioning from a matrix with no apparent structure to a block-wise homogeneous distribution after reorganizing the rows and columns.

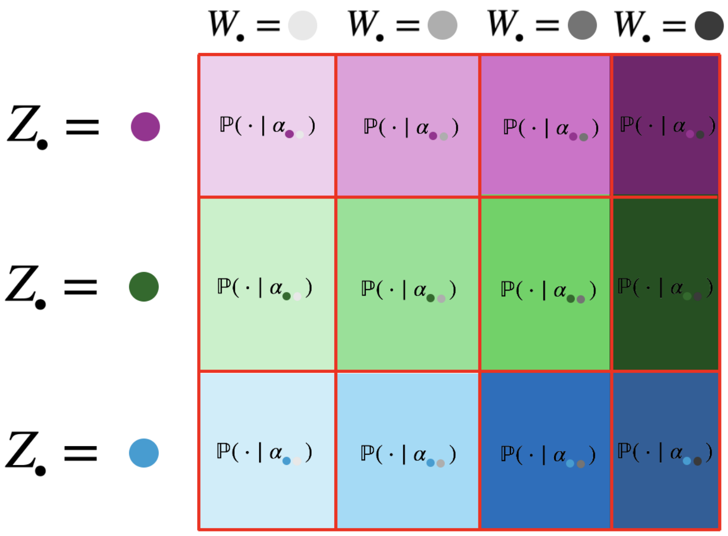


Figure 1.12: Representation of an LBM. This matrix shows the relationships between different latent variables (indicated by Z and W) and the observed data blocks. The colors of the squares correspond to different pairs of community/component within this matrix, as well as to the underlying distribution.

numerous observations and variables because it provides a sparse representation of the data by simultaneously grouping both dimensions (Keribin et al., 2017). However, the model faces several numerical challenges: maximum likelihood estimation can be compromised by computational complexity issues, convergence of estimators to a local maximum depending on the optimization chosen, and the complexity of selecting an appropriate model due to the model's dependence on hyperparameters (Govaert and Nadif, 2008).

To mitigate these problems, Bayesian inference is often used to achieve numerical stability and consistency of estimators, and the Integrated Completed Likelihood (ICL) is recommended for selecting a coherent LBM with lower complexity (Brault and Mariadassou, 2015).

Many of the extensions described above also apply to LBMs.

Ordinal Data An ordinal variable is a categorical variable with logically ordered categories, though the intervals between values may not be equal.

Jacques and Biernacki (2018) extended the LBM to handle ordinal data by incorporating the Binary Ordinal Search (BOS) distribution (Biernacki and Jacques, 2016). For community k, component s, and with m ordinal levels, this distribution is parameterized by a position parameter $\mu_{ks} \in \{1, \dots, m\}$ and a precision parameter $\xi_{ks} \in [0, 1]$. These parameters provide a flexible way to model ordinal data, ranging from a uniform distribution, with $\xi_{ks} = 0$, to a Dirac distribution centered at μ_{ks}, with $\xi_{ks} = 1$.

Model inference is carried out using a stochastic EM algorithm coupled with Gibbs sampling method (SEM-Gibbs). The optimal number of communities and components is selected using the ICL-BIC criterion.

Functional Data Functional Latent Block Models (FunLBM) uncover latent structures in functional data, such as curves or time series, by partitioning both observations and features representing the variables.

The model developed by Bouveyron et al. (2018) is based on decomposing time series into a finite set of basis functions (Fourier, Legendre, B-spline, etc.). Let us consider a dataset represented by a matrix of N individuals and p features:

$$\mathbf{X}(t) = \left(X_{ij}(t)\right)_{i=1:N, j=1:p} \ , \tag{1.7}$$

where $t \in [0, T]$. This model assumes the following form:

$$X_{ij}(t) = \sum_{h=1}^{m} a_{ijh} \phi_h(t), \quad t \in [0, T] \ . \tag{1.8}$$

For observation i and feature j, the signal is decomposed over the basis $(\phi_1(t), \ldots, \phi_m(t))$, with the associated coefficients $(a_{ijh})_{h=1:m}$.

The model focuses on applying the LBM to the coefficients $(a_{ijh})_{i=1:N, j=1:p, h=1:m}$:

$$\mathbb{P}(\mathbf{a} \mid \boldsymbol{\Theta}) = \sum_{\mathbf{Z} \in \mathcal{Z}} \sum_{\mathbf{W} \in \mathcal{W}} \mathbb{P}(\mathbf{Z} \mid \boldsymbol{\Theta}) \mathbb{P}(\mathbf{W} \mid \boldsymbol{\Theta}) \mathbb{P}(\mathbf{a} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\Theta}), \tag{1.9}$$

where $\mathbf{Z}_i$ represents the latent variable indicating the community of individual i, and $\mathbf{W}_j$ represents the latent variable indicating the component of feature j.

The coefficients $\mathbf{a}_{ij} \in \mathbb{R}^m$ are modeled as an m-variate Gaussian vector, conditioned on the community k and component s, with parameters $(\boldsymbol{\mu}_{ks}, \boldsymbol{\Sigma}_{ks})$, where $\boldsymbol{\mu}_{ks}$ is the mean vector, and $\boldsymbol{\Sigma}_{ks}$ is the covariance matrix:

$$\mathbb{P}(\mathbf{a}_{ij} \mid Z_{ik} = 1, W_{js} = 1, \boldsymbol{\theta}_{ks}) = \mathcal{N}(\mathbf{a}_{ij}; \boldsymbol{\mu}_{ks}, \boldsymbol{\Sigma}_{ks}). \tag{1.10}$$

Bouveyron et al. (2018) further refine the modeling of $(\boldsymbol{\mu}_{ks}, \boldsymbol{\Sigma}_{ks})$ by incorporating a low-rank structure based on Principal Component Analysis.

Model inference is performed using a SEM-Gibbs algorithm, and the ICL criterion is employed to select optimal hyperparameters.

Two extensions of this model have been proposed: (i) Goffinet et al. (2020) extend the FunLBM framework by incorporating conditional partitioning between communities and components. In this model, the partition of variables into components remains stable across communities, but the communities vary depending on the components. (ii) Goffinet et al. (2021) introduce a nonparametric approach for the co-clustering of multivariate time series, which does not make prior assumptions about the number of communities or components. The model selection process is thus naturally integrated within the method.

Multiway Clustering Multiway clustering can be seen as a natural extension of co-clustering to simultaneously group observations, variables, as well as additional dimensions or features. This approach proves particularly useful for analyzing complex tensor-structured datasets, where simple data aggregation or independent analysis of each dimension fails to capture the full richness of information.

Robert et al. (2015) propose a novel statistical model for pharmacovigilance, called the Multiple Latent Block Model (MLBM). This model performs simultaneous clustering of the rows and columns of two binary data matrices, enforcing the same row partition across both. The objective is to identify groups of individuals with similar medication profiles, as well as subgroups of interacting drugs and adverse effects. MLBM can be considered as a form of multiway clustering because it jointly analyzes two related datasets (drugs and adverse effects) connected through a third entity (individuals).

The model assumes that the binary variables follow Bernoulli distributions, with parameters dependent on the latent blocks to which both rows and columns belong. Moreover, Robert et al. (2015) provide sufficient conditions to ensure model identifiability, along with an estimation algorithm that combines Gibbs sampling with a variational Bayes (V-Bayes) approach. The Bayesian framework helps avoid degenerate solutions and facilitates the estimation of partitions.

Marchello et al. (2022) introduce the dynamic Latent Block Model (dLBM), a generative model that extends the classic LBM to the case of dynamic count data. The primary objective of the dLBM is to simultaneously cluster the rows, columns, and time slices of an evolving count matrix. The model assumes that the number of interactions between rows and columns follows an inhomogeneous Poisson process, with intensity λ_{ksh} depending on the latent clusters : the communities k, components s, and time intervals h.

The dLBM allows for the identification of structural changes in group interactions (communities/components) over time. The SEM-Gibbs algorithm is used for model inference, and the ICL criterion is employed to select the optimal number of clusters.

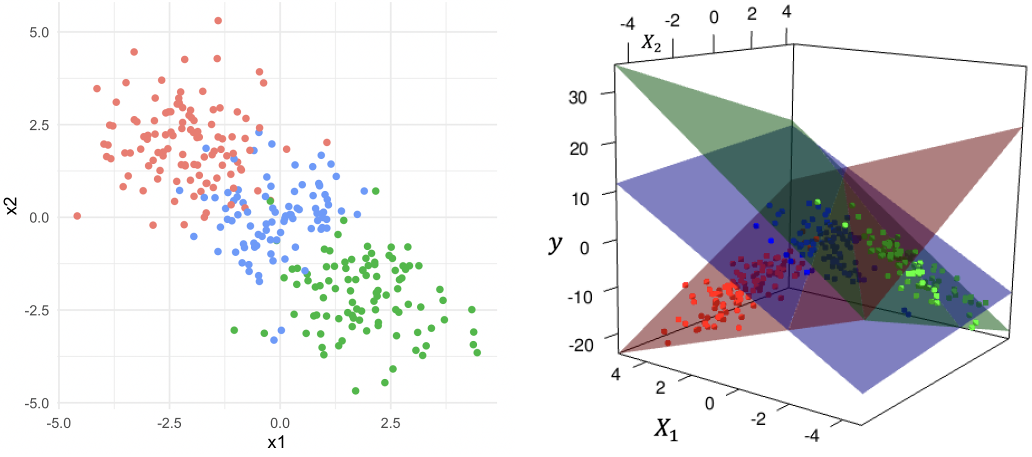
For a comprehensive overview on this topic, we encourage the reader to consult the survey by Biernacki et al. (2023).

1.2.3 Mixture of experts

Now, we turn our attention to the predictive task, where the objective is to model the relationship between the feature matrix $\mathbf{X}$ and the target vector $\mathbf{y}$. Mixtures of Experts (MoEs) are particularly suited for effective prediction while simultaneously identifying subgroups of individuals. Indeed, stratification of individuals is crucial for leveraging underlying structures and improve predictive accuracy by considering the specific characteristics of subgroups.

The concept of MoEs dates back to the early 1990s, originating with the work of Jacobs et al. (1991). Their approach aimed to improve the generalization capability of neural networks by combining multiple specialized networks instead of relying on a single monolithic model (Masoudnia and Ebrahimpour, 2014). This methodology fits within a broader trend in supervised learning, where complex tasks are decomposed into smaller, more manageable sub-tasks, each handled by a separate model. Such a divide-and-conquer strategy allows experts to focus on distinct aspects of the problem, ultimately leading to a more accurate and robust overall solution (Yuksel et al., 2012).

The core mechanism behind MoEs rests on two main components: the experts and the gating network (Gormley and Frühwirth-Schnatter, 2019). The experts consist of independant models — often neural networks or regressions — each trained to specialize in a specific subset of the data. The gating network evaluates the input data to assign relevance weights to each expert, which are then used to combine their outputs into a global prediction that leverages the strengths of all experts.



(a)

(b)

Figure 1.13: Example of a regression mixture model. (a) The graph shows three distinct groups of data (red, blue, and green). Each point is characterized by the variables X^1 and X^2. (b) Each group is associated with a hyperplane relating y to X^1 and X^2, as well as the group membership. The regression parameters $(\boldsymbol{\beta}_{\bullet}, \boldsymbol{\beta}_{\bullet}, \boldsymbol{\beta}_{\bullet})$ will need to be estimated.

Let us consider a dataset $\mathcal{D}_N = \{(\mathbf{X}_i, y_i)_{i=1:N}\}$, where $\mathbf{y}$ represents the target vector. We also have a set of K *experts*, denoted as $(f_k)_{k=1:K}$, each parameterized by $(\boldsymbol{\beta}_k)_{k=1:K}$, as depicted in Figure 1.13.

The general model formulation can be expressed as:

$$\mathbb{P}\left(y_i \mid \mathbf{X}_i\right) = \sum_{k=1}^{K} g(\mathbf{X}_i)_k f_k(y_i; \mathbf{X}_i, \boldsymbol{\beta}_k) \ , \tag{1.11}$$

where $g(\mathbf{X}_i)_k$ represents the k-th output of the gating network g, which gives the probability that the observation $\mathbf{X}_i$ is assigned to expert k for prediction.

Alternatively, we can model this problem using latent variables $(Z_i)_{i=1:N}$, which serve the same role as g:

$$Z_i \mid \mathbf{X}_i \sim \mathcal{M}\left(1; g(\mathbf{X}_i) = \left(g_1(\mathbf{X}_i), \cdots, g_K(\mathbf{X}_i)\right)\right), \tag{1.12}$$

where the gating function $g : \mathcal{X} \to]0,1[^K$ satisfies $\sum_{k=1}^{K} g(\mathbf{X}_i)_k = 1$.

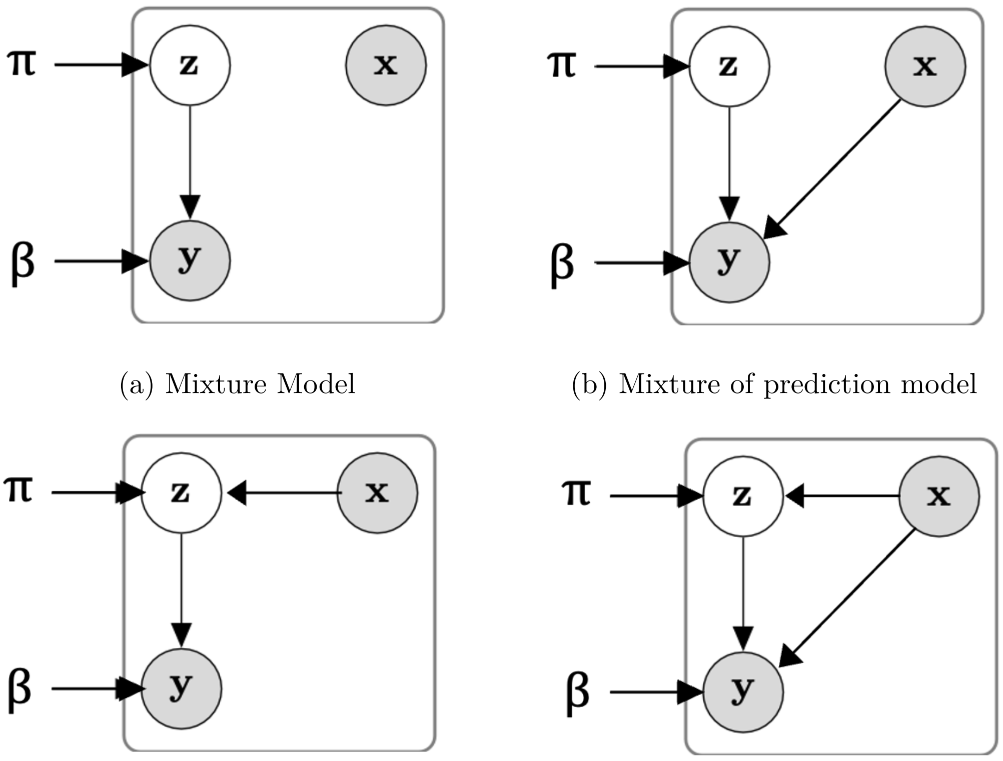
Remark 1 *The function g is typically modeled as a sigmoid function when $K = 2$, or a softmax function when $K > 2$.*

The recent surge in transformer models (Vaswani, 2017) and the widespread success of ChatGPT have reignited interest in MoEs. On one hand, MoEs facilitate the creation of large-scale, flexible models by activating only a subset of experts during inference, thus reducing computational complexity compared to fully dense networks. However, they still face challenges due to their high number of parameters, and training the gating network for expert selection, can be complex and computational demanding.

A notable advancement in scaling MoEs was presented by Lepikhin et al. (2020) with GShard, which introduced an automated mechanism for distributing workloads across multiple experts. This innovation enabled the creation of models with hundreds of billions of parameters while retaining computational efficiency. Building upon this, Fedus et al. (2022) demonstrated further reductions in computational costs compared to GShard, showing that activating a single expert could still achieve comparable performance.

Recent research has also explored various strategies to improve MoE architectures, such as expert specialization (Zadouri et al., 2023), partial activation schemes (Clark et al., 2022), and advanced optimization techniques (Shen et al., 2023). These developments continue to push the boundaries of MoEs' performance and efficiency.

In Chapter 3, we present an interpretable MoE model based on a structure derived from the latent block model combined with experts. The assumption



(a) Mixture Model

(b) Mixture of prediction model

(c) Simple mixture of experts model

(d) Standard mixture of experts model

Figure 1.14: Different types of expert models, based on the representation from Gormley and Frühwirth-Schnatter (2019). The variables x and y are known, z (group membership) needs to be estimated, as well as the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$. Model (a) is $\mathbb{P}(y, z \mid x) = \mathbb{P}(y \mid z)\,\mathbb{P}(z)$, model (b) is $\mathbb{P}(y, z \mid x) = \mathbb{P}(y \mid z, x)\,\mathbb{P}(z)$, model (c) is $\mathbb{P}(y, z \mid x) = \mathbb{P}(y \mid z)\,\mathbb{P}(z \mid x)$ and model (d) is $\mathbb{P}(y, z \mid x) = \mathbb{P}(y \mid z, x)\,\mathbb{P}(z \mid x)$.

of LBM structure of the data facilitates both individual stratification, acting as the gating network, and variable grouping into components. This approach lowers computational cost by reducing the number of variables utilized by the experts.

1.3 Optimization methods

We have explored the fundamental concepts and motivations behind the presented models. However, to maximize their effectiveness, it is necessary to optimize their parameters based on the data used during training.

In the context of estimating latent variable models, the Expectation-Maximization

(EM) algorithm is frequently used (Dempster et al., 1977). The general idea of the EM algorithm is to alternate between two steps:

1. The E-step : Compute the conditional expectation of the complete log-likelihood based on the current parameter estimates to evaluate the latent variables.

2. The M-step : Maximize this expectation to update the parameter estimates.

However, estimating the latent variables and optimizing the parameters often require computing complex integrals. To address this issue, two main approaches are used:

1. Maximization of a lower bound on the model likelihood: This approach, known as variational methods, involves optimizing an approximation of the posterior distribution of the latent variables. It transforms the inference problem into an optimization problem, which is often easier to solve (Blei et al., 2017; Fox and Roberts, 2012).

2. Direct estimation of the integral using Markov Chain Monte Carlo (MCMC) methods: This approach utilizes sampling techniques to approximate the integral that is difficult to compute analytically (Gilks et al., 1995).

This section provides a review of various inference methods for model estimation, highlighting their strengths and limitations in different contexts. However, our focus will shift to two specific approaches that are central to the models developed in the following chapters. In Chapter 2, model inference will be based on the variational Bayesian approach to the EM algorithm while in Chapter 3, the inference for the developed model will rely on SEM-Gibbs algorithm, a stochastic EM approach combined with a Gibbs sampling algorithm.

1.3.1 Variational approach

The direct EM approach has certain limitations, especially in the presence of complex models where exact posterior distributions are inaccessible. To address these limitations, a variational approach to EM has been developed (Jordan et al., 1999; Blei et al., 2017). It relies on the idea of approximating the posterior distribution of the latent variables using a simpler distribution.

More formally, in the context of the variational approach to the EM algorithm, the goal is to maximize the likelihood of the parameters $\mathbf{\Theta}$, which is equivalent to maximizing a lower bound on the marginal log-likelihood of:

$$\mathbb{P}\left(\mathbf{X} \mid \mathbf{\Theta}\right) = \sum_{\mathbf{Z} \in \mathcal{Z}} \mathbb{P}(\mathbf{X} \mid \mathbf{Z}, \mathbf{\Theta})\, \mathbb{P}(\mathbf{Z} \mid \mathbf{\Theta}) \ , \tag{1.13}$$

where $\mathbf{Z}$ represents the set of latent variables.

This sum is often intractable due to the exponential number of possible configurations for $\mathbf{Z}$, especially in network models where the number of summations grows combinatorially with the size of the network (Nowicki and Snijders, 2001; Latouche et al., 2011). It is precisely to solve this problem that variational approaches have been proposed, allowing this intractable sum to be approximated by a variational lower bound, thus facilitating parameter estimation in a more efficient manner (Latouche et al., 2012).

To achieve this, the posterior distribution $\mathbb{P}\left(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta}\right)$ of the latent variables $\mathbf{Z}$ is approximated by a distribution $q\left(\mathbf{Z}\right)$ belonging to a simpler family, often chosen as a factorizable distribution such as mean field approximation (Bernardo et al., 2003). This approximation is then refined by minimizing the Kullback-Leibler divergence between $q\left(\mathbf{Z}\right)$ and $\mathbb{P}\left(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta}\right)$, denoted $\mathbf{KL}\left[q\left(\mathbf{Z}\right) \mid \mathbb{P}\left(\mathbf{Z} \mid \mathbf{X}\right)\right]$.

In the variational step, the goal is to find the distribution $q\left(\mathbf{Z}\right)$ that minimizes this divergence while maximizing a variational lower bound called the *Evidence Lower Bound* (ELBO):

$$\log \mathbb{P}\left(\mathbf{X}\right) = \underbrace{\mathbb{E}_{\mathbf{Z} \sim q}\left[\log \frac{\mathbb{P}\left(\mathbf{X}, \mathbf{Z}\right)}{q\left(\mathbf{Z}\right)}\right]}_{\text{ELBO}} + \mathbf{KL}\left[q\left(\mathbf{Z}\right) \mid \mathbb{P}\left(\mathbf{Z} \mid \mathbf{X}\right)\right]. \tag{1.14}$$

Proof

$$
\begin{aligned}
\log \mathbb{P}(\mathbf{X}) &= \mathbb{E}_{\mathbf{Z}\sim q}\left[\log \mathbb{P}(\mathbf{X})\right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \mathbb{P}(\mathbf{X}) \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{\mathbb{P}(\mathbf{X},\mathbf{Z})\, q(\mathbf{Z})}{\mathbb{P}(\mathbf{Z}\mid\mathbf{X})\, q(\mathbf{Z})}\right] \\
&= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{\mathbb{P}(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]}_{\mathbb{E}_{\mathbf{Z}\sim q}\left[\log \frac{\mathbb{P}(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{q(\mathbf{Z})}{\mathbb{P}(\mathbf{Z}\mid\mathbf{X})}\right]}_{\mathbf{KL}[q(\mathbf{Z})|\mathbb{P}(\mathbf{Z}|\mathbf{X})]} . \qquad (1.15)
\end{aligned}
$$

■

Maximizing this lower bound allows for more efficient optimization of the model parameters compared to the traditional EM algorithm, especially for complex models with a large number of parameters or observations, such as Bayesian graphical models (Wainwright et al., 2008).

The variational approach provides a reasonable approximation in polynomial time, at the cost of a slight loss of precision (Bishop, 2006). Additionally, this method offers a way to manage uncertainties through density estimation by approximate distributions, which is crucial for parameter estimation in Bayesian contexts (Blei et al., 2017).

1.3.2 MCMC approaches

MCMC methods are a category of algorithms used to sample probability distributions by constructing a Markov chain whose stationary distribution is the desired distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. Here are some of the most commonly used methods.

Gibbs Sampling

This technique, introduced by Geman and Geman (1984), is particularly useful in contexts where the target distribution is multidimensional and complex, such as in Bayesian models with a large number of parameters (Casella and

George, 1992; Gelman et al., 1995).

The fundamental principle of Gibbs sampling is that, although directly sampling from a complex joint distribution may be challenging, it is often simpler to sample from the corresponding univariate conditional distributions. Assume a joint distribution $\mathbb{P}(\mathbf{X}_1, \ldots, \mathbf{X}_N)$. Gibbs sampling operates by iterating over each variable $\mathbf{X}_i$ in the sequence, sampling each X_i from its conditional distribution $\mathbb{P}(\mathbf{X}_i \mid \mathbf{X}_{-i})$, where $\mathbf{X}_{-i}$ represents the set of all other variables (Robert et al., 1999).

Convergence is ensured if the underlying Markov chain is irreducible, aperiodic, and reversible. Irreducibility guarantees that any state can be reached from any other state, while aperiodicity prevents systematic cycles. Reversibility, on the other hand, is a property that facilitates proving convergence to the stationary distribution (Tierney, 1994).

Various optimizations and variants have been proposed to improve the efficiency of Gibbs sampling, such as block Gibbs sampling, which allows for the simultaneous updating of multiple variables to reduce the correlation between successive samples and accelerate convergence (García-Cortés and Sorensen, 1996).

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm was developed to solve problems in statistical physics (Metropolis et al., 1953). It is designed to sample from a probability distribution that is difficult to sample from directly, such as the Boltzmann distribution in statistical physics.

The algorithm generates a sequence of random samples from a proposal distribution, accepting or rejecting each sample based on a defined acceptance probability. The proposal distribution is typically chosen to be symmetric with respect to the current state, allowing the algorithm to efficiently explore the space. The acceptance probability is calculated as the minimum of 1 and the ratio of the target distribution at the proposed state to the target distribution at the current state, multiplied by the ratio of the proposal distribution at the current state to the proposal distribution at the proposed state (Hastings, 1970).

Theoretically, the algorithm is guaranteed to converge to the stationary distribution with an infinite number of iterations, though in practice, it may need many iterations to achieve this (Robert et al., 1999).

The Metropolis-Hastings algorithm is particularly useful because it does not require the normalization constant of the probability distribution, which is often difficult to compute (Brooks et al., 2011).

Hamiltonian Monte Carlo (HMC)

Hamiltonian Monte Carlo (HMC) is a sampling method based on the principles of Hamiltonian mechanics, designed to overcome some limitations of traditional MCMC methods. HMC is particularly effective for sampling complex probability distributions in high-dimensional spaces (Pal and Coates, 2019; Duane et al., 1987), using gradients to guide the exploration of the parameter space. This method was popularized in the 1980s by Neal and is now widely used in Bayesian contexts (Neal, 2012; Betancourt, 2017).

The core of the HMC method relies on simulating the motion of a particle in phase space, where the state of the particle is defined by its positions and momenta. The goal is to use Hamiltonian dynamics to efficiently explore the target distribution $\mathbb{P}(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ represents the model parameters. To do this, a potential energy function $U(\boldsymbol{\Theta}) = -\log \mathbb{P}(\boldsymbol{\Theta})$ and a kinetic energy function $K(p)$ are defined, where p is the momentum associated with $\boldsymbol{\Theta}$. The joint distribution of the parameters and momenta is then given by

$$\mathbb{P}\left(\boldsymbol{\Theta}, p\right) = \exp\left[-H\left(\boldsymbol{\Theta}, p\right)\right], \tag{1.16}$$

with $H\left(\boldsymbol{\Theta}, p\right) = U(\boldsymbol{\Theta}) + K(p)$, which is the total Hamiltonian of the system.

The HMC algorithm alternates between two steps: sampling the momenta and evolving the positions using Hamilton's equations of motion. First, the momenta p are sampled from a centered Gaussian distribution, which simulates a random "boost" to the particle (Betancourt and Girolami, 2013). Next, Hamilton's equations are integrated to simulate the system's evolution in the parameter space. This integration often uses the leapfrog scheme, which conserves energy and ensures more efficient exploration of the parameter space (Duane et al., 1987).

Advantages and Disadvantages

Each of these methods has its own strengths and weaknesses. For example, the Metropolis-Hastings algorithm is simple to implement and can be used for a wide range of problems, but it can be slow to converge for high-dimensional distributions. Gibbs sampling is efficient for sampling distributions with strong correlations between variables, but it requires that the conditional distributions can be easily sampled. HMC is also effective for sampling distributions with strong correlations between variables, but it demands more complex computations than the Metropolis-Hastings algorithm. Additionally, while HMC significantly reduces autocorrelation between samples and speeds up convergence, its performance heavily depends on the ability to calculate the gradients of the target probability function. This can be a limitation in cases where these gradients are difficult to obtain or expensive to compute. Finally, the choice between these methods must also take into account the nature of the data and the available computational resources, especially in contexts where precision is crucial but resources are limited.

2 MIxture of Multilayer Integrator Stochastic Block Model

2.1 Introduction 52

2.2 Background 54

2.2.1 Consensus Clustering 54

2.2.2 Block Models for Multiview Clustering 55

2.3 Mixture of Multilayer Stochastic Block Models . 57

2.3.1 A Mixture of Observations across a Mixture of Views 58

2.3.2 Identifiability of the Model 59

2.3.3 Variational Expectation-Maximization for mimi-SBM 60

2.4 Bayesian Framework 62

2.4.1 Evidence Lower Bound 63

2.4.2 Lower Bound Optimization 65

2.5 Initialization and model selection 68

2.5.1 Initialization 68

2.5.2 Model selection 69

2.6 Experiments 71

2.6.1 Simulation study 71

2.6.2 Worldwide Food Trading Networks 81

2.7 Bilan 85

2.1 Introduction

In the clustering framework, the output of algorithms often consists of a partition or a membership matrix **Z**. While this information is useful, direct use of the Z matrix for meta-clustering is a source of disruption, due to possible label-switching. To mitigate this issue, **Z** can be transformed into an adjacency matrix **A** as follows:

$$A_{ij} = \begin{cases} 1, & \text{if individuals } i, j \text{ belong to the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

The adjacency matrix carries the same information about the clusters, grouping the data into blocks. And, when the position of individuals in the matrix is fixed, this representation enables individuals to be characterized by the clusters to which they belong.

Consensus clustering is the process of combining multiple clusterings, generated from different algorithms or approaches, to find commonalities and insights that may not be apparent when examined separately (Monti et al., 2003; Li et al., 2015; Liu et al., 2018). Model-based consensus clustering offers advantages, such as identifying redundancies and complementarities across information sources, producing a final clustering from previously generated outputs, and enabling the best grouping of individuals by utilizing multiple information sources. Moreover, model-based approaches provide evaluation criteria (e.g., log-likelihood, evidence) and, in the Bayesian framework, criteria for model selection (Biernacki et al., 2010).

The learning models differ in their fusion strategies, with three main categories: early, intermediate, and late fusion of views. Late fusion is particularly well-suited for clustering since each view is often associated with specialized clustering algorithms.

Contribution. In this work, we propose estimating a coordinated representation produced by learning separate clustering for each view, by coordination through a probabilistic model: the MIxture of Multilayer Integrator Stochastic Block Model (*mimi-SBM*). Our model is a Bayesian mixture of multilayer SBMs that accounts for multiple information sources, with a shared clustering structure across views, as illustrated in Figure 2.1.

In simpler terms, each individual is assigned to a single group, rather than belonging to multiple groups across views. This meta-clustering approach enables the model to capture common information for each group by emphasizing clustering redundancy across sources. By applying a mixture model to the views, the particularities of each source are accounted for, allowing the model to differentiate between redundant and complementary sources of information, thereby extracting the maximum insight from the data.

Finally, within the Bayesian framework, it is possible to develop a model selection criterion for both the mixture of views and the number of clusters by deriving it from the evidence lower bound. The identifiability of the model parameters is established, and we propose a variational Bayesian EM algorithm for parameter estimation.

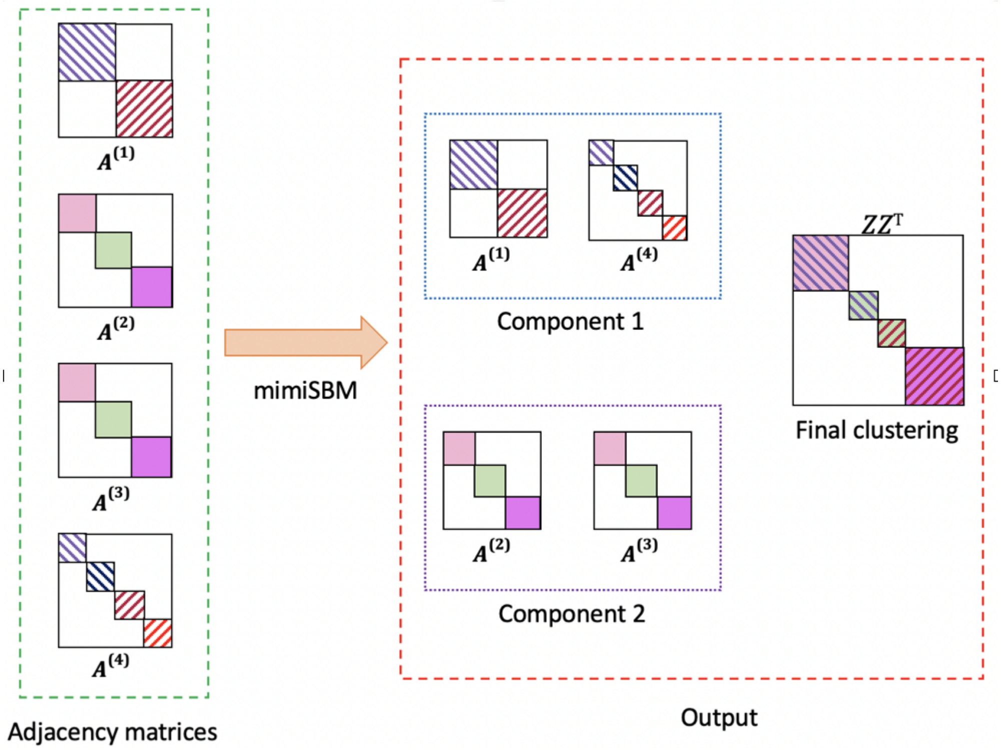


Figure 2.1: Illustration of the mimi-SBM. Left: Four adjacency matrices $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(4)}$ from four different views, organized into two components. Right: Identification of the two components from the views (within: redundant information; between: complementary information), and clustering of the observations described by the classification matrix $\mathbf{Z}$ (global and consensus information).

Organization of the Chapter. This chapter is organized as follows: First, we delve into related works, providing a selective overview of the literature. Then, we describe the innovative *mimi-SBM* model, detailing its key components and parameter estimation. Subsequently, we discuss model selection and variational parameter initialization, comparing different criteria to identify the most effective approach. To assess the performance of our approach, we conduct synthetic experiments. Finally, we conclude with a discussion of the results and potential future research directions.

2.2 Background

In multiview clustering, various fusion strategies have been developed to efficiently combine information from multiple data sources. These strategies are commonly categorized based on the stage at which the fusion occurs: early, intermediate, or late (as discussed in Section 1.1.2). This chapter focuses on the late fusion approach, where the different layers, represented as adjacency matrices, collectively form a tensor.

The advantage of late fusion lies in the ability to apply tailored clustering techniques for each view, leveraging both their unique features and complementary information. The integration of the clustering results in a subsequent phase further enhances robustness and interpretability. In this context, consensus clustering serves as a foundational technique, and will be introduced in this section.

Additionally, stochastic block models offer valuable advantages for multiview clustering, particularly in terms of model selection. In this section, we also present a focused review of these approaches, situating our method within the late fusion clustering framework.

2.2.1 Consensus Clustering

Consensus clustering (Monti et al., 2003; Fred and Jain, 2005; John et al., 2020), also referred to as cluster ensemble (Strehl and Ghosh, 2002; Golalipour et al., 2021), is a technique designed to derive a single partition from multiple clustering solutions.

It is used to aggregate and analyze diverse clustering results obtained from different algorithms, parameter configurations, or subsets of data. The objective is to find a consensus among the individual clustering solutions, leading to a more robust and reliable clustering outcome. This process is akin to soliciting multiple opinions (clusters) and then synthesizing a common consensus that captures the overall agreement.

Each clustering run generates a set of clusters that can be represented as a partition matrix, where each entry indicates the cluster assignment of a data point. From all partition matrices, an agreement matrix is constructed, where each entry reflects how often a pair of data points is assigned to the same cluster across all solutions. A clustering algorithm is then applied to the agreement matrix to determine the final partition.

A specific example is the Monte Carlo reference-based consensus clustering (John et al., 2020, M3C), which generates multiple clustering solutions by applying various algorithms and parameter settings to the same dataset, often incorporating random (re)sampling techniques.

2.2.2 Block Models for Multiview Clustering

In the context of multiview clustering, block models typically define the views as a collection of V *graphs*, often referred to as V *layers* within a network. Terminologies such as *multigraph*, *multilayer*, or *multiplex* networks are also frequently used.

Given the extensive literature on this topic, we narrow our focus to studies where different views represent distinct types of interactions among a common set of N observations. Importantly, we exclude works that focus on overlapping partitions, mixed memberships, or dependencies across views, such as spatial or temporal dependencies, which are addressed in other studies.

Multilayer SBM

Multilayer stochastic block models aim to identify a partition $\mathbf{Z}$ with K blocks of observations, consistent across the layers. Variational Expectation Maximization algorithms are commonly used to infer these models (Daudin et al., 2008). In the MLSBM framework, various estimation methods can be applied to recover the partition, with spectral clustering being widely used (Von Luxburg, 2007; Von Luxburg et al., 2008).

VEM Inference. Several approaches based on SBM have been proposed for multilayer (Han et al., 2015; Paul and Chen, 2016) and multiplex (Barbillon et al., 2017) networks. Han et al. (2015) introduces a consistent maximum likelihood estimator (MLE) and investigates the asymptotic behavior of class memberships as the number of layers increases. Similarly, Paul and Chen (2016) studies the consistency of MLEs when both the number of nodes and types of edges grow. Barbillon et al. (2017) proposes an Erdös-Rényi model that incorporates covariates related to pairs of observations, employing the Integrated Completed Likelihood (ICL) criterion Biernacki et al. (2010) for model selection. Boutalbi et al. (2021), on the other hand, proposes a VEM-based approach grounded in Latent Block Models (LBM).

Spectral Clustering. Significant attention has been devoted to spectral clustering in the context of multilayer SBM. Han et al. (2015) explores its asymptotic properties, while Chen and Hero (2017) proposes a multilayer spectral clustering framework with adaptive layer weighting, providing theoretical guarantees for the clustering's reliability. Mercado et al. (2018) develops a spectral clustering algorithm for multilayer graphs using the matrix power mean of Laplacians. Additionally, Paul and Chen (2020) proves the consistency of co-regularized spectral clustering and orthogonal linked matrix factorization. Finally, Huang et al. (2022) presents integrated spectral clustering methods based on convex layer aggregations.

Multiway Block Models

Unlike the previous approaches that focus on partitions across observations, multiway block models aim to identify between- and within-layer structures. Depending on the specific framework, multilayer SBM can extend into Mixture of Multilayer SBM (MMLSBM) or Tensor Block Models (TBM).

Mixture of Multilayer SBM. Stanley et al. (2016) introduces an early approach combining multilayer SBM with layer mixtures using a two-step greedy algorithm. Initially, SBM is applied to each layer, grouping SBMs with similar parameters. These results serve as an initialization for an iterative algorithm that simultaneously identifies Q strata across V layers. Each stratum s contains K_s blocks of observations, leading to Q membership matrices $\{\mathbf{Z}^1, \cdots, \mathbf{Z}^Q\}$.

In a similar vein, Fan et al. (2022) proposes an alternating minimization algorithm with theoretical guarantees for clustering errors across layers and observations. Rebafka (2023) presents a Bayesian approach for a finite mixture of MLSBM, employing hierarchical agglomerative clustering based on the ICL criterion for model selection.

Additionally, Pensky and Wang (2021) proposes a versatile model for multiplex networks, encompassing both MLSBM and MMLSBM. In this model, the assumption is that the number of blocks K_s is consistent across groups of layers ($K_s = K$, $\forall s$). Based on this, Noroozi and Pensky (2022) introduces a sparse subspace clustering algorithm, demonstrating its effectiveness for between-layer clustering.

Tensor Block Models. An alternative approach to late fusion multiview clustering involves modeling data as tensors, several models are detailed in section 1.2. Wang and Zeng (2019) formulates a least-squares estimator for sparse Tensor Block Models, proving consistency in recovering block structures as the tensor's dimensionality grows. Han et al. (2022) employs high-order spectral clustering as an initialization for a high-order Lloyd algorithm, providing statistical guarantees under sub-Gaussian noise assumptions. Jing et al. (2021, TWIST) introduces a Tucker decomposition-based approach for regularized low-rank tensor approximations, achieving consensus clustering across layers via k-means applied to the local membership matrix.

2.3 Mixture of Multilayer Stochastic Block Models

In this work, we propose a novel approach that builds upon the Stochastic Block Model framework by incorporating two distinct sets of latent variables: one that governs the structure of observations and another that governs the structure of the views. This model bridges the gap between the Multilayer SBM, which identifies a shared community membership across multiple layers, and the Mixture Multilayer SBM, which uncovers distinct latent structures within each layer.

Observations. We represent the observed data as a tensor $\mathbf{A} \in \{0,1\}^{N\times N\times V}$, where N denotes the number of observations (vertices) and V the number of views. Each of the V slices of $\mathbf{A}$ corresponds to an adjacency matrix representing a graph $\mathcal{G}^v$, leading to a stack of adjacency matrices for multiple-view graphs $(\mathcal{G}^1, \ldots, \mathcal{G}^V)$, all sharing the same vertices. An edge between two observations i and j in view v is represented by $A_{ijv} = \mathbb{I}[(i,j) \in E^v]$, where E^v is the edge set of graph $\mathcal{G}^v$.

Latent Structures. Let $\mathbf{Z} \in \{0,1\}^{N\times K}$ represent the community membership matrix of the observations, where K is the number of communities shared across views. Specifically, $Z_{ik} = \mathbb{1}_{i\in k}$, where i is an observation and k represents a community across all views.

Similarly, let $\mathbf{W} \in \{0,1\}^{V\times Q}$ represent the community membership matrix for the views, where Q is the number of view clusters. Here, $W_{vs} = \mathbb{1}_{v\in s}$, where v denotes a view and s a cluster of views.

2.3.1 A Mixture of Observations across a Mixture of Views

We assume that the V views are generated by a mixture of Q components, with each component representing an SBM. The membership of views across these components is modeled as a multinomial distribution $\mathbf{W}_v \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q))$, with the likelihood given by:

$$\mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho}) = \prod_{v=1}^{V} \prod_{s=1}^{Q} \rho_s^{W_{vs}}. \tag{2.2}$$

Despite the use of multiple views with distinct cluster structures, we assume the existence of a shared community structure across all views, represented by the latent variable $\mathbf{Z}$. This structure allows us to integrate information from all available views, resulting in more consistent community assignments across the views. We assume that individuals are drawn from one of K latent communities.

Each observation's membership vector $\mathbf{Z}_i$ follows a multinomial distribution, $\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \dots, \pi_K))$, with the probability:

$$\mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{Z_{ik}}. \tag{2.3}$$

Conditional Likelihood. Each entry A_{ijv}, conditioned on the latent variables $\mathbf{Z}$ and $\mathbf{W}$, follows a Bernoulli distribution

$$A_{ijv} \mid Z_{ik} = 1, Z_{jl} = 1, W_{vs} = 1 \sim \mathcal{B}(\alpha_{kls}). \tag{2.4}$$

The likelihood of the observed data given the latent variables and model parameters is:

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\Theta}) = \prod_{\substack{i=1,\\ i<j}}^{N} \prod_{\substack{k=1\\ l=1}}^{K} \prod_{v=1}^{V} \prod_{s=1}^{Q} \left(\alpha_{kls}^{A_{ijv}} (1-\alpha_{kls})^{1-A_{ijv}}\right)^{Z_{ik}Z_{jl}W_{vs}}. \tag{2.5}$$

2.3.2 Identifiability of the Model

We now turn to the identifiability of the model parameters. Identifiability ensures that the model parameters are uniquely determined by the observed data. We aim to prove that if $\mathbb{P}(\mathbf{A}; \boldsymbol{\Theta}) = \mathbb{P}(\mathbf{A}; \boldsymbol{\Theta}')$, then $\boldsymbol{\Theta} = \boldsymbol{\Theta}'$. Our proof builds on the identifiability results for SBMs as established by Celisse et al. (2012), which are valid up to a permutation of block labels.

Theorem 1 *Let $N \geq \max(2K, 4Q)$ and $V \geq 2K$. Assume that for all $k, l \in \{1, \dots, K\}$ and $s \in \{1, \dots, Q\}$, the coordinates of $\boldsymbol{\pi}^\top \boldsymbol{\alpha}_{k..}\boldsymbol{\rho}$ are distinct, $(\boldsymbol{\pi}^\top \boldsymbol{\alpha}_{..s}\boldsymbol{\pi})_{s=1:Q}$ are distinct, and each $(\boldsymbol{\alpha}_{kl.}\boldsymbol{\rho})_{k,l=1:K}$ differs. Then, the model parameters $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ are identifiable.*

Proof The proof is detailed in Appendix A. A key element in the proof is the identifiability of the $\boldsymbol{\alpha}$ parameters, which is derived by drawing a parallel between our model and the identifiability results for Latent Block Models demonstrated by Keribin et al. (2015). ■

2.3.3 Variational Expectation-Maximization for mimi-SBM

One of the main challenges when working with Stochastic Block Models is computing the marginal likelihood:

$$\mathbb{P}\left(\mathbf{A}\right) = \sum_{\mathbf{Z}} \sum_{\mathbf{W}} \mathbb{P}\left(\mathbf{A}, \mathbf{Z}, \mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right). \tag{2.6}$$

This summation over latent variables $\mathbf{Z}$ and $\mathbf{W}$ becomes computationally prohibitive as the number of parameters or observations increases. To address this, we employ the Variational Expectation-Maximization algorithm, which offers a computationally efficient alternative by approximating the posterior distribution (Brault, 2014).

Evidence Lower Bound (ELBO). The goal of the variational approach is to maximize the Evidence Lower Bound (ELBO), defined as:

$$\mathcal{L}(q) = \mathbb{E}_{\mathbf{Z},\mathbf{W}\sim q(\cdot)}\left[\log \mathbb{P}\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}\right)\right] - \mathbb{E}_{\mathbf{Z},\mathbf{W}\sim q(\cdot)}\left[\log q\left(\mathbf{Z}, \mathbf{W}\right)\right], \tag{2.7}$$

where q represents a variational distribution over the latent variables. Following the work of Tabouy et al. (2020), we model the latent variables $\mathbf{Z}$ and $\mathbf{W}$ with multinomial priors, as:

$$q(\mathbf{Z}, \mathbf{W}) = \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{v=1}^{V} q(\mathbf{W}_v) = \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\tau}_i) \prod_{v=1}^{V} \mathcal{M}(\mathbf{W}_v; 1, \boldsymbol{\nu}_v), \tag{2.8}$$

where τ_{ik} and ν_{vs} are variational parameters that estimate the probability of individual i belonging to cluster k, and view v belonging to component s, respectively.

The VEM Algorithm. The VEM algorithm alternates between two steps:

- Variational Expectation (VE): In this step, we maximize the ELBO with respect to the variational parameters $(\tau_{ik})_{i=1:N,k=1:K}$ and $(\nu_{vs})_{v=1:V,s=1:Q}$.
- Maximization (M): We optimize the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, and $\boldsymbol{\alpha}$ by maximizing the expectation computed during the VE step.

After deriving the update rules, which are detailed in Appendix B, we obtain the following optimal estimates for the model parameters at each iteration of the algorithm:

Community Membership Probability. The optimal update for τ_{ik}, representing the probability that observation i belongs to cluster k, is given by:

$$\tau_{ik} = \frac{\exp(T_{ik})}{\sum_{k'=1}^{K} \exp(T_{ik'})}, \tag{2.9}$$

where

$$T_{ik} = \sum_{j \neq i}^{N} \sum_{l=1}^{K} \sum_{v=1}^{V} \sum_{s=1}^{Q} \tau_{jl} \nu_{vs} \left[A_{ijv} \log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right) \log\left(1 - \alpha_{kls}\right) \right] + \log\left(\pi_k\right). \tag{2.10}$$

The optimization problem is as a fixed-point problem, requiring successive iterations until convergence to a solution is achieved.

Component Membership Probability. The optimal update for ν_{vs}, representing the probability that view v belongs to component s, is given by:

$$\nu_{vs} = \frac{\exp\left(R_{vs}\right)}{\sum_{s'=1}^{Q} \exp\left(R_{vs'}\right)}, \tag{2.11}$$

where

$$R_{vs} = \sum_{\substack{i=1, \\ i<j}}^{N} \sum_{\substack{k=1, \\ l=1}}^{K} \tau_{ik} \tau_{jl} \left[A_{ijv} \log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right) \log\left(1 - \alpha_{kls}\right) \right] + \log\left(\rho_s\right). \tag{2.12}$$

Estimation of Model Parameters. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, and $\boldsymbol{\alpha}$ are updated as follows:

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{ik}}{N}, \qquad \forall k \in \{1, \ldots, K\}, \tag{2.13}$$

$$\rho_s = \frac{\sum_{v=1}^{V} \nu_{vs}}{V}, \qquad \forall s \in \{1, \ldots, Q\}. \tag{2.14}$$

Finally, the optimal estimate for the interaction probability α_{kls} is given by:

For $k \neq l$,

$$\alpha_{kls} = \frac{\sum_{i=1}^{N} \sum_{j \neq i} \sum_{v=1}^{V} \tau_{ik} \tau_{jl} \nu_{vs} A_{ijv}}{\sum_{i=1}^{N} \sum_{j \neq i} \sum_{v=1}^{V} \tau_{ik} \tau_{jl} \nu_{vs}}. \tag{2.15}$$

For $k = l$,

$$\alpha_{kks} = \frac{\sum_{i=1}^{N} \sum_{j > i} \sum_{v=1}^{V} \tau_{ik} \tau_{jk} \nu_{vs} A_{ijv}}{\sum_{i=1}^{N} \sum_{j > i} \sum_{v=1}^{V} \tau_{ik} \tau_{jk} \nu_{vs}}. \tag{2.16}$$

The parameter estimates are similar to those obtained in the classical SBM. The contribution to parameter estimation depends on the probability that individuals i and j belong to communities k and l, respectively, as well as the probability that view v is part of component s.

2.4 Bayesian Framework

Bayesian modeling offers a robust framework for integrating prior knowledge, enhancing the accuracy of the estimated block structure, especially in scenarios where the available data is limited or noisy.

In this context, we adopt the approach proposed by Latouche et al. (2012) and define the selected conjugate distributions for both the mixture proportions and the block proportions. Utilizing conjugate priors allows for the derivation of closed-form posterior distributions.

$$\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{\beta}^0 = (\beta_1^0, \dots, \beta_K^0)) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}^0), \tag{2.17}$$

$$\mathbb{P}(\boldsymbol{\rho} \mid \boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_Q^0)) = \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}^0), \tag{2.18}$$

where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution.

$$\mathbb{P}(\boldsymbol{\alpha} \mid \boldsymbol{\eta}^0 = (\eta_{kls}^0), \boldsymbol{\xi}^0 = (\xi_{kls}^0)) = \prod_{k,k<l} \prod_{s} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}^0, \xi_{kls}^0). \tag{2.19}$$

The parameters $\boldsymbol{\beta}^0, \boldsymbol{\theta}^0, \boldsymbol{\eta}^0, \boldsymbol{\xi}^0$ require careful selection, as they can significantly influence the optimization process (Kass and Wasserman, 1996). Generally, priors with lower values (close to zero) tend to encourage sparser distributions, which can be beneficial for promoting model parsimony. Conversely, priors with higher values yield more uniform distributions, which may be preferable in situations where prior knowledge about the data structure is minimal.

In our study, we select priors according to Jeffreys priors, often regarded as

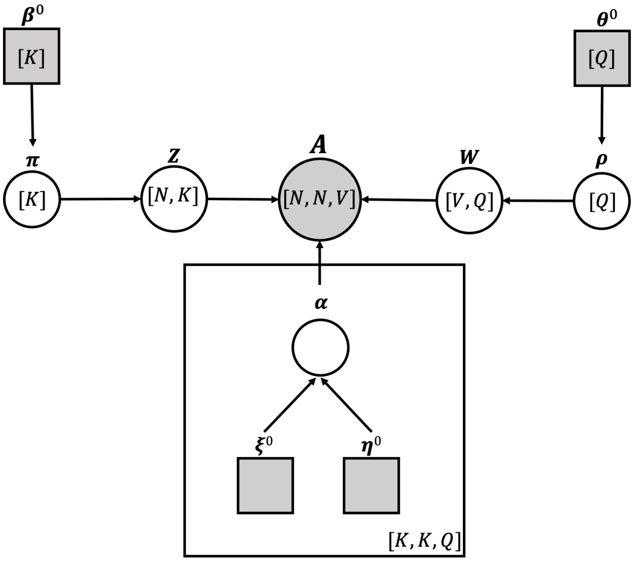


Figure 2.2: Illustration of the mimi-SBM with Bayesian notations

non-informative or weakly informative (Jeffreys, 1946), as they do not impose strong prior assumptions or biases on the analysis.

For the Dirichlet distribution, a suitable choice for β_k^0 and θ_s^0 is to set them both to $\frac{1}{2}$, corresponding directly to an objective Jeffreys prior distribution. Likewise, for the Beta distribution, we can choose η_{kls}^0 and ξ_{kls}^0 to be $\frac{1}{2}$ for all relevant indices k, l, and s.

A more comprehensive examination of the priors was conducted in the thesis of Brault (2014).

2.4.1 Evidence Lower Bound

The approximation of complex posterior distributions is typically achieved through either sampling methods (such as Markov Chain Monte Carlo) or Variational Bayes inference, as introduced by Attias (1999). The Variational Bayes Expectation Maximization algorithm exhibits favorable statistical properties for estimators derived from the mean-field approximation (Brault, 2014; Keribin, 2010).

In the context of the Stochastic Block Model (SBM), the distribution $\mathbb{P}(\mathbf{Z}, \mathbf{W}|\mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})$ is intractable. Therefore, we approximate the entire distribution $\mathbb{P}(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho} \mid \mathbf{A})$ by a distribution q. Given a variational distribution q over the variables $\{\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\}$, we can decompose the marginal log-likelihood into two parts: the Evidence Lower Bound (ELBO) and the

Kullback-Leibler (KL) divergence between the variational and posterior distributions:

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

The KL divergence is expressed as follows:

$$\mathbf{KL}\left(q\|\mathbb{P}\right) = -\mathbb{E}_q[\log \frac{\mathbb{P}}{q}] \geq -\log \mathbb{E}_q[\frac{\mathbb{P}}{q}] \geq 0,$$

which follows from Jensen's inequality.

The ELBO is defined as:

$$\mathcal{L}\left(q(\cdot)\right) = \sum_{\mathbf{Z},\mathbf{W}} \int\int\int q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho}) \log \frac{p\left(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho}\right)}{q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})}\ d\boldsymbol{\alpha}\ d\boldsymbol{\pi}\ d\boldsymbol{\rho}. \tag{2.20}$$

The variational distribution is typically chosen from a more tractable family of distributions, such as the exponential family. The parameters of the variational distribution are adjusted to minimize the KL divergence from the posterior distribution. If $q(.)$ precisely equals $p(.|\mathbf{A})$, then the KL term becomes zero, and the ELBO is maximized.

We adopt a mean-field approximation for $q(\cdot)$:

$$\begin{aligned}q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho}) &= \prod_{i=1}^{N} q(\mathbf{Z}_i)\ \prod_{v=1}^{V} q(\mathbf{W}_v)\ \prod_{s=1}^{Q}\prod_{k,k\leq l}^{K} q(\alpha_{kls})\ q(\boldsymbol{\pi})\ q(\boldsymbol{\rho}) \\ &= \mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\beta})\ \mathrm{Dir}(\boldsymbol{\rho};\boldsymbol{\theta}) \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z}_i;1,\boldsymbol{\tau}_i)\ \prod_{v=1}^{V} \mathcal{M}(\mathbf{W}_v;1,\boldsymbol{\nu}_v) \\ &\prod_{s=1}^{Q}\prod_{k,k\leq l}^{K} \mathrm{Beta}(\alpha_{kls};\eta_{kls},\xi_{kls}),\end{aligned} \tag{2.21}$$

where τ_{ik} (and ν_{vs}) are variational parameters representing the probability that individual i (and view v) belongs to cluster k (and component s, respectively).

According to Equation 2.20, given the distribution $q(.)$, the ELBO can be expressed as:

$$\begin{aligned}\mathcal{L}\left(q(.)\right) = \log\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\}\\ + \sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\}\\ - \sum_{i}^{N}\sum_{k}^{K}\tau_{ik}\log\tau_{ik} \ - \sum_{v}^{V}\sum_{s}^{Q}\nu_{vs}\log\nu_{vs},\end{aligned} \tag{2.22}$$

where $\Gamma(.)$ denotes the Gamma function. This expression is also known as the Integrated Likelihood variational Bayes (Latouche et al., 2012, ILvb)), as it can be employed for model selection. Detailed calculations can be found in Appendix D.

2.4.2 Lower Bound Optimization

We consider a Variational Bayes EM algorithm for parameter estimation. The algorithm begins by initializing the model parameters and then iteratively executes two steps: the Variational Bayes Expectation step (VBE-step) and the Maximization step (M-step) (see Algorithm 1).

In the VBE-step, the variational distributions are optimized over the latent variables $q(\mathbf{Z}_i)$ for all $i \in \{1,\dots,N\}$ and $q(\mathbf{W}_v)$ for all $v \in \{1,\dots,V\}$ to approximate the true posterior distribution.

During the M-step, the model parameters are updated to maximize a lower bound on the log-likelihood concerning the parameters computed in the VBE-step: $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\xi}$.

Variational Parameters of Clustering τ_{ik} The optimal approximation for $q(\mathbf{Z}_i)$ is given by:

$$q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i; (\tau_{i1},\dots,\tau_{iK})), \tag{2.23}$$

where τ_{ik} represents the probability of node i belonging to community k. It satisfies the relation:

$$\tau_{ik} \propto e^{\psi(\beta_k)-\psi(\sum_{k'}\beta_{k'})} \prod_{j\neq i}^{N}\prod_{l=1}^{K}\prod_{v=1}^{V}\prod_{s=1}^{Q} e^{\tau_{jl}\nu_{vs}[A_{ijv}(\psi(\eta_{kls})-\psi(\xi_{kls}))+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})]}, \tag{2.24}$$

where ψ denotes the digamma function. The optimization problem is a fixed-point problem, allowing the distribution $q(\mathbf{Z})$ to be optimized using a fixed-point algorithm.

Variational Parameters of Component Membership ν_{vs} The optimal approximation for $q(\mathbf{W}_v)$ is:

$$q(W_v) = \mathcal{M}(W_v; (\nu_{v1}, \dots, \nu_{vQ})), \tag{2.25}$$

with

$$\begin{aligned}\nu_{vs} \propto e^{\psi(\theta_s)-\psi(\sum_{s'}\theta_{s'})} &\prod_{i\neq j}^{N}\prod_{k\neq l}^{K} e^{\tau_{ik}\tau_{jl}[A_{ijv}(\psi(\eta_{kls})-\psi(\xi_{kls}))+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})]} \\ &\prod_{k}^{K}\prod_{i<j}^{N} e^{\tau_{ik}\tau_{jk}[A_{ijv}(\psi(\eta_{kks})-\psi(\xi_{kks}))+\psi(\xi_{kks})-\psi(\eta_{kks}+\xi_{kks})]}.\end{aligned} \tag{2.26}$$

Optimization of $q(\boldsymbol{\pi})$ (β_k) Due to the choice of prior distributions, the distribution $q(\boldsymbol{\pi})$ remains within the same family as the prior distribution $\mathbb{P}(\boldsymbol{\pi})$:

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}), \tag{2.27}$$

with

$$\beta_k = \beta_k^0 + \sum_{i=1}^{N} \tau_{ik}. \tag{2.28}$$

Optimization of $q(\boldsymbol{\rho})$ (θ_k) Given the choice of prior distributions, the distribution $q(\boldsymbol{\rho})$ stays within the same family as the prior $p(\boldsymbol{\rho})$:

$$q(\boldsymbol{\rho}) = \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}), \tag{2.29}$$

with

$$\theta_s = \theta_s^0 + \sum_{v=1}^{V} \nu_{vs}. \tag{2.30}$$

Optimization of $q(\boldsymbol{\alpha})$ (η_{kls} and ξ_{kls}) Once again, the form of the prior distribution $\mathbb{P}(\boldsymbol{\alpha})$ is preserved through the variational optimization process:

$$q(\alpha_{kls}) = \mathrm{Beta}(\alpha_{kls}; \eta_{kls}, \xi_{kls}). \tag{2.31}$$

When $k \neq l$, the parameters η_{kls} and ξ_{kls} are given by:

$$\begin{aligned} \eta_{kls} &= \eta^0_{kls} + \sum_{i \neq j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}, \\ \xi_{kls} &= \xi^0_{kls} + \sum_{i \neq j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}\left(1 - A_{ijv}\right). \end{aligned} \tag{2.32}$$

Otherwise, when $k = l$, the parameters η_{kks} and ξ_{kks} are determined by:

$$\begin{aligned} \eta_{kks} &= \eta^0_{kks} + \sum_{i < j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}, \\ \xi_{kks} &= \xi^0_{kks} + \sum_{i < j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}\left(1 - A_{ijv}\right). \end{aligned} \tag{2.33}$$

Full details of parameter optimization are available in Appendix C.

Algorithm 1 mimi-SBM

```
Require: Tensor of adjacency matrices $\mathbf{A}$, Number of clusters $K$, Number of
    components of the views $Q$, precision $eps$.
    Initialization: $\tau_{ik}^{(old)}$ and $\nu_{ik}^{(old)}$
    while $\|\mathcal{L}\left(q^{new}(.)\right) - \mathcal{L}\left(q^{old}(.)\right)\| < eps$ do
        VBE-step
        Compute $\tau_{ik}^{(new)}$ for all $i \in \{1, \dots, N\}$ and $k \in \{1, \dots, K\}$
        Compute $\nu_{vs}^{(new)}$ for all $v \in \{1, \dots, V\}$ and $s \in \{1, \dots, Q\}$
        M-step
        Optimize $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ with respect to $(\tau_{ik}^{(new)})$ and $(\nu_{vq}^{(new)})$
        ELBO
        Compute $\mathcal{L}\left(q^{new}(.)\right)$
    end while
```

2.5 Initialization and model selection

2.5.1 Initialization

There exist multiple techniques for initializing the EM algorithm. One prevalent approach involves using random initial values, where the model parameters are assigned random values drawn from a designated distribution. Nevertheless, this method may lack reliability and fails to provide satisfactory starting values for the algorithm. It is necessary to initialize our parameters as close to the optimum as possible in order to ensure that the local maximum is as high as possible and to avoid degeneracy (Baudry and Celeux, 2015). According to the initialization method proposed in Stanley et al. (2016), the parameters (τ_{ik}) and (ν_{vs}) are initialized based on the outcomes of a stochastic block model applied separately to each view. The objective is to capture the overall structure of the data from each view using SBM, combine this information using K-means clustering, and subsequently refine the obtained results using our model.

Let define $\mathbf{A}^{(v)}$ the v-th adjacency matrix,

$$\forall v \in \{1, \dots, V\}, \quad \text{SBM}(\mathbf{A}^{(v)}) \to \left(\mathbf{Z}^{(v)}, \boldsymbol{\tau}^{(v)}, \boldsymbol{\alpha}^{(v)}\right), \tag{2.34}$$

with $\mathbf{Z}^{(v)}$ the optimal clustering on the view v, $\boldsymbol{\tau}^{(v)}$ the fuzzy clustering matrix and $\boldsymbol{\alpha}^{(v)}$ the matrix of component-connection probability from the v-th view.

Initialization of $\boldsymbol{\tau}$ By applying a K-means with K centers on the concatenation of the fuzzy clustering matrices on each view, this result is used as an initialization of $\boldsymbol{\tau}$.

Initialization of $\boldsymbol{\nu}$ First, let's define a new $N \times N$ matrix $\mathbf{M}^{(v)}$ such as :

$$(\mathbf{M}^{(v)})_{ij} = (\alpha^{(v)})_{\mathbf{Z}_i \mathbf{Z}_j}. \tag{2.35}$$

After treating each $\mathbf{M}^{(v)}$ matrices, it is necessary to vectorize each of them to consider this information as a feature vector specific to each view. Then, By applying a K-means with Q centers, it is possible to identify a prestructure of the views membership. Again, this result is used as an initialization of $\boldsymbol{\nu}$.

2.5.2 Model selection

In the context of clustering, model selection often refers to the process of determining the ideal number of clusters for a given dataset. In our situation, the key decision lies in selecting appropriate values for K and Q to strike a balance between data attachment and model complexity. To achieve this, several criteria based on penalized log-likelihood can be employed, such as the Akaike Information Criterion (AIC) (Akaike, 1998), Bayesian Information Criterion (BIC) (Schwarz, 1978) and more recently the Integrated Completed Likelihood (ICL) (Biernacki et al., 2000). We specifically consider the ICL criterion and its associated penalties as they frequently yield good trade-offs in the selection of mixture models (Biernacki et al., 2010).

The ICL is based on the log-likelihood integrated over the parameters of the complete data. Furthermore, if we assume that parameters of component-connection probability $\boldsymbol{\alpha}$, parameters for mixture of communities $\boldsymbol{\pi}$ and parameters for views mixture $\boldsymbol{\rho}$ are independent, we have:

$$\begin{aligned}\mathrm{ICL}(\mathbf{A}, K, Q) &= \log \mathbb{P}(\mathbf{A}, \mathbf{Z}, \mathbf{W} \mid K, Q) \\ &= \log \int_{\boldsymbol{\alpha}} \mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha})\, \mathbb{P}(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\ &\quad + \log \int_{\boldsymbol{\pi}} \mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi})\, \mathbb{P}(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &\quad + \log \int_{\boldsymbol{\rho}} \mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho})\, \mathbb{P}(\boldsymbol{\rho}) d\boldsymbol{\rho}.\end{aligned} \tag{2.36}$$

In our variational framework, $\mathbf{Z}$ and $\mathbf{W}$ must be estimated. $\hat{\mathbf{Z}}$ (resp. $\hat{\mathbf{W}}$) can be chosen as the variational parameters $\boldsymbol{\tau}$ (resp. $\boldsymbol{\nu}$) directly or by a Maximum a Posteriori (MAP):

$$\hat{\mathbf{Z}}_i = \underset{k \in 1:K}{\operatorname{argmax}}\, \tau_{ik}. \tag{2.37}$$

By using approximations, such as Stirling's approximation formula on $\mathbb{P}(\boldsymbol{\pi})$ and $\mathbb{P}(\boldsymbol{\rho})$ and the Laplace asymptotic approximation on $\mathbb{P}(\boldsymbol{\alpha})$, we can define an *approximate ICL*:

$$\begin{aligned}\mathrm{ICL}(\mathbf{A}, K, Q) &\approx \log \mathbb{P}(\mathbf{A}, \hat{\mathbf{Z}}, \hat{\mathbf{W}} \mid K, Q) - \mathrm{pen}(K, Q) \\ &\approx \mathcal{L}\left(q(.)\right) - \mathrm{pen}(K, Q),\end{aligned} \tag{2.38}$$

where

$$\text{pen}(K,Q) = \frac{1}{2}\frac{K(K+1)}{2}Q\log(V\frac{N(N-1)}{2}) + \frac{1}{2}(K-1)\log(N) + \frac{1}{2}(Q-1)\log(V). \tag{2.39}$$

The penalization in this *approximate ICL* is composed of a part depending on the number of parameters of component-connection probability tensor $\boldsymbol{\alpha}$ and the number of edges taken into account, and a part that takes into account the number of degree of freedom in mixture parameters and the number of variables related to them. Recall that our model is based on undirected (symmetric) adjacency matrices, so we only consider the upper triangular matrices (without the diagonal).

However, in the Bayesian framework with conjugate priors, it is possible to define an exact ICL (Côme and Latouche, 2015). Moreover, it can be obtained from the previously defined ILvb (2.22) when the entropy of the latent variables is zero and the Expectation-Maximization algorithm is a Classification EM (CEM, Celeux and Govaert, 1992). In other words, variational parameters are equal to 1 if it is the MAP and 0 otherwise. Thus, this *exact ICL* can be defined as:

$$\begin{aligned}\text{ICL}_{\text{exact}}(\mathbf{A},K,Q) = \log&\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\}\\ +\log&\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\}\\ +\sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log&\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\}.\end{aligned} \tag{2.40}$$

Instead of using a CEM, it is possible to use the v Variational parameters directly, and to derive a *variational ICL* from the previous criterion. Figure 2.3 summarizes the links between the various selection criteria and clearly shows that in a particular context *exact ICL*, *Variational ICL*, and *ILvb* criteria are identical.

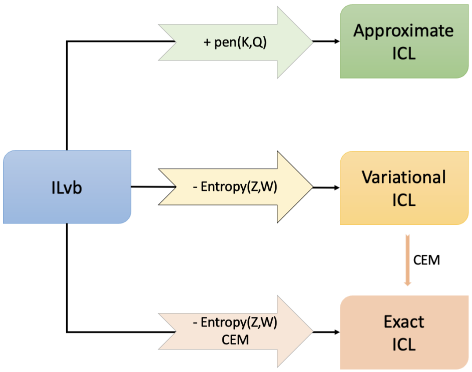


Figure 2.3: Diagram of links between different model selection criteria.

2.6 Experiments

2.6.1 Simulation study

To ensure that *mimi-SBM* behaves consistently, we have developed a simulation scheme, as depicted in Figure 2.4. The simulated data have been designed to reflect the complexity found in real-world data clustering challenges. In particular, this scheme allows for diverse clustering patterns across different view components. Also, it includes the possibility of controlling clustering errors, with observations being inaccurately assigned to an incorrect group, implying inconsistencies in the adjacency matrices.

Simulated data. Artificial adjacency matrices are generated from observations and views. We aim to establish a link between the simulated adjacency matrices and the final clustering that most accurately represents a problem of meta-clustering.

Various parameter values are tested for N (observations), V (views), K (clusters), and Q (components) in different scenarios. Besides, $\boldsymbol{\pi}, \boldsymbol{\rho}$ correspond to an equiprobability of belonging to a cluster or component, thus $\{\pi_k\}_{k=1}^K = 1/K$ and $\{\rho_s\}_{s=1}^Q = 1/Q$ (Figure 2.4, *Parameters*).

First of all, it is assumed that the number of real clusters (K) will always be equal or higher than the number of clusters coming from each component. Also, each component has a precise number of clusters (K^q), and each view belonging to this component will have this number of clusters $K^q \sim \mathcal{U}(\{2,\dots,K\})$, where $\mathcal{U}$ is the discrete uniform distribution (Figure 2.4, *Clusters per component*).

Now, for each component, we randomly associate a link between the final consensus clusters $\mathbf{Z}$ and clusters coming from the component $\mathbf{Z}^q$, ensuring that no component cluster remains empty (Figure 2.4, *Links between clusters*).

Eventually, for each pair of nodes (i, j) and each layer v, an edge is generated with probability $\alpha_{Z_i Z_j W_v}$, leading to set the corresponding entry in the multilayer adjacency tensor $\mathbf{A}$ to 1 or 0 (Figure 2.4, *Generation of edges*).

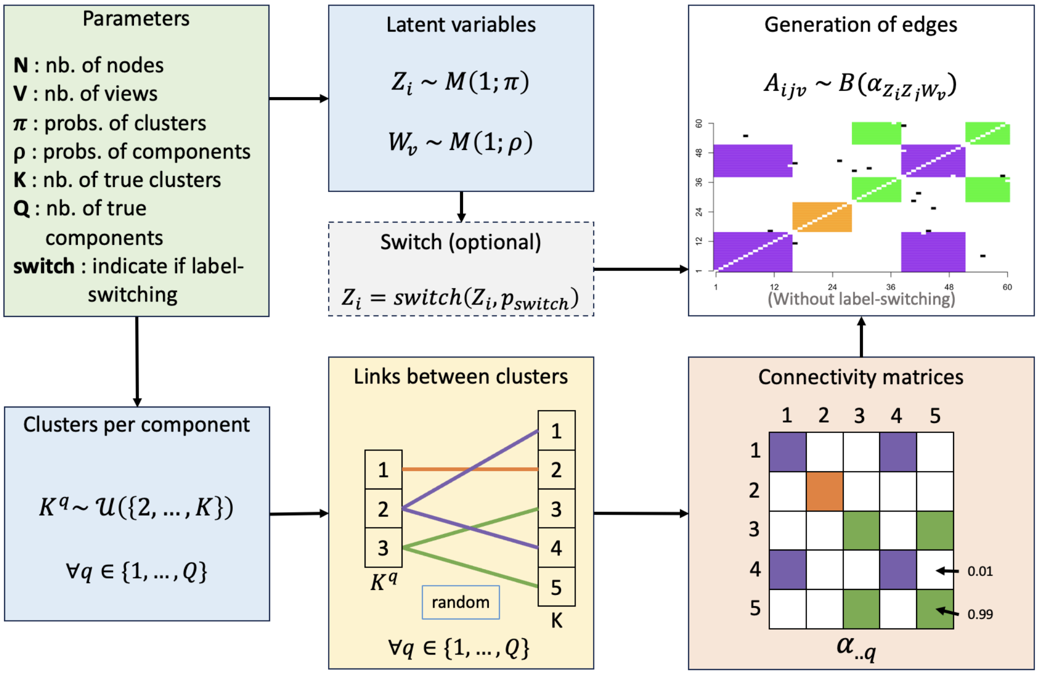


Figure 2.4: Diagram of the simulation process. Example of adjacency matrices resulting from mixing and traversing clusters across views, with potentially label-switching, for $K = 5$. For each view component, a number of clusters K^q is randomly drawn (discrete uniform distribution). Each cluster in the q^{th} component is then linked to certain clusters in the final partition. For this component, $K^q = 3$, and final clusters 1 (respectively 3) and 4 (resp. 5) are merged into cluster 2 (resp. 3) of the component, and the first cluster of the component corresponds perfectly to the final consensus cluster 2. Afterwards, these links are represented by a very strong connectivity within the $\boldsymbol{\alpha}_{..q}$ matrix ($p = 0.99$) and a very weak one ($p = 0.01$) for the others.

The simulated data are used to assess three aspects:

1. ***Model selection.*** In Section 2.6.1, various criteria are examined to recover the true parameters K and Q.

2. ***Clustering ability.*** In Sections 2.6.1 and 2.6.1, *mimi-SBM* is evaluated against other state-of-the-art techniques regarding the clustering of observations and the clustering of views using ARI scores (see below). We compare our approach to three different models: (i) *TWIST* (Jing et al., 2021), a tensor-based SBM model that captures both local and global community structures, as well as a partition of the views. We will use the global partition obtained by TWIST to evaluate community performance, and view partitions for component performances. (ii) *Graphclust* (Rebafka, 2023), a hierarchical SBM-based classification model that groups views into components and proposes clustering for each component. To preserve the integrity of this method, we will focus solely on the component clustering aspect for comparison. (iii) *Monte Carlo Reference-based Consensus Clustering* (John et al., 2020, M3C), a consensus clustering algorithm that improves partition robustness by comparing the obtained solutions to Monte Carlo-generated reference partitions. It statistically evaluates cluster stability, thereby reducing false positives and identifying significant structures in complex data. This method will be applied to the tensor $\mathbf{A}$, unfolded along the columns, to assess community performances.

3. ***Robustness.*** In Section 2.6.1, we investigate further the model ability to handle noisy configurations inherent in real-world clustering problems.

The code for the simulations is available on the CRAN (De Santiago et al., 2024b), and on GitHub in the repository *mimiSBM.* [1]

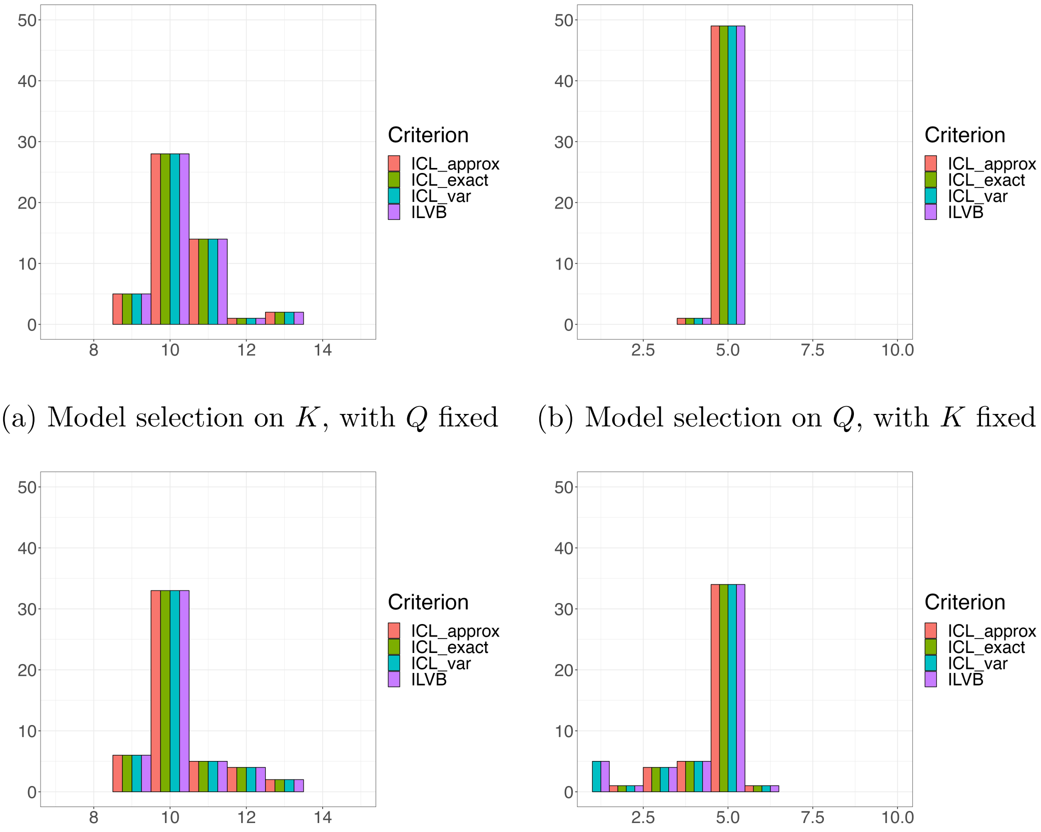
Adjusted Rand Index. The Adjusted Rand Index (ARI, Hubert and Arabie, 1985) quantifies the similarity between two partitions. In the simulation to follow, it quantifies the similarity between the prediction by our clustering models and the true partition. It corresponds to the proportion of pairs (i, j) of observations jointly grouped or separated. The more similar the partitions, the closer the ARI is to 1.

Comparing model selection criteria

In this section, our goal is to undertake a comparative analysis of model selection criteria to determine the optimal choice of criterion.

[1] `https://github.com/Kdesantiago/mimiSBM`.

The use of simulations gives us a complete control over the hyperparameters that generated the data. To do this, we generated 50 different datasets with hyperparameters $K = 10$ and $Q = 5$. The model selected for each criterion is the one that maximizes its value.



(a) Model selection on K, with Q fixed

(b) Model selection on Q, with K fixed

(c) Model selection on K, with Q free

(d) Model selection on Q, with K free

Figure 2.5: Bar plots of model selection criteria on 50 simulations with 10 true clusters for observations and 5 true components for views. Figure (a) (resp. (b)) indicates the number of times the K (resp. Q) value selected while the other parameters is set to the true value. Figures (c) and (d) show the same information when hyperparameters are optimized at the same time.

Simulation results in Figure 2.5, clearly show that, in all scenarios, each criterion delivers consistent and comparable performance. Without exception, the criteria consistently produce the same selection of clusters and number of view components.

In Figure 2.5a, only the hyperparameter K varies. This parameter was mostly well estimated because, during model selection, the true number of clusters was typically identified in the majority of cases. However, it was crucial to note that in the context of individual clustering, the criteria tended to overestimate the number of clusters.

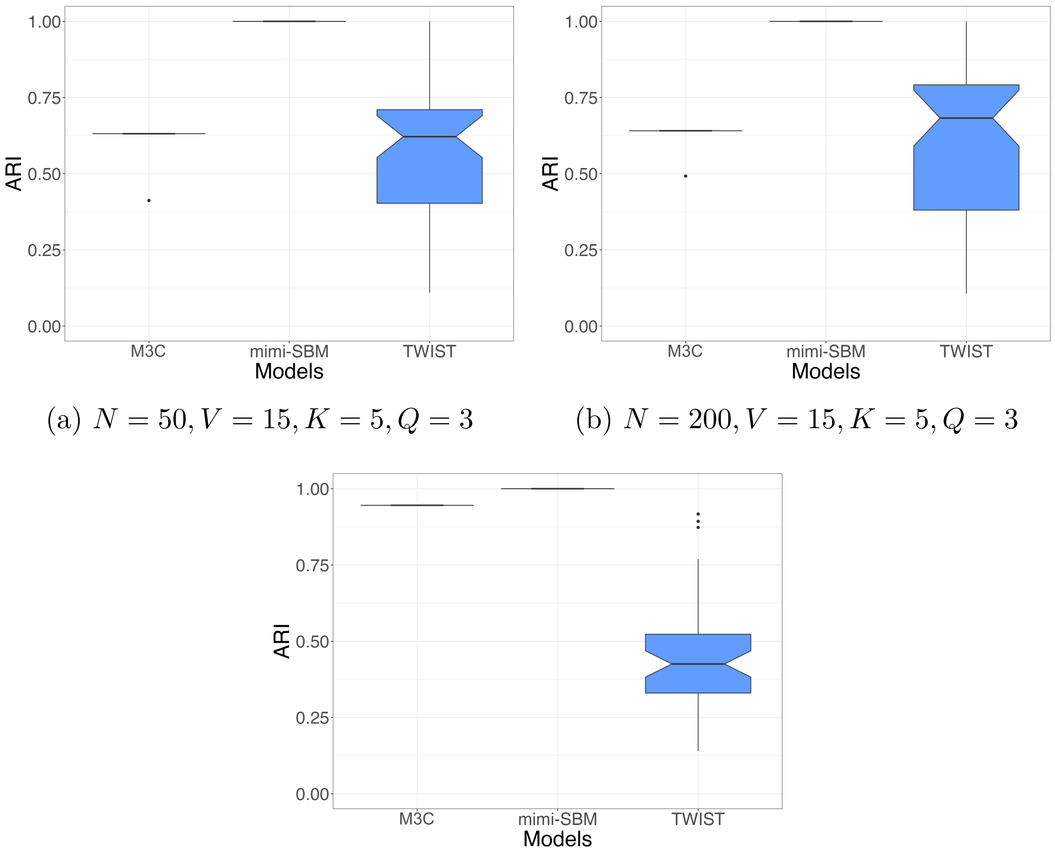
In Figure 2.5b, the number of components parameter Q is variable, while K remains constant. In the majority of scenarios, it was observed that the number of components was accurately estimated. Furthermore, when a fixed parameter for clustering was considered, the task inherently became more tractable due to the use of abundant information for the estimation of view components.

In Figures 2.5c and 2.5d, the selection of hyperparameters is aligned with fixed-parameter results. The criteria consistently demonstrate an aptitude for identifying the optimal cluster and view component quantities. Nonetheless, akin to previous instances, the model occasionally exhibits errors in hyperparameter estimation, often underestimating the number of components while overestimating the number of classes.

Simulations without label-switching

Comparison of clustering. In Figure 2.6, it has been observed that *mini-SBM* achieved the best clustering results for each considered experimental configuration. Indeed, *mini-SBM* recorded the highest ARI score for all data sizes, number of clusters and sources. On the other hand, the *M3C* model improved as the number of views, clusters and sources increased. In contrast, the *TWIST* model showed poorer performances as the clustering problem became more complex, suggesting that this model may be less suitable for difficult clustering problems.

Comparison of view components. In Figure 2.7, as the number of observations increases, the performances of the models generally tends to improve. However, the *graphclust* model appears to identify the true sources less frequently than the *mimi-SBM* and *TWIST* models. While *TWIST* often identifies the true members of the sources perfectly, it does make some errors, visible as outliers on the boxplot.



(a) $N = 50, V = 15, K = 5, Q = 3$

(b) $N = 200, V = 15, K = 5, Q = 3$

(c) $N = 200, V = 50, K = 10, Q = 10$

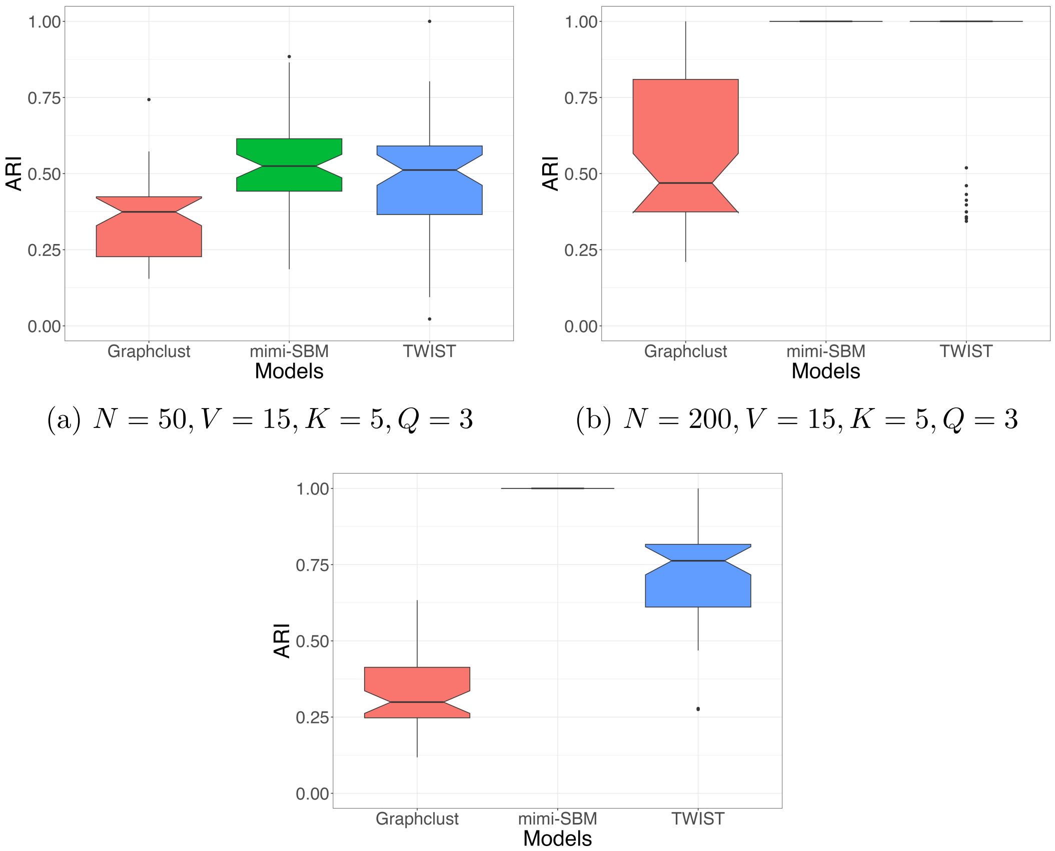
Figure 2.6: Boxplot of ARI measure between true partition and output partition of *M3C*, *mimi-SBM* and *TWIST* models.

Simulations with label-switching

In this section, we revisit the analyses from the previous section, but with a focus on a more challenging issue: label-switching. The idea of perturbing the cluster labels within the generation process simulates the fact that an individual has been associated with another cluster during the process of creating adjacency matrices.

In our context, we simulate the fact that an individual belongs to the real cluster, and then we simulate the representation of this clustering by the link between the final clustering and the one specific to each view component, to obtain the different affinity matrices.

The perturbation occurs during the generation of individual component-based clusters. For each view, a perturbation is introduced for each individual



(a) $N = 50, V = 15, K = 5, Q = 3$

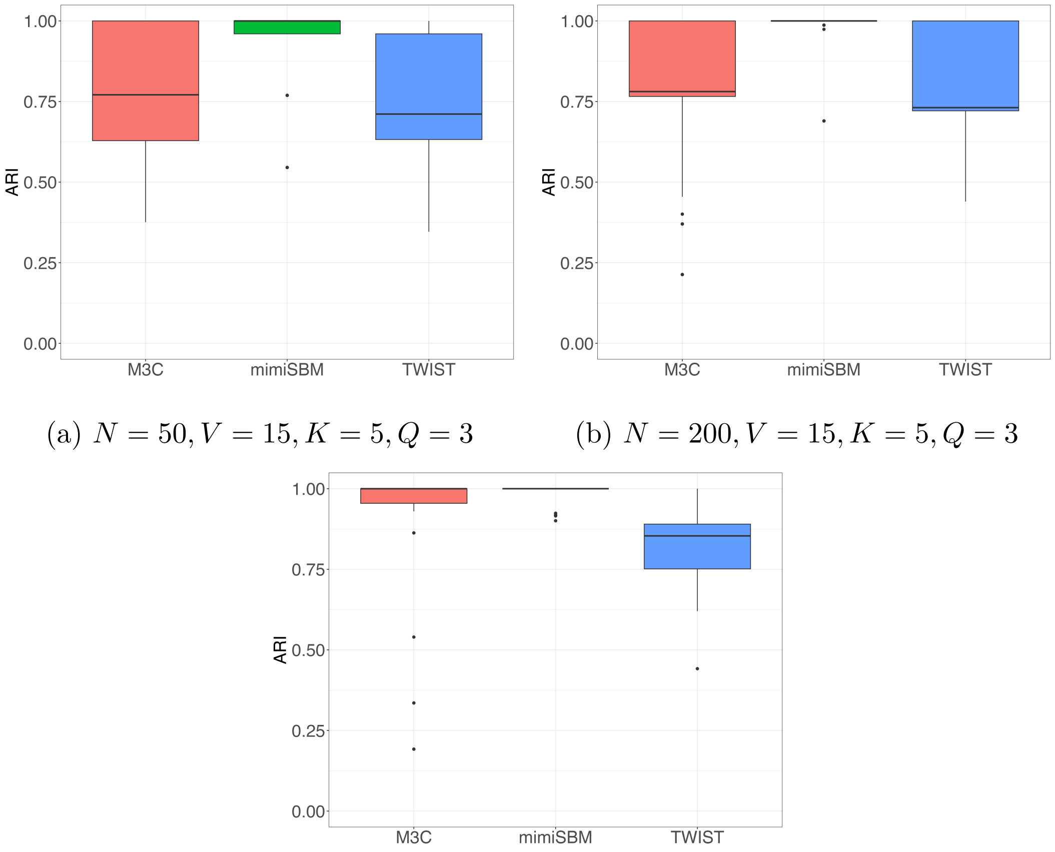
(b) $N = 200, V = 15, K = 5, Q = 3$

(c) $N = 200, V = 50, K = 10, Q = 10$

Figure 2.7: Boxplot of ARI measure between true view clustering and output clustering of *graphclust*, *mimi-SBM* and *TWIST* models.

with a probability of $p_{\text{switch}} = 0.1$. In such perturbation, the respective individual is then associated with one of the other available clusters. As a result, the probability of creating a link between individuals is influenced.

Comparison of clustering. The analysis summarized in Figure 2.8 reveals that across all examined experimental setups, the *mimi-SBM* consistently attained the most favorable clustering outcomes. Furthermore, it is noteworthy that the score variability associated with *mimi-SBM* is notably lower than that observed for other models. The effectiveness of the *M3C* and *TWIST* model showed improvement as the number of views, clusters, and sources increased, yet it maintained a relatively high level of variance. The number of individuals to be clustered plays a crucial role in minimizing errors. This effect stems from the fact that a larger number of individuals subject to clustering contributes to a more robust estimation of the parameters.



(a) $N = 50, V = 15, K = 5, Q = 3$

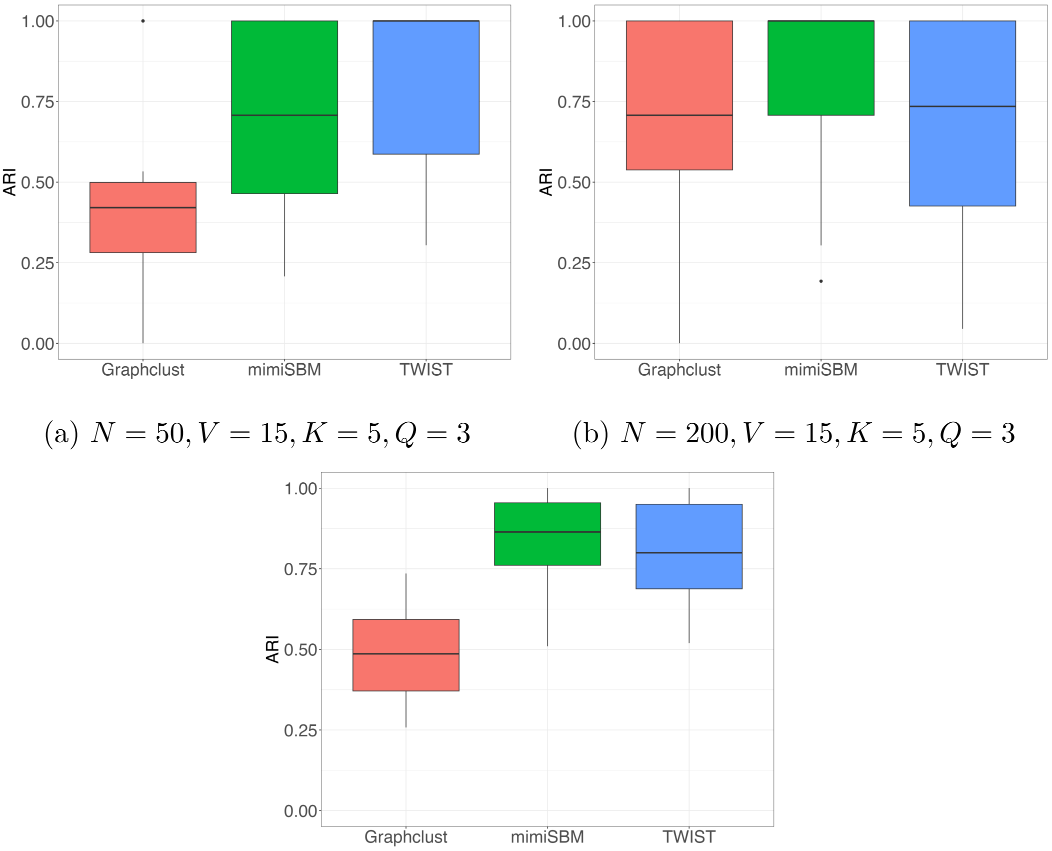
(b) $N = 200, V = 15, K = 5, Q = 3$

(c) $N = 200, V = 50, K = 10, Q = 10$

Figure 2.8: Boxplot of ARI measure between true partition and output partition of *M3C*, *mimi-SBM* and *TWIST* models.

Comparison of view components. In Figure 2.9, as the quantity of observations increases, the models typically exhibit enhanced performance. Nevertheless, the *graphclust* model seems to less frequently pinpoint the actual sources compared to the *mimi-SBM* and *TWIST* models. When faced with a small number of perspectives, *TWIST* model displays significant variability. While it consistently delivers good results, it remains vulnerable to unfavorable initializations, which can lead to notably suboptimal clustering outcomes. Moreover, when label-switching is introduced, the model's performance is observed to be slightly less effective compared to the precedent scenario.

Similar to the scenario without label-switching, the model experiences considerable variability in its estimation when dealing with a limited number of individuals and perspectives. However, as the number of individuals and views increases, the variance of ARI decreases noticeably, accompanied by an improvement in performance. *Mimi-SBM* model consistently demonstrates efficacy across all cases, even including perturbations in the adjacency matrices used for clustering.



(a) $N = 50, V = 15, K = 5, Q = 3$

(b) $N = 200, V = 15, K = 5, Q = 3$

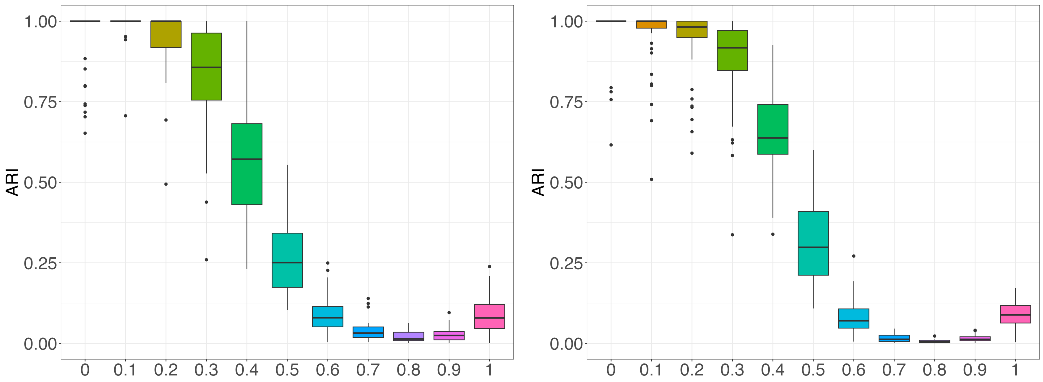
(c) $N = 200, V = 50, K = 10, Q = 10$

Figure 2.9: Boxplot of ARI measure between true view clustering and output clustering of *graphclust*, *mimi-SBM* and *TWIST* models.

Robustness to label-switching

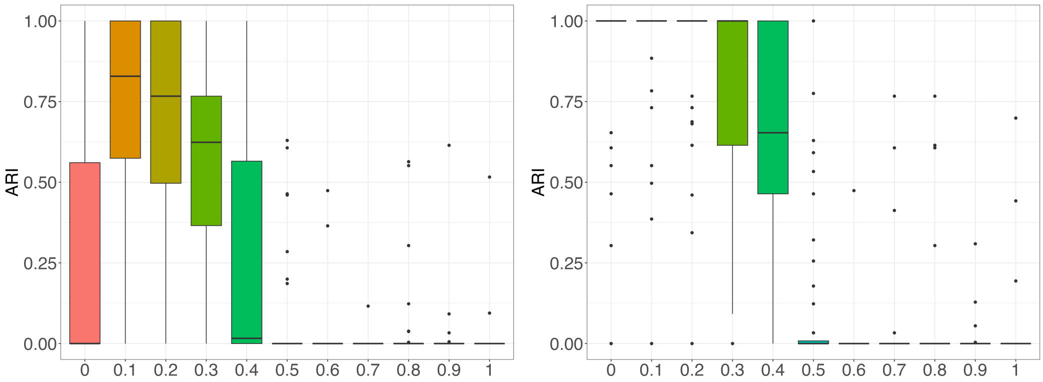
Given that the *mimi-SBM* showed a satisfactory performance level in the previous section, even under the influence of label-switching perturbation, this section aims to further assess the robustness and limitations of our model concerning this criterion.

By varying the label switching rate, from 0 to 1 in steps of 0.10, in order to see the evolution of clustering capacities on individuals and views.



(a) $N = 50, V = 15, K = 5, Q = 3$ (b) $N = 200, V = 15, K = 5, Q = 3$

Figure 2.10: Performances of *mimi-SBM* on individual clustering through the evolution of label-switching rate.



(a) $N = 50, V = 15, K = 5, Q = 3$ (b) $N = 200, V = 15, K = 5, Q = 3$

Figure 2.11: Performances of *mimi-SBM* on view clustering through the evolution of label-switching rate.

For Figures 2.10 and 2.11, clustering performances demonstrate a significant level of efficacy when the label-switching rate is low.

As the rate of switched labels exceeds 40%, the stability of the individual clustering process progressively diminishes. This trend continues until the clustering process becomes entirely arbitrary when the switch-labeling rate surpasses 60%, as contrasted with the true partition. An observable improvement in performance becomes evident as the switched label rate approaches 1. This outcome is logically anticipated, as the reassignment of all individuals from one cluster to another results in their distribution across $K - 1$ clusters instead of the initial K clusters.

In the context of view-based clustering, we encounter a similar set of observations, albeit with a much more pronounced decline in performance. When the label-switching rate surpasses 20%, the ability of *mimi-SBM* to effectively identify view components experiences a drastic reduction. Furthermore, when this rate exceeds 40%, the feasibility and relevance of conducting clustering based on these views are severely compromised. One plausible explanation for this phenomenon is that, due to the perturbation, each adjacency matrix becomes highly noisy, lacking any discernible structure. Consequently, the model struggles to distinguish any specific connections within the mixtures, leading to a notably diminished clustering performance score.

The results shown in Figure 2.11a are surprising, with 0% label switching. The initialization phase accurately identifies the communities but often struggles with the components. Since the algorithm is trapped in a local minimum, it does not iterate further, leaving the estimate unchanged. In other cases, the algorithm requires multiple iterations, allowing it to update the initialization of the components.

Summary. The *mimi-SBM* model has shown its capability in successfully recovering the stratification of individuals and the components of the mixture of views, even when the data is perturbed. However, like any statistical model, its performance, especially regarding the mixture of views, benefits from larger sample sizes. The accurate modeling of mixture components is crucial in various applications, making the mimi-SBM model highly valuable in a wide range of contexts.

2.6.2 Worldwide Food Trading Networks

Data. This section delves into the analysis of a global food trading dataset initially assembled by De Domenico et al. (2015), accessible at `http://www.fao.org`. The dataset includes economic networks covering a range of products, where countries are represented as nodes and the edges indicate trade links for particular food products. Following the same preprocessing steps as Jing et al. (2021), we prepared the data to establish a common ground for comparing clustering outcomes. The original directed networks were simplified by omitting their directional features, thereby converting them into undirected networks.

Subsequently, to effectively filter out less significant information from the dataset, we eliminate links with a weight of less than 8 and layers containing limited information (less than 150 nodes). Finally, the intersections of the biggest networks of the preselected layers are then extracted. Each layer reflects the international trade interactions involving 30 distinct food products among 99 different countries and regions (nodes).

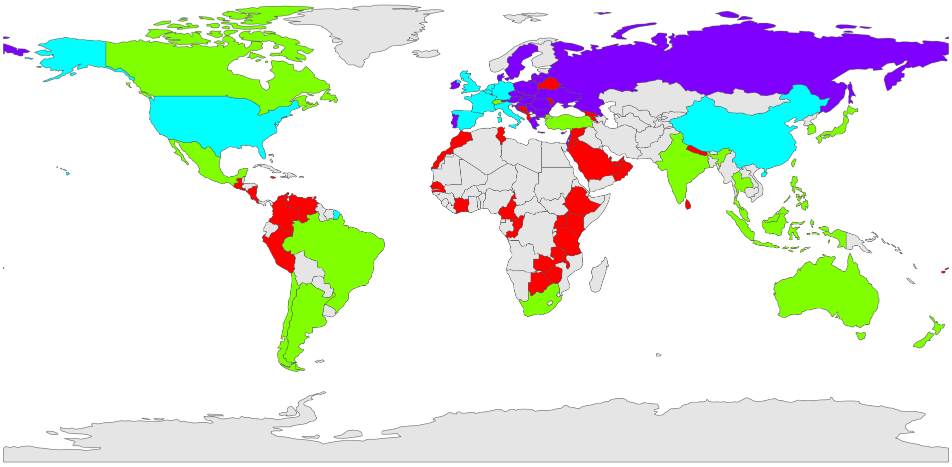


Figure 2.12: World Map of Clusters: Countries are color-coded based on the clusters identified by the model. The cyan cluster (cluster 1) encompasses the West and China; The violet cluster (cluster 2) consists of Russia and some parts of Western Europe; The red cluster (cluster 3) includes countries from Africa and Central America; The green cluster (cluster 4) covers Mexico, Canada, India, Australia, South Africa, Japan, among others; Countries depicted in grey are not included in the database analyzed.

TWIST analysis. In our research, we followed the analytical process described in Jing et al. (2021), to facilitate reliable comparison of results. Consistent with this methodology, we fixed the number of clusters at $K = 4$ for individuals and $Q = 2$ for views.

In Figure 2.12, clusters have their own interaction patterns:

- Cluster 1 serves as a hub due to its centralization of exchanges, exhibiting a high intra-connectivity ($> 90\%$) and substantial inter-connectivity ($> 70\%$), as revealed by the multilayer adjacency probability analysis.

- Cluster 2 displays a robust intra-connectivity, with notable interactions observed with both clusters 1 and 4. Conversely, exchanges with cluster 3 are infrequent for the commodities comprising the database.

- Cluster 3 and Cluster 4 exhibit both intra-cluster and inter-cluster interactions, with a preference for inter-cluster interactions with cluster 1. However, while Cluster 3 predominantly interacts with Cluster 1, Cluster 4 demonstrates partial interaction with Cluster 2.

View component 1	Beverages_non_alcoholic , Food_prep_nes, Chocolate_products_nes , Crude_materials, Fruit_prepared_nes, Beverages_distilled_alcoholic, Pastry, Sugar_confectionery, Wine
View component 2	Cheese_whole_cow_milk, Cigarettes, Flour_wheat Beer_of_barley, Cereals_breakfast, Coffee_green, Milk_skimmed_dried, Juice_fruit_nes, Maize, Macaroni, Oil_palm, Milk_whole_dried, Oil_essential_nes, Rice_milled, Sugar_refined, Tea Spices_nes, Vegetables_preserved_nes, Water_ice_etc, Vegetables_fresh_nes, Tobacco_unmanufactured

Table 2.1: Table of members in view components.

Exploration of the view components in Table 2.1 reveals a marked tendency to distinguish between "processed products" and "unprocessed products", although there are some notable exceptions. In addition, it should be noted that Component 1 displays more important connections than Component 2, suggesting that the main flow of transactions is mainly concentrated on products included in Component 1. This observation reinforces Cluster 1's position as a central hub, remaining a predominant actor in the concentration of trade within the various components.

The analysis carried out in this study is reflected in a striking correlation with the steps taken in the precedent analysis. Firstly, we found that the same partitions of individuals were present, with only minor variations in clustering. The links forged within these groups proved to be consistent with market dynamics, highlighting, in particular, the hub role played by cluster 1 in global trade. Furthermore, the overall partitioning of food types persisted, illustrating the persistent distinction between processed and unprocessed products, although a few exceptions were noted, similar to those observed in the previous analysis. In sum, our results largely converge with those of the Jing et al. (2021) study, although a few discrepancies remain, underlining the importance of continuing research in this area to refine our understanding and approach.

Our optimization. First, the criterion for choosing the optimal model was employed to guide the selection of hyperparameters. A grid search was conducted over a range of values, spanning from 1 to 20 for the hyperparameter K and from 1 to 10 for Q, in concordance with parameters of the first core in Jing et al. (2021) experimentation. The model selection process led to the choice of hyperparameters $K = 20$ and $Q = 1$ as the most suitable configuration. The model found it excessively costly to introduce additional components across the views compared to the information gain achieved, so $Q = 1$ was selected. For individual clustering, $K = 20$ was based on the model's discovery of numerous micro-clusters representing countries based on their interaction habits. This indicates that the model successfully identified fine-grained distinctions among countries, revealing intricate subgroups within the data.

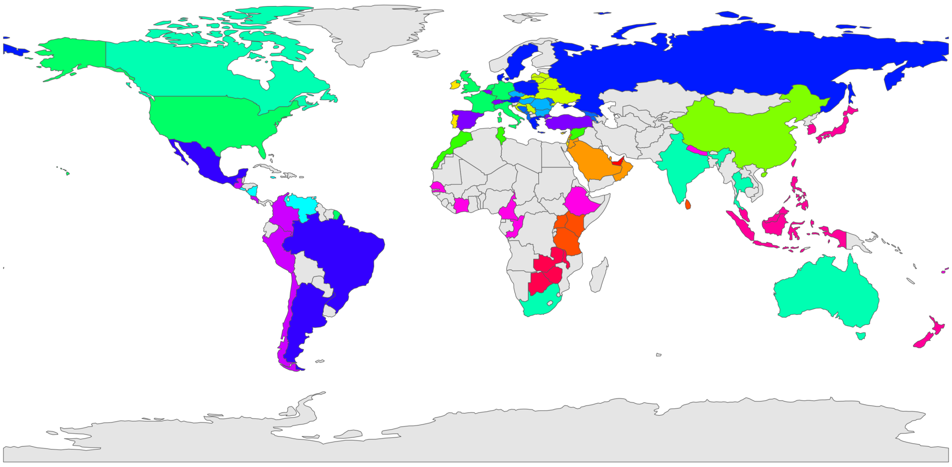


Figure 2.13: Clustering world map: countries are colored according to the clusters, and parameters defined by the model ($K = 20$).

The results show that certain clusters have been substantially preserved, in particular Cluster 1, which remains virtually intact, as does Cluster 4. However, there has been a significant fragmentation of existing clusters, with China in particular remaining isolated. There have also been significant changes in the configuration of clusters, notably the inclusion of Russia along with other South American countries. This change could be explained by the fact that we are now considering only one view component, Russia is closer, in the sense of trading more with, the countries of South America than the rest of the world.

2.7 Bilan

This paper proposes a new framework for Mixture of Multilayer SBM *mimi-SBM* that stratifies individuals as well as views. In order to get a manageable lower bound on the observed log-likelihood, a variational Bayesian approach has been devised. Each model parameter has been estimated using a Variational Bayes EM algorithm. The advantage of such a Bayesian framework consists in allowing the development of an efficient model selection strategy. Moreover, we have provided the proof of model identifiability for the *mimi-SBM* parameters.

In our simulation setting, the *mimi-SBM* related algorithm has been shown to compete with methods based on tensor decomposition, hierarchical model-based SBM, and reference model in consensus clustering in two critical aspects of data analysis: individual clustering and view component identification. Specifically, our algorithm reliably recovered the primary sources of information in the majority of investigated cases. These remarkable performances attest to the efficacy of our approach, underscoring its potential for diverse applications requiring a profound understanding of complex data structures. In real-world application on *Worldwide Food Trading Networks*, when considering the paradigm provided by Jing et al. (2021), we obtain consistent results. However, upon optimizing our model using our metric, a distinctly different clustering emerges. This alternative clustering not only diverges significantly but also reflects much finer nuances in transactional natures.

3 Mixture of Experts models × Conditional LBM

3.1 Introduction 89
3.2 Material and Methods 93
3.2.1 General modeling 93
3.2.2 Mixture Of Experts and BIclustering Unified Strategy model 95
3.3 Optimization via SEM-Gibbs algorithm 98
3.3.1 Gibbs-sampling Expectation step 99
3.3.2 Maximization step 99
3.4 Model Selection 101
3.5 Experiments 102
3.5.1 Simulation process 103
3.5.2 Co-clustering and regression performance 105
3.5.3 Parameter estimations 107
3.5.4 Model selection 107
3.6 Bilan 110

Our approach with *mimi-SBM* focused on late fusion, a method designed to condense information after preprocessing multimodal data to extract coherent and synthetic clusters.

In this chapter, we shift towards a supervised learning framework, where the goal is to predict the target variable $\mathbf{y}$ based on the covariates $\mathbf{X}$. In this context, we aim to leverage models that utilize early or intermediate fusion, allowing us to incorporate multi-source information at earlier stages of the learning process, and therefore an interpretation directly linked to the raw data (Stahlschmidt et al., 2022). Our focus lies particularly on Mixtures of Experts, owing to their general framework and versatility in handling heterogeneous data (Gormley and Frühwirth-Schnatter, 2019).

One of our primary objectives is to utilize MoEs for data stratification, enabling us to identify and characterize subpopulations by forming distinct communities. This stratification is achieved through the gating network, which dynamically assigns observations to specialized experts. However, a notable limitation of MoEs is their potential to function as a black-box model, making the interpretation of variables challenging (Ismail et al., 2022). This complexity is influenced by the type of gating network and expert models used. A detailed discussion of these challenges will be presented in section 3.1.

Thus, our ultimate goal is to develop an interpretable MoE framework that not only allows for the characterization of subpopulations but also captures the redundancy and complementarity of the information specific to each community.

Contribution. This work extends Mixtures of Experts models by adding a conditional biclustering structure to the data. This modeling results in interpretable MoEs that reveal the redundancy and complementarity of information by clustering the variables into components. This fusion also enhances computational efficiency by allowing specialized experts to focus on the redescription of the variables, with representative variables, characterizing the components.

Organization of the Chapter. In this chapter, we will delve into the mathematical modeling, parameter estimation, model selection, and performance evaluation of a new model entitled Mixture of Experts and BIclustering Unified Strategy (MoEBIUS), based on the combination of an original conditional co-clustering LBM (Co-coLBM, see appendix F) and Mixture of experts. Section 3.1 outlines interpretability and parsimony for Mixture of Experts, and justifies their combination with a conditional biclustering approach for supervised learning.Section 3.2 provides the mathematical formulation of the *MoEBIUS* model, describing the latent variable distributions and the target variable modeling for regression and classification tasks. It integrates block structures and expert predictions into a unified framework. In Section 3.3, the Gibbs Sampling Expectation-Maximization algorithm is detailed for estimating the latent variables and model parameters. In Section 3.4, we introduce the BIC_ICL criterion, balancing model performance and complexity to select the optimal hyperparameters. Finally, Section 3.5 summarize simulations evaluating the *MoEBIUS*'s performance in co-clustering and regression parameter estimation, and model selection. Finally, Section 3.6 reviews the results, highlights the strengths and limitations of the model.

3.1 Introduction

In Mixture of Experts, although each expert receives all input variables, only a subset is truly relevant to the specific task assigned to the expert. This variable selection is frequently implicit, as each expert focuses on the most useful characteristics for its task. However, this selection can be made explicit through regularization mechanisms, such as L_1 (Courbariaux et al., 2022), L_2 (Jordan and Jacobs, 1994), or elastic net (Chamroukhi and Huynh, 2019; Ma et al., 2018). In accordance with Section 1.2.3, for K experts, this would be equivalent to maximize the following quantity:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\pi};\mathbf{y},\mathbf{X}) = \prod_{i=1}^{N}\sum_{k=1}^{K} \underbrace{g\left(\mathbf{X}_i;\boldsymbol{\pi}\right)_k}_{\text{Gating Network}} \underbrace{f_k\left(y_i;\mathbf{X}_i,\boldsymbol{\beta}_k\right)}_{\text{Expert}} + \underbrace{\lambda\left(\alpha\sum_{k=1}^{K}\sum_{j=1}^{p}|\pi_{kj}| + \frac{1-\alpha}{2}\sum_{k=1}^{K}\sum_{j=1}^{p}\pi_{kj}^2\right)}_{\text{Regularization terms}}, \tag{3.1}$$

where $g\left(\mathbf{X}_i;\boldsymbol{\pi}\right)_k$ is the k-th output of the gating network detailed in Equation (3.3), $\boldsymbol{\pi}$ are gating network parameters, $(\boldsymbol{\beta}_k)_{k=1:K}$ are experts parameters, $\lambda > 0$ is an hyperparameter linked to regularization strength and $\alpha > 0$ is an hyperparameter for elastic-net regularization.

These regularizations, primarily applied to the gating network, enhance interpretability by controlling expert selection directly. The gating network plays a central role in variable and expert selection, and different approaches can be classified into three main categories Cai et al. (2024): dense (Wu et al., 2023; Pan et al., 2024), sparse (Jiang et al., 2024; Tan et al., 2023; Zhou et al., 2022), and soft (Puigcerver et al., 2023; Zadouri et al., 2023).

In sparse methods, regularizations are designed to activate a limited number of experts or variables, which improves the interpretability by aligning predictions with the variables involved (Zhou et al., 2022). The main approach is to use the top-L experts, meaning the L experts with the highest values according to the gating network: in general, the gating network g is composed of $(g_k)_{k=1:K}$ functions, and is defined as

$$g(\mathbf{X}_i; \boldsymbol{\pi}) = \operatorname{softmax}\left(g_1\left(\mathbf{X}_i; \boldsymbol{\pi}_1\right), \cdots, g_K\left(\mathbf{X}_i; \boldsymbol{\pi}_K\right)\right). \tag{3.2}$$

Typically, to limit the number of activated experts, according to Cai et al. (2024), the gating network output $g(\mathbf{X}_i)_k$ is replaced by:

$$\tilde{g}(\mathbf{X}_i)_k = \operatorname{softmax}\left(\operatorname{TopL}\left(g_1\left(\mathbf{X}_i; \boldsymbol{\pi}_1\right)\right), \cdots, \operatorname{TopL}\left(g_K\left(\mathbf{X}_i; \boldsymbol{\pi}_K\right)\right)\right)_k, \tag{3.3}$$

where

$$\operatorname{TopL}(g_k(\mathbf{X}_i; \boldsymbol{\pi})) = \begin{cases} g_k(\mathbf{X}_i; \boldsymbol{\pi}) & \text{if } g(\mathbf{X}_i; \boldsymbol{\pi})_k \in \text{top-L elements of } g(\mathbf{X}_i; \boldsymbol{\pi}), \\ -\infty & \text{else.} \end{cases} \tag{3.4}$$

Currently, sparse MoEs typically refers to limiting the number of active experts, thereby reducing computational cost. However, an increased specialization among a large pool of experts does not necessarily ensure sparsity in terms of variable selection. On the other hand, soft approaches allow more gradual and partial expert activation, offering a balance between sparsity and flexibility. All MoEs with regularization mechanisms on the gating network fall into this class of models. While these models achieve sparsity in terms of variable selection, they remain soft in terms of expert activation. Indeed, it is not always possible to ensure that only a minimal subset of experts will be activated.

The integration of these penalties structures the gating network's behavior and encourages a parsimonious selection of variables. This simplifies interpretation by reducing the number of variables and experts involved in decision-making, while maintaining computational efficiency (Shazeer et al., 2017). However, the separation between variables used by the gating network and those leveraged by experts in the final prediction can complicate the interpretability, as the most influential variables for selection are not always the most relevant for prediction. This disjunction introduces a potential bias in model interpretation. Several efforts have been made to improve the interpretability of MoEs, for instance:

- Ismail et al. (2022) developed an interpretable MoE model, introducing an approach that makes expert decisions more transparent by employing explanation techniques such as assignment modules and studying expert contributions.

- Since 2017, the integration of attention mechanisms in MoEs, as presented by Shazeer et al. (2017), has made the decision process more explicit by focusing the model's attention on specific aspects of the data. Additionally, this work introduced the concept of the MoE-layer, which has been widely reused in recent Deep Learning architectures (Lepikhin et al., 2020; Fedus et al., 2022; Zhou et al., 2022).

- In the context of Multitask Learning, Ma et al. (2018) proposed using multiple gating networks, linked to the associated prediction tasks, providing a clearer connection between the experts utilized, the gating network predictions, and the final outcomes.

Ensuring interpretability is an important aspect in machine learning applied to sensitive areas. In Mixture Of Experts framework, which can be achieved not only through the design of the gating mechanism but also by employing interpretable experts. Experts based on linear regression are a common choice in this regard, as they allow for a clear justification of the decisions made by each expert. Moreover, experts based on linear regression are often favored due to their computational efficiency.

These type of Mixture of Experts models generally fall within the category of Latent Regression Models (LRMs), where the experts are based on latent subpopulations, and the regression coefficients capture subgroup-specific effects (Vermunt, 2002; DeSarbo and Cron, 1988). More specifically, the general model is often defined as follows:

$$\mathbf{Z}_i \sim \mathcal{M}\left(1; \boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)\right), \tag{3.5}$$

$$y_i \mid \mathbf{X}_i, Z_{ik} = 1 \sim \mathcal{N}\left(\mathbf{X}_i \boldsymbol{\beta}_k, \sigma_k^2\right), \tag{3.6}$$

where $\mathbf{Z}_i$ is a latent variable for expert allocation for observation i and $\boldsymbol{\beta}_k$ are regression parameters for the k-th expert (or community).

In this framework, the role of the gating network is associated with the variable Z. This variable refers to community membership, it is assumed that individuals from the same cluster are studied by the same expert.

Nevertheless, several extensions have been developed, including longitudinal data models and co-clustering regression models.

Longitudinal data models: Individuals within the same community share similar dynamics over time. Each latent class has its own risk profile for the event under study or its own longitudinal pattern (Proust-Lima et al., 2014; Courbariaux et al., 2022). Following the model of Courbariaux et al. (2022), and with our notations, we would have :

$$\mathbf{Z}_i \sim \mathcal{M}\left(1; \mathrm{softmax}\left(\mathbf{X}_i \boldsymbol{\pi}\right)\right), \tag{3.7}$$

$$y_{iv} \mid \mathbf{X}_i, \mathbf{T}_i, Z_{ik} = 1 \sim \mathcal{N}\left(\sum_{r=1}^{R} \beta_{kvr} T_{iv}^{r}, \sigma_{vk}^2\right). \tag{3.8}$$

In this context, $\mathbf{X}$ is considered as a $N \times p$ matrix. The latent variable is obtained by a multiclass regression parameterized by a $p \times K$ matrix $\boldsymbol{\pi}$. On the v-th visit and for the community k, the polynomial regression is parameterized by $\boldsymbol{\beta}_{kv}$, and T_{iv} is the time metric (the patient's age or time since the disease was first diagnosed, for instance).

Co-clustering Regression Model: Vu and Aitkin (2015) proposed a multitask regression algorithm based on an *LBM* partitioning of the $N \times p$ label matrix $\mathbf{y}$. Each block of the matrix is associated with regression parameters for the covariates $\mathbf{X}$. The Latent Block Regression Model (*LBRM*) developed by Boutalbi et al. (2022) extends this by partitioning the $N \times p \times d$-tensor $\mathbf{X}$ based on the same stratification as the label matrix $\mathbf{y}$. The last dimension of the tensor $\mathbf{X}$ corresponds to the covariates used in the regression. More precisely, and with our notations :

$$\mathbf{Z}_i \sim \mathcal{M}\left(1; \boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)\right), \tag{3.9}$$

$$\mathbf{W}_j \sim \mathcal{M}\left(1; \boldsymbol{\rho} = (\rho_1, \cdots, \rho_Q)\right), \tag{3.10}$$

$$y_{ij} \mid \mathbf{X}_{ij}, Z_{ik} = 1, W_{js} = 1 \sim \mathcal{N}\left(\mathbf{X}_{ij} \boldsymbol{\beta}_{ks}, \sigma_{ks}^2\right). \tag{3.11}$$

For an observation i belonging to community k and a problem j in category s, a regression is performed along the third dimension, on $(\mathbf{X}_{ijr})_{r=1:d}$, parameterized by a coefficient vector $\boldsymbol{\beta}_{ks} \in \mathbb{R}^d$.

In the supervised framework, we aim to achieve both interpretability and predictive performance. The objective of our approach is to incorporate an additional structure into the data used for prediction. While the stratification of individuals is already induced by the Mixture of Experts framework, we aim to extend this by also integrating a structure on the variables. Through a mixture model applied to covariates, the goal is to capture the emergence of both redundant information within components and complementary information between components. This approach mirrors the work on *mimi-SBM*, which handles similar challenges but for multi-view clustering.

As illustrated in Figure 3.1, to obtain more specific partitions compared to a standard mixture model on covariates, a conditional stratification of variables, depending on the community structure, is needed. The resulting model, named Mixture Of Experts and BIclustering Unified Strategy (*MoEBIUS*), merges MoEs with a conditinal biclustering algorithm to offer precise predictions alongside clear interpretation of latent communities and components.

More specifically, in Section 3.2, the model summarizes the components into a representative variable for each, then performs a regression on these representations, where the regression parameters depend on the community to which the observation belongs, an illustration of MoEBIUS is provided in Figure 3.2.

3.2 Material and Methods

3.2.1 General modeling

In this section, we focus on incorporating a conditional variable stratification mechanism within Mixtures of Experts. To achieve this, we introduce a new random variable, denoted as $\mathbf{W}$. The purpose of the variable $\mathbf{W}$ is to determine which covariables are considered and how they are utilized for prediction.

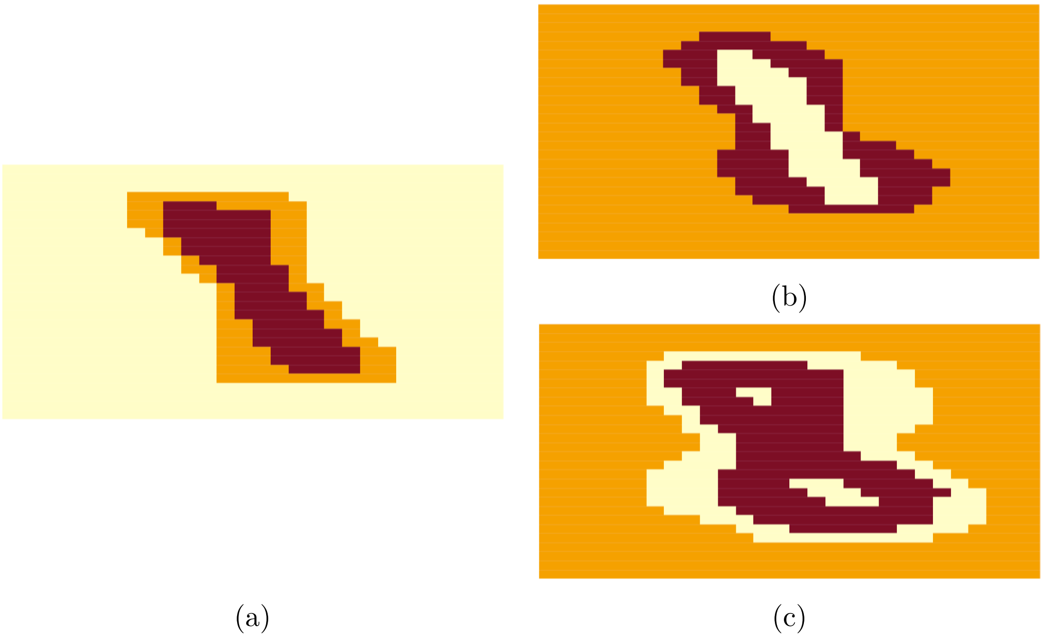


Figure 3.1: Different component partitions on MNIST data for digits 1 and 8. Figure (a) shows a partition defined by an *LBM*, where the overall shape of both digits is captured. Figures (b) and (c) are derived from the *Conditional LBM* (Goffinet et al., 2020), where the partitioning of variables depends on the observations (i.e., the digits). For each partition, the key pixels of interest specific to each digit are more clearly identified.

This leads us to the following general formulation:

$$\mathbb{P}\left(\mathbf{y}, \mathbf{W}, \mathbf{Z} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\pi}\right) = \underbrace{\mathbb{P}\left(\mathbf{y} \mid \mathbf{W}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\beta}\right)}_{\text{Experts}} \underbrace{\mathbb{P}\left(\mathbf{W} \mid \mathbf{Z}, \mathbf{X}, \boldsymbol{\rho}\right)}_{\text{Variable stratification}} \underbrace{\mathbb{P}\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\pi}\right)}_{\text{Gating network}}. \tag{3.12}$$

At this stage, the specific dimensions of variables are not crucial; these details will be provided in Section 3.2.2.

The definition of experts and the law associated with $\mathbf{y}$ are intrinsically dependent on the nature of the data. However, they can be formalized through a function f such that $f(\mathbf{y}, \mathbf{W}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\beta})$, with $\boldsymbol{\beta}$ the expert parameters. The variable $\mathbf{W}$ is influenced by both $\mathbf{Z}$ and $\mathbf{X}$, allowing for a stratification that depends simultaneously on the communities and the individual-specific characteristics. Finally, the modeling of communities via $\mathbf{Z}$ is itself conditioned by $\mathbf{X}$; however, a marginal law could be considered.

3.2.2 Mixture Of Experts and BIclustering Unified Strategy model

As in Courbariaux et al. (2022), the assignment to a community is obtained through a multiclass regression.

$$\mathbf{Z}_i \mid \mathbf{X}_i \sim \mathcal{M}\left(1; \mathrm{softmax}\left(\mathbf{X}_i \boldsymbol{\pi}\right)\right), \tag{3.13}$$

where $\boldsymbol{\pi}$ is a matrix of dimension $p \times K$.

The goal is to predict the latent variable $\mathbf{Z}_+$ for a new observation $\mathbf{X}_+$, leveraging information obtained during model training. In this context, a regression-based approach is more appropriate than a marginal distribution, as it directly incorporates the relationships learned between the input variables and the latent variables.

Now, based on the Conditional Latent block Model (CLBM) approach developed by Goffinet et al. (2020), we define the $K \times p \times Q$ tensor $\mathbf{W}$, which models the partition of the p variables into Q components.

These partitions are conditioned by the K communities, meaning that for each community k, the partition of the variables may differ, hence introducing the conditional aspect. Conditionally to the cluster k, each row of the component membership matrix follows:

$$\mathbf{W}_{kj} \sim \mathcal{M}\left(1; \boldsymbol{\rho}_k = (\rho_{k1}, \ldots, \rho_{kQ})\right). \tag{3.14}$$

One could have considered incorporating the contribution of $\mathbf{X}_i$ in the modeling process, allowing for a partition dependent on the individual rather than the community. However, this would have complicated the interpretability. The rationale behind this approach is to partially summarize the behavior of individuals through the communities to which they belong, making this modeling choice preferable.

In addition, an analysis of the effectiveness of the CLBM under the assumption of constant communities and conditional variable partitions has been conducted, through a study of the Co-Conditional Latent Block Model (*Co-CoLBM*). Parameter estimation, model selection and performance on simulated data have been studied. In addition, the code is available in my GitHub profil (*Kdesantiago*) [1]. The detailed results of this study are presented in Appendix F.

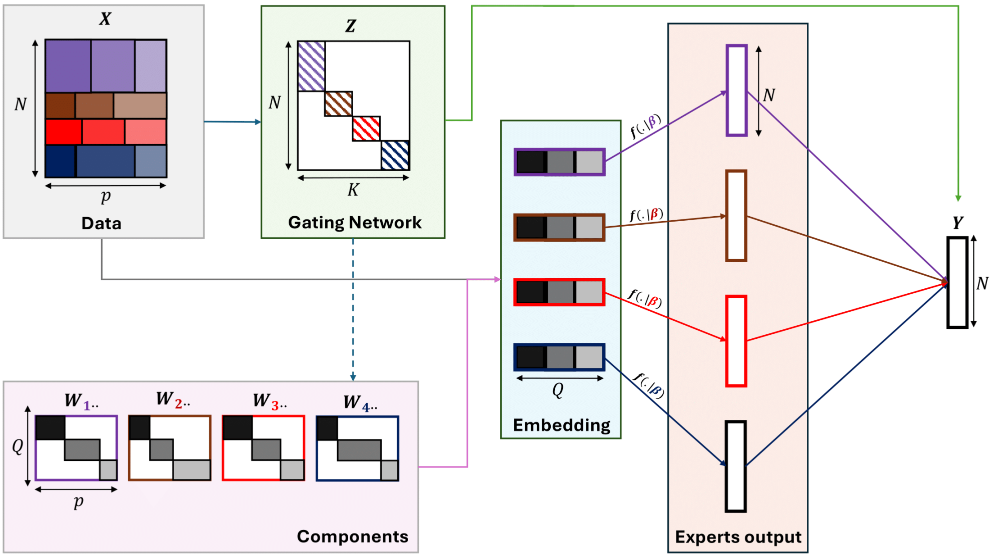


Figure 3.2: Schematic representation of *MoEBIUS*. The input data $\mathbf{X}$, of dimension $N \times p$, is used to define selection weights for the K experts (gating network $\mathbf{Z}$, with K communities). Each expert is associated with a projection matrix $\mathbf{W}_k$, of dimension $p \times Q$, that groups the covariates into components. The data, coupled with the projection matrices, are transformed into matrices of size $N \times Q$, which are then passed through a regression function $f_k(. \mid \boldsymbol{\beta}_k)$ specific to each expert. The outputs of the experts are combined according to the gating network's weights (based on community membership) to produce a final prediction $\mathbf{y}$, representing the target variable.

In modeling the response variable $\mathbf{y}$ conditionally on covariates $\mathbf{X}$ and latent variables $\mathbf{Z}$ and $\mathbf{W}$, different strategies can be employed depending on the nature of the response variable's support. In this chapter, we focus on two primary settings: regression for continuous outcomes and multinomial logistic regression for categorical outcomes.

[1] https://github.com/Kdesantiago.

Regression When the support of $\mathbf{y}$ lies in $\mathbb{R}^N$, we adopt the following regression model:

$$y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_k \sim \mathcal{N}\left(y_i; \mathbf{X}_i \mathbf{W}_k \boldsymbol{\beta}_k, \sigma_k^2\right). \tag{3.15}$$

In this formulation, the regression parameters $\boldsymbol{\beta}_k \in \mathbb{R}^Q$ for each latent class k must be estimated, and σ_k^2, representing the class-specific noise variance, too.

Multinomial logistic regression For the case where the support of $\mathbf{y}$ is $\{1, \dots, C\}^N$, we use a multinomial logistic regression type model:

$$y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_k \sim \mathcal{M}\left(1; \mathrm{softmax}\left(\mathbf{X}_i \mathbf{W}_k \boldsymbol{\beta}_k\right)\right). \tag{3.16}$$

In this case, the classification parameters $\boldsymbol{\beta}_k \in \mathbb{R}^{Q \times C}$ are parameter matrices corresponding to each latent class k.

For a given community k, this model can be interpreted as a regression on the vector $(\mathbf{X}_i \mathbf{W}_k) \in \mathbb{R}^Q$. The components of this vector represent synthesized features of individual i, conditioned on the partition of community k, and reduced to Q variables. In other words, the p original variables are summarized into Q representative variables, each corresponding to a specific component. The graphical representation of the proposed model is provided in Figure 3.3.

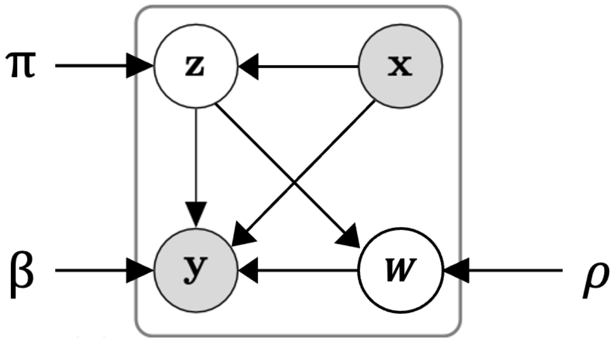


Figure 3.3: Graphical model for *MoEBIUS*. The joint distribution can be factorized as follows:
$\mathbb{P}(\mathbf{y}, \mathbf{Z}, \mathbf{W} \mid \mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\beta}) = \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\beta})\, \mathbb{P}(\mathbf{W} \mid \mathbf{Z}, \boldsymbol{\rho})\, \mathbb{P}(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\pi})$.

The complete likelihood can be expressed as:

$$\begin{aligned}\mathcal{L}^c\left(\boldsymbol{\beta},\boldsymbol{\rho},\boldsymbol{\pi};\mathbf{y},\mathbf{Z},\mathbf{W}\mid\mathbf{X}\right)&=\mathbb{P}\left(\mathbf{y},\mathbf{Z},\mathbf{W}\mid\mathbf{X},\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\beta}\right)\\&=\prod_{i=1}^{N}\prod_{k=1}^{K}\mathbb{P}\left(y_i\mid\mathbf{Z}_{ik},\mathbf{W}_k,\mathbf{X}_i,\boldsymbol{\beta}_k\right)\prod_{k=1}^{K}\prod_{j=1}^{p}\mathbb{P}\left(\mathbf{W}_{kj}\mid\mathbf{Z},\boldsymbol{\rho}\right)\\&\prod_{i=1}^{N}\mathbb{P}\left(\mathbf{Z}_i\mid\mathbf{X}_i,\boldsymbol{\pi}\right).\end{aligned}\tag{3.17}$$

3.3 Optimization via SEM-Gibbs algorithm

The Variational EM, as applied in the *mimiSBM* model (Section 2.3.3), is not utilized in this context due to the increased computational complexity. This complexity stems from the dependency between the latent variables forming the tensor $\mathbf{W}$, which is introduced by the regression component.

However, Gibbs sampling EM emerges as a viable alternative. This approach relies on sampling from the conditional distributions of each variable given the others, which, in our case, are known. This makes Gibbs sampling a practical and implementable strategy for optimizing the model while managing the latent structure effectively.

Through a stochastic EM formulation based on Gibbs sampling, we can estimate latent variables and model parameters for Latent Block Models (Keribin et al., 2012). Moreover, Gibbs sampling is a widely used Monte Carlo Markov Chain method for Bayesian inference in complex statistical models (Gelfand et al., 1990; Yildirim, 2012; Tobin et al., 2024).

As in classical EM algorithms, this optimization consists of two steps: the *Stochastic Expectation Gibbs-sampling step* (SE-Gibbs step) and the *Maximization step* (M step). We begin by defining the objective function to be optimized at each iteration of the algorithm:

$$\mathcal{J}\left(\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\beta}\mid\boldsymbol{\pi}^{(t)},\boldsymbol{\rho}^{(t)},\boldsymbol{\beta}^{(t)}\right)=\mathbb{E}_{\mathbf{Z},\mathbf{W}\sim\mathbb{P}\left(.\mid\mathbf{y},\mathbf{X},\boldsymbol{\pi}^{(t)},\boldsymbol{\rho}^{(t)},\boldsymbol{\beta}^{(t)}\right)}\left[\log\mathbb{P}\left(\mathbf{y},\mathbf{Z},\mathbf{W}\mid\mathbf{X},\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\beta}\right)\right].\tag{3.18}$$

3.3.1 Gibbs-sampling Expectation step

At step (t) of the algorithm, in Gibbs-sampling Expectation step, we first compute the following probabilities:

$$\tau_{ik}^{(t+1)} = \frac{\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}^{(t)}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \hat{\mathbf{W}}_k^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)}{\sum_{k'} \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik'}=1, \hat{\mathbf{W}}_{k'}^{(t)}, \boldsymbol{\beta}_{k'}^{(t)}\right)}, \tag{3.19}$$

with $\boldsymbol{\pi}_{\bullet k} = (\pi_{ik})_{i=1:N}$.

Next, we perform a random draw from a multinomial distribution:

$$\hat{\mathbf{Z}}_i^{(t+1)} \sim \mathcal{M}\left(1; \left(\tau_{i1}^{(t+1)}, \dots, \tau_{iK}^{(t+1)}\right)\right). \tag{3.20}$$

Similarly, the same operations are applied to $\boldsymbol{\nu}$ and $\mathbf{W}$:

$$\nu_{kjs}^{(t+1)} = \frac{\rho_{ks}^{(t)} \prod_i^N \mathbb{P}\left(y_i \mid \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, W_{kjs}=1, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}{\sum_{s'} \rho_{ks'}^{(t)} \prod_i^N \mathbb{P}\left(y_i \mid \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, W_{kjs'}=1, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}, \tag{3.21}$$

where the tensor $\hat{\mathbf{W}}_{-kj}^{(t)}$ corresponds to tensor $\hat{\mathbf{W}}^{(t)}$ with the third dimension element associated with the k-th entry of the first dimension and the j-th entry of the second removed.

$$\hat{\mathbf{W}}_{kj}^{(t+1)} \sim \mathcal{M}\left(1; \left(\nu_{kj1}^{(t+1)}, \dots, \nu_{kjQ}^{(t+1)}\right)\right). \tag{3.22}$$

3.3.2 Maximization step

Using the results from the SE-Gibbs step, we now estimate the model parameters $\boldsymbol{\pi}, \boldsymbol{\rho}, (\boldsymbol{\beta}_k)_{k=1:K}$.

For the component parameters $\boldsymbol{\rho}$, The estimation follows a natural approach, where, conditionally on community k, the estimation is based on counting the number of variables within the component s and dividing by the total number of variables p.

$$\rho_{ks}^{(t+1)} = \frac{\sum_{j=1}^{p} \hat{\mathbf{W}}_{kjs}^{(t+1)}}{p}, \qquad \forall k \in \{1, \dots, K\}, \forall s \in \{1, \dots, Q\}. \tag{3.23}$$

The parameters $\boldsymbol{\pi}$ and $(\boldsymbol{\beta}_k)_{k=1:K}$ (for multiclass classification) are estimated using a gradient ascent approach, where a step is performed at each iteration of the *SEM-gibbs* algorithm.

Regarding the logistic regression parameters $\boldsymbol{\pi}$ on the latent variables $(\mathbf{Z}_i)_{i=1:N}$:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\pi}} \left(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\beta} \mid \boldsymbol{\pi}^{(t)}, \boldsymbol{\rho}^{(t)}, \boldsymbol{\beta}^{(t)}\right) = \mathbf{X}^T \left(\hat{\mathbf{Z}}^{(t+1)} - \mathbf{S}^{\boldsymbol{\pi}^{(t)}}\right), \tag{3.24}$$

with

$$S_{ik}^{\pi} = \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}. \tag{3.25}$$

Here, $\hat{\mathbf{Z}}^{(t+1)}$ represents the "true" community membership matrix, while $\mathbf{S}^{\pi}$ is an estimate. In other words, the parameters $\boldsymbol{\pi}$ are optimized to produce predictions consistent with the posterior.

The parameters are updated as follows:

$$\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)} + h_t \frac{\partial \mathcal{J}}{\partial \boldsymbol{\pi}} \left(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\beta} \mid \boldsymbol{\pi}^{(t)}, \boldsymbol{\rho}^{(t)}, \boldsymbol{\beta}^{(t)}\right), \tag{3.26}$$

where h_t is the gradient ascent step size for iteration t.

The estimation of parameters $(\boldsymbol{\beta}_k)_{k=1:K}$ depends on the problem at hand.

Regression The estimation of $(\boldsymbol{\beta}_k)_{k=1:K}$ is given by:

$$\boldsymbol{\beta}_k^{(t+1)} = \left(W_k^{(t+1)^T} \mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right) \mathbf{X} \mathbf{W}_k^{(t+1)}\right)^{-1} \mathbf{W}_k^{(t+1)^T} \mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right) \mathbf{y}, \tag{3.27}$$

with $\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} = \left(\hat{\mathbf{Z}}_{ik}^{(t+1)}\right)_{i=1:N}$. This is a weighted least squares estimator, where the weighting is provided by $\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}$ and the data matrix is $\mathbf{X}\mathbf{W}_k^{(t+1)}$.

The estimation of $(\sigma_k^2)_{k=1:K}$ is given by:

$$\sigma_k^{2(t+1)} = \frac{\sum_{i=1}^N \hat{Z}_{ik}^{(t+1)} \left(y_i - \mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k^{(t+1)}\right)^2}{\sum_{i=1}^N \hat{Z}_{ik}^{(t+1)}}. \tag{3.28}$$

Once again, we find the weighted least squares estimator for σ_k^2.

Multiclass Classification The update for the parameters $(\boldsymbol{\beta}_k)_{k=1:K}$ is defined as:

$$\boldsymbol{\beta}_k^{(t+1)} = \boldsymbol{\beta}_k^{(t)} + h_t \frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k} \left(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\beta} \mid \boldsymbol{\pi}^{(t)}, \boldsymbol{\rho}^{(t)}, \boldsymbol{\beta}^{(t)}\right), \tag{3.29}$$

where the gradient is given by:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k}\left(\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\beta} \mid \boldsymbol{\pi}^{(t+1)},\boldsymbol{\rho}^{(t+1)},\boldsymbol{\beta}^{(t)}\right) = \left[\left(\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)}\right)^T \odot \mathbf{1}_{Q,1}\hat{\mathbf{Z}}_{\bullet k}^{(t+1)^T}\right]\left(\mathbf{y}-\mathbf{S}^{\beta_k^{(t)}}\right), \tag{3.30}$$

with the probability of variable y_i belonging to class c, for community k and given observation $\mathbf{X}_i$, defined as:

$$S_{ic}^{\beta_k} = \frac{e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\beta_{k\bullet c}}}{\sum_{c'} e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\beta_{k\bullet c'}}}. \tag{3.31}$$

Here, $\mathbf{1}_{Q,1}$ is a matrix of size $Q \times 1$ filled with ones,$\odot$ represents the Hadamard (element-wise) product and $\boldsymbol{\beta}_{k\bullet c} = (\boldsymbol{\beta}_{ksc})_{s=1:Q}$.

Further details on the optimization procedure are provided in Appendix E.

3.4 Model Selection

The model selection process for *MoEBIUS* is based on the ICL criterion proposed by Goffinet et al. (2020) for conditional co-clustering and the BIC criterion introduced by Schwarz (1978) to penalize the regression component.

$$\begin{aligned}\text{BIC\_ICL}(\mathbf{y},\mathbf{X},K,Q) = &\log \mathbb{P}\left(\mathbf{y},\hat{\mathbf{Z}},\hat{\mathbf{W}} \mid \mathbf{X},\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\rho}},\hat{\boldsymbol{\beta}}\right) - p\frac{K-1}{2}\log(N) \\ &- K\frac{Q-1}{2}\log(p) - \frac{O_{prob}}{2}\log(N),\end{aligned} \tag{3.32}$$

where O_{prob} represents the number of parameters depending on the type of model used for $\mathbf{y}$.

The criterion penalizes the log-likelihood based on the number of free parameters, thus preventing overfitting:

- For communities $\boldsymbol{\pi}$: $p \times (K-1)$ free parameters.

- For components $\boldsymbol{\rho}$: $K \times (Q-1)$ free parameters.

- For regression $\boldsymbol{\beta}$: $O_{prob} = 2 \times K \times Q$ free parameters.

- For multiclass classification $\boldsymbol{\beta}$: $O_{prob} = (C-1) \times K \times Q$ free parameters.

The best model is the one that maximizes the BIC_ICL criterion, achieving a balance between predictive performance and model complexity.

3.5 Experiments

The simulations will assess the model according to three main criteria:

- **Co-clustering and Regression Performance**: These will be evaluated on both training and test datasets. The test data will be generated using the same procedure as the training data, with $N_{test} = 1000$.

  For **co-clustering**, the Adjusted Rand Index (Steinley, 2004, ARI) and the Normalized Mutual Information (Strehl and Ghosh, 2002, NMI) scores will again be computed for rows and columns on the training set, but only for the rows in the test set. This is because we retain the same partitioning of variables between the training and test datasets, focusing on predicting the community membership of new observations. The ARI measures clustering accuracy by quantifying the agreement between estimated partitions and a reference partition, while the NMI quantifies the amount of shared information between the estimated clusters and the true underlying data structure. Both metrics are normalized, reaching a maximum value of 1 when the estimated clustering perfectly matches the reference clustering. Additionally, to ensure optimal alignment between the obtained and true clusters, the Hungarian algorithm will be applied to realign the generated clusters.

  For **regression**, the Mean Squared Error (MSE) and Mean Absolute Error (MAE) will be computed on both training and test data.

- **Regression Parameter Estimation**: The accuracy of the regression coefficients $\boldsymbol{\beta}$ will be analyzed.

- **Model Selection**: Different hyperparameter selection methods will be compared, including the ELBO, the BIC_ICL criterion, and cross-validation based on minimizing the MSE (CV).

3.5.1 Simulation process

The simulations are designed using a generative framework based on the Conditional Latent Block Model (*CLBM*). This framework simulates a set of N observations belonging to K communities, characterized by p variables following Q components. Each observation is drawn from normal distributions parameterized by $(\mu_{ks}, \sigma^2_{ks})$, where μ_{ks} and σ^2_{ks} represent the mean and variance, respectively, for community k and component s.

Four distinct simulation scenarios were designed to evaluate the model's performance across different contexts, as illustrated in Figure 3.4:

1. The first scenario corresponds to an ideal setting with $N = 1000$, $p = 30$, $K = 2$, and $Q = 3$, where the parameters $(\mu_{ks}, \sigma^2_{ks})$ are well-separated, ensuring a clear differentiation between communities and components.

$$\boldsymbol{\mu} = \begin{bmatrix} 1 & 7 & 5 \\ -3 & -2 & 1 \end{bmatrix}, \tag{3.33}$$

2. The second scenario introduces added complexity with $K = 5$ and $Q = 7$, while maintaining $N = 1000$ and $p = 30$, but the distributions become more blurred.

3. The third scenario maintains $p = 30$, $K = 2$, and $Q = 3$, but with very close parameters $(\mu_{ks}, \sigma^2_{ks})$ between communities, simulating more challenging-to-distinguish distributions.

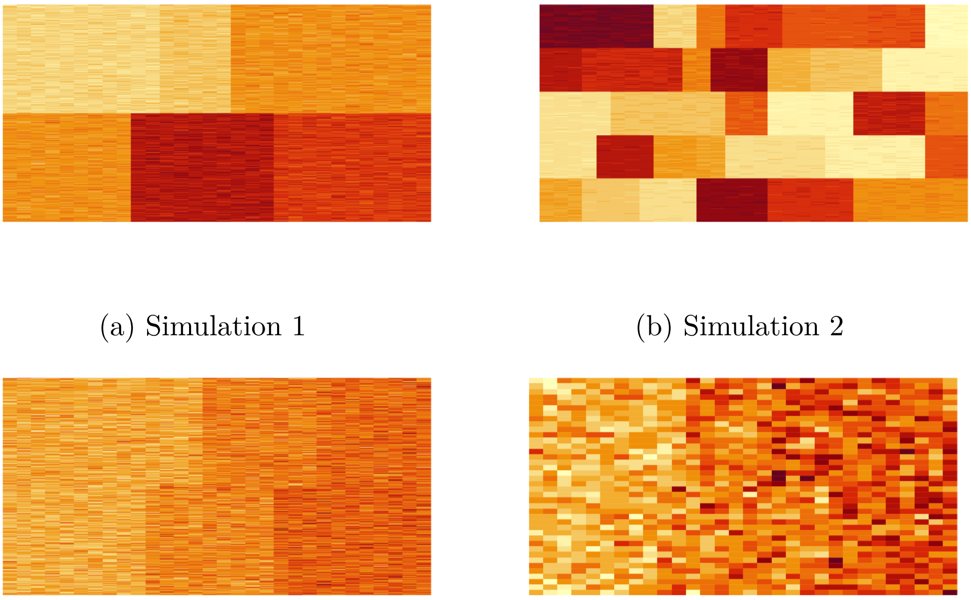
$$\mu_{k's} = \mu_{ks} \pm 0.2, \qquad \forall k, k' \in \{1, \dots, K\}. \tag{3.34}$$

$$\sigma^2_{ks} = 1, \qquad \forall (k, s) \in \{1, \dots, K\} \times \{1, \dots, Q\} \tag{3.35}$$

4. Finally, the fourth scenario combines the close parameters from scenario 3 with a reduced number of observations ($N = 40$), while keeping $p = 30$, $K = 2$, and $Q = 3$.

For each scenario, 100 repetitions were conducted to ensure robust and reliable performance evaluation of the model.

In addition, we introduce a regression component by defining matrices $\boldsymbol{\beta}$, regression parameters on the components, as in the *MoEBIUS* model. The target variable $\mathbf{y}$ thus generated depends on the communities and the variables



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure 3.4: Examples of generated data $\mathbf{X}$ according to the four simulation scenarios. As the simulation process progresses, the co-clustering task becomes increasingly complex.

representing the components, and therefore ultimately on the grouping of the variables.

For simulations 1, 3, and 4, the regression matrix $\boldsymbol{\beta}$ is defined as:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & -2 & 2 \\ -2 & 1 & 3 \end{bmatrix},$$

and for simulation 2, it is:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 & 3 & 5 & 0 & 1 \\ -1 & 0 & 1 & 3 & 0 & 1 & 10 \\ 4 & 3 & -2 & -2 & 1 & 0 & -1 \\ 0 & 2 & -1 & -1 & -1 & 1 & 0 \\ -3 & -3 & 2 & -2 & 5 & 0 & -7 \end{bmatrix}.$$

3.5.2 Co-clustering and regression performance

In this section, we compare the *MoEBIUS* model with two other regression models. The first comparison is with the *Global Model*, which corresponds to a standard linear regression, in order to evaluate the benefit of modeling both communities and components within the framework. The second comparison is with the *K-Means + Reg.* model, which employs *K-Means* applied on rows and columns to identify communities and components. A linear regression is then applied within each community, using a representative variable for each component. These representative variables are derived by summing all the variables within the relevant component.

The use of *K-Means* is motivated by the unsupervised performance observed in Section F.5, where it also serves as the initialization method for *MoEBIUS*. Furthermore, *K-Means* facilitates the prediction of community memberships for new data, enabling a direct comparison of clustering performance between *K-Means + Reg.* and *MoEBIUS* on the test set. Finally, this approach allows us to compare the unified *MoEBIUS* model with its two-step counterpart, *K-Means + Reg.*

The results of these comparisons are compiled in Table 3.1.

In simulation 1 (ideal case), the *Global Model* fails to effectively capture the variability in the data compared to the *K-Means + Reg.* and *MoEBIUS* models. The *K-Means + Reg.* model enhances regression by effectively detecting communities, both during training and prediction phases. However, *MoEBIUS* achieves the best regression performance, despite a slight decrease in clustering performance during prediction.

For simulations 2 to 4, the *Global Model* shows average performance on both training and test datasets. The *K-Means + Reg.* model struggles to generalize the latent structures, leading to significantly degraded predictions. In contrast, *MoEBIUS* excels in accurately identifying communities in the test data, although it performs less well in variable clustering. This explains its substantial advantage in predictive performance over the *K-Means + Reg.* model. However, a performance drop is observed in simulation 4, which is attributed to the very small sample size ($N = 40$).

Overall, *MoEBIUS* consistently outperforms the other models by capturing both the underlying data structures and providing accurate predictions. Its flexibility in increasingly complex scenarios, combined with stable perfor-

Examples	Algorithms	Regression				Clustering					
		Train		Test		Row (train)		Col (train)		Row (test)	
		MSE	MAE	MSE	MAE	ARI	NMI	ARI	NMI	ARI	NMI
Simulation 1	Global model	60.053	6.101	63.043	6.232	–	–	–	–	–	–
		(17.573)	(0.844)	(19.595)	(0.885)	–	–	–	–	–	–
	K-Means + Reg.	15.295	1.668	15.614	1.676	**1.000**	**1.000**	**0.838**	**0.881**	**1.000**	**1.000**
		(17.803)	(1.835)	(18.294)	(1.854)	(0.000)	(0.000)	(0.172)	(0.125)	(0.000)	(0.000)
	MoEBIUS	**6.244**	**0.120**	**8.294**	**0.147**	**1.000**	**1.000**	**0.838**	**0.881**	0.994	0.986
		(7.613)	(0.101)	(10.025)	(0.138)	(0.000)	(0.000)	(0.172)	(0.125)	(0.006)	(0.013)
Simulation 2	Global model	333.240	14.098	355.662	14.581	–	–	–	–	–	–
		(112.896)	(2.348)	(115.405)	(2.331)	–	–	–	–	–	–
	K-Means + Reg.	76.944	5.605	122507.430	221.904	0.771	0.854	**0.525**	**0.726**	0.437	0.511
		(58.178)	(2.075)	(106625.273)	(112.638)	(0.176)	(0.119)	(0.143)	(0.096)	(0.196)	(0.194)
	MoEBIUS	**14.711**	**1.379**	**36.157**	**1.514**	**0.980**	**0.985**	**0.525**	**0.726**	**0.980**	**0.984**
		(21.923)	(1.138)	(70.954)	(1.215)	(0.062)	(0.049)	(0.143)	(0.096)	(0.063)	(0.050)
Simulation 3	Global model	115.641	7.980	122.957	8.234	–	–	–	–	–	–
		(75.623)	(2.408)	(80.130)	(2.461)	–	–	–	–	–	–
	K-Means + Reg.	**18.496**	**1.611**	1289.846	21.333	**0.989**	**0.976**	**0.843**	**0.884**	0.033	0.027
		(21.253)	(1.544)	(957.637)	(9.159)	(0.011)	(0.022)	(0.173)	(0.129)	(0.046)	(0.037)
	MoEBIUS	63.844	1.953	**71.233**	**2.090**	0.972	0.972	**0.843**	**0.884**	**0.856**	**0.782**
		(77.511)	(2.114)	(82.702)	(2.148)	(0.160)	(0.160)	(0.173)	(0.129)	(0.146)	(0.139)
Simulation 4	Global model	31.162	4.197	726.877	20.057	–	–	–	–	–	–
		(25.293)	(1.688)	(661.567)	(7.741)	–	–	–	–	–	–
	K-Means + Reg.	23.613	2.783	2368.281	37.004	**0.978**	**0.971**	**0.691**	**0.736**	0.044	0.036
		(36.566)	(1.620)	(1575.762)	(12.194)	(0.058)	(0.074)	(0.139)	(0.107)	(0.060)	(0.047)
	MoEBIUS	**9.916**	**1.927**	**287.727**	**7.644**	**0.978**	0.970	**0.691**	**0.736**	**0.657**	**0.562**
		(11.279)	(0.879)	(268.338)	(3.758)	(0.065)	(0.084)	(0.139)	(0.107)	(0.131)	(0.130)

Table 3.1: Comparison of regression and co-clustering performance on simulated data. The mean (standard deviation) of the scores obtained over 100 simulations.

mance in both clustering and regression, makes it particularly well-suited for applications requiring fine segmentation and robust predictive modeling, even with limited observations.

3.5.3 Parameter estimations

The results presented in Figure 3.5 show that in simulations 1 and 3, the parameters $\boldsymbol{\beta}$ are generally estimated with good accuracy. However, simulation 2, which involves a larger number of parameters, makes the estimation process more complex. Depending on the repetitions, certain parameters are very accurately estimated. Nonetheless, the vast majority of parameters are recovered.

Simulation 4 provides an interesting perspective on the model's robustness. Despite a drastic reduction in the number of observations, the parameter estimates, while less accurate, remain within an acceptable range. This resilience in the face of reduced input data highlights the model's robustness due to its statistical framework.

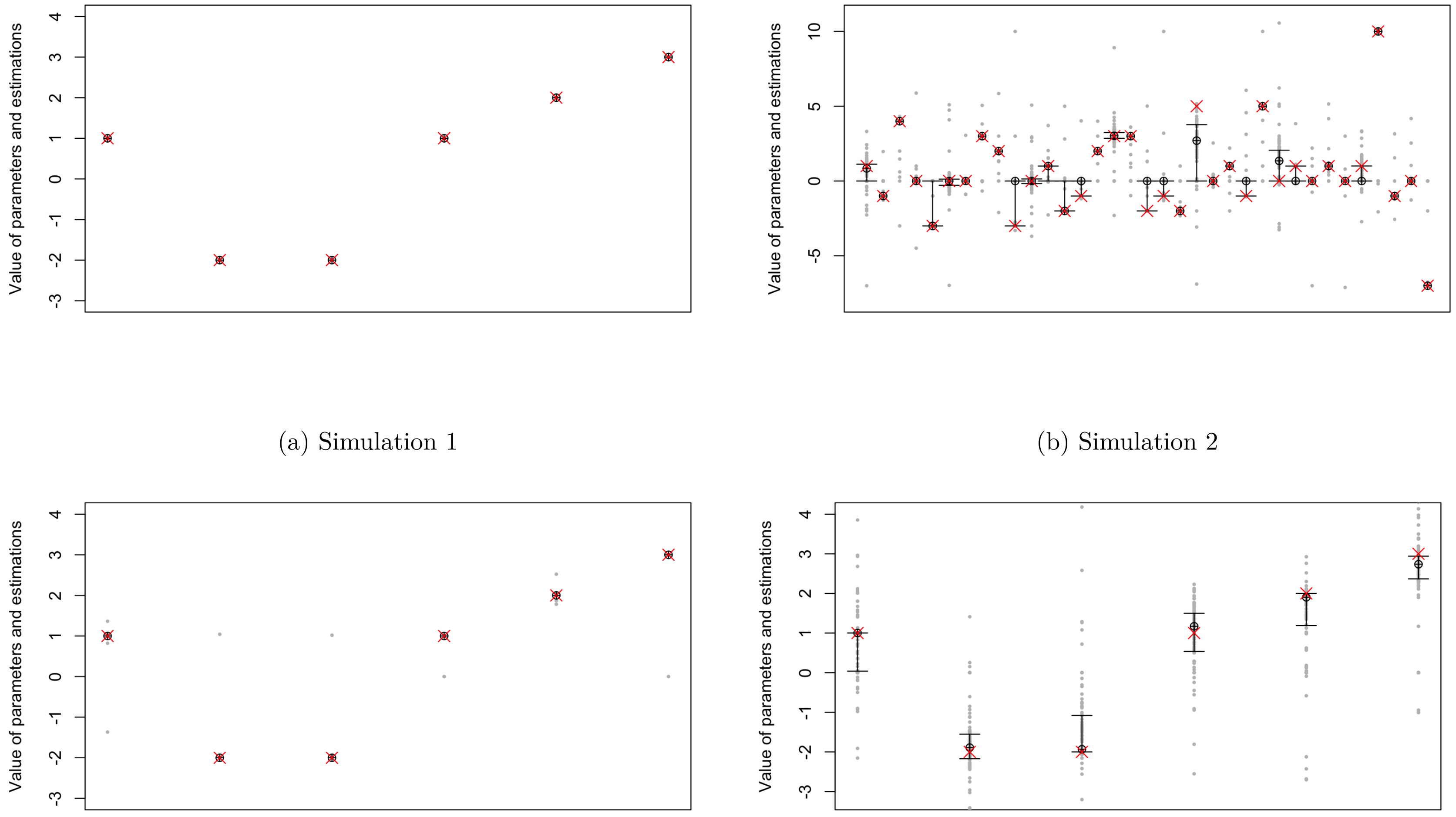
3.5.4 Model selection

Figures 3.6 and 3.7 compare the effectiveness of three hyperparameter selection methods: BIC_ICL, ELBO, and cross-validation (CV).

From Figures 3.6 and 3.7, we observe that the BIC_ICL criterion (yellow) consistently identifies the true values of K across all four simulations. BIC_ICL significantly outperforms ELBO and CV in terms of robustness for community detection across different simulation settings. However, when focusing on Q, the performance is reduced. Although BIC_ICL often selects the correct value, it occasionally chooses other values, particularly in more complex simulations (Figures 3.7b and 3.7c). This suggests that a clear distinction between components and a larger number of variables may be necessary to fully leverage this criterion for model selection.

The CV method (blue) is the least effective of the three, struggling to identify the correct values of K and Q in most simulations, and tending to overestimate both parameters, even under ideal conditions (Figures 3.6a and 3.7a).

The ELBO criterion (green), while occasionally effective, yields inconsistent results. For example, it nearly identifies the correct value of K in sim-



(a) Simulation 1

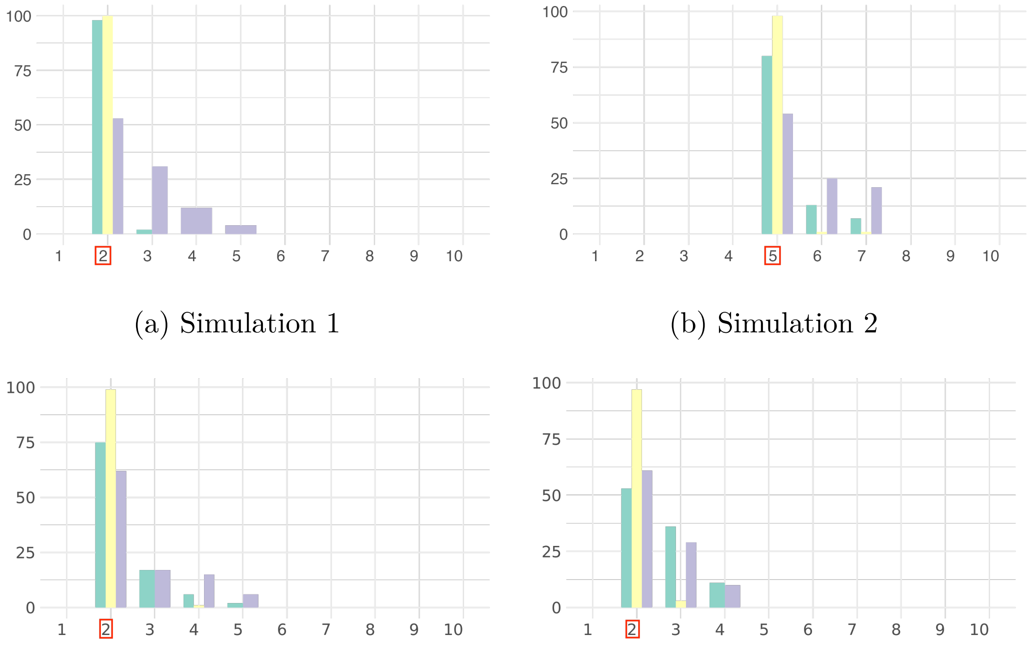
(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure 3.5: Parameter estimates across the four simulation settings. The x-axis represents the progression of different variables, while the y-axis indicates the estimated parameter values. The red cross marks the true parameter value. Black circles represent the median observed values for each variable, with the error bars indicating the first and third quartiles. Grey dots illustrate the estimates for each repetition.

ulations 1 and 2 (Figures 3.6a and 3.6b), but struggles when the component parameters are closely related (Figures 3.6c and 3.6d). Similarly, for the hyperparameter Q, ELBO performs acceptably in Figures 3.7a and 3.7d, with the correct parameter being selected in the majority of cases. However, in other scenarios, its performance is much more erratic.



(a) Simulation 1

(b) Simulation 2

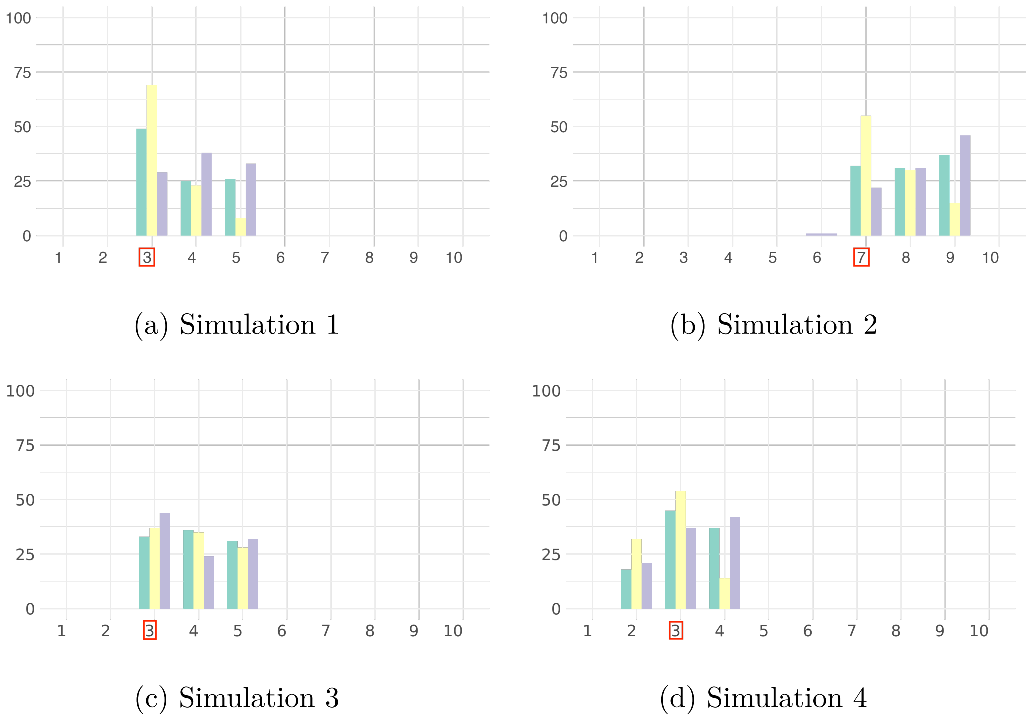
(c) Simulation 3

(d) Simulation 4

Figure 3.6: Comparison of hyperparameter K selection across BIC_ICL (yellow), ELBO (green), and CV (blue) across the four simulation settings.

The simulations suggest that BIC_ICL is the most robust and effective model selection method for identifying hyperparameters K and Q. While ELBO provides an interesting alternative, particularly for the parameter K, it does not reliably select the hyperparameter Q. The CV method under the MSE criterion is the least reliable and should be avoided if possible. One possible explanation is that the criterion becomes easier to minimize as the number of communities and components increases, allowing for more detailed regressions within the considered communities. By increasing the number of components, the model gains flexibility in capturing variations within each community.

Furthermore, when a component is split into two (considered as two distinct components), in the context of the linear model, it is sufficient to assign



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure 3.7: Comparison of hyperparameter Q selection across BIC_ICL (yellow), ELBO (green), and CV (blue) across the four simulation settings.

them the same coefficient. This ensures that their combined effect behaves as if it were a single representative variable, thus maintaining the overall influence unchanged. The representative variable is simply the sum of the component variables, in this context.

3.6 Bilan

This chapter explores the integration of conditional biclustering in mixture of experts models to enhance interpretability and predictive performance within the context of multimodal machine learning. Rather than treating variables uniformly across all individuals, the objective was to uncover finer, more community-specific relationships, while characterizing both redundancy and complementarity of the information.

MoEBIUS enables the construction of more accurate and interpretable regression and classification models by adapting the model parameters to the specificities of each community (Figure 3.2). Furthermore, grouping variables into components allows for the characterization of both redundant and complementary information. This approach, by using representative variables for each component, thereby reducing computational complexity. Experiments conducted on simulated data confirmed the effectiveness of *MoEBIUS*, significantly outperforming both a global linear regression model and a two-step model using K-means followed by linear regression. *MoEBIUS* demonstrated superiority in terms of regression performance and co-clustering accuracy, particularly in complex scenarios with a limited number of observations (Table 3.1). Moreover, the BIC_ICL criterion proved to be the most reliable model selection method for *MoEBIUS*, enabling precise selection of hyperparameters K and Q in most simulation scenarios (Figures 3.6 and 3.7).

4 Conclusion and perspectives

Cette thèse explore l'importance de la modélisation conjointe des individus et des variables pour l'analyse de données multimodales. Deux contributions sont présentées: le modèle *mimi-SBM* pour le clustering multi-vues, et le modèle *MoEBIUS* pour l'intégration d'un biclustering conditionnel au sein d'un modèle d'expert.

Le chapitre 1 traite l'apprentissage multimodal en caractérisant les trois schémas principaux d'intégration, à savoir l'intégration verticale, horizontale et diagonale. Il examine également les problématiques liées à l'intégration de données multimodales, en abordant la représentation et le stade de fusion de l'information dans la modélisation, ainsi que la nature de l'information recherchée qui peut être redondante ou complémentaire.

Le chapitre 2 introduit *mimi-SBM*, un modèle bloc stochastique multicouches (*MMLSBM* pour Mixture of MultiLayer Stochastic Block Model) fréquentiste et bayésien

dans un contexte de clustering multi-vues, avec une stratification d'individus en communautés transverses à des vues homogènes regroupées en composantes.

Le chapitre 3 aborde la fusion entre les méthodes de biclustering conditionnel et les modèles d'experts, à travers le modèle *MoEBIUS*. L'objectif est de tirer parti de la flexibilité du Conditional Latent Block Model (*CLBM*) pour modéliser des relations entre les variables conditionnellement aux communautés, tout en exploitant la capacité des MoEs à réaliser des prédictions précises en présence d'hétérogénéité dans les données. La modélisation du *CLBM* est utilisé pour partitionner les individus (gating network) et les variables. Les experts, des modèles de régression, sont spécialisés sur chaque communauté, utilisant les variables représentatives issues de chaque composante.

Les paragraphes suivants seront d'abord consacrés à des pistes visant à approfondir les méthodes développées au cours de cette thèse. Ensuite, j' explorerai une extension de l'approche supervisée, en intégrant une fusion plus étroite entre les Conditional LBM et les MoEs.

Perspectives : mimi-SBM

Un des axes d'améliorations possibles pour *mimi-SBM* serait d'intégrer l' information realtive aux noeuds. Une première piste consisterait à poursuivre les travaux de Zanghi et al. (2010) en dans le cadre multicouches, avec une partition des données caractérisant les noeuds en composantes. Une autre piste serait associée aux travaux de Mariadassou et al. (2010) qui modélisent ces informations directement dans la loi d'émission des SBMs. Cette stratégie pourrait intégrer une stratification pour relier les composantes aux variables concomitantes.

Sur les aspects relatifs au passage à l'échelle, les approches de type Deep Learning permettrait de modéliser ce type de relations complexes. Par exemple, en étendant le Latent Position Model (Hoff et al., 2002), où la proximité dans l'espace latent reflète la probabilité de connexion, il serait possible de capturer des structures plus riches. Des travaux récents, comme ceux de Boutin et al. (2023) ou encore Liang et al. (2024), ont montré l'efficacité de ces méthodes pour structurer des données et identifier des relations latentes, offrant des pistes prometteuses pour combiner Deep Learning et modélisation probabiliste des réseaux.

Enfin, l'étude des réseaux multiplexes dynamiques, basée sur les travaux de Matias and Miele (2017), offrirait une opportunité de modéliser l'évolution simultanée de plusieurs réseaux pour par exemple comprendre l'évolution de comportements ou d'interactions dans différents contextes.

Perspectives : MoEBIUS

Dans les perspectives d'extension de *MoEBIUS*, une approche temporelle permettrait de capturer la dynamique de données longitudinales obtenues dans le cadre d'examens répétés par exemple. L'évolution des données au cours du temps peut être prise en compte dans le modèle développé initialement, mais aussi dans une nouvelle modélisation. Ainsi, les données receuillies seraient un tenseur et le problème de machine learning se rapprocherait du Multi-task learning. Cette représentation plus complexe du problème se rapprocherait des travaux de Boutalbi et al. (2020) auxquels une structure temporelle et conditionelle seraient ajoutées. Cette extension serait particulièrement pertinente pour des applications dans des domaines comme la médecine personnalisée où les données longitudinales permettent de suivre l'évolution des patients et d'adapter les interventions thérapeutiques.

Une autre amélioration pourrait se porter sur la variable représentative de chaque composante. En effet, dans le modèle actuel, il s'agit simplement d'une somme des valeurs des variables. L'ajout de paramètres permettant une combinaison linéaire des variables représentatives, en passant de $\mathbf{X}\mathbf{W}_k^T$ à $\mathbf{X}\boldsymbol{\alpha}_k\mathbf{W}_k^T$, apporterait une flexibilité supplémentaire au modèle. Cette modification offrirait la possibilité de pondérer différemment chaque variable d'entrée.

Une autre approche consisterait à introduire une nouvelle variable $\tilde{\mathbf{W}}$ qui suivrait une loi conditionnellement à $\mathbf{W}$, par exemple relative à une approche de type spike-and-slab (Mitchell and Beauchamp, 1988). Cela ouvrirait la voie à des explorations sur le choix optimal de la distribution de $\tilde{\mathbf{W}}$, en fonction du type de données et de la structure du problème. Concernant le rôle du vecteur de régression $\boldsymbol{\beta}$ dans ce cadre, une réflexion plus approfondie serait nécessaire pour déterminer comment $\tilde{\mathbf{W}}$ interagirait avec lui, ainsi que sa pertinence.

Conditional Latent Block Mixture of Experts

Le modèle présenté en section 3.2 peut-être étendu avec une variable cible issue d'un modèle de mélange. Cette approche permettrait de tenir compte des informations sous-jacentes au sein des composantes et mettrait en lumière la contribution des sources d'informations dans la qualité de prédiction.

Concernant les variables latentes, celles-ci sont maintenues avec des distributions identiques. Pour cela, les variables latentes définies aux équations (3.13) et (3.14) seraient maintenues. Cependant, il est nécessaire d'introduire une nouvelle variable:

$$\mathbf{G}_i \mid Z_{ik} = 1 \sim \mathcal{M}\left(1; \mathrm{softmax}\left(\mathbf{X}_i \boldsymbol{\gamma}_k\right)\right), \tag{4.1}$$

avec $\boldsymbol{\gamma} \in \mathbb{R}^{p \times Q}$. Elle a pour fonction d'indiquer quelle composante (source d'information) a été utilisée pour la prédiction. Plus précisement, on suppose que la prédiction peut être donnée à partir d'une contribution pondérée de modèles d'experts spécialisés sur chaque composante. De plus, à l'instar des variables $(\mathbf{Z}_i)_{i=1:N}$, sa distribution paramétrique permet la prédiction de l'expert sollicité pour de nouvelles observations.

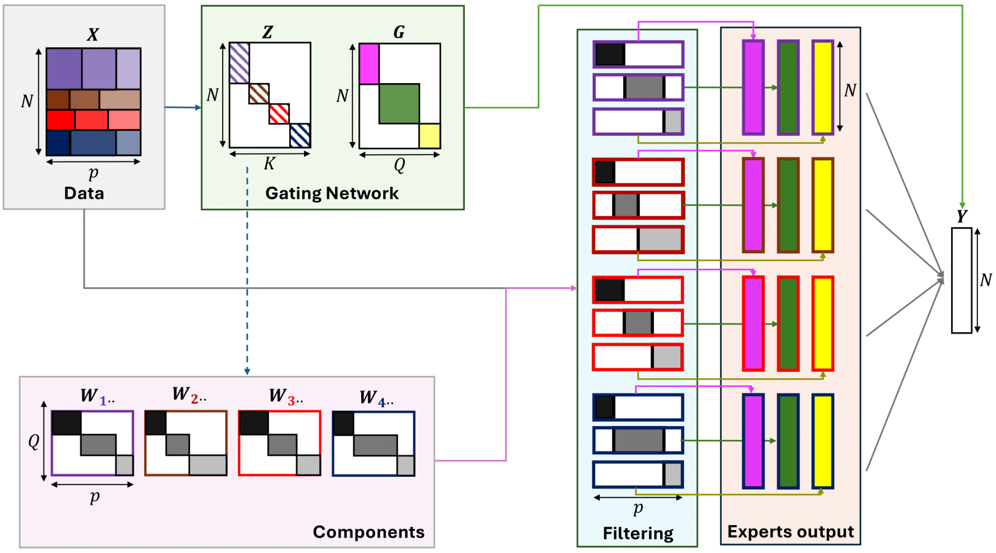


Figure 4.1: Co-Conditional Latent Block Mixture of Experts (*Co-CoLBMoE*). Le modèle repose sur un Gating Network déterminé par la combinaison des variables $\mathbf{Z}$ et $\mathbf{G}$, qui indiquent respectivement la communauté et la composante activées. Les partitions des covariables $\mathbf{W}$ permettent de filtrer les données transmises aux experts, les spécialisant ainsi sur un groupe spécifique. Par conséquent, chaque expert est dédié à une communauté et à une composante précises, renforçant ainsi leur spécialisation.

Nous souhaitons à nouveau modéliser $\mathbf{y} \in \mathbb{R}^N$ (ou $\{1, \ldots, C\}^N$), conditionnellement aux observations $(\mathbf{X}_i)_{i=1:N}$ et aux variables latentes $\mathbf{Z}, \mathbf{W}, \mathbf{G}$. Selon le cadre considéré, on aurait pour la régression

$$y_i \mid \mathbf{X}_i, \mathbf{Z}_{ik} = 1, \mathbf{G}_{is} = 1, \mathbf{W}_{k\bullet s}, \sim \mathcal{N}\left(\left(\mathbf{X}_i \odot \mathbf{W}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet}, \sigma^2_{ks}\right), \qquad (4.2)$$

avec $\boldsymbol{\beta}_{ks} \in \mathbb{R}^p$, et pour la classification

$$y_i \mid \mathbf{X}_i, \mathbf{Z}_{ik} = 1, \mathbf{G}_{is} = 1, \mathbf{W}_{k \bullet s} \sim \mathcal{M}\left(y_i; \operatorname{softmax}\left(\left(\mathbf{X}_i \odot \mathbf{W}_{k \bullet s}\right) \boldsymbol{\beta}_{ks \bullet}\right)\right), \quad (4.3)$$

avec $\boldsymbol{\beta}_{ks} \in \mathbb{R}^{p \times C}$.

Contrairement aux équations 3.15 et 3.16, où les paramètres de régression sont modélisés par des régressions sur des variables représentatives, ici les paramètres de régression sont directement associés aux p variables, masqués selon $W_{k \bullet s}$. Ainsi, on obtient une paramétrisation à la fois par source d' information et par communauté.

L'estimation des paramètres de ce modèle est disponible en Annexe H. Une comparaison des avantages de l'intégration de la variable de selection d'expert G avec le modèle précédemment développé reste à faire.

A Identifiability of mimiSBM

This appendix is dedicated to the proof of the theorem of Section 2.3.2 related to the identifiability of the parameters of *mimi-SBM*, recalled below. The proof is very similar to the one of Celisse et al. (2012) and make use of algebraic properties to prove that the parameters depend solely on the marginal distribution of our data.

Theorem 2 *Let* $N \geq \max(2K, 4Q)$ *and* $V \geq 2K$*. Assume that for any* $1 \leq k, l \leq K$ *and every* $1 \leq s \leq Q$*, the coordinates of* $\boldsymbol{\pi}^T \boldsymbol{\alpha}_{k..}\boldsymbol{\rho}$ *are all different,* $(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{..s}\boldsymbol{\pi})_{s=1:Q}$ *are distinct, and each* $(\boldsymbol{\alpha}_{kl.}\boldsymbol{\rho})_{k,l=1:K}$ *differs. Then, the mimi-SBM parameter* $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *is identifiable.*

A.1 Assumptions

$\mathcal{A}1$: $(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{k..}\boldsymbol{\rho})_{k=1:K}$ are all different.

$\mathcal{A}2$: $(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{..s}\boldsymbol{\pi})_{s=1:Q}$ are all different.

$\mathcal{A}3$: $N, V \geq 2K$.

$\mathcal{A}4$: $N \geq 4Q$.

$\mathcal{A}5$: $(\boldsymbol{\alpha}_{kl.}\boldsymbol{\rho})_{k,l=1:K}$ are all different.

A.2 Identifiability of $\boldsymbol{\pi}$

To prove the identifiability of $\boldsymbol{\pi}$, we first need to establish some correspondences. For any $1 \leq k \leq K$, $\forall(i, j, v)$, let r_k be the probability that an edge between i and j in layer v given individual i is in the cluster k:

$$
\begin{aligned}
r_k &= \mathbb{P}(A_{ijv} = 1 \mid Z_i = k) \\
&= \sum_l \sum_s \mathbb{P}(A_{ijv} = 1 \mid Z_i = k, Z_j = l, W_v = s)\ \pi_l\ \rho_s \\
&= \sum_l \sum_s \alpha_{kls}\ \pi_l\ \rho_s \\
&= \boldsymbol{\pi}^T \boldsymbol{\alpha}_{k..} \boldsymbol{\rho}\,. \qquad \text{(A.1)}
\end{aligned}
$$

Proposition 1 (Invertibility of R) *Let* $\mathbf{R}$ *denote a Vandermonde matrix of size* $K \times K$ *such as* $R_{ik} = (r_k)^{i-1},\ 1 \leq i, k \leq K$*.* $\mathbf{R}$ *is invertible, since the coordinates of* r *are all different according to Assumption* $\mathcal{A}1$*.*

Furthermore, for $2 \leq i \leq K$, the joint probability of having $(i-1)$ edges is given by:

$$
\begin{aligned}
&\mathbb{P}(A_{121} = 1, A_{132} = 1, \ldots, A_{1i(i-1)} = 1 \mid Z_1 = k) \\
&\quad = \sum_l \sum_s \mathbb{P}(A_{121} = 1, A_{132} = 1, \ldots, A_{1i(i-1)} = 1 \mid Z_1 = k, Z_2 = l, W_1 = s)\ \pi_l\ \rho_s \\
&\quad = \mathbb{P}(A_{132} = 1, \ldots, A_{1i(1-1)} = 1 \mid Z_1 = k) \\
&\qquad\quad \times \sum_l \sum_s \mathbb{P}(A_{121} = 1 \mid Z_1 = k, Z_2 = l, W_1 = s)\ \pi_l\ \rho_s \\
&\quad = \mathbb{P}(A_{132} = 1, \ldots, A_{1i(1-1)} = 1 \mid Z_1 = k)\ r_k \\
&\quad = (r_k)^{i-1} \\
&\quad = R_{ik}\,. \qquad \text{(A.2)}
\end{aligned}
$$

Now, we define $u_0 = 1$ and for $1 \leq i \leq 2K-1$:

$$
\begin{aligned}
u_i &= \mathbb{P}(A_{121} = 1, A_{132} = 1, \ldots, A_{1i(1-1)} = 1, A_{1(i+1)i} = 1) \\
&= \sum_k \mathbb{P}(A_{121} = 1, A_{132} = 1, \ldots, A_{1(1+1)i} = 1 \mid Z_1 = k)\ \pi_k \\
&= \sum_k (r_k)^i\ \pi_k\,. \qquad \text{(A.3)}
\end{aligned}
$$

By Assumption $\mathcal{A}3$, $(u_i)_{i=1:(2K-1)}$ are well defined. Hence, $u_0 = 1$ and $(u_i)_{i=1:(2K-1)}$

are known and defined from the marginal $\mathbb{P}_A$. As a consequence, $(u_i)_{i=1:(2K-1)}$ are identifiable.

Also, let $\mathbf{M}$ of size $(K+1) \times K$ be the matrix given by $M_{ij} = u_{i+j-2}$ for $1 \leq i \leq K+1$ and $1 \leq j \leq K$, and let $\mathbf{M}_{-i}$ denote the square matrix obtained by removing the row i from $\mathbf{M}$. The coefficients of $\mathbf{M}_{-(K+1)}$, for $1 \leq i, j \leq K$, are:

$$\begin{aligned} M_{ij} &= \sum_{k=1}^{K} (r_k)^{i-1} \; \pi_k \; (r_k)^{j-1}, \text{ and} \\ \mathbf{M}_{-(K+1)} &= \mathbf{R}\,\mathrm{Diag}(\boldsymbol{\pi})\mathbf{R}^T . \end{aligned} \tag{A.4}$$

Proposition 2 (Relations between R, M and $\boldsymbol{\pi}$) *From Proposition 1 and Equation* (A.4)*, we can define*

$$\mathbf{M}_{-(K+1)} = \mathbf{R} \;\; \mathrm{Diag}(\boldsymbol{\pi}) \;\; \mathbf{R}^T . \tag{A.5}$$

The correspondence of the different terms being established, we now need to prove the identifiability of $\boldsymbol{\pi}$, which means showing that $\mathbf{M}_{-(K+1)}$ and $\mathbf{R}$ are identifiable.

First, for the identifiability of r_k, with $\delta_k = \mathrm{Det}(\mathbf{M}_{-k})$, we define a polynomial function B such as:

$$B(x) = \sum_{k=0}^{K} (-1)^{K+k} \; \delta_{k+1} \; x^k . \tag{A.6}$$

This polynomial function has two important properties.

Proposition 3 *Let* $\deg(B)$ *denote the degree of* B*. We have* $\deg(B) = K$*.*

Proof Let $\delta_{K+1} = \mathrm{Det}(\mathbf{M}_{-(K+1)})$, with $M_{-(K+1)} = \mathbf{R} \;\; \mathrm{Diag}(\boldsymbol{\pi}) \;\; \mathbf{R}^T$ as stated in Proposition 2, and $\mathbf{R}$ being invertible as stated in Proposition 1. In consequence, $\mathbf{M}_{-(K+1)}$ is the product of invertible matrices, $\delta_{K+1} = \mathrm{Det}(\mathbf{M}_{-(K+1)}) \neq 0$ and, moreover, $\deg(B) = K$. ■

Proposition 4 *For* $1 \leq k \leq K$*,* $B(r_k) = 0$*.*

Proof Let $\mathbf{N}_k$ of size $(K+1) \times (K+1)$ be the concatenation in columns of the matrix $\mathbf{M}$ with the vector $V_k = [1, r_k, r_k^2, \ldots, r_k^K]^T$.

Now let's calculate the determinant of $\mathbf{N}_k$ developed by the last column:

$$\begin{aligned}
\det(\mathbf{N}_k) &= \sum_{l=0}^{K} (-1)^{K+1+l+1} \, (r_k)^l \, \det(\mathbf{M}_{l+1}) \\
&= \sum_{l=0}^{K} (-1)^{K+l} \, \delta_{l+1} \, (r_k)^l \\
&= B(r_k). \qquad \text{(A.7)}
\end{aligned}$$

In addition, the jth column of the $\mathbf{M}$ matrix can be written as $M_{.j} = \sum_{k=1}^{K} r_k^{j-1} \pi_k V_k$. Therefore, $\text{rank}(\mathbf{N}_k) < K+1$ and $\det(\mathbf{N}_k) = 0$ for $1 \le k \le K$. In consequence, $B(r_k) = 0$ for $1 \le k \le K$. ■

With $(r_k)_{k=1:K}$ being the roots of B (proposition 4), they are functions of $(\delta_k)_{k=1:K+1}$ which are themselves derived from $\mathbb{P}_A$. Also, $(r_k)_{k=1:K}$ can be expressed in a unique way (up to label switching) from $\mathbb{P}_A$, thus $(r_k)_{k=1:K}$ are identifiable. In consequence, $\mathbf{R}$ is also identifiable by definition. Finally, since $\mathbf{M}_{-(K+1)}$ and $\mathbf{R}$ are identifiable and invertible, $\text{Diag}(\boldsymbol{\pi}) = \mathbf{R}^{-1}\mathbf{M}_{-(K+1)}(\mathbf{R}^T)^{-1}$. In conclusion, $\boldsymbol{\pi}$ is identifiable.

A.3 Identifiability of $\boldsymbol{\rho}$

Identifiability of $\boldsymbol{\rho}$ is similar to $\boldsymbol{\pi}$, the main difference lies in the assumptions made and the quantities defined.

For any $1 \le s \le Q$, $\forall(i,j,v)$, let t_s be the probability of an edge between i and j in layer v given view v is in the component s:

$$\begin{aligned}
t_s &= \mathbb{P}(A_{ijv} = 1 \mid W_v = s) \\
&= \sum_{l} \sum_{k} \mathbb{P}(A_{ijv} = 1 \mid Z_i = k, Z_j = l, W_v = s) \; \pi_l \; \pi_k \\
&= \sum_{l} \sum_{k} \alpha_{kls} \; \pi_l \; \pi_k \\
&= \boldsymbol{\pi}^T \boldsymbol{\alpha}_{..s} \boldsymbol{\pi} \, . \qquad \text{(A.8)}
\end{aligned}$$

Proposition 5 (Invertibility of T) *Let* $\mathbf{T}$ *denote a Vandermonde matrix of size* $Q \times Q$ *such as* $T_{is} = (t_s)^{i-1}$, $1 \le i, s \le Q$. $\mathbf{T}$ *is invertible, since the coordinates of* (t_s) *are all different according to Assumption* $\mathcal{A}2$.

Let's define the joint probability of $i-1$ edges given the latent component of the view 1:

$$\begin{aligned}
&\mathbb{P}(A_{121}=1, A_{341}=1, \ldots, A_{2i-1\,2i\,1}=1 \mid W_1=s)\\
&\quad=\sum_l \sum_k \mathbb{P}(A_{121}=1, A_{341}=1, \ldots, A_{2i-1\,2i\,1}=1 \mid Z_1=k, Z_2=l, W_1=s)\ \pi_l\ \pi_k\\
&\quad=\mathbb{P}(A_{341}=1, \ldots, A_{2i-1\,2i\,1}=1 \mid W_1=s)\\
&\qquad\times \sum_l \sum_k \mathbb{P}(A_{121}=1 \mid Z_1=k, Z_2=l, W_1=s)\ \pi_l\ \pi_k\\
&\quad=\mathbb{P}(A_{341}=1, \ldots, A_{2i-1\,2i\,1}=1 \mid W_1=s)\times\ t_s\\
&\quad=(t_s)^{i-1}\,. \qquad\qquad (A.9)
\end{aligned}$$

Now, we define $v_0=1$ and for $1 \leq i \leq 2Q-1$:

$$\begin{aligned}
v_i&=\mathbb{P}(A_{121}=1, A_{341}=1, \ldots, A_{2i\,2i+1\,1}=1)\\
&=\sum_s \mathbb{P}(A_{121}=1, A_{341}=1, \ldots, A_{2i\,2i+1\,1}=1 \mid W_1=s)\ \rho_s\\
&=\sum_s (t_s)^i\ \rho_s. \qquad\qquad (A.10)
\end{aligned}$$

By Assumption $\mathcal{A}4$, $(v_i)_{i=1:(2Q-1)}$ are well defined. Hence, $v_0=1$ and $(v_i)_{i=1:(2Q-1)}$ are known and defined from the marginal $\mathbb{P}_A$. As a consequence, $(v_i)_{i=1:(2Q-1)}$ are identifiable.

Also, let $\tilde{\mathbf{M}}$ be the matrix of size $(Q+1)\times Q$ given by $\tilde{M}_{ij}=v_{i+j-2}$ for $1 \leq i \leq Q+1$ and $1 \leq j \leq Q$, and let $\tilde{\mathbf{M}}_{-i}$ denote the square matrix obtained by removing the row i from $\tilde{\mathbf{M}}$. The coefficients of $\tilde{\mathbf{M}}_{-(K+1)}$, for $1 \leq i,j \leq Q$, are:

$$\begin{aligned}
\tilde{M}_{ij}&=\sum_{s=1}^{Q} (t_s)^{i-1}\ \rho_s\ (r_s)^{j-1}\,,\ \text{and}\\
\tilde{\mathbf{M}}_{-(Q+1)}&=\mathbf{T}\ \operatorname{Diag}(\boldsymbol{\rho})\ \mathbf{T}^T\,. \qquad\qquad (A.11)
\end{aligned}$$

Proposition 6 (Relations between T, $\tilde{\mathbf{M}}$ and $\boldsymbol{\rho}$) *From Proposition 5 and Equation* (A.11)*, we can define*

$$\tilde{\mathbf{M}}_{-(Q+1)}=\mathbf{T}\ \operatorname{Diag}(\boldsymbol{\rho})\ \mathbf{T}^T\,. \qquad (A.12)$$

The correspondence of the different terms being established, we now need to prove the identifiability of $\boldsymbol{\rho}$, which means showing that $\tilde{\mathbf{M}}_{-(Q+1)}$ and $\mathbf{T}$ are identifiable.

As for the previous proof regarding the identifiability of t_s, with $\delta_s = \mathrm{Det}(\tilde{\mathbf{M}}_{-s})$, we define a polynomial function $\tilde{B}$ such as:

$$\tilde{B}(x) = \sum_{s=0}^{Q} (-1)^{Q+s}\ \delta_{s+1}\ x^s \tag{A.13}$$

This polynomial function has again two important properties summarized in the following proposition.

Proposition 7 *Let* $\deg(\tilde{B})$ *denote the degree of* $\tilde{B}$*. We have* $\deg(\tilde{B}) = Q$ *and* $\tilde{B}(t_s) = 0$*, for* $1 \leq s \leq Q$*.*

Proof The proof follow the same lines as those of Proposition 3 and Proposition 4. ■

With $(t_s)_{s=1:Q}$ being the roots of $\tilde{B}$ (proposition 7), they are functions of $(\delta_s)_{s=1:Q+1}$ which are themselves derived from $\mathbb{P}_A$. Also, $(t_s)_{s=1:Q}$ can be expressed in a unique way (up to label switching) from $\mathbb{P}_A$, thus $(t_s)_{s=1:Q}$ are identifiable. In consequence, $\mathbf{T}$ is also identifiable by definition. Finally, since $\tilde{\mathbf{M}}_{-(Q+1)}$ and $\mathbf{T}$ are identifiable and invertible, $\mathrm{Diag}(\boldsymbol{\rho}) = \mathbf{T}^{-1}\tilde{\mathbf{M}}_{-(Q+1)}(\mathbf{T}^T)^{-1}$. In conclusion, $\boldsymbol{\rho}$ is identifiable.

A.4 Identifiability of $\boldsymbol{\alpha}$

To establish the identifiability of $\boldsymbol{\alpha}$, the initial proof relies on matrix inversion. However, within our framework, tensor inversion is not as straightforward as uniqueness may not be inherently guaranteed. To overcome this issue, we will reparametrize our problem to revert to a matrix-based formulation. To do this, we shift from utilizing the reference frame of nodes (individuals) to that of edges (connections).

First, the tensor $\mathbf{A}$ is transformed into a matrix $\tilde{\mathbf{A}}$ of size $\tilde{N} \times V$, with $\tilde{N} = N(N-1)/2$ in an undirected framework. Each column corresponds to a vectorization of the upper triangular matrix of each layer of $\mathbf{A}$. Thus, the edge $\mathbf{A}_{ijv}$ will be described by $\tilde{\mathbf{A}}_{\tilde{i}v}$, with $\tilde{i}$ being the index corresponding to the edge (i, j) between nodes i and j.

Then, we can map the clustering of observations into a clustering of edges, which results in a matrix $\tilde{\mathbf{Z}}$ of size $\tilde{N} \times \tilde{K}$, with $\tilde{K} = K(K+1)/2$. Each row of the matrix corresponds to the clustering of the pair of nodes making up the edges $1 \leq \tilde{i} \leq \tilde{N}$.

Also, we denote $\tilde{\boldsymbol{\pi}}$ the proportion vector of pairs such that $\tilde{\boldsymbol{\pi}}_{\tilde{k}} = \boldsymbol{\pi}_k \boldsymbol{\pi}_l$, for $1 \leq \tilde{k} \leq \tilde{K}$, where $1 \leq k, l \leq K$ are the initial clusters corresponding to the index of the $\tilde{k}$ in the reparametrization.

Finally, $\tilde{\boldsymbol{\alpha}}$ is a $\tilde{K} \times Q$ matrix whose rows represents the clusters related to the pairs of nodes while the columns are the components. The terms of $\tilde{\boldsymbol{\alpha}}$ represent the probabilities of connection between these clusters and components.

Now, let's define a function ϕ such as:

$$\phi(\mathbf{A}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\alpha}) = (\tilde{\mathbf{A}}, \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\alpha}})\,. \tag{A.14}$$

The function is bijective for $\mathbf{A}$ and $\boldsymbol{\alpha}$ and injective for $\mathbf{Z}$ and $\boldsymbol{\pi}$; The bijective relationship involving the parameter $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ enables the establishment of identifiability. The aim is therefore to show the identifiability of $\tilde{\boldsymbol{\alpha}}$.

Remark 2 *These transformations map our problem into a LBM framework (see Figure A.1). Hence, the identifiability of $\tilde{\boldsymbol{\alpha}}$ will be developed accordingly.*

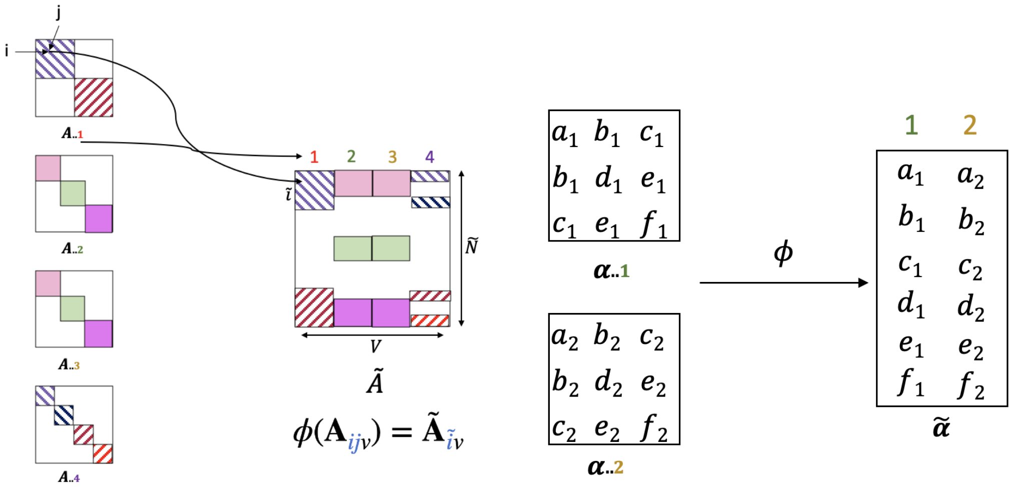


Figure A.1: Illustration of transformation on $\mathbf{A}$ and $\boldsymbol{\alpha}$.

The proof is identical to the ones of Section A.2 For any $1 \leq \tilde{k} \leq \tilde{K}$ and

$\forall i, j$, let's define:

$$\begin{aligned}
\tilde{r}_{\tilde{k}} &= \mathbb{P}(\tilde{A}_{ij} = 1 \mid \tilde{Z}_i = \tilde{k}) \qquad \text{(A.15)} \\
&= \sum_s \mathbb{P}(\tilde{A}_{ij} = 1 \mid \tilde{Z}_i = \tilde{k}, W_v = s)\ \rho_s \\
&= \sum_s \tilde{\alpha}_{\tilde{k}s} \rho_s \\
&= (\tilde{\boldsymbol{\alpha}} \boldsymbol{\rho})_{\tilde{k}}\,. \qquad \text{(A.16)}
\end{aligned}$$

Proposition 8 (Invertibility of $\tilde{\mathbf{R}}$) *Let* $\tilde{\mathbf{R}}$ *denote a Vandermonde matrix of size* $\tilde{K} \times \tilde{K}$ *such as* $\tilde{R}_{i\tilde{k}} = (\tilde{r}_{\tilde{k}})^{i-1}$*, for* $1 \leq i \leq \tilde{K}$ *and* $1 \leq \tilde{k} \leq \tilde{K}$*.* $\tilde{\mathbf{R}}$ *is invertible, since the coordinates of* r *are all different according to Assumption* $\mathcal{A}5$*.*

The rest of the proof is identical to the one of Section A.2 so that it can be show that $\tilde{\mathbf{R}}$ is identifiable and so $\tilde{\boldsymbol{\pi}}_{\tilde{k}}$.

We now focus on the the identifiability of $\tilde{\boldsymbol{\alpha}}$. Let $\mathbf{U}$ be a matrix of size $\tilde{K} \times Q$ such that the (i, j) entry of is the joint probability of having i connections in the first row and $j-1$ connections in the first column:

$$\begin{aligned}
\mathbf{U}_{ij} &= \mathbb{P}(\tilde{\mathbf{A}}_{11} = 1, \tilde{\mathbf{A}}_{12} = 1, \ldots, \tilde{\mathbf{A}}_{1i} = 1, \tilde{\mathbf{A}}_{21} = 1, \ldots, \tilde{\mathbf{A}}_{j1} = 1) \\
&= \sum_{\tilde{k}} \sum_q \tilde{\pi}_{\tilde{k}}\, \rho_s\, \mathbb{P}(\tilde{\mathbf{A}}_{11} = 1, \tilde{\mathbf{A}}_{12} = 1, \ldots, \tilde{\mathbf{A}}_{1i} = 1, \tilde{\mathbf{A}}_{21} = 1, \ldots, \tilde{\mathbf{A}}_{j1} = 1 \mid \tilde{\mathbf{Z}}_1 = \tilde{k}, \mathbf{W}_1 = s) \\
&= \sum_{\tilde{k}} \sum_q \tilde{\pi}_{\tilde{k}}\, \rho_s\, \tilde{\alpha}_{\tilde{k}s}\, \mathbb{P}(\tilde{\mathbf{A}}_{12} = 1, \ldots, \tilde{\mathbf{A}}_{1i} = 1, \tilde{\mathbf{A}}_{21} = 1, \ldots, \tilde{\mathbf{A}}_{j1} = 1 \mid \tilde{\mathbf{Z}}_1 = \tilde{k}, \mathbf{W}_1 = s) \\
&= \sum_{\tilde{k}} \sum_q \tilde{\pi}_{\tilde{k}}\, \rho_s\, \tilde{\alpha}_{\tilde{k}s}\, \tilde{r}_{\tilde{k}}^{i-1} t_s^{j-1}. \qquad \text{(A.17)}
\end{aligned}$$

Proposition 9 (Relations between $\tilde{\mathbf{R}}$, $\mathbf{T}$, $\mathbf{U}$, $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\pi}}$ and $\boldsymbol{\rho}$) *From Proposition 8 and Equation* (A.17)*, we can define* $\mathbf{U} = \tilde{\mathbf{R}}\, \mathrm{Diag}(\tilde{\boldsymbol{\pi}})\, \tilde{\boldsymbol{\alpha}}\ \mathrm{Diag}(\boldsymbol{\rho})\, \mathbf{T}^T$*, with* $\mathbf{U}$*,* $\tilde{\mathbf{R}}$*,* $\mathrm{Diag}(\tilde{\boldsymbol{\pi}})$*,* $\mathrm{Diag}(\boldsymbol{\rho})$ *and* $\mathbf{T}$ *being invertible. Therefore,*

$$\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{R}}^{-1}\, \mathrm{Diag}(\tilde{\boldsymbol{\pi}})^{-1}\, \mathbf{U}\ \mathrm{Diag}(\rho)^{-1}\, (\mathbf{T}^T)^{-1}\,. \tag{A.18}$$

In addition to Proposition 9, $\mathbf{U}$ being defined from $\mathbb{P}_{\tilde{\mathbf{A}}}$, all its coefficients are identifiable. As a consequence, $\tilde{\boldsymbol{\alpha}}$ is identifiable. In conclusion, $\boldsymbol{\alpha} = \phi^{-1}(\tilde{\boldsymbol{\alpha}})$ is identifiable.

B Details of VEM algorithm for mimiSBM - frequentist framework

B.1 Variational parameters of clustering τ_{ik}

The optimal estimation of τ_{ik}, indicated at Equation 2.9, is given by

$$\tau_{ik} = \frac{\exp(T_{ik})}{\sum_{k'=1} exp(T_{ik'})}, \tag{B.1}$$

with

$$T_{ik} = \sum_{j \neq i}^{N} \sum_{l=1}^{K} \sum_{v=1}^{V} \sum_{s=1}^{Q} \tau_{jl} \nu_{vs} \left[A_{ijv} \log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right) \log\left(1 - \alpha_{kls}\right) \right] + \log\left(\pi_k\right) - 1. \tag{B.2}$$

We see that we have a fixed-point problem in the optimization of $\boldsymbol{\tau}$. To overcome this problem, we iterate until convergence.

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$\begin{aligned} J_{\boldsymbol{\tau}}(.) = & \sum_{\substack{i=1, \\ i<j}}^{N} \sum_{\substack{k=1, \\ l=1}}^{K} \sum_{v=1}^{V} \sum_{s=1}^{Q} \tau_{ik} \tau_{jl} \nu_{vs} \left[A_{ijv} \log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right) \log\left(1 - \alpha_{kls}\right) \right] \\ & + \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log\left(\pi_k\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log\left(\tau_{ik}\right) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} \tau_{ik}\right), \end{aligned} \tag{B.3}$$

where $(\lambda_i)_{i=1:N}$ are Lagrange parameters. Now, the aim is to get the gradient in τ_{ik} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_\tau}{\partial \tau_{ik}}(.) = \sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\nu_{vs}\left[A_{ijv}\log\left(\alpha_{kls}\right)+\left(1-A_{ijv}\right)\log\left(1-\alpha_{kls}\right)\right] + \log\left(\pi_k\right) - \log\left(\tau_{ik}\right) - 1 - \lambda_i \tag{B.4}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\tau_{ik}\right) = \underbrace{\sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\nu_{vs}\left[A_{ijv}\log\left(\alpha_{kls}\right)+\left(1-A_{ijv}\right)\log\left(1-\alpha_{kls}\right)\right] + \log\left(\pi_k\right) - 1}_{T_{ik}} - \lambda_i \tag{B.5}$$

Moreover, by constraints on τ_{ik},

$$\sum_{k'}\tau_{ik} = 1 \iff \sum_{k'}\exp\left(T_{ik'}\right)\exp\left(-\lambda_i\right) = 1 \iff \exp\left(-\lambda_i\right) = \frac{1}{\sum_{k'}\exp\left(T_{ik'}\right)} \tag{B.6}$$

Finally,

$$\tau_{ik} = \frac{\exp\left(T_{ik}\right)}{\sum_{k'}\exp\left(T_{ik'}\right)} \tag{B.7}$$

■

B.2 Variational parameters of component membership ν_{vs}

The optimal estimation of ν_{vs}, indicated at Equation 2.11, is given by

$$\nu_{vs} = \frac{\exp\left(R_{vs}\right)}{\sum_{s'=1} exp\left(R_{vs'}\right)}, \tag{B.8}$$

with

$$R_{vs} = \sum_{\substack{i=1,\\ i<j}}^{N}\sum_{\substack{k=1,\\ l=1}}^{K}\tau_{ik}\tau_{jl}\left[A_{ijv}\log\left(\alpha_{kls}\right)+\left(1-A_{ijv}\right)\log\left(1-\alpha_{kls}\right)\right] + \log\left(\rho_s\right) - 1. \tag{B.9}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$J_{\boldsymbol{\nu}}(.) = \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{\substack{k=1,\\ l=1}}^{K} \sum_{v=1}^{V} \sum_{s=1}^{Q} \tau_{ik}\tau_{jl}\nu_{vs}\left[A_{ijv}\log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right)\log\left(1 - \alpha_{kls}\right)\right] + \sum_{v=1}^{V}\sum_{s=1}^{Q}\nu_{vs}\log\left(\rho_s\right) - \sum_{v=1}^{V}\sum_{s=1}^{Q}\nu_{vs}\log\left(\nu_{vs}\right) + \sum_{v=1}^{V}\lambda_v\left(1 - \sum_{s=1}^{Q}\nu_{vs}\right), \tag{B.10}$$

where $(\lambda_v)_{v=1:V}$ are Lagrange parameters. Now, the aim is to get the gradient in ν_{vs} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_{\boldsymbol{\nu}}}{\partial \nu_{vs}}(.) = \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{\substack{k=1,\\ l=1}}^{K} \tau_{ik}\tau_{jl}\left[A_{ijv}\log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right)\log\left(1 - \alpha_{kls}\right)\right] + \log\left(\rho_s\right) - \log\left(\nu_{vs}\right) - 1 - \lambda_v \tag{B.11}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\nu_{vs}\right) = \underbrace{\sum_{\substack{i=1,\\ i<j}}^{N} \sum_{\substack{k=1,\\ l=1}}^{K} \tau_{ik}\tau_{jl}\left[A_{ijv}\log\left(\alpha_{kls}\right) + \left(1 - A_{ijv}\right)\log\left(1 - \alpha_{kls}\right)\right] + \log\left(\rho_s\right) - 1}_{R_{vs}} - \lambda_v \tag{B.12}$$

Moreover, by constraints on ν_{vs},

$$\sum_{s'}\nu_{vs} = 1 \iff \sum_{s'}\exp\left(R_{vs'}\right)\exp\left(-\lambda_v\right) = 1 \iff \exp\left(-\lambda_v\right) = \frac{1}{\sum_{s'}\exp\left(R_{vs'}\right)} \tag{B.13}$$

Finally,

$$\nu_{vs} = \frac{\exp\left(R_{vs}\right)}{\sum_{s'}\exp\left(R_{vs'}\right)} \tag{B.14}$$

■

B.3 Optimization of $\boldsymbol{\pi}$

The optimal estimation of π_k, indicated at Equation 2.14, is given by

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{ik}}{N}. \tag{B.15}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$J_{\boldsymbol{\pi}}(.) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log(\pi_k) + \lambda_{\boldsymbol{\pi}} \left(1 - \sum_{k=1}^{K} \pi_k\right), \tag{B.16}$$

where $\lambda_{\boldsymbol{\pi}}$ is Lagrange parameter.

$$\frac{\partial J_{\boldsymbol{\pi}}}{\partial \pi_k}(.) = \frac{\sum_{i=1}^{N} \tau_{ik}}{\pi_k} - \lambda_{\boldsymbol{\pi}}. \tag{B.17}$$

Moreover, by constraints on $\boldsymbol{\pi}$,

$$\sum_k \pi_k = 1 \iff \sum_k \frac{\sum_{i=1}^{N} \tau_{ik}}{\lambda_{\boldsymbol{\pi}}} = 1 \iff \lambda_{\boldsymbol{\pi}} = \sum_k \sum_{i=1}^{N} \tau_{ik} \iff \lambda_{\boldsymbol{\pi}} = N. \tag{B.18}$$

Finally,

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{ik}}{N}. \tag{B.19}$$

■

B.4 Optimization of $\boldsymbol{\rho}$

The optimal estimation of ρ_s, indicated at Equation 2.14, is given by

$$\rho_s = \frac{\sum_{v=1}^{V} \nu_{vs}}{V}. \tag{B.20}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$J_{\boldsymbol{\rho}}(.) = \sum_{v=1}^{V}\sum_{s=1}^{Q} \nu_{vs} \log(\rho_s) + \lambda_{\boldsymbol{\rho}} \left(1 - \sum_{s=1}^{Q} \rho_s\right), \tag{B.21}$$

where $\lambda_{\boldsymbol{\rho}}$ is Lagrange parameter.

$$\frac{\partial J_{\boldsymbol{\rho}}}{\partial \rho_s}(.) = \frac{\sum_{v=1}^{V} \nu_{vs}}{\rho_s} - \lambda_{\boldsymbol{\rho}}. \tag{B.22}$$

Moreover, by constraints on $\boldsymbol{\rho}$,

$$\sum_s \rho_s = 1 \Longleftrightarrow \sum_s \frac{\sum_{v=1}^{V} \nu_{vs}}{\lambda_{\boldsymbol{\rho}}} = 1 \Longleftrightarrow \lambda_{\boldsymbol{\rho}} = \sum_s \sum_{v=1}^{V} \nu_{vs} \Longleftrightarrow \lambda_{\boldsymbol{\rho}} = V. \tag{B.23}$$

Finally,

$$\rho_s = \frac{\sum_{v=1}^{V} \nu_{vs}}{V}. \tag{B.24}$$

■

B.5 Optimization of $\boldsymbol{\alpha}$

The optimal estimation of α_{kls}, indicated at Equations 2.15 and 2.16 is given by, for $k \neq l$

$$\alpha_{kls} = \frac{\sum_{i=1}^{N} \sum_{j \neq i} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}}{\sum_{i=1}^{N} \sum_{j \neq i} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}}. \tag{B.25}$$

and, for $k = l$:

$$\alpha_{kks} = \frac{\sum_{i=1}^{N} \sum_{j > i} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}}{\sum_{i=1}^{N} \sum_{j > i} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}}. \tag{B.26}$$

Proof For $k \neq l$:

$$\frac{\partial \mathcal{L}}{\partial \alpha_{kls}}(.) = \sum_{\substack{i=1, \\ i \neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs} \left[\frac{A_{ijv}}{\alpha_{kls}} - \frac{1 - A_{ijv}}{1 - \alpha_{kls}}\right]. \tag{B.27}$$

We want to solve $\frac{\partial \mathcal{L}}{\partial \alpha_{kls}}(.) = 0$, so

$$\begin{aligned}
\sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs} \left[\frac{A_{ijv}}{\alpha_{kls}} - \frac{1-A_{ijv}}{1-\alpha_{kls}}\right] &= 0 \\
\sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs} \left(1-\alpha_{kls}\right) A_{ijv} &= \sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}\alpha_{kls} \left(1-A_{ijv}\right) \\
\sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} - \alpha_{kls} \sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} &= \sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs} - \alpha_{kls} \sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} \\
\sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} &= \sum_{\substack{i=1,\\ i\neq j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs} \qquad \text{(B.28)}
\end{aligned}$$

Finally,

$$\alpha_{kls} = \frac{\sum_{i=1}^{N} \sum_{j\neq i} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}}{\sum_{i=1}^{N} \sum_{j\neq i} \sum_{v=1}^{V} \tau_{ik}\tau_{jl}\nu_{vs}}. \qquad \text{(B.29)}$$

For $k = l$:

$$\frac{\partial \mathcal{L}}{\partial \alpha_{kks}}(.) = \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs} \left[\frac{A_{ijv}}{\alpha_{kks}} - \frac{1-A_{ijv}}{1-\alpha_{kks}}\right]. \qquad \text{(B.30)}$$

We want to solve $\frac{\partial \mathcal{L}}{\partial \alpha_{kks}}(.) = 0$, so

$$\begin{aligned}
\sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs} \left[\frac{A_{ijv}}{\alpha_{kks}} - \frac{1-A_{ijv}}{1-\alpha_{kks}}\right] &= 0 \\
\sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs} \left(1-\alpha_{kks}\right) A_{ijv} &= \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}\alpha_{kks} \left(1-A_{ijv}\right) \\
\sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} - \alpha_{kks} \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} &= \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs} - \alpha_{kks} \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} \\
\sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} &= \sum_{\substack{i=1,\\ i<j}}^{N} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs} \qquad \text{(B.31)}
\end{aligned}$$

Finally,

$$\alpha_{kks} = \frac{\sum_{i=1}^{N} \sum_{j<i} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}}{\sum_{i=1}^{N} \sum_{j<i} \sum_{v=1}^{V} \tau_{ik}\tau_{jk}\nu_{vs}}. \qquad \text{(B.32)}$$

$\blacksquare$

C Details of VBEM algorithm for mimiSBM - Bayesian framework

C.1 Variational parameters of clustering τ_{ik}

The optimal approximation for $q(\mathbf{Z}_i)$ is

$$q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i; (\tau_{i1}, \dots, \tau_{iK})), \tag{C.1}$$

where τ_{ik} is the probability of node i to belong to class k. It satisfies the relation indicated at Equation 2.9 :

$$\tau_{ik} \propto e^{\psi(\beta_k) - \psi(\sum_{k'} \beta_{k'})} \prod_{j \neq i}^{N} \prod_{l=1}^{K} \prod_{v=1}^{V} \prod_{s=1}^{Q} e^{\tau_{jl}\, \nu_{vs} \left[A_{ijv} \left(\psi(\eta_{kls}) - \psi(\xi_{kls}) \right) + \psi(\xi_{kls}) - \psi(\eta_{kls} + \xi_{kls}) \right]}, \tag{C.2}$$

where ψ is digamma function. Distribution $q(\mathbf{Z})$ is optimized with a fixed point algorithm.

Proof According to the model, the optimal distribution $q(\mathbf{Z}_i)$ is given by

$$\begin{aligned}
&\log q(\mathbf{Z}_i)\\
&= \mathbb{E}_{\mathbf{Z}^{\setminus i},\alpha,\pi,W,\rho}\left[\log \mathbb{P}(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})\right]\\
&\propto \mathbb{E}_{\mathbf{Z}^{\setminus i},\boldsymbol{\alpha},\mathbf{W}}[\log \mathbb{P}(\mathbf{A}|\mathbf{Z},\mathbf{W},\boldsymbol{\alpha})] + \mathbb{E}_{\mathbf{Z}^{\setminus i},\boldsymbol{\pi}}[\log \mathbb{P}(\mathbf{Z}|\pi)]\\
&\propto \mathbb{E}_{\mathbf{Z}^{\setminus i},\boldsymbol{\alpha},\mathbf{W}}\left[\sum_{i'=1,j>i'}^{N}\sum_{k,l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\mathbb{1}_{\mathbf{Z}_{i'},\mathbf{Z}_j,\mathbf{W}_v}\left(\log \mathbb{P}(A_{i'jv}|Z_{i'k},Z_{jl},W_{vs},\boldsymbol{\alpha})\right)\right]\\
&\quad + \mathbb{E}_{\mathbf{Z}^{\setminus i},\boldsymbol{\pi}}\left[\sum_{i'=1}^{N}\sum_{k}^{K}\log \mathbb{P}(Z_{i'}=k|\boldsymbol{\pi})\right]\\
&\propto \sum_k \mathbb{1}_{Z_i=k}\Big\{\mathbb{E}_{\boldsymbol{\pi}}[\log(\pi_k)] \; + \sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\,\nu_{vs}\;\mathbb{E}_{\boldsymbol{\alpha}}\Big[A_{ijv}\log(\alpha_{kls}) +\\
&\quad (1-A_{ijv})\log(1-\alpha_{kls})\Big]\Big\}.
\end{aligned} \tag{C.3}$$

Remember that :

- $\boldsymbol{\pi} \sim Dir(\boldsymbol{\pi};\boldsymbol{\beta})$, so $\pi_k \sim Beta(\pi_k;\beta_k,\sum_{k'}\beta_{k'}-\beta_k)$;
- $\mathbb{E}_{\boldsymbol{\pi}}[\log(\pi_k)] = \psi(\beta_k) - \psi(\sum_{k'}\beta_{k'})$;
- $q(\alpha_{kls}) = Beta(\alpha_{kls};\eta_{kls},\xi_{kls})$;
- $\mathbb{E}_{\alpha_{kls}}[\log(\alpha_{kls})] = \psi(\eta_{kls}) - \psi(\xi_{kls}+\eta_{kls})$;
- $\mathbb{E}_{\alpha_{kls}}[\log(1-\alpha_{kls})] = \psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})$.

In consequence,

$$\begin{aligned}
&\log q(Z_i)\\
&\propto \sum_k \mathbb{1}_{Z_i=k}\Big\{\psi\left(\beta_k\right) - \psi\left(\sum_{k'}\beta_{k'}\right) + \sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\,\nu_{vs}\Big[A_{ijv}\Big(\,(\psi(\eta_{kls})\\
&\quad -\psi\left(\xi_{kls}+\eta_{kls}\right)) - \left(\psi\left(\xi_{kls}\right) - \psi\left(\eta_{kls}+\xi_{kls}\right)\right)\Big) + \psi\left(\xi_{kls}\right)\\
&\quad -\psi\left(\eta_{kls}+\xi_{kls}\right)\Big]\Big\}\\
&= \sum_k \mathbb{1}_{Z_i=k}\Big\{\psi(\beta_k) - \psi(\sum_{k'}\beta_{k'}) \; + \sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\,\nu_{vs}\Big[A_{ijv}\Big(\psi(\eta_{kls}) - \psi(\xi_{kls})\Big)\\
&\quad + \psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\Big]\Big\}.
\end{aligned} \tag{C.4}$$

We can therefore deduce that, by applying the exponential :

$$\begin{aligned} & q(Z_i = k) \\ & \propto e^{\psi(\beta_k)-\psi(\sum_{k'}\beta_{k'})+\sum_{j\neq i}^{N}\sum_{l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\tau_{jl}\,\nu_{vs}\left[A_{ijv}\left(\psi(\eta_{kls})-\psi(\xi_{kls}\right)+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})\right]} \qquad (C.5). \\ & = e^{\psi(\beta_k)-\psi(\sum_{k'}\beta_{k'})}\prod_{j\neq i}^{N}\prod_{l=1}^{K}\prod_{v=1}^{V}\prod_{s=1}^{Q} e^{\tau_{jl}\,\nu_{vs}\left[A_{ijv}\left(\psi(\eta_{kls})-\psi(\xi_{kls})\right)+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})\right]} \end{aligned}$$

Therefore,

$$\tau_{ik} \propto e^{\psi(\beta_k)-\psi(\sum_{k'}\beta_{k'})}\prod_{j\neq i}^{N}\prod_{l=1}^{K}\prod_{v=1}^{V}\prod_{s=1}^{Q} e^{\tau_{jl}\,\nu_{vs}\left[A_{ijv}\left(\psi(\eta_{kls})-\psi(\xi_{kls})\right)+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})\right]}. \tag{C.6}$$

So,

$$q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i; (\tau_{i1}, \ldots, \tau_{iK})). \tag{C.7}$$

■

C.2 Variational parameters of component membership ν_{vs}

The optimal approximation for $q(\mathbf{W}_v)$ is

$$q(W_v) = \mathcal{M}(W_v; (\nu_{v1}, \ldots, \nu_{vQ})), \tag{C.8}$$

with

$$\begin{aligned} \nu_{vs} \propto e^{\psi(\theta_s)-\psi(\sum_{s'}\theta_{s'})}\prod_{i\neq j}^{N}\prod_{k\neq l}^{K} e^{\tau_{ik}\,\tau_{jl}\left[A_{ijv}\left(\psi(\eta_{kls})-\psi(\xi_{kls}\right)+\psi(\xi_{kls})-\psi(\eta_{kls}+\xi_{kls})\right]}\times \\ \prod_{k}^{K}\prod_{i<j}^{N} e^{\tau_{ik}\,\tau_{jk}\left[A_{ijv}\left(\psi(\eta_{kks})-\psi(\xi_{kks}\right)+\psi(\xi_{kks})-\psi(\eta_{kks}+\xi_{kks})\right]}. \end{aligned} \tag{C.9}$$

ν_{vs} is the probability of layer v to belong to component s, as indicated at Equation 2.26.

Proof As previously mentioned, in accordance with the principles of variational Bayes, the optimal probability distribution can be expressed as follows:

$$\begin{aligned}
&\log q(\mathbf{W}_v)\\
&= \mathbb{E}_{\mathbf{W}^{\setminus v},\boldsymbol{\alpha},\boldsymbol{\pi},\mathbf{Z},\boldsymbol{\rho}}[\log \mathbb{P}(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})]\\
&\propto \mathbb{E}_{\mathbf{W}^{\setminus v},\boldsymbol{\alpha},\mathbf{Z}}[\log \mathbb{P}(\mathbf{A}|\mathbf{Z},\mathbf{W},\boldsymbol{\alpha})] + \mathbb{E}_{\mathbf{W}^{\setminus v},\boldsymbol{\rho}}\,[\ln \mathbb{P}(\mathbf{W}|\rho)]\\
&\propto \mathbb{E}_{W^{\setminus v},\boldsymbol{\alpha},\mathbf{Z}}\left[\sum_{i=1,j>i}^{N}\sum_{k,l=1}^{K}\sum_{v=1}^{V}\sum_{s=1}^{Q}\mathbb{1}_{\mathbf{Z}_i,\mathbf{Z}_j,W_v}\Big(\log \mathbb{P}(A_{ijv}|Z_{ik},Z_{jl},W_{vs},\boldsymbol{\alpha})\Big)\right]\\
&\quad + \mathbb{E}_{\mathbf{W}^{\setminus v},\boldsymbol{\rho}}\left[\sum_{v=1}^{V}\sum_{s}^{Q}\log \mathbb{P}(W_v = s|\boldsymbol{\rho})\right]\\
&\propto \sum_s \mathbb{1}_{W_{vs}}\Big\{\sum_{k\neq l}^{K}\sum_{i=1,j\neq i}^{N}\tau_{ik}\,\tau_{jl}\;\mathbb{E}_{\boldsymbol{\alpha}}\Big[A_{ijv}\log(\alpha_{kls}) + (1-A_{ijv})\log(1-\alpha_{kls})\Big]\\
&\quad + \sum_{k}^{K}\sum_{i=1,i<j}^{N}\tau_{ik}\,\tau_{jk}\;\mathbb{E}_{\boldsymbol{\alpha}}\Big[A_{ijv}\log(\alpha_{kks}) + (1-A_{ijv})\log(1-\alpha_{kks})\Big]\\
&\quad + \mathbb{E}_{\rho}[\log(\rho_s)]\Big\}.
\end{aligned} \tag{C.10}$$

Reminder :

- $\rho \sim Dir(\pi;\theta)$, so $\rho_s \sim Beta(\rho_s;\theta_s,\sum_{s'}\theta_{s'} - \theta_s)$;
- $\mathbb{E}_\rho[\log(\rho_s)] = \psi(\theta_s) - \psi(\sum_{s'}\theta_{s'})$.

Hence,

$$\begin{aligned}
&\log q(W_v)\\
&\propto \sum_s \mathbb{1}_{W_v=q}\Big\{\psi(\theta_s) - \psi(\sum_{s'}\theta_{s'})\; + \sum_{i=1,j>i}^{N}\sum_{k,l=1}^{K}\tau_{ik}\tau_{jl}\Big[A_{ijv}\Big((\psi(\eta_{kls}) - \psi(\xi_{kls}+\eta_{kls}))\\
&\quad - (\psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls}))\Big) + \psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\Big]\Big\} \qquad (C.11)\\
&= \sum_s \mathbb{1}_{W_v=q}\Big\{\psi(\theta_s) - \psi(\sum_{s'}\theta_{s'})\; + \sum_{i=1,j>i}^{N}\sum_{k,l=1}^{K}\tau_{ik}\,\tau_{jl}\Big[A_{ijv}\Big(\psi(\eta_{kls}) - \psi(\xi_{kls})\Big)\\
&\quad + \psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\Big]\Big\}.
\end{aligned}$$

Consequently,

$$\begin{aligned} & q(W_v = s) \\ & \propto e^{\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'}) \ + \sum_{i=1, j>i}^{N} \sum_{k,l=1}^{K} \tau_{ik}\, \tau_{jl} \left[A_{ijv} \left(\psi(\eta_{kls}) - \psi(\xi_{kls}) \right) + \psi(\xi_{kls}) - \psi(\eta_{kls} + \xi_{kls}) \right]} \\ & = e^{\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})} \prod_{k \neq l}^{K} \prod_{i=1, j \neq i}^{N} e^{\tau_{ik}\, \tau_{jl} \left[A_{ijv} \left(\psi(\eta_{kls}) - \psi(\xi_{kls}) \right) + \psi(\xi_{kls}) - \psi(\eta_{kls} + \xi_{kls}) \right]} \\ & \quad \prod_{k}^{K} \prod_{i<j}^{N} e^{\tau_{ik}\, \tau_{jk} \left[A_{ijv} \left(\psi(\eta_{kks}) - \psi(\xi_{kks}) \right) + \psi(\xi_{kks}) - \psi(\eta_{kks} + \xi_{kks}) \right]}. \end{aligned} \tag{C.12}$$

So,

$$\begin{aligned} \nu_{vs} \propto e^{\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})} \prod_{i \neq j}^{N} \prod_{k \neq l}^{K} e^{\tau_{ik}\, \tau_{jl} \left[A_{ijv} \left(\psi(\eta_{kls}) - \psi(\xi_{kls}) \right) + \psi(\xi_{kls}) - \psi(\eta_{kls} + \xi_{kls}) \right]} \times \\ \prod_{k}^{K} \prod_{i<j}^{N} e^{\tau_{ik}\, \tau_{jk} \left[A_{ijv} \left(\psi(\eta_{kks}) - \psi(\xi_{kks} \right) + \psi(\xi_{kks})) - \psi(\eta_{kks} + \xi_{kks}) \right]} \end{aligned} , \tag{C.13}$$

and

$$q(W_v) = \mathcal{M}(W_v; (\nu_{v1}, \dots, \nu_{vQ})). \tag{C.14}$$

■

C.3 Optimization of $q(\boldsymbol{\pi})$ (β_k)

Due to the selection of prior distributions, the distribution $q(\boldsymbol{\pi})$ remains within the same family of distributions as the prior distribution $\mathbb{P}(\boldsymbol{\pi})$.

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}), \tag{C.15}$$

with

$$\beta_k = \beta_k^0 + \sum_{i}^{N} \tau_{ik}, \tag{C.16}$$

as indicated at Equation 2.28.

Proof The optimal probability distribution can be formulated in the following manner:

$$\begin{aligned}\log q(\boldsymbol{\pi}) &\propto \mathbb{E}_{\mathbf{W},\boldsymbol{\alpha},\mathbf{Z},\boldsymbol{\rho}}[\log \mathbb{P}(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})] \\ &\propto \mathbb{E}_{\mathbf{Z}}[\log \mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi})] + \log p(\boldsymbol{\pi}) \\ &\propto \sum_{i}^{N}\sum_{k}^{K} \tau_{ik} \log \pi_k + \sum_{k=1}^{K} \left(\beta_k^0 - 1\right) \log \pi_k \ \cdot \\ &\propto \sum_{k}^{K} \left(\beta_k^0 + (\sum_{i}^{N} \tau_{ik}) - 1\right) \log \pi_k\end{aligned} \tag{C.17}$$

After exponentiation and normalization, we obtain:

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}), \tag{C.18}$$

with

$$\beta_k = \beta_k^0 + \sum_{i}^{N} \tau_{ik}. \tag{C.19}$$

■

C.4 Optimization of $q(\boldsymbol{\rho})$ (θ_s)

As previously mentioned, the selection of prior distributions enables us to remain within the same family of distributions.

$$q(\boldsymbol{\rho}) = \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}), \tag{C.20}$$

with

$$\theta_s = \theta_s^0 + \sum_{v=1}^{V} \nu_{vs}, \tag{C.21}$$

as indicated at Equation 2.30.

Proof According to variational Bayes, the optimal probability distribution can be expressed as follows:

$$\begin{aligned}\log q(\boldsymbol{\rho}) &\propto \mathbb{E}_{\mathbf{W},\boldsymbol{\alpha},Z}[\log \mathbb{P}(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})] \\ &\propto \mathrm{E}_{\mathbf{W}}[\log p(\mathbf{W} \mid \boldsymbol{\rho})] + \log \mathbb{P}(\boldsymbol{\rho}) \\ &\propto \sum_{v}^{V}\sum_{s}^{Q} \nu_{vs} \log \rho_s + \sum_{q=1}^{Q} \left(\theta_s^0 - 1\right) \log \rho_s \cdot \\ &\propto \sum_{s}^{Q} \left(\theta_s^0 + (\sum_{v}^{V} \nu_{vs}) - 1\right) \log \rho_s\end{aligned} \tag{C.22}$$

After exponentiation and normalization, we have

$$q(\boldsymbol{\rho}) = \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}), \tag{C.23}$$

with

$$\theta_s = \theta_s^0 + \sum_{v=1}^{V} \nu_{vs}. \tag{C.24}$$

■

C.5 Optimization of $q(\boldsymbol{\alpha})$ (η_{kls} and ξ_{kls})

Once again, the distribution form of the prior distribution $\mathbb{P}(\boldsymbol{\alpha})$ is preserved through the variational optimization process.

$$q(\alpha_{kls}) = \mathrm{Beta}(\alpha_{kls}; \eta_{kls}, \xi_{kls}). \tag{C.25}$$

When $k \neq l$, parameters η_{kls} and ξ_{kls} are, as indicated at Equation 2.32, given by:

$$\begin{aligned} \eta_{kls} &= \eta_{kls}^0 + \sum_{i \neq j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} \\ \xi_{kls} &= \xi_{kls}^0 + \sum_{i \neq j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}\left(1 - A_{ijv}\right) \end{aligned} . \tag{C.26}$$

Otherwise, when k equals l, the parameters η_{kks} and ξ_{kks} are, as indicated at Equation 2.33, determined by:

$$\begin{aligned} \eta_{kks} &= \eta_{kks}^0 + \sum_{i<j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} \\ \xi_{kks} &= \xi_{kks}^0 + \sum_{i<j}^{N} \sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}\left(1 - A_{ijv}\right) \end{aligned} . \tag{C.27}$$

Proof In accordance with the principles of variational Bayes, the optimal probability distribution can be formulated as follows:

$$\begin{aligned}
\log q(\boldsymbol{\alpha}) &\propto \mathrm{E}_{\mathbf{Z},\mathbf{W}}[\log \mathbb{P}(\mathbf{A},\mathbf{Z},\boldsymbol{\alpha},\mathbf{W})] \\
&\propto \mathrm{E}_{\mathbf{Z},\mathbf{W}}[\log p(\mathbf{A} \mid \mathbf{Z},\mathbf{W},\boldsymbol{\alpha})] + \log \mathbb{P}(\boldsymbol{\alpha}) \\
&= \sum_{i<j}^{N}\sum_{k,l}^{K}\sum_{v}^{V}\sum_{s}^{Q} \tau_{ik}\tau_{jl}\nu_{vs}\left(A_{ijv}\log(\alpha_{kls}) + (1-A_{ijv})\log\left(1-\alpha_{kls}\right)\right) \\
&\quad + \sum_{k\leq l}^{K}\sum_{s}^{Q}\left(\left(\eta^0_{kls}-1\right)\log(\alpha_{kls}) + \left(\xi^0_{kls}-1\right)\log\left(1-\alpha_{kls}\right)\right) \\
&= \sum_{k<l}^{K}\sum_{i\neq j}^{N}\sum_{v}^{V}\sum_{s}^{Q} \tau_{ik}\tau_{jl}\nu_{vs}\left(A_{ijv}\log(\alpha_{kls}) + (1-A_{ijv})\log\left(1-\alpha_{kls}\right)\right) \\
&\quad + \sum_{k=1}^{K}\sum_{i<j}^{N}\sum_{v}^{V}\sum_{s=1}^{Q} \tau_{ik}\tau_{jk}\nu_{vs}\left(A_{ijv}\log(\alpha_{kks}) + (1-A_{ijv})\log\left(1-\alpha_{kks}\right)\right) \\
&\quad + \sum_{k\leq l}^{K}\sum_{s}^{Q}\left(\left(\eta^0_{kls}-1\right)\log(\alpha_{kls}) + \left(\xi^0_{kls}-1\right)\log\left(1-\alpha_{kls}\right)\right)
\end{aligned} . \tag{C.28}$$

By factoring the terms :

$$\begin{aligned}
\log q(\boldsymbol{\alpha}) \propto \sum_{k<l}^{K}\sum_{s}^{Q} &\left(\eta^0_{kls} - 1 + \sum_{i\neq j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}\right)\log(\alpha_{kls}) + \\
&\left(\xi^0_{kls} - 1 + \sum_{i\neq j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jl}\nu_{vs}\left(1-A_{ijv}\right)\right)\log\left(1-\alpha_{kls}\right) \\
+ \sum_{k=1}^{K}\sum_{s}^{Q} &\left(\eta^0_{kks} - 1 + \sum_{i<j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}\right)\log\alpha_{kks} + \\
&\left(\xi^0_{kks} - 1 + \sum_{i<j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jk}\nu_{vs}\left(1-A_{ijv}\right)\right)\log\left(1-\alpha_{kks}\right)
\end{aligned} . \tag{C.29}$$

Therefore,

$$q(\alpha_{kls}) = \mathrm{Beta}(\alpha_{kls};\eta_{kls},\xi_{kls}), \tag{C.30}$$

if $k \neq l$,

$$\begin{aligned}
\eta_{kls} &= \eta^0_{kls} + \sum_{i\neq j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jl}\nu_{vs}A_{ijv} \\
\xi_{kls} &= \xi^0_{kls} + \sum_{i\neq j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jl}\nu_{vs}\left(1-A_{ijv}\right)
\end{aligned} ; \tag{C.31}$$

otherwise,

$$\begin{aligned}\eta_{kks} &= \eta^0_{kks} + \sum_{i<j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jk}\nu_{vs}A_{ijv} \\ \xi_{kks} &= \xi^0_{kks} + \sum_{i<j}^{N}\sum_{v}^{V}\tau_{ik}\tau_{jk}\nu_{vs}\left(1 - A_{ijv}\right)\end{aligned}. \tag{C.32}$$

■

D Evidence Lower Bound

The lower bound assumes a simplified form after the variational Bayes M-step. It relies solely on the posterior probabilities τ_{ik} and ν_{vs} and the normalizing constants of the Dirichlet and Beta distributions.

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

Proof The lower bound can be expressed as:

$$\begin{aligned}\mathcal{L}\left(q(.)\right) &= \sum_{\mathbf{Z}}\sum_{\mathbf{W}}\int\int\int q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})\log\frac{\mathbb{P}\left(\mathbf{A},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho}\right)}{q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{\rho})}\, d\boldsymbol{\alpha}\ d\boldsymbol{\pi}\ d\boldsymbol{\rho}\\ &= \mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\rho},\boldsymbol{\pi}}[\log\mathbb{P}(\mathbf{A},\mathbf{Z},\boldsymbol{\alpha},\mathbf{W},\boldsymbol{\rho},\boldsymbol{\pi})] - \mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\rho},\boldsymbol{\pi}}[\log q(\mathbf{Z},\boldsymbol{\alpha},\mathbf{W},\boldsymbol{\rho},\boldsymbol{\pi})]\end{aligned} \tag{D.2} .$$

We can decompose the following terms as:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\rho},\boldsymbol{\pi}}[\log p(\mathbf{A},\mathbf{Z},\boldsymbol{\alpha},\mathbf{W},\boldsymbol{\rho},\boldsymbol{\pi})] = \mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha}}&[\log\mathbb{P}(\mathbf{A}\mid\mathbf{Z},\mathbf{W},\boldsymbol{\alpha})] + \mathbb{E}_{\boldsymbol{\alpha}}[\log p(\boldsymbol{\alpha})]\\ &+ \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[\log p(\mathbf{Z}\mid\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\pi}}[\log p(\boldsymbol{\pi})]\\ &+ \mathbb{E}_{\mathbf{W},\boldsymbol{\rho}}[\log p(\mathbf{W}\mid\boldsymbol{\rho})] + \mathbb{E}_{\boldsymbol{\rho}}[\log p(\boldsymbol{\rho})]\end{aligned} \tag{D.3} ,$$

and

$$\begin{aligned}\mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\rho},\boldsymbol{\pi}}[\log q(\mathbf{Z},\boldsymbol{\alpha},\mathbf{W},\boldsymbol{\rho},\boldsymbol{\pi})] = \ &\mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] + \ \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})]\\ &+ \mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{W})] + \mathbb{E}_{\boldsymbol{\rho}}[\log q(\boldsymbol{\rho})].\\ &+ \mathbb{E}_{\boldsymbol{\alpha}}[\log q(\boldsymbol{\alpha})]\end{aligned} \tag{D.4}$$

Now, the next step involves developing each of these terms and simplifying them as much as possible.

Terms dependent on the parameter $\boldsymbol{\alpha}$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha}}[\log \mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha})] + \mathbb{E}_{\boldsymbol{\alpha}}[\log \mathbb{P}(\boldsymbol{\alpha})] \\
=&\sum_{i<j}^{N}\sum_{k,l}^{K}\sum_{v}^{V}\sum_{s}^{Q} \tau_{ik}\tau_{jl}\nu_{vs}\Big\{A_{ijv}\Big(\psi(\eta_{kls}) - \psi(\xi_{kls})\Big) + \psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\Big\} \\
&+\sum_{k\leq l}^{K}\sum_{s}^{Q}\Big\{\log\Gamma(\eta^0_{kls}+\xi^0_{kls}) - \log\Gamma(\eta^0_{kls}) - \log\Gamma(\xi^0_{kls}) + \Big(\eta^0_{kls}-1\Big)\left(\psi(\eta_{kls})\right. \\
&\left.-\,\psi(\xi_{kls}+\eta_{kls})\right) + \Big(\xi^0_{kls}-1\Big)\left(\psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\right)\Big\},
\end{aligned} \tag{D.5}$$

and

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\alpha}}[\log q(\boldsymbol{\alpha})] \\
=&\sum_{k\leq l}^{K}\sum_{s}^{Q}\Big\{\log\Gamma(\eta_{kls}+\xi_{kls}) - \log\Gamma(\eta_{kls}) - \log\Gamma(\xi_{kls}) + \\
&(\eta_{kls}-1)\left(\psi(\eta_{kls}) - \psi(\xi_{kls}+\eta_{kls})\right) + (\xi_{kls}-1)\left(\psi(\xi_{kls}) - \psi(\eta_{kls}+\xi_{kls})\right)\Big\}.
\end{aligned} \tag{D.6}$$

Terms dependent on the parameter $\boldsymbol{\pi}$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}[\log p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\pi}}[\log p(\boldsymbol{\pi})] \\
=&\sum_{i}^{N}\sum_{k}^{K}\tau_{ik}\left(\psi(\beta_k) - \psi(\sum_{k'}\beta_{k'})\right) + \log\Gamma(\sum_{k'}\beta^0_{k'}) - \log\left(\sum_{k'}\Gamma(\beta^0_{k'})\right) \\
&+\sum_{k=1}^{K}\Big(\beta^0_k - 1\Big)\left(\psi(\beta_k) - \psi(\sum_{k'}\beta_{k'})\right),
\end{aligned} \tag{D.7}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] + \ \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})] =&\sum_{i}^{N}\sum_{k}^{K}\tau_{ik}\log\tau_{ik} + \log\Gamma(\sum_{k'}\beta_{k'}) - \log\left(\sum_{k'}\Gamma(\beta_{k'})\right) \\
&+\sum_{k=1}^{K}(\beta_k - 1)\left(\psi(\beta_k) - \psi(\sum_{k'}\beta_{k'})\right).
\end{aligned} \tag{D.8}$$

Terms dependent on the parameter $\boldsymbol{\rho}$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W},\boldsymbol{\rho}}[\log p(\mathbf{W} \mid \boldsymbol{\rho})] + \mathbb{E}_{\boldsymbol{\rho}}[\log p(\boldsymbol{\rho})] \\
&= \sum_{v}^{V}\sum_{s}^{Q} \nu_{vs} \left(\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})\right) + \log\Gamma(\sum_{s'} \theta^0_{s'}) - \log\left(\sum_{s'} \Gamma(\theta^0_{s'})\right) \\
&\quad + \sum_{s=1}^{Q} \left(\theta^0_s - 1\right) \left(\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})\right).
\end{aligned} \tag{D.9}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{W})] + \mathbb{E}_{\boldsymbol{\rho}}[\log q(\boldsymbol{\rho})] = &\sum_{v}^{V}\sum_{s}^{Q} \nu_{vs} \log \nu_{vs} \log\Gamma(\sum_{s'} \theta_{s'}) - \log\left(\sum_{s'} \Gamma(\theta_{s'})\right) \\
&+ \sum_{s=1}^{Q} (\theta_s - 1) \left(\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})\right).
\end{aligned} \tag{D.10}$$

ELBO: Now that all the terms have been developed, it's just a matter of grouping them together, to obtain the ELBO defined at Equation 2.22.

$$\begin{aligned}
&\mathcal{L}\left(q(.)\right) = \\
&\sum_{k<l}^{K}\sum_{s}^{Q} \left(\eta^0_{kls} + \left(\sum_{i\neq j}^{N}\sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}\right) - \eta_{kls}\right) \left(\psi(\eta_{kls}) - \psi(\eta_{kls} + \xi_{kls})\right) \\
&+ \sum_{k=1}^{K}\sum_{s}^{Q} \left(\eta^0_{kks} + \left(\sum_{i<j}^{N}\sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}\right) - \eta_{kks}\right) \left(\psi(\eta_{kks}) - \psi(\eta_{kks} + \xi_{kks})\right) \\
&+ \sum_{k<l}^{K}\sum_{s}^{Q} \left(\xi^0_{kls} + \left(\sum_{i\neq j}^{N}\sum_{v}^{V} \tau_{ik}\tau_{jl}\nu_{vs}(1 - A_{ijv})\right) - \eta_{kls}\right) \left(\psi(\xi_{kls}) - \psi(\eta_{kls} + \xi_{kls})\right) \\
&+ \sum_{k=1}^{K}\sum_{s}^{Q} \left(\xi^0_{kks} + \left(\sum_{i<j}^{N}\sum_{v}^{V} \tau_{ik}\tau_{jk}\nu_{vs}(1 - A_{ijv})\right) - \xi_{kks}\right) \left(\psi(\xi_{kks}) - \psi(\eta_{kks} + \xi_{kks})\right) \\
&+ \sum_{k=1}^{K} \left(\beta^0_k + \sum_{i=1}^{N} \tau_{ik} \;\; - \beta_k\right) \left(\psi(\beta_k) - \psi(\sum_{k'} \beta_{k'})\right) \\
&+ \sum_{q=1}^{Q} \left(\theta^0_s + \sum_{v=1}^{V} \nu_{vs} \;\; - \theta_s\right) \left(\psi(\theta_s) - \psi(\sum_{s'} \theta_{s'})\right) \\
&- \sum_{i}^{N}\sum_{k}^{K} \tau_{ik} \log \tau_{ik} \;\; - \sum_{v}^{V}\sum_{s}^{Q} \nu_{vs} \log \nu_{vs} \\
&+ \log\left\{\frac{\Gamma\left(\sum_{k=1}^{K} \beta^0_k\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^0_k\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q} \theta^0_s\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^0_s\right)}\right\} \\
&+ \sum_{k\leq l}^{K}\sum_{s=1}^{Q} \log\left\{\frac{\Gamma\left(\eta^0_{kls} + \xi^0_{klq}\right) \Gamma\left(\eta_{kls}\right) \Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls} + \xi_{kls}\right) \Gamma\left(\eta^0_{kls}\right) \Gamma\left(\xi^0_{kls}\right)}\right\}
\end{aligned} \tag{D.11}$$

However, by definition of the parameters, we have many terms that cancel each other out:

- $\eta_{kls} = \eta^0_{kls} + \left(\sum_{i \neq j}^N \sum_v^V \tau_{ik}\tau_{jl}\nu_{vs}A_{ijv}\right)$,
- $\eta_{kks} = \eta^0_{kks} + \left(\sum_{i<j}^N \sum_v^V \tau_{ik}\tau_{jk}\nu_{vs}A_{ijv}\right)$,
- $\eta_{kls} = \xi^0_{kls} + \left(\sum_{i \neq j}^N \sum_v^V \tau_{ik}\tau_{jl}\nu_{vs}(1 - A_{ijv})\right)$,
- $\xi_{kks} = \xi^0_{kks} + \left(\sum_{i<j}^N \sum_v^V \tau_{ik}\tau_{jk}\nu_{vs}(1 - A_{ijv})\right)$,
- $\beta_k = \beta^0_k + \sum_{i=1}^N \tau_{ik}$,
- $\theta_s = \theta^0_s + \sum_{v=1}^V \nu_{vs}$.

Hence:

$$\begin{aligned}
\mathcal{L}\left(q(.)\right) = \log \left\{ \frac{\Gamma\left(\sum_{k=1}^K \beta^0_k\right) \prod_{k=1}^K \Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^K \beta_k\right) \prod_{k=1}^K \Gamma\left(\beta^0_k\right)} \right\} + \log \left\{ \frac{\Gamma\left(\sum_{s=1}^Q \theta^0_s\right) \prod_{s=1}^Q \Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^Q \theta_s\right) \prod_{s=1}^Q \Gamma\left(\theta^0_s\right)} \right\} \\
+ \sum_{k \leq l}^K \sum_{s=1}^Q \log \left\{ \frac{\Gamma\left(\eta^0_{kls} + \xi^0_{klq}\right) \Gamma\left(\eta_{kls}\right) \Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls} + \xi_{kls}\right) \Gamma\left(\eta^0_{kls}\right) \Gamma\left(\xi^0_{kls}\right)} \right\} \\
- \sum_i^N \sum_k^K \tau_{ik} \log \tau_{ik} \;\; - \sum_v^V \sum_s^Q \nu_{vs} \log \nu_{vs}
\end{aligned} \tag{D.12}$$

■

Appendix A
Appendix B
Appendix C
Appendix D
Appendix E

E Details of Gibbs Sampling - EM with gradient ascent algorithm for MoEBIUS model

E.1 Variational parameters of clustering τ_{ik}

The optimal estimation of $(\tau_{ik})_{i=1:N,k=1:K}$, as shown in Equation 3.19, is obtained by

$$\tau_{ik}^{(t+1)} = \frac{\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}^{(t)}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \hat{\mathbf{W}}_k^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)}{\sum_{k'} \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik'}=1, \hat{\mathbf{W}}_{k'}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)}. \tag{E.1}$$

Proof

$$\begin{aligned}
\tau_{ik} &= \mathbb{P}\left(Z_{ik}=1 \mid y_i, \mathbf{X}_i, \mathbf{W}_k, \boldsymbol{\pi}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right) \\
&= \frac{\mathbb{P}\left(Z_{ik}=1 \mid \mathbf{X}_i, \boldsymbol{\pi}^{(t)}\right) \mathbb{P}\left(y_i \mid Z_{ik}=1, \mathbf{X}_i, \hat{\mathbf{W}}_k, \boldsymbol{\beta}_k^{(t)}\right)}{\sum_{k'=1}^{K} \mathbb{P}\left(Z_{ik'}=1 \mid \mathbf{X}_i, \boldsymbol{\pi}^{(t)}\right) \mathbb{P}\left(y_i \mid Z_{ik'}=1, \mathbf{X}_i, \hat{\mathbf{W}}_k, \boldsymbol{\beta}_k^{(t)}\right)} \\
&= \frac{\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}^{(t)}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \hat{\mathbf{W}}_k^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)}{\sum_{k'} \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}^{(t)}}} \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik'}=1, \hat{\mathbf{W}}_{k'}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)}.
\end{aligned} \tag{E.2}$$

■

E.2 Variational parameters of component membership ν_{vs}

The optimal estimation of $(\nu_{kjs})_{k=1:K,j=1:p,s=1:Q}$, as shown in Equation 3.21, is obtained by

$$\nu_{kjs}^{(t+1)} = \frac{\rho_{ks}^{(t)} \prod_i^N \mathbb{P}\left(y_i \mid \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, W_{kjs}=1, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}{\sum_{s'} \rho_{ks'}^{(t)} \prod_i^N \mathbb{P}\left(y_i \mid \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, W_{kjs'}=1, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}, \tag{E.3}$$

Proof

$$\begin{aligned}
&\nu_{kjs}^{(t+1)} \\
&= \mathbb{P}\left(W_{kjs}=1 \mid \mathbf{y}, \mathbf{X}, \hat{\mathbf{Z}}_{\bullet k}^{(t+1)}, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\Theta}^{(t)}\right) \\
&= \frac{\mathbb{P}\left(W_{kjs}=1 \mid \hat{\mathbf{Z}}_{\bullet k}^{(t+1)}, \boldsymbol{\rho}_k^{(t)}\right) \prod_{i=1}^N \mathbb{P}\left(y_i \mid W_{kjs}=1, \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}{\sum_{s'=1}^Q \mathbb{P}\left(W_{kjs'}=1 \mid \hat{\mathbf{Z}}_{\bullet k}^{(t+1)}, \boldsymbol{\rho}_k^{(t)}\right) \prod_{i=1}^N \mathbb{P}\left(y_i \mid W_{kjs'}=1, \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}} \\
&= \frac{\rho_{ks}^{(t)} \prod_{i=1}^N \mathbb{P}\left(y_i \mid W_{kjs}=1, \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}{\sum_{s'=1}^Q \rho_{ks'}^{(t)} \prod_{i=1}^N \mathbb{P}\left(y_i \mid W_{kjs'}=1, \mathbf{X}_i, \hat{Z}_{ik}^{(t+1)}, \hat{\mathbf{W}}_{-kj}^{(t)}, \boldsymbol{\beta}_k^{(t)}\right)^{\hat{Z}_{ik}^{(t+1)}}}.
\end{aligned} \tag{E.4}$$

■

E.3 Optimization of $\boldsymbol{\rho}$

The optimal estimation of ρ_s, indicated at Equation 3.23, is given by

$$\rho_{ks}^{(t+1)} = \frac{\sum_{j=1}^p \hat{\mathbf{W}}_{kjs}^{(t+1)}}{p}, \qquad \forall k \in \{1, \ldots, K\}, \forall s \in \{1, \ldots, Q\}. \tag{E.5}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$J_{\boldsymbol{\rho}}(.) = \sum_{k=1}^K \sum_{j=1}^p \sum_{s=1}^Q \hat{\mathbf{W}}_{kjs}^{(t+1)} \log\left(\rho_{ks}\right) + \sum_{k=1}^K \lambda_{\boldsymbol{\rho}_k} \left(1 - \sum_{s=1}^Q \rho_{ks}\right), \tag{E.6}$$

where $\left(\lambda_{\boldsymbol{\rho}_k}\right)_{k=1:K}$ are Lagrange parameters.

$$\frac{\partial J_{\boldsymbol{\rho}}}{\partial \rho_{ks}}(.) = \frac{\sum_{j=1}^{p} \hat{\mathbf{W}}_{kjs}^{(t+1)}}{\rho_{ks}} - \lambda_{\boldsymbol{\rho}_k}. \tag{E.7}$$

Moreover, by constraints on $\boldsymbol{\rho}_k$,

$$\sum_{s=1}^{Q} \rho_{ks} = 1 \Longleftrightarrow \sum_{s=1}^{Q} \frac{\sum_{j=1}^{p} \hat{\mathbf{W}}_{kjs}^{(t+1)}}{\lambda_{\boldsymbol{\rho}_k}} = 1 \Longleftrightarrow \lambda_{\boldsymbol{\rho}_k} = \sum_{s=1}^{Q}\sum_{j=1}^{p} \hat{\mathbf{W}}_{kjs}^{(t+1)} \Longleftrightarrow \lambda_{\boldsymbol{\rho}} = p. \tag{E.8}$$

Finally,

$$\rho_{ks}^{(t+1)} = \frac{\sum_{j=1}^{p} \hat{\mathbf{W}}_{kjs}^{(t+1)}}{p}. \tag{E.9}$$

■

E.4 Optimization of $\boldsymbol{\pi}$

We aim to prove that the derivative of the function $\mathcal{J}$ corresponds to

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\pi}}(\cdot) = \mathbf{X}^T \left(\hat{\mathbf{Z}}^{(t+1)} - \mathbf{S}^{\boldsymbol{\pi}}\right), \tag{E.10}$$

as shown in Equation 3.24.

Proof

$$\begin{aligned}
\mathcal{J}(\cdot) &\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \log \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \mathbf{X}_i \boldsymbol{\pi}_{\bullet k} - \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \log\left(\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \mathbf{X}_i \boldsymbol{\pi}_{\bullet k} - \sum_{i=1}^{N} \log\left(\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}\right).
\end{aligned} \tag{E.11}$$

The gradient of the function $\mathcal{J}$ with respect to the parameter π_{jk} is obtained by

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \pi_{jk}}(\cdot) &= \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)} X_{ij} - \sum_{i=1}^{N} X_{ij} \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}} \\ &= \sum_{i=1}^{N} X_{ij} \left(\hat{Z}_{ik}^{(t+1)} - S_{ik}^{\boldsymbol{\pi}} \right). \qquad \text{(E.12)}\end{aligned}$$

Finally, writing this in matrix form :

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\pi}}(\cdot) = \mathbf{X}^T \left(\hat{\mathbf{Z}}^{(t+1)} - \mathbf{S}^{\boldsymbol{\pi}} \right). \qquad \text{(E.13)}$$

■

E.5 Optimization of $\boldsymbol{\beta}$

E.5.1 Optimization of $\boldsymbol{\beta}$ - Regression

The optimal estimation of $(\beta_{ks})_{k=1:K,s=1:Q}$, as shown in Equation 3.27, is obtained by

$$\boldsymbol{\beta}_k^{(t+1)} = \left(\mathbf{W}_k^{(t+1)^T} \mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} \right) \mathbf{X} \mathbf{W}_k^{(t+1)} \right)^{-1} \mathbf{W}_k^{(t+1)^T} \mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} \right) \mathbf{y}. \qquad \text{(E.14)}$$

Proof

$$\begin{aligned}\mathcal{J}(\cdot) &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \hat{\mathbf{W}}_k^{(t+1)}, \boldsymbol{\beta}_k \right) \\ &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \left(-\frac{1}{2\sigma_k^{2^{(t)}}} \left(y_i - \mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k \right)^2 \right). \qquad \text{(E.15)}\end{aligned}$$

Writing this in matrix form:

$$\begin{aligned}\mathcal{J}(\cdot) \propto \sum_{k=1}^{K} &-\frac{1}{2\sigma_k^{2^{(t)}}} \mathbf{y}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} \right) \mathbf{y} + \frac{1}{2\sigma_k^{2^{(t)}}} \times 2\mathbf{y}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} \right) \mathbf{X} \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k \\ &- \frac{1}{2\sigma_k^{2^{(t)}}} \left(\mathbf{X} \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k \right)^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)} \right) \mathbf{X} \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k. \qquad \text{(E.16)}\end{aligned}$$

The gradient of the function $\mathcal{J}$ with respect to the parameter $\boldsymbol{\beta}_k$ is obtained by:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k}(\cdot) = \frac{1}{\sigma_k^{2(t)}} \hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{y} - \frac{1}{\sigma_k^{2(t)}}\hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T\operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_k. \tag{E.17}$$

. We aim to find the optimal parameter by setting the gradient to zero:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k}(\cdot) &= 0 \\ \hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T \operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{y} &= \hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T\operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_k \end{aligned} \tag{E.18}$$

. ■

Hence,

$$\boldsymbol{\beta}_k^{(t+1)} = \left(\hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T\operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)}\right)^{-1}\hat{\mathbf{W}}_k^{(t+1)^T}\mathbf{X}^T\operatorname{diag}\left(\hat{\mathbf{Z}}_{\bullet k}^{(t+1)}\right)\mathbf{y}. \tag{E.19}$$

E.5.2 Optimization of σ_k^2

The optimal estimation of $(\sigma_k^2)_{k=1:K}$, as shown in Equation 3.28, is obtained by

$$\sigma_k^{2(t+1)} = \frac{\sum_{i=1}^N \hat{Z}_{ik}^{(t+1)}\left(y_i - \mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_k^{(t+1)}\right)^2}{\sum_{i=1}^N \hat{Z}_{ik}^{(t+1)}}. \tag{E.20}$$

Proof

$$\begin{aligned} \mathcal{J}(\cdot) &\propto \sum_{i=1}^N\sum_{k=1}^K \hat{Z}_{ik}^{(t+1)}\log\mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \hat{\mathbf{W}}_k^{(t+1)}, \boldsymbol{\beta}_k^{(t+1)}\right) \\ &= \sum_{i=1}^N\sum_{k=1}^K \hat{Z}_{ik}^{(t+1)}\log\left[\frac{1}{\sqrt{2\pi}\sqrt{\sigma_k^2}}\exp\left(-\frac{1}{2\sigma_k^2}\left(y_i - \mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_k^{(t+1)}\right)^2\right)\right]. \end{aligned} \tag{E.21}$$

The gradient of the function $\mathcal{J}$ with respect to the parameter σ_k^2 is obtained by:

$$\frac{\partial \mathcal{J}}{\partial \sigma_k^2}(\cdot) = \sum_{i=1}^N \hat{Z}_{ik}^{(t+1)}\left[-\frac{1}{2\sqrt{\sigma_k^2}} + \frac{\left(y_i - \mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_k^{(t+1)}\right)^2}{2(\sigma_k^2)^2}\right]. \tag{E.22}$$

We aim to find the optimal parameter by setting the gradient to zero:

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \sigma_k^2}(\cdot) &= 0 \\ \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)} &= \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)} \frac{\left(y_i - \mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k^{(t+1)}\right)^2}{\sigma_k^2}.\end{aligned} \tag{E.23}$$

Finally,

$$\sigma_k^{2(t+1)} = \frac{\sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)} \left(y_i - \mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_k^{(t+1)}\right)^2}{\sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)}}. \tag{E.24}$$

■

E.5.3 Optimization of $\boldsymbol{\beta}$ - multiclass classification

Estimation of $(\beta_{ksc})_{k=1:K,s=1:Q,c=1:C}$ is obtained by gradient ascent. To do this, we'll prove that the gradient corresponds to

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k}(\cdot) = \left[\left(\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)^T}\right) \odot \mathbf{1}_{Q,1} \hat{\mathbf{Z}}_{\bullet k}^{(t+1)^T}\right] \left(\mathbf{y} - \mathbf{S}^{\boldsymbol{\beta}_k}\right). \tag{E.25}$$

as shown in Equation 3.30.

Proof

$$\begin{aligned}\mathcal{J}(\cdot) &\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \hat{\mathbf{W}}_k^{(t+1)}, \boldsymbol{\beta}_k\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{c=1}^{C} y_{ic} \hat{Z}_{ik}^{(t+1)} \log\left(\frac{e^{\mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \beta_{k\bullet c}}}{\sum_{c'} e^{\mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \beta_{k\bullet c'}}}.\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{c=1}^{C} y_{ic} \hat{Z}_{ik}^{(t+1)} \mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \boldsymbol{\beta}_{k\bullet c} - \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{Z}_{ik}^{(t+1)} \log\left(\sum_{c'} e^{\mathbf{X}_i \hat{\mathbf{W}}_k^{(t+1)} \beta_{k\bullet c'}}\right).\end{aligned} \tag{E.26}$$

The gradient of the function $\mathcal{J}$ with respect to the parameter β_{ksc} is obtained by:

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \beta_{ksc}}(\cdot) &= \sum_{i=1}^{N} y_{ic}\hat{Z}_{ik}^{(t+1)}\left(\sum_{j=1}^{p} X_{ij}W_{kjs}^{(t+1)}\right) - \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)}\left(\sum_{j=1}^{p} X_{ij}W_{kjs}^{(t+1)}\right)\left(\frac{e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_{k\bullet c}}}{\sum_{c'} e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_{k\bullet c'}}}\right) \\
&= \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)}\left(\sum_{j=1}^{p} X_{ij}W_{kjs}^{(t+1)}\right)\left(y_{ic} - \frac{e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_{k\bullet c}}}{\sum_{c'} e^{\mathbf{X}_i\hat{\mathbf{W}}_k^{(t+1)}\boldsymbol{\beta}_{k\bullet c'}}}\right) \\
&= \sum_{i=1}^{N} \hat{Z}_{ik}^{(t+1)}\left(\sum_{j=1}^{p} X_{ij}W_{kjs}^{(t+1)}\right)\left(y_{ic} - \mathbf{S}_{ic}^{\boldsymbol{\beta}_k}\right). \qquad \text{(E.27)}
\end{aligned}$$

Writing this in matrix form :

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}_k}(\cdot) = \left[\left(\mathbf{X}\hat{\mathbf{W}}_k^{(t+1)^T}\right) \odot \mathbf{1}_{Q,1}\hat{\mathbf{Z}}_{\bullet k}^{(t+1)^T}\right]\left(\mathbf{y} - \mathbf{S}^{\boldsymbol{\beta}_k}\right). \qquad \text{(E.28)}$$

■

F Co-conditional Latent Block Model

F.1 Introduction

The Conditional Latent Block Model (*CLBM*), introduced by Goffinet et al. (2020), is an extension of latent block models that incorporates a conditional component into the constructed blocks. Initially, the *CLBM* was designed for co-clustering of temporal data, enabling the simultaneous partitioning of individuals and variables. In this model, the conditioning of individuals is directly influenced by the conditioning of variables, thereby creating an interdependent relationship.

More specifically, the variable components remain consistent across all individual blocks, ensuring homogeneity in the representation of variables. However, the defined communities of individuals vary depending on the block of variable components considered. In other words, while the structure of the variables remains constant, the way individuals group into communities depends on the specific characteristics of the variable blocks.

In our approach, we adopt a reverse conditioning strategy, where the communities of individuals remain constant, but the components change based on the communities. That is, individuals are stably grouped into communities, but the variable partitions adjust according to these communities. To distinguish this model, we refer to it as the Co-Conditional Latent Block Model (*Co-CoLBM*).

This form of conditioning allows us to model structures where the way variables group together varies according to the characteristics of the individual communities, offering greater flexibility to capture dynamics specific to each group, as illustrated in Figure F.1 and Figure 3.1.

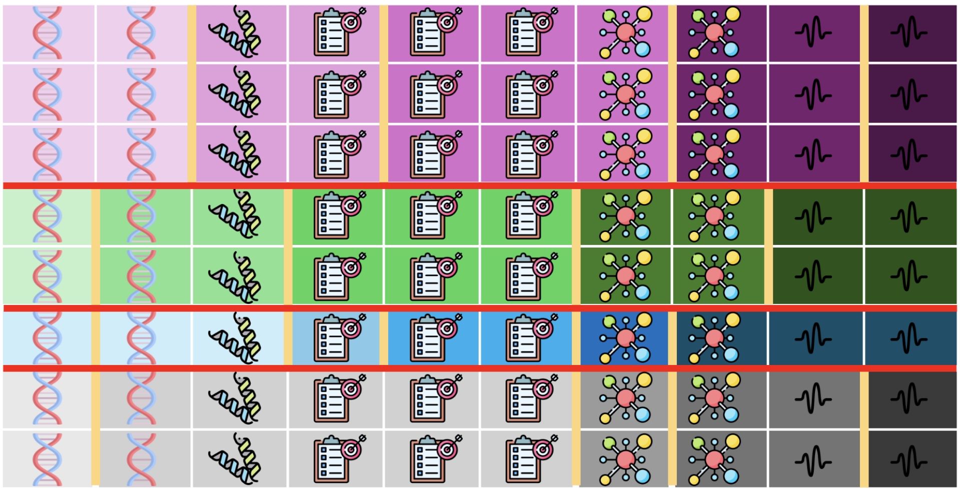


Figure F.1: Example of data partitioning using multimodal information: (i) Considering only the red rows results in a classic mixture model with four communities. (ii) Taking into account the variable types, the model becomes a Latent Block Model, where observed variables are grouped into latent blocks capturing underlying structures. (iii) When both the red and orange rows are considered, the model becomes a Co-Conditional Latent Block Model, integrating aspects of both mixture models and conditional latent block models, where the communities are stable, and the orange lines indicate the conditional stratification of variables based on the community.

Conditioning based on groups of individuals is particularly important in contexts like image analysis. For example, within a set of images, a specific pixel may provide no relevant information for one group of images but may be crucial for another, as shown in Figure 3.1.

More generally, the strength of the *Co-CoLBM* lies in its ability to capture situations where the same variable may exhibit entirely different behaviors, and therefore meanings, depending on the observed community. This makes it possible to model complex relationships where variables are not interpreted uniformly across all individuals, but are instead adjusted based on the specific characteristics of each group.

F.2 Materials and Methods

In this analysis, we directly consider the feature matrix $\mathbf{X}$, of dimension $N \times p$. The primary goal is to characterize and cluster the variables that share common information, conditioned on the group structure defined by the individuals. This approach leverages the intrinsic structure of the data to identify correlations and meaningful patterns among the variables, while accounting for the similarities and differences between the various groups of individuals.

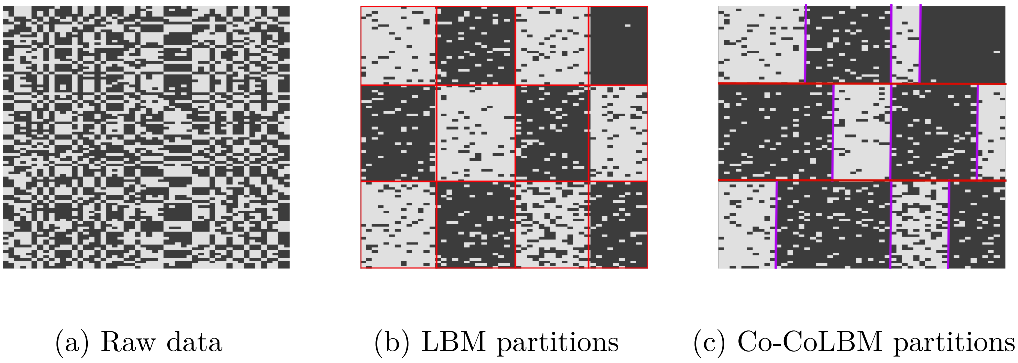
Once again, we define:

$$\mathbf{Z}_i \sim \mathcal{M}\left(1; \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)\right). \tag{F.1}$$

We now define the tensor $\mathbf{W}$, which represents the partition of the variables into Q components, conditioned by the communities:

$$\mathbf{W}_{kj\bullet} \mid \mathbf{Z}_{\bullet k} \sim \mathcal{M}\left(1; \boldsymbol{\rho}_k = (\rho_{k1}, \ldots, \rho_{kQ})\right). \tag{F.2}$$

Using these latent variables, we can generally define the distribution of our observations as follows:



(a) Raw data (b) LBM partitions (c) Co-CoLBM partitions

Figure F.2: Comparison of data organization using different modeling methods. Figure (a) shows the raw data, arranged without any prior sorting or clustering. Figure (b) displays the data sorted according to the partitions obtained via an LBM. Figure (c) presents the data arranged by the *Co-CoLBM*, where the individual partitions are stable, but the variable partitions depend on the chosen community.

$$X_{ij} \mid Z_{ik} = 1, W_{kjs} = 1 \sim \mathbb{P}\left(X_{ij}; \boldsymbol{\Theta}_{ks}\right) \tag{F.3}$$

This modeling approach leads to the following marginal likelihood:

$$\mathbb{P}(\mathbf{X} \mid \mathbf{\Theta}) = \sum_{\mathcal{Z}\times\mathcal{W}} \left[\prod_i \prod_j \mathbb{P}\left(X_{ij} \mid Z_{ik} = 1, W_{kjs} = 1, \mathbf{\Theta}_{ks}\right) \right.$$
$$\left. \prod_k \prod_j \prod_s \mathbb{P}\left(W_{kjs} \mid Z_{\bullet k}, \mathbf{\Theta}\right) \prod_i \prod_k \mathbb{P}\left(Z_{ik} \mid \mathbf{\Theta}\right) \right]. \tag{F.4}$$

The marginal likelihood of the model corresponds to the one initially obtained with the Latent Block Model, up to the conditioning on $\mathbf{W}$. Moreover, the same computational constraints apply, making direct computation highly challenging. Therefore, we rely on a VEM approach.

F.3 Optimization by Variational EM

As in Section 2.3.3, we aim to maximize an evidence lower bound, which in our context is expressed as:

$$\mathcal{L}\left(q(\cdot)\right) = \sum_{\mathbf{Z},\mathbf{W}} q(\mathbf{Z}, \mathbf{W}) \log \frac{\mathbb{P}\left(\mathbf{X}, \mathbf{Z}, \mathbf{W}\right)}{q(\mathbf{Z}, \mathbf{W})}. \tag{F.5}$$

By applying a mean-field approximation, we define q as:

$$\begin{aligned} q\left(\mathbf{Z}, \mathbf{W}\right) &= \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{k=1}^{K} \prod_{j=1}^{p} q(\mathbf{W}_{kj}) \\ &= \prod_{i=1}^{N} \mathcal{M}\left(\mathbf{Z}_i; (\tau_{i1}, \dots, \tau_{iK})\right) \prod_{k=1}^{K} \prod_{j=1}^{p} \mathcal{M}\left(\mathbf{W}_{kj}; (\nu_{kj1}, \dots, \nu_{kjQ})\right) \end{aligned} \tag{F.6}$$

To optimize all parameters, we alternate between the Variational Expectation (VE) and Maximization (M) steps until convergence.

Variational Expectation step During the VE step, we optimize the latent variable parameters. After constrained maximization, the following results are obtained:

$$\tau_{ik} = \frac{\exp\Big(\log(\pi_k) + \sum_j^p \sum_s^Q \nu_{kjs} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right) - 1\Big)}{\sum_{k'} \exp\Big(\log(\pi_{k'}) + \sum_j^p \sum_s^Q \nu_{k'js} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{k's}\right) - 1\Big)}, \tag{F.7}$$

$$\nu_{kjs} = \frac{\exp\left(\log(\rho_{ks}) + \sum_{i=1}^{N} \tau_{ik} \log \mathbb{P}\left(X_{ij} \mid \boldsymbol{\Theta}_{ks}\right) - 1\right)}{\sum_{s'} \exp\left(\log(\rho_{ks'}) + \sum_{i=1}^{N} \tau_{ik} \log \mathbb{P}\left(X_{ij} \mid \boldsymbol{\Theta}_{ks}\right) - 1\right)}. \tag{F.8}$$

The parameters obtained during the variational law estimation q resemble those from classical variational distributions in the Latent Block Model (Brault and Mariadassou, 2015). However, the conditioning by $Z_{\bullet k}$ involves parameters (τ_{ik}), which weight individual's contributions to component construction proportionally to their community membership.

Maximization step The updates for the proportion parameters $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ correspond to standard estimates:

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{ik}}{N}, \qquad \forall k \in \{1, \dots, K\}. \tag{F.9}$$

$$\rho_{ks} = \frac{\sum_{j=1}^{p} \nu_{kjs}}{p}, \qquad \forall k \in \{1, \dots, K\}, \forall s \in \{1, \dots, Q\}. \tag{F.10}$$

If we assume that $(X_{ij})_{i=1:N, j=1:p}$ are normally distributed, conditional on Z_{ik}, W_{kjs}, according to $\mathcal{N}\left(X_{ij}; \mu_{ks}, \sigma^2_{ks}\right)$. By deriving the evidence lower bound, we obtain estimators $\hat{\mu}_{ks}$ and $\hat{\sigma}^2_{ks}$, where:

$$\hat{\mu}_{ks} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} X_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs}}. \tag{F.11}$$

$$\hat{\sigma}^2_{ks} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} \left(X_{ij} - \hat{\mu}_{ks}\right)^2}{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs}}. \tag{F.12}$$

These estimators are similar to those from the Gaussian Latent Block Model's emission law, except that the variational parameters of the components depend on the community k. Further optimization details are provided in Appendix G.

F.4 Model selection

The model selection for the *Co-CoLBM* is based on the ICL criterion, as proposed by Biernacki et al. (2000). More specifically, it relies on the extensions developed by Lomet (2012) for biclustering and by Goffinet et al. (2020) for conditional classes extension. In this extension, the ICL takes into account both the number of communities K and the number of components Q, allowing for a more adapted model selection in line with the structure of conditional clustering.

$$\text{ICL}(\mathbf{X}, K, Q) = \log \mathbb{P}\left(\mathbf{X}, \hat{\mathbf{Z}}, \hat{\mathbf{W}} \mid \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\rho}}\right) - \frac{K-1}{2} \log (N) - K\frac{Q-1}{2} \log (p) - \frac{KQ}{2} \operatorname{Card}\left(\boldsymbol{\Theta}_{ks}\right) \log (Np). \tag{F.13}$$

The ICL penalizes the log-likelihood based on the number of free parameters, thus avoiding overfitting. The best model is the one that maximizes the ICL, achieving an optimal balance between the precision of the fit and the complexity of the model.

F.5 Experiments

The capabilities of the *Co-CoLBM* model will be evaluated based on three key criteria:

1. **Co-clustering performance:** To assess the quality of the co-clustering produced by *Co-CoLBM*, two standard metrics will be employed: the Adjusted Rand Index (Steinley, 2004, ARI) and the Normalized Mutual Information (Strehl and Ghosh, 2002, NMI). The ARI measures clustering accuracy by quantifying the agreement between estimated partitions and a reference partition, while the NMI quantifies the amount of shared information between the estimated clusters and the true underlying data structure. Both metrics are normalized, reaching a maximum value of 1 when the estimated clustering perfectly matches the reference clustering. Additionally, to ensure optimal alignment between the obtained and true clusters, the Hungarian algorithm will be applied to realign the generated clusters. *Co-CoLBM*'s performance will be compared to traditional clustering methods such as K-Means, Spectral Clustering, and Hierarchical Agglomerative Clustering. Each method will first be applied to rows to identify observation communities, followed by application to each community's columns to segment associated variables.

2. **Estimation of emission distribution parameters:** Another crucial aspect of evaluation is the model's ability to accurately estimate the

emission distribution parameters $(\mu_{ks}, \sigma^2_{ks})$, which characterize the communities and components across different simulations.

3. **Model selection:** Finally, *Co-CoLBM* will be evaluated in terms of model selection using two criteria: the ELBO and the ICL.

The simulations are designed using a generative framework based on the *Co-CoLBM* model, and described in Section 3.5.

F.5.1 Co-clustering Performance

The co-clustering performance results are reported in Table F.1.

In the first and third simulations, the K-means, Hierarchical Clustering, and *Co-CoLBM* models achieve near-perfect results for both rows and columns. In contrast, Spectral Clustering shows excellent performance for rows but struggles to accurately capture the structure of the columns.

For the second simulation, K-means, Hierarchical Clustering, and *Co-CoLBM* continue to perform well, although a slight overall performance decline is observed. Spectral Clustering, on the other hand, shows poor results for the columns and a noticeable drop in performance for the rows, indicating difficulty in correctly segmenting the data in this configuration.

The fourth simulation reveals a slight performance degradation across all algorithms. K-means shows a notable decline in column clustering accuracy, whereas *Co-CoLBM* remains relatively stable. Hierarchical Clustering also demonstrates a significant decrease in column performance, while maintaining good results for rows. Spectral Clustering continues to underperform for columns, though it remains competitive for row clustering.

F.5.2 Co-clustering Performance

Overall, *Co-CoLBM* and *K-means* are the most effective algorithms, achieving scores close to the optimal in most simulations. Both methods appear to handle the structures of rows and columns well. However, *Co-CoLBM* demonstrates greater stability in scenarios where the number of observations is reduced. *Hierarchical Clustering* remains effective but shows a tendency to slightly underperform in certain simulations, particularly regarding the columns. Finally, *Spectral Clustering* seems to be the least suitable algorithm

Examples	Algorithms	Clustering			
		ARI		NMI	
		Row	Col	Row	Col
Simulation 1	K-means	**1.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(0.000)	(0.000)	(0.000)
	Hierarchical Clustering	**1.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(0.000)	(0.000)	(0.000)
	Spectral Clustering	**1.000**	0.456	**1.000**	0.505
		(0.000)	(0.479)	(0.000)	(0.448)
	Co-CoLBM	**1.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(0.000)	(0.000)	(0.000)
Simulation 2	K-means	**1.000**	0.909	**1.000**	0.953
		(0.000)	(0.028)	(0.000)	(0.013)
	Hierarchical Clustering	**1.000**	0.915	**1.000**	0.956
		(0.000)	(0.026)	(0.00)	(0.012)
	Spectral Clustering	0.670	0.145	0.785	0.449
		(0.214)	(0.138)	(0.149)	(0.098)
	Co-CoLBM	**1.000**	**0.922**	**1.000**	**0.957**
		(0.000)	(0.029)	(0.000)	(0.012)
Simulation 3	K-means	0.986	**1.000**	0.969	**1.000**
		(0.014)	(0.000)	(0.028)	(0.000)
	Hierarchical Clustering	0.949	0.999	0.907	0.999
		(0.040)	(0.004)	(0.061)	(0.005)
	Spectral Clustering	0.986	0.468	0.968	0.509
		(0.015)	(0.485)	(0.030)	(0.452)
	Co-CoLBM	**0.987**	**1.000**	**0.971**	**1.000**
		(0.014)	(0.000)	(0.028)	(0.000)
Simulation 4	K-means	0.978	0.737	0.970	0.769
		(0.055)	(0.127)	(0.073)	(0.101)
	Hierarchical Clustering	0.926	0.650	0.901	0.693
		(0.099)	(0.117)	(0.121)	(0.089)
	Spectral Clustering	0.979	0.357	0.970	0.417
		(0.053)	(0.340)	(0.076)	(0.324)
	Co-CoLBM	**0.981**	**0.833**	**0.973**	**0.848**
		(0.049)	(0.100)	(0.068)	(0.083)

Table F.1: Comparison of co-clustering performance on simulated data. The mean (standard deviation) of the scores obtained over 100 simulations.

for these data, especially for column segmentation, where its performance consistently falls below that of the other algorithms.

F.5.3 Parameter Estimations

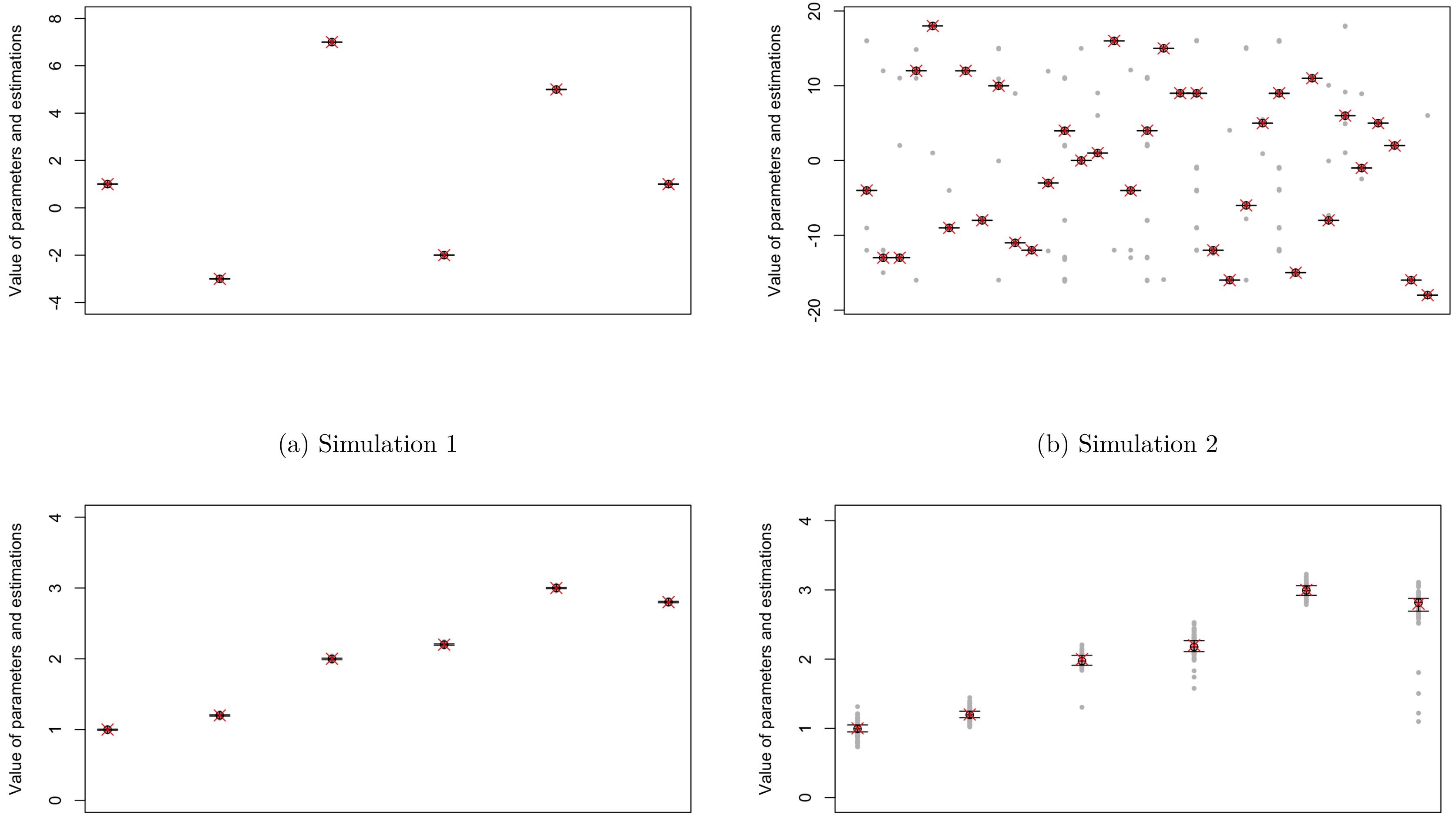
From the simulations presented in Figure F.3, it is observed that in simulations 1 and 3, the parameters are generally estimated with high precision. In contrast, during simulation 2, where the number of parameters was increased, estimating the parameters becomes noticeably more complex. Although the estimates are generally successful, the task becomes more challenging as the number of parameters increases for a fixed number of observations. In simulation 4, where the number of observations was drastically reduced, the estimates converge less accurately toward the true values. However, they still remain within a reasonable range, suggesting that despite the reduced number of observations, the method remains relatively robust.

F.5.4 Model Selection

Figures F.4 and F.5 illustrate how many times the corresponding hyperparameter value was selected based on the ELBO and ICL criteria.

Regarding the hyperparameter K, which determines the number of communities, it is generally estimated with high accuracy, except in rare cases where the number of observations is reduced (Simulation 4), where exceptions are observed.

Regarding the number of components Q, it is observed that the ELBO tends to overestimate the number of components compared to the ICL. The ICL is generally more accurate in parameter selection, although it may sometimes overestimate (Simulation 2) or slightly underestimate (Simulation 4) the number of components.



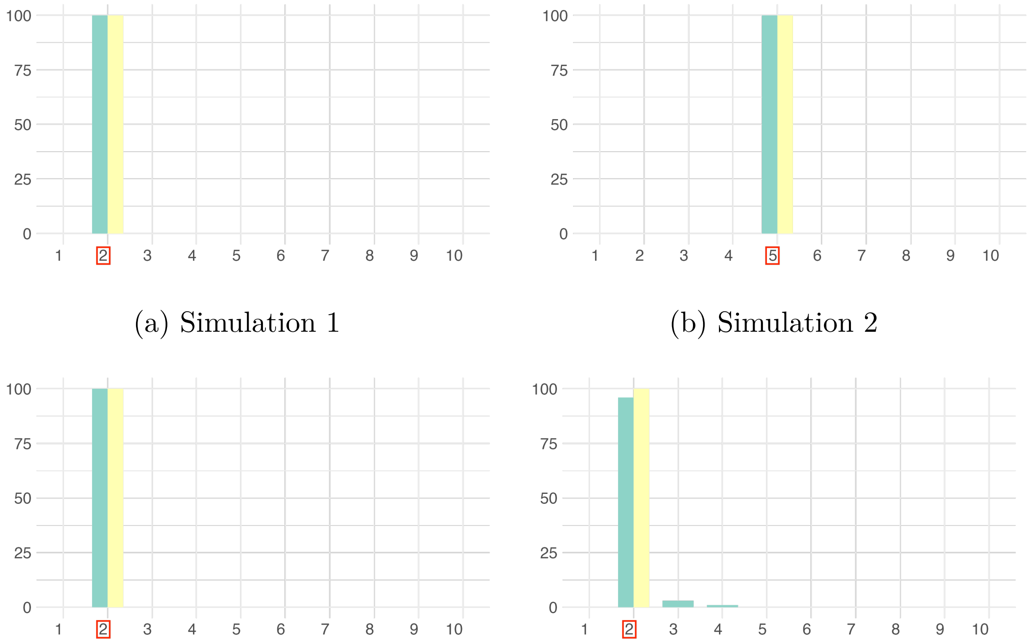
(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure F.3: Parameter estimations based on the four simulation scenarios. The x-axis represents different variables, while the y-axis indicates their corresponding values. The red cross denotes the true value of the parameter. The black circles mark the observed median values for each variable, and the bars represent the first and third quartiles. The gray points illustrate the estimated values for each repetition.



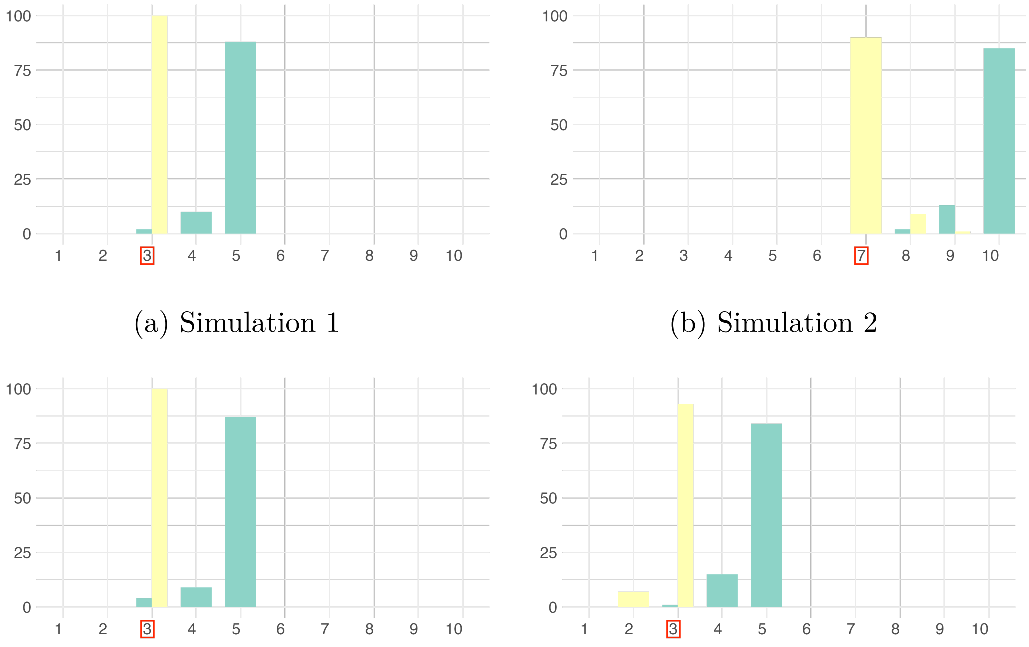
(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure F.4: Comparison of hyperparameter K selection between ICL (yellow) and ELBO (green) criteria across the four simulation scenarios. The red square indicates the true value of the hyperparameter.



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure F.5: Comparison of hyperparameter Q selection between ICL (yellow) and ELBO (green) criteria across the four simulation scenarios. The red square indicates the true value of the hyperparameter.

G Details of Variation EM algorithm for Co-conditional Latent Block Model

G.1 Variational parameters of clustering τ_{ik}

The optimal estimation of τ_{ik} is given by

$$\tau_{ik} = \frac{\exp\Big(\log(\pi_k) + \sum_j^p \sum_s^Q \nu_{kjs} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right) - 1\Big)}{\sum_{k'} \exp\Big(\log(\pi_{k'}) + \sum_j^p \sum_s^Q \nu_{k'js} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{k's}\right) - 1\Big)}. \tag{G.1}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$\begin{aligned} J_{\boldsymbol{\tau}}(.) = &\sum_{i=1}^N \sum_{j=1}^p \sum_{k=1}^K \sum_{s=1}^Q \tau_{ik}\nu_{kjs} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \sum_{i=1}^N \sum_{k=1}^K \tau_{ik} \log\left(\pi_k\right) \\ &- \sum_{i=1}^N \sum_{k=1}^K \tau_{ik} \log\left(\tau_{ik}\right) + \sum_{i=1}^N \lambda_i \left(1 - \sum_{k=1}^K \tau_{ik}\right), \end{aligned} \tag{G.2}$$

where $(\lambda_i)_{i=1:N}$ are Lagrange parameters. Now, the aim is to get the gradient in τ_{ik} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_{\boldsymbol{\tau}}}{\partial \tau_{ik}}(.) = \sum_{j=1}^p \sum_{s=1}^Q \nu_{kjs} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \log\left(\pi_k\right) - \log\left(\tau_{ik}\right) - 1 - \lambda_i. \tag{G.3}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\tau_{ik}\right) = \underbrace{\sum_{j=1}^p \sum_{s=1}^Q \nu_{kjs} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \log\left(\pi_k\right) - 1}_{T_{ik}} - \lambda_i. \tag{G.4}$$

Moreover, by constraints on τ_{ik},

$$\sum_{k'} \tau_{ik'} = 1 \Longleftrightarrow \sum_{k'} \exp\left(T_{ik'}\right) \exp\left(-\lambda_i\right) = 1 \Longleftrightarrow \exp\left(-\lambda_i\right) = \frac{1}{\sum_{k'} \exp\left(T_{ik'}\right)} \tag{G.5}$$

Finally,

$$\tau_{ik} = \frac{\exp\left(T_{ik}\right)}{\sum_{k'} \exp\left(T_{ik'}\right)}. \tag{G.6}$$

■

G.2 Variational parameters of component membership ν_{vs}

$$\nu_{kjs} = \frac{\exp\Big(\log(\rho_{ks}) + \sum_{i=1}^{N} \tau_{ik} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right) - 1\Big)}{\sum_{s'} \exp\Big(\log(\rho_{ks'}) + \sum_{i=1}^{N} \tau_{ik} \log \mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right) - 1\Big)}. \tag{G.7}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$\begin{aligned} J_{\boldsymbol{\nu}}(.) = &\sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik}\nu_{kjs} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{kjs} \log\left(\rho_{ks}\right) \\ &- \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{kjs} \log\left(\nu_{kjs}\right) + \sum_{k=1}^{K}\sum_{j=1}^{p} \lambda_{kj} \left(1 - \sum_{s=1}^{Q} \rho_{ks}\right), \end{aligned} \tag{G.8}$$

where $(\lambda_{kj})_{k=1:K, j=1:p}$ are Lagrange parameters. Now, the aim is to get the gradient in ν_{vs} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_{\boldsymbol{\nu}}}{\partial \nu_{kjs}}(.) = \sum_{i=1}^{N} \tau_{ik} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \log\left(\rho_{ks}\right) - \log\left(\nu_{kjs}\right) - 1 - \lambda_{kj}. \tag{G.9}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\nu_{kjs}\right) = \underbrace{\sum_{i=1}^{N} \tau_{ik} \log\left(\mathbb{P}\left(X_{ij} \mid \mathbf{\Theta}_{ks}\right)\right) + \log\left(\rho_{ks}\right) - 1}_{R_{kjs}} - \lambda_{kj}. \tag{G.10}$$

Moreover, by constraints on ν_{kjs},

$$\sum_{s'} \nu_{kjs} = 1 \Longleftrightarrow \sum_{s'} \exp\left(R_{kjs'}\right) \exp\left(-\lambda_{kj}\right) = 1 \Longleftrightarrow \exp\left(-\lambda_{kj}\right) = \frac{1}{\sum_{s'} \exp\left(R_{kjs'}\right)}. \tag{G.11}$$

Finally,

$$\nu_{kjs} = \frac{\exp\left(R_{kjs}\right)}{\sum_{s'} \exp\left(R_{jks'}\right)}. \tag{G.12}$$

■

G.3 Parameters of distribution μ_{ks}

The partial derivative of the ELBO function with respect to μ_{ks} is given by:

$$\frac{\partial \mathcal{L}}{\partial \mu_{ks}}(.) = \sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} \frac{1}{2\sigma_{ks}^2} \left(X_{ij} - \mu_{ks}\right). \tag{G.13}$$

Setting the derivative to zero to find the estimator:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_{ks}}(.) &= 0 \\ \sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} \frac{1}{2\sigma_{ks}^2} \left(X_{ij} - \mu_{ks}\right) &= 0 \\ \sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} X_{ij} &= \sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} \mu_{ks}. \end{aligned} \tag{G.14}$$

Finally,

$$\hat{\mu}_{ks} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} X_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs}}. \tag{G.15}$$

G.4 Parameters of distribution σ_{ks}^2

The partial derivative of the ELBO function with respect to σ_{ks}^2 is given by:

$$\frac{\partial \mathcal{L}}{\partial \sigma_{ks}^2}(.) = \sum_{i=1}^{N} \sum_{j=1}^{p} \tau_{ik} \nu_{kjs} \left[-\frac{1}{2(\sigma_{ks}^2)^2} + \frac{1}{2\sigma_{ks}^2} \left(X_{ij} - \mu_{ks}\right)^2 \right]. \tag{G.16}$$

Setting the derivative to zero to find the estimator:

$$\frac{\partial \mathcal{L}}{\partial \sigma_{ks}^2}(.) = 0$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p} \tau_{ik}\nu_{kjs}\left[-\frac{1}{2\sigma_{ks}^2} + \frac{1}{2(\sigma_{ks}^2)^2}\left(X_{ij} - \mu_{ks}\right)^2\right] = 0$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p} \tau_{ik}\nu_{kjs} = \frac{1}{\sigma_{ks}^2}\sum_{i=1}^{N}\sum_{j=1}^{p} \tau_{ik}\nu_{kjs}\left(X_{ij} - \mu_{ks}\right)^2 . \tag{G.17}$$

Thus, the estimate of σ_{ks}^2 is obtained as:

$$\hat{\sigma}_{ks}^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{p} \tau_{ik}\nu_{kjs}\left(X_{ij} - \mu_{ks}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{p} \tau_{ik}\nu_{kjs}}. \tag{G.18}$$

H Details of Variation EM algorithm for Co-CoLBMoE

H.1 Objective function

Again, we want to maximise a lower bound evidence, which in our context is expressed as :

$$\mathcal{L}\left(q(\cdot)\right) = \sum_{\mathbf{G},\mathbf{Z},\mathbf{W}} q(\mathbf{G},\mathbf{Z},\mathbf{W}) \log \frac{\mathbb{P}\left(\mathbf{y},\mathbf{G},\mathbf{Z},\mathbf{W} \mid \mathbf{X}\right)}{q(\mathbf{G},\mathbf{Z},\mathbf{W})}. \tag{H.1}$$

and by a mean-field approximation, q is defined as :

$$\begin{aligned} q\left(\mathbf{G},\mathbf{Z},\mathbf{W}\right) &= \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{k=1}^{K}\prod_{j=1}^{p} q(\mathbf{W}_{kj}) \prod_{i=1}^{N} q(\mathbf{G}_i) \\ &= \prod_{i=1}^{N} \mathcal{M}\left(\mathbf{Z}_i; (\tau_{i1},\dots,\tau_{iK})\right) \prod_{k=1}^{K}\prod_{j=1}^{p} \mathcal{M}\left(\mathbf{W}_{kj}; (\nu_{kj1},\dots,\nu_{kjQ})\right) \\ &\prod_{i=1}^{N} \mathcal{M}\left(\mathbf{G}_i; (g_{i1},\dots,g_{iQ})\right) \end{aligned} \tag{H.2}$$

$$
\begin{aligned}
\mathcal{L}\left(q(\cdot)\right) = & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{W\sim q}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, G_{is}=1, \mathbf{W}_{k\bullet s}, \beta_{ks}\right)\right] \\
& + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \log\left(\frac{e^{\mathbf{X}_i \gamma_{k\bullet s}}}{\sum_{k'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}}\right) - \sum_{i=1}^{N}\sum_{s=1}^{Q} \tau_{ik} g_{is} \log(g_{is}) \\
& + \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{kjs} \log(\rho_{ks}) - \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{kjs} \log(\nu_{kjs}) \\
& + \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log\left(\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}^{(t)}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}^{(t)}}}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log(\tau_{ik})
\end{aligned}
\tag{H.3}
$$

To optimize all parameters, the Variational Expectation (VE) and Maximization (M) steps alternate successively until convergence is reached.

H.2 Community membership $\boldsymbol{\tau}$:

The optimal estimation of τ_{ik} is given by

$$
\hat{\tau}_{ik} = \frac{\exp\left(T_{ik}\right)}{\sum_{k'} \exp\left(T_{ik'}\right)}, \tag{H.4}
$$

with

$$
\begin{aligned}
T_{ik} = & \sum_{s=1}^{Q} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] \\
& + \log\left(\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}\right).
\end{aligned}
\tag{H.5}
$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$
\begin{aligned}
J_{\boldsymbol{\tau}}(.) = & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] \\
& + \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log\left(\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log\left(\tau_{ik}\right) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} \tau_{ik}\right),
\end{aligned}
\tag{H.6}
$$

where $(\lambda_i)_{i=1:N}$ are Lagrange parameters. Now, the aim is to get the gradient in τ_{ik} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_{\boldsymbol{\tau}}}{\partial \tau_{ik}}(.) = \sum_{s=1}^{Q} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] + \log\left(\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}\right) - \log\left(\tau_{ik}\right) - 1 - \lambda_i \tag{H.7}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\tau_{ik}\right) = \underbrace{\sum_{s=1}^{Q} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] + \log\left(\frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}\right) - 1}_{T_{ik}} - \lambda_i \tag{H.8}$$

Moreover, by constraints on τ_{ik},

$$\sum_{k'} \tau_{ik} = 1 \Longleftrightarrow \sum_{k'} \exp\left(T_{ik'}\right) \exp\left(-\lambda_i\right) = 1 \Longleftrightarrow \exp\left(-\lambda_i\right) = \frac{1}{\sum_{k'} \exp\left(T_{ik'}\right)} \tag{H.9}$$

Finally,

$$\hat{\tau}_{ik} = \frac{\exp\left(T_{ik}\right)}{\sum_{k'} \exp\left(T_{ik'}\right)} \tag{H.10}$$

■

H.3 Component membership $\boldsymbol{\nu}$:

$$\hat{\nu}_{kjs} = \frac{\exp\left(R_{kjs}\right)}{\sum_{s'} \exp\left(R_{kjs'}\right)}, \tag{H.11}$$

with

$$R_{kjs} = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{s=1}^{Q} \tau_{ik} g_{is} \frac{\partial}{\partial \nu_{kjs}} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] + \log\left(\rho_{ks}\right). \tag{H.12}$$

Optimization of (ν_{kjs}) is a fixed-point problem, and requires iterations until convergence to be optimal.

Proof

$$J_{\boldsymbol{\nu}}(.) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] \\ + \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{jks}\log\left(\rho_{ks}\right) - \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{s=1}^{Q} \nu_{jks}\log\left(\nu_{jks}\right) + \sum_{k=1}^{K}\sum_{j=1}^{p}\lambda_{kj}\left(1-\sum_{s=1}^{Q}\nu_{kjs}\right), \tag{H.13}$$

where $(\lambda_{ks})_{k=1:K,j=1:p}$ are Lagrange parameters.

Now, the aim is to get the gradient in ν_{kjs} to 0 and find an estimation while preserving the constraints.

$$\frac{\partial J_{\boldsymbol{\nu}}}{\partial \nu_{kjs}}(.) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \frac{\partial}{\partial \nu_{kjs}} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] \\ + \log\left(\rho_{ks}\right) - \log\left(\nu_{kjs}\right) - 1 - \lambda_{ks} \tag{H.14}$$

By setting the derivative to 0, the following equality is obtained

$$\log\left(\nu_{kjs}\right) = \\ \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \frac{\partial}{\partial \nu_{kjs}} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(.|\boldsymbol{\Theta})}\left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, G_{is}=1, \boldsymbol{\beta}_{ks}\right)\right] + \log\left(\rho_{ks}\right) - 1}_{R_{kjs}} \\ - \lambda_{ks}. \tag{H.15}$$

Moreover, by constraints on ν_{kjs},

$$\sum_{s'} \nu_{kjs} = 1 \Longleftrightarrow \sum_{s'} \exp\left(R_{kjs'}\right)\exp\left(-\lambda_{kj}\right) = 1 \Longleftrightarrow \exp\left(-\lambda_{kj}\right) = \frac{1}{\sum_{s'}\exp\left(R_{kjs'}\right)}. \tag{H.16}$$

Finally,

$$\hat{\nu}_{kjs} = \frac{\exp\left(R_{kjs}\right)}{\sum_{s'}\exp\left(R_{kjs'}\right)}. \tag{H.17}$$

Optimization of (ν_{kjs}) is a fixed-point problem, and requires iterations until convergence to be optimal. ■

H.4 Expert selection g:

$$\hat{g}_{is} = \frac{\exp\left(C_{is}\right)}{\sum_{s'} \exp\left(C_{is'}\right)}, \tag{H.18}$$

with

$$\begin{aligned} C_{is} = &\sum_{k=1}^{K} \tau_{ik} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(.|\mathbf{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_{k\bullet s}, G_{is} = 1, \boldsymbol{\beta}_{ks}\right)\right] \\ &+ \sum_{k=1}^{K} \tau_{ik} \log\left(\frac{e^{\mathbf{X}_i \gamma_{k\bullet s}}}{\sum_{k'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}}\right). \end{aligned} \tag{H.19}$$

Proof

$$\begin{aligned} J_{\mathbf{g}}(.) = &\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(.|\mathbf{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_{k\bullet s}, G_{is} = 1, \boldsymbol{\beta}_{ks}\right)\right] \\ &+ \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \log\left(\gamma_{ks}\right) - \sum_{i=1}^{N}\sum_{s=1}^{Q} g_{is} \log\left(g_{is}\right) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{s=1}^{Q} g_{is}\right), \end{aligned} \tag{H.20}$$

where $(\lambda_i)_{i=1:N}$ are Lagrange parameters. Now, the aim is to get the gradient in g_{is} to 0 and find an estimation while preserving the constraints.

$$\begin{aligned} \frac{\partial J_{\mathbf{g}}}{\partial g_{is}}(.) = &\sum_{k=1}^{K} \tau_{ik} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(.|\mathbf{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_{k\bullet s}, G_{is} = 1, \boldsymbol{\beta}_{ks}\right)\right] \\ &+ \sum_{k=1}^{K} \tau_{ik} \log\left(\gamma_{ks}\right) - \log\left(g_{is}\right) - 1 - \lambda_i \end{aligned} \tag{H.21}$$

By setting the derivative to 0, the following equality is obtained

$$\begin{aligned} &\log\left(g_{is}\right) \\ &= \underbrace{\sum_{k=1}^{K} \tau_{ik} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(.|\mathbf{\Theta})} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_{k\bullet s}, G_{is} = 1, \boldsymbol{\beta}_{ks}\right)\right] + \sum_{k=1}^{K} \tau_{ik} \log\left(\gamma_{ks}\right) - 1}_{C_{is}} \\ &- \lambda_i \end{aligned} \tag{H.22}$$

Moreover, by constraints on g_{is},

$$\sum_{s'} g_{is} = 1 \Longleftrightarrow \sum_{s'} \exp\left(C_{is'}\right) \exp\left(-\lambda_i\right) = 1 \Longleftrightarrow \exp\left(-\lambda_i\right) = \frac{1}{\sum_{s'} \exp\left(C_{is'}\right)} \tag{H.23}$$

Finally,

$$\hat{g}_{is} = \frac{\exp\left(C_{is}\right)}{\sum_{s'} \exp\left(C_{is'}\right)} \tag{H.24}$$

■

H.5 Community proportions $\boldsymbol{\rho}$:

The optimal estimation of ρ_{ks} is given by :

$$\hat{\rho}_{ks} = \frac{\sum_{j=1}^{p} \nu_{kjs}}{p}. \tag{H.25}$$

Proof First, we start by defining the Lagrangian function of the optimization problem :

$$J_{\boldsymbol{\rho}}(.) = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{s=1}^{Q} \nu_{kjs} \log\left(\rho_{ks}\right) + \lambda_k \left(1 - \sum_{s=1}^{Q} \rho_{ks}\right), \tag{H.26}$$

where $(\lambda_k)_{k=1:K}$ is Lagrange parameter.

$$\frac{\partial J_{\boldsymbol{\rho}}}{\partial \rho_{ks}}(.) = \frac{\sum_{j=1}^{p} \nu_{kjs}}{\rho_{ks}} - \lambda_k. \tag{H.27}$$

Moreover, by constraints on $\boldsymbol{\rho}$,

$$\sum_{s=1}^{Q} \rho_{ks} = 1 \Longleftrightarrow \sum_{s=1}^{Q} \frac{\sum_{j=1}^{p} \nu_{kjs}}{\lambda_k} = 1 \Longleftrightarrow \lambda_k = \sum_{s=1}^{Q} \sum_{j=1}^{p} \nu_{kjs} \Longleftrightarrow \lambda_k = p. \tag{H.28}$$

Finally,

$$\rho_{ks} = \frac{\sum_{j=1}^{p} \nu_{kjs}}{p}. \tag{H.29}$$

■

H.6 Community membership regression parameters $\boldsymbol{\pi}$:

Regarding the logistic regression parameters, denoted as $\boldsymbol{\pi}$, on the latent variables $(\mathbf{Z}_i)_{i=1:N}$, we have the following:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\pi}}(.) = \mathbf{X}^T (\boldsymbol{\tau} - \mathbf{S}^{\boldsymbol{\pi}}), \tag{H.30}$$

where

$$S_{ik}^{\boldsymbol{\pi}} = \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_{k'} e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k'}}}. \tag{H.31}$$

The parameters are then updated according to the following rule:

$$\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)} + h_t \frac{\partial \mathcal{L}}{\partial \boldsymbol{\pi}}(.), \tag{H.32}$$

where h_t is the gradient ascent step size at iteration t.

Proof

$$\begin{aligned}
\mathcal{L}(\cdot) &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \mathbf{X}_i \boldsymbol{\pi}_{\bullet k} - \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log \left(\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \mathbf{X}_i \boldsymbol{\pi}_{\bullet k} - \sum_{i=1}^{N} \log \left(\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}} \right)
\end{aligned} \tag{H.33}$$

Next, to derive the gradient of $\mathcal{J}(\cdot)$ with respect to the logistic regression parameters π_{jk}, we compute the partial derivative:

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \pi_{jk}}(\cdot) &= \sum_{i=1}^{N} \tau_{ik} X_{ij} - \sum_{i=1}^{N} X_{ij} \frac{e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet k}}}{\sum_l e^{\mathbf{X}_i \boldsymbol{\pi}_{\bullet l}}} \\
&= \sum_{i=1}^{N} X_{ij} \left(\tau_{ik} - S_{ik}^{\boldsymbol{\pi}} \right)
\end{aligned} \tag{H.34}$$

Finally, writing this in matrix form, we obtain the gradient with respect to the entire parameter matrix $\boldsymbol{\pi}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\pi}}(\cdot) = \mathbf{X}^T (\boldsymbol{\tau} - \mathbf{S}^{\boldsymbol{\pi}}). \tag{H.35}$$

■

H.7 Expert selection regression parameters $\boldsymbol{\gamma}$:

For the logistic regression parameters $\boldsymbol{\gamma}$ on the latent variables $(\mathbf{G}_i)_{i=1}$, the gradient of the log-likelihood function with respect to $\boldsymbol{\gamma}_k$ is given by:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}_k}(.) = \left[\mathbf{X} \odot (\boldsymbol{\tau}_{\bullet k} \mathbf{1}_{1,p})\right]^T (\mathbf{g} - \mathbf{S}^{\gamma_k}), \tag{H.36}$$

where $\mathbf{1}1, p$ is a $p \times 1$ matrix filled with ones, and $S_{is}^{\gamma_k}$ is the softmax probability for the i-th individual and s-th class, given by:

$$S_{is}^{\gamma_k} = \frac{e^{\mathbf{X}_i \gamma_{k\bullet s}}}{\sum_{k'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}}. \tag{H.37}$$

This gradient represents the difference between the posterior probability $\mathbf{g}$ and the predicted probabilities from the softmax function, weighted by the latent responsibilities $\boldsymbol{\tau}_{\bullet k}$. Next, the parameter update is performed as follows:

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + h_t \frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}}(.), \tag{H.38}$$

where h_t is the gradient ascent step size at iteration t.

Proof

$$\begin{aligned}
\mathcal{L}(\cdot) \propto & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \log\left(\frac{e^{\mathbf{X}_i \gamma_{k\bullet s}}}{\sum_{k'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}}\right) \\
= & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \left(\mathbf{X}_i \boldsymbol{\gamma}_{k\bullet s}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \log\left(\sum_{s'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}\right) \\
= & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \left(\mathbf{X}_i \boldsymbol{\gamma}_{k\bullet s}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log\left(\sum_{s'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}\right)
\end{aligned} \tag{H.39}$$

Next, to derive the gradient of $\mathcal{J}(\cdot)$ with respect to the logistic regression parameters γ_{kjs}, we compute the partial derivative:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \gamma_{kjs}}(\cdot) &= \sum_{i=1}^{N} \tau_{ik} g_{is} X_{ij} - \sum_{i=1}^{N} X_{ij} \frac{e^{\mathbf{X}_i \gamma_{k\bullet s}}}{\sum_{s'} e^{\mathbf{X}_i \gamma_{k\bullet s'}}} \\ &= \sum_{i=1}^{N} \tau_{ik} X_{ij} \left(g_{is} - S_{is}^{\gamma_k} \right) \end{aligned} \tag{H.40}$$

Finally, writing this in matrix form, we obtain the gradient with respect to the entire parameter matrix $\boldsymbol{\gamma}_k$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}_k}(.) = \left[\mathbf{X} \odot \left(\boldsymbol{\tau}_{\bullet k} \mathbf{1}_{1,p}\right)\right]^T \left(\mathbf{g} - \mathbf{S}^{\gamma_k}\right). \tag{H.41}$$

■

H.8 Regression parameters : $\boldsymbol{\beta}$

The estimate of the parameter $\boldsymbol{\beta}_{ks}$ is derived using the following formula:

$$\hat{\boldsymbol{\beta}}_{ks} = \left[\left(\mathbf{X}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k} \odot \mathbf{g}_{\bullet s}\right) \mathbf{X}\right) \odot \mathbf{U}_{ks}\right]^{-1} \operatorname{diag}\left(\boldsymbol{\nu}_{k\bullet s}\right) \mathbf{X}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k} \odot \mathbf{g}_{\bullet s}\right) \mathbf{y}, \tag{H.42}$$

where $\mathbf{U}_{ks}$ is given by:

$$\mathbf{U}_{ks} = \boldsymbol{\nu}_{k\bullet s} \boldsymbol{\nu}_{k\bullet s}^T - \operatorname{diag}(\boldsymbol{\nu}_{k\bullet s}^2) + \operatorname{diag}(\boldsymbol{\nu}_{k\bullet s}). \tag{H.43}$$

In this expression, $\hat{\boldsymbol{\beta}}_{ks}$ represents the coefficient estimates for the k-th community and s-th components. The term $\boldsymbol{\tau}_{\bullet k} \odot \mathbf{g}_{\bullet s}$ captures the element-wise interaction between the latent community and the expert selection, while $\mathbf{U}_{ks}$ encodes the covariance structure of the latent components.

Proof The objective function $\mathcal{L}$ is proportional to:

$$\begin{aligned}\mathcal{L}(\cdot) &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(\cdot)} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik} = 1, \mathbf{W}_{k\bullet s}, \boldsymbol{\beta}_k^{(t)}\right)\right] \\ &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s} \sim q(\cdot)} \left[-\frac{1}{2\sigma_{ks}^2} \left(y_i - \mathbf{X}_i \operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right) \boldsymbol{\beta}_{ks}\right)^2\right]. \end{aligned} \tag{H.44}$$

By rewriting this in matrix form, we get:

$$\begin{aligned}\mathcal{L}(\cdot) \propto \sum_{k=1}^{K}\sum_{s=1}^{Q} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)} \Bigg[& -\frac{1}{2\sigma_{ks}^2}\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{y} \\ & +\frac{1}{2\sigma_{ks}^2} \times 2\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks} \\ & -\frac{1}{2\sigma_{ks}^2}\left(\mathbf{X}\operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks}\right)^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks}\Bigg]\end{aligned} \tag{H.45}$$

. Replacing the expectation term $\mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)}$ by its known form, we obtain:

$$\begin{aligned}\mathcal{L}(\cdot) \propto & \sum_{k=1}^{K}\sum_{s=1}^{Q} \Bigg[-\frac{1}{2\sigma_{ks}^2}\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{y} \\ & +\frac{1}{2\sigma_{ks}^2} \times 2\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\operatorname{diag}\left(\boldsymbol{\nu}_{k\bullet s}\right)\boldsymbol{\beta}_{ks} \\ & -\frac{1}{2\sigma_{ks}^2}\boldsymbol{\beta}_{ks}^T\mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)}\left[\operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\mathbf{X}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\right]\boldsymbol{\beta}_{ks}\Bigg] \\ \propto & \sum_{k=1}^{K}\sum_{s=1}^{Q} \Bigg[-\frac{1}{2\sigma_{ks}^2}\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{y} + \frac{1}{2\sigma_{ks}^2} \times 2\mathbf{y}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\operatorname{diag}\left(\boldsymbol{\nu}_{k\bullet s}\right)\boldsymbol{\beta}_{ks} \\ & -\frac{1}{2\sigma_{ks}^2}\boldsymbol{\beta}_{ks}^T\left\{\left(\mathbf{X}^T \operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\right)\odot\underbrace{\left(\boldsymbol{\nu}_{k\bullet s}\boldsymbol{\nu}_{k\bullet s}^T - \operatorname{diag}(\boldsymbol{\nu}_{k\bullet s}^2) + \operatorname{diag}(\boldsymbol{\nu}_{k\bullet s})\right)}_{\mathbf{U}_{ks}}\right\}\boldsymbol{\beta}_{ks}\Bigg]\end{aligned} \tag{H.46}$$

. Now, taking the derivative of the objective function with respect to $\boldsymbol{\beta}_{ks}$ yields:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}_{ks}}(\cdot) = \frac{1}{\sigma_{ks}^2}\operatorname{diag}\left(\boldsymbol{\nu}_{k\bullet s}\right)\mathbf{X}^T\operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{y} - \frac{1}{\sigma_{ks}^2}\left[\left(\mathbf{X}^T\operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\right)\odot\mathbf{U}_{ks}\right]\boldsymbol{\beta}_{ks} \tag{H.47}$$

Setting this gradient equal to zero leads to the update for $\hat{\boldsymbol{\beta}}_{ks}$:

$$\hat{\boldsymbol{\beta}}_{ks} = \left[\left(\mathbf{X}^T\operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{X}\right)\odot\mathbf{U}_{ks}\right]^{-1}\operatorname{diag}\left(\boldsymbol{\nu}_{k\bullet s}\right)\mathbf{X}^T\operatorname{diag}\left(\boldsymbol{\tau}_{\bullet k}\odot \mathbf{g}_{\bullet s}\right)\mathbf{y} \tag{H.48}$$

It is important to note that we would have obtained the standard weighted formula for linear regression estimation if the second moment of a categorical distribution equaled the parameter itself, rather than its square. ■

H.9 Noise σ^2_{ks}

Proof The objective function $\mathcal{L}$ is proportional to:

$$\begin{aligned}\mathcal{L}(\cdot) &\propto \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)} \left[\log \mathbb{P}\left(y_i \mid \mathbf{X}_i, Z_{ik}=1, \mathbf{W}_{k\bullet s}, \boldsymbol{\beta}_k^{(t)}\right)\right] \\ &\propto \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \left(-\frac{1}{2}\log(\sigma^2_{ks}) - \frac{1}{2\sigma^2_{ks}} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)} \left[\left(y_i - \mathbf{X}_i \operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks}\right)^2\right]\right).\end{aligned} \tag{H.49}$$

■

Now, taking the derivative of the objective function with respect to σ^2_{ks} yields:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2_{ks}}(\cdot) = \sum_{i=1}^{N} \tau_{ik} g_{is} \left(-\frac{1}{2\sigma^2_{ks}} - \frac{1}{2\sigma^4_{ks}} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)} \left[\left(y_i - \mathbf{X}_i \operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks}\right)^2\right]\right). \tag{H.50}$$

Setting this gradient equal to zero leads to the update for $\hat{\sigma}^2_{ks}$:

$$\hat{\sigma}^2_{ks} = \frac{\sum_{i=1}^{N} \tau_{ik} g_{is} \mathbb{E}_{\mathbf{W}_{k\bullet s}\sim q(\cdot)} \left[\left(y_i - \mathbf{X}_i \operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right)\boldsymbol{\beta}_{ks}\right)^2\right]}{\sum_{i=1}^{N} \tau_{ik} g_{is}}. \tag{H.51}$$

H.10 Multinomial logistic regression by SEM-Gibbs algorithm W $\boldsymbol{\beta}$

In this context, applying a Variational Expectation Maximization algorithm is not feasible due to the complexity it introduces in optimizing the problem. Therefore, we opted for a SEM-Gibbs algorithm to perform the optimization, which allows us to sample from the posterior distribution and compute the expectation over the latent variables.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_{ks}} = \left[\mathbf{X} \odot \left(\left[\boldsymbol{\tau}_{\bullet k} \odot \mathbf{g}_{\bullet s}\right] \hat{\mathbf{W}}^T_{k\bullet s}\right)\right]^T \left(\mathbf{y} - S^{\beta_{ks}}\right), \tag{H.52}$$

where

$$S_{ic}^{\boldsymbol{\beta}_{ks}} = \frac{\exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c}\right]}{\sum_{c'=1}^{C} \exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c'}\right]}, \tag{H.53}$$

and we sample $\mathbf{W}_{kj}$ as follows:

$$\hat{\mathbf{W}}_{kj} \leftsquigarrow \mathcal{M}\left(1; \left(\nu_{kj1}, \ldots, \nu_{kjQ}\right)\right). \tag{H.54}$$

Proof We begin by expressing the objective function in proportion to $\boldsymbol{\beta}$:

$$\begin{aligned}\mathcal{L}(\cdot) &\propto \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \sum_{c=1}^{C} y_{ic} \log\left[\frac{\exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c}\right]}{\sum_{c'=1}^{C} \exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c'}\right]}\right]\\ &\propto \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{Q} \tau_{ik} g_{is} \left\{\sum_{c=1}^{C} y_{ic}\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c} - \log\left[\sum_{c'=1}^{C} \exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c'}\right]\right]\right\}\end{aligned} \tag{H.55}$$

Next, we derive the gradient of the log-likelihood with respect to β_{ksjc}:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_{ksjc}}(\cdot) &= \sum_{i=1}^{N} \tau_{ik} g_{is} \left(y_{ic} X_{ij} \hat{W}_{kjs} - X_{ij} \hat{W}_{kjs} \frac{\exp\left[\mathbf{X}_i \operatorname{diag}\left(\hat{\mathbf{W}}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c}\right]}{\sum_{c'=1}^{C} \exp\left[\mathbf{X}_i \operatorname{diag}\left(\mathbf{W}_{k\bullet s}\right) \boldsymbol{\beta}_{ks\bullet c'}\right]}\right)\\ &= \sum_{i=1}^{N} \tau_{ik} g_{is} X_{ij} \hat{W}_{kjs} \left(y_{ic} - S_{ic}^{\boldsymbol{\beta}_{ks}}\right).\end{aligned} \tag{H.56}$$

Finally, we obtain the expression for the gradient with respect to $\boldsymbol{\beta}_{ks}$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_{ks}} = \left[\mathbf{X} \odot \left(\left[\boldsymbol{\tau}_{\bullet k} \odot \mathbf{g}_{\bullet s}\right] \hat{\mathbf{W}}_{k\bullet s}^{T}\right)\right]^{T} \left(\mathbf{y} - S^{\boldsymbol{\beta}_{ks}}\right). \tag{H.57}$$

The parameters are then updated according to the following rule:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + h_t \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}}(.), \tag{H.58}$$

where h_t is the gradient ascent step size at iteration t. ■

Bibliography

Acosta, J. N., G. J. Falcone, P. Rajpurkar, and E. J. Topol (2022). Multimodal biomedical ai. Nature Medicine 28(9), 1773–1784. 17

Airoldi, E. M., D. Blei, S. Fienberg, and E. Xing (2008). Mixed membership stochastic blockmodels. Advances in neural information processing systems 21. 34

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike, pp. 199–213. Springer. 69

Athaya, T., R. C. Ripan, X. Li, and H. Hu (2023). Multimodal deep learning approaches for single-cell multi-omics data integration. Briefings in Bioinformatics 24(5), bbad313. 22

Attias, H. (1999). A variational baysian framework for graphical models. Advances in neural information processing systems 12. 63

Baltrušaitis, T., C. Ahuja, and L.-P. Morency (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443. 17, 22, 23, 28

Barbillon, P., S. Donnet, E. Lazega, and A. Bar-Hen (2017). Stochastic block models for multiplex networks: an application to a multilevel network of researchers. Journal of the Royal Statistical Society Series A: Statistics in Society 180(1), 295–314. 56

Baudry, J.-P. and G. Celeux (2015). Em for mixtures: Initialization requires special care. Statistics and computing 25, 713–726. 68

Bernardo, J., M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West, et al. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian statistics 7(453-464), 210. 45

Bertsimas, D. and Y. Ma (2024). M3h: Multimodal multitask machine learning for healthcare. arXiv preprint arXiv:2404.18975. 18

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. arXiv preprint arXiv:1701.02434. 48

Betancourt, M. J. and M. Girolami (2013). Hamiltonian monte carlo for hierarchical models, arxiv. arXiv preprint arXiv:1312.0906 3. 48

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. 36

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and newman–girvan and other modularities. Proceedings of the National Academy of Sciences 106(50), 21068–21073. 36

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence 22(7), 719–725. 29, 36, 69, 161

Biernacki, C., G. Celeux, and G. Govaert (2010). Exact and monte carlo calculations of integrated likelihoods for the latent class model. Journal of Statistical Planning and Inference 140(11), 2991–3002. 52, 56, 69

Biernacki, C. and J. Jacques (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing 26, 929–943. 38

Biernacki, C., J. Jacques, and C. Keribin (2023). A survey on model-based co-clustering: high dimension and estimation challenges. Journal of Classification 40(2), 332–381. 40

Biffi, A., C. D. Anderson, R. S. Desikan, M. Sabuncu, L. Cortellini, N. Schmansky, D. Salat, J. Rosand, A. D. N. I. (ADNI, et al. (2010). Genetic variation and neuroimaging measures in alzheimer disease. Archives of neurology 67(6), 677–685. 18

Billot, D. (2010, September). Résultats de la prise en charge des traumatismes sonores aigus : à propos de 225 patients. Master's thesis, UHP - Université Henri Poincaré. 2

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer google schola 2, 1122–1128. 46

Bishop, C. M. (2013). Model-based machine learning. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371(1984), 20120222. 29

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. Journal of the American statistical Association 112(518), 859–877. 44, 46

Boehm, K. M., E. A. Aherne, L. Ellenson, I. Nikolovski, M. Alghamdi, I. Vázquez-García, D. Zamarin, K. Long Roche, Y. Liu, D. Patel, et al. (2022). Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer 3(6), 723–733. 19

Boehm, K. M., P. Khosravi, R. Vanguri, J. Gao, and S. P. Shah (2022). Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer 22(2), 114–126. 18

Borenstein, M., L. V. Hedges, J. P. Higgins, and H. R. Rothstein (2021). Introduction to meta-analysis. John Wiley & Sons. 20

Boutalbi, R., L. Labiod, and M. Nadif (2020). Tensor latent block model for co-clustering. International Journal of Data Science and Analytics 10(2), 161–175. 115

Boutalbi, R., L. Labiod, and M. Nadif (2021). Implicit consensus clustering from multiple graphs. Data Mining and Knowledge Discovery 35, 2313–2340. 56

Boutalbi, R., L. Labiod, and M. Nadif (2022). Latent block regression model. In Conference of the International Federation of Classification Societies, pp. 73–81. Springer International Publishing Cham. 92

Boutin, R., P. Latouche, and C. Bouveyron (2023). The deep latent position topic model for clustering and representation of networks with textual edges. arXiv preprint arXiv:2304.08242. 114

Bouveyron, C., L. Bozzi, J. Jacques, and F.-X. Jollois (2018). The functional latent block model for the co-clustering of electricity consumption curves. Journal of the Royal Statistical Society Series C: Applied Statistics 67(4), 897–915. 38, 39

Brault, V. (2014). Estimation et sélection de modèle pour le modèle des blocs latents. Ph. D. thesis, Université Paris Sud-Paris XI. 60, 63

Brault, V., C. Keribin, and M. Mariadassou (2020). Consistency and asymptotic normality of latent block model estimators. 36

Brault, V. and M. Mariadassou (2015). Co-clustering through latent bloc model: A review. Journal de la Société Française de Statistique 156(3), 120–139. 38, 161

Briere, G. (2022). Développements méthodologiques pour l'intégration de données omiques: applications à l'oncologie et aux neurosciences. Ph. D. thesis, Bordeaux. 19, 22, 24

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (2011). Handbook of markov chain monte carlo. CRC press. 48

Cai, W., J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang (2024). A survey on mixture of experts. arXiv preprint arXiv:2407.06204. 89, 90

Calaon, M., T. Chen, and G. Tosello (2024). Integration of multimodal data and explainable artificial intelligence for root cause analysis in manufacturing processes. CIRP Annals. 21

Cantini, L., P. Zakeri, C. Hernandez, A. Naldi, D. Thieffry, E. Remy, and A. Baudot (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature communications 12(1), 124. 19

Casella, G. and E. I. George (1992). Explaining the gibbs sampler. The American Statistician 46(3), 167–174. 46

Celeux, G. and G. Govaert (1992). A classification em algorithm for clustering and two stochastic versions. Computational statistics & Data analysis 14(3), 315–332. 70

Celisse, A., J.-J. Daudin, and L. Pierre (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. 34, 36, 59, 119

Chamroukhi, F. and B.-T. Huynh (2019). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. Journal de la société française de statistique 160(1), 57–85. 89

Chen, P.-Y. and A. O. Hero (2017). Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms. IEEE Transactions on Signal and Information Processing over Networks 3(3), 553–567. 56

Chen, S., B. Zhu, S. Huang, J. W. Hickey, K. Z. Lin, M. Snyder, W. J. Greenleaf, G. P. Nolan, N. R. Zhang, and Z. Ma (2024). Integration of spatial and single-cell data across modalities with weakly linked features. Nature Biotechnology 42(7), 1096–1106. 22

Clark, A., D. de Las Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. Hechtman, T. Cai, S. Borgeaud, et al. (2022). Unified scaling laws for routed language models. In International conference on machine learning, pp. 4057–4086. PMLR. 42

Côme, E. and P. Latouche (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Statistical Modelling 15(6), 564–589. 36, 70

Cornuéjols, A., C. Wemmert, P. Gançarski, and Y. Bennani (2018). Collaborative clustering: Why, when, what and how. Information Fusion 39, 81–95. 17

Courbariaux, M., K. De Santiago, C. Dalmasso, F. Danjou, S. Bekadar, J.-C. Corvol, M. Martinez, M. Szafranski, and C. Ambroise (2022). A sparse mixture-of-experts model with screening of genetic associations to guide disease subtyping. Frontiers in Genetics 13, 859462. 89, 92, 95

Cui, C., H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering 5(2), 022001. 25

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. Statistics and computing 18(2), 173–183. 34, 36, 55

De Domenico, M., V. Nicosia, A. Arenas, and V. Latora (2015). Structural reducibility of multilayer networks. Nature communications 6(1), 6864. 81

De Santiago, K., M. Szafranski, and C. Ambroise (2023). Mixture of stochastic block models for multiview clustering. In ESANN 2023-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 151–156. 7

De Santiago, K., M. Szafranski, and C. Ambroise (2024a). Intégration tardive de données multimodales par modèles à blocs stochastiques. In 55e Journées de Statistique de la SFdS. 7

De Santiago, K., M. Szafranski, and C. Ambroise (2024b). mimiSBM: Mixture of Multilayer Integrator Stochastic Block Models. R package version 0.0.1.3. 7, 73

De Santiago, K., M. Szafranski, and C. Ambroise (2024c). Mixture of multilayer stochastic block models for multiview clustering. arXiv preprint arXiv:2401.04682. 7

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39(1), 1–22. 30, 44

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. Journal of classification 5, 249–282. 91

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. Physics letters B 195(2), 216–222. 48

El-Amraoui, A. and C. Petit (2010). Thérapie cellulaire dans l'oreille interne-nouveaux développements et perspectives. médecine/sciences 26(11), 981–985. 6

Fan, X., M. Pensky, F. Yu, and T. Zhang (2022). Alma: alternating minimization algorithm for clustering mixture multilayer network. The Journal of Machine Learning Research 23(1), 14855–14900. 57

Fedus, W., B. Zoph, and N. Shazeer (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research 23(120), 1–39. 42, 91

Fienberg, S. E., E. P. Xing, and T. Jaakkola (2008). Mixed membership stochastic blockmodels. 36

Fox, C. W. and S. J. Roberts (2012). A tutorial on variational bayesian inference. Artificial intelligence review 38, 85–95. 44

Franke, A., D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. Nature genetics 42(12), 1118–1125. 20

Fred, A. L. and A. K. Jain (2005). Combining multiple clusterings using evidence accumulation. IEEE transactions on pattern analysis and machine intelligence 27(6), 835–850. 54

García-Cortés, L. and D. Sorensen (1996). On a multivariate implementation of the gibbs sampler. Genetics Selection Evolution 28(1), 121–126. 47

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. Smith (1990). Illustration of bayesian inference in normal data models using gibbs sampling. Journal of the American Statistical Association 85(412), 972–985. 98

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). Bayesian data analysis. Chapman and Hall/CRC. 47

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (6), 721–741. 46

Gilks, W. R., S. Richardson, and D. Spiegelhalter (1995). Markov chain Monte Carlo in practice. CRC press. 44

Glattke, T. J. and S. G. Kujawa (1991). Otoacoustic emissions. American Journal of Audiology 1(1), 29–40. 5

Goffinet, E., A. Coutant, M. Lebbah, H. Azzag, and L. Giraldi (2020). Conditional latent block model: a multivariate time series clustering approach for autonomous driving validation. arXiv preprint arXiv:2008.00946. 39, 94, 95, 101, 157, 161

Goffinet, E., M. Lebbah, H. Azzag, A. Coutant, and L. Giraldi (2021). A new multivariate time series co-clustering non-parametric model applied to driving-assistance systems validation. In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases-Workshop on Advanced Analytics and Learning on Temporal Data A modern web site. 39

Golalipour, K., E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar (2021). From clustering to clustering ensemble selection: A review. Engineering Applications of Artificial Intelligence 104, 104388. 54

Gormley, I. C. and S. Frühwirth-Schnatter (2019). Mixture of experts models. In Handbook of mixture analysis, pp. 271–307. Chapman and Hall/CRC. 41, 43, 88

Govaert, G. and M. Nadif (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis 52(6), 3233–3245. 38

Govaert, G. and M. Nadif (2010). Latent block model for contingency table. Communications in Statistics—Theory and Methods 39(3), 416–425. 36

Guarrasi, V., F. Aksu, C. M. Caruso, F. D. Feola, A. Rofena, F. Ruffini, and P. Soda (2024). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. 26

Gurevitch, J., J. Koricheva, S. Nakagawa, and G. Stewart (2018). Meta-analysis and the science of research synthesis. Nature 555(7695), 175–182. 20

Han, Q., K. Xu, and E. Airoldi (2015). Consistent estimation of dynamic and multi-layer block models. In International Conference on Machine Learning, pp. 1511–1520. PMLR. 56

Han, R., Y. Luo, M. Wang, and A. R. Zhang (2022). Exact clustering in tensor block model: Statistical optimality and computational limit. Journal of the Royal Statistical Society Series B: Statistical Methodology 84(5), 1666–1698. 57

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. 47

Hinton, G. E. and R. Zemel (1993). Autoencoders, minimum description length and helmholtz free energy. Advances in neural information processing systems 6. 26

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. Journal of the american Statistical association 97(460), 1090–1098. 35, 114

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. Social networks 5(2), 109–137. 31

Huang, F., X. Zhang, and Z. Li (2018). Learning joint multimodal representation with adversarial attention networks. In Proceedings of the 26th ACM international conference on Multimedia, pp. 1874–1882. 29

Huang, S., H. Weng, and Y. Feng (2022). Spectral clustering via adaptive layer aggregation for multi-layer networks. Journal of Computational and Graphical Statistics, 1–15. 56

Hubert, L. and P. Arabie (1985). Comparing partitions. Journal of classification 2, 193–218. 73

Ismail, A. A., S. Ö. Arik, J. Yoon, A. Taly, S. Feizi, and T. Pfister (2022). Interpretable mixture of experts. arXiv preprint arXiv:2206.02107. 88, 91

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. Neural computation 3(1), 79–87. 41

Jacques, J. and C. Biernacki (2018). Model-based co-clustering for ordinal data. Computational Statistics & Data Analysis 123, 101–115. 38

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 186(1007), 453–461. 63

Jiang, A. Q., A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088. 89

Jing, B.-Y., T. Li, Z. Lyu, and D. Xia (2021). Community detection on mixture multilayer networks via regularized tensor decomposition. The Annals of Statistics 49(6), 3181–3205. 57, 73, 81, 82, 84, 85

John, C. R., D. Watson, D. Russ, K. Goldmann, M. Ehrenstein, C. Pitzalis, M. Lewis, and M. Barnes (2020). M3c: Monte carlo reference-based consensus clustering. Scientific reports 10(1), 1816. 54, 55, 73

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

Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. Neural computation 6(2), 181–214. 89

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics 83(1), 016107. 35

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. Journal of the American statistical Association 91(435), 1343–1370. 62

Keribin, C. (2010). Méthodes bayésiennes variationnelles: concepts et applications en neuroimagerie. Journal de la Société Française de Statistique 151(2), 107–131. 63

Keribin, C., V. Brault, G. Celeux, and G. Govaert (2015). Estimation and selection for the latent block model on categorical data. Statistics and Computing 25(6), 1201–1216. 59

Keribin, C., V. Brault, G. Celeux, G. Govaert, et al. (2012). Model selection for the binary latent block model. In Proceedings of COMPSTAT, Volume 2012. 98

Keribin, C., G. Celeux, and V. Robert (2017). The latent block model: a useful model for high dimensional data. In ISI 2017-61st world statistics congress, pp. 1–6. 36, 37

Khan, M., W. Gueaieb, A. El Saddik, and S. Kwon (2024). Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications 245, 122946. 29

Kingma, D. P. and M. Welling (2022). Auto-encoding variational bayes. 27

Kline, A., H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo (2022). Multimodal machine learning in precision health: A scoping review. npj Digital Medicine 5(1), 171. 22

Lachaux, J., P. A. Giéré, Q. Vuillemin, T. Colléony, A. Crambert, S. Siegrist, C. Parietti-Winkler, P.-É. Schwartzbrod, and G. Andéol (2024). Long-term hearing loss after acute acoustic trauma in the french military: a retrospective study. Military medicine 189(3-4), e698–e704. 2

Latouche, P., E. Birmelé, and C. Ambroise (2011). Overlapping stochastic block models with application to the french political blogosphere. 35, 45

Latouche, P., E. Birmele, and C. Ambroise (2012). Variational bayesian inference and complexity control for stochastic block models. Statistical Modelling 12(1), 93–115. 45, 62, 65

Lauriola, I., C. Gallicchio, and F. Aiolli (2020). Enhancing deep neural networks via multiple kernel learning. Pattern Recognition 101, 107194. 26

Lee, C. and D. J. Wilkinson (2019). A review of stochastic block models and extensions for graph clustering. Applied Network Science 4(1), 1–50. 36

Leger, J.-B., J.-J. Daudin, and C. Vacher (2015). Clustering methods differ in their ability to detect patterns in ecological networks. Methods in Ecology and Evolution 6(4), 474–481. 31

Lepikhin, D., H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668. 42, 91

Li, Y., F. Nie, H. Huang, and J. Huang (2015). Large-scale multi-view spectral clustering via bipartite graph. In Proceedings of the AAAI conference on artificial intelligence, Volume 29. 52

Liang, D., M. Corneli, C. Bouveyron, and P. Latouche (2024). Clustering by deep latent position model with graph convolutional network. Advances in Data Analysis and Classification, 1–34. 114

Liang, P. P., A. Zadeh, and L.-P. Morency (2022). Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. arXiv preprint arXiv:2209.03430. 22, 24

Liu, L., F. Nie, A. Wiliem, Z. Li, T. Zhang, and B. C. Lovell (2018). Multimodal joint clustering with application for unsupervised attribute discovery. IEEE Transactions on Image Processing 27(9), 4345–4356. 52

Lomet, A. (2012). Sélection de modèle pour la classification croisée de données continues. Ph. D. thesis, Compiègne. 161

Ma, J., Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1930–1939. 89, 91

Marchello, G., A. Fresse, M. Corneli, and C. Bouveyron (2022). Co-clustering of evolving count matrices with the dynamic latent block model: application to pharmacovigilance. Statistics and Computing 32(3), 41. 40

Mariadassou, M. and C. Matias (2015). Convergence of the groups posterior distribution in latent or stochastic block models. 34

Mariadassou, M., S. Robin, and C. Vacher (2010). Uncovering latent structure in valued graphs: a variational approach. 114

Mariadassou, M. and T. Tabouy (2020). Consistency and asymptotic normality of stochastic block models estimators from sampled data. 36

Mariette, J. and N. Villa-Vialaneix (2018). Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics 34(6), 1009–1015. 28

Masoudnia, S. and R. Ebrahimpour (2014). Mixture of experts: a literature survey. Artificial Intelligence Review 42, 275–293. 41

Matias, C. and V. Miele (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society Series B: Statistical Methodology 79(4), 1119–1141. 34, 35, 114

McLachlan, G. J. and D. Peel (2000). Finite mixture models, Volume 299. John Wiley & Sons. 29

Mercado, P., A. Gautier, F. Tudisco, and M. Hein (2018). The power mean laplacian for multilayer graph clustering. In International Conference on Artificial Intelligence and Statistics, pp. 1828–1838. PMLR. 56

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. The journal of chemical physics 21(6), 1087–1092. 47

Miele, V. and C. Matias (2017). Revealing the hidden structure of dynamic ecological networks. Royal Society open science 4(6), 170251. 31

Mihaylov, I., M. Kańduła, M. Krachunov, and D. Vassilev (2019). A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. Biology direct 14, 1–17. 19

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. Journal of the american statistical association 83(404), 1023–1032. 115

Monti, S., P. Tamayo, J. Mesirov, and T. Golub (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52, 91–118. 52, 54

Moshawrab, M., M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad (2023). Reviewing multimodal machine learning and its use in cardiovascular diseases detection. Electronics 12(7), 1558. 24

Neal, R. M. (2012). Mcmc using hamiltonian dynamics. arXiv preprint arXiv:1206.1901. 48

Noroozi, M. and M. Pensky (2022). Sparse subspace clustering in diverse multiplex network model. arXiv preprint arXiv:2206.07602. 57

Nottet, J.-B., A. Moulin, A. Crambert, D. Bonete, and A. Job (2009). Traumatismes sonores aigus. Oto-rhino-laryngologie. 3

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. Journal of the American statistical association 96(455), 1077–1087. 31, 45

Pal, S. and M. Coates (2019). Scalable mcmc in degree corrected stochastic block model. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5461–5465. IEEE. 48

Pan, B., Y. Shen, H. Liu, M. Mishra, G. Zhang, A. Oliva, C. Raffel, and R. Panda (2024). Dense training, sparse inference: Rethinking training of mixture-of-experts language models. arXiv preprint arXiv:2404.05567. 89

Paul, S. and Y. Chen (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. Electronic Journal of Statistics 10(2), 3807 – 3870. 56

Paul, S. and Y. Chen (2020). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. 56

Peixoto, T. P. (2014). Hierarchical block structures and high-resolution model selection in large networks. Physical Review X 4(1), 011047. 35

Pensky, M. and Y. Wang (2021). Clustering of diverse multiplex networks. arXiv preprint arXiv:2110.05308. 57

Pezza, L., A. Alonso-Ojembarrena, Y. Elsayed, N. Yousef, L. Vedovelli, F. Raimondi, and D. De Luca (2022). Meta-analysis of lung ultrasound scores for early prediction of bronchopulmonary dysplasia. Annals of the American Thoracic Society 19(4), 659–667. 18

Phillips, S. P., S. Spithoff, and A. Simpson (2022). L'intelligence artificielle et les algorithmes prédictifs en médecine: Promesses et problèmes. Canadian Family Physician 68(8), e230–e233. 18

Proust-Lima, C., M. Séne, J. M. Taylor, and H. Jacqmin-Gadda (2014). Joint latent class models for longitudinal and time-to-event data: a review. Statistical methods in medical research 23(1), 74–90. 92

Puigcerver, J., C. Riquelme, B. Mustafa, and N. Houlsby (2023). From sparse to soft mixtures of experts. arXiv preprint arXiv:2308.00951. 89

Rebafka, T. (2023). Model-based clustering of multiple networks with a hierarchical algorithm. 57, 73

Robert, C. P., G. Casella, and G. Casella (1999). Monte Carlo statistical methods, Volume 2. Springer. 47, 48

Robert, V., G. Celeux, and C. Keribin (2015). Un modèle statistique pour la pharmacovigilance. In 47emes Journées de Statistique de la SFdS. 39, 40

Rohe, K., S. Chatterjee, and B. Yu (2011, August). Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics 39(4). 34

Saria, S. and A. Goldenberg (2015). Subtyping: What it is and its role in precision medicine. IEEE Intelligent Systems 30(4), 70–75. 19

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461–464. 69, 101

Sharifi-Noghabi, H., O. Zolotareva, C. C. Collins, and M. Ester (2019). Moli: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 35(14), i501–i509. 28

Shazeer, N., A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538. 90, 91

Shen, K. and G. C. Tseng (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. Bioinformatics 26(10), 1316–1323. 20

Shen, S., L. Hou, Y. Zhou, N. Du, S. Longpre, J. Wei, H. W. Chung, B. Zoph, W. Fedus, X. Chen, et al. (2023). Mixture-of-experts meets instruction tuning: A winning combination for large language models. arXiv preprint arXiv:2305.14705. 42

Shivahare, B. D., J. Singh, V. Ravi, R. R. Chandan, T. J. Alahmadi, P. Singh, and M. Diwakar (2024). Delving into machine learning's influence on disease diagnosis and prediction. The Open Public Health Journal 17(1). 18

Singh, A., C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao (2019). Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 35(17), 3055–3062. 28

Song, B., S. Miller, and F. Ahmed (2023). Attention-enhanced multimodal learning for conceptual design evaluations. Journal of Mechanical Design 145(4), 041410. 29

Stahlschmidt, S. R., B. Ulfenborg, and J. Synnergren (2022). Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics 23(2), bbab569. 26, 88

Stanley, N., S. Shai, D. Taylor, and P. J. Mucha (2016). Clustering network layers with the strata multilayer stochastic block model. IEEE transactions on network science and engineering 3(2), 95–105. 56, 68

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. Psychological methods 9(3), 386. 102, 162

Strehl, A. and J. Ghosh (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research 3(Dec), 583–617. 54, 102, 162

Subramanian, I., S. Verma, S. Kumar, A. Jere, and K. Anamika (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights 14, 1177932219899051. 20

Tabouy, T., P. Barbillon, and J. Chiquet (2020). Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association 115(529), 455–466. 60

Tan, S., Y. Shen, Z. Chen, A. Courville, and C. Gan (2023). Sparse universal transformer. arXiv preprint arXiv:2310.07096. 89

Tierney, L. (1994). Markov chains for exploring posterior distributions. the Annals of Statistics, 1701–1728. 47

Tobin, J., M. Black, J. Ng, D. Rankin, J. Wallace, C. Hughes, L. Hoey, A. Moore, J. Wang, G. Horigan, et al. (2024). Co-clustering multi-view data using the latent block model. arXiv preprint arXiv:2401.04693. 98

Tseng, G. C., D. Ghosh, and E. Feingold (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic acids research 40(9), 3785–3799. 20

Tzenios, N., M. E. Tazanios, and M. Chahine (2024). The impact of bmi on breast cancer–an updated systematic review and meta-analysis. Medicine 103(5), e36831. 20

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems. 42

Vermunt, J. K. (2002). Latent class cluster analysis. Applied latent class analysis/Cambridge UniversityPress. 91

Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing 17, 395–416. 55

Von Luxburg, U., M. Belkin, and O. Bousquet (2008). Consistency of spectral clustering. The Annals of Statistics, 555–586. 55

Vu, D. and M. Aitkin (2015). Variational algorithms for biclustering models. Computational Statistics & Data Analysis 89, 12–24. 92

Wainwright, M. J., M. I. Jordan, et al. (2008). Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning 1(1–2), 1–305. 46

Wang, H., P. Guo, P. Zhou, and L. Xie (2024). Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8150–8154. IEEE. 27

Wang, M. and Y. Zeng (2019). Multiway clustering via tensor block models. Advances in neural information processing systems 32. 57

Wang, T., L. Zhang, and W. Hu (2021). Bridging deep and multiple kernel learning: A review. Information Fusion 67, 3–13. 26

Wikipédia (2024). Perte d'audition due au bruit — wikipédia, l'encyclopédie libre. [En ligne; Page disponible le 13-août-2024]. 2

Wu, M., H. Yi, and S. Ma (2021). Vertical integration methods for gene expression data analysis. Briefings in Bioinformatics 22(3), bbaa169. 19

Wu, X., S. Huang, and F. Wei (2023). Mole: Mixture of lora experts. In The Twelfth International Conference on Learning Representations. 89

Xing, E. P., W. Fu, and L. Song (2010). A state-space mixed membership blockmodel for dynamic network tomography. 36

Xu, Y. and R. P. McCord (2022). Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. Nature Communications 13(1), 3505. 20

Yildirim, I. (2012). Bayesian inference: Gibbs sampling. Technical Note, University of Rochester. 98

Yuksel, S. E., J. N. Wilson, and P. D. Gader (2012). Twenty years of mixture of experts. IEEE transactions on neural networks and learning systems 23(8), 1177–1193. 41

Zadouri, T., A. Üstün, A. Ahmadian, B. Ermiş, A. Locatelli, and S. Hooker (2023). Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. arXiv preprint arXiv:2309.05444. 42, 89

Zanghi, H., S. Volant, and C. Ambroise (2010). Clustering based on random graph model embedding vertex features. Pattern Recognition Letters 31(9), 830–836. 34, 114

Zhao, J., X. Xie, X. Xu, and S. Sun (2017). Multi-view learning overview: Recent progress and new challenges. Information Fusion 38, 43–54. 17

Zhou, T., M. Liu, K.-H. Thung, and D. Shen (2019). Latent representation learning for alzheimer's disease diagnosis with incomplete multimodality neuroimaging and genetic data. IEEE transactions on medical imaging 38(10), 2411–2422. 18

Zhou, Y., T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon, et al. (2022). Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems 35, 7103–7114. 89, 90, 91

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pp. 2223–2232. 28