# Actes des 28es Rencontres de la Société Francophone de Classification

Pascal Préa

AfIA

Association française
pour l'Intelligence Artificielle

# SFC

*Rencontres de la Société Francophone de Classification*

# PFIA 2023

Les rencontres de la SFC sont soutenues par
l'École Centrale Méditerranée

# Table des matières

Rafik Abdesselam, Véronique Cariou, Ndèye Niang, Pascal Préa & Allou Samé.
**Éditorial** ..... 7

**Comité de programme** ..... 9

**Conférences Invitées** ..... 11

Christophe Biernacki
**Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering** ..... 13

Anne-Laure Boulesteix
**Towards reliable empirical evidence in methodological computational research : recent developments and remaining challenges** ..... 15

**Prix Simon Régnier** ..... 17

K. LeGall, L. Bellanger, A. Stamm, D.A. Laplaud
**Génération de données synthétiques de marche : application au cas de patients atteints de sclérose en plaques** ..... 19

**Communications** ..... 23

L. Labiod, M. Nadif
**Approximation matricielle bi-stochastique de k-means et ses variante** ..... 25

P. Riverain, A. Samé, L. Oukhellou
**Classification et segmentation conjointes de données fonctionnelles, une approche par blocs latents** ..... 27

P. Bertrand, J. Diatta
**Classifications multi-niveaux et convexités d'intervalle : cas de la hiérarchie du lien simple** ..... 31

R. Abdesselam
**Classification Topologique sur Données Evolutives** ..... 35

L. Labiod, M. Nadif
**Clustering sur l'hypersphère unitaire via NMF** ..... 41

M.L. Ndao, N. Niang, G. Youness, G. Saporta
**Consensus de partitions en NLP pour une revue systématique de la littérature autour de l'XAI du biais et de l'équité** ..... 43

T. Ranvier, H. Elghazel, E. Coquery, K. Benabdeslem
**Considération de l'Incertitude d'Imputation pour l'Apprentissage des Réseaux de Neurones** ..... 49

S. Dominique, M. Hanafi, F. Llobell, J.M. Ferrandi, V. Cariou
**Deux variantes à la méthode de classification FIMIX-PLS dans le cadre des modèles d'équations structurelles** ..... 55

I. Keraghel, S. Morbieu, M. Nadif
**Étude sur la classification d'entitées nommées** ..... 61

J.-C. Lamirel, F. Lareau, C. Malaterre
**La méthode de modélisation thématique CFMf basée sur le clustering neuronal avec maximisation des traits : Comparaison avec LDA sur des études scientifiques** ..... 67

D. Desbois
**Méthode de classification divisive sur intervalles d'estimation duale des quantiles de coûts spécifiques et de marges brutes** ..... 73

A. Ferdjaoui, S. Affeldt, M. Nadif
**Modèles graphiques causaux interactifs pour les données textuelles** ..... 81

M. Carmona, V. Chepoi, G. Naves, P. Préa
**Modules dans les Espaces de Robinson** ..... 87

M. Hennequin, K. Benabdeslem, H. Elghazel
**PAC-Bayesian bornes pour l'adaptation de domaine non supervisé dans un cadre d'apprentissage multi-vue** ..... 93

R. Khoufache, M.D. Dilmi, H. Azzag, É. Goffinet, M. Lebbah
**Propriétés émergentes du *multi-clustering* bayésien non paramétrique : Application aux données images multivues** ..... 101

Z. Tighidet, L. Labiod, M. Nadif
**Réduction de la Dimension et Classification : approche jointe** ..... 107

J. Ah-Pine
**Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering** ..... 109

M. Carmona, V. Chepoi, G. Naves, P. Préa
**Un algorithme simple et efficace pour la sériation circulaire** ..... 113

N. Niang, M. Ouattara, G. Saporta
**Une comparaison de quelques méthodes de classification de variables mixtes** ..... 115

# Éditorial

## Rencontres de la Société Francophone de Classification

Depuis plus de trente ans, les rencontres de la SFC ont pour objectif de présenter des résultats récents et des applications originales en classification sous toutes ses formes, mathématique, informatique et statistique, de favoriser les échanges scientifiques entre ces trois communautés autour de la thématique commune de la classification et de faire connaitre à divers partenaires extérieurs les travaux de ses membres.

Cette année est malheureusement très particulière :

Edwin DIDAY s'est éteint le 28 avril dernier, à l'âge de 83 ans. Au nom de la société francophone de classification, nous tenons à lui rendre hommage, lui qui a tant œuvré pour la classification et l'analyse de données.

Edwin Diday était Professeur émérite à l'Université Paris-Dauphine et membre du laboratoire de recherche CEREMADE. Il a aussi été Directeur de Recherche à l'INRIA (Rocquencourt). Edwin Diday a eu un impact profond en classification et plus généralement en analyse des données, tant en France qu'à l'étranger. Convaincu très tôt de la valeur des données, de leur complexité potentielle et des possibilités de calcul et de stockage offertes par l'évolution technologique, il a développé de nouvelles méthodes, introduit des paradigmes conduisant vers de nouveaux champs de recherche. Son nom reste associé aux nuées dynamiques, méthode de classification qu'il a développée en 1971 et qui a ouvert la voie aux modèles locaux, ainsi qu'à la classification pyramidale. À partir de la fin des années 1980, il a développé l'Analyse des Données Symboliques (ADS) pour traiter des unités statistiques plus complexes pouvant correspondre à des concepts, des classes, des agrégats, etc. Par l'ADS, il proposait ainsi le cadre formel de représentation des données avec leur variabilité, sous forme d'ensembles de valeurs, d'intervalles ou, plus généralement, de distributions et l'extension des méthodes existantes à de telles structures. Ce champ de recherche fut pour lui l'occasion de nombreuses collaborations en particulier au travers de projets européens comme "SODAS" et "ASSO" dont il était le Directeur Scientifique (17 équipes de 9 pays européens) jusqu'en 2003. L'héritage scientifique qu'il laisse à la communauté est sans mesure avec près de 200 publications incluant 14 livres en tant qu'auteur et éditeur. Ses deux derniers livres parus en 2018 et 2019 sont consacrés à l'ADS. Lauréat du prix Montyon de l'Académie des Sciences, Edwin Diday a reçu en juillet 2022 à Porto la médaille de recherche IFCS pour sa très grande contribution à la recherche en classification et en analyse des données.

Edwin aimait partager et faire partager la recherche. Il a formé un très grand nombre d'étudiants et dirigé plus d'une cinquantaine de thèses. Il a contribué avec une constante motivation au rayonnement de la classification à travers les congrès et les sociétés savantes. Durant les années 70 et 80, les premières conférences d'analyse de données avaient lieu puis dans les années 80, la création des journées « symboliques – numériques » réunissant des chercheurs en apprentissage symbolique et en analyse des données offrait l'opportunité de croiser les approches. Edwin Diday a été le co-fondateur de la Société francophone de classification (SFC), dont il a été président dans les années 1990. Les assemblées annuelles de la SFC, telles que celle de Strasbourg cette année, ont vu le jour à cette époque, sous l'égide de sa présidence. L'an dernier à Lyon, nous avions eu le plaisir de l'écouter, lors d'une session plénière, nous parler de ses derniers travaux en analyse des concordances et discordances. A l'international, Edwin Diday a contribué à la création de l'IFCS. Il en a d'ailleurs organisé la 3e édition à Paris en 1993. Ces dernières années, Edwin était à pied d'œuvre pour organiser un Workshop annuel à l'Université Paris-Dauphine sur l'analyse de données complexes et la Science des Données, avec le même enthousiasme, la même curiosité et une égale bonne humeur.

La recherche n'a pas de frontière et Edwin Diday l'a illustré à de maintes reprises. Edwin Diday a eu de multiples collaborations avec des chercheurs étrangers de partout dans le monde : Brésil, Canada, Etats-Unis, Espagne, Inde, Japon, Portugal, etc . Grâce à Edwin, des relations d'échange et d'amitié sincère se sont renforcées entre la société francophone de classification et ses homologues en Europe, tout particulièrement la Société Italienne et la Société Portugaise : la CLAD. Aussi il nous paraissait ô combien évident de demander à Paula Brito une présentation en l'hommage d'Edwin. Paula est présidente de la CLAD, elle est aussi une

ancienne doctorante d'Edwin avant de devenir une fidèle collègue de recherche jusqu'à ses tous derniers travaux. Au nom de la communauté francophone de classification, merci Paula de nous faire la gentillesse de ta présence.

Rafik Abdesselam, Véronique Cariou, Ndèye Niang, Pascal Préa & Allou Samé.

# Comité de programme

## Présidence

— Pascal Préa, École Centrale Méditerranée, Marseille.

## Membres

— Rafik Abdesselam, Université Lumière Lyon ;
— Séverine Affeldt, Université de Paris ;
— Alexandre Bazin, LIRMM, Montpellier ;
— Patrice Bertrand, Université Paris Dauphine ;
— Paula Brito, Université de Porto, Portugal ;
— François Brucker, École Centrale Méditerranée ;
— Véronique Cariou, ONIRIS Nantes ;
— Christian Derquenne, EDF R&D ;
— Dominique Desbois, INRAE-Paris-Saclay ;
— Jean Diatta, Université de La Réunion ;
— Nadia Ghazalli, Université du Québec à Trois-Rivières ;
— Pascale Kuntz, Université de Nantes ;
— Lazhar Labiod, Université Paris Descartes ;
— Mustapha Lebbah, Université Paris 13 ;
— Ahmed Moussa, ENSA Tanger, Maroc ;
— Mohamed Nadif, Université Paris Descartes ;
— Amedeo Napoli, LORIA, Nancy ;
— Ndèye Niang, CNAM Paris ;
— Allou Samé, Université Gustave Eiffel ;
— Rosanna Verde, Université della Campania, Caserta, Italie.

# Conférences Invitées

# Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering

Christophe Biernacki[1]

[1] INRIA & Université de Lille I

christophe.biernacki@inria.fr

**Résumé**

A generic method is introduced to visualize in a "Gaussian-like way," and onto R2, results of Gaussian or non-Gaussian–based clustering. The key point is to explicitly force a visualization based on a spherical Gaussian mixture to inherit from the within cluster overlap that is present in the initial clustering mixture. The result is a particularly user-friendly drawing of the clusters, providing any practitioner with an overview of the potentially complex clustering result.

An entropic measure provides information about the quality of the drawn overlap compared with the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional, and network) and is implemented on the R package CLUSVIS.

# A replication crisis in methodological research? Recent developments and remaining challenges towards reliable empirical evidence in methodological computational research

Anne-Laure Boulesteix[1]

[1] Ludwig-Maximilians Universität, Munich

bouleste@ibe.med.uni-muenchen.de

**Résumé**

Statisticians are often keen to analyze the statistical aspects of the so-called "replication crisis in science". They condemn fishing expeditions and publication bias across empirical scientific fields applying statistical methods, such as health sciences. But what about good practice issues in their own - methodological - research, i.e. research considering statistical (or more generally, computational) methods as research objects ? When developing and evaluating new statistical methods and data analysis tools, do statisticians and data scientists adhere to the good practice principles they promote in fields which apply statistics and data science ?

I argue that methodological researchers should make substantial efforts to address what may be called the replication crisis in the context of methodological research in statistics and data science, in particular by trying to avoid bias in comparison studies based on simulated or real data. I discuss topics such as publication bias, cherry-picking, and the design and necessity of neutral comparison studies, and review recent positive developments towards more reliable empirical evidence in the context of methodological computational research.

## Prix Simon Régnier

# Génération de données synthétiques de marche : application au cas de patients atteints de sclérose en plaques

K. Le Gall[1], L. Bellanger[1], A. Stamm[1], D.A. Laplaud[2]

[1] Laboratoire de Mathématique Jean Leray, UMR CNRS 6629, Nantes Université, France
[2] CR2TI, INSERM U1064, CHU de Nantes, Nantes Université, France

Klervi.Legall@univ-nantes.fr

## Résumé

*L'objectif de ce travail est de générer des séries temporelles de quaternions synthétiques (QTS). L'approche proposée permet de construire un jeu de données synthétiques en mêlant ACP fonctionnelle et proches voisins et permet une bonne conservation de la géométrie des données. Nous montrerons la pertinence de notre approche à l'aide d'un échantillon de séries temporelles de quaternions issu de données de marche de 27 patients atteints de sclérose en plaques issu d'une étude clinique menée en collaboration avec l'équipe de neurologie du CHU de Nantes.*

## Mots-clés

*Analyse en Composantes Principales fonctionnelle, Séries temporelles de Quaternions unitaire, Marche, Sclérose en plaques, Données Synthétiques.*

## Abstract

*The objective of this work is to generate synthetic quaternion time series (QTS). The proposed approach allows to build a synthetic dataset by mixing functional PCA and nearest neighbours and allows a good preservation of the data geometry. We will show the relevance of our approach using a sample time series of quaternions from the gait data of 27 multiple sclerosis patients from a clinical study conducted in collaboration with the neurology team of the Nantes University Hospital.*

## Keywords

*Functional Principal Component Analysis, Unit Quaternion Time Series, Walking, Multiple Sclerosis, Synthetic Data.*

# 1 Introduction

Les séries temporelles de quaternions unitaires (QTS) permettent de caractériser les rotations et sont donc présentes dans de nombreux domaines tels que la robotique, les jeux vidéo ou la santé. L'un des principaux défis dans certains domaines tels que la médecine est la taille réduite des échantillons en raison de la difficulté à mener de nombreuses expériences coûteuses. A ce jour, il n'existe pas de méthode pour générer des séries temporelles de quaternions unitaires proches des données originales. La création de données synthétiques permet remédier à ce problème. Les données synthétiques sont "toute donnée de production applicable à une situation donnée qui n'est pas obtenue par mesure directe", selon le McGraw-Hill Dictionary of Scientific and Technical Terms [7].

L'approche que nous proposons pour générer des QTS synthétique est inspirée de la méthode avatar [6], une méthode d'anonymisation de données basée sur l'individu et ses proches voisins, qui nous permet de respecter la géométrie de nos données. La méthode proposée peut également servir de méthode d'anonymisation des données afin de respecter le Règlement général sur la protection des données (RGPD) et ainsi protéger les données personnelles des individus.

Nous illustrerons notre approche en utilisant un échantillon de 27 patients atteints de sclérose en plaques issus d'une étude clinique menée en collaboration avec l'équipe de neurologie du CHU de Nantes pour lesquels nous avons mesuré un biomarqueur appelé Individual Gait Pattern qui caractérise la rotation de la hanche d'un individu au cours d'un cycle de marche moyen en utilisant des séries temporelles de quaternions unitaires. L'augmentation de notre ensemble de données nous permettra également de tester la robustesse des algorithmes de classification précédemment mis en œuvre.

# 2 Méthode de génération de séries temporelles de quaternions unitaires synthétiques

## 2.1 Séries temporelles de quaternions unitaires

Un quaternion est un élément (w,x,y,z) $\in \mathbb{R}^4$, qui est une extension des nombres complexe et qui s'écrit comme suit :

$$\mathbf{q} = (w, x, y, z)^\top = w + ix + jy + kz \in \mathbb{R}^4 \quad (1)$$

où $i$, $j$ et $k$ respectent $i^2 = j^2 = k^2 = ijk = -1$. [12]

**Les quaternions unitaires** sont ceux pour lesquels $||\mathbf{q}|| = w^2 + x^2 + y^2 + z^2 = 1$ de sorte que leur norme soit égale à 1. Le groupe des quaternions unitaires $\mathbb{H}_u$ forme un groupe de Lie isomorphe au groupe unitaire spécial $SU(2)$

qui couvre deux fois le groupe des matrices de rotations en 3-Dimensions [2].
Les quaternions unitaires décrivent une rotation d'un angle $\theta$ autour d'un axe $\mathbf{u} = (u_x, u_y, u_z)^\top \in \mathbb{S}^2$ où $\mathbb{S}^2$ est la 2-sphère qui peut être exprimée comme suit :

$$\mathbf{q} = (cos\ \frac{\theta}{2} + \mathbf{u}_x\ sin\ \frac{\theta}{2} i + \mathbf{u}_y\ sin\ \frac{\theta}{2} j + \mathbf{u}_z\ sin\ \frac{\theta}{2} k)^\top \quad (2)$$

**Série temporelle de quaternion unitaires**
Une série temporelle de quaternions unitaires (QTS) est un ensemble de quaternions unitaires suivant une grille temporelle $t_{i,1}, \ldots, t_{i,n}$. On note une QTS comme : $\mathbf{Q}_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,n})$. Elle représente des rotations 3D consécutives dans le temps.

## 2.2 Génération de QTS synthétiques

Le logarithme d'un quaternion unitaire est déterminé à partir de sa forme polaire :

$$ln(\mathbf{q}) = ln(exp(\tilde{\mathbf{u}}\frac{\theta}{2})) = \tilde{\mathbf{u}}\frac{\theta}{2} = (0, u_x\frac{\theta}{2}, u_y\frac{\theta}{2}, u_z\frac{\theta}{2})^\top \quad (3)$$

Cette transformation logarithmique est une application entre l'espace des quaternions unitaires et l'espace tangent au point $q = (1, 0, 0, 0)$ [8]. Cela signifie que si les données sont préalablement centrées, on peut les traiter dans $\mathbb{R}^3$. Cet outil est particulièrement intéressant lorsque les points dans $\mathbb{H}_u$ sont proches les uns des autres, cela garanti une conservation suffisante des distances entre les points dans l'espace tangent.
En centrant les séries temporelles de quaternions et appliquant cette transformation logarithmique, nous obtenons trois séries temporelles qui peuvent être traitées comme des *données fonctionnelles*.
Lorsque l'on travaille avec des données fonctionnelles, les observations sont des fonctions qui varient selon certaines variables continues, par exemple le temps. L'ensemble de données est alors une collection de $n$ fonctions $X_i(t), \quad i = 1, \ldots, n$.
La méthode que nous proposons repose sur la possibilité de décomposer des données fonctionnelles sous forme de fonctions et de scores à l'aide d'une *analyse en composantes principales fonctionnelle multivariée* (MFCPA) [9] tel que :

$$X_i(t) = \sum_{k=1}^{K} F_i^k \mathbf{u}_k(t) \quad (4)$$

Ainsi, pour chaque individu $i$, à chaque temps $t$, on peut décomposer une log-série temporelle de quaternions selon trois fonctions $\mathbf{u}_k(t) \in \mathbb{R}^3$ et un score $F_i^k \in \mathbb{R}$ pour chaque composante principale $k \in 1 : K$ .
Cette formule nous permet donc également d'obtenir de nouvelles données, en changeant le score $F_i^k$, mais en conservant les fonctions principales.
Dans la méthode proposée, de nouveaux scores synthétiques $F_{i\,new}^k$ sont calculés pour chaque individu $i$. A partir des scores obtenus par MFPCA, on détermine les $L$ plus proches voisins de l'individu $i$ sur les $T \leq K$ premières

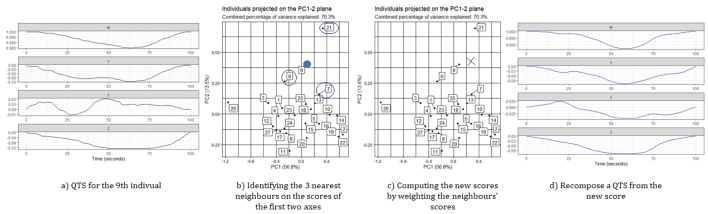



FIGURE 1 – Génération d'une série temporelle de quaternions unitaires synthétique

composantes. Les scores de ces voisins sont ensuite pondérés en prenant compte d'une distribution aléatoire et la distance à l'individu $i$ et ses $L$ voisins afin d'obtenir ce score synthétique.
Les données sont ensuite recomposées pour obtenir des séries temporelles de quaternions unitaires $\mathbf{Q}_{i\,new}$ tel que :

$$\mathbf{Q}_{i\,new} = exp(\sum_{k=1}^{K}(F_{i\,new}^k \mathbf{u}_k(t))) \quad (5)$$

Enfin, les séries temporelles de quaternions unitaires sont décentrées.
Nous proposons également des métriques qui permettent de contrôler la qualité des QTS synthétiques obtenues. Afin de vérifier le respect de la géométrie, nous calculons des distances de Frobenius entre les matrices d'adjacences des graphes des plus proches voisins [1, 5]. La conservation des des informations est indiquée par le coefficient RV [10], et si la méthode est utilisée pour de l'anonymisation, le risque de ré-identification est évaluée par le local cloaking [6].
La figure 1 illustre la méthode proposée.

# 3 Application aux données de marche

Les données sur lesquelles nous nous appuyons sont issues d'une étude menée en partenariat avec l'équipe de neurologie du Centre d'Investigation Clinique de Nantes, comprenant le Pr. Laplaud, (Neurologue et PU-PH au CHU de Nantes) et le Pr. Gourraud (PU-PH au CHU de Nantes). L'étude MYO porte sur le signal nerveux des patients atteints par la sclérose en plaques (SEP) mesuré par le bracelet électronique MYO en 2018 et un amendement a permis d'ajouter à cette étude principale une étude ancillaire pour mesurer la marche des patients via un capteur positionné à la hanche droite. Nous avons pu recueillir des données pour 27 patients.
Nous travaillons sur l'analyse des données de ce capteur afin de comprendre les troubles de la marche dans le contexte de la sclérose en plaques. Le système de capteurs transmet le vecteur d'orientation absolue du dispositif sous la forme d'un quaternion unitaire toutes les 0,01 secondes, une série temporelle de quaternions unitaires est donc recueillie pour chaque patient. Ces données brutes sont ensuite segmentées en cycles de marche et la signature de marche (SdM) d'un individu correspond au centre

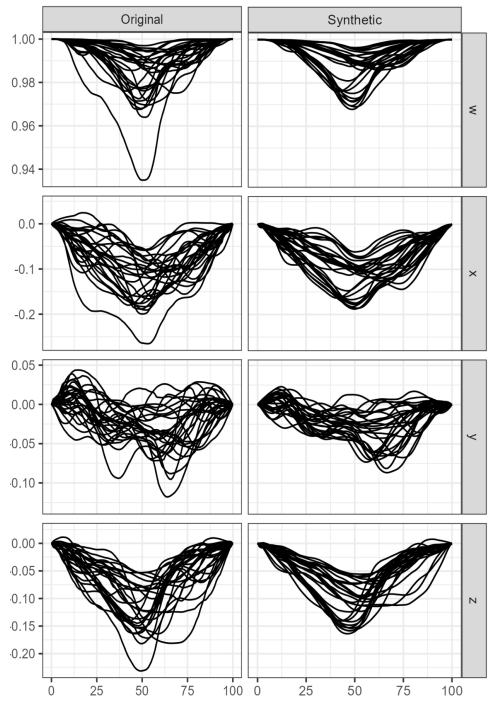



FIGURE 2 – Signatures de marche des 27 patients de l'étude ancillaire MYO et signatures de marches synthétiques obtenues à partir de ces données avec $L = 4$ et $T = 10$

des cycles de marche détectés, ce centre est obtenu avec une méthode de k-means alignement [11] pour lequel le nombre de cluster est égal à 1. La série temporelle est ensuite exprimée en pourcentage de la durée totale. Le temps 1 ($0\%$) est la référence à partir de laquelle les orientations observées pendant le cycle de marche sont calculées. Cela permet d'obtenir une SdM dont le premier et le dernier élément sont des quaternions identité $\mathbf{q} = (1, 0, 0, 0)$. Ces SdM sont des QTS unitaires et décrivent les rotations "moyennes" de la hanche pendant un cycle de marche [4].

La figure 2 représente les SdM des patients de l'étude ancillaire MYO, ainsi qu'une génération de données synthétiques possible à partir de ces SdM.

Deux éléments se dégagent de cette figure, on remarque tout d'abord que les données sont davantage centrées, cela est lié à la pondération des proches voisins pour créer les scores synthétiques qui permet de donner moins d'importance à de potentiels cas extrêmes. On observe également des groupes d'amplitudes différentes, en partie expliqués par la pathologie. Le coefficient RV est de 0.75, ce qui montre une bonne conservation des données, la géométrie a également été vérifiée.

Cette approche pourra aussi permettre de tester la robustesse des algorithmes de classification précédemment mis en œuvre [3].

## Remerciements

Les auteurs remercient la fondation ARSEP (fondation pour l'Aide à la Recherche sur la Sclérose En Plaques) et l'AMIES (Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société) pour le financement des études cliniques qui ont menées à l'obtention des données ainsi que le groupe de l'Observatoire Français de la Sclérose en Plaques (OFSEP) et les CHU de Rennes et de Nantes. Ce travail s'inscrit dans une thèse co-financée par l'ANR AIBY4 (ANR-20-THIA-0011) ainsi que Nantes Université.

## Références

[1] R. Balakrishnan, K. Ranganathan, *A Textbook of Graph Theory,* Springer New York, 2012.

[2] M.S. Dijkhuizen, *The double covering of the quantum group SOq(3),* Proceedings of the Winter School "Geometry and Physics". Circolo Matematico di Palermo, Vol. 37, pp. 47-57, 1994.

[3] P. Drouin, A. Stamm, L. Chevreuil et al, *Semi-supervised clustering of quaternion time series : application to gait analysis in multiple sclerosis using motion sensor data,* accepted, 2022.

[4] P. Drouin, A. Stamm, L. Chevreuil et al, *Gait impairment monitoring in multiple sclerosis using a wearable motion sensor,* Medical Case reports and Reviews, Vol. 5, pp. 1-5, 2022.

[5] D.Eppstein, M.S Paterson, F.F Yao, *Discrete and Computational Geometry,* Springer New York, pp. 263-282, 1997.

[6] M. Guillaudeux, O. Rousseau, J. Petot et al, *Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymised data,* PREPRINT(V1), 2022.

[7] S.P. Parker, *McGraw-Hill Dictionary of Scientific and Technical Terms. 7e éd.,* Cambridge University Press, 2009.

[8] M. Piórek, *« Analysis of Chaos for Quaternion Time Series » In : Analysis of Chaotic Behavior in Non-linear Dynamical Systems,* Springer, pp. 73-88, 2019.

[9] J.O. Ramsay , B.W. Silverman, *« Principal components analysis for functional data» In : Functional Data Analysis,* Springer, pp. 147-172, 2005.

[10] P. Robert, Y. Escoufier, *A Unifying Tool for Linear Multivariate Statistical Methods : The RV- Coefficient,* Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol.25, 1976.

[11] L.M Sangalli, P. Secchi, S. Vantini et al, *k-mean alignment for curve clustering,* Computational Statistics Data Analysis, Vol. 54, pp. 1219-1233, 2010.

[12] J. Voight, *Quaternion Algebras,* Springer Nature, 2005.

# Communications

# Approximation matricielle bi-stochastique de k-means et ses variantes

Lazhar Labiod, Mohamed Nadif

Centre Borelli UMR 9010, Université Paris Cité, 45 rue des Saints-Pères, 75006 Paris

prénom.nom@u-paris.fr

## Résumé

*L'algorithme k-means et certaines de ses variantes se sont révélés utiles et efficaces pour résoudre le problème de clustering. Dans cet article, nous intégrons de telles variantes dans un cadre d'approximation matricielle bi-stochastique (BMA). De la fonction objectif k-means nous dérivons une nouvelle formulation du critère. En particulier, nous montrons que certaines variantes sont équivalentes au problème algébrique d'approximation de matrice bi-stochastique sous certaines contraintes appropriées. Pour optimiser la fonction objectif dérivée, nous développons deux algorithmes ; le premier consiste à apprendre une matrice de similarité bi-stochastique tandis que la seconde recherche la partition optimale qui est l'état d'équilibre d'un processus en chaîne de Markov. Des expérimentations numériques sur des jeux de données réels montrent l'intérêt de notre approche.*

## Mots-clés

*Classification non supervisée, factorisation, k-means.*

## 1 Introduction

Ces dernières décennies, l'apprentissage non supervisé et plus particulièrement le clustering, ont reçu une attention significative en tant que problème important avec de nombreuses applications en science des données. Soit $A = (a_{ij})$ une matrice de données continue $n \times m$ où l'ensemble des lignes (objets, individus) est noté $I$ et l'ensemble des colonnes (attributs, caractéristiques) par $J$ . De nombreuses méthodes de clustering visent à construire une partition optimale de $I$ ou, parfois de $J$.

Dans cet article, nous montrons comment certaines variantes de k-means peuvent être présentées comme un problème d'approximation de matrice bi-stochastique sous certaines contraintes appropriées générées par les propriétés de la solution recherchée. Pour atteindre cet objectif, nous démontrons d'abord que certaines variantes de k-means sont équivalentes à l'apprentissage d'une matrice de similarité bi-stochastique ayant une structure diagonale par blocs. Sur la base de cette formulation, appelée BMA, nous dérivons deux algorithmes itératifs, le premier algorithme apprend une matrice de similarité bi-stochastique $n \times n$ tandis que le second cherche directement une solution de clustering optimale.

Notre principale contribution est d'établir la connexion théorique des k-means conventionnels et de certaines de ses variantes au cadre BMA. Les conséquences de la reformulation des k-means en tant que problème BMA sont multiples :

- Elle permet d'établir des liens avec les méthodes de clustering récentes comme le *spectral clustering* et le *subspace clustering*.
- Elle apprend une matrice de similarité bien normalisée (normalisation bi-stochastique), bénéfique pour le *spectral clustering* [5]
- Contrairement aux méthodes spectrales et *subspace* existantes qui combinent de manière séquentielle les étapes d'apprentissage de similarité et de dérivation de clustering, notre méthode proposée apprend conjointement une matrice d'affinité bistochastique diagonale par blocs exprimant naturellement une structure de clustering.

## 2 Variantes de k-means

Étant donné une matrice de données $A = (a_{ij}) \in \mathbb{R}^{n \times m}$, le but du clustering est de regrouper les lignes ou les colonnes de $A$, de manière à optimiser l'écart (à définir) entre $A = (a_{ij})$ et la matrice réorganisée révélant ainsi une structure en blocs. Plus formellement, on cherche à partitionner l'ensemble des lignes $I = \{1, \ldots, n\}$ en $k$ clusters $C = \{C_1, \ldots, C_l, \ldots, C_k\}$. Le partitionnement induit naturellement la matrice d'indices de clustering $R = (r_{il}) \in \mathbb{R}^{n \times k}$, définie comme matrice de classification binaire telle que nous avons $r_{il} = 1$, si la ligne $a_i \in C_l$, et 0 sinon. D'autre part, on note $S \in \mathbb{R}^{m \times k}$ une matrice correspondant à la représentation de cahque classe. La détection de classes homogènes d'objets peut être atteinte en recherchant les deux matrices $R$ et $S$ minimisant la partie résiduelle notée $\mathcal{J}_{KM}$.

$$\mathcal{J}_{KM}(R, S) = ||A - RS^\top||^2 \tag{1}$$

Le terme $RS^\top$ caractérise les informations de $A$ qui peuvent être décrites par la structure des classes. Le problème de clustering peut ainsi être formulé comme un problème d'approximation matricielle où le clustering vise à minimiser l'erreur d'approximation entre les données d'origine $A$ et la matrice reconstruite s'appuyant sur les classes.

Par ailleurs, rappelons que *Factoriel k-means* (FKM) [3] et *Reduced k-means* (RKM) [1] sont des méthodes de clustering qui visent à obtenir simultanément un clustering des objets et une réduction de dimension des caractéristiques. L'avantage de ces méthodes est que le regroupement d'objets et le sous-espace de faible dimension capturant la structure en classes sont obtenus simultanément. Pour atteindre cet objectif, RKM est défini par le problème de minimisation du critère suivant

$$\mathcal{J}_{RKM}(R,S,Q) = ||A - RS^\top Q^\top||^2 \tag{2}$$

et FKM est défini par le problème de minimisation du critère suivant

$$\mathcal{J}_{FKM}(R,S,Q) = ||AQ - RS^\top||^2 \tag{3}$$

où $S \in \mathbb{R}^{p\times k}$ avec RKM et FKM, et $Q \in \mathbb{R}^{m\times p}$ matrice orthonormée.

# 3 Approximation bi-stochastique

## 3.1 Factorisation de rang inférieur (MF)

En considérant k-means comme une factorisation matricielle de rang inférieur avec contraintes, plutôt qu'une méthode de clustering, nous pouvons formuler des contraintes à imposer à la formulation MF. Soit $D_r^{-1} \in \mathbb{R}^{k\times k}$ une matrice diagonale $D_r^{-1} = Diag(r_1^{-1},\ldots,r_k^{-1})$. En utilisant les matrices $D_r$, $A$ et $R$, le résumé de la matrice $S$ peut être exprimé comme $S^T = D_r^{-1}R^\top A$. En introduisant $S$ dans la fonction objectif de l'équation, (1) conduit à optimiser

$$\mathcal{J}_{MF-KM}(\mathbf{R}) = ||A - \mathbf{R}\mathbf{R}^\top A||^2, \text{ where } \mathbf{R} = RD_r^{-0.5}.$$

D'autre part, il est facile de vérifier que l'approximation $\mathbf{R}\mathbf{R}^\top A$ de $A$ est formée par la même valeur dans chaque bloc $A_{l,(l=1,\ldots,k)}$. Concrètement, la matrice $\mathbf{R}^\top A$, égale à $S^T$, joue le rôle d'un résumé de $A$ et absorbe les différentes échelles de $A$ et $\mathbf{R}$. De la même manière, nous pouvons dériver des formulation MF de FKM et RKM,

$$\mathcal{J}_{MF-FKM}(\mathbf{R}) = ||AQ - \mathbf{R}\mathbf{R}^\top AQ||^2, \tag{4}$$

$$\mathcal{J}_{MF-RKM}(\mathbf{R}) = ||A - \mathbf{R}\mathbf{R}^\top AQQ^\top||^2. \tag{5}$$

## 3.2 Formulation BMA

Soit $\mathbf{\Pi} = \mathbf{R}\mathbf{R}^\top$ une matrice de similarité bi-stochastique. A noter que par construction, $\mathbf{\Pi}$ qui a plusieurs propriétés est nonnegative, symmetric, et idempotente. Le problème de clustering peut être reformulé comme l'apprentissage d'une matrice de similarité bi-stochastique structurée $\mathbf{\Pi}$ en minimisant les critères associés aux différentes variantes de k-means avec $Q^\top Q = I$,

$$\mathcal{J}_{BMA-kM}(\mathbf{\Pi}) = ||A - \mathbf{\Pi}A||^2, \tag{6}$$

$$\mathcal{J}_{BMA-FKM}(\mathbf{\Pi}) = ||AQ - \mathbf{\Pi}AQ||^2, \tag{7}$$

$$\mathcal{J}_{BMA-RKM}(\mathbf{\Pi}) = ||A - \mathbf{\Pi}AQQ^\top||^2, \tag{8}$$

Dans la suite de l'article, nous ne considérerons que les contraintes de non-négativité, de symétrie et bi-stochastiques.

## 3.3 Equivalence entre BMA et k-means

Le théorème ci-dessous démontre que l'optimisation de l'objectif k-means et de l'objectif BMA sous certaines contraintes appropriées sont équivalentes. L'équation (9) établit l'équivalence entre k-means et la formulation BMA sous les contraintes $\{\mathbf{\Pi} \geq 0, \mathbf{\Pi} = \mathbf{\Pi}^\top, \mathbf{\Pi}\mathbb{1} = \mathbb{1}, Tr(\mathbf{\Pi}) = k, \mathbf{\Pi}\mathbf{\Pi}^\top = \mathbf{\Pi}\}$

**Theorem 1**

$$\arg\min_{R,S} ||A - RS^\top||^2 \Leftrightarrow \arg\min ||A - \mathbf{\Pi}A||^2 \tag{9}$$

Cette nouvelle formulation donne quelques points saillants intéressants sur k-means et ses variantes :

- Tout d'abord, elle montre que k-means équivaut à apprendre une matrice de similarité bi-stochastique diagonale par blocs.
- Deuxièmement, elle permet d'établir des connexions très intéressantes de k-means avec de nombreuses méthodes de *subspace clustering* [4, 2]. De plus, cette formulation combine le processus traditionnel en deux étapes utilisées par les méthodes de *subspace clustering*, qui consiste à construire d'abord une matrice d'affinité entre les points de données, puis à appliquer un *spectral clustering* à cette matrice. Cela permet l'apprentissage conjoint d'une matrice de similarité qui reflète mieux la structure de clustering par sa forme diagonale par blocs.
- Enfin, elle permet d'appliquer l'esprit des k-means pour les données de graphe ou de similarité.

# 4 Conclusion

Dans ce travail nous proposons une autre formulation matricielle de k-means et ses variantes. Cette formulation est intéressante à divers niveaux. Sur des applications réelles, cette approche montre son intérêt.

# Références


[1] Geert De Soete and J Douglas Carroll. K-means clustering in a low-dimensional euclidean space. E. Diday et al. (Eds.), New approaches in classification and data analysis, Berlin, Springer-Verlag, pages 212–219. 1994.

[2] Derek Lim, René Vidal, and Benjamin D. Haeffele. Doubly stochastic subspace clustering. *ArXiv*, abs/2011.14859, 2020.

[3] Maurizio Vichi and Henk AL Kiers. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37(1) :49–64, 2001.

[4] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2) :52–68, 2011.

[5] Ron Zass and Amnon Shashua. A unifying approach to hard and probabilistic clustering. In *ICCV*, pages 294–301, 2005.

# Classification et segmentation conjointes de données fonctionnelles, une approche par blocs latents

Paul Riverain[1], Allou Samé[1], Latifa Oukhellou[1]

[1] Université Gustave Eiffel, COSYS-GRETTIA, 77420, Champs-sur-Marne, France

{paul.riverain,allou.same,latifa.oukhellou}@univ-eiffel.fr

## Résumé

*Cet article aborde le problème de la classification et de la segmentation simultanées de courbes. Nous proposons un modèle qui s'inspire à la fois d'un mélange de lois dédié à la classification de courbes à changements de régimes, et du modèle des blocs latents utilisé pour la classification croisée. Cette nouvelle formulation mène à une estimation de paramètres par un algorithme EM variationnel maximisant une borne inférieure de la log-vraisemblance. Des expériences numériques menées sur des données synthétiques permettent d'évaluer la méthode proposée.*

## Mots-clés

*Données fonctionnelles, clustering et segmentation, modèles de mélange, blocs latents*

## Abstract

*This paper addresses the problem of curve clustering and segmentation. We propose a model inspired by a mixture model dedicated to clustering curves with regime changes and the latent block model used in co-clustering. For this new formulation, a variational EM algorithm maximizing a lower bound of the log-likelihood is used for parameter estimation. Numerical experiments on synthetic data are used to evaluate the proposed method.*

## Keywords

*Functional data, clustering and segmentation, mixture model, latent block*

## 1 Introduction

Dans de nombreux domaines applicatifs, les observations à analyser se présentent sous la forme d'un ensemble de courbes discrétisées. Les méthodes de clustering de données fonctionnelles [1, 4] constituent dans ce cadre des outils de référence. Elles s'appuient généralement sur une représentation des données dans des bases de fonctions adaptées (ex. splines, ondelettes, Fourier) [4].

Outre le partitionnement de courbes, il peut être également utile de fournir une vue synthétique des régimes temporels locaux caractérisant les classes de courbes. Dans cette optique, les travaux initiés dans [5] ont proposé une approche dénommée ClustSeg dédiée à la classification de courbes à changements de régime, qui opère simultanément le partitionnement et la segmentation de courbes. Celle-ci repose sur un modèle de mélange à deux niveaux hiérarchiques : un mélange global associé au partitionnement de courbes, dont les classes sont formalisées par un mélange local à proportions variables traduisant la segmentation du temps.

Nous abordons dans cet article le même problème de classification et de segmentation de courbes, mais à travers une formulation générative plus contrainte faisant intervenir un nombre réduit de variables latentes. Cette réduction du nombre de variables latentes s'effectue néanmoins au prix d'une estimation variationnelle des paramètres, l'inférence exacte par maximum de vraisemblance devenant infaisable d'un point de vue combinatoire. La méthode proposée possède également des liens avec le modèle des blocs latents qui est dédié à la classification croisée [2]. Elle a l'avantage d'être plus rapide que l'approche ClustSeg et de rendre plus robuste l'estimation des transitions entre segments.

L'article est organisé comme suit. La partie 2 décrit la nouvelle approche de classification et de segmentation proposée, dénommée BlockSeg, en la positionnant par rapport à l'approche de référence ClustSeg [5]. La partie 3 est dédiée à des expérimentations numériques permettant d'évaluer les performances de cette méthode.

## 2 Modèle à blocs latents pour la classification et la segmentation

On désigne par $(\boldsymbol{x}_i)_{i=1,\dots,n}$ $n$ courbes observées sur la même grille temporelle indicée par $j \in \{1,\dots,T\}$, avec $\boldsymbol{x}_i = (x_{ij})_{j=1,\dots,T}$, où $x_{ij} \in \mathbb{R}$. Ces courbes sont supposées être formées de $K$ classes désignées par les variables latentes $(z_1,\dots,z_n)$, avec $z_i = k$ si $\boldsymbol{x}_i$ appartient à la classe $k$. On utilisera de manière équivalente les variables binaires $z_{ik}$ valant 1 si $z_i = k$ et 0 sinon. Au sein d'une même classe, les courbes sont supposées être formées de $L$ régimes définissant une segmentation temporelle des points $x_{ij}$, dont les labels sont notés $w_{ij} \in \{1,\dots,L\}$. Avant de présenter la nouvelle approche, nous commençons par rappeler l'approche ClustSeg [5] qui a servi de référence à ce travail, les deux approches se différenciant par leur mode de génération des variables latentes $w_{ij}$.

(a) Modèle ClustSeg (b) Modèle BlockSeg proposé

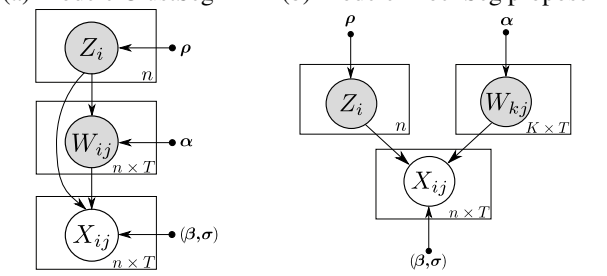



FIGURE 1 – Modèles graphiques probabilistes associés aux deux approches de classification-segmentation ; les variables latentes sont représentées en fond gris

## 2.1 Rappel sur l'approche ClustSeg

La méthode de classification et de segmentation proposée dans [5] est basée sur le modèle génératif suivant :

$$\left\{\begin{array}{lcl} z_i & \sim & \mathrm{Cat}\big((\rho_k)_{k=1,\dots,K}\big) \\ w_{ij}|z_i=k & \sim & \mathrm{Cat}\big((\pi_{j\ell}(\boldsymbol{\alpha}_k))_{\ell=1,\dots,L}\big) \\ x_{ij}|z_i=k,w_{ij}=\ell & \sim & \mathcal{N}\big(\boldsymbol{\beta}'_{k\ell}\boldsymbol{u}_j,\sigma^2_{k\ell}\big), \end{array}\right. \quad (1)$$

qui est décrit sous forme graphique par la figure 1(a). La notation Cat fait ici référence à une loi catégorielle, les $\rho_k$ sont les proportions du mélange vérifiant $\sum_k \rho_k = 1$ et les $\pi_{j\ell}(\boldsymbol{\alpha}_k)$ sont des transformations logistiques de fonctions linéaires du temps définies par

$$\begin{aligned} \pi_{j\ell}(\boldsymbol{\alpha}_k) &= p(w_{ij}=\ell|z_i=k) \\ &= \frac{\exp(\alpha_{k\ell 0}+\alpha_{k\ell 1}\,j)}{\sum_h \exp(\alpha_{kh0}+\alpha_{kh1}\,j)}, \end{aligned}$$

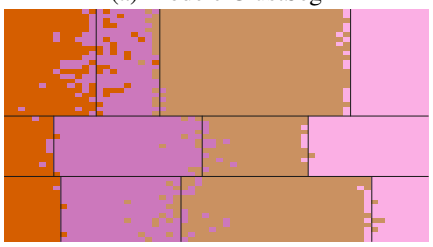
avec $\boldsymbol{\alpha}_k = (\alpha_{k\ell 0},\alpha_{k\ell 1})_{\ell=1,\dots,L}$ (pour des questions d'identifiabilité, on pose $(\alpha_{kL0},\alpha_{kL1}) = (0,0)$). Notons que l'usage de telles fonctions logistiques mène à un partitionnement local des points de chaque classe en $L$ segments temporel contigus [5]. Pour prendre en compte le caractère fonctionnel des données, chaque segment $(k,\ell)$ est modélisé par une régression linéaire par rapport à une base de fonctions donnée par les covariables $\boldsymbol{u}_j \in \mathbb{R}^p$, dont le vecteur des coefficients de régression est $\boldsymbol{\beta}_{k\ell} \in \mathbb{R}^p$ et la variance $\sigma^2_{k\ell}$. La figure 2(a) illustre la structure des classes et des segments latents issus de ce modèle.

On peut montrer que les courbes $\boldsymbol{x}_i$ sont générées indépendamment suivant le mélange de lois suivant :

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_k \rho_k \Big(\prod_j \sum_\ell \pi_{j\ell}(\boldsymbol{\alpha}_k)\mathcal{N}(x_{ij},\boldsymbol{\beta}'_{k\ell}\boldsymbol{u}_j,\sigma^2_{k\ell})\Big)$$

de paramètre $\boldsymbol{\theta} = \big((\rho_k)_k,(\boldsymbol{\alpha}_k)_k,(\boldsymbol{\beta}_{k\ell})_{k\ell},(\sigma^2_{k\ell})_{k\ell}\big)$ ; ce qui mène naturellement à une estimation du paramètre $\boldsymbol{\theta}$ par la méthode du maximum de vraisemblance via l'algorithme EM [5]. La partition ainsi que la segmentation finales sont déduites de cette estimation.

(a) Modèle ClustSeg



(b) Modèle BlockSeg proposé

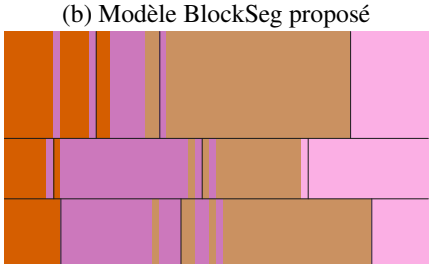


FIGURE 2 – Exemples de partition en $K = 3$ classes et $L = 4$ segments simulées à partir du modèle de référence ClustSeg et du modèle proposé ; les lignes horizontales délimitent les classes de courbes et les lignes verticales définissent la segmentation

## 2.2 Modèle BlockSeg proposé

La nouvelle formalisation proposée se distingue du modèle initial par le mode de génération des variables latentes relatives à la segmentation. Celles-ci ne sont plus spécifiques à chaque courbe mais plutôt spécifiques à chaque classe. Partant du même paramètre $\boldsymbol{\theta}$, le modèle est défini comme suit :

$$\left\{\begin{array}{lcl} z_i & \sim & \mathrm{Cat}((\rho_k)_{k=1,\dots,K}) \\ w_{kj} & \sim & \mathrm{Cat}\big((\pi_{j\ell}(\boldsymbol{\alpha}_k))_{\ell=1,\dots,L}\big) \\ x_{ij}|z_i=k,w_{kj}=\ell & \sim & \mathcal{N}(\boldsymbol{\beta}'_{k\ell}\boldsymbol{u}_j,\sigma^2_{k\ell}) \end{array}\right. \quad (2)$$

Ce modèle fait intervenir un nombre réduit de variables latentes, à savoir, pour chaque classe $k$, un unique processus latent défini par $(w_{kj})_{j=1,\dots,T}$, avec $w_{kj} \in \{1,\dots,L\}$, et dont la loi a priori est définie indépendamment des variables $z_i$. On utilisera de manière équivalente les variables binaires $w_{kj\ell}$. La figure 1(b) fournit le modèle graphique correspondant à cette nouvelle formulation et la figure 2(b) illustre la structure des variables latentes de ce modèle qui sera appelé BlockSeg. Nous verrons dans les expérimentations numériques que le modèle BlockSeg peut s'avérer plus rapide et plus robuste que ClustSeg, notamment dans l'estimation des changements de régime.

Pour ClustSeg, une transition entre deux segments peut être lente pour deux raisons : si le passage d'un régime à un autre a lieu à différents instants pour chaque courbe, ou si, à la frontière des deux segments, les courbes se comportent selon un mélange du régime précédent et du régime suivant. Les transitions de BlockSeg ne seront affectées que par la deuxième situation.

Si on considère la variante du modèle où la segmentation est commune à l'ensemble des classes ($\boldsymbol{\alpha}_k = \boldsymbol{\alpha}$ et

$w_{kj} = w_j \sim \mathrm{Cat}\big((\pi_{j\ell}(\boldsymbol{\alpha}))_{\ell=1,\dots,L}\big)$, on obtient ainsi un modèle similaire au modèle des blocs latents [2], où le partitionnement en colonne (segmentation) est structuré par les proportions logistiques.

### 2.3 Estimation variationnelle des paramètres

La structure en blocs latents du modèle proposé ne permet pas de réaliser l'inférence de manière exacte. Comme pour le modèle des blocs latents [2], nous nous sommes appuyés dans cet article sur une stratégie variationnelle d'estimation des paramètres. Dans notre situation, cela consiste à maximiser l'approximation suivante de la log-vraisemblance :

$$\widetilde{L}(\boldsymbol{\theta}) = \max_{\widetilde{\mathbf{z}},\widetilde{\mathbf{w}}} \Big[ L_C(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}};\boldsymbol{\theta}) + H(\widetilde{\mathbf{z}}) + H(\widetilde{\mathbf{w}}) \Big], \quad (3)$$

où $\widetilde{\mathbf{z}} = (\widetilde{z}_{ik})$ et $\widetilde{\mathbf{w}} = (\widetilde{w}_{kj\ell})$ sont des distributions des variables latentes, $L_C$ est la vraisemblance complétée du modèle, et $H(\widetilde{\mathbf{z}})$ et $H(\widetilde{\mathbf{w}})$ sont les entropies respectives des distributions $\widetilde{\mathbf{z}}$ et $\widetilde{\mathbf{w}}$. Ces dernières quantités s'écrivent

$$\begin{aligned} L_C(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}};\boldsymbol{\theta}) = & \textstyle\sum_{i,k} \widetilde{z}_{ik} \log \rho_k + \sum_{k,j,\ell} \widetilde{w}_{kj\ell} \log \pi_{j\ell}(\boldsymbol{\alpha}_k) \\ & + \textstyle\sum_{i,j,k,\ell} \widetilde{z}_{ik}\widetilde{w}_{kj\ell} \log \mathcal{N}(x_{ij}; \boldsymbol{\beta}'_{k\ell}\mathbf{u}_j, \sigma^2_{k\ell}) \end{aligned}$$

$$H(\widetilde{\mathbf{z}}) = -\sum_{i,k} \widetilde{z}_{ik} \log \widetilde{z}_{ik}, \quad H(\widetilde{\mathbf{w}}) = -\sum_{j,k,\ell} \widetilde{w}_{kj\ell} \log \widetilde{w}_{kj\ell}$$

La maximisation du critère variationnel défini par l'équation (3) est finalement obtenue par un algorithme Variational Expectation Maximization (VEM) qui alterne jusqu'à la convergence les deux étapes suivantes, en démarrant de paramètres initiaux $\boldsymbol{\theta}, \widetilde{\mathbf{z}}, \widetilde{\mathbf{w}}$. Les étapes VE et M sont effectuées conditionnellement aux classifications des lignes ou des colonnes, dans le même ordre que celui défini par l'algorithme BEM2 [3].

**Etape VE** : maximisation alternée par rapport à $\widetilde{\mathbf{z}}$ et $\widetilde{\mathbf{w}}$

$$\widetilde{z}_{ik} \propto \rho_k \prod_{j,\ell} \left[\mathcal{N}(x_{ij}; \boldsymbol{\beta}'_{k\ell}\mathbf{u}_j; \sigma^2_{k\ell})\right]^{\widetilde{w}_{kj\ell}} \quad (4)$$

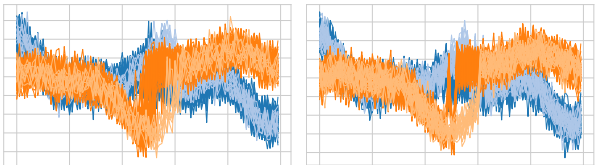
$$\widetilde{w}_{kj\ell} \propto \pi_{j\ell}(\boldsymbol{\alpha}_k) \prod_{i} \left[\mathcal{N}(x_{ij}; \boldsymbol{\beta}'_{k\ell}\mathbf{u}_j; \sigma^2_{k\ell})\right]^{\widetilde{z}_{ik}} \quad (5)$$

**Etape M** : maximisation par rapport à $\boldsymbol{\theta}$

$$\begin{aligned} \rho_k &= \frac{\sum_i \widetilde{z}_{ik}}{n} \\ \boldsymbol{\alpha}_k &= \mathrm{argmax}_{\boldsymbol{\alpha}} \sum_{i,j,\ell} \widetilde{z}_{ik}\widetilde{w}_{kj\ell} \log \pi_{j\ell}(\boldsymbol{\alpha}) \\ \boldsymbol{\beta}_{k\ell} &= \Big[\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{kj\ell}\mathbf{u}_j\mathbf{u}'_j\Big]^{-1} \Big[\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{kj\ell}x_{ij}\mathbf{u}_j\Big] \\ \sigma^2_{k\ell} &= \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{kj\ell}\left(x_{ij} - \boldsymbol{\beta}'_{k\ell}\mathbf{u}_j\right)^2}{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{kj\ell}}, \end{aligned}$$

où l'équation (6) est un problème de régression logistique pondéré, résolu par l'algorithme IRLS [5].

A l'issue de l'estimation des paramètres, la partition des courbes est obtenue en rangeant chaque courbe dans la classe maximisant les probabilités a posteriori $(\widetilde{z}_{ik})_{k=1,\dots,K}$. La segmentation des classes peut être obtenue de deux manières : en maximisant les probabilités a posteriori $(\widetilde{w}_{kj\ell})_{\ell=1,\dots,L}$, ou en maximisant les probabilités a priori $(\pi_{j\ell}(\boldsymbol{\alpha}_k))_{\ell=1,\dots,L}$. Dans ce dernier cas, les segments obtenus sont contigus.



(a) ClustSeg (b) BlockSeg

FIGURE 3 – Données simulées suivant les modèles ClustSeg et BlockSeg pour 10% d'outliers

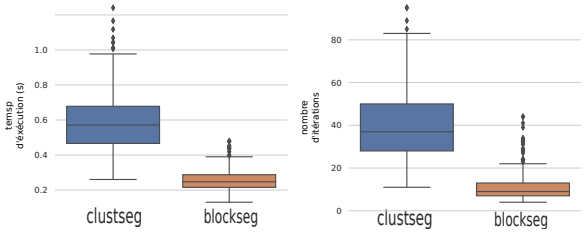



FIGURE 4 – Temps d'exécution et nombre d'itérations pour ClustSeg et BlockSeg et 0% d'outliers

## 3 Expérimentations numériques

**Protocole expérimental** On compare ici les performances de BlockSeg à celles de ClustSeg en terme de temps d'exécution, de nombre d'itérations, de classification et de segmentation. Pour cela, on considère des données simulées à partir de ces deux modèles.

On cherche notamment à comparer la sensibilité des deux algorithmes au bruit dans la segmentation. Pour cela, on échantillonne $K = 4$ classes et $L = 2$ segments, (de même variance $\sigma_{kl} = \sigma$), où 2 classes ne diffèrent que par leur structure en segments : $\boldsymbol{\beta}_{1l} = \boldsymbol{\beta}_{2l}$, $\boldsymbol{\beta}_{3l} = \boldsymbol{\beta}_{4l}$ et $\boldsymbol{\alpha}$ est tel que toutes les vitesses de transition sont identiques et tel que les classes 1 et 3 ont un point de changement à $T/2$ et les classes 2 et 4 à $3T/5$. Les proportions de mélange $\rho_k$ sont égales pour les classes 1 et 3, qui correspondent aux deux classes majoritaires, et pour les classes minoritaires 2 et 4 qui jouent le rôle d'outliers. La figure 3 montre un exemple de courbes simulées à partir de chacun des modèles, avec une proportion de 10% d'outliers.

On considère $n = 200$ courbes sur $T = 200$ instants et l'on échantillonne 200 jeux de données complètes pour chaque proportions d'outliers dans $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. La base de fonction $(\mathbf{u}_j)_j$ est composée des 4 premières harmoniques de Fourier. Pour chaque jeu de données, les paramètres $\boldsymbol{\beta}$ sont échantillonnés à partir d'une loi normale centrée

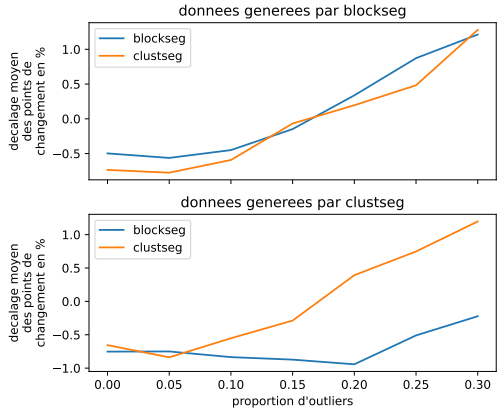



FIGURE 5 – Écart moyen entre les points de changement estimés et le point de changement de la classe majoritaire pour BlockSeg et ClustSeg en fonction de la proportion d'outliers

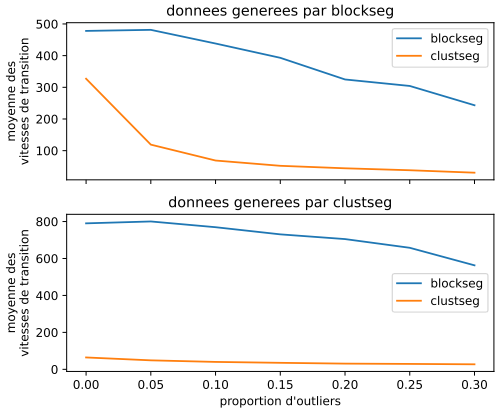



FIGURE 6 – Vitesse moyenne de transition pour BlockSeg et ClustSeg en fonction de la proportion d'outliers

réduite. On exécute les algorithmes pour $K = 2$ classes et $L = 2$ segments de façon à ce que les courbes de chaque classe majoritaire soient classifiées avec celles de la classe minoritaire correspondante.

**Résultats** Tout d'abord, on mesure des ARI en classification, qui sont en moyenne supérieurs à $99,8\%$ pour toutes les configurations. On note toutefois quelques valeurs extrêmes, de l'ordre de $50\%$ pour BlockSeg. Sur la figure 4, on compare les temps d'exécution et nombre d'itérations pour les deux approches. On observe que BlockSeg converge en un nombre d'itérations plus faible que ClustSeg, ce qui rend en général son exécution plus rapide.

On compare, sur la figure 5, la façon dont le point de changement des classes majoritaires est affecté par les outliers. On observe que lorsque les données sont générées par BlockSeg, les deux modèles renvoient des segmentations dont les points de changement se décalent vers la droite de façon similaire. Lorsque les données sont générées par ClustSeg, les points de changement des segmentations se décalent plus pour ClustSeg que pour BlockSeg. BlockSeg semble donc plus robuste que ClustSeg à la variabilité dans le point de changement si celle-ci n'est pas commune à toutes les courbes.

Enfin, on observe sur la figure 6 que dans les deux configurations, les vitesses de transition estimées par BlockSeg sont plus grandes que celles estimées par ClustSeg. BlockSeg tend donc à mieux séparer les régimes que ClustSeg, comme l'illustre la figure 7.

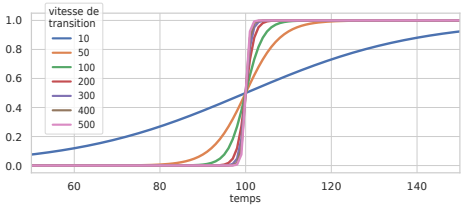



FIGURE 7 – Probabilités $j \mapsto \pi_{j2}(\boldsymbol{\alpha}_k)$ pour différentes vitesses de transition

## 4 Conclusion et perspectives

Ainsi, nous avons présenté BlockSeg, un modèle de classification et de segmentation de courbes, à travers une formulation générative contrainte étroitement liée au modèle des blocs latents et avons proposé un algorithme variationnel pour l'estimation des paramètres.

Nos expériences numériques ont mis en évidence que BlockSeg, grâce à un espace latent réduit, converge plus rapidement que ClustSeg, bien qu'il semble marginalement plus soumis à des maxima locaux de la fonction de vraisemblance, et ce, probablement en raison de l'approche variationnelle. Aussi, BlockSeg apparaît plus robuste que ClustSeg à une variabilité dans les points de changement et permet d'obtenir des segments mieux séparés.

## Références

[1] Emilie Devijver, Yannig Goude, and Jean-Michel Poggi. Clustering electricity consumers using high-dimensional regression mixture models. *Applied Stochastic Models in Business and Industry*, 36(1) :159–177, 2020.

[2] Gerard Govaert and Mohamed Nadif. *Co-clustering : models, algorithms and applications*. Wiley, 2013.

[3] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6) :3233–3245, 2008.

[4] Julien Jacques and Cristian Preda. Functional data clustering : a survey. *Advances in Data Analysis and Classification*, 8 :231–255, 2014.

[5] Allou Samé, Faicel Chamroukhi, Gérard Govaert, and Patrice Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5(4) :301–321, 2011.

# Classifications multi-niveaux et convexités d'intervalle : cas de la hiérarchie du lien simple

P. Bertrand[1], J. Diatta[2]

[1] Université Paris Dauphine-PSL, CEREMADE
[2] Université de La Réunion, LIM

patrice.bertrand@ceremade.dauphine.fr, jean.diatta@univ-reunion.fr

## Résumé

*Il existe plusieurs caractérisations de certaines classifications multi-niveaux, l'une d'elles les identifiant à des collections de sous-ensembles qui sont non vides et convexes selon un type de fonction d'intervalle. Nous proposons : (a) de nouvelles caractérisations des hiérarchies et des hiérarchies faibles en tant que convexités d'intervalle, (b) des fonctions d'intervalle qui induisent des classifications hiérarchiques connues, (c) une suite de convexités d'intervalle emboitées qui croît progressivement de la hiérarchie d'Apresjan à la hiérarchie du lien simple.*

## Mots-clés

*Hiérarchie d'Apresjan, hiérarchie faible, dissimilarités basées sur des chemins*

## Abstract

*There are several ways to characterize some types of multilevel clustering, one being as collections of nonempty subsets that are convex according to a type of interval function. We propose : (a) New characterizations of hierarchies and weak hierarchies as interval convexities, (b) Interval functions which induce known hierarchical clustering schemes, (c) A sequence of nested families of interval convexities that is gradually increasing from the Apresjan hierarchy to the Single-Link hierarchy.*

## Keywords

*Apresjan hierarchy, weak hierarchy, path-based dissimilarities*

# 1 Introduction

La classification est une approche classique de fouille de données qui vise à révéler une structure de l'ensemble de données, généralement exprimée sous la forme d'une collection de sous-ensembles homogènes appelés "classes". Selon Mirkin et Muchnick [15], il existe trois approches principales pour déterminer une classe, qui sont fondées sur, respectivement, une définition précise de la notion de classe, un algorithme de classification et un critère de classification à optimiser. Dans ce qui suit, nous nous intéressons principalement aux classes qui sont définies comme étant les parties convexes (non vides) au sens d'une fonction d'intervalle. Une fonction d'intervalle $I$ sur un ensemble de données $S$ est une fonction symétrique qui associe à chaque couple $(x, y) \in S \times S$ un sous-ensemble $I(x, y)$ de $S$, contenant $x$ et $y$. Le sous-ensemble $I(x, y)$ est appelé $I$-intervalle d'extrémités $x$ et $y$. Un sous-ensemble $A$ de $S$ est dit convexe au sens de la fonction d'intervalle $I$, ou $I$-convexe, s'il contient chaque $I$-intervalle dont les extrémités appartiennent à $A$. La collection $\mathscr{C}_I$ des sous-ensembles $I$-convexes de $S$ forme une convexité appelée convexité d'intervalle induite par $I$. Il s'avère que la notion de convexité définie au sens le plus abstrait (par exemple [10, 16]), coïncide avec celle de classification multi-niveaux, lorsque cette classification est supposée fermée par intersections arbitraires. Une classification multi-niveaux de $S$ est une collection de sous-ensembles non vides de S, contenant $S$ lui-même et dont au moins deux membres sont strictement emboîtés. Une structure de classification multi-niveaux très connue est la structure hiérarchique. Une classification hiérarchique sur $S$ peut être définie comme une union de partitions de $S$ de moins en moins fines, allant de la partition discrète (i.e. l'ensemble des singletons $\{x\}$ avec $x \in S$) à la partition grossière (i.e. la partition réduite au singleton $\{S\}$). Une hiérarchie est aussi une classification multi-niveaux pour laquelle l'intersection de deux classes quelconques est soit vide, soit égale à l'une d'entre elles. Une hiérarchie faible est une classification multi-niveaux pour laquelle l'intersection de trois classes quelconques se réduit à l'intersection de deux d'entre elles [2, 3, 12]. Il s'agit donc d'une extension directe de la structure de classification hiérarchique, qui permet un type de recouvrement des classes. Au cours des dernières décennies, plusieurs auteurs ont étudié diverses généralisations du modèle hiérarchique, telles que les pyramides [13, 14], les hiérarchies par paires [5, 6] et les hypergraphes totalement équilibrés [9, 8]. Ces modèles sont des sous-modèles du modèle de hiérarchie faible.

# 2 Caractérisation des hiérarchies et des hiérarchies faibles comme convexités d'intervalle

Dans la lignée des recherches menées par [7, 11], nous proposons de nouvelles caractérisations flexibles des mo-

dèles hiérarchique et faiblement hiérarchique en tant que convexités d'intervalle. Plus précisément, nous considérons une application arbitraire $g : S \times S \longrightarrow 2^S$ telle que $\{x, y\} \subseteq g(x, y)$ pour tout $x, y \in S$. En notant $g(y, x)$ par $\overline{g}(x, y)$ pour tout $x, y \in S$, il en résulte que $J_g = g \cup \overline{g}$ (resp. $M_g = g \cap \overline{g}$) est une fonction d'intervalle. De plus, nous dirons que $g$ satisfait (H) et (W), respectivement, si :

(H) pour tout $x_1, x_2, x_3 \in S$,
$g(x_1, x_2) \subseteq g(x_1, x_3)$ ou $g(x_1, x_3) \subseteq g(x_1, x_2)$.

(W) pour tout $x_1, x_2, x_3 \in S$, il existe $i, j, k$ avec $\{i, j, k\} = \{1, 2, 3\}$, tel que :
$g(x_i, x_j) \subseteq g(x_i, x_k)$ et $g(x_k, x_j) \subseteq g(x_k, x_i)$.

Nous montrons alors que $J_g$ (resp. $M_g$) induit une structure hiérarchique (resp. faiblement hiérarchique), si et seulement si $g$ satisfait la propriété $(H)$ (resp. (W)).
Considérons alors une dissimilarité $\delta$ définie sur $S$.
Soit $g_{B_\delta} : S \times S \longrightarrow 2^S$ l'application qui est definie pour tout $x, y \in S$, par :

$$g_{B_\delta}(x, y) = B_\delta(x, \delta(x, y)) = \{s \in S \mid \delta(x, s) \leq \delta(x, y)\}.$$

S'il n'y a pas d'ambiguité sur le choix de la dissimilarité $\delta$, l'application $g_{B_\delta}$ est notée $g_B$. Etant donné une dissimilarité $\delta$ sur $S$, Apresjan [1] a considéré les parties $C$ de $S$ tels que pour tout $x, y \in C$ :

$$\delta(x, y) < \min_{z \notin C} \big(\min\{\delta(x, z), \delta(y, z)\}\big). \tag{1}$$

Un sous-ensemble non vide est appelé *classe d'Apresjan* de $\delta$, s'il vérifie (1). Les classes d'Apresjan d'une dissimilarité $\delta$ forment une hiérarchie [1], appelée la *hiérarchie d'Apresjan de* $\delta$ [4]. Bandelt and Dress [3] et Bandelt [2], utilisent un critère plus faible que (1), pour définir la notion de classe faible : une *classe faible* est une partie $C$ non vide de $S$ telle que pour tout $x, y \in C$,

$$\delta(x, y) < \min_{z \notin C} \big(\max\{\delta(x, z), \delta(y, z)\}\big). \tag{2}$$

Par la suite, une partie $C$ non vide de $S$ est appelée *classe de Bandelt et Dress de* $\delta$, si elle vérifie (2). Les classes d'Apresjan d'une dissimilarité $\delta$ forment une hiérarchie faible [3], appelée la *hiérarchie faible de Bandelt et Dress de* $\delta$.
Avec les définitions précédentes, nous montrons que la fonction d'intervalle $J_{g_B}$ (resp. $M_{g_B}$) induit la hiérarchie d'Apresjan (resp. la hiérarchie faible de Bandelt et de Dress) de $\delta$.

# 3 Filtrations de la hiérarchie du lien simple et de la hiérarchie faible de Bandelt et Dress

Nous considérons les dissimilarités, notées $\delta_\ell$ avec $\ell \geq 1$, qui sont basées sur des chemins, également connues sous le nom de distances transitives d'ordre $\ell$ [17, 19, 18]. Pour tous $x, y \in S$, la valeur $\delta_\ell(x, y)$ est égale au plus petit saut maximum de $\delta$ le long de tous les chemins de longueur au plus $\ell$, joignant $x$ et $y$ dans le graphe complet induit par $\delta$. Tout d'abord, nous avons déterminé une méthode récursive pour calculer la séquence $(\delta_\ell)_{1 \leq \ell \leq n-1}$ qui s'avère être strictement décroissante jusqu'à un certain rang $r(\delta)$ de $\ell$, puis stationnaire et coïncidant avec l'ultramétrique sous-dominante de $\delta$ à partir du rang $r(\delta)$. Enfin, nous définissons $g_\ell$ comme l'application boule de $\delta_\ell$, i.e. $g_\ell = g_{B_\ell}$ avec $B_\ell = B_{\delta_\ell}$, nous avons prouvé que les séquences dont les termes généraux sont, d'une part, $\mathcal{H}_\ell(\delta) = \mathrm{conv}(J_{g_\ell})$ et, d'autre part, $\mathcal{W}_\ell(\delta) = \mathrm{conv}(M_{g_1}) \cap \Big[\bigcup_{k \leq \ell} \mathrm{conv}(M_{g_{n-k}})\Big]$, avec $1 \leq \ell < n - 1$, sont respectivement une filtration de la hiérarchie du lien simple de $\delta$ et une filtration de la hiérarchie faible de Bandelt et Dress de $\delta$.
Ces résultats théoriques peuvent être considérés comme une étape vers l'introduction d'outils fondés sur la convexité d'intervalle pour réaliser et interpréter une classification multi-niveaux basée sur les dissimilarités. Nous concluons cette présentation en soulignant quelques remarques finales et en discutant des extensions potentielles de cette approche, en particulier pour la pratique de l'exploration de données impliquant la classification hiérarchique du lien simple.

# Références


[1] J. Apresjan. An algorithm for constructing clusters from a distance matrix. *Mashinnyi perevod : prikladnaja lingvistika*, 9 :3–18, 1966.

[2] H. J. Bandelt. Four point characterization of the dissimilarity functions obtained from indexed closed weak hierarchies. Technical report, Mathematiche Seminar der Universität, Hamburg, 1992.

[3] H.-J. Bandelt and A.W.M. Dress. Weak hierarchies associated with similarity measures : an additive clustering technique. *Bull. Math. Biology*, 51 :133–166, 1989.

[4] J.-P. Benzécri. *L'Analyse des Données*. Dunod, Paris, 1973.

[5] P. Bertrand. Set systems for which each set properly intersects at most one other set - Application to cluster analysis. *Discrete Applied Mathematics*, 156(8) :1220–1236, 2008.

[6] P. Bertrand and F. Brucker. On lower-maximal paired-ultrametrics. In P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho, editors, *Selected Contributions in Data Analysis and Classification*, pages 455–464. Springer-Verlag, 2007.

[7] P. Bertrand and J. Diatta. Multilevel clustering models and interval convexities. *Discrete Applied Mathematics*, 222 :54–66, 2017.

[8] F. Brucker and A. Gély. Crown-free lattices and their related graphs. *Order*, 28(3) :443–454, 2011.

[9] F. Brucker, P. Préa, and C. Châtel. Totally balanced dissimilarities. *Journal of Classification*, 2019 (accessed March 30, 2019). https ://doi.org/10.1007/s00357-019-09320-w.

[10] J. Calder. Some elementary properties of interval convexities. *J. Lond. Math. Soc.*, s2-3(3) :422—428, 1971.

[11] M. Changat, P.-G. Narasimha-Shenoi, and P.-F. Stadler. Axiomatic characterization of transit functions of weak hierarchies. *The Art of Discrete and Applied Mathematics*, 2 :#P1.01, 2019.

[12] J. Diatta and B. Fichet. Quasi-ultrametrics and their 2-ball hypergraphs. *Discrete Math.*, 192 :87–102, 1998.

[13] E. Diday. Orders and overlapping clusters in pyramids. In Jan De Leeuw et al., editor, *Multidimensional Data Analysis*, pages 201–234. DSWO Press, 1986.

[14] C. Durand and B. Fichet. One-to-one correspondence in pyramidal representation : a unified approach. In H. H. Bock, editor, *Classification and Related Methods of Data Analysis*, pages 85–90. North-Holland, 1988.

[15] B. Mirkin and I. Muchnik. Combinatorial optimization in clustering. In D.-Z. Du and P. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 2, pages 261–329. Kluwer Academic Publishers, 1998.

[16] M. Van de Vel. *Theory of Convex Structures*. Elsevier, Amsterdam, North-Holland, 1993.

[17] Chunjing Xu, Jianzhuang Liu, and Xiaoou Tang. Clustering with transitive distance and k-means duality. *ArXiv*, abs/0711.3594, 2007.

[18] Zhiding Yu, Weiyang Liu, Wenbo Liu, Yingzhen Yang, Ming Li, and B. V. K. Vijaya Kumar. On order-constrained transitive distance clustering. In *AAAI*, 2016.

[19] Zhiding Yu, Chunjing Xu, Deyu Meng, Zhuo Hui, Fanyi Xiao, Wenbo Liu, and Jianzhuang Liu. Transitive distance clustering with k-means duality. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 987–994, 06 2014.

# Classification Topologique sur Données Evolutives

R. Abdesselam

Université Lumière Lyon 2, Laboratoires ERIC & COACTIS

rafik.abdesselam@univ-lyon2.fr

## Résumé

*L'objectif de cet article est de proposer une approche de classification topologique sur données évolutives. Nous nous intéressons à la classification qui résulte de méthodes exploratoires pour l'analyse conjointe de plusieurs tableaux de données ; plus spécifiquement, les méthodes appliquées aux données temporelles.*

*La classification est l'une des approches les plus largement utilisées pour explorer des données multidimensionnelles. Deux stratégies de classifications non supervisées courantes sont la classification ascendante hiérarchique (CAH) et la méthode k-means, utilisés pour identifier des groupes d'objets similaires dans un jeu de données afin de le partitionner en groupes homogènes. L'approche proposée, dite classification topologique sur données évolutives (CTDE), est basée sur la notion de graphes de voisinage dans un contexte de données évolutives. Elle permet d'explorer simultanément plusieurs tableaux de données collectées à des moments différents sur les mêmes lignes-individus, même dans les cas où les variables sont différentes dans les tableaux considérés. Les variables de chaque tableau sont plus ou moins corrélées ou liées selon le type de variables. Dans chaque tableau, la CTDE permet d'analyser la structure des corrélations ou associations observées entre les variables selon qu'elles soient de type quantitatif, qualitatif ou un mélange des deux.*

*L'approche CTDE proposée est présentée et illustrée ici à l'aide d'un ensemble de données réelles avec des variables quantitatives. Les résultats sont comparés à ceux issus de la classification sur les résultats de l'analyse factorielle multiple (AFM) sur données évolutives.*

## Mots-clés

*Données évolutives, mesure de proximité, graphe de voisinage, matrice d'adjacence, classification hiérarchique, indices de comparaison de classifications.*

## Abstract

*The objective of this paper is to propose a topological approach to clustering in evolutionary data analysis. We are interested in clustering that results from exploratory methods for the joint analysis of several data tables ; more specifically, methods that can be applied to temporal data.*

*Clustering is one of the most widely used approaches for exploring multidimensional data. Two common unsupervised clustering strategies are hierarchical ascending clustering (HAC) and k-means partitioning, used to identify groups of similar objects in a dataset in order to divide it into homogeneous groups.*

*The proposed approach, known as topological clustering on evolutionary data (TCED), is based on the notion of neighborhood graphs in an evolutionary data context. It makes it possible to simultaneously explore several tables of data collected at different times on the same individual rows, even in cases where the variables are different in the tables considered. The columns–variables of each table are more-or-less correlated or linked according to the variable type. In each table, TCED is used to analyze the structure of the correlations or associations observed between the variables according to whether they are of a quantitative or qualitative type or a mixture of both.*

*The proposed TCED approach is presented and illustrated here using a real dataset with quantitative variables. The results are compared with those resulting from the clustering on the results of the multiple factorial analysis (MFA) on evolutionary data.*

## Keywords

*Evolutionary data cluster, proximity measure, neighborhood graph, adjacency matrix, hierarchical clustering, clustering comparison indexes.*

## 1 Introduction

L'objectif de cet article est de proposer une approche topologique d'analyse de données appliquée à des tableaux de données croisant les mêmes individus avec éventuellement des variables différentes, quantitatives, qualitatives ou mixtes.

L'approche proposée CTDE est différente de celles qui existent déjà, en particulier la classification sur les résultats de l'Analyse Factorielle Multiple (AFM)[6, 7] avec laquelle elle est comparée , ou encore sur les résultats de la méthode des Tableaux Structurants à Trois Indices de la Statistique (STATIS) [10, 11] ou de la méthode de l'Analyse en Composantes Principales Doubles (DPCA)[5].

Il existe des approches topologiques spécifiquement dédiées au clustering [3, 12] mais à notre connaissance, aucune de ces approches n'a été proposée pour analyser plusieurs tableaux de données simultanément. On peut également citer l'approche de clustering évolutive des données proposée dans [4] mais pas dans un contexte topologique.

Le choix de la mesure de proximité parmi les nombreuses mesures existantes, joue un rôle important dans l'analyse multidimensionnelle des données [16], les résultats de toute opération de structuration, de regroupement ou de classification d'objets dépendent fortement de la mesure de proximité choisie.

Cette étude propose une classification topologique évolutive des individus, généralement dans le temps, quel que soit le type de variables considérées : quantitatives, qualitatives ou un mélange des deux.

La structure de corrélation ou de dépendance des variables quantitatives ou qualitatives de chaque tableau de données évolutives ou temporelles, dépend des données considérées. Les résultats peuvent changer selon la mesure de proximité choisie pour chaque tableau de données. Une mesure de proximité est une fonction qui mesure la similitude ou la dissemblance entre deux objets ou variables au sein d'un ensemble.

## 2 Données évolutives dans un contexte topologique

L'analyse de données topologiques est une approche basée sur le concept de graphe de voisinage. L'idée de base est en fait assez simple : pour une mesure de proximité donnée pour des données continues ou binaires et pour une structure topologique choisie, on peut faire correspondre un graphe topologique induit sur l'ensemble des objets.

L'analyse topologique sur des données évolutives consiste à analyser simultanément plusieurs tableaux de données $(X_t)_{t=1,T}$ collectées à des moments différents sur les mêmes individus, les variables pouvant être identiques ou différentes selon les tableaux.

On considère au temps $t$, $E_t = \{x^1, \cdots, x^j, \cdots, x^{p_t}\}$ un ensemble de $p_t$ variables quantitatives du tableau de données $X_t$ . On peut voir dans [1,2,3] des cas de variables qualitatives ou même mixtes.

Nous pouvons, au moyen d'une mesure de proximité $u_t$, définir une relation de voisinage, $V_{u_t}$, comme étant une relation binaire basée sur $E_t \times E_t$. Il existe de nombreuses possibilités pour construire cette relation binaire de voisinage.

Ainsi, pour une mesure de proximité donnée $u_t$, nous pouvons construire un graphe de voisinage sur $E_t$, où les sommets sont les variables et les arêtes sont définies par une relation de voisinage. De nombreuses définitions sont possibles pour construire cette relation de voisinage binaire. On peut choisir par exemple, l'Arbre de Longueur Minimale (ALM), le Graphe de Gabriel (GG), ou encore le Graphe des Voisins Relatifs (GVR) [13].

Étant donné un ensemble $E_t$ de $p_t$ variables du tableau de données $X_t$ et une mesure de proximité $u_t$, pour des données continues ou binaires, on peut construire la matrice symétrique binaire d'adjacence associée $V_{u_t}$ d'ordre $p_t$, où, toutes les paires de variables voisines dans $E_t$ satisfont la propriété GVR suivante :

$$V_{u_t}(x^k, x^l) = \begin{cases} 1 & \text{if } u_t(x^k, x^l) \leq \max[u_t(x^k, x^t), u_t(x^t, x^l)]; \\ & \forall x^k, x^l, x^t \in E_t,\ x^t \neq x^k \ \ and \ \ x^t \neq x^l \\ 0 & \text{otherwise.} \end{cases}$$

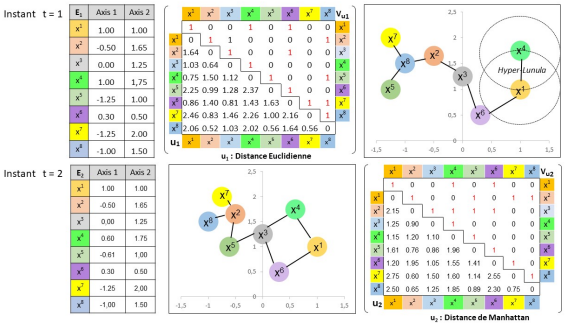



FIGURE 1 – Données - GVR structure - Distances - Matrices d'adjacences associées

La Figure 1 montre un exemple simple dans $\mathbb{R}^2$ de deux ensembles de variables $E_1$ et $E_2$ des mêmes huit variables quantitatives, qui vérifient la structure GVR avec la distance euclidienne $u_1(x^k\ ,\ x^l) = \sqrt{\sum_{j=1}^{2}(x_j^k - x_j^l)^2}$ pour le tableau de données $X_1$ au temps $t = 1$ et distance de Manhattan $u_2(x^k, x^l) = \sum_{j=1}^{p} |x_j^k - x_j^l|$ pour le tableau de données $X_2$ au temps $t = 2$, ainsi que les matrices d'adjacence binaires associées $V_{u_1}$ et $V_{u_2}$.

## 3 Analyse et Classification des données évolutives

### 3.1 Matrices d'adjacence évolutives

L'objectif est dans un premier temps, d'analyser de manière topologique et évolutive les structures de corrélation des variables des tableaux de données considérés, puis d'établir sur cette analyse, une classification des individus.

Au temps $t$, on construit la matrice d'adjacence de référence notée $V_{u_\star t}$, dans le cas de variables quantitatives, à partir de la matrice de corrélation du tableau de données $X_t$. Les expressions des matrices de référence d'adjacence appropriées dans le cas de variables qualitatives ou de variables mixtes sont données dans en bibliographie.

Pour examiner la structure de corrélation entre les variables du tableau de données $X_t$, nous examinons la signification de leur coefficient de corrélation linéaire. Cette matrice d'adjacence peut s'écrire comme suit en utilisant le test t de Student du coefficient de corrélation linéaire de Bravais-Pearson. Pour des variables quantitatives, la matrice de d'adjacence de référence $V_{u_t^\star}$ associée à la mesure de référence $u_t^\star$ est définie comme suit :

$$V_{u_t^\star}(x_t^k, x_t^l) = \begin{cases} 1 & \text{si p-value} = P[|\ T_{n-2}\ | > \text{t-value}\ ] \leq \alpha; \forall k, l = 1, p \\ 0 & \text{sinon.} \end{cases}$$

### 3.2 Notations & Définition

Soient $T$ tableaux des données évolutives $X_t$ avec les mêmes $n$ lignes-individus et $p_t$ différentes colonnes-variables ou les mêmes mesurées à des instants différents $t, t = 1, \cdots, T$. Nous utilisons les notations suivantes :
- $X_{t_{(n,p_t)}}$ est la matrice des données à $n$ individus et $p_t$ variables à l'instant $t$,
- $X_{(n,p)} = [X_1|\cdots|X_t|\cdots|X_T]$ est la matrice globale à $n$ individus et $p = \Sigma_{t=1}^{t=T} p_t$ variables, concatenation en colonnes des $T$ tableaux de données $X_t$.
- $V_{u_t^\star(p_t)}$ est la matrice symétrique d'adjacence d'ordre $p_t$, associée à la mesure de proximité de référence $u_{\star t}$ qui structure au mieux les corrélations ou dépendances des variables du tableau de données $X_t$,
- $V_{u^\star(p)} = Diag[V_{u_t^\star}]_{t=1,T}$ est la matrice diagonale globale des matrices d'adjacence d'ordre $p$, associée à la matrice des données globale $X$,
- $\widehat{X}_{(n,p)} = XV_{u^\star}$ est la matrice de données projetées à $n$ individus et $p$ variables,
- $M_p$ est la matrice de distance d'ordre $p$ dans l'espace des individus,
- $D_n = \frac{1}{n} I_n$ est la matrice diagonale des poids d'ordre $n$ dans l'espace des variables.

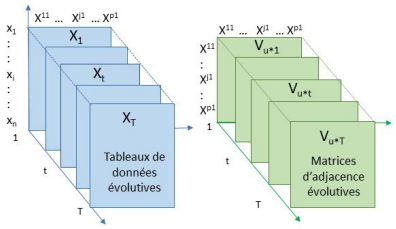



FIGURE 2 – Données évolutives & Matrices d'adjacence

La CTDE consiste à effectuer une CAH basée sur le critère Ward [1], sur les facteurs significatifs de ACP topologique du triplet $(\widehat{X}, M_p, D_n)$.

## 4 Exemple illustratif

Pour illustrer l'approche CTDE, nous utilisons les données d'Eurostat sur l'état des finances publiques des 28 pays de l'Union Européenne (UE) sur la période homogène de quatre ans, de 2016 à 2019.
Nous examinons ici l'évolution des principales caractéristiques des finances publiques de l'UE-28 durant la période $2016 - 2019$, qui sont plus précisément, la dette publique brute, le déficit, les dépenses et les recettes publiques. Des statistiques sommaires des variables considérées sont données dans le Tableau 1.

1. Agrégation basée sur le critère de la perte d'inertie minimale.

TABLE 1 – Statistiques sommaires des finances publiques de l'UE-28 - Période 2016-2019

| | | 2016 | | | |
|---|---|---|---|---|---|
| Variable | Ident | Moyenne | Ecart-type(N) | Min | Max |
| Dépenses | EXPE | 43.54 | 6.88 | 28.10 | 56.70 |
| Déficit | DEFI | -0.98 | 1.59 | -4.30 | 1.90 |
| Recettes | REVE | 42.55 | 6.58 | 27.30 | 53.90 |
| Dette | DEBT | 70,90 | 37.52 | 10.00 | 180.50 |
| | | 2017 | | | |
| Variable | Ident | Moyenne | Ecart-type(N) | Min | Max |
| Dépenses | EXPE | 42.72 | 6.87 | 26.20 | 56.50 |
| Déficit | DEFI | -0.29 | 1.76 | -3.10 | 3.30 |
| Recettes | REVE | 42.43 | 6.67 | 25.90 | 53.50 |
| Dette | DEBT | 68.13 | 37.21 | 9.10 | 179.50 |
| | | 2018 | | | |
| Variable | Ident | Moyenne | Ecart-type(N) | Min | Max |
| Dépenses | EXPE | 43.10 | 6.48 | 25.30 | 55.60 |
| Déficit | DEFI | -0.27 | 1.63 | -3.60 | 3.00 |
| Recettes | REVE | 42.84 | 6.46 | 25.50 | 53.40 |
| Dette | DEBT | 66.29 | 38.51 | 8.20 | 186.40 |
| | | 2019 | | | |
| Variable | Ident | Moyenne | Ecart-type(N) | Min | Max |
| Dépenses | EXPE | 42.94 | 6.45 | 24.30 | 55.40 |
| Déficit | DEFI | -0.11 | 1.81 | -4.30 | 4.10 |
| Recettes | REVE | 42.81 | 6.54 | 24.70 | 53.80 |
| Dette | DEBT | 64.05 | 37.53 | 8.50 | 180.60 |

TABLE 2 – Matrice globale d'adjacence de référence

$$V_{u^*} = \begin{pmatrix} V_{u*2016} & 0 & 0 & 0 \\ 0 & V_{u*2017} & 0 & 0 \\ 0 & 0 & V_{u*2018} & 0 \\ 0 & 0 & 0 & V_{u*2019} \end{pmatrix}$$

$$= \left(\begin{array}{cccc|cccc|cccc|cccc}
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1
\end{array}\right)$$

La figure 3 montre que les deux premiers facteurs de l'ACP topologique expliquent $71.46\%$ et $17.94\%$, respectivement, ce qui représente $89.40\%$ de la variation totale de l'ensemble de données ; cependant, les deux premiers facteurs de l'AFM totalisent $82.35\%$. Ainsi, les deux premiers facteurs fournissent une synthèse adéquate des données, c'est-à-dire des finances publiques de l'UE-28 sur la période 2016-2019.

Les corrélations significatives entre les variables initiales et les principaux facteurs des deux analyses sont assez différentes.

A titre de comparaison, la Figure 4 donne les dendrogrammes des classifications topologique CTDE et AFM des 28 pays de l'UE selon leurs finances publiques.

A noter que les partitions choisies en 4 classes sont sensiblement différentes, tant par leur composition que par leur

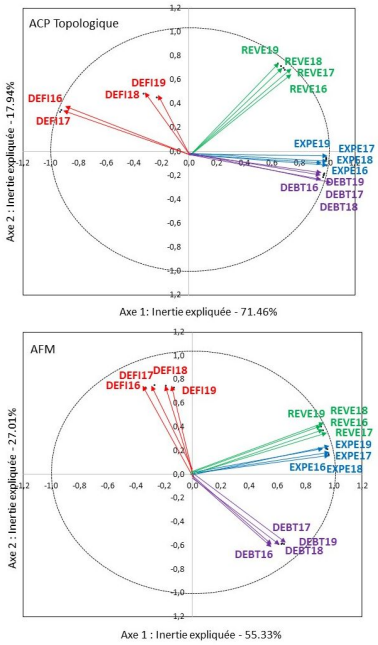



FIGURE 3 – ACP topologique & AFM des finances publiques de l'UE-28

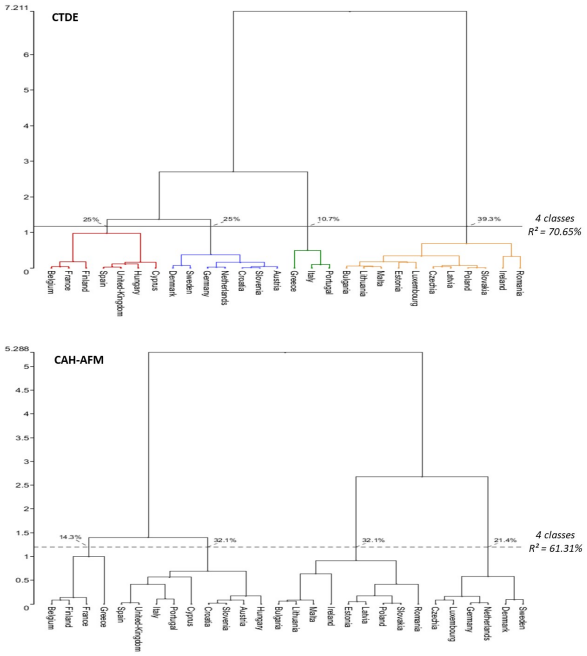



FIGURE 4 – Dendrogrammes des pays de l'UE-28 - Topologique & CAH-AFM

caractérisation. Le pourcentage de la variance totale expliquée par l'approche CTDE, $R^2 = 70.65\%$, est supérieure à celui de l'approche CAH-AFM, $R^2 = 61.31\%$, indiquant ainsi que les classes de la CTDE sont plus homogènes.

Enfin, divers indices et mesures [8,9,14,15] sont utilisés pour comparer les deux classifications.

| **CTDE** | | | | |
|---|---|---|---|---|
| Cluster | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Frequency (%) | 7 (25.00%) | 7 (25.00%) | 3 (10.71%) | 11 (39.29%) |
| Composition | Belgium, Spain, Hungary, Finland, France, Cyprus, United-Kingdom | Denmark,Sweden, Germany, Croatia, Austria Netherlands | Greece, Italy, Portugal Slovenia | Bulgaria, Malta, Czechia, Poland, Estonia, Romania, Lithuania, Luxembourg, Latvia, Ireland Slovakia |
| Profile (+) | EXPE16 to 19 DEBT18 to 19 REVE18 | REVE16 to 19 DEFI18,19 | DEBT16 to 19 EXPE16 to 19 | DEFI16,17 |
| Anti-profile(-) | DEFI16 to 19 | | DEFI16,17 | DEBT16 to 19 REVE16 to 19 EXPE16 to 19 |
| **CAH-AFM** | | | | |
| Cluster | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Frequency (%) | 6 (21.43%) | 9 (32.14%) | 7 (25.00%) | 6 (21.43%) |
| Composition | Italy, Greece, Belgium, France, Finland Austria | Spain,Hungary , United-Kingdom, Cyprus Slovenia, Poland, Slovakia Portugal | Bulgaria, Ireland, Latvia, Lithuania, Malta, Estonia Romania | Netherlands, Germany, Sweden, Denmark, Luxembourg Croatia |
| Profile (+) | EXPE16 to 19 REVE16 to 19 DEBT16 to 19 | | | DEFI16 to 19 REVE19 |
| Anti-profile (-) | | DEFI16 to 18 | DEBT16 to 19 REVE16 to 19 EXPE16 to 19 | DEBT16 to 19 |

TABLE 3 – Caracterisation des classes

Le tableau 3 résume les profils significatifs (+) et anti-profils (-) des deux typologies ; avec un risque d'erreur inférieur ou égal à $5\%$, ils sont bien différents.

## 5 Conclusion & Perspective

Une nouvelle approche de classification topologique d'individus dans un contexte de données évolutives est proposée, elle vient ainsi enrichir les méthodes conventionnelles de classification de données quantitatives, qualitatives ou encore mixtes. Il serait intéressant d'étendre cette approche topologique à d'autres méthodes d'analyse de données, notamment dans le cadre des modèles décisionnels.

## Références

[1] R. Abdesselam, Analyse en composantes principales mixte. *Revue des Nouvelles Technologies de l'Information, Classification : points de vue croisés*. Cépaduès Ed., pp. 31–41, 2008.

[2] R. Abdesselam, A topological multiple correspondence analysis. *Journal of Mathematics and Statistical Science* 5, 8, pp. 175–192, 2019.

[3] R. Abdesselam, A topological clustering of variables. *Journal of Mathematics and System Science* 11, 2, 1–17, 2021.

[4] I. Aljarah, et al. Evolutionary data clustering : algorithms and applications. *Springer*, 2021.

[5] . J.M. Bouroche, Analyse des données ternaires : la double analyse en composantes principales. *Thèse*, 1975.

[6] F. Dazy, J.L. Barzic, G. Saporta, et F. Lavallard. L'analyse des données évolutives – Méthodes et applications. *Editions TECHNIP.*, 1996

[7] B. Escofier, J. Pagès Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes. *Publication interne de l'lIRISA*, 1985.

[8] E. Fowlkes, B. Mallows A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78, 383, pp. 53–569, 1983.

[9] L. Hubert, P. Arabie Comparing partitions. *Journal of Classification*, pp. 193–218, 1985.

[10] C. Lavit, Analyse conjointe de tableaux quantitatifs. *Editions Masson*, 1988.

[11] H. L'Hermier des planttes, Structuration des tableaux à trois indices de la statistique. *Thèse de 3ème cycle*. Université de Montpellier, 1976.

[12] D. Panagopoulos, Topological data analysis and clustering. *Chapter for a book, Algebraic Topology, Machine Learning*, 2022.

[13] W. Rand, Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, American Statistical Association* 846–850, pp. 53–569, 1971.

[14] G.T. Toussaint, The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, pp. 261-268, 1980.

[15] G. Youness, G. Saporta, Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée* 52, 1, pp. 97–120, 2004.

[16] D. Zighed, R. Abdesselam, A. Hadgu, Topological comparisons of proximity measures. 16th PAKDD Conference (Part I, LNAI 7301), pp. 379–391, 2012.

# Clustering sur l'hypersphère unitaire via NMF

Lazhar Labiod, Mohamed Nadif

Centre Borelli UMR 9010, Université Paris Cité, 45 rue des Saints-Pères, 75006 Paris

prénom.nom@u-paris.fr

## Résumé

*Nous proposons un nouveau cadre de factorisation matricielle non négative (NMF) inspiré de la fonction objectif du* spherical $k$-means*. Nous dérivons un nouveau critère dont l'optimisation est réalisé par un algorithme basé sur une seule règle de mise à jour multiplicative. En particulier, nous montrons que le critère optimisé par* spherical $k$-means *est approximativement équivalent au problème algébrique de NMF sous certaines contraintes appropriées. La simplicité et l'efficacité sont deux caractéristiques majeures de notre approche. Des expérimentations numériques sur des jeux de données de type documents $\times$ mots démontrent son intérêt.*

## Mots-clés

*$k$-means sphérique, factorisation de matrice nonnégative.*

## 1 Introduction

Le clustering a fait l'objet d'une attention considérable en tant que problème important avec de nombreuses applications, et un certain nombre de méthodes différentes ont émergé au fil des ans. En général, la parcimonie et la haute dimensionnalité sont les problèmes rencontrés par les différents algorithmes de clustering existants. C'est le cas des matrices documents$\times$termes où chaque cellule représente la fréquence d'un mot dans un document. Par conséquent, le choix d'une mesure de distorsion appropriée peut être crucial pour la performance d'un algorithme de clustering de documents/termes. Ainsi, dans le clustering de documents, considérer les données comme directionnelles est une bonne alternative. En effet, pour des données de grande dimension éparses ou non, la similarité cosinus s'est avérée être une mesure supérieure à la distance euclidienne [1] ; la direction d'un vecteur document est plus importante que son ordre de grandeur. Cela conduit à une représentation vectorielle unitaire, c'est-à-dire que chaque vecteur document est normalisé et de norme 1. Par conséquent, il est montré que l'algorithme *Spherical kmeans* (SPKM) qui est un k-means dont le critère est basé sur une similarité cosinus et où les centres sont normalisés pour être unitaires, est l'un des algorithmes de clustering les plus efficaces [11, 8]. Cependant, malgré les avantages de SPKM, dans [2] les auteurs ont montré que le critère de SPKM est associé au mélange restreint de distributions de von Mises-Fisher où les proportions de composants sont supposées égales et le coefficient de concentration (ou dispersion) de chaque cluster est la même pour tous les clusters [4, 2]. Par conséquent, SPKM présente certains inconvénients lorsque les clusters ne sont pas bien séparés. Ensuite, les auteurs ont proposé d'utiliser des modèles de mélange vMF avec moins de contraintes et pour l'estimation des paramètres et le clustering ils ont effectué des algorithmes de clustering *hard* et *soft* dérivés de l'algorithme EM [3]. Cependant, il convient de noter que l'approximation proposée utilisée pour l'estimation des concentrations de classes souffre de la grande dimensionnalité. Cette difficulté est due au paramètre de concentration qui est non trivial en grande dimension du fait de l'inversion fonctionnelle des rapports des fonctions de Bessel. Dans leur conclusion, les auteurs ont souligné que les investigations sur le compromis entre la complexité du modèle et la complexité de l'échantillon méritent d'être étudiées dans le contexte des données directionnelles. Dans la suite, nous ne considérerons pas l'approche mélange mais, nous proposons d'envisager une autre approche simple et efficace permettant de dépasser les limites de SPKM.

Même si apporter une solution au problème de clustering n'est pas l'objectif principal de la matrice de factorisation non négative (NMF) [7], cette approche a séduit de nombreux auteurs pour le clustering de données et particulièrement pour le clustering de documents. Différents auteurs [9] ont souligné que l'approche NMF surpasse $k$means sur la plupart des ensembles de données puisque NMF semble modéliser différentes distributions en raison de la flexibilité de la factorisation matricielle.

D'autre part, concernant SPKM, notons que la version dure de EM proposée dans [2] est en fait un algorithme de classification EM optimisant la log-vraisemblance classifiante d'un modèle de mélange restreint de distributions v-MFs. Il est naturel de penser qu'un NMF approprié est susceptible de modéliser des distributions variables par rapport au modèle sous-jacent que la fonction objectif de SPKM tente de capturer. Pour cette raison, dans ce travail, nous avons choisi de considérer les objectifs de SPKM dans un cadre NMF.

## 2 Algorithmes *Spherical NMF*

Soit une matrice de données $A = (a_{ij}) \in \mathcal{R}^{M\times N}$ de type documents $\times$termes. Chaque ligne $i$ représente le document $\mathbf{a}_i$ décrit par un ensemble de mots. Le but du clustering de documents est de partitionner l'ensemble des documents en classes homogènes. Cet objectif peut être réalisé de manière

à optimiser la *cohérence* entre $A$ et une matrice révélant une structure en classes. Plus précisément et selon l'approche NMF, le but est de factoriser $A$ par la matrice $M \times K$ non négative $\mathbf{U}$ et la matrice $N \times K$ non négative $\mathbf{V}$ en minimisant la fonction objectif suivante.

$$||A - \mathbf{U}\mathbf{V}^T||^2. \quad (1)$$

Chaque colonne de $\mathbf{V}$ est un vecteur de base, un codage d'un espace sémantique ou d'un concept de $A$ et chaque ligne de $\mathbf{U}$ contient un codage de la combinaison linéaire des vecteurs de base qui se rapproche des lignes correspondantes de $A$. Les dimensions de $\mathbf{U}$ et $\mathbf{V}$ sont $M \times K$ et $N \times K$ respectivement, où $K$ est le rang réduit ou encore le nombre de classes de documents souhaité. Habituellement, $K$ est choisi pour être beaucoup plus petit que $N$, mais plus précisément $K \leq \min(M, N)$. Tous les algorithmes de type NMF sont itératifs et peuvent être différenciés par les règles de mise à jour de deux matrices dues à la méthode d'optimisation choisie ou aux contraintes supplémentaires imposées aux deux matrices. L'approximation de $A$ peut être résolue par une procédure d'optimisation itérative. Dans [5] et [10] les auteurs ont souligné l'importance de la contrainte d'orthogonalité. Ils l'ont introduite sur $\mathbf{U}$ et ont proposé différents algorithmes différenciés par les règles de mise à jour des facteurs.

Dans ce travail, nous proposons un nouveau cadre de clustering basé sur la formulation NMF. L'objet de cette contribution est triple :

- Tout d'abord, contrairement aux approches NMF précédentes liées à $k$means, nous intégrons l'objectif de clustering par SPKM dans le cadre de la NMF dès le début. De cette manière, nous fixons l'objectif de la factorisation en termes de clustering.
- Deuxièmement, étant donné une matrice de données $A$, l'approche proposée optimise une formulation relaxée de SPKM ; un critère dans un style NMF. La procédure d'optimisation recherche alors la meilleure approximation $A \approx \mathbf{R}\mathbf{R}^\top A$ selon un critère à définir par rapport à des contraintes appropriées générées par les propriétés de la matrice $\mathbf{R}$.
- Enfin, nous justifions la souplesse de notre approche et sa simplicité. Nous montrons que son objectif est mathématiquement équivalent aux objectifs de clustering de type *spectral* ou *kernel*. L'intérêt de cette approche est illustré par des expériences numériques sur des jeux de données réels.

Comme notre formulation est issue d'une reformulation de SPKM, nous l'avons appelée *spherical non-negative matrix factorization*. Nous proposons deux algorithmes (NMF1) et (NMF2) où la contrainte d'orthogonalité sur $\mathbf{R}$ est requise pour les deux formulations. Ensuite, dans ce cadre, nous développons un nouvel algorithme de clustering pour des données non négatives, calculant itérativement un seul facteur basé sur une règle de mise à jour multiplicative. La convergence des deux algorithmes utilisés est garantie. Des expériences numériques démontrent l'efficacité et le potentiel des algorithmes proposés.

# 3 Conclusion

Nous proposons une nouvelle approche via NMF pour surmonter les difficultés de sperical $k$means. Deux algorithmes sont proposés et évalués sur des données réelles montre l'intérêt de cette approche en terme de clustering. Une extension à la version *Semantic NMF* [6] est désormais possible.

# Références


[1] Melissa Ailem, François Role, and Mohamed Nadif. Sparse poisson latent block model for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 29(7) :1563–1576, 2017.

[2] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6 :1345–1382, 2005.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[4] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2) :143–175, 2001.

[5] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[6] Mickael Febrissy, Aghiles Salah, Melissa Ailem, and Mohamed Nadif. Improving nmf clustering by leveraging contextual relationships among words. *Neurocomputing*, 495 :105–117, 2022.

[7] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[8] Aghiles Salah and Mohamed Nadif. Directional coclustering. *Advances in Data Analysis and Classification*, 13 :591–620, 2019.

[9] Dingding Wang, Tao Li, and Chris Ding. Weighted feature subset non-negative matrix factorization and its applications to document understanding. In *2010 IEEE International Conference on Data Mining*, pages 541–550. IEEE, 2010.

[10] Jiho Yoo and Seungjin Choi. Orthogonal nonnegative matrix tri-factorization for co-clustering : Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5) :559–570, 2010.

[11] Shi Zhong. Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 3180–3185. IEEE, 2005.

# Consensus de partitions en NLP pour une revue systématique de la littérature autour de l'XAI du biais et de l'équité

M.L. Ndao[1,2], N. Niang[2], G. Youness[1,2], G. Saporta[2]

[1] Laboratoire LINEACT CESI, Nanterre, IDFC
[2] Laboratoire Cedric-MSDMA, Paris, France

mlndao@cesi.fr ; gyouness@cesi.fr ; ndeye.niang_keita@cnam.fr, gilbert.saporta@cnam.fr

## Résumé

*Ce travail présente une analyse comparative d'une bibliographie autour du biais de l'équité et de l'explicabilité des algorithmes de l'IA entre 2015 et 2022. Par trois approches de Traitement Automatique du Langage Naturel (LDA, NMF et $k$-SVD), nous avons extrait différents sujets traités par cette bibliographie. Ces trois approches nous ont également fourni trois partitions. Dans l'optique d'éviter de faire un choix entre ces partitions, nous avons proposé une synthèse de ces trois partitions par une approche de consensus pondérée.*

## Mots-clés

*Intelligence Artificielle eXplicable (XAI), Traitement Automatique du Langage Naturel (TAL), Non-negative matrix factorization (NMF), Consensus de partitions pondéré*

## Abstract

*This work provides a comparative analysis of a bibliography around fairness bias and explainability of AI algorithms between 2015 and 2022. Through three approaches of Natural Language Processing (LDA, NMF et $k$-SVD), we extracted different topics covered by this bibliography. These three approaches also provided us with three partitions. In order to avoid making a choice between these partitions, we proposed a synthesis of these three partitions by a weighted consensus approach.*

## Keywords

*eXplainable Artificial Intelligence (XAI), Natural Language Processing (NLP), Non-negative matrix factorization (NMF), Weighted consensus*

## Introduction

La problématique de l'équité, du biais et de l'équité en apprentissage automatique (Machine Learning ML) est de plus en plus présent. Ceci est lié à de nombreuses failles dans ces algorithmes qui sont souvent source de discrimination dans plusieurs domaines comme en reconnaissance faciale, en justice, en recommandation, en recrutement, en banque, en santé, etc. (Google photo [1], COMPAS [2], logiciel de recrutement chez Amazon [3]). Étant donné que la plupart des algorithmes d'apprentissage automatique (Machine Learning ML) établissent des règles sur la base des données d'apprentissage susceptibles de présenter un biais, il en est de même des prédictions issues de ces algorithmes. Ce contexte a provoqué une vague de recommandations de la part de certains organismes tels que la DARPA (Defense Advanced Research Projects Agency). On assiste à cet effet à l'annonce du concept d'XAI (eXplainable Artificial Intelligent) en 2016 (D. Gunning et al. 2019) [5]. Ce concept met en avant la compréhension par l'humain des décisions prises sur la base des algorithmes de l'IA.
Depuis cette annonce, on note une forte multiplication des recherches et publications sur l'équité, l'explicabilité et le biais des algorithmes de l'IA. C'est ce qu'on observe en analysant les données de Google Trends sur les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » (FIGURE 1).

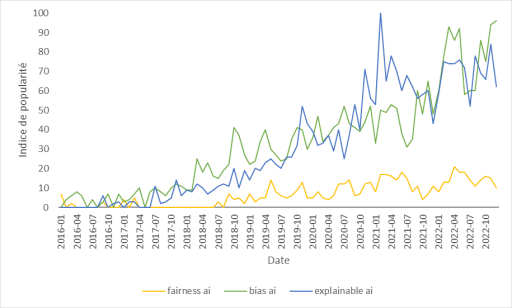



FIGURE 1 – Les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » dans le monde depuis 2016 selon Google Trends.

Aujourd'hui, une des problématiques autour de la littérature du biais, de l'explicabilité et de l'équité est le nombre important de propositions de modèles d'XAI et de métriques d'équité (plus de 400 références citées par Barredo Arrieta et al., 2020 [1]) dont certaines sont contradictoires (Mitchell et al., 2021 [11]). Ainsi, une réorganisation et une recherche de la structure sous-jacente de la bibliographie de

1. https://www.dailymail.co.uk/sciencetech/article
2. ProPublica. 23 mai 2016 ajouter l'article dans ref
3. https://www.assessfirst.com/fr/algorithme-sexiste-amazon/

l'explicabilité, du biais et de l'équité en IA est nécessaire. C'est l'objectif de ce travail.

Nous proposons une analyse de la structure sous-jacente de la bibliographie autour du biais de l'équité et de XAI à l'aide de approche de Traitement Automatique du Langage Naturel (Natural Language Processing ou NLP) non supervisée. Notre approche consistera a utilisé 3 modèles d'analyse : Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]); NMF et $k$-SVD. Ensuite, par une approche de consensus de partitions pondérés WNMF, on regardera le compromis entre ces trois modèles d'analyse.

Le reste du papier est organisé comme suit : la première section est consacrée à une brève présentation d'une part des travaux antérieurs qui ont utilisé une approche NLP pour synthétiser un ensemble d'archives et d'autre part les travaux sur le consensus de partitions. Ensuite, la section 2 est dédiée à la présentation de l'ensemble de notre démarche allant de la collecte des données à la modélisation. La deuxième partie de cette section portera sur l'analyse et la discussion des résultats obtenus.

# 1 Méthodologie

## 1.1 Topic Modeling

Le Topic modeling est une approche d'apprentissage automatique non supervisée qui est souvent utilisée dans différents domaines selon divers contextes afin de synthétiser, d'organiser ou d'analyser des collections de documents ou d'archives. C'est une approche pertinente dans un contexte de données massives ou big data. En effet, elle permet de retrouver une structure sous-jacente d'une collection de documents (partition) en extrayant les sujets liés à chacun des sous-ensembles de cette structure. Il existe de nombreuses approches de topic modeling. Dans le cadre de ce travail, nous nous sommes particulièrement intéressé à trois modèles : le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]); la SVD tronqué (appelé également $k$-SVD) [6] et la Non-negative matrix factorization [9]. Ces trois approches permettent d'avoir : d'une part une relation sujets-mots et d'autre part une relation documents sujets qui conduit à une partition des documents. Dans cette section on se limite à la présentation de l'approche LDA qui est notre modèle de référence.

### 1.1.1 Principe de LDA

Le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]) est une des techniques de NLP non supervisées les plus connues qui cherchent à découvrir des thématiques ou sujets cachés dans un ensemble de $M$ documents appelé corpus noté $D$. C'est un modèle probabiliste génératif permettant de trouver la structure sous-jacente d'un ensemble de documents en termes de sujets. Il considère le corpus comme un mélange de $K$ sujets décrits chacun par un ensemble de mots auxquels sont associés une probabilité.

L'ensemble des $M$ documents ou encore corpus est représenté par une matrice dite document-mots, souvent sparse, notée $D_{M,N}$ de dimension $(M, N)$ où la cellule $(D_i, w_j)$ correspond à la fréquence du mot $w_j$ dans le document $D_i$, par exemple :

$$D_{M,N} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0.3 & & 0 & & 0.2 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \\ \vdots & \cdots & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0.01 \end{bmatrix} \end{array}$$

Le nombre de sujets $K$ est choisi *a priori* ou au regard d'un indicateur comme le score de cohérence que l'on définira dans la section suivante.

Partant de $D_{M,N}$, toutes les trois approches estiment les matrices $\theta_{M,K}$ (documents-sujets) et $\phi_{K,N}$ (sujets-mots).

Dans la matrice $\theta_{M,K}$, $\theta_{m,k}$ correspond à la probabilité que le sujet $z_k$ soit traité dans le document $D_m$ ($\theta_{i.} = \sum_{k=1}^{K} \theta_{ik}$=1).

Le résultat est une classification floue en $K$ clusters où chaque cluster correspond à un sujet. Nous utilisons dans la suite les deux termes sujet ou cluster indifféremment. À partir de $\theta_{M,K}$, on retrouve une partition des documents en $K$ clusters, en affectant chaque document au sujet pour lequel sa probabilité d'appartenance est maximale.

La matrice $\phi_{K,N}$ correspond à la matrice sujets-mots, où $\phi_{kj}$ correspond à la probabilité que le mot $w_j$ soit dans le sujet $z_k$. Chaque sujet $z_k$ est décrit par les $n$ mots ayant les plus fortes probabilités $\phi_{kj}$, nous les notons $(w_j^k)_{1 \leq j \leq n}$.

La matrice $\phi_{K,N}$ est initialisée par une distribution de Dirichlet $Dir(\beta)$. Des exemples de matrices $\theta_{M,K}$ et $\phi_{K,N}$ sont données ci-après :

$$\theta_{M,K} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{ccc} z_1 & \cdots & z_K \end{array} \\ \begin{bmatrix} 0 & & 0.2 \\ \vdots & & \vdots \\ 0.1 & & 0.5 \\ & \vdots & \\ 0.6 & \cdots & 0.0 \end{bmatrix} \end{array}$$

$$\phi_{K,N} = \begin{array}{c} \\ z_1 \\ \vdots \\ z_K \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0 & & 0 & & 0.3 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \end{bmatrix} \end{array}$$

Pour évaluer la qualité des sujets obtenues plusieurs mesures sont classiquement utilisées. Il s'agit des indices : Umass (Université du Massachusetts) [2], CV (Coherence value) [10], UCI (Ultra-Compactness Index) [10] et NPMI (Normalized Pointwise Mutual Information) [3]. Parmi ces métriques, nous allons utiliser le score de cohérence CV pour choisir le nombre de sujets $K$ optimal.

En plus de l'approche LDA, nous utiliserons deux autres méthodes : NMF et k-SVD obtenant ainsi 3 partitions éventuellement de qualité différentes. Dans ce travail, nous proposons de synthétiser ces trois à travers une approche de type ensemble afin d'avoir une seule partition des documents.

## 1.2 Consensus de partitions

Le problème de la combinaison de plusieurs partitions d'un ensemble d'objets (ou d'individus) en une seule partition,

connu également sous le nom de consensus de partitions ou agrégation de partitions, consiste à identifier une partition compromis d'un ensemble de partitions obtenues sur le même ensemble d'observations [4, 12].

Le principe des méthodes de consensus est de trouver la partition compromis des partitions séparées appelées partitions contributives. Cette dernière partition doit être la plus similaire aux partitions contributives. Plusieurs méthodes ont été proposées. Elles peuvent être regroupées en trois grandes familles : celle basée la sur maximisation d'un indice (par exemple l'indice de Rand) ; celle par vote majoritaire, et celle basée sur les matrices association des partitions. C'est à cette dernière qu'on s'intéresse ici.

Dans cette approche, on considère un ensemble de $T$ partitions $P = \{P_1, P_1, P_2, ..., P_T\}$ différentes d'un même ensemble de $M$ observations. Ces partitions sont les résultats des multiples partitionnements pouvant provenir de plusieurs applications (initialisations) d'un même algorithme de classification, différents algorithmes sur le même jeu de données (notre cas) ou d'un même algorithme sur différents ensembles de variables décrivant les $M$ individus.

Pour chaque partition $P_t$, on définit un tableau disjonctif contenant les indicatrices $H(P_t)$ des classes de la partition et une matrice d'association ou d'adjacence $M(P_t)$. La matrice d'association est une matrice $(M \times M)$ qui contient 1 si les deux individus $i$ et $j$ se trouvent dans la même classe, 0 sinon.

La matrice de connectivité $M(P_t)$ est obtenue par $M(P_t) = H(P_t)H(P_t)^{'}$ où $H(P_t)^{'}$ désigne la transposée de $H(P_t)$. On remarque qu'il s'agit d'une matrice symétrique positive contenant que des 1 en diagonale. Le nombre $K$ de classes peut être différent d'une partition à une autre.

On définit la matrice d'association $\tilde{M}$ qui est une simple moyenne des matrices d'association par :

$$\tilde{M_{ij}} = \sum_{t=1}^{T} w_t M_{ij}(P_t) \tag{1}$$

Elle représente l'association moyenne entre deux observations $(i, j)$. Dans une approche consensus simple (NMF), les poids sont donnés par $w_t = \frac{1}{T}$ pour toutes les partitions contributives. Par contre, l'approche pondérée repose sur la détermination et la recherche des poids $w_t$, en fonction de la qualité et la particularité de chaque partition. La matrice d'association s'exprime donc comme une moyenne pondérée des matrices d'association en fonction de ces poids. C'est le cas de "Weighted Nonnegative Matrix Factorization (WNMF) proposé par Ding et al. [4].

Cette approche est pertinente dans la mesure où certaines partitions peuvent paraître particulières par rapport aux autres. Dans ce cas, accorder le même poids à toutes les partitions peut biaiser la partition compromis obtenue.

Dans le cadre de ce travail, nous allons nous intéresser à l'approche WNMF. Cette approche permet la recherche simultanée de la partition compromis et les poids associés aux partitions contributives.

# 2 Application

Dans cette section, nous commencerons par expliquer notre processus de modélisation depuis la collecte des données. Ensuite, nous présenterons les résultats obtenus à l'issue de cette analyse.

## 2.1 Processus d'analyse

### 2.1.1 Les données

Cette étude est basée sur les articles publiés sur les quatre plateformes de bases données suivantes : arXiv, Springer, ScienceDirect et IEEE-Explorer. Sur chaque base de données, nous avons considéré les articles publiés entre 2015 et 2022 avec une recherche séparée sur les méta-données des termes suivants : bias AND (machine learning OR data) ; XAI AND (machine learning OR data) et ; fairness AND (machine learning OR data). Au total, 31 860 articles ont été obtenus. Ensuite, les tâches suivantes ont été réalisées :

- Suppression des duplications : articles ayant les mêmes auteurs, le même titre et le même résumé ;
- Suppression des publications sans résumé ;
- Suppression des articles en d'autres langues que l'anglais.

Par la suite, trois variables binaires ont été créées permettant de vérifier que la publication traite au moins un des trois thèmes : XAI, biais et équité (1 si oui, 0 sinon). Pour chaque thème, les termes suivants ont été considérés :

- Pour XAI : XAI, explainable, explainability, interpretable et interpretability ;
- Pour Biais : bias, harm et disparate ;
- Pour Fairness : fair.

Cette recherche a été faite sur le résumé, le titre et les mots clés de chaque article. Par la suite, seuls les articles ayant traité au moins, un des thèmes a été retenu. Au final, 10 237 publications ont été considérées pour l'étude.

### 2.1.2 Pré-traitement

Un pré-traitement a été fait sur les données. Il s'agit de :

- la suppression des 'stopword' qui consiste à supprimer tous les articles, pronoms et autres mots qui n'ont pas de sens pour notre analyse ;
- la tokenization qui consiste à découper chaque document en une liste de mots appelés tokens. Cette étape conduit à l'obtention d'une matrice documents-termes (matrice d'occurrence).
- la lemmatisation qui consiste à regrouper tous les mots en leur forme de base. (par exemple transformer tous les verbes conjugués en forme infinitif) ;
- la normalisation qui consiste à pondérer chaque terme de la matrice d'occurrence. Dans notre cas, nous avons utilisé l'approche tf-idf (Joachims, T. et al., 1996[8]) qui permet d'évaluer l'importance d'un terme dans un document relativement à tous les autres documents.

Ce processus conduit à l'obtention d'une matrice sparse où chaque ligne correspond à un document et chaque colonne correspond à un mot. Suite à ce processus, une analyse du corpus a été faite pour choisir les paramètres optimaux.

### 2.1.3 Choix des paramètres et du corpus

Dans cette analyse, nous nous sommes basé sur le résumé de chaque article. Par ailleurs, puisqu'il s'agit d'une analyse non supervisée, et que nous n'avons pas l'information sur le nombre de sujets *a priori*, nous avons choisi un $K$ optimal pour chaque modèle d'analyse au regard d'une mesure de cohérence CV [10] des sujets extraits.

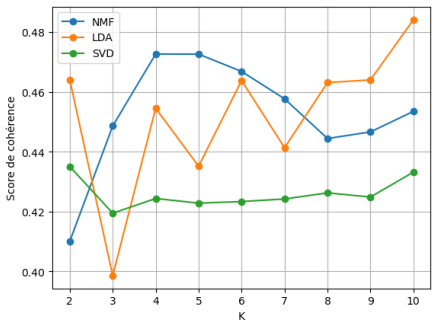



FIGURE 2 – Variation du score de cohérence en fonction du nombre de sujets $K$ pour chaque modèle d'analyse

L'analyse de ce score en fonction du nombre de sujets $K$ (FIGURE 2) pour chaque modèle nous permet de retenir les valeurs de $K$ suivantes (4, 10, 2) respectivement pour NMF, LDA et $k$-SVD.

## 2.2 Résultats

Dans cette section, nous ferons, dans un premier temps, une analyse des sujets obtenus. Ensuite, nous évaluerons la qualité des partitions obtenues en nous basant sur la matrice $\theta$. Une comparaison entre ces partitions sera faite en terme d'indices de Rand.

Une dernière analyse portera sur l'évaluation quantitative de la partition consensus obtenues à partir de l'approche WNMF.

### 2.2.1 Analyse des sujets

Une première analyse des sujets obtenus basée sur les TABLES 3, 1 et 5 permet de voir qu'on arrive à extraire des sujets cohérents traitant les thèmes biais, équité et explicabilité des algorithmes de ML. Par exemple, on peut voir que le sujet 3 extrait par la LDA concerne l'équité algorithmique dans un contexte de prise de décision où le besoin d'explicabilité et de compréhension de cette décision par l'humain se pose. Un autre constat est le fait que certains sujets sont extraits à la fois par les trois modèles d'analyse.

La TABLE 2 montre les résultats sur l'analyse quantitative de la qualité des sujets extraits au moyen des mesures de cohérences classiques. On note que la LDA a souvent une meilleure qualité. Par exemple pour l'indice UMASS, plus celui est faible, meilleure est la qualité des sujets obtenus en terme cohérence. Sur cet indice, on note que la LDA a la plus faible valeur. De même, lorsqu'on regarde les valeurs de CV, on note que l'approche LDA a une plus grande valeur (0.48). Ce qui signifie que ses résultats ont une meilleure qualité en termes de cohérence au regard de cet indice.

| Sujet1 | Sujet2 |
|---|---|
| bias | bias |
| feature | attentional |
| fairness | cognitive |
| propose | negative |
| performance | participant |
| prediction | exchange |
| image | stimulus |
| classification | magnetic |
| analysis | find |
| base | positive |

TABLE 1 – Description des sujets par les 10 mots les plus significatifs. Il s'agit des résultats de l'approche $k$-SVD.

| | | Indices de cohérence | | | |
|---|---|---|---|---|---|
| | | UMASS | CV | UCI | NPMI |
| Modèles | NMF | -10.70 | 0.47 | -7.16 | -0.26 |
| | LDA | -11.14 | 0.48 | -7.51 | -0.27 |
| | $k$-SVD | -8.21 | 0.44 | -6.09 | -0.22 |

TABLE 2 – Qualité des sujets extraits selon les indices de cohérence classiques.

| Sujet1 | Sujet2 | Sujet3 | Sujet4 |
|---|---|---|---|
| feature | bias | explanation | fairness |
| image | attentional | user | fair |
| propose | cognitive | explainable | algorithmic |
| classification | negative | decision | group |
| performance | find | explainability | metric |
| accuracy | patient | human | bias |
| problem | participant | prediction | problem |
| neural | attention | research | discrimination |
| prediction | risk | explain | framework |
| base | positive | trust | privacy |

TABLE 3 – Description des sujets par les 10 mots les plus significatifs pour l'approche NMF

### 2.2.2 Comparaison des 3 partitions

Dans cette analyse, on s'est intéressé aux partitions de documents fournies par les trois approches d'analyse. L'objectif est de voir si on arrive à retrouver des structures similaires de partitions des trois approches. Pour analyser cette similarité, nous avons utilisé l'indice de Rand ajusté (ARI) [7] qui permet de quantifier la similarité entre deux partitions d'une même population.

L'analyse de la matrice des ARI fournit par la FIGURE 3 permet de constater une similarité entre les partitions de NMF et de la LDA. Cette similarité est plus faible pour la partition de la $k$-SVD.

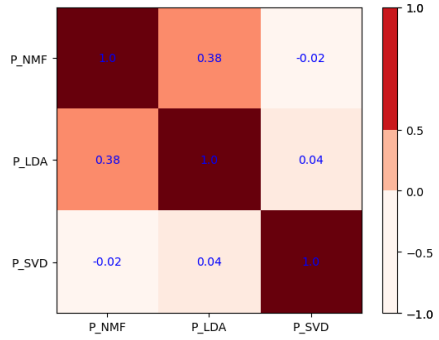



FIGURE 3 – Similarité des partitions au sens de l'indice de rand ajusté.

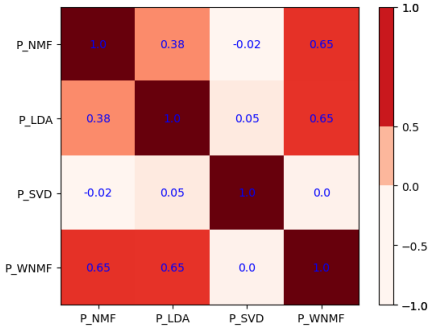



FIGURE 4 – Similarité entre les partitions contributives et la partition consensus au sens de l'indice de rand ajusté.

### 2.3 Résultats du consensus de partitions

Dans cette section, nous allons décrire les résultats fournis par le modèle de consensus de partitions en les comparant avec les partitions contributives obtenues grâce aux trois modèles de l'analyse initiale. Puisq'on considère l'approche LDA comme modèle de base, le nombre de clusters $K = 8$ sera considéré. Tout d'abord on constate que dans le consensus pondéré, le poids accordé à la partition de $k$-SVD est très faible en raison de sa forte différence avec les deux autres comme déjà soulignée TABLE 4. Ainsi, cette partition aura une plus faible contribution dans le processus de consensus.

| Partitions | NMF | LDA | $k$-SVD |
|---|---|---|---|
| Poids | 0.50 | 0.49 | 0.01 |

TABLE 4 – Poids accordé à chaque partition dans le processus de consensus WNMF. Ces poids sont fournis par l'algorithme de WNMF.

Une analyse de la similarité entre la partition consensus et les partitions contributives permet de noter une plus forte ressemblance entre la partition consensus et les partitions fournies par NMF et LDA. Ceci peut être expliqué par la différence de poids accordés aux différentes partitions. En effet, on a noté que ce poids est très faible pour la partition obtenue à partir de la $k$-SVD.

## Conclusion et perspectives

Le travail proposé illustre l'intérêt des approches de traitement de langage naturel pour synthétiser, résumer, et même organiser une bibliographie dans un contexte des données massives (big data) où un besoin d'analyse systématique se pose de plus en plus. En effet, cela peut être utile pour organiser une bibliographie en permettant d'aborder de manière directe les principaux sujets d'intérêt. En pratique, on est souvent emmené à faire un choix entre les modèles existants. Dans cette situation, nous proposons de faire recours à une approche consensus des résultats des modèles permettant ainsi d'éviter ce dilemme.

Cependant, cette approche permet uniquement de partitionner les documents et non les mots décrivant chaque classe comme le fait la LDA par exemple. Ainsi, dans nos futurs travaux, il serait intéressant de proposer une approche de consensus qui fournit également une relation entre chaque classe de documents et l'ensemble des mots. Ceci pourrait permettre d'interpréter facilement chaque classe de documents.

## Références

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58 :82–115, 2020.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[3] Gosse Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, 2009.

[4] Chris H. Q. Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8) :3913–3927, 2008.

[5] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019.

[6] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2) :217–288, 2011.

[7] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.

[8] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[9] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, 1999.

[10] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[11] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness : Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8 :141–163, 2021.

[12] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3 :583–617, 2002.

# A Description des 10 sujets extraits par l'approche LDA

| **sujet1** | **sujet2** | **sujet3** | **sujet4** |
|---|---|---|---|
| bias | cell | explanation | fault |
| magnetic | receptor | fairness | history |
| exchange | signal | explainability | contain |
| substrate | protein | decision | threat |
| field | agonist | explainable | particle |
| device | drug | user | stress |
| measurement | activation | research | dependence |
| film | bind | human | science |
| phase | distinct | algorithmic | resistance |
| property | efficacy | technology | coverage |

| sujet5 | sujet6 | sujet7 |
|---|---|---|
| patient | attack | bias |
| healthcare | security | cognitive |
| clinical | threat | participant |
| medical | adversarial | attentional |
| disease | judgment | gender |
| diagnosis | lime | risk |
| health | sensor | group |
| cancer | dnn | patient |
| care | behaviour | individual |
| treatment | resistance | find |

| sujet8 | sujet9 | sujet10 |
|---|---|---|
| feature | recommendation | privacy |
| propose | item | fairness |
| classification | sentiment | traffic |
| image | recommender | federate |
| performance | user | resource |
| prediction | news | protocol |
| bias | social_medium | user |
| accuracy | medium | scheduling |
| problem | political | communication |
| neural | rating | throughput |

TABLE 5 – Description des sujets par les 10 mots les plus significatifs pour l'approche LDA.

# Considération de l'Incertitude d'Imputation pour l'Apprentissage des Réseaux de Neurones

Thomas Ranvier, Haytham Elghazel, Emmanuel Coquery, Khalid Benabdeslem

Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205
43 bd du 11 Novembre 1918, 69622 Villeurbanne, France

{thomas.ranvier,haytham.elghazel,emmanuel.coquery,khalid.benabdeslem}@univ-lyon1.fr

## Résumé

*Dans cet article, nous nous intéressons à l'entraînement de réseaux de neurones dans un contexte de données incomplètes. Nous cherchons à améliorer l'entraînement des réseaux de neurones en réduisant les biais potentiels pouvant survenir durant la phase d'apprentissage sur des jeux de données artificiellement complétés. Nous proposons deux frameworks d'imputation, S-HOT et M-HOT, pouvant être utilisés pour entraîner des réseaux neuronaux sur des données complétées de manière moins biaisée. Nous réalisons des expérimentations comparatives approfondies et évaluons statistiquement les résultats. Nous montrons que les frameworks proposés sont compétitifs et même plus performants que d'autres frameworks d'imputation existants.*

## Mots-clés

*Imputation, Incertitude d'Imputation, Variance Inter, Optimisation, Réseaux de Neurones.*

## Abstract

*In this paper we are interested in dealing with missing values when training a neural network.We focus on improving neural network training by reducing the potential biases that can occur during the training phase on artificially imputed datasets. We propose two new imputation frameworks, S-HOT and M-HOT, that can be used to train neural networks on completed data in a less biased way.We perform extensive comparative experiments and statistically assess the results.We show that our frameworks compete against and even outperform existing imputation frameworks.*

## Keywords

*Data Imputation, Imputation Uncertainty, Between-variance, Optimization, Neural Networks.*

## 1 Introduction

Deux types de variances apparaissent lorsque l'on impute plusieurs fois un jeu de données incomplet : la variance intra et la variance inter. La variance intra correspond à la variance au sein de chaque jeu de données complété. La variance inter correspond à la variance entre chaque jeu de données complété. Dans la réalité, il n'est généralement pas possible de connaître la variance intra, puisque la plupart des méthodes d'imputation estiment une valeur fixe à la place des valeurs manquantes sans fournir de mesure de probabilité. Cependant, la variance inter peut facilement être calculée entre deux jeux de données complétés. Dans la suite, nous appelons "incertitude d'imputation" cette variance inter.

Dans cet article, nous différencions explicitement les méthodes d'imputation, visant à imputer les valeurs manquantes dans un jeu de données, et les frameworks d'imputation tels que l'imputation simple (*SI*) ou l'imputation multiple (*MI*), qui reposent sur n'importe quelle méthode d'imputation pour traiter les données manquantes selon une méthodologie définie. Nos contributions sont des frameworks d'imputation, visant à entraîner des réseaux de neurones en tenant compte de l'incertitude d'imputation pouvant s'appuyer sur n'importe quelle méthode d'imputation. Il est à noter que nous ne sommes pas intéressés par la comparaison des performances des-dites méthodes d'imputation.

Il a été démontré que, lors de l'utilisation de modèles d'inférence forts tels que des réseaux de neurones, quasiment n'importe quelles imputations conduisent asymptotiquement à une prédiction optimale [5]. C'est probablement l'une des principales raisons pour lesquelles la prise en compte de l'incertitude d'imputation n'a jamais fait l'objet de recherches approfondies. Cependant, même si des modèles forts obtiennent de bons résultats dans de telles situations, ils n'en restent pas moins biaisés par l'incertitude d'imputation. Nous montrons que la prise en compte de cette incertitude pendant la phase d'apprentissage permet d'obtenir de meilleurs résultats de prédiction. Dans cet article, nous proposons deux frameworks d'imputation, *S-HOT* et *M-HOT*. Ils visent à entraîner des réseaux de neurones sur des jeux de données imputées en tenant compte de l'incertitude d'imputation de manière à réduire le biais naturel apparaissant lors de l'entraînement sur des jeux de données complétés. Ces frameworks sont destinés à être utilisés dans des situations différentes : *S-HOT* est adapté à l'entraînement d'un large réseau de neurones unique, *M-HOT* entraîne plusieurs modèles de manière ensembliste et permet d'obtenir d'excellents résultats de prédiction au prix d'un coût de calcul plus élevé. Nous menons des expé-

rimentations approfondies pour comparer nos deux frameworks avec les frameworks existants : l'imputation unique et l'imputation multiple. Nous effectuons une analyse statistique afin d'évaluer les résultats obtenus sur différents jeux de données. Nous montrons que les frameworks proposés rivalisent avec les frameworks d'imputation existants, voire les surpassent. Cet article est un premier pas vers la recherche de meilleurs moyens de gérer les valeurs manquantes dans le domaine de l'apprentissage automatique. Nous espérons qu'il suscitera l'intérêt d'autres chercheurs en apprentissage automatique sur cette question importante et pourtant largement négligée dans la littérature.

Dans la suite de cet article, nous présentons d'abord les travaux connexes sur les frameworks et méthodes d'imputation. Nous présentons et décrivons ensuite nos propositions dans la section 3. La section 4 présente nos expérimentations et les résultats obtenus. Enfin, nous concluons par un résumé de nos contributions.

## 2 Travaux Connexes

L'imputation simple (*SI*) est probablement le framework le plus couramment utilisé pour gérer les valeurs manquantes en pratique [10]. La méthodologie est simple : une méthode d'imputation est choisie et appliquée au jeu de données incomplet, permettant d'obtenir un jeu de données artificiellement complété où les valeurs manquantes ont été remplacées par de nouvelles valeurs, qui peut être utilisé et exploité comme tout autre jeu de données complet. L'obtention d'un jeu de données complet est un avantage considérable, car il est alors possible d'intégrer *SI* dans n'importe quel pipeline ou logiciel existant afin de les rendre utilisables en présence de valeurs manquantes [4]. Cependant, il est problématique de traiter les valeurs imputées comme de vraies valeurs [10], la variabilité due aux valeurs manquantes inconnues ne peut pas être prise en compte, les inférences en se basant sur ces données imputées surestimeront la précision [8].

L'imputation multiple (*MI*) a été originellement proposée par Rubin [10]. Ce framework consiste à remplacer chaque valeur manquante par au moins deux valeurs de substitution représentant une distribution de possibilités, ce qui représente l'incertitude quant à la valeur à imputer [13]. Dans un contexte d'apprentissage automatique, un modèle est formé sur chaque jeu de données complété et les résultats de tous les modèles sont ensuite regroupés de manière ensembliste. Un avantage de *MI* par rapport à *SI* est que chaque valeur manquante est représentée par un échantillon de valeurs d'imputation possibles, ce qui donne lieu à des inférences qui reflètent mieux le niveau d'incertitude associé à chaque valeur manquante [13]. Par conséquent, les résultats obtenus par l'ensemble des modèles seront moins biaisés que ceux de chaque modèle pris indépendamment. Un inconvénient évident est le coût de calcul d'une telle méthodologie, l'entraînement de multiples modèles multiplie le temps de calcul nécessaire. Bien qu'il s'agisse d'un framework assez ancien, *MI* est rarement utilisé en pratique, les scientifiques et utilisateurs de méthodes d'imputation se reposent encore généralement sur l'imputation simple. Cela peut s'expliquer en partie par le coût de calcul de *MI*, qui est élevé en raison du paradigme ensembliste.

Il existe de nombreuses méthodes permettant de traiter les valeurs manquantes en les remplaçant par des valeurs plausibles. Une méthode très simple qui peut être utilisée pour traiter les valeurs manquantes est la substitution par la moyenne, où les valeurs manquantes de chaque caractéristique sont remplacées par la valeur moyenne pour cette caractéristique. Cette méthode a l'avantage d'être facile à mettre en œuvre et à utiliser, tout en conservant toutes les informations non manquantes. Des méthodes d'imputation plus avancées peuvent être utilisées pour obtenir de meilleurs résultats d'inférence sur les données imputées. Une méthode populaire est l'algorithme SOFTIMPUTE, introduit par Mazumder et al. [6], qui fonctionne de manière itérative, à chaque étape les valeurs manquantes sont remplacées en utilisant une décomposition en valeur singulières. En 2012, Stekhoven et Bühlmann ont présenté l'algorithme MISSFOREST, une méthode d'imputation itérative basée sur les forêts aléatoires [11]. Ils ont montré que MISSFOREST peut traiter avec succès les valeurs manquantes, en particulier dans les jeux de données comprenant des types de variables mixtes. En 2018, Gondara et Wang ont présenté MIDA : Multiple-Imputation Using Denoising Autoencoders (DAEs) [3]. Cette méthode est basée sur les modèles d'autoencodeurs, un modèle de réseau neuronal utilisé pour reconstruire sa propre entrée à partir d'une représentation latente réduite. Plus récemment, Yoon et al. ont introduit GAIN : Generative Adversarial Imputation Nets [12], qui s'appuie sur des modèles adversaires pour imputer les valeurs manquantes. En 2020, Muzellec et al. ont présenté SINKHORN OT, une méthode basée sur le transport optimal pour l'imputation des données [7]. Toutes ces méthodes d'imputation peuvent être utilisées différemment en fonction du framework d'imputation que l'on souhaite appliquer.

L'algorithme MICE, pour Multivariate Imputation by Chained Equations [1], traite l'incertitude d'imputation en imputant de manière itérative l'ensemble de données au sein de son propre algorithme. Les valeurs manquantes sont d'abord imputées par une méthode de substitution par la moyenne, puis MICE entraîne itérativement plusieurs modèles de régression linéaire pour imputer chaque valeur manquante de chaque caractéristique en utilisant toutes les autres caractéristiques pour entraîner les régressions jusqu'à convergence. En ce sens, MICE produit un jeu de données complété unique mais qui prend en compte l'incertitude d'imputation.

## 3 Contributions

*MI* est un premier pas vers la prise en compte de l'incertitude d'imputation. Il prend naturellement en compte l'incertitude des valeurs imputées grâce à sa nature ensembliste, mais chaque modèle d'inférence prit à part est tout de même biaisé du fait qu'il a été entraîné sur un jeu de données complété arbitrairement [10].

Nous proposons deux frameworks qui prennent en compte cette incertitude d'imputation et montrons que les réseaux de neurones entraînés à l'aide de ces frameworks ont une meilleure capacité de généralisation. Ces frameworks sont basés sur le calcul de l'incertitude entre les imputations, qui correspond à l'écart type entre les valeurs imputées de tous les jeux de données complétés. Cette incertitude est utilisée comme échelle pour ajouter de la stochasticité à l'imputation des valeurs manquantes directement lors de l'extraction des batchs pendant l'apprentissage. Il en résulte une sorte de régularisation par le bruit qui prend en compte l'incertitude d'imputation, ce qui améliore la capacité de généralisation et, par conséquent, les résultats d'inférence sur des données de test.

## 3.1 Single-Hotpatching

Notre framework Single-Hotpatching (*S-HOT*) est similaire à *MI*, mais présente l'avantage de n'entraîner qu'un seul réseau de neurones. les valeurs manquantes sont imputées dynamiquement lors de l'extraction de batches lors de la phase d'entraînement du modèle.

L'entraînement d'un large réseau de neurones nécessite beaucoup de temps et de ressources. L'entraînement de plusieurs modèles de ce type de manière ensembliste n'est pas toujours une option viable. *S-HOT* vise à former un seul modèle tout en s'appuyant sur des imputations multiples telles que *MI*. Il forme un modèle moins biaisé que s'il avait été entraîné à l'aide de *SI* sans nécessiter autant de temps de calcul que *MI*. Nous montrons expérimentalement que *S-HOT* obtient des résultats nettement meilleurs que *SI* pour des temps d'exécutions identiques. Il est donc intéressant d'utiliser *S-HOT* dans toutes les situations où l'on cherche à former un modèle large et unique sur des données complétées.

Lorsque l'on entraîne un réseau de neurones sur un jeu de données complété dans lequel les valeurs imputées sont très certainement non optimales, le modèle apprend de manière répétée sur des données erronées et imprécises, conduisant à un modèle biaisé manquant de capacité de généralisation. Au lieu de cela, *S-HOT* effectue des imputations multiples et calcule l'écart type entre les imputations de chaque valeur manquante, que nous appelons le niveau d'incertitude. Ce niveau d'incertitude est utilisé pour tirer des valeurs aléatoires d'une distribution normale paramétrée à l'aide de la moyenne et de l'écart type calculés entre les imputations. En entraînant le modèle sur des batches dans lesquelles les valeurs manquantes sont tirées dynamiquement sur les distributions d'imputation, nous nous assurons que le modèle apprend sur une multitude d'imputations plausibles. Le modèle est entraîné sur un éventail de valeurs possibles à la place de chaque valeur manquante, conduisant à un modèle moins biaisé avec une plus grande capacité de généralisation. La figure 1 illustre les phases d'entraînement et de test d'un réseau de neurones lors de l'utilisation de *S-HOT*. Une fois le modèle entraîné, les résultats de prédiction sont obtenus en tirant aléatoirement les valeurs manquantes dans le jeu de test de la même manière pour $p$ itérations, aboutissant à $p$ prédictions. La moyenne des $p$ probabilités de sortie du modèle est utilisée comme prédiction finale, ce qui est plus robuste qu'une itération unique de ce processus.

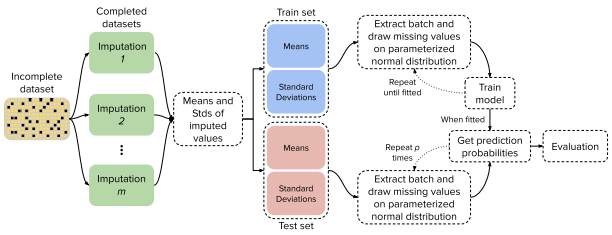



FIGURE 1 – Single-Hotpatching. Nous effectuons $m$ imputations et calculons les moyennes et les écarts types des valeurs imputées. Celles-ci sont séparées entre les jeux d'entraînement et de test. Pendant l'entraînement, chaque fois qu'une batch est extraite, les valeurs manquantes sont tirées d'une distribution normale paramétrée à l'aide des moyennes et des écarts-types calculés précédemment. Une fois le modèle entraîné, nous obtenons les probabilités de prédiction en appliquant le même processus $p$ fois pour chaque instance de test, menant à $p$ prédictions. La prédiction finale est calculée comme la moyenne des $p$ probabilités de prédiction. Le paramètre $p$ est fixé à quelques dizaines pour obtenir des résultats de prédiction robustes.

Mathématiquement, on note $X \in \mathbb{R}^{n \times d}$ le jeu de données originel, où chaque valeur $X_{ij}$ est soit observée ou manquante, avec $n$ et $d$ le nombre d'instances et de variables dans $X$. On effectue $m$ imputations, menant à $m$ différents jeux de données complétés $\tilde{X}^{1\dots m}$, avec $\tilde{X}^k \in \mathbb{R}^{n \times d}$ le $k$-ième jeu complété. Donc, une valeur manquante $X_{ij}$ est imputée avec $m$ valeurs différentes $\tilde{X}_{ij}^{1\dots m}$. Ensuite, on calcule les moyennes $\mu$ et écarts types $\sigma$ de chaque valeur des $m$ jeux complétés, avec $\mu_{ij} = \frac{1}{m}\sum_{k=1}^{m} \tilde{X}_{ij}^k$ (1) et $\sigma_{ij} = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(\tilde{X}_{ij}^k - \mu_{ij})^2}$ (2). On note que les valeurs de $\tilde{X}^{1\dots m}$ qui sont non manquantes dans $X$ ont une moyenne de $\mu_{ij} = X_{ij}$ et un écart type de $\sigma_{ij} = 0$, seul les valeurs manquantes dans $X$ ont une valeur $\sigma_{ij} > 0$. Le réseau de neurones est ensuite entraîné avec des batches calculées à partir de $\mu$ et $\sigma$. Pour extraire une batch $B \in \mathbb{R}^{b \times d}$, avec $b$ le nombre d'éléments dans la batch, chaque valeur de la batch est tirée à partir d'une distribution normale paramétrisée avec la moyenne $\mu_{ij}$ et l'écart type $\alpha \cdot \sigma_{ij}$, tel que $B_{ij} \sim \mathcal{N}(\mu_{ij}, \alpha \cdot \sigma_{ij})$. Où $\alpha$ est un hyper-paramètre d'échelle pouvant être défini à 1 dans la plupart des cas et peut être fixé à une valeur plus basse selon le taux d'incertitude moyen observé. Si la méthode d'imputation utilisée produit des valeurs imputées avec une grande incertitude, l'échelle $\alpha$ doit être empiriquement fixée à une valeur inférieure à 1 afin de limiter l'impact stochastique induit par l'approche. Nous n'avons jamais trouvé de situation où l'augmentation de la valeur de $\alpha$ était bénéfique. Lorsqu'une instance est présenté au réseau de neurones, les valeurs observées sont définies par $X_{ij}$ et les valeurs manquantes par une valeur aléatoire suivant les distributions normales des $m$ imputations. Ainsi, le réseau de neurones n'est pas entraîné de manière répétitive sur des imputations

arbitrairement fixées (et très probablement non optimales), comme ce serait le cas avec *SI* ou *MI*. Ce processus fonctionne comme une régularisation par le bruit prenant en compte l'incertitude entre les imputations, ce qui permet d'obtenir un réseau neuronal moins biaisé et plus généralisé.

### 3.2 Multiple-Hotpatching

Multiple-Hotpatching (*M-HOT*) étend *S-HOT* à l'aide du paradigme ensembliste. Ce framework entraîne autant de modèles d'inférence que d'imputations effectuées, ces modèles sont entraînés en tenant compte de l'incertitude entre les imputations, ce qui permet d'obtenir des modèles individuellement moins biaisés. Nous montrons empiriquement que *M-HOT* donne des résultats systématiquement meilleurs que *MI* sans pour autant être plus coûteux en termes de calcul. Il est donc avantageux d'utiliser *M-HOT* dans les situations où il est possible de se permettre d'entraîner plusieurs réseaux de neurones de manière ensembliste.

Le framework est similaire à celui de *S-HOT*, à la différence que *M-HOT* repose sur le paradigme ensembliste comme pour *MI*. Nous utilisons les notations mathématiques définies précédemment, $X \in \mathbb{R}^{n \times d}$ est le jeu de données initial incomplet, où chaque valeur $X_{ij}$ est soit observée, soit manquante. Nous effectuons $m$ imputations qui conduisent à $m$ différents jeux de données complétés $\tilde{X}^{1 \cdots m}$, avec $\tilde{X}^k \; dans \mathbb{R}^{n \times d}$ le $k$-ième jeu de données complété. Un modèle est défini par jeu de données complété, conduisant à $m$ modèles. Les écarts types $\sigma$ sont calculé de la même manière que dans l'équation 2, il n'est pas utile de calculer les moyennes. L'ensemble des modèles est ensuite entraîné. Pour extraire une batch $B^k \; dans \mathbb{R}^{b \times d}$ qui sera donnée en entrée au $k$-ième modèle, chaque valeur de la batch est tiré d'une distribution normale paramétrée avec une moyenne $\tilde{X}^k_{ij}$ et un écart type $\alpha \cdot \sigma_{ij}$, tel que $B^k_{ij} \sim \mathcal{N}(\tilde{X}^k_{ij}, \alpha \cdot \sigma_{ij})$. Ainsi, tous les modèles sont entraînés en parallèle, les valeurs manquantes présentées au $k$-ième modèle sont remplacées par des valeurs imputées tirées d'une distribution normale centrée sur la $k$-ième imputation calculée, menant à une diversité accrue des données d'entraînement utilisées. Il en résulte un ensemble de modèles moins biaisés et capables d'une plus grande généralisation que dans *MI*, puisqu'ils prennent en compte l'incertitude entre les imputations. *M-HOT* peut être utilisé comme substitut à *MI* dans toutes les situations où *MI* est viable.

## 4 Expérimentations

### 4.1 Protocole Expérimental

Dans nos expérimentations, nous ne cherchons pas à comparer les performances des méthodes d'imputation utilisées. Nous nous intéressons plutôt aux résultats que nous pouvons observer pour une même méthode d'imputation lorsque nous utilisons chacun des frameworks d'imputation comparés.

Nous avons mené nos expérimentations sur cinq jeux de données tabulaires de classification : le célèbre jeu de données IRIS [1], le jeu de données STATLOG [2], le jeu de données WINE [2], le jeu de données PIMA [3] et le jeu de données ABALONE [2].

Les jeux de données utilisés ne contenant pas de valeurs manquantes, nous avons donc ajouter artificiellement différents taux de valeurs manquantes. Nous simulons trois mécanismes de valeurs manquantes différents au sein des jeux de données, MCAR, MAR et MNAR [9]. Avec le mécanisme MCAR, nous introduisons des valeurs manquantes de manière totalement aléatoire en masquant au hasard un certain taux de valeurs. Pour le mécanisme MAR, un sous-ensemble aléatoire de caractéristiques est choisi pour ne pas être masqué, ce sous-ensemble est utilisé comme entrée d'un modèle logistique et la sortie du modèle est utilisée pour masquer les valeurs des caractéristiques restantes. Pour le mécanisme MNAR, la mise en oeuvre est proche de celle de MAR, mais l'entrée du modèle logistique est masquée à l'aide d'un mécanisme MCAR. Ainsi, la sortie du modèle dépend de valeurs qui sont indifféremment connues ou masquées.

Nous comparons les frameworks à l'aide des méthodes d'imputation décrites plus tôt : MISSFOREST, SOFTIMPUTE, GAIN, MIDA et SINKHORN. Nous appliquons chaque framework d'imputation en utilisant chaque méthode d'imputation et nous nous intéressons uniquement à la comparaison des résultats obtenus entre chaque framework d'imputation. Dans nos expérimentations, nous considérons MICE comme un framework d'imputation auquel nous comparons les résultats de nos frameworks.

Nos expériences visent à évaluer les performances des réseaux neuronaux dans un contexte d'apprentissage supervisé afin de mesurer le biais et la capacité de généralisation du modèle. Nous utilisons un réseau de neurone simple sous la bibliothèque scikit-learn [4], paramétré à l'aide d'hyperparamètres menant à de bons résultats et identiques pour chaque jeu de données afin de garantir une comparaison juste et impartiale. Dans tous les cas nous utilisons l'optimiseur Adam avec un taux d'apprentissage fixé à $0.001$. Les modèles sont composés de deux couches cachées : pour les jeux de données IRIS, STATLOG, PIMA et ABALONE les deux couches sont de dimensions $32$, pour le jeu de données WINE les couches sont de dimensions $64$ et $32$ respectivement. Les performances de prédiction sont évaluées à l'aide de la métrique AUC (aire sous la courbe ROC), l'accuracy équilibrée et le score F1.

Afin de mieux évaluer les résultats obtenus, nous basons nos comparaisons sur les tests statistiques de Friedman et de Nemenyi tels que décrits dans [2]. Le test de Friedman est d'abord utilisé pour vérifier si l'hypothèse nulle selon laquelle tous les frameworks comparés sont statistiquement équivalents pour une $p$-value donnée est rejetée ou non. Si

1. https://scikit-learn.org/stable/modules/generated/sklearets.load_iris.html
2. https://archive.ics.uci.edu
3. https://rioultf.users.greyc.fr/uci/files/pima-indians-diabetes
4. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

l'hypothèse nulle est rejeté, le test de Nemenyi est utilisé pour comparer les frameworks par paire. Le test de Nemenyi peut être facilement visualisé à l'aide d'un simple diagramme, ce qui facilite l'analyse des résultats.

## 4.2 Résultats

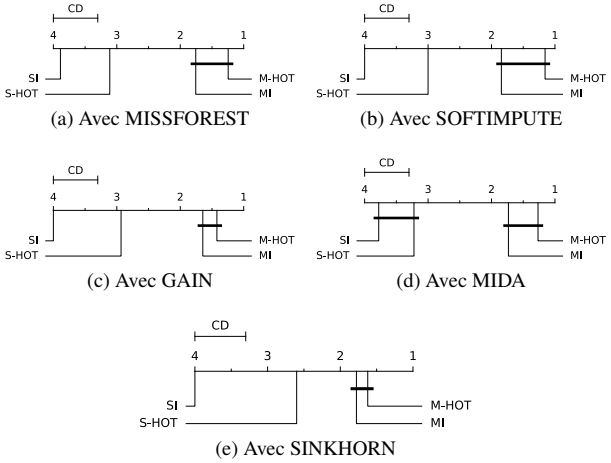
### 4.2.1 Étude Comparative Entre les Frameworks d'Imputation.

Les $m$ imputation sont calculées, à la suite de quoi nous exécutons les quatre frameworks comparés *SI*, *MI*, *S-HOT* et *M-HOT*, en suivant le protocole expérimental décrit précédemment. Le tableau 1 présente une partie des résultats obtenus avec les quatre frameworks comparés lors de l'utilisation de la méthode d'imputation MISSFOREST, pour des raisons de places nous ne pouvons ajouter les résultats complets. Nous n'observons aucune influence du mécanisme ou du taux de valeurs manquantes sur les résultats obtenus, ce qui semble montrer que nos frameworks ne sont pas sensibles à ces paramètres et peuvent être utilisés en toutes circonstances. Nous constatons que le frameworks *M-HOT* obtient les meilleurs résultats dans la grande majorité des cas, tandis que *S-HOT* obtient systématiquement de meilleurs résultats que *SI*.

| Dataset | Pattern | | *SI* | *MI* | *S-HOT* | *M-HOT* |
|---|---|---|---|---|---|---|
| WINE | MCAR | 10% | 0.9987736 (4) | 0.9989672 (2) | 0.9988008 (3) | **0.9989811** (1) |
| | | 15% | 0.9955454 (4) | 0.9960062 (2) | 0.9956407 (3) | **0.9960307** (1) |
| | | 25% | 0.9910498 (4) | 0.9919927 (2) | 0.9914148 (3) | **0.9923485** (1) |
| | MAR | 10% | 0.9961058 (4) | 0.9965056 (2) | 0.9961142 (3) | **0.9965060** (1) |
| | | 15% | 0.9977720 (4) | **0.9982010** (1) | 0.9978677 (3) | 0.9981809 (2) |
| | | 25% | 0.9952116 (4) | 0.9965157 (2) | 0.9959576 (3) | **0.9968702** (1) |
| | MNAR | 10% | 0.9987205 (3) | **0.9988244** (1) | 0.9987058 (4) | 0.9988131 (2) |
| | | 15% | 0.9974746 (4) | 0.9976465 (2) | 0.9974850 (3) | **0.9976683** (1) |
| | | 25% | 0.9808498 (4) | 0.9841291 (2) | 0.9822719 (3) | **0.9844587** (1) |
| PIMA | MCAR | 10% | 0.8193054 (4) | **0.8211340** (1) | 0.8196586 (3) | 0.8211193 (2) |
| | | 15% | 0.8073739 (4) | 0.8095505 (2) | 0.8078378 (3) | **0.8095824** (1) |
| | | 25% | 0.8029002 (4) | 0.8060367 (2) | 0.8043589 (3) | **0.8065611** (1) |
| | MAR | 10% | 0.8238900 (4) | **0.8257089** (1) | 0.8242472 (3) | 0.8256414 (2) |
| | | 15% | 0.8045918 (4) | 0.8080830 (2) | 0.8061503 (3) | **0.8083194** (1) |
| | | 25% | 0.8017568 (4) | **0.8041203** (1) | 0.8025144 (3) | 0.8040062 (2) |
| | MNAR | 10% | 0.8280685 (4) | 0.8302729 (2) | 0.8284115 (3) | **0.8303076** (1) |
| | | 15% | 0.8279577 (4) | 0.8298929 (2) | 0.8283792 (3) | **0.8300464** (1) |
| | | 25% | 0.8005008 (4) | 0.8043448 (2) | 0.8021749 (3) | **0.8047250** (1) |
| ABAL | MCAR | 10% | 0.8737739 (4) | 0.8748059 (2) | 0.8740180 (3) | **0.8749393** (1) |
| | | 15% | 0.8714861 (4) | 0.8725539 (2) | 0.8717831 (3) | **0.8726250** (1) |
| | | 25% | 0.8663833 (4) | 0.8674332 (2) | 0.8666186 (3) | **0.8675645** (1) |
| | MAR | 10% | 0.8742551 (4) | 0.8751972 (2) | 0.8743399 (3) | **0.8753671** (1) |
| | | 15% | 0.8720722 (4) | 0.8731502 (2) | 0.8721856 (3) | **0.8731739** (1) |
| | | 25% | 0.8697060 (4) | 0.8708046 (2) | 0.8699619 (3) | **0.8709102** (1) |
| | MNAR | 10% | 0.8760444 (4) | 0.8768033 (2) | 0.8760447 (3) | **0.8769350** (1) |
| | | 15% | 0.8753217 (4) | 0.8763271 (2) | 0.8754605 (3) | **0.8764456** (1) |
| | | 25% | 0.8683507 (4) | 0.8696780 (2) | 0.8690281 (3) | **0.8697714** (1) |
| Average rank | | | 3.8889 | 1.7556 | 3.1111 | 1.2444 |

TABLE 1 – Résultats de prédictions obtenue après l'application de chaque framework, en utilisant la méthode d'imputation MISSFOREST.

Le test de Friedman permet de rejeter l'hypothèse nulle selon laquelle tous les frameworks seraient équivalent. Nous appliquons le test de Nemenyi et présentons les résultats sous forme graphique dans la Figure 2. Deux frameworks peuvent être considérés comme significativement différents s'ils ne sont pas reliés par une barre noire, c'est à dire si leurs rangs moyens diffèrent de plus que la distance critique indiquée en haut de chaque graphique. Dans tous les cas nous constatons que l'ordre des frameworks, du pire au meilleur, est le même : *SI* est le moins performant, suivi de *S-HOT*, puis de *MI* et enfin de *M-HOT*. *S-HOT* obtient des résultats nettement meilleurs que *SI*, ce qui montre qu'il s'agit d'une bonne alternative à *SI* lorsque l'on souhaite entraîner un large et unique modèle. Nous notons que *M-HOT* obtient systématiquement de meilleurs résultats que *MI*, ce qui semble montrer qu'il s'agit toujours d'une bonne alternative viable à *MI*.




(a) Avec MISSFOREST (b) Avec SOFTIMPUTE

(c) Avec GAIN (d) Avec MIDA

(e) Avec SINKHORN

FIGURE 2 – Tests de Nemenyi comparant les quatre frameworks avec chaque méthode d'imputation, $CD \approx 0.6992$.

### 4.2.2 Étude Comparative Entre MICE et les Frameworks d'Imputation.

Nous comparons *S-HOT* et *M-HOT* à MICE, nous observons que le framework *M-HOT* obtient les meilleurs résultats dans la plupart des cas. Le test de Friedman nous permet à nouveau de rejetter l'hypothèse nulle. Nous appliquons le test de Nemenyi, la Figure 3 présente les résultats. Nous constatons que les performances de MICE sont largement supérieures à celles de *SI*, même si elles ne sont pas significativement meilleures d'un point de vue statistique. *S-HOT* obtient des résultats nettement meilleurs que *SI* et légèrement meilleurs que MICE. Les frameworks *MI* et *M-HOT* sont nettement meilleurs que les autres frameworks et méthodes testés, une fois encore, *M-HOT* obtient de meilleurs résultats que *MI*.

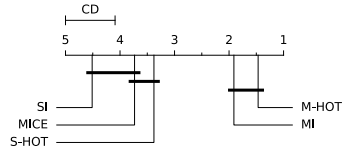



FIGURE 3 – Tests de Nemenyi comparant la méthode d'imputation MICE avec les meilleurs résultats précédemment obtenus sur chaque jeu de données, $CD \approx 0.9093$.

### 4.2.3 Comparaison des Temps d'Exécution.

Enfin, nous avons comparé le temps de calcul nécessaire à l'exécution de chaque framework dans chaque scénario testé. Dans tous les cas, la majeure partie du temps de calcul provient du calcul des $m$ imputations. *SI* bénéficie largement de ce point, les trois autres frameworks nécessitent tous le même temps pour calculer les multiples imputations.

Nous n'observons aucune différence dans le temps d'entraînement requis entre *SI* et *S-HOT*. Dans le cas de *MI* et de *M-HOT*, *M-HOT* est un peu plus lent de moins d'une seconde à quelques secondes pour les modèles et les jeux de données les plus larges en comparaison à *MI*. La différence globale de temps d'exécution entre *MI* et *M-HOT* est négligeable. Étant donné que nous avons montré que *M-HOT* obtient systématiquement de meilleurs résultats que *MI*, la plupart des scénarios bénéficieraient de l'utilisation de *M-HOT* plutôt que de *MI*.

# 5 Discussion and Conclusion

La prise en compte de l'incertitude d'imputation lors de l'apprentissage d'un réseau de neurones ne fait pas l'objet de nombreuses recherches. En effet, les réseaux de neurones sont naturellement capables d'une généralisation suffisante pour négliger les conséquences du biais induit par la non prise en compte de l'incertitude d'imputation. Dans cet article, nous avons étudié et proposé deux frameworks d'imputation qui peuvent être utilisés pour entraîner des modèles en tenant compte de ce niveau d'incertitude, permettant d'obtenir des modèles capables d'une plus grande généralisation et de meilleurs résultats d'inférence.

Les deux frameworks proposés, *S-HOT* et *M-HOT*, visent à remplacer respectivement l'imputation simple (*SI*) et l'imputation multiple (*MI*). Nous réalisons différentes expérimentations pour comparer les frameworks d'imputation et nous montrons qu'ils rivalisent avec d'autres frameworks d'imputation, voire les surpassent dans de nombreuses situations. Nous évaluons statistiquement les résultats à l'aide des tests de Friedman et de Nemenyi et montrons que nos frameworks conduisent à des réseaux neuronaux moins biaisés, ce qui améliore les résultats de l'inférence. Nous comparons également les temps d'exécution requis pour chaque framework et concluons que la différence totale de temps d'exécution entre *MI* et *M-HOT* est négligeable, tandis que *SI* est plus rapide que *S-HOT* mais obtient des résultats nettement moins bons que *S-HOT*. Nous avons montré que *S-HOT* obtient des résultats nettement meilleurs que *SI* dans tous les scénarios testés, et nous en concluons qu'il est avantageux d'utiliser *S-HOT* lorsque l'on doit former un réseau de neurones unique et de grande taille. Nos expériences montrent que *M-HOT* obtient systématiquement de meilleurs résultats que *MI* dans tous les scénarios testés pour un temps d'exécution comparable.

Dans ce travail, nous supposons une distribution normale des valeurs imputées. Nous obtenons de bons résultats empiriques, mais une distribution normale n'est pas forcément toujours pertinente selon la nature de la caractéristique manquante ou de la méthode d'imputation utilisée, de futurs travaux se concentreront sur cette question.

# Remerciements

This research is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 875171, project QUALITOP (Monitoring multidimensional aspects of QUAlity of Life after cancer ImmunoTherapy - an Open smart digital Platform for personalized prevention and patient management).

# Références

[1] Stef van Buuren and Karin Groothuis-Oudshoorn. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3) :1–67, 2011.

[2] Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, pages 1–30, 2006.

[3] Lovedeep Gondara and Ke Wang. MIDA : Multiple Imputation Using Denoising Autoencoders. *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018*, pages 260–272, 2018.

[4] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. *On the consistency of supervised learning with missing values*. ArXiv, February 2019.

[5] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with missing values ? In *Advances in Neural Information Processing Systems*, volume 34, pages 11530–11540. Curran Associates, Inc., 2021.

[6] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research : JMLR*, 11 :2287–2322, March 2010.

[7] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing Data Imputation using Optimal Transport. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7130–7140. PMLR, November 2020. ISSN : 2640-3498.

[8] D. B. Rubin and N. Schenker. Multiple imputation in health-care databases : an overview and some applications. *Statistics in Medicine*, 10(4) :585–598, April 1991.

[9] Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3) :581–592, 1976. Publisher : [Oxford University Press, Biometrika Trust].

[10] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, June 2004. Google-Books-ID : bQBtw6rx_mUC.

[11] Daniel J. Stekhoven and Peter Bühlmann. MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, 28(1), January 2012.

[12] Jinsung Yoon, James Jordon, and Mihaela Schaar. GAIN : Missing Data Imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning*, page 5689. PMLR, July 2018. ISSN : 2640-3498.

[13] Yang Yuan. Multiple Imputation for Missing Data : Concepts and New Development. *SAS Institute Inc.*, January 2005.

# Deux variantes à la méthode de classification FIMIX-PLS dans le cadre des modèles d'équations structurelles

S. Dominique[1,2], M. Hanafi[1], F. Llobell[2], J.M. Ferrandi[3], V. Cariou[1]

[1] Oniris, INRAE, STATSC, 44300 Nantes, France, StatSC
[2] Lumivero, XLSTAT, Paris, France
[3] LEMNA, Oniris, 44300 Nantes, France

sophie.dominique@oniris-nantes.fr

**Résumé**

*Dans le cadre de PLS-SEM, nous proposons deux variantes à la méthode de classification FIMIX-PLS qui vise à déterminer simultanément une partition des individus et la meilleure adéquation entre chaque classe et le modèle PLS-SEM qui lui est associé. Une première variante conduit à imposer une contrainte pour restreindre certains paramètres structurels d'être communs aux classes. Une seconde prend en compte les différences entre classes des scores moyens sur les proxies. Ces deux variantes sont illustrées sur la base de simulations.*

**Mots-clés**

*modèle d'équations structurelles, classification, FIMIX-PLS.*

**Abstract**

*In the context of PLS-SEM, two variants of the FIMIX-PLS clustering approach are proposed. FIMIX-PLS aims to simultaneously determine a partition of the observations and the best adequation between each cluster and its associated PLS-SEM model. A first variant leads to imposing a constraint to restrict some structural parameters from being common to all the clusters. A second takes into account the fact that average scores of the proxies may differ from one cluster to another. These two variants are illustrated on the basis of simulations.*

**Keywords**

*Structural equation modeling, clustering, FIMIX-PLS.*

## 1 Introduction

Les modèles d'équations structurelles (SEM) se sont imposés dans de nombreux domaines d'applications par leur capacité à modéliser un ensemble de relations définies a priori entre différents construits, souvent unidimensionnels, déterminés à partir de blocs de variables mesurées, et à en estimer les paramètres. Parmi les approches de modélisation proposées, la méthode PLS-SEM, encore appelée PLS-PM, a gagné en popularité depuis ces dernières décennies pour sa versatilité et la possibilité de déterminer un proxy reflétant pour chaque individu son score associé à chacun des construits [7].

Afin de prendre en compte l'éventuelle hétérogénéité des individus pouvant entâcher l'estimation des paramètres du modèle PLS-SEM, plusieurs approches classificatoires ont été proposées ([4, 2, 1, 6, 8, 3]). Ces dernières reposent sur une approche clusterwise dans la mesure où elles visent à partitionner les individus en classes, avec l'estimation de paramètres spécifiques à chaque classe, de manière à maximiser l'adéquation entre les individus qui la composent et le modèle qui leur est associé.

Parmi les méthodes classificatoires proposées, FIMIX-PLS ([4, 5]), est celle qui reste la plus utilisée dans le cadre de PLS-SEM. Dans FIMIX-PLS, seuls les paramètres correspondant au modèle structurel de PLS-SEM peuvent varier d'une classe à une autre, impliquant de fait que les proxies, obtenus comme combinaison linéaire des variables manifestes du bloc auquel chacune est respectivement associée, soient communes à l'ensemble des classes.

Partant de ce postulat, nous proposons ici deux variantes à FIMIX-PLS. Dans la première, nous introduisons une contrainte conduisant à restreindre certains paramètres structurels à être communs à l'ensemble des classes, reflétant des hypothèses définies a priori par l'analyste et ainsi faciliter l'interprétation des modèles. Il s'agit de Local FIMIX-PLS. Dans une seconde variante, nommée Moving FIMIX-PLS, nous proposons d'associer à chaque classe un vecteur de paramètres supplémentaires reflétant le score moyen de la classe pour chaque proxy.

## 2 Variantes développées sur la base de FIMIX-PLS

Adoptant l'approche des modèles de mélange, la méthode FIMIX-PLS suppose que les individus sont issus de différentes sous-populations, appelées classes. Ces dernières diffèrent les unes des autres par les paramètres de leur distribution et sont présentes selon une certaine proportion dans la population. Dans FIMIX-PLS, les proxies sont supposés communs aux différentes classes, les paramètres

des distributions à estimer sont donc strictement liés au modèle structurel.

## 2.1 Rappel de FIMIX-PLS

La méthode FIMIX-PLS consiste en deux étapes pour déterminer la partition et les paramètres spécifiques à chaque classe. Dans une première étape, les proxies et leurs scores associés sont calculés pour l'ensemble de la population à l'aide de l'algorithme PLS-SEM. Dans une deuxième étape, les paramètres spécifiques aux différentes classes ainsi que la probabilité d'appartenance d'un individu à celles-ci sont déterminés.

Etant donné un individu $i$ de la population, on note $\eta_i$ (resp. $\xi_i$ ) le vecteur des variables endogènes (resp. exogènes) du modèle structurel. Soient $\mathbf{B}$ de dimension $J \times J$ (resp. $\mathbf{\Gamma}$ de dimension $J \times P$), la matrice des coefficients structurels établissant les relations entre les variables endogènes (resp. entre les variables endogènes et exogènes). Soient enfin $\zeta_i$, le vecteur des résidus associés au modèle structurel et $\mathbf{\Psi}$, de dimension $J \times J$, la matrice diagonale des variances résiduelles associées aux équations de régression du modèle structurel. Dans le cadre général, celles-ci s'écrivent :

$$\eta_i = \mathbf{B}\eta_i + \mathbf{\Gamma}\xi_i + \zeta_i \tag{1}$$

$$\Leftrightarrow (\mathbf{I} - \mathbf{B})\,\eta_i - \mathbf{\Gamma}\xi_i = \zeta_i \tag{2}$$

Supposant que la population est constituée de $K$ classes, la méthode FIMIX-PLS vise ainsi à estimer les paramètres $\{\mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k | k = 1, \ldots, K\}$ spécifiques au modèle structurel de chaque sous-population. Sous l'hypothèse de normalité, les variables endogènes suivent la loi de densité $f_{i|k}$ telle que :

$$\eta_i \sim \sum_{k=1}^{K} \pi_k f_{i|k}\left(\eta_i | \xi_i, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right) \tag{3}$$

$$\eta_i = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{J/2}\left|\mathbf{\Psi}_k\right|^{1/2}} \times e^{-\frac{1}{2}((\mathbf{I}-\mathbf{B}_k)\eta_i + (-\mathbf{\Gamma}_k)\xi_i)^T \mathbf{\Psi}_k^{-1}((\mathbf{I}-\mathbf{B}_k)\eta_i + (-\mathbf{\Gamma}_k)\xi_i)} \tag{4}$$

où $\pi_k$ représente la proportion de la classe $G_k$ dans la population ($\pi_k > 0\ \forall k$ et $\sum_{k=1}^{K} \pi_k = 1$).

En considérant les vecteurs $\eta_i$ indépendants et identiquement distribués, la fonction de vraisemblance s'écrit :

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f_{i|k}\left(\eta_i | \xi_i, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right) \tag{5}$$

Les probabilités d'appartenance d'un individu $i$ aux classes sont déterminées en utilisant le théorème de Bayes conditionnellement aux paramètres de chaque classe :

$$p_{ik} = \frac{\pi_k f_{i|k}\left(\eta_i | \xi_i, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right)}{\sum_{k=1}^{K} \pi_k f_{i|k}\left(\eta_i | \xi_i, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right)} \tag{6}$$

Enfin, les paramètres $(\pi_k, \mathbf{B}_k, \mathbf{\Gamma}_k)_{k=1,\ldots,K}$ sont estimés en optimisant la fonction de vraisemblance par un algorithme EM (Expectation-Maximization). La log-vraisemblance s'écrit alors :

$$\ln L_C = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln\left(f_{i|k}\left(\eta_i | \xi_i, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right)\right) + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln\left(\pi_k\right) \tag{7}$$

avec $z_{ik} = 1$ si l'individu appartient au groupe $k$, 0 sinon.

Cette dernière formule correspond au critère de convergence de la méthode FIMIX-PLS.

## 2.2 Intégration d'une contrainte sur les paramètres des modèles structurels associés aux classes

L'approche locale adopte les mêmes étapes que FIMIX-PLS, c'est-à-dire 1) la détermination des proxies associés à $\eta$ et $\xi$ avec PLS-SEM puis 2) l'estimation des paramètres des classes et des probabilités d'appartenance par la maximisation de la fonction de vraisemblance grâce à un algorithme d'espérance-maximisation (EM) et enfin 3) la détermination de la partition à l'aide des probabilités d'appartenance obtenues. L'intégration d'une contrainte, appelée Local FIMIX-PLS, induit un modèle restreint où seul un sous-ensemble des variables endogènes, notées $\eta^{(s)}$ suit une loi de probabilité correspondant à un mélange de distributions. Par la suite, notons $\mathbf{B}^{(s)}$ et $\mathbf{\Gamma}^{(s)}$, les matrices des coefficients structurels associés à ce sous-ensemble des variables endogènes. Le sous-modèle structurel correspondant est égal à :

$$\eta_i^{(s)} = \mathbf{B}_k^{(s)}\eta_i + \mathbf{\Gamma}_k^{(s)}\xi_i + \zeta_i \tag{8}$$

FIMIX-PLS est donc appliqué sur une sous partie du modèle avec :

$$\eta_i \sim \sum_{k=1}^{K} \pi_k f_{i|k}\left(\eta_i \middle| \xi_i, \mathbf{B}_k^{(s)}, \mathbf{\Gamma}_k^{(s)}, \mathbf{\Psi}_k^{(s)}\right) \tag{9}$$

$$\eta_i^{(s)} = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{J^{(s)}/2}\left|\mathbf{\Psi}_k^{(s)}\right|^{1/2}} \times e^{-\frac{1}{2}\left(\left(\mathbf{I}^{(s)}-\mathbf{B}_k^{(s)}\right)\eta_i + \left(-\mathbf{\Gamma}_k^{(s)}\right)\xi_i\right)^T \mathbf{\Psi}_k^{(s)-1}\left(\left(\mathbf{I}^{(s)}-\mathbf{B}_k^{(s)}\right)\eta_i + \left(-\mathbf{\Gamma}_k^{(s)}\right)\xi_i\right)} \tag{10}$$

avec $J^{(s)}$, le nombre de variables endogènes spécifiques au modèle local et $I^{(s)}$, la matrice identité ($J^{(s)} \times J^{(s)}$).

### 2.3 Prise en compte d'éventuelles différences dans les scores moyens des variables synthétiques associées aux classes

Dans la méthode FIMIX-PLS, les coefficients de régression associés aux constantes des équations du modèle structurels sont supposés être égaux sur l'ensemble des classes. Les proxies déterminés par PLS-SEM étant centrés, il en résulte qu'ils sont donc tous définis comme égaux à zéro. Dans la pratique, nous pouvons néanmoins supposer que les sous populations diffèrent non seulement du fait de leurs coefficients de régression associés aux variables endogènes et exogènes ; mais également du fait des coefficients de régression relatifs aux constantes du modèle. Si on note $\alpha_k$ le vecteur de ces coefficients, on a les équations suivantes qui décrivent la variante Moving FIMIX-PLS :

$$\eta_i \sim \sum_{k=1}^{K} \pi_k f_{i|k}\left(\eta_i|\xi_i, \alpha_k, \mathbf{B}_k, \mathbf{\Gamma}_k, \mathbf{\Psi}_k\right) \tag{11}$$

$$\eta_i = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{J/2}\left|\mathbf{\Psi}_k\right|^{1/2}} \times e^{-\frac{1}{2}((\mathbf{I}-\mathbf{B}_k)\eta_i+(-\mathbf{\Gamma}_k)\xi_i-\alpha_k)^T\mathbf{\Psi}_k^{-1}((\mathbf{I}-\mathbf{B}_k)\eta_i+(-\mathbf{\Gamma}_k)\xi_i-\alpha_k)} \tag{12}$$

## 3 Simulations

### 3.1 Local FIMIX-PLS

Le modèle utilisé pour l'étude de simulation inclut trois variables exogènes et trois variables endogènes, dont les relations sont représentées dans la Figure 1. Deux classes sont construites en définissant une relation forte parmi les variables exogènes, différente entre chaque classe.
Le partitionnement local est appliqué sur la première partie du modèle de structure qui se concentre sur les relations des variables exogènes.

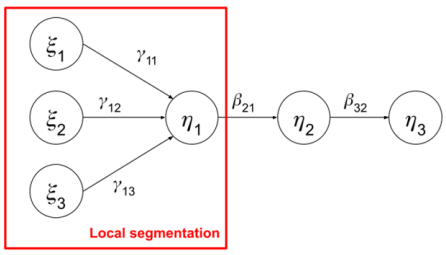



FIGURE 1 – Modèle étudié.

Les différents scénarios, variant suivant la taille de la population, la taille relative des classes et les écarts entre classes des différents coefficients de régression, sont représentés dans le tableau 1.

| Case | Nombre d'individus | Taille relative des classes | Ecart entre les coefficients |
|---|---|---|---|
| 1 | Grand | Équilibrée | Grand |
| 2 | Grand | Équilibrée | Petit |
| 3 | Grand | Déséquilibrée | Grand |
| 4 | Grand | Déséquilibrée | Petit |
| 5 | Petit | Équilibrée | Grand |
| 6 | Petit | Équilibrée | Petit |
| 7 | Petit | Déséquilibrée | Grand |
| 8 | Petit | Déséquilibrée | Petit |

TABLE 1 – Liste des cas étudiés dans l'étude de simulation.

Les performances des deux méthodes (FIMIX-PLS, Local FIMIX-PLS) sont évaluées suivant la moyenne des différences absolues entre les coefficients générés et estimés (voir Table 2).

| | $\gamma_{11}$ (spécifique) | | $\gamma_{12}$ (spécifique) | | $\gamma_{13}$ (spécifique) | |
|---|---|---|---|---|---|---|
| Cas | Local | FIMIX-PLS | Local | FIMIX-PLS | Local | FIMIX-PLS |
| 1 | 0.025 | 0.025 | 0.035 | 0.040 | 0.025 | 0.020 |
| 2 | 0.180 | 0.165 | 0.220 | 0.165 | 0.235 | 0.190 |
| 3 | 0.030 | 0.030 | 0.035 | 0.030 | 0.025 | 0.030 |
| 4 | 0.255 | 0.260 | 0.195 | 0.235 | 0.220 | 0.240 |
| 5 | 0.040 | 0.035 | 0.040 | 0.040 | 0.035 | 0.035 |
| 6 | 0.265 | 0.295 | 0.280 | 0.220 | 0.225 | 0.210 |
| 7 | 0.055 | 0.070 | 0.065 | 0.075 | 0.060 | 0.065 |
| 8 | 0.285 | 0.255 | 0.260 | 0.230 | 0.295 | 0.215 |

| | $\beta_{21}$ (commun) | | $\beta_{23}$ (commun) | |
|---|---|---|---|---|
| Cas | Local | FIMIX-PLS | Local | FIMIX-PLS |
| 1 | 0.025 | 0.035 | 0.020 | 0.045 |
| 2 | 0.020 | 0.150 | 0.020 | 0.225 |
| 3 | 0.030 | 0.055 | 0.030 | 0.060 |
| 4 | 0.030 | 0.190 | 0.030 | 0.250 |
| 5 | 0.030 | 0.045 | 0.040 | 0.070 |
| 6 | 0.030 | 0.200 | 0.030 | 0.215 |
| 7 | 0.045 | 0.100 | 0.040 | 0.115 |
| 8 | 0.040 | 0.215 | 0.055 | 0.205 |

TABLE 2 – Moyennes des différences entre les coefficients estimés par FIMIX-PLS et par Local FIMIX-PLS suivant chaque cas.

Les résultats sur les coefficients spécifiques sont relativement similaires entre les deux méthodes. Le fait de contraindre FIMIX-PLS n'améliore pas les résultats mais ne les détériore pas non plus quant à l'estimation des paramètres structurels de la partie locale. Dans la mesure où Local FIMIX-PLS est un modèle davantage parcimonieux, il est par conséquent à privilégier dans ce cas de figure.

L'aspect notable et qui peut paraître surprenant est le comportement de FIMIX-PLS sur la partie commune du modèle où il ne doit pas y avoir de différence entre les coefficients de régression des classes. En effet, les écarts entre les coefficients simulés et les estimations par la méthode FIMIX-PLS sont plus importants qu'avec Local FIMIX-PLS. C'est le cas en particulier pour les scénarii 2, 4, 6 et 8 qui correspondent à des écarts plus faibles entre classes sur les coefficients de régression générés pour la partie locale. Il semble donc que FIMIX-PLS tend à forcer les différences entre les coefficients associées aux classes, cette stratégie vient alors détériorer la solution lorsque cet écart est plus faible et réside seulement sur une partie du modèle structurel. Cette différence perturbe ainsi les résultats et peut amener à une mauvaise interprétation. En effet, elle indique des coefficients différents entre les classes alors que, comme dans nos simulations, il n'y en a pas pour une partie du modèle

structurel.

### 3.2 Moving FIMIX-PLS

Dans cette partie, nous considérons un modèle plus simple, comprenant deux variables exogènes et une variable endogène (voir Figure 2).

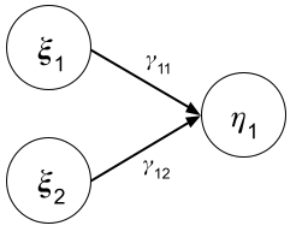



FIGURE 2 – Modèle étudié.

Nous simulons un jeu de données en intégrant des scores moyens des variables endogènes différents entre les classes, comme l'indique la Figure 3.

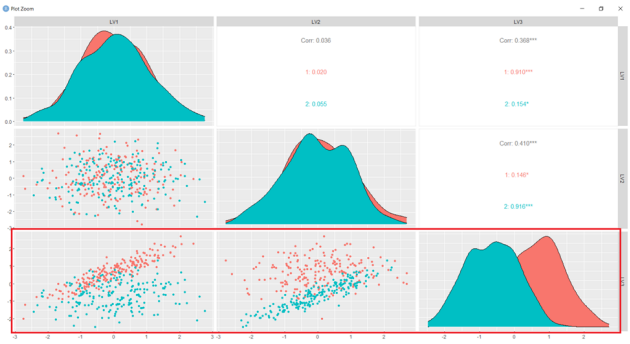


FIGURE 3 – Description des scores des variables latentes.

Les estimations des paramètres par la méthode FIMIX-PLS avec l'intégration de la constante sont de fait plus proches des paramètres estimés et nous retrouvons de manière naturelle les coefficients de détermination $R^2$ calculés par l'application de PLS-SEM sur chaque classe qui avait été générée. Il est à noter que la prise en compte des constantes, si elle s'avère naturelle dans des méthodes comme REBUS, n'est pas considérée dans FIMIX-PLS.

| | Moving FIMIX-PLS | | FIMIX-PLS | | Paramètres simulés | |
|---|---|---|---|---|---|---|
| | Classe 1 (N = 195) | Classe 2 (N = 205) | Classe 1 (N = 326) | Classe 2 (N = 74) | Classe 1 (N = 200) | Classe 2 (N = 200) |
| $\gamma_{11}$ | 0.73 | 0.07 | 0.60 | -0.13 | 0.91 | 0.11 |
| $\gamma_{12}$ | 0.10 | 0.70 | 0.19 | 0.86 | 0.13 | 0.91 |
| Constante | 0.70 | -0.64 | - | - | - | - |
| $R^2(\eta_1)$ | 0.86 | 0.86 | 0.43 | 0.78 | 0.84 | 0.85 |

TABLE 3 – Estimation des coefficients par la méthode FIMIX-PLS et la variante intégrant une constante.

## 4 Conclusion

La notion de classes dans le cadre de modèles d'équations structurelles avec PLS-SEM s'avère difficile à définir rigoureusement. FIMIX-PLS, en imposant les proxies d'être les mêmes pour les classes, simplifie ce problème en supposant une hétérogénéité concentrée sur le modèle de structure. Dans certains domaines, tels qu'en Marketing, cette hypothèse paraît pleinement justifiée. Sur la base des proxies obtenus par PLS-SEM, FIMIX-PLS vise donc à déterminer une partition où les classes diffèrent les unes des autres par leurs coefficients associés au modèle structurel. Nous avons vu ici les limites d'une telle approche lorsque l'écart entre ces coefficients est plus faible et n'apparaît que sur une partie des équations structurelles, FIMIX-PLS va chercher à amplifier ces différences et échoue à retrouver correctement les classes initiales. La variante Local FIMIX-PLS vise à pallier à ce premier problème. Indépendamment, nous avons vu, à travers le jeu de données simulé, l'importance de l'introduction des constantes lors des étapes d'estimation des paramètres par la variante Moving FIMIX-PLS. Cet ajout très simple permet en effet de gagner significativement en qualité des modèles générés et s'avère également intéressant à prendre en compte, in fine, dans l'interprétaion des classes obtenues.

## Références

[1] Jan-Michael Becker, Arun Rai, Christian M Ringle, and Franziska Völckner. Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS quarterly*, pages 665–694, 2013.

[2] Vincenzo Esposito Vinzi, Laura Trinchera, Silvia Squillacciotti, and Michel Tenenhaus. Rebus-pls : A response-based procedure for detecting unit segments in pls path modelling. *Applied Stochastic Models in Business and Industry*, 24(5) :439–458, 2008.

[3] Mario Fordellone and Maurizio Vichi. Finding groups in structural equation modeling through the partial least squares algorithm. *Computational Statistics & Data Analysis*, 147 :106957, 2020.

[4] Carsten Hahn, Michael D Johnson, Andreas Herrmann, and Frank Huber. Capturing customer heterogeneity using a finite mixture pls approach. *Schmalenbach Business Review*, 54 :243–269, 2002.

[5] Christian M Ringle, Marko Sarstedt, and Erik A Mooi. Response-based segmentation using fimix-pls : Theoretical foundations and an application to american customer satisfaction index data. *Annals of Information Systems*, 8(1) :19–49, 2010.

[6] Christian M Ringle, Marko Sarstedt, and Rainer Schlittgen. Genetic algorithm segmentation in partial least squares structural equation modeling. *OR spectrum*, 36 :251–276, 2014.

[7] Marko Sarstedt, Joseph F Hair, Mandy Pick, Benjamin D Liengaard, Lăcrămioara Radomir, and Christian M Ringle. Progress in partial least squares structural equation modeling use in marketing research in the last decade. *Psychology & Marketing*, 39(5) :1035–1064, 2022.

[8] Rainer Schlittgen, Christian M Ringle, Marko Sarstedt, and Jan-Michael Becker. Segmentation of pls path mo-

dels by iterative reweighted regressions. *Journal of Business Research*, 69(10) :4583–4592, 2016.

# Étude sur la classification d'entités nommées

I. Keraghel[1, 2], S. Morbieu[2], M. Nadif[1]

[1] Centre Borelli UMR9010, Université Paris Cité, 75006 Paris
[2] Kernix, 75014, Paris

## Résumé

*La reconnaissance et la classification d'entités nommées au sein de documents est un processus permettant d'extraire des informations pertinentes pour une tâche donnée. Cependant, les méthodes disponibles pour effectuer ce processus varient selon leur nature et leur mode de fonctionnement. Dans cet article, nous nous intéressons principalement aux performances des frameworks de reconnaissance d'entité nommée. Une étude comparative entre les frameworks les plus connus aujourd'hui est proposée afin de sélectionner l'approche qui s'adapte le mieux aux différents jeux de données.*

## Mots-clés

*Reconnaissance d'entités nommées, classification, apprentissage profond, transformers.*

## Abstract

*Named entity recognition and classification is a process for extracting relevant information for a given task. However, the methods available to perform this process vary in nature and mode of operation. In this paper, we are mainly interested in the performance of named entity recognition frameworks. A comparative study between the most known frameworks is proposed in order to select the approach that best fits the different datasets.*

## Keywords

*Named entity recognition, classification, machine learning, deep learning, transformers.*

## 1 Introduction

La reconnaissance et la classification d'entités nommées (NERC) est une technique qui consiste à localiser des concepts clés susceptibles de représenter des entités nommées dans des textes, afin de les classer dans un ensemble prédéfini de classes dépendant de la problématique à traiter. Cette technique apparaît aujourd'hui comme une composante principale dans plusieurs domaines de traitement automatique du langage naturel dont la traduction automatique [1], les systèmes de questions-réponses [2] et la recherche d'information [3].
De nombreux systèmes de NERC [4] ont été développés en particulier dans le monde anglo-saxon, puis dans d'autres langues comme le français. Dans [5], les auteurs présentent un système basé sur des règles permettant d'extraire des noms d'entreprises. Les premiers systèmes de NERC utilisaient des algorithmes basés sur des règles manuscrites, des lexiques et des caractéristiques orthographiques. Ces systèmes ont été d'abord suivis par des algorithmes basés sur l'apprentissage automatique comme [6], et ensuite par des réseaux de neurones [7] et récemment des transformers [8].

Pour la NERC, cinq familles de méthodes sont souvent évoquées :

- celles basées sur des règles, qui ne nécessitent pas l'acquisition d'un jeu données annoté car les règles sont élaborées par un expert ;
- l'apprentissage automatique non supervisé ;
- l'apprentissage automatique supervisé, avec une ingénierie pour créer les caractéristiques ;
- l'apprentissage profond, qui s'appuit sur l'utilisation des réseaux de neurones comme les LSTM et CNN ;
- et enfin celles basées sur les transformers qui s'appuient sur l'utilisation du mécanisme d'attention.

Dans la littérature, il existe un bon nombre d'études dans le domaine des NERC. En particulier, l'étude [6] présente une vue d'ensemble des techniques de NERC, depuis l'apparition des systèmes à base de règles jusqu'aux systèmes basés sur l'apprentissage automatique. Les auteurs dans [9] ont analysé les travaux pertinents de la NERC dans les textes biomédicaux, sur la période de 2007-2009. Marrero et al. [10] ont résumé les travaux sur la NERC d'un point de vue théorique et pratique, et ont démonté que cette tâche est loin d'être résolue en 2013. Des études récentes ont inclus dans leurs analyses les méthodes basées sur des réseaux de neurones et les transformers [11].

Dans cet article, nous nous intéressons principalement aux performances des frameworks de NERC sur différents jeux de données. Une étude comparative est proposée afin de sélectionner l'approche qui s'adapte le mieux sur les données testées. Pour ce faire, nous présentons dans un premier temps les approches et les frameworks de NERC les plus connus que nous avons retenus (Hugging Face, Stanford CoreNLP, NLTK, OpenNLP, spaCy et Flair) afin de dresser notre étude comparative. Ensuite, nous présenterons nos données à tester ainsi que notre méthodologie d'expérimentation que nous évaluerons en se basant les métriques appropriées.

# 2 Méthodes

Plusieurs types de méthodes permettent la reconnaissance et l'extraction des entités nommées. Nous discutons dans cette partie des différentes approches.

## 2.1 Règles et base de connaissances

Les systèmes de reconnaissance d'entités nommées à base de connaissances s'appuient sur des ressources lexicales et des connaissances spécifiques au domaine [12]. Ces approches impliquent de confectionner des règles d'extraction, à la manière d'une expression rationnelle. Elles se fondent sur l'extraction des entités nommées en utilisant des marqueurs lexicaux [13] et des ressources terminologiques de noms propres [12].

Les marqueurs lexicaux sont des indices qui encadrent l'entité nommée et qui permettent de dévoiler sa présence. Par exemple, le système CasEN [14] est destiné à reconnaître les organisations politiques.

Les ressources terminologiques regroupent généralement une collection d'entités nommées les plus fréquentes dans un domaine d'application. Par exemple, le système *ProMiner* [15] utilise un dictionnaire de noms des protéines et de gènes.

## 2.2 Apprentissage automatique

Les méthodes classiques d'apprentissage automatique permettent de résoudre les lacunes des méthodes basées sur des règles. Généralement, ces méthodes sont divisées en trois catégories : non supervisées, semi-supervisées et supervisées.

Les méthodes non supervisées exploitent les ressemblances dans les données afin d'inférer le bon modèle. Elles regroupent des syntagmes selon les propriétés communes qu'ils présentent [16].

Les méthodes semi-supervisées apprennent à partir d'un petit nombre d'exemples annotés, les données non annotées étant annotées en fonction de la similitude entre les échantillons non annotés et ceux annotés. Ceci augmente efficacement le volume de données d'apprentissage [17].

En utilisant les méthodes supervisées, la tâche de NERC a été transformée en deux sous problèmes : la classification et l'étiquetage de séquences. On cherche alors à reproduire un schéma d'annotation appris d'un corpus déjà annoté. Les prédictions sont faites en utilisant le modèle déjà formé sur les données annotées. Les algorithmes les plus courants dans cette catégorie sont le modèle de Markov caché (HMM) [18, 19], le modèle d'entropie maximale (ME) [20], les machines à vecteurs de support (SVM) [21] et les champs aléatoires conditionnels (CRF) [22].

## 2.3 Apprentissage profond

Ces dernières années, les méthodes d'apprentissage profond ont été largement utilisées pour la NERC dans plusieurs domaines, et ont donné de bons résultats sur la plupart des corpus en dépassant les modèles traditionnels [23]. Cette famille de méthodes résout la tâche de NERC en exploitant (1) la sémantique en s'appuyant sur des word embeddings tels que Word2vec [24], GloVe [25], fastText [26], (2) les caractéristiques des entités nommées en s'appuyant sur les embeddings de caractères tels que les réseaux neuronaux convolutifs (CNN) [27], (3) l'ordre des mots par le biais des réseaux récurrents (RNN) [28], (4) l'encodage du contexte en vue de la détection des entités nommées en utilisant des RNN ou des CNN, (5) la modélisation probabiliste des étiquettes en utilisant un CRF [28].

## 2.4 Modèles de langage

Depuis 2017, le traitement automatique du langage a fait un grand pas en avant avec l'avènement des *transformers* [29]. Dans le domaine de NERC, ces modèles peuvent être utilisés afin d'extraire des vecteurs de caractéristiques qui servent d'entrée pour d'autres modèles tels que le CRF ou le LSTM, comme ils peuvent être entraînés directement pour la tâche de classification d'entités nommées. Plusieurs auteurs ont étudié l'influence de ces méthodes dans des tâches de NERC [30], et ont observé que leur utilisation, que ce soit comme modèle de représentation de caractères ou comme classifieur, améliore considérablement les performances de leur modèle.

# 3 Frameworks

Nous listons dans cette partie les principaux frameworks de NERC.

**spaCy**[1] : une bibliothèque libre et open source dédiée au traitement automatique du langage naturel en Python. Parmi les fonctionnalités disponibles dans SpaCy, on peut citer : tokenisation, classification, Parts-of-Speech tagging et NER. Il existe plusieurs modèles pré-entraînés dans spaCy que nous pouvons utiliser pour des tâches telles que NER, extraction d'informations, etc.

**NLTK**[2] : une suite de modules Python dédiés au traitement naturel du langage [31]. NLTK intègre plus de 50 corpus et ressources lexicales telles que WordNet. Au contraire de spaCy qui regroupe des algorithmes adaptés à différentes problématiques, et qui sont gérés par la bibliothèque, NLTK laisse la liberté aux utilisateurs de choisir parmi une large gamme d'algorithmes.

**Stanford CoreNLP**[3] : une bibliothèque qui implémente en Java des modèles de reconnaissance d'entités nommées basés sur des CRF [32]. Malheureusement, les modèles proposés ne sont pas disponibles pour toutes les langues.

**Apache OpenNLP**[4] : une bibliothèque Apache [33] qui prend en charge les tâches NLP les plus courantes, comme l'extraction d'entités nommées, la détection de langue, l'étiquetage morpho-syntaxique, le découpage, etc. Au contraire des autres frameworks qui détectent les entités nommées avec le même modèle, OpenNLP propose un modèle spécialisé pour chaque type d'entité nommée.

**Flair**[5] : une bibliothèque open source qui permet de mettre en place un pipeline de traitement automatique de langage

1. https://spacy.io/
2. https://www.nltk.org/
3. https://stanfordnlp.github.io/CoreNLP/
4. https://opennlp.apache.org/
5. https://github.com/flairNLP/flair

naturel supportant des applications multilingues [34]. Flair permet d'utiliser de nombreux modèles de langage, comme Flair, BERT et CamemBERT. Elle offre aussi la possibilité de les combiner.

**Hugging Face** : une bibliothèque créée en 2015[6]. Elle fournit des technologies de traitement du langage naturel open source. Hugging Face propose deux types de services : des services gratuits et des services payants destinés aux entreprises. Hugging Face doit sa popularité à sa bibliothèque transformers, qui offre une API permettant d'accéder à plusieurs modèles pré-entraînés notamment pour la reconnaissance d'entités nommées comme BERT, CamemBERT, etc. Hugging Face propose également une plateforme de collaboration permettant aux utilisateurs de créer, entraîner et partager leurs modèles de deep learning.

# 4 Évaluation

Nos expériences ont été menées sur deux jeux de données issues de sources diverses, et permettant de rechercher des entités nommées dans des classes telles que les noms de personnes, les noms d'organisations, les noms de lieux.

## 4.1 Jeux de données

**CoNLL-2003** : un jeu de données constitué principalement d'articles de presse provenant de Reuters, et se concentre sur quatre types d'entités nommées : les personnes (PER), les lieux (LOC), les organisations (ORG) et les noms d'entités diverses (MISC). CoNLL-2003 couvre deux langues (anglais et allemand), et est utilisé pour étalonner les modèles sur la tâche de NERC. Dans notre étude, nous n'avons utilisé que l'anglais. Pour chaque langue, nous disposons de trois ensembles de données : apprentissage, test, et validation.

**WikiANN** : un jeu de données composé d'articles Wikipédia annotés avec trois types d'entités nommées : les personnes, les lieux et les organisations. Il a été construit en utilisant les entités liées dans les pages Wikipédia pour 282 langues différentes. Nous avons utilisé la version de [35], qui prend en charge 176 des 282 langues du corpus WikiANN original. Cette version contient 40 000 exemples dans sa base anglaise, mais nous nous sommes limités a 10 000 exemples (6 000 pour l'apprentissage, 2 000 pour la validation et 2 000 pour les tests) afin d'évaluer les algorithmes sur un jeu de données moins fourni.
Les tables 1 et 2 résument les caractéristiques de ces jeux de données.

TABLE 1 – Description des jeux de données

| Données | Ensemble | Articles | Phrases | Tokens |
|---|---|---|---|---|
| CoNLL | apprentissage | 946 | 14 041 | 203 621 |
| | validation | 216 | 3 250 | 51 362 |
| | test | 231 | 3 453 | 46 435 |
| WikiANN | apprentissage | - | 20 000 | 160 394 |
| | validation | - | 10 000 | 80 536 |
| | test | - | 10 000 | 80 326 |

TABLE 2 – Nombres d'entités nommées

| | | Catégories d'entités nommées | | | |
|---|---|---|---|---|---|
| Données | Ensemble | PER | LOC | ORG | MISC |
| CoNLL | apprentissage | 6 600 | 7 140 | 6 321 | 3 438 |
| | validation | 1 842 | 1 837 | 1 341 | 922 |
| | test | 1 617 | 1 668 | 1 661 | 702 |
| WikiANN | apprentissage | 9 164 | 9 345 | 9 422 | - |
| | validation | 4 635 | 4 834 | 4 677 | - |
| | test | 4 556 | 4 657 | 4 745 | - |

## 4.2 Mesures

L'évaluation des systèmes de reconnaissance d'entités nommées repose sur la comparaison des prédictions avec une base de référence en utilisant une évaluation exacte ou relâchée. Dans l'évaluation exacte, les contours et la classe de l'entité nommée doivent correspondre à la base de référence. En revanche, l'évaluation relâchée repose sur un système de score où chaque entité avec la bonne classe est créditée, même si les contours ne sont pas exacts. Les entités sont aussi créditées si les contours sont bons, malgré l'inexactitude de la classe.
Les campagnes d'évaluation de MUC [36] et ACE [37] reposent sur des méthodes d'évaluation relachée. Cependant, CoNLL-2003 [38] adopte l'évaluation exacte, qui représente la méthode la plus utilisée aujourd'hui afin d'évaluer les systèmes de reconnaissance d'entités nommées.
Pour l'évaluation des entités nommées, on utilise souvent des métriques classiques comme la précision, le rappel et le F1-score que nous rappelons ci-après.

— La précision donne le nombre d'entités nommées bien reconnues par le modèle rapporté au nombre total d'entités nommées

$$Precision = \frac{VP}{VP + FP} \tag{1}$$

où VP désigne le nombre de Vrais Positifs et FP de Faux Positifs.

— Le rappel mesure le nombre d'entités nommées pertinentes retrouvées par le modèle au regard du nombre total d'entités nommées pertinentes :

$$Rappel = \frac{VP}{VP + FN} \tag{2}$$

où FN est le nombre de Faux Négatifs.

— Le f1-score permet d'évaluer la capacité du modèle à détecter efficacement les entités nommées, en faisant un compromis entre la précision et le rappel.

$$F1 = 2 \times \frac{precision \times rappel}{precision + rappel}. \tag{3}$$

# 5 Expériences

Les trois étapes suivantes ont été réalisées :

— **Mise sous un format commun des jeux de données :** Chaque portion de texte est décrite par sa

6. https ://huggingface.co/

TABLE 3 – Comparaison des frameworks de reconnaissance d'entités nommées

| | | CoNLL-2003 | | | WikiANN | | |
|---|---|---|---|---|---|---|---|
| **Frameworks** | **Algorithmes** | **Precision** | **Rappel** | **F1** | **Precision** | **Rappel** | **F1** |
| Spacy | en_core_web_sm | 82.72 | 81.94 | 82.33 | 73.44 | 72.71 | 73.07 |
| | en_core_web_md | 84. 96 | 84.54 | 84.75 | 80.12 | 79.05 | 79.58 |
| | en_core_web_lg | 86.53 | 86.63 | 86.58 | 81.25 | 80.16 | 80.70 |
| | en_core_web_trf | 89.47 | 89.70 | 89.58 | **85.05** | **85.44** | **85.24** |
| Hugging Face | xlm-roberta-large | **92.10** | **93.50** | **92.80** | 84.27 | 84.29 | 84.28 |
| | electra-large-discriminator | 92.02 | 92.12 | 92.07 | 83.21 | 84.11 | 83.66 |
| | bert-large-cased | 91.23 | 92.37 | 91.80 | 84.80 | 84.19 | 84.49 |
| | bert-base-NER | 90.65 | 91.93 | 91.29 | 83.58 | 85.40 | 84.48 |
| NLTK | | 88.10 | 87.00 | 87.55 | 75.19 | 78.02 | 76.58 |
| Flair | | 91.03 | 90.53 | 90.72 | 81.67 | 79.36 | 80.29 |
| Stanford CoreNLP | | 88.12 | 87.20 | 87.66 | 74.50 | 77.81 | 76.00 |
| Apache OpenNLP | | 91.57 | 74.84 | 81.64 | 83.16 | 68.08 | 74.85 |

classe (type d'entité nommée ou classe autre) ainsi que sa position dans le texte. Chaque framework étant accompagné de sa propre représentation de données une conversion de ce format pivot vers chacun des formats spécifiques à chaque framework est effectué.

— **Recherche des valeurs d'hyper-paramètres :** Les valeurs par défaut, celles des modèles pré-entraînés librement disponibles ainsi qu'une grille ont été considérées. Les valeurs des modèles pré-entraînés se sont avérées être les meilleures excepté pour les modèles Hugging Face sur les données WikiANN. Les valeurs retenues sont learning_rate = 2e-5, batch_size = 16, number_of_epochs = 20 et weight_decay = 0.01.

— **Évaluation :** Les métriques discutées précédemment ont été utilisées en comparant les labels prédits avec la base de référence. Le protocole exact basé sur l'évaluation CoNLL-2003 a été retenu, c'est-à-dire que les contours et la classe de l'entité nommée doivent correspondre à la base de référence (cf. la section 4.2). Les résultats sont donnés dans Table 3.

Pour le corpus CoNLL-2003, nous constatons que les modèles issus du framework Hugging Face, essentiellement basés sur des transformers, surpassent nettement les autres méthodes. Contrairement à ce qui a été observé dans l'étude [39], nous avons remarqué que OpenNLP surpasse Stanford CoreNLP.

Sur un jeu de données bien équilibré comme wikiANN où les proportions des entités nommées sont égales, nous constatons aussi que les architectures basées sur des transformers donnent toujours les meilleurs résultats, avec une petite différence en faveur du modèle transformer de spaCy, qui donne des résultats légèrement meilleurs que ceux d'Hugging Face. Il convient également de préciser que les hyper-paramètres utilisés pour entraîner un modèle sur un jeu de données spécifique ne donnent pas nécessairement les meilleurs résultats sur un autre jeu de données, même si les deux semblent proches et partagent les mêmes classes. Par exemple pour le modèle *lm-roberta-large*, nous avons remarqué que l'optimisation de certains hyper-paramètres a un impact important sur les performances de ce modèle. Cette optimisation nous a permis de gagner environ 4 points en F1-score.

# 6 Conclusion et perspectives

Cet article présente les principales méthodes souvent utilisées pour la reconnaissance d'entités nommées. Nous avons évalué les frameworks les plus populaires permettant de faire de la reconnaissance d'entités nommées sur deux jeux de données. Les résultats de nos expériences ont montré que les méthodes basées sur des transformers sont significativement meilleures que les autres méthodes basées sur les réseaux de neurones ou les méthodes probabilistes. Cependant, les performances dépendent des valeurs des hyper-paramètres. Les valeurs d'hyper-paramètres des modèles pré-entraînés donnent de bons résultats même sur un jeu de données (WikiANN) sur lequel ils n'ont pas été réglés. Une étude plus complète des valeurs des hyper-paramètres et une évaluation sur d'autres jeux de données constituent une piste de futurs travaux.

# Références

[1] Bogdan Babych and Anthony Hartley. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, 2003.

[2] Diego Mollá, Menno Van Zaanen, and Daniel Smith. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58, 2006.

[3] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274, 2009.

[4] Ralph Grishman and Beth M Sundheim. Message understanding conference-6 : A brief history. In *CO-*

*LING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*, 1996.

[5] Lisa F Rau. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society, 1991.

[6] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26, 2007.

[7] Ronan Collobert. Deep learning for efficient discriminative parsing. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 224–232. JMLR Workshop and Conference Proceedings, 2011.

[8] Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th conference on natural language processing, Erlangen, Germany*, pages 8–11, 2019.

[9] Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17 :103–116, 2011.

[10] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition : fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5) :482–489, 2013.

[11] Xiaole Li, Tianyu Wang, Yadan Pang, Jin Han, and Jin Shi. Review of research on named entity recognition. In *Advances in Artificial Intelligence and Security : 8th International Conference on Artificial Intelligence and Security, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part II*, pages 256–267. Springer, 2022.

[12] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[13] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition : Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6) :1088–1098, 2013. Special Section : Social Media Environments.

[14] Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, and Damien Nouvel. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Revue TAL*, 52(1) :69–96, 2011.

[15] Bmc Bioinformatics, Daniel Hanisch, Katrin Fundel, Heinz theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer : rule-based protein and gene entity recognition, 2005.

[16] Yusuke Shinyama and Satoshi Sekine. Named entity discovery using comparable news articles. In *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, pages 848–853, Geneva, Switzerland, aug 23–aug 27 2004. COLING.

[17] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[18] Sudha Morwal, Nusrat Jahan, and Deepti Chopra. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1, 2012.

[19] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Mach. Learn.*, 34(1–3) :211–231, feb 1999.

[20] Andrew Eliot Borthwick. *A maximum entropy approach to named entity recognition*. New York University, 1999.

[21] Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. In *COLING 2002 : The 19th International Conference on Computational Linguistics*, 2002.

[22] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 107–110, Geneva, Switzerland, August 28th and 29th 2004. COLING.

[23] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14) :i37–i48, 2017.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[26] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*, 2016.

[27] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[28] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding, 2018. cite arxiv :1810.04805Comment : 13 pages.

[31] Steven Bird. Nltk : the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[32] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, pages 55–60, 2014.

[33] Ted Kwartler. *The OpenNLP Project*, pages 237–269. 05 2017.

[34] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair : An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.

[35] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for ner. *arXiv preprint arXiv :1902.00193*, 2019.

[36] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*, 1996.

[37] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[38] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[39] Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. A replicable comparison study of ner software : Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343. IEEE, 2019.

# La méthode de modélisation thématique CFMf basée sur le clustering neuronal avec maximisation des traits : Comparaison avec LDA sur des études scientifiques

J.-C. Lamirel[1], F. Lareau[2], C. Malaterre[3]

[1] Université de Strasbourg, SYNALP-LORIA
[2] Université du Québec à Montréal, Computer Science Dept.
[3] Université du Québec à Montréal, Dept. of Philosophy & CIRST

lamirel@loria.fr, lareau.francis@courrier.uqam.ca, malaterre.christophe@uqam.ca

## Résumé

*L'amélioration des méthodes de modélisation thématique reste une préoccupation majeure pour l'analyse non supervisée des données textuelles. Nous proposons ici une approche de modélisation thématique basée sur le clustering neuronal et la maximisation des traits. Nous comparons ses performances à celles de LDA en appliquant les deux méthodes à un large corpus de référence d'articles de philosophie des sciences en texte intégral. Les résultats montrent des améliorations très significatives des mesures de performance quantitatives clés telles que la cohérence, ainsi que des résultats qualitatifs.*

## Mots-clés

*Modélisation thématique, apprentissage non supervisé, LDA, clustering, maximisation des traits.*

## Abstract

*The improvement of topic modeling methods remains a major concern for unsupervised analysis of textual data. We propose here a topic modeling approach based on neural clustering and feature maximization. We compare its performance to that of LDA by applying both methods to a large reference corpus of full-text philosophy of science articles. The results show very significant improvements in key quantitative performance measures such as coherence, as well as qualitative results.*

## Keywords

*Topic modeling, unsupervised learning, LDA, clustering, feature maximization.*

## 1 Introduction

En tant que résultats privilégiés de la recherche scientifique, les articles et leur contenu offrent des perspectives uniques pour comprendre la science. L'exploration par des méthodes informatiques du contenu textuel non structuré de ces articles peut être une solution efficace pour étudier des corpus de textes scientifiques trop volumineux pour une lecture manuelle. À cet égard, la modélisation thématique peut être utilisée pour inférer de manière fiable le contenu sémantique des publications, ce qui permet d'identifier les thèmes de recherche dominants dans des disciplines scientifiques spécifiques, y compris leur évolution dans le temps (par exemple, [1]; [2]; [3]). L'une de ces approches bien établies est le modèle Latent Dirichlet Allocation (LDA) [4] et ses variantes [5]. Des méthodes alternatives ont été récemment conçues, notamment certaines qui utilisent une combinaison de clustering neuronal et de maximisation des traits au moyen du contraste (« CFMc » pour neural Clustering and Feature Maximization with Contrast) [6]. Ces dernières ont montré des améliorations qualitatives prometteuses par rapport à LDA. Par contre, des tests approfondis sur un corpus de référence d'articles de philosophie des sciences en texte intégral (N=16917) qui avait déjà été analysé en détail au moyen d'un modèle thématique LDA [7] ont révélé des limites en termes d'interprétabilité des thèmes. Ces limites nous ont amenés à concevoir une nouvelle approche toujours basée sur le clustering neuronal avec maximisation des traits, mais qui s'appuie désormais sur la mesure F1 (« CFMf » pour neural Clustering and Feature Maximization with F1-measure). Dans la présente recherche en cours, nous décrivons l'approche CFMf. Pour évaluer sa performance par rapport à CFMc et LDA, nous appliquons ces méthodes au corpus de référence et comparons les modèles. D'abord quantitativement en termes de cohérence [8] pour plusieurs valeurs $w$ du nombre de mots principaux et $k$ du nombre de thèmes. Ensuite, nous évaluons également la performance qualitative en examinant l'interprétabilité des thèmes à $k = 25$ du point de vue de la connaissance experte. Les résultats montrent des améliorations très significatives apportées par CFMf, à la fois en termes de mesures de performance et d'évaluations qualitatives.

## 2 Méthodes

Alors que les approches CFMc avaient conduit à des résultats prometteurs lorsque testées sur un corpus d'articles de recherche chinois dans le domaine des « sciences de la science », nos expériences préliminaires sur le jeu de don-

nées plus complexe d'articles de philosophie des sciences - précédemment analysé avec la LDA [7] - ont mis en évidence trois limites. Premièrement, la représentation binaire des mots dans les documents, telle que mise en œuvre dans CFMc, semblait trop restrictive, encourageant ainsi l'utilisation des fréquences de mots. Deuxièmement, le contraste semblait ne pas être applicable dans les corpus dont les documents ne contenaient pas de thèmes discriminants clairement définis, d'où la nécessité de modéliser les documents comme intégrants plusieurs thèmes. Troisièmement, aucune comparaison quantitative avec la LDA n'avait été réalisée. La nouvelle approche que nous proposons ci-après (CFMf) répond à ces limitations. Elle s'appuie toujours sur le clustering neuronal et la maximisation des traits FMax, mais utilise désormais la mesure F1 au lieu du contraste. L'approche s'appuie également sur la représentation des documents en sac de mots (« BoW » pour Bag of Words), mais en exploitant l'information fréquentielle des mots. Les documents sont partitionnés en utilisant le clustering neuronal GNG [10]. Il s'agit d'une approche de type « winner-take-most » basée sur l'apprentissage Hebbien, moins sujette aux problèmes connus du clustering concernant la sensibilité aux valeurs aberrantes et à l'initialisation (comme c'est notamment le cas pour des méthodes classiques telles que k-means). Dans une étape ultérieure, les mots-clés représentatifs des clusters qui représenteront les thèmes sont extraits des documents associés à chaque cluster à l'aide de l'approche FMax (associée à la mesure F1), qui est un schéma générique de comparaison de données pouvant être utilisé comme une alternative aux métriques usuelles telles que le chi2, la métrique euclidienne ou la similarité cosinus lorsqu'il s'agit de traiter des données éparses et fortement multidimensionnelles, comme c'est le cas des données textuelles quand elles sont représentées en mode BoW. L'approche FMax offre des capacités de sélection et de pondération de variables sans nécessiter d'utiliser de paramètre [9] et s'est avérée très utile dans de nombreuses tâches d'exploration de données, y compris pour le plongement de mots et le plongement de graphes [11]. Dans le cas que nous traitons, les variables (c.àd. les traits) sont des mots et les données sont les documents associés à chaque cluster. FMax est basée ici sur l'estimation de la mesure F1 qui représente la moyenne harmonique (1) du rappel de trait, qui estime le pouvoir de discrimination d'un mot vis-à-vis d'un cluster ; et (2) de la prédominance de trait, qui évalue la capacité de généralisation du mot vis-à-vis de ce même cluster.

Considérons une partition $C$ qui résulte d'une méthode de clustering appliquée à un ensemble de documents $D$ représenté par un ensemble de mots $F$. Les mesures de rappel de traits, de prédominance de traits et la mesure F1 sont respectivement définies comme suit :

$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c \in C} \Sigma_{d \in c} W_d^f} \tag{1}$$

$$FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{2}$$

avec

$$F1_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \tag{3}$$

où $W_d^f$ représente le poids du mot $f$ pour le document $d$ et $F_c$ représente l'ensemble des mots présents dans l'ensemble des documents associés au cluster $c$.

Pour réduire le bruit, nous éliminons au sein d'un cluster donné les mots qui répondent à au moins une des conditions suivantes : a) une mesure F1 inférieure à la moyenne des mesures F1 de ce même mot pour tous les clusters dans lesquels le mot est présent ou b) une mesure F1 inférieure à la moyenne des mesures F1 de tous les mots de tous les documents.

Ainsi, l'ensemble $S_c$ des mots qui sont caractéristiques d'un cluster $c$ issu d'une partition $C$ est défini par :

$$S_c = \left\{ f \in F_c \mid F1_c(f) > \overline{F1}(f) \text{ et } F1_c(f) > \overline{F1}_D \right\} \tag{4}$$

avec

$$\overline{F1}(f) = \Sigma_{c' \in C} \frac{F1_{c'}(f)}{|C_{/f}|} \text{ et } \overline{F1}_D = \Sigma_{f \in F} \frac{\overline{F1}(f)}{|F|} \tag{5}$$

où $C_{/f}$ représente le sous-ensemble des clusters de $C$ dans lequel le mot $f$ est présent.

Il en résulte un profil de mesure F1 (sur un lexique réduit) pour chaque cluster, ce dernier étant alors considéré comme un thème. D'où la possibilité d'extraire les mots les plus importants sur la base de leur classement selon la mesure F1.
De plus amples détails sur les mesures ci-dessus mentionnées sont également donnés dans [9].

# 3 Protocole expérimental

Le corpus est constitué de tous les articles de recherche en texte intégral provenant de huit revues majeures de philosophie des sciences qui ont été rassemblées dans le cadre de l'étude de la philosophie des sciences [13] : le *British Journal for the Philosophy of Science*, le *European Journal for Philosophy of Science*, *Erkenntnis*, *International Studies in the Philosophy of Science*, le *Journal for General Philosophy of Science*, *Philosophy of Science*, *Studies in History and Philosophy of Science Part A* et *Synthese*. Il s'étend de 1930 à 2017 et comprend 16 917 documents. Le corpus a été nettoyé et prétraité de manière standard (les textes en langue étrangère ont été traduits mécaniquement en anglais). Seuls les noms, les verbes, les adverbes et les adjectifs ont été conservés après étiquetage POS et lemmatisation (TreeTagger package [12] avec les jeux d'étiquettes de Penn TreeBank [13]) et les mots apparaissant dans moins de 50 phrases du corpus ont été supprimés. Tous les documents ont ensuite été vectorisés, ce qui a permis d'obtenir une matrice termes-documents

avec fréquences de mots.

La matrice termes-documents a ensuite été soumise à CFMc, CFMf et LDA. Pour comparer quantitativement CFMf et CFMc, les deux méthodes ont été utilisées pour construire des modèles à $k$ = 25. Ces modèles ont ensuite été comparés en termes de cohérence calculée avec différents nombres de mots principaux (de $w$ = 5 à 100). L'objectif était ici d'évaluer quelle méthode donnait de meilleurs résultats pour des valeurs de $w$ relativement petites (étant donné que les petits ensembles de mots principaux sont généralement plus faciles à interpréter à condition qu'ils soient bien formés). Dans une étape ultérieure, des modèles de thèmes ont été construits à la fois avec CFMf et LDA pour différents nombres de thèmes allant de $k$ = 5 à 50 (par incréments de 5 de 5 à 20, et par incréments de 1 au-delà de 20). La modélisation LDA a été réalisée conformément à [1] et [4] par l'intermédiaire d'une API Python. Les modèles thématiques obtenus ont ensuite été comparés à l'aide de trois mesures permettant d'estimer la consistance ou la cohérence thématique. Tout d'abord $C_{PMI}$, (également appelée $C_{UCI}$) suivant [15] qui ont proposé d'évaluer la qualité des thèmes en termes de cohérence telle qu'elle est comprise par les lecteurs humains ; cette mesure compte la cooccurrence des mots dans une fenêtre glissante et calcule, pour chaque paire de mots, son PMI (information mutuelle ponctuelle) ; $C_{PMI}$ est la somme (ou la moyenne arithmétique selon les implémentations) des valeurs PMI. Deuxièmement, $C_{NPMI}$, tel que proposé par [14] est une version améliorée de $C_{PMI}$ utilisant l'information mutuelle ponctuelle normalisée (NPMI). Troisièmement, la mesure de cohérence souvent utilisée $C_V$ proposée par [8] et mise en œuvre dans le package populaire Gensim en Python ; $C_V$ compte les cooccurrences d'un certain nombre de mots principaux (généralement 10 à 20) dans une fenêtre glissante (généralement de taille 110) ; les cooccurrences sont utilisées pour calculer la NPMI entre les mots principaux, produisant des vecteurs pour chacun d'entre eux ; la moyenne arithmétique des similitudes cosinus entre chaque vecteur de mots principaux et la somme de tous les vecteurs de mots principaux est ensuite calculée. Les noms de thèmes de l'étude [7] ont été utilisés pour étiqueter les thèmes LDA, tandis que les thèmes CFMf ont été nommés à l'aide de leurs premiers mots et de la connaissance des experts, et les thèmes des deux modèles ont été comparés qualitativement. Pour comparer davantage les résultats des modèles LDA et CFMf, la distance Hellinger entre les 25 thèmes de chaque modèle (représentés sous forme de vecteurs de mots) a été calculée. Les thèmes d'un modèle ont ensuite été alignés sur ceux de l'autre.

# 4 Résultats

Les résultats de cohérence comparant les performances de CFMf et CFMc en fonction du nombre de mots principaux (pour un nombre donné de thèmes $k$=25) montrent que les modèles CFMf avec moins de mots principaux sont nettement plus performants que les modèles CFMc (figure 1A). Cela signifie que la mesure F1 donne des ensembles de mots principaux plus cohérents que le contraste. Compte tenu de l'objectif d'interprétabilité du sujet (basé sur un ensemble relativement restreint de mots principaux ordonnés), l'utilisation de la mesure F1 au lieu du contraste apporte une amélioration méthodologique significative, ce qui justifie l'utilisation de CFMf par rapport à CFMc dans des contextes similaires.

Lorsqu'il s'agit de comparer CFMf avec LDA, les résultats montrent que CFMf surpasse largement LDA en termes de mesures de cohérence (Fig. 1B, C, D). C'est le cas pour les trois mesures de cohérence que nous avons testées ($C_V$, $C_{NMPI}$ et $C_{PMI}$), et pour une large gamme de modèles avec différents nombres de thèmes (de $k$ = 5 à 50 thèmes). L'amélioration de la cohérence apportée par CFMf par rapport à LDA est très significative, puisqu'elle va d'environ 50 % pour $C_V$ à plus de 200 % pour $C_{PMI}$. Dans tous les cas, la cohérence augmente considérablement de 5 à 20 thèmes, puis plus lentement de 20 à 40 thèmes, et approche un plateau au-delà de 40 thèmes. Un nombre idéal semble donc se situer autour de 30-35 thèmes, en fonction des objectifs du modèle thématique. L'aspect le plus significatif des résultats est la surperformance quantitative constante du modèle CFMf sur le modèle LDA en termes de cohérence. Les résultats de l'analyse qualitative effectuée sur les 10 premiers mots des thèmes du CFMf montrent un type de couverture thématique similaire à celui de la LDA. Rappelons que, pour des raisons de commodité, cette comparaison a été effectuée pour $k$ = 25 thèmes, étant donné qu'un modèle LDA antérieur a été examiné en détail pour $k$ = 25 [7]. A l'époque, $k$ = 25 avait été choisi pour des raisons pragmatiques, afin d'avoir un modèle thématique avec un nombre de thèmes relativement faible. Or, comme nous venons de le voir, des valeurs de $k$ plus élevées montrent des mesures de cohérence plus élevées, ce qui implique que $k$ = 25 n'est pas optimal à cet égard. Néanmoins, les 25 thèmes résultant du modèle CFMf semblent facilement interprétables sur la base des 10 premiers mots et de la connaissance du domaine par les experts.

Les distances Hellinger entre ces thèmes et ceux du modèle LDA montrent une assez bonne correspondance des sujets, mais l'appariement est loin d'être parfait, ce qui montre que les thèmes des deux modèles ont encore des différences notables. Nous avons examiné et comparé manuellement les 10 premiers mots des thèmes des deux modèles. Des exemples sont présentés dans la table 1 : certains sujets ont des mots du top-10 très similaires, tandis que d'autres semblent avoir été en quelque sorte fusionnés ou divisés.

# 5 Discussion

Cette première expérience de comparaison quantitative et qualitative donne des résultats préliminaires intéressants. La méthode CFMf, qui est une extension de la méthode CFMc utilisant désormais la mesure F1 plutôt que le contraste, semble bien plus efficace que la méthode LDA au regard de trois mesures de performance ainsi qu'en termes d'interprétation (au moins en ce qui concerne les

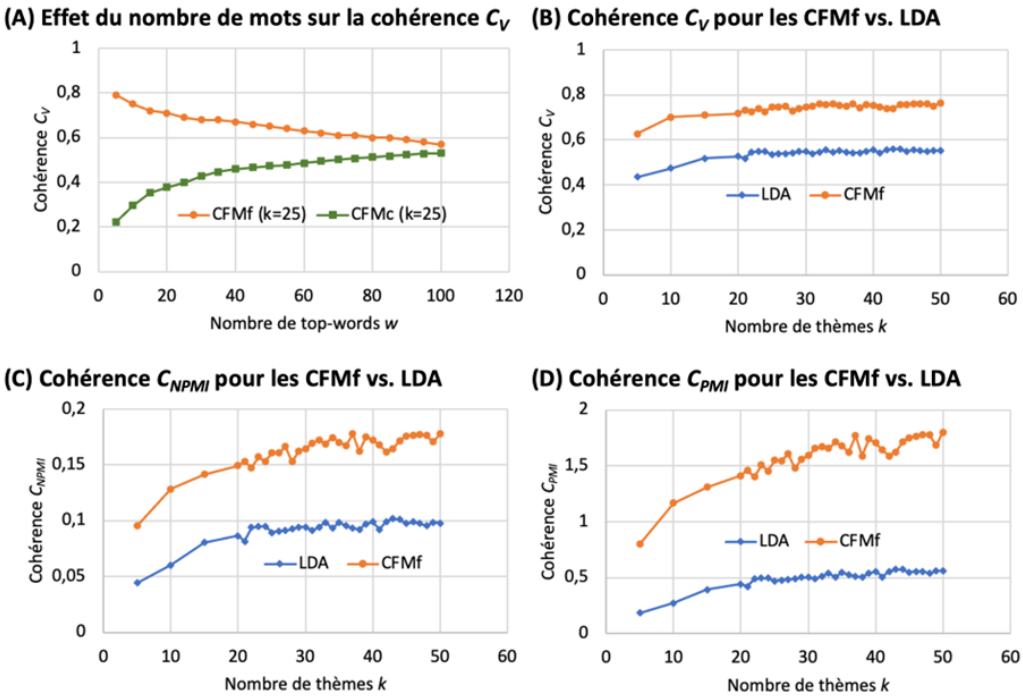



FIGURE 1 – Comparaisons entre modèles thématiques. (A) Cohérence $C_V$ pour CFMf et CFMc en fonction du nombre de top-words $w$ (pour $k$ = 25 thèmes). (B, C, D) Cohérences $C_V$, $C_{NPMI}$ et $C_{PMI}$ pour les modèles CFMf et LDA en fonction du nombre de thèmes $k$ (pour $w$ = 10 top-words).

| *CFMf* | *Top-10 words* | *LDA topics* | *Top-10 words* |
|---|---|---|---|
| Knowledge (11) | belief; epistemic; justified; justification; epistemically; doxastic; reliable; testimony; proposition; agent | Knowledge (21) | belief; knowledge; epistemic; believe; know; case; evidence; reason; justification; true |
| Neurosciences (20) | neural; brain; processing; cognitive; input; computational; neuron; cognition; mechanism; visual | Neurosciences (13) | system; information; process; cognitive; level; mechanism; state; representation; structure; function |
| Causation (7) | causation; causal; cause; counterfactual; intervention; causally; woodward; probabilistic; event; counterfactuals | Causation (19) | causal; cause; event; effect; causation; condition; case; variable; time; occur |
| | | Explanation (16) | model; explanation; explain; account; explanatory; phenomenon; use; case; system; provide |
| Quantum mechanics (18) | quantum; measurement; particle; mechanic; observables; spin; operator; wave; probability; bell | Quantum (14) | time; state; space; quantum; system; theory; particle; physical; field; point |
| Relativity (23) | spacetime; relativity; inertial; metric; einstein; velocity; motion; coordinate; frame; tensor | | |

TABLE 1 – Comparaison des top-10 mots pour un échantillon de thèmes (CFMf et LDA).

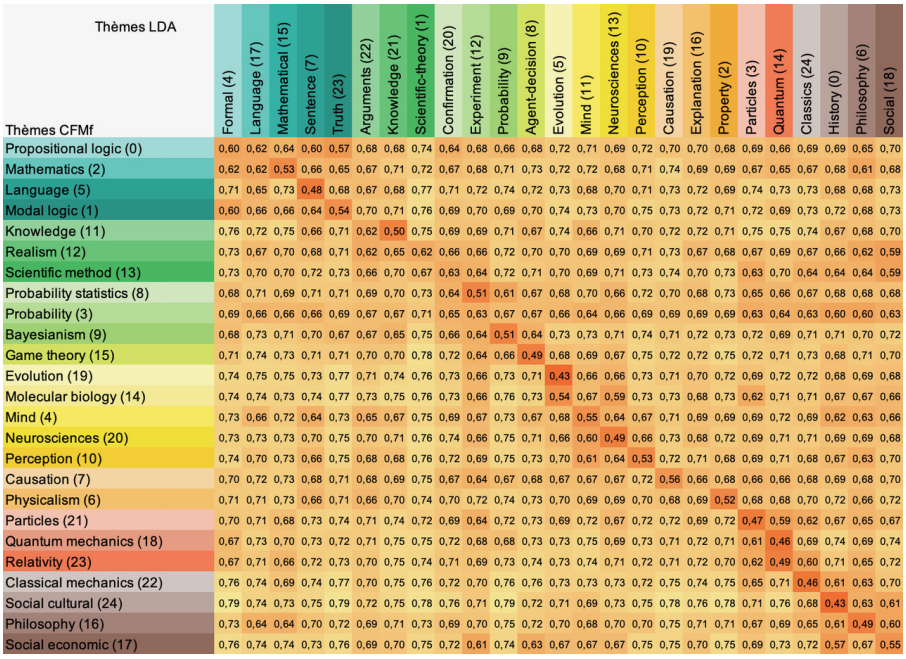



FIGURE 2 – Distances Hellinger entre thèmes issus du modèle LDA (rangée du haut) et ceux du modèle CFMf (colonne de gauche). Distances mesurées entre thèmes représentés comme des vecteurs de probabilités sur le lexique du corpus.

10 premiers mots du modèle $k = 25$). Il serait intéressant d'évaluer ces méthodes avec d'autres mesures de performance, notamment la perplexité [16], UMass [17], la KL-divergence symétrique [18], la densité [19] ou les statistiques bayésiennes [1]. Les résultats peuvent également être comparés à d'autres implémentations de LDA (par exemple, les approches LDA avec Bayes variationnel [20] au lieu de l'échantillonnage de Gibbs), et à des modèles alternatifs (par exemple, Structural Topic Models (STM) [21]). Le comportement de CFMf sur les documents (comparé à LDA) devrait également être exploré, afin d'évaluer les changements dans les distributions de probabilité des thèmes. CFMf pourrait par ailleurs être testé sur d'autres corpus : il serait intéressant d'étudier si certains corpus se prêtent mieux à une méthode qu'à l'autre. Bien que CFMf suppose que le corpus se prête au clustering et que les clusters identifiés correspondent à des thèmes, elle ne requiert pas initialement l'hypothèse (faite dans LDA) que les documents sont des distributions de thèmes. En outre, en dehors du nombre de thèmes, CFMf ne nécessite aucune paramétrisation. Il serait également intéressant de vérifier la robustesse des résultats en répétant la même méthode sur le même corpus avec les mêmes paramètres. Nous pensons que CFMf est plus robuste que LDA, car cette méthode est moins sensible à l'ensemencement aléatoire et implique une coopération entre les prototypes pour opérer des regroupements thématiques.

# 6 Conclusion

Nous avons présenté une nouvelle approche prometteuse pour la construction d'un modèle thématique basé sur une combinaison de clustering neuronal, de maximisation des traits et de mesure F1 : CFMf. Nous avons également exposé les résultats de nos expériences comparatives. Cependant, malgré le potentiel évident de CFMf pour la réalisation d'études à grande échelle, telles que les études scientifiques présentées ici, plusieurs adaptations et expériences supplémentaires sont encore possibles. Nous prévoyons tout d'abord d'étendre notre comparaison à des méthodes connues pour être des alternatives efficaces à LDA, comme la méthode STM. Nous examinerons également si LDA peut être amélioré en utilisant des composants-clés de la présente méthode, notamment en exploitant la sélection de variables et l'approche FMax pour pondérer les termes dans les thèmes de LDA. Nous examinerons aussi l'effet de la réduction du lexique sur les deux types de méthodes, tout en comparant notre modèle avec des approches de modélisation de thèmes basées sur différentes représentations du corpus (binaire, fréquence, enchâssements). Nous envisageons également d'étudier la faisabilité du choix d'un nombre optimal $k$ de thèmes, notamment avec une approche similaire à celle de [22]. Des stratégies spécifiques combinant la mesure F1 et le classement par contraste pourront être explorées afin d'optimiser davantage encore les descriptions de thèmes.

# Remerciements

J.-C.L. remercie l'ANRT pour son soutien financier. F.L. remercie le Fonds de recherche du Québec Société et culture (FRQSC-276470) et la Chaire de recherche du Canada en philosophie des sciences de la vie de l'UQAM. C.M. remercie le Conseil de recherches en sciences humaines du Canada (subvention 430-2018-00899) et le programme des Chaires de recherche du Canada (CRC-950-230795) pour

leur soutien financier.

## Références

[1] T.L. Griffiths et M. Steyvers, "Finding scientific topics.," *Proceedings of the National Academy of Sciences*. vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[2] E.M. Talley, D. Newman, D. Mimno, et al., "Database of NIH grants using machine-learned categories and graphical clustering.," *Nature methods*. vol. 8, no. 6, pp. 443–444, 2011.

[3] K. Börner, F.N. Silva, et S. Milojević, "Visualizing big science projects.," *Nature Reviews Physics*. vol. 3, no. 11, pp. 753–761, 2021.

[4] D.M. Blei, A.Y. Ng, et M.I. Jordan, "Latent dirichlet allocation.," *Journal of Machine Learning Research*. vol. 3, no. Jan, pp. 993–1022, 2003.

[5] J.L. Boyd-Graber, Y. Hu, et D. Mimno, *Applications of topic models*. now Publishers Incorporated, 2017.

[6] J.-C. Lamirel, Y. Chen, P. Cuxac, S. Al Shehabi, N. Dugué, et Z. Liu, "An overview of the history of Science of Science in China based on the use of bibliographic and citation data : a new method of analysis based on clustering with feature maximization and contrast graphs.," *Scientometrics*. vol. 125, no. 3, pp. 2971–2999, 2020.

[7] C. Malaterre et F. Lareau, "The early days of contemporary philosophy of science : novel insights from machine translation and topic-modeling of non-parallel multilingual corpora.," *Synthese*. vol. 200, no. 3, p. 242, 2022.

[8] M. Röder, A. Both, et A. Hinneburg, "Exploring the Space of Topic Coherence Measures.," In : *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. pp. 399–408. ACM Press, Shanghai, China (2015).

[9] J.-C. Lamirel, P. Cuxac, A.S. Chivukula, et K. Hajlaoui, "Optimizing text classification through efficient feature selection based on quality metric.," *Journal of Intelligent Information Systems*. vol. 45, no. 3, pp. 379–396, 2015.

[10] B. Fritzke, "A growing neural gas network learns topologies.," *Advances in neural information processing systems*. vol. 7, p. 1994.

[11] T. Prouteau, V. Connes, N. Dugué, et al., "SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin !," In : P.H. Abreu, P.P. Rodrigues, A. Fernández, and J. Gama, Eds. Advances in Intelligent Data Analysis XIX. pp. 325–337. *Springer International Publishing*, Cham (2021).

[12] H. Schmid, "Probabilistic part-of-speech tagging using decision trees.," In : *Proceedings of International Conference on New Methods in Language Processing*. pp. 44–49. , Manchester (1994).

[13] M.P. Marcus, M.A. Marcinkiewicz, et B. Santorini, "Building a Large Annotated Corpus of English : The Penn Treebank.," *Computational Linguistics*. vol. 19, no. 2, pp. 313–330, 1993.

[14] N. Aletras et M. Stevenson, "Evaluating topic coherence using distributional semantics.," In : *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. pp. 13–22 (2013).

[15] D. Newman, J. Han Lau, K. Grieser, et T. Baldwin, "Automatic evaluation of topic coherence.," In : *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 100–108 (2010).

[16] H.M. Wallach, I. Murray, R. Salakhutdinov, et D. Mimno, "Evaluation methods for topic models.," In : *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. pp. 1–8. ACM Press, Montreal, Quebec, Canada (2009).

[17] D. Mimno, H. Wallach, E. Talley, M. Leenders, et A. McCallum, "Optimizing semantic coherence in topic models.," In : *Proceedings of the 2011 conference on empirical methods in natural language processing*. pp. 262–272 (2011).

[18] R. Arun, V. Suresh, C.E. Veni Madhavan, et N. Murthy, "On finding the natural number of topics with latent dirichlet allocation : Some observations.," In : *Pacific-Asia conference on knowledge discovery and data mining*. pp. 391–402. Springer (2010).

[19] J. Cao, T. Xia, J. Li, Y. Zhang, et S. Tang, "A density-based method for adaptive LDA model selection.," *Neurocomputing*. vol. 72, no. 7–9, pp. 1775–1781, 2009.

[20] M. Hoffman, F. Bach, et D. Blei, "Online learning for latent dirichlet allocation.," *advances in neural information processing systems*. vol. 23, p. 2010.

[21] M.E. Roberts, B.M. Stewart, D. Tingley, et E.M. Airoldi, "The structural topic model and applied social science.," In : *Advances in neural information processing systems workshop on topic models : computation, application, and evaluation*. pp. 1–20. Harrahs and Harveys, Lake Tahoe (2013).

[22] N. Dugué, J.-C. Lamirel, et Y. Chen, "Evaluating clustering quality using features salience : a promising approach.," *Neural Computing and Applications*. p. 2021.

# Méthode de classification divisive sur intervalles d'estimation duale des quantiles de coûts spécifiques et de marges brutes

D. Desbois[1]

[1] INRAE-AgroParisTech, Paris Saclay Applied Economics

dominique.desbois@inrae.fr

## Résumé

*Cette communication utilise la classification des données symboliques pour explorer les similitudes entre distributions d'estimations quantiles conditionnelles, en l'appliquant au problème de l'allocation des coûts spécifiques et des marges brutes en agriculture. Après avoir rappelé le cadre conceptuel de l'estimation des coûts et des marges en production agricole, la première partie présente le modèle empirique, l'approche de régression quantile et la technique de classification des données d'intervalle utilisée. La seconde partie présente l'analyse comparative entre les régions de douze États membres de l'UE des résultats issus de la classification hiérarchique divisive des estimations par intervalle, appliquée au blé.*

## Mots-clés

*données d'intervalle, classification divisive, coûts spécifiques, marges brutes, régions européennes.*

## Abstract

*This paper uses symbolic data clustering to explore the similarities between distributions of conditional quantile estimates, applying it to the problem of allocating specific costs and gross margins in agriculture. After recalling the conceptual framework of cost and margin estimation in agricultural production, the first part presents the empirical model, the quantile regression approach and the divisive clustering technique on interval data used. The second part presents the comparative analysis between the regions of twelve EU Member States of the results of the hierarchical divisive classification of interval estimates, applied to wheat.*

## Keywords

*interval data, divisive clustering, specific costs, gross margins, European regions.*

## 1 Introduction

L'intégration de l'agriculture dans les 28 États membres résultant de l'élargissement de l'Union européenne (UE) a suscité des besoins récurrents d'estimation des coûts de production des principaux produits agricoles, tout au long des réformes de la politique agricole commune (PAC), sur les marchés concurrentiels comme réglementés. L'analyse des coûts de production agricole est un outil d'analyse des marges des agriculteurs : elle permet d'évaluer la compétitivité prix des exploitations agricoles, l'un des éléments majeurs du développement et de la durabilité des chaînes alimentaires dans les régions européennes. Pour répondre aux besoins de simulations et d'analyses d'impact pour les différentes organisations communes de marchés, nous devons fournir des informations sur l'ensemble de la répartition des coûts de production afin d'évaluer les options de politique agricole publique. En se basant sur le constat de l'asymétrie et de l'hétérogénéité au sein de la distribution empirique des intrants agricoles, nous avons proposé une méthodologie adaptée à l'estimation de la distribution empirique des coûts de production spécifiques des principaux produits agricoles dans un contexte européen où les exploitations agricoles restent principalement multiproductives [6]. À partir de cette approche, nous présentons le modèle empirique d'estimation des coûts de production spécifiques, inspirée d'une approche micro-économétrique de répartition des coûts pour construire une matrice entrées-sorties au niveau national [8]. Puis, nous rappelons la méthodologie d'estimation permettant de prendre en compte l'hétérogénéité des exploitations agricoles, selon l'approche du quantile conditionnel proposée par [12]. Ensuite, pour explorer les distributions empiriques des intervalles d'estimation de quantiles conditionnels, nous présentons la procédure de classification utilisée [10] dans le cadre conceptuel de l'analyse symbolique de données [1]. Nous introduisons alors le graphique des résultats de la procédure de classification appliquée aux intervalles d'estimation des quantiles conditionnels. Enfin, nous concluons sur la pertinence de cette approche appliquée à l'estimation du coût des fertilisants pour les productions végétales.

## 2 Cadre conceptuel et aspects méthodologiques

Nous présentons d'abord la méthodologie d'estimation des coûts spécifiques. Puis, nous introduisons l'outil de classification des intervalles d'estimation dans le formalisme de l'analyse symbolique de données.

## 2.1 Le modèle d'estimation des coûts spécifiques de production

Inspiré de [8], l'affectation de la somme $x_i$ des coûts des intrants pour l'exploitation agricole est réalisée par décomposition linéaire selon les produits bruts $Y_i^j$ de l'exploitation agricole $i$ pour chaque production $j$, où $u_i$ est un vecteur aléatoire d'espérance nulle :

$$x_i = \sum_{j=1}^{p} \beta_j Y_i^j + u_i \quad (1)$$

Comme [2], nous supposons que le processus générateur de données est un modèle linéaire à hétéroscédasticité multiplicative caractérisé par :

$$x = Y'\beta + u \text{ avec } u = Y'\alpha \times \varepsilon \text{ et } Y'\alpha > 0 \quad (2)$$

où $\varepsilon \sim$ iid$\{0;\sigma\}$ est un vecteur aléatoire identique et indépendant à moyenne nulle et variance constante $\sigma^2$. Sous cette hypothèse, $\mu_q(x|Y,\beta,\alpha)$, le $q^e$ quantile conditionnel du coût de production $x$, conditionné par $Y$ et les paramètres, $\alpha$ et $\beta$, se déduit analytiquement comme suit :

$$\mu_q(x \mid Y,\beta,\alpha) = Y'[\beta + \alpha \times F_\varepsilon^{-1}(q)] = Y'\gamma \quad (3)$$

où $F_\varepsilon^{-1}$ est la distribution cumulée des erreurs. Le coefficient technique du $q^e$ quantile de coûts spécifiques pour le $j^e$ produit est défini par le $j^e$ composant du vecteur de pente :

$$\beta^j(q) = [\beta + \alpha \times F_\varepsilon^{-1}(q)]^j \quad (4)$$

Au moins deux types de modèle peuvent être dérivés de cette spécification [9] :

— $x = Y'\beta + u$ avec $u = K\varepsilon$, à résidus homoscédastiques ( $V(\varepsilon \mid Y) = \sigma^2$), dénommé *modèle à translation simple*, *i.e.* un modèle linéaire à pentes homogènes ; puisque $Y'\alpha = K$ est constant, les quantiles conditionnels $\mu_q(x \mid Y,\beta,\alpha) = Y'\beta + K \times F_\varepsilon^{-1}(q)$ ont tous la même pente, mais diffèrent seulement d'un écart constant, croissant à mesure que l'ordre $q$ du quantile augmente ;

— $x = Y'\beta + (Y'\alpha)\varepsilon$ avec $Y'\alpha > 0$ à résidus hétéroscédastiques, dénommé *modèle à changement d'échelle*, *i.e.* le modèle linéaire de quantiles conditionnels à pentes hétérogènes.

Suivant le modèle d'estimation des quantiles conditionnels pondérés par [13], la pondération $\Omega_I$ des observations, définie par $\{\omega_i; i = 1,\ldots,n\}$, est introduite dans la fonction de perte du problème de minimisation de la régression quantile comme suit :

$$\begin{aligned} &\Sigma_{(i;x_i \geq \beta y_i)}[\omega_i q \|x_i - \beta y_i\|] \\ &+ \\ &\Sigma_{(i;x_i < \beta y_i)}[\omega_i (1-q) \|x_i - \beta y_i\|] \end{aligned} \quad (5)$$

conduisant à l'estimation des paramètres du modèle (2) comme solution optimale du problème de minimisation de la fonction de perte (5), soit :

$$\begin{aligned} \widehat{\beta_{\omega(q)}} = \\ argmin_{(\beta \in R^p)}\{\Sigma_{(i;x_i \geq y_i'\beta)}[\omega_i q \|x_i - y_i'\beta\|] \\ + \\ \Sigma_{(i;x_i < y_i'\beta)}[\omega_i (1-q) \|x_i - y_i'\beta\|]\} \end{aligned} \quad (6)$$

Les estimations pondérées des quantiles conditionnelles sont fournies par la procédure QUANTREG du logiciel $SAS$, version 9.2.

## 2.2 L'estimation duale, complète ou partielle

Le $q^e$ quantile conditionnel possède la propriété d'équivariance, spécifique aux transformations monotones, impliquant les deux règles conditionnelles suivantes :

— si $\lambda \in [0;\infty]$ alors

$$\mu_q(\lambda \times X + C \mid Y) = C + \lambda \times \mu_q(X \mid Y) \quad (7)$$

— si $\lambda \in [-\infty;0]$ alors

$$\mu_q(\lambda \times X + C \mid Y) = C + \lambda \times \mu_{1-q}(X \mid Y) \quad (8)$$

Par re-paramétrisation en $X$ de $M = Y - X$, la seconde règle permet de déduire l'estimation unitaire de marge brute à partir de l'estimation unitaire de coûts spécifique, suivant la séquence de transformations ci-après :

$$\mu_q(M \mid Y) = \mu_q(Y - X \mid Y) = 1 - \mu_{(1-q)}(X \mid Y) \quad (9)$$

L'estimation duale $\mu_q^{mb} = \mu_q(\widehat{M \mid Y})$ correspond à l'estimation $\mu_{(1-q)}^{cs} = \mu_{(1-q)}(\widehat{X \mid Y})$.

Au terme de ce processus d'estimation, les distributions de paramètres sont *complètement estimées* si l'ensemble de leurs différentes estimations obtenues pour les $p$ différents paramètres quantiles peuvent être considérées sur la base de leur variabilité comme significativement non-nulles. Dans le cas contraire, les distributions de paramètres sont *partiellement estimées*.

## 2.3 Analyse factorielle des distributions empiriques duales

Soit $\Delta = \{\delta_1,\ldots,\delta_i,\ldots,\delta_n\}$, l'ensemble des distributions empiriques de coûts spécifiques et de marges brutes, décrites par un ensemble de $2p$ estimations quantiles conditionnelles : $\hat{\Gamma} = \{\widehat{\mu_1^{cs}},\ldots,\widehat{\mu_j^{cs}},\ldots,\widehat{\mu_p^{cs}},\widehat{\mu_1^{mb}},\ldots,\widehat{\mu_j^{mb}}j,\ldots,\widehat{\mu_p^{mb}}\}$. L'analyse factorielle des distributions empiriques est conduite par analyse en composantes principales normées (ACPn).En raison du centrage et de la standardisation des estimations, l'analyse montre que l'ACPn du tableau complet $\hat{\Gamma}$ des quantiles estimés (coûts spécifiques et marges brutes) et l'ACPn d'un sous-tableau (soit celui des coûts spécifiques, soit celui des marges brutes) sont structurellement équivalentes, à une constante signée près dans la définition des composantes principales. Cependant, l'utilisation conjointe des coûts et des marges permet d'enrichir conceptuellement l'interprétation.

Certains quantiles conditionnels des régions insuffisamment représentées n'étant pas estimables ou non significatifs, une affectation au plus proche barycentre, selon une norme quadratique des écarts, permet de décider de l'appartenance des régions partiellement estimées aux classes de la partition *P*, retenue comme référentiel typologique. La procédure d'imputation des estimations quantiles non significatives pour les régions partiellement estimées s'apparente aux méthodes de hot-deck métrique utilisées pour le traitement de la non-réponse.
Pour visualiser les distributions de coûts et de marges du référentiel typologique retenu (partition $P$), nous utilisons les intervalles d'estimation $[Inf; Sup]$, selon l'extension de l'analyse en composantes principales normalisée d'intervalles (ACPni) proposée par [3]. La localisation d'hyper-rectangles y est construite à partir de la projection des intervalles de confiance, arêtes des hyper-rectangles, et renseigne sur les différences significatives de niveau et de forme.

## 2.4 Classification par intervalles des distributions de coûts spécifiques

Pour un produit donné tel que le blé, le coûts spécifique ou la marge brute ($j_0$) et une région européenne ($l$), l'intervalle d'estimation des coefficients techniques de coût spécifique ou de marge brute

$$z_l^q = [Inf\{\widehat{\gamma_l^{j_0}(q)}\}; Sup\{\widehat{\gamma_l^{j_0}(q)}\}] \tag{10}$$

est obtenu par bootstrap marginal en chaînes de Markov [11]. Objets symboliques, les $L$ distributions régionales $\Omega = \{\omega_1, \ldots, \omega_l, \ldots, \omega_L\}$,sont décrites par un ensemble de $Q = 5$ descripteurs qui sont les intervalles d'estimation des coefficients techniques pour les quantiles conditionnels $Z = \{z_1^{cs}, \ldots, z_q^{cs}, \ldots, z_Q^{cs}, z_1^{mb}, \ldots, z_q^{mb}, \ldots, z_Q^{mb}\}$.
Le choix d'un petit nombre de descripteurs, soit $Q = 5$, $\{z_{0,1}^{cs}, z_{0,25}^{cs}, z_{0,5}^{cs}, z_{0.75}^{cs}, z_{0,9}^{cs}, z_{0,1}^{mb}, z_{0,25}^{mb}, z_{0,5}^{mb}, z_{0,75}^{mb}, z_{0,9}^{mb}\}$ a été fait pour des raisons de comparabilité avec des approches graphiques plus classiques [5]. Cependant, si les objectifs de l'analyse l'exigeaient, il pourrait être étendu sans inconvénient aux ensembles de descripteurs de cardinalité supérieure : déciles ($Q = 9$), voire centiles ($Q = 99$).
Les dissimilarités locales entre la région $l$ et la région $l'$, associées aux intervalles d'estimation des coefficients techniques pour le $q^e$ quantile conditionnel, sont calculées selon la distance euclidienne :

$$\begin{aligned} &\delta_M^2(z_l^q, z_{l'}^q) = \\ &(Inf\{\widehat{\gamma_l^{j_0}(q)}\} - Inf\{\widehat{\gamma_{l'}^{j_0}(q)}\})^2 \\ &+(Sup\{\widehat{\gamma_l^{j_0}(q)}\} - Sup\{\widehat{\gamma_{l'}^{j_0}(q)}\})^2 \end{aligned} \tag{11}$$

Pour cette métrique $M$, une dissimilarité globale entre le pays $l$ et le pays $l'$ basée sur les différences entre distributions nationales des intervalles d'estimation des coefficients techniques est calculée selon le critère quadratique suivant :

$$d_M(\omega_l, \omega_{l'}) = \left(\sum_{q=1}^{Q} \delta_M^2(z_l^q, z_{l'}^q)\right)^{\frac{1}{2}} \tag{12}$$

Étant donné la matrice des dissimilarités entre distributions nationales de coûts spécifiques issues des calculs précédents, nous pouvons utiliser les méthodes de classification non supervisée. De façon similaire à la méthode de Ward, [4] propose un algorithme divisif de classification descendante hiérarchique sur données symboliques (DIVCLUST), valable pour les données d'intervalle et les données catégorielles.
Par la suite, nous détaillons pour les données d'intervalle les principes opérationnels de cette procédure de classification non supervisée. L'algorithme divisif de classification hiérarchique partage récursivement chaque classe en deux sous-classes, à partir de l'ensemble des pays en tant qu'objets symboliques $\Omega = \{\omega_1, \ldots, \omega_l, \ldots, \omega_L\}$. À chaque partition $P_K = \{C_1, \ldots, C_k, \ldots, C_K\}$ en $K$ classes symboliques, une classe doit être scindée pour obtenir une partition $P_{K+1}$, à $K + 1$ classes, optimisant le critère de sélection basé sur l'inertie. L'inertie de la $k^e$ classe est définie par $I(C_k) = \sum_{l \in C_k} \mu_l d_M^2(z_l, g(C_k))$ où $\mu_l$ est le poids du $l^e$ pays et $g(C_k$ est le barycentre de classe définie par :

$$g(C_k) = \frac{1}{\sum_{l \in C_k} \mu_l} \sum_{l \in C_k} \mu_l z_l \tag{13}$$

L'inertie intra-classes est définie par la somme des inerties des classes à leurs barycentres :

$$W(P_k) = \sum_{k=1,\ldots,K} I(C_k) \tag{14}$$

L'inertie inter-classes est définie par l'inertie des barycentres relatives au barycentre global $g$ de l'ensemble $\Omega$, comme suit :

$$\begin{aligned} W(P_k) = &\sum_{k=1,\ldots,K} \mu_k d_M^2(g(C_k, g)) \\ &\text{où } \mu_k = \sum_{l=1,\ldots,L} \mu_l \end{aligned} \tag{15}$$

Pour une partition , l'inertie totale regroupe l'inertie intra-classes avec l'inertie inter-classes :

$$I(\Omega) = W(P_K) + B(P_K) \tag{16}$$

Ainsi, minimiser l'hétérogénéité (mesurée par $W$) est équivalent à maximiser l'homogénéité (mesurée par $B$).
Générée par la réponse binaire (*oui/non*) à une question $\Psi = [z^q \leq c\,?]$, notons $\{A_k, \overline{A_k}\}$ la bipartition induite de la classe $C_k$ formée de $n_k$ objets. Afin de choisir parmi les $n_k - 1$ bipartitions possibles de la classe $C_k$, le critère discriminant est défini par le ratio suivant :

$$D(\Psi) = \frac{B^q(A_k, \overline{A_k})}{I^j(C_k)} = 1 - \frac{W^j(A_k, \overline{A_k})}{I^q(C_k)} \tag{17}$$

où l'inertie inter-classes $B^q(A_k, \overline{A_k})$ et l'inertie $I^q(C_k)$ sont calculées par rapport au $q^e$ quantile conditionnel. Aussi, minimiser l'inertie intra-classes $W(A_k, \overline{A_k})$ équivaut à maximiser l'inertie inter-classes $B(A_k, \overline{A_k})$ et, par

conséquence, le critère discriminant $D(\Psi)$.
Comme dans la méthode de Ward, la « hiérarchie supérieure » [14] à la partition $P_K$ est indexée par l'indice $h$ de la classe $C_K$, définie par son inertie inter-classes comme suit :

$$h(C_k) = B(A_k, \overline{A_k}) = \frac{\mu(A_k)\mu(\overline{A_k})}{\mu(A_k) + \mu(\overline{A_k})} d^2(g(A_k), g(\overline{A_k})) \quad (18)$$

L'algorithme DIVCLUS-T scinde la classe $C_K^*$ qui maximise $h(C_K)$, en assurant que $P_{K+1} = P_K \cup \{A_k, \overline{A_k}\} - C_K^*$, la partition suivante, présente la valeur minimum de l'inertie intra-classes, conformément à l'équation :

$$W(P_{K+1}) = W(P_K) - h(C_K^*) \quad (19)$$

D'après les deux règles conditionnelles de l'équivariance (§2.2), le critère discriminant (15) appliqué aux quantiles conditionnels demeure invariant par la transformation monotone correspondant à une reparamétrisation de $X$ de $M = Y - X$. Ainsi, les hiérarchies de classification divisive obtenues sont en dualité au sens de (7) par leurs seuils divisifs si l'on passe des quantiles conditionnels de coûts spécifiques aux quantiles conditionnels de marges brutes, et inversement. Comme pour le *biplot*, cette propriété justifie d'étiqueter le dendrogramme résultant de façon soit alternative, soit simultanée par les seuils estimées de coûts spécifiques ou de marges brutes pour en faciliter l'interprétation.

# 3 Application aux régions de l'Union européenne

L'estimation des quantiles de marge brute et de coûts spécifiques pour p=5 quantiles, déciles extrêmes (D1 et D9), quartiles (Q1 et Q3) et médiane (Q2) est effectuée conditionnellement au produit brut (cf. figure 1).

## 3.1 Estimations duales des coûts spécifiques et des marges brutes

Ces estimations vérifient empiriquement la relation de dualité (9), déduite de l'équivariance conditionnelle des estimateurs quantiles conditionnels par application des propriétés (7) et (8). Ainsi, pour le Schleswig-Holstein ($A010$), le premier décile estimé la marge brute unitaire ($D1mb$) est le complément à 1 000 € du neuvième décile estimé des coûts spécifiques ($D9cs$) : $\widehat{\mu_{0,1}^{mb}} = 458 = 1000 - \widehat{\mu_{0.9}^{cs}} = 1000 - 542$. La relation de complémentarité similaire entre le quartile inférieur de marge brute ($Q1mb$) et le quartile supérieur de coûts spécifiques ($Q3cs$) est également vérifiée quel que soit la région, par exemple pour le Schleswig-Holstein : $\widehat{\mu_{0,2}^{mb}} = 646 = 1000 - \widehat{\mu_{0,75}^{cs}} = 1000 - 354$. Cette propriété de dualité des estimations a été vérifiée par l'application conjointe de procédures d'estimation portant sur les coûts spécifiques d'une part et d'autre part sur les marges brutes.
Signalons le caractère très asymétrique des distributions,

| Blé | | Coûts spécifiques (€) pour 1 000 € de produit brut | | | | | | | Marges brutes (€) pour 1 000 € de produit brut | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Code | Régions RICA | D1cs | Q1cs | Q2cs | Q3cs | D9cs | MCOcs | | D1mb | Q1mb | Q2mb | Q3mb | D9mb | MCOmb | |
| A010 | Schleswig-Holstein | 389 | 407 | 364 | 354 | 542 | 464 | | 458 | 646 | 636 | 593 | 611 | 536 | |
| A030 | Niedersachsen | 214 | 229 | 295 | 351 | 327 | 305 | | 673 | 649 | 705 | 771 | 786 | 695 | |
| A050 | Nordrhein-Westfalen | 325 | 245 | 222 | 224 | 233 | 204 | | 767 | 776 | 778 | 755 | 675 | 796 | |
| A060 | Hessen | 383 | 410 | 343 | 356 | 333 | 428 | | 667 | 644 | 657 | 590 | 617 | 572 | |
| A090 | Bayern | 283 | 354 | 418 | 561 | 669 | 487 | | 331 | 439 | 582 | 646 | 717 | 513 | |
| A115 | Sachsen-Anhalt | 245 | 214 | 290 | 284 | 289 | 188 | | 711 | 716 | 710 | 786 | 755 | 812 | |
| F131 | Champagne-Ardenne | 379 | 514 | 537 | 530 | 752 | 492 | | 248 | 470 | 463 | 486 | 621 | 508 | |
| F132 | Picardie | 314 | 299 | 376 | 465 | 465 | 332 | | 535 | 535 | 624 | 701 | 686 | 668 | |
| F134 | Centre | 303 | 317 | 376 | 490 | 568 | 367 | | 432 | 510 | 624 | 683 | 697 | 633 | |
| F135 | Basse-Normandie | 699 | 646 | 617 | 525 | 687 | 507 | | 313 | 475 | 383 | 354 | 301 | 493 | |
| F136 | Bourgogne | 475 | 442 | 543 | 657 | 779 | 491 | | 221 | 343 | 457 | 598 | 525 | 509 | |
| F141 | Nord-Pas-de-Calais | 321 | 370 | 452 | 508 | 687 | 498 | | 313 | 492 | 548 | 630 | 679 | 502 | |
| F151 | Lorraine | 524 | 508 | 488 | 398 | 531 | 452 | | 469 | 602 | 512 | 492 | 476 | 548 | |
| F152 | Alsace | 589 | 721 | 817 | 886 | 1001 | 864 | | -1 | 114 | 183 | 279 | 411 | 136 | |
| F153 | Franche-Comté | 689 | 759 | 680 | 625 | 832 | 706 | | 168 | 375 | 320 | 241 | 311 | 294 | |
| F162 | Pays de la Loire | 428 | 421 | 435 | 562 | 555 | 403 | | 445 | 438 | 565 | 579 | 572 | 597 | |
| F163 | Bretagne | 368 | 400 | 456 | 421 | 483 | 396 | | 517 | 579 | 544 | 600 | 632 | 604 | |
| F192 | Rhône-Alpes | 609 | 640 | 904 | 935 | 911 | 903 | | 89 | 65 | 96 | 360 | 391 | 97 | |
| F193 | Auvergne | 336 | 384 | 528 | 792 | 804 | 434 | | 196 | 208 | 472 | 616 | 664 | 566 | |
| I260 | Emilia-Romagna | 187 | 270 | 374 | 437 | 481 | 346 | | 519 | 563 | 626 | 730 | 813 | 654 | |
| I282 | Umbria | 316 | 305 | 327 | 398 | 359 | 376 | | 641 | 642 | 673 | 695 | 684 | 624 | |
| B341 | Flandre | 305 | 363 | 446 | 496 | 1085 | 552 | | -85 | 504 | 554 | 637 | 695 | 448 | |
| B343 | Wallonie | 397 | 306 | 342 | 407 | 407 | 404 | | 593 | 593 | 658 | 694 | 603 | 596 | |
| D370 | Danemark | 213 | 261 | 363 | 462 | 552 | 426 | | 448 | 538 | 637 | 739 | 787 | 574 | |
| U411 | England-North | 332 | 348 | 346 | 471 | 564 | 375 | | 436 | 529 | 654 | 652 | 668 | 625 | |
| U412 | England-East | 264 | 302 | 336 | 421 | 475 | 323 | | 525 | 579 | 661 | 698 | 736 | 677 | |
| U413 | England-West | 363 | 343 | 294 | 308 | 406 | 349 | | 594 | 692 | 706 | 657 | 637 | 651 | |
| U431 | Scotland | 338 | 405 | 522 | 443 | 490 | 478 | | 510 | 557 | 478 | 595 | 662 | 522 | |
| E570 | Extremadura | 182 | 206 | 218 | 373 | 498 | 299 | | 502 | 627 | 782 | 794 | 818 | 701 | |
| O660 | Österreich | 201 | 240 | 266 | 261 | 344 | 200 | | 656 | 739 | 734 | 760 | 799 | 800 | |
| S710 | Slattyugdalan | 341 | 331 | 398 | 522 | 567 | 323 | | 433 | 478 | 612 | 669 | 659 | 677 | |
| H761 | Közép-Dunántúl | 382 | 465 | 419 | 348 | 638 | 563 | | 362 | 652 | 581 | 535 | 618 | 437 | |
| H762 | Nyugat-Dunántúl | 361 | 357 | 393 | 388 | 509 | 759 | | 491 | 612 | 607 | 643 | 639 | 241 | |
| H765 | Eszak-Alfoid | 302 | 238 | 339 | 371 | 341 | 197 | | 659 | 629 | 661 | 762 | 698 | 803 | |
| P785 | Pomoze & Mazury | 300 | 353 | 395 | 505 | 562 | 418 | | 438 | 495 | 605 | 647 | 700 | 582 | |
| P790 | Wielko Polska &Slesk | 252 | 280 | 344 | 430 | 536 | 424 | | 464 | 570 | 656 | 720 | 748 | 576 | |
| P795 | Mazowsze & Podlasie | 266 | 263 | 302 | 349 | 458 | 336 | | 542 | 651 | 698 | 737 | 734 | 664 | |
| P800 | Malopolska & Pogórze | 254 | 290 | 372 | 331 | 383 | 347 | | 617 | 699 | 628 | 710 | 746 | 653 | |

FIGURE 1 – Estimations des coûts spécifiques et des marges brutes pour 1000 € de produit brut.

ainsi les écarts d'estimation de coûts spécifiques les plus importants (supérieurs à 20% en valeur absolue) entre l'estimation conditionnelle de la médiane ($Q2cs$ et $Q2mb$) et celle de la moyenne ($MCOcs$ et $MCOmb$) sont pour les coûts spécifiques l'Extremadure (128%) et la Flandre (143%), tandis que pour les marges brutes, il s'agit de la Basse-Normandie (29%) et de la région hongroise du Nyugat-Dunàntùl (60%).
La distribution des valeurs propres de l'ACPn conduit à retenir les deux premières composantes principales selon la règle de Kayser. Ces deux premières composantes principales rassemblent 91% de l'information disponible. Le *biplot* de ces deux premières composantes principales est reproduit en figure 2. La première composante principale ($Dim1$) représente 80% de l'inertie. L'ensemble des quantiles conditionnels de coûts spécifiques ($cs$) est corrélé négativement ($Dim1 < 0$) tandis que l'ensemble des quantiles conditionnels de marge brute ($mb$) est corrélé positivement ($Dim1 > 0$). Les médianes conditionnelles ($Q2cs$ et $Q2mb$ de signes opposés) sont les quantiles estimés les mieux corrélés à cette première composante principale, qui reflète ainsi le niveau global de la distribution des quantiles estimés. L'angle entre l'orientation du vecteur $Q2cs$ et du vecteur $MCOcs$, identique à celui entre $Q2mb$ et $MCOmb$, fournit une indication intéressante sur l'asymétrie des distributions : plus cet angle est ouvert, plus l'asymétrie est forte. L'asymétrie des estimations quantiles de coûts supérieurs est portée par les quantiles d'ordre supérieur ($Q3cs$ et $D9cs$) tandis que celle des estimations quantiles de marge brute par les quantiles d'ordre inférieur ($Q1mb$, $D1mb$).
La seconde composante principale ($Dim2$), représentant 11% de l'inertie, restitue la dispersion de la distribution en opposant, dans le demi-plan $Dim1 < 0$, les quantiles su-

périeurs de coûts $Q3cs$ et $D9cs$ ($Dim2 < 0$) aux quantiles inférieurs de coûts spécifiques $Q1cs$ et $D1cs$ (demi-plan $Dim2 > 0$). De façon similaire, la seconde composante principale oppose dans le demi-plan $Dim1 > 0$ les quantiles inférieurs de marges brutes $D1mb$ et $Q1mb$ ($Dim2 > 0$) aux quantiles supérieurs de marge brute $Q3mb$ et $D9mb$ ($Dim2 < 0$).

## 3.2 Organisation de la variabilité des régions totalement estimées

Le *biplot* permet d'étiqueter chaque quadrant des plans factoriels de projection des observations par des orientations croissantes (e.g. $Q2^{+}_{cs}$ pour la zone d'estimations élevée des coûts spécifiques) ou décroissantes (e.g. $Q2^{-}_{mb}$ pour la zone d'estimations faibles des marges brutes) pour faciliter la lecture et l'interprétation des graphiques factoriels (cf. figures 4 et 5). La hiérarchie divisive obtenue avec l'option de distance euclidienne (figure 3) montre que, à l'exception du quantile *D9*, l'ensemble des estimations des quantiles *D1*, *Q1*, *Q2* et *Q3* est utilisé par les valeurs discriminantes, ce qui conduit à conserver ces paramètres pour décrire la distribution.

L'algorithme divisif permet d'interpréter les différences les

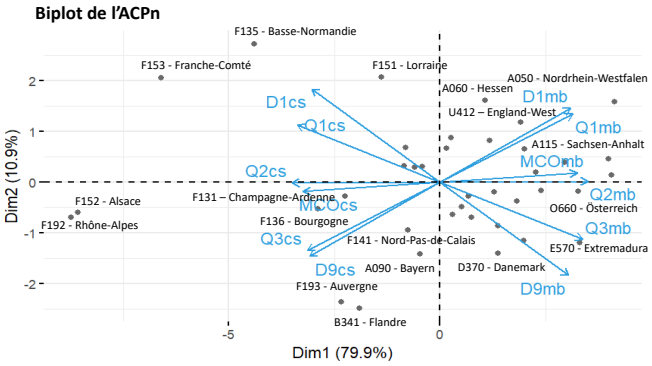



FIGURE 2 – Biplot de l'ACPN des estimations duales par quantiles de coût et de marge, régions de l'UE12. Source : traitement de l'auteur, RICA-UE 2006.

plus marquantes entre classes au sien d'une même partition en fonction de seuils de coûts et de marges. La première partition $P2$ en deux classes $C1$ et $C2$ où les régions se répartissent selon les valeurs du quartile inférieur de coûts ($Q1c$) et du quartile supérieur de marges ($Q3m$). La classe $C1$ rassemble les régions pour lesquelles $Q1c$ est supérieur ou égal à 577 € ($Q1c \geq 577$) et $Q3m$ est inférieur à 423€ ($Q3m < 423$). La classe $C2$ regroupe les régions pour lesquelles $Q1c$ est inférieur à 577€ ($Q1c < 577$) et $Q3m$ est supérieur ou égal à 423€ ($Q3m \geq 423$). Ainsi, au sommet de la hiérarchie divisive, la procédure de classification permet d'identifier deux modèles contrastés pour les distributions empiriques des coûts et des marges unitaires du blé pour 1€ de produit brut :

— d'une part, le modèle à résidus homoscédastiques identifiable à la classe $C1$ (coûts élevés et faibles marges aux distributions relativement homogènes) ;
— d'autre part, le modèle à résidus hétéroscédastiques identifiable à la classe $C2$ (coûts plus faibles et marges plus élevées aux distributions relativement hétérogènes).

La seconde partition $P3$ en trois classes divise la classe $C2$ en deux agrégats $C2.1$ et $C2.2$ selon les quantiles médian de coûts ($Q2c$) et de marge ($Q2m$). L'agrégat $C2.1$ rassemble les régions dont le quantile médian de coûts est supérieur ou égal à $609,5$ € ($Q2c \geq 609,5$) et le quantile médian de marge est inférieur à $390,5$ € ($Q2m < 390,5$). L'agrégat $C2.1$ regroupe les régions dont le quantile médian de coûts est inférieur à $609,5$€ ($Q2c < 609,5$) et le quantile médian de marge est supérieur ou égal à $390,5$ € ($Q2m \geq 390,5$). Cette césure correspond à une opposition de leurs projections selon la première composante principale ($Dim1$).

Affichée en figure 3, la partition $P4$ en quatre classes est optimale pour la différence minimale dans le logarithme du rapport des déterminants, fournie par le paquet *clusterCrit* [7], qui constitue une règle cohérente avec le critère de l'algorithme DIVCLUS-T [10]. Au niveau de cette partition $P4$, la classe $C1$ se scinde en deux agrégats $C1.1$ et $C1.2$ selon le décile supérieur de marge ($D9m$) et le décile inférieur de coûts ($D1c$). L'agrégat $C1.1$ regroupe les régions dont le neuvième décile est supérieur ou égal à 351 € ($D9m \geq 351$) et le premier décile de coûts est inférieur à 649€ ($D1c < 649$). L'agrégat $C1.2$ rassemble les régions dont le neuvième décile est inférieur à 351€ ($D9m < 351$) et le premier décile de coûts est supérieur ou égal à 649€ ($D1c \geq 649$). Cette césure correspond à une opposition de leurs projections selon la seconde composante principale ($Dim2$).

La figure 4 projette la partition en treize classes (*P13*) sur le plan principal de l'ACPn distributionnelle et permet de prendre en compte davantage d'information structurelles apportées par les quantiles conditionnels. En effet, l'axe *F2* constitue le facteur de dispersion intraclasse lié aux niveaux relatifs des quantiles conditionnels supérieurs (*D9* et *Q3*) vis-à-vis des quantiles conditionnels inférieurs (*Q1* et *D1*). Selon l'axe *F2*, la classe *C1* aux estimations quantiles les plus extrêmes est scindée en deux agrégats bien distincts. D'une part dans le quadrant $F1 > 0$ & $F2 > 0$, l'aîné $C1.1 =$ {*Rhône-Alpes*, *Alsace*} aux quantiles supérieurs de coûts parmi les plus élevés ($Q3^{+}_{cs} = 911$€, $D9^{+}_{cs} = 956$€) et aux quantiles inférieurs de marge parmi les plus faibles ($Q1^{-}_{mb} = 89$ €, $D9^{-}_{mb} = 44$€). D'autre part, dans le quadrant $F1 > 0$ & $F2 < 0$, le benjamin $C1.2 =$ {*Franche-Comté*, *Basse-Normandie*} présentant un décile supérieur de coûts équivalent mais dont les autres estimations quantiles sont des extrema de second rang, avec des quantiles inférieurs de coûts parmi les plus élevés ($Q1^{+}_{cs} = 703$€, $D1^{+}_{mb} = 694$€) et des quantiles supérieurs de marges parmi les plus faibles ($Q3^{-}_{mb} = 297$€, $D9^{-}_{mb} = 306$€).

Selon l'axe $F2$, la classe $C2.1$ est formée par la réunion de deux agrégats :

— d'une part, situé dans le quadrant $F1 < 0$ & $F2 > 0$), l'agrégat $C2.1.1 =$ {*Extremadura*, *Wielkopolska*

*& Slask*, *Danemark*, *Emilia-Romagna*, *England-East*, *Picardie*} présente des estimations de quantiles inférieurs de coûts plus faibles ($Q1^-_{cs} = 273$ €, $D1^-_{cs} = 238$ €) et de quantiles supérieurs de marges plus élevées ($Q3^-_{mb} = 727$€, $D9^-_{cs} = 762$€) que l'ensemble des régions actives ;

— d'autre part, situé essentiellement dans le quadrant $F1 < 0$ &$F2 < 0$ l'agrégat $C2.1.2$ ={*Mazowsze & Podlasie*, *Malopolska & Pogörze*, *Eszak-Alfoid*, *Umbria*, *England West*, *Wallonie*, *Hessen*, *Nordrhein-Westfalen*, *Österreich*, *Sachsen-Anhalt*, *Niedersachsen*} se distingue du précédent par des estimations de quantiles inférieurs de marges plus fortes ($Q1^+_{mb} = 673$€, $D9^+_{cs} = 647$€) et des estimations de quantiles supérieurs de coûts plus faibles ($Q3^-_{cs} = 327$€, $D9^-_{cs} = 353$€).

Au niveau des coûts faibles (demi-plan $F1 < 0$), les agrégats $C2.1.1.2.1$ ={*Emilia-Romagna*, *Danemark*, *Wiekopolska & Slask*} et $C2.1.2.1.2$ ={*Hessen*, *Wallonie*, *England-West*}, situés au même niveau médian de coûts spécifiques ($Q2^-_{cs} = 360$€ versus $Q2^-_{cs} = 326$€ respectivement), se différencient selon l'axe $F2$ par leurs profils de distribution en quantiles inférieurs de coûts (soit $D1^-_{cs} = 217$€ versus $D1^-_{cs} = 383$€, et respectivement $D1^-_{cs}$ =217€ versus $D1^-_{cs} = 383$€) relativement moins élevés pour l'agrégat $C2.1.1.2.1$ par rapport à ceux de l'agrégat $C2.1.2.1.2$ alors qu'ils appartiennent tous les deux à la classe $C2.1$ de la partition $P4$. Enfin, à un niveau plus élevé de coûts (demi-plan $F1 > 0$), les agrégats $C2.2.2.1.2$ ={Bayern, Nord-Pas-de-Calais} et $C2.2.2.2.1$ ={Lorraine, Bretagne, Pays de la Loire, Scotland}, issus de l'éclatement de la classe $C2.2.2$ de la partition $P4$, se distinguent tant selon les quantiles supérieurs de marge (avec des estimations plus élevées pour $C2.2.2.1.2$, soit $Q3^+_{mb} = 638$€ et $D9^+_{mb} = 698$€, que pour $C2.2.2.1.2$, soit $Q3^+_{mb} = 567$€ et $D9^+_{mb} = 586$€) que selon les quantiles inférieurs de coûts (avec des estimations plus faibles pour $C2.2.2.2.1$ soit $Q1^-_{cs}$ =362€ et $D1^-_{cs}$ =302€, que pour $C2.2.2.1.2$, soit $Q1^+_{cs}$ =434€ et $D1^+_{cs} = 415$€).

Ainsi, dans l'allocation spécifique des charges aux produits, l'estimation en quantiles conditionnels nous permet de conserver l'information distributionnelle relative à l'hétérogénéité des coûts et des marges pour un niveau donné de dépenses spécifiques au blé, contrairement à l'estimation des moindres carrés ordinaires.

Le graphique suivant (figure 5) visualise les résultats de l'analyse en composantes normée sur intervalles (ACPni) sur la base des intervalles d'estimations par intervalles $[Inf, Sup]$ de quantiles conditionnels des barycentres des 13 agrégats du référentiel typologique $P13$ choisi pour l'affectation pseudobarycentrique.

Le premier axe factoriel $F1$ de l'ACPni donne un gradient décroissant (inversé par rapport aux axes factoriels des ACP classiques précédentes) des estimations de coûts spécifiques allant des estimations supérieures de coûts médians et inférieures de marges médianes (agrégats $C1.1$ et $C1.2$ projetés aux extrêmes du pôle $F1 < 0$, $Q2^+_{cs} = 755$€ et $Q2^-_{mb} = 246$€, en moyenne) aux estimations inférieures

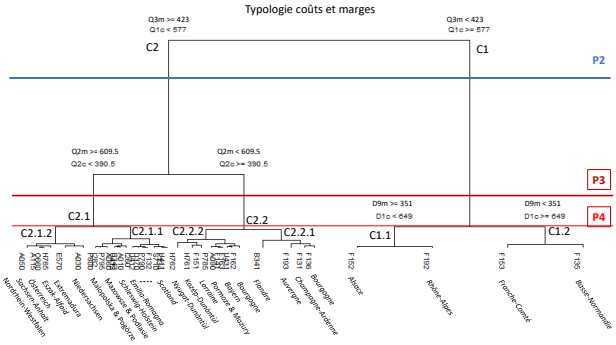



FIGURE 3 – Classification symbolique divisive en distance euclidienne pour les estimations duales en quantiles de coûts et de marges, régions de l'UE12.

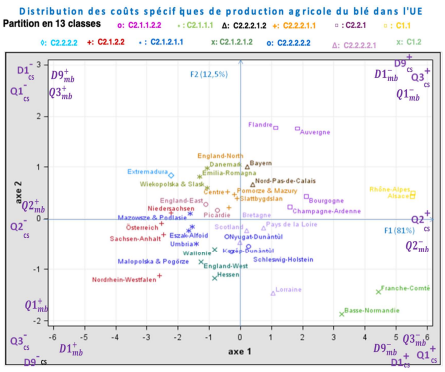



FIGURE 4 – blé, projection de la partition en treize classes des régions complètement estimées.

de coûts médians et aux estimations supérieures de marges médianes (agrégats $C2.1.2.2$ et $C2.1.1.1$ projetés aux extrêmes du pôle $F1 > 0$, $Q2^-_{cs} = 243$€ et $Q2^+_{mb} = 757$€, en moyenne).

Le second axe factoriel $F2$ de l'ACPni oppose l'agrégat $C1.1$ ={*F192-Rhône-Alpes*, *F152-Alsace*} dans le quadrant $F1 < 0$ & $F2 > 0$, présentant un écart d'estimation élevé entre le premier et le dernier décile de coûts ($[D9\text{-}D1]^+_{cs} = 357$€, en moyenne). Dans le quadrant $F1 < 0$ & $F2 < 0$, l'agrégat $C1.2$ ={*F135-Basse-Normandie*, *F153-Franche-Comté*} dont les écarts d'estimation entre le premier et le dernier décile sont parmi les plus faibles ($[D9\text{-}D1]^-_{cs} = 66$€, en moyenne). Le second axe constitue également l'axe majeur de dispersion intra-classe des autres agrégats de la partition $P13$, en particulier celui de l'agrégat $C2.2.1$ ={*B341-Flandre*, *F193-Auvergne*, *F136-Bourgogne*, *F131-Champagne-Ardenne*}. Cette analyse peut servir de test graphique de séparation : les agrégats $C1.1$ et $C1.2$ de la classe $C1$ présentent entre

eux des différences de coûts et de marge à la fois en termes de niveau (selon l'axe $F1$ corrélé au niveau médian $Q2$) et de structure (selon l'axe $F2$, corrélé à l'écart inter-décile $D9 - D1$).

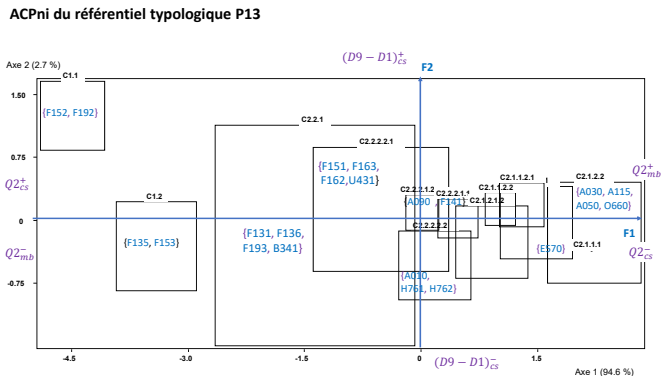



FIGURE 5 – ACPni du référentiel typologique en 13 classes, estimations duales en quantiles de coûts et de marges, régions de l'UE12.

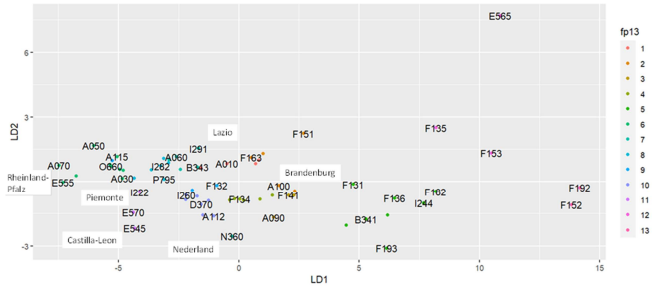



FIGURE 6 – blé, ADL de la partition P13 pour les régions complètement estimées et classement des régions partiellement estimées

Une affectation au plus proche barycentre, selon une norme quadratique des écarts, permet de décider de l'appartenance des régions partiellement estimées aux treize classes de la partition $P13$, retenue comme référentiel typologique. Quasiment équivalente à une analyse discriminante linéaire (ADL, figure 6) réalisée à partir des estimations complètes de distributions régionales (échantillon d'apprentissage) et appliquée aux régions partiellement estimées (échantillon-test), l'affectation réalisée par la procédure FASTCLUS, à partir des estimations barycentriques de quantiles conditionnels, est projetée dans le premier plan factoriel (figure 7) de l'ACPn des barycentres d'agrégats représentant 98% de l'inertie interclasses de la partition $P13$.

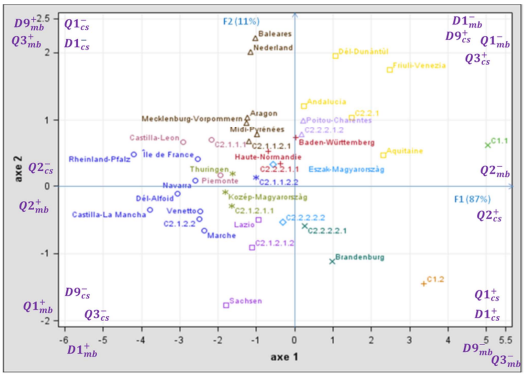



FIGURE 7 – ACPn des barycentres de la partition P13 et imputation des régions partiellement estimables.

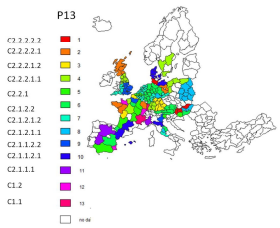



FIGURE 8 – blé, projection cartographique de la partition P13 des régions complètement et partiellement estimées.

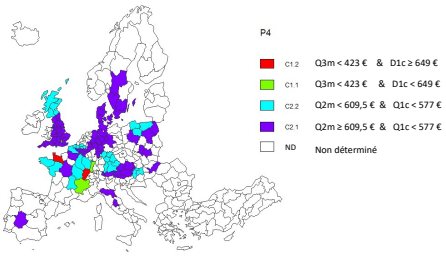



FIGURE 9 – blé, projection cartographique de la partition P4 des régions complètement estimées.

Une cartographie des régions complètement estimées et de l'affectation des régions partiellement estimées (figure 8) situe la localisation des classes du référentiel typologique en treize classes et leur répartition au sein de l'espace agricole de l'Union européenne.Cette carte typologique en treize classes précise et complète, la cartographie effectuée à partir de la partition $P4$ en quatre classes (figure 9) par une catégorisation plus fine et respectivement une couverture territoriale plus étendue. Comme affiché en figure 9, il est loisible de communiquer la genèse duale de chaque classe fournie par la hiérarchie divisive en termes de seuils de coûts et de marges pour en faciliter l'interprétation.

# 4 Conclusion

Sur la base du Réseau d'information comptable européen, nous avons testé sur une base régionale la faisabilité et la pertinence pour le blé de notre méthodologie d'estimation micro-économétrique des coûts spécifiques de production et des marges brutes selon les quantiles conditionnels. Cette méthodologie est complétée par une procédure d'imputation pour les régions partiellement estimées.
Compte-tenu de la nature duale des estimations de coûts spécifiques et de marge brute, la représentation en *biplot* constitue une aide à l'interprétation pertinente. En cohérence, les noeuds du dendrogramme de la classification divisive basée sur le pourcentage de variabilité interclasses sont étiquetés de façon duale par les seuils estimés de coûts et de marge facilitant l'interprétation de la hiérarchie des partitions.
L'analyse en composantes principales sur intervalle permet d'interpréter les composantes de la variabilité de la distribution des agrégats de la typologie choisie comme référentiel pour l'imputation pseudo-barycentrique des régions partiellement estimées.
Grâce à ce type d'analyse, nous confirmons qu'il n'y a pas *un* coût spécifique national de production qui pourrait être estimé par une moyenne conditionnelle mais des classes régionales européennes de distribution des coûts spécifiques et des marges brutes qui peuvent être positionnées dans un schéma bidimensionnel stable selon un nombre déterminé d'estimations quantiles conditionnelles.
Pour mieux distinguer les différences entre certaines des distributions régionales, il est loisible d'étendre l'analyse à une échelle de quantiles plus fine si nécessaire.
**In Memoriam** : l'auteur dédie ce travail à la mémoire d'***Edwin Diday***, Professeur émérite de l'Université Paris Dauphine, récemment disparu.

# Références

[1] L. Billard et E. Diday, *Symbolic Data Analysis : Conceptual Statistics and Data Mining,* Wiley–Blackwell, 2006.

[2] A. Cameron et P. K. Trivedi, *Microeconometric. Methods and Applications,* University Press, 2005.

[3] P. Cazes, A. Chouakria, E. Diday et Y. Schektman, Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de Statistique Appliquée*, Vol. 45(3), pp. 5-24, 1997.

[4] M. Chavent, Y. Lechevallier et O. Briant, Divclus-t : A Monothetic Divisive Hierarchical Clustering Method, *Comput. Statist. Data Anal.*, Vol. 52(2), pp. 687-701, 2007.

[5] D. Desbois, *Estimation des coûts de production agricoles : approches économétriques*, Thèse de doctorat, Université Paris Saclay, 2015.

[6] D. Desbois, J.P. Butault et Y. Surry, Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Économie rurale*, Vol. 361, pp. 3-22, 2017.

[7] B. Desgraupes, Clustering indices, *Vignette R* , CRAN, 2017.

[8] J.F. Divay et F. Meunier, Deux méthodes de confection du tableau entrées-sorties, *Annales de l'INSEE*, Vol. 37, pp. 59-109, 1980.

[9] X. D'Haultfoeuille et P. Givord, La régression quantile en pratique, *Économie et Statistique*, Vol. 471(1), pp. 85-111, 2014.

[10] M. Fuentes et M. Chavent, Clustering divisif monothétique, *Vignette R*, 4e Rencontre *R*, 2015.

[11] X. He et F. Hu, Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association*, Vol. 97(459), pp. 783-795, 2002.

[12] R. Koenker et G. Bassett, Regression quantiles, *Econometrica*, Vol. 46, pp. 33-50, 1978.

[13] R. Koenker et Q. Zhao, L-estimation for linear heteroscedastic models, *Journal of Nonparametric Statistics*, Vol. 3, pp. 223-235, 1994.

[14] B. Mirkin, *Clustering for Data Mining. A Data Recovery Approach,* CRC Press, 2005.

# Modèles graphiques causaux interactifs pour les données textuelles

A. Ferdjaoui[1,2], S. Affeldt[2], M. Nadif[2]

[1] SogetiLabs
[2] Centre Borelli UMR 9010, Université Paris Cité, France

amine.ferdjaoui@etu.u-paris.fr | severine.affeldt@u-paris.fr | mohamed.nadif@u-paris.fr

## Résumé

*Nous proposons de reconstruire des modèles graphiques causaux à partir de données textuelles via un nouveau package Python,* `WordGraph`*. Ce package facilite l'exploration de grands corpus de documents par des visualisations interactives sous forme de modèles graphiques de mots. Le pipeline WordGraph exploite à la fois les widgets jupyter et le notebook jupyter pour aider les utilisateurs qui n'ont pas d'expérience en Python à maîtriser un pipeline* `WordGraph`*, qui est entièrement personnalisable.* `WordGraph` *est disponible via un dépôt GitHub*[1] *qui fournit également une courte vidéo présentant l'utilisation de notre système.*

## Mots-clés

*co-clustering, réseaux causaux, données textuelles.*

## Abstract

*We propose to reconstruct causal graphical models from textual data via a new Python package,* `WordGraph`*. This package facilitates the exploration of large document corpora through interactive visualizations in the form of graphical word models. The WordGraph pipeline leverages both jupyter widgets and jupyter notebook technologies to help users with no Python experience master a fully customizable WordGraph pipeline.* `WordGraph` *is available via a GitHub*[1] *repository, which also provides a short video demonstrating the usage of our system.*

## Keywords

*co-clustering, causal network reconstruction, text data.*

# 1 Introduction

## 1.1 Co-clustering et données textuelles

La tâche consistant à répartir simultanément les objets et les caractéristiques dans des blocs homogènes pour une matrice permet d'obtenir des regroupements de lignes et de colonnes plus précis et plus faciles à interpréter qu'un regroupement unidimensionnel. Depuis les travaux fondateurs de Hartigan [19], le co-clustering (également connu sous le nom de biclustering) a trouvé des applications dans de nombreux domaines tels que la bio-informatique [18], le web mining [14] et le text mining [9]. Le co-clustering est particulièrement bien adapté aux matrices de données documents×termes, qui sont par essence de haute dimension, peu denses et présentent des caractéristiques directionnelles.

Les algorithmes de co-clustering exploitables sur ces matrices de cooccurrences sont issus de différentes approches. Les méthodes de co-clustering de type spectral, qui traitent la matrice d'entrée comme un graphe bipartite entre les documents et les mots, approximent la coupe normalisée de ce graphe à l'aide d'une relaxation réelle [8]. Les méthodes basées sur des modèles appropriés dérivés de blocs latents ou *Latent Block Models* (LBM) [15], reposent sur des algorithmes de type Expectation-Maximisation [16, 34]. Le co-clustering peut également s'appuyer sur des méthodes basées sur la factorisation matricielle telles que la Factorisation Matricielle Non négative (NMF) [25, 4, 12] ou la Tri-Factorisation (NMTF) [10, 33]. En revanche, les méthodes qui utilisent la théorie de l'information, utilisées avec les tableaux de contingence à deux voies, visent à minimiser la perte d'information mutuelle en regroupant les lignes et les colonnes de la matrice en fonction du co-clustering [9] ; il convient de noter que le critère optimisé est associé à un modèle de Poisson de type LBM contraint [27, 17]. Enfin, la copartition est également possible via une version adaptée de la mesure de *modularité* habituellement utilisée pour les réseaux [24, 3].

Parmi les packages populaires mis à disposition pour le co-clustering de données [23, 5, 11], le package Python `CoClust` est devenu une référence bien connue dans le domaine du co-clustering des matrices de cooccurrence, en particulier pour les matrices document-terme utilisées dans les applications de text mining [32]. `CoClust` fournit les implémentations de trois algorithmes conçus pour traiter efficacement de telles matrices[2]. L'un de ces algorithmes utilise le critère d'information mutuelle, et les deux autres effectuent le co-clustering en maximisant la modularité des graphes bipartites.

Dans les sections suivantes, nous montrons comment le package Python `WordGraph` que nous proposons tire parti de `CoClust` pour permettre l'exploration de grands corpus et du vocabulaire associé à l'aide de modèles graphiques.

1. https://github.com/MLDS-software/WordGraph

2. https ://coclust.readthedocs.io/en/v0.2.1/install.html

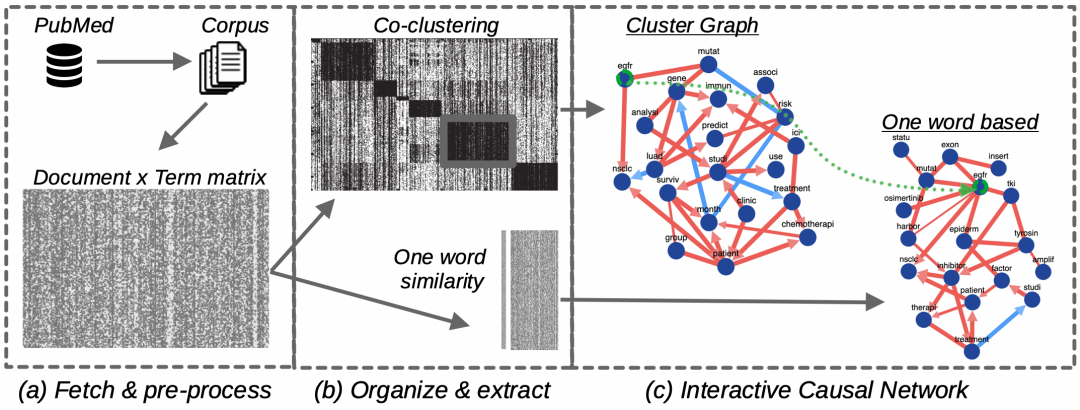



FIGURE 1 – Process WordGraph : de la recherche de documents aux réseaux causaux interactifs entre les mots.

### 1.2 Inférence de réseaux causaux

La reconstruction de modèles graphiques, pour analyser facilement la quantité toujours croissante des données massives, est devenue une pratique bien connue dans de nombreux domaines. Ces modèles peuvent être appris à partir de données de séries temporelles, d'expériences de perturbation contrôlée ou de données d'observation, comme c'est le cas dans de nombreux contextes biologiques. Les méthodes traditionnelles de reconstruction de modèles graphiques, qui exploitent des données non perturbées, comprennent le *search-and-score* bayésiens [20], la *sparse inverse covariance estimation* [13], l'entropie maximale [22] ou les *diffusion maps* [6].

Au-delà de la reconstruction de modèles graphiques, la découverte de la causalité entre plusieurs variables est d'un grand intérêt. Par exemple, elle peut conduire à l'identification de facteurs de transcription ayant un rôle de régulateurs majeurs pour des maladies humaines complexes. Appliquée à des données textuelles, ces reconstructions peuvent fournir des *résumés graphiques* des interactions entre les différents sujets d'un large corpus. Les réseaux de causalité qui en résultent indiqueraient ainsi le flux organisationnel entre les mots par le biais de réseaux dirigés. Cependant, les approches traditionnelles supposent que le modèle sous-jacent est soit un réseau complètement orienté, soit un réseau non orienté, car ces approches ne peuvent pas découvrir la causalité à partir de données d'observation. En revanche, les méthodes basées sur les contraintes [37, 28, 7, 31], qui peuvent identifier les contraintes structurelles correspondant à des arêtes inutiles dans un graphe, peuvent découvrir la causalité à partir de données non perturbées. L'inconvénient des méthodes basées sur les contraintes est qu'elles présentent des problèmes de complexité algorithmique en présence de variables non observées, et qu'elles ne sont généralement pas robustes sur les petits ensembles de données.

Pour contourner ces difficultés, l'algorithme `MIIC` (Multivariate Information-based Inductive Causation), s'appuie sur une méthode dérivée de la théorie de l'information qui combine l'apprentissage basé sur les contraintes et le principe de maximisation de la vraisemblance [1, 2, 38]. Plus précisément, `MIIC` est basé sur l'analyse de l'information multivariée [26, 39], qui étend le concept d'information mutuelle à plus de deux variables. En pratique, l'intégration d'une approche basée sur la découverte des contraintes structurelles dans le cadre de la théorie de l'information améliore considérablement la précision de la prédiction, le temps d'exécution et les capacités de passage à l'échelle en termes de taille d'échantillon et de taille de réseau par rapport aux approches traditionnelles.

Ainsi, alors que `MIIC` a démontré son intérêt sur une grande variété de données génomiques, à différentes échelles biologiques d'espace et de temps [38], il semble également être particulièrement bien adapté pour traiter les données textuelles de grande dimension, facilitant ainsi l'exploration des thématiques de large corpus. `MIIC` a été implémenté et mis à la disposition de la communauté dans un package R[3]. Il est également disponible via un serveur web accessible en ligne[4] [35].

Dans ce qui suit, nous détaillons l'interface de programmation Python proposée par `WordGraph` et montrons comment elle combine naturellement les approches de co-clustering, qui facilitent la découverte des thématiques d'un corpus, et la reconstruction de modèles graphiques causaux, qui identifient les interactions dirigées entre les termes.

## 2 Description générale du pipeline

Notre pipeline se décompose en trois parties. Tout d'abord, il propose la création d'un corpus de documents directement extraits de la base de données en ligne PubMed[5] (Fig. 1, *a*), qui comprend plus de 35 millions de citations pour la littérature biomédicale. La création du corpus inclut le prétraitement des données textuelles (par exemple, le *stemming* ou l'élimination des mots peu informatifs)

3. https ://github.com/miicTeam/miic_R_package
4. https ://miic.curie.fr
5. https ://pubmed.ncbi.nlm.nih.gov

et la construction d'une matrice documents×termes, avec une procédure de pondération TF-IDF (Term Frequency-Inverse Document Frequency).
Ensuite, `WordGraph` permet le co-clustering de la matrice documents×termes pondérée pour explorer les thématiques sous-jacentes. Cette partie s'appuie sur le package `Coclust` et fournit le regroupement simultané des mots et des documents (Fig. 1, *b*, haut). Dans cette étape, on peut également obtenir un extrait de la matrice documents×termes avec les colonnes correspondant aux termes les plus similaires à un mot donné, sur la base du score de *similarité cosine* (Fig. 1, *b*, bas).
Enfin, `WordGraph` fournit une interface de programmation Python du package R `MIIC` qui permet la reconstruction des réseaux causaux à partir de données d'observation. Un modèle graphique peut être obtenu à partir des termes les plus représentatifs - typiquement, les mots les plus fréquents - d'un co-cluster (Fig. 1, *c*, haut). Les visualisations de graphes, qui utilisent la bibliothèque Python `ipycytoscape`[6], sont interactives. L'utilisateur peut réorganiser les nœuds, obtenir des vues agrandies et cliquer sur un sommet pour construire un graphe secondaire, *One word similarity*, basé sur un extrait de la matrice document×term (Fig. 1, *c*, bas). On peut également obtenir ce dernier type de graphe en entrant un mot du vocabulaire dans le champ de saisie *widget*.

Dans ce qui suit, nous fournissons les détails du code source de `WordGraph` et expliquons les exploitations possibles à plusieurs niveaux. En effet, le package proposé peut être utilisé tel quel, dans une source Python ordinaire, ou bien l'utilisateur peut personnaliser un *notebook jupyter* interactif. Enfin, une application web peut être instantanément générée à partir du notebook pour faciliter le partage en ligne.

# 3 Un package à plusieurs niveaux

`WordGraph` permet d'accéder aux fonctionnalités de co-clustering de `Coclust` et à la méthode de reconstruction de réseaux causaux de `MIIC` suivant trois niveaux différents afin d'explorer les données textuelles. Dans cette section, nous présentons `WordGraph` comme *(i)* une interface de programmation Python légère pour le package R `MIIC`, *(ii)* un *notebook* Python interactif pour la reconstruction de modèles graphiques de mots, et *(iii)* une application web.

## 3.1 Une interface de programmation Python pour les réseaux causaux

`WordGraph` est le premier package Python qui propose une interface de programmation au package R `MIIC`. La figure 2 présente les principales fonctionnalités de `WordGraph` et de ses modules.
L'interface de programmation pour `MIIC` est encapsulée dans la méthode *build_graph* (*class* **WordGraph**) et repose sur les fonctions *create_miic* et *preproc_graphML* (module *utils*). Les autres fonctions de *utils* sont dédiées au prétraitement du corpus, tandis que la fonction *parse_pubmed_api* offre un accès facile à PubMed (module *io*).

6. https ://ipycytoscape.readthedocs.io/en/latest/

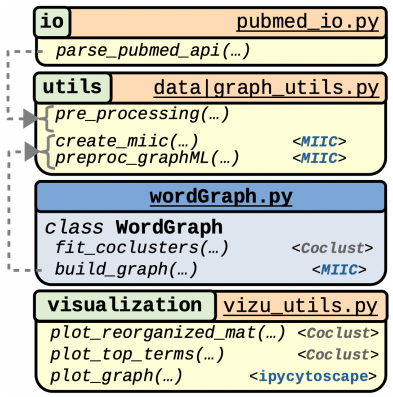



FIGURE 2 – Functionnalités principales de WordGraph.

```
1  from wordGraph.wordGraph import WordGraph
2  from wordGraph.io.pubmed_io import
       parse_pubmed_api
3  from wordGraph.visualization.viz_utils import
       plot_top_terms, plot_graph
4
5  # 1. Definition des parametres
6  pm_query = 'cancer'; n_doc = 2500 # Corpus
7  n_clust = 3; n_terms = 10         # Co-clustering
8  g_shf = 100; g_thresh = 0.05      # MIIC
9  # 2. Recuperation des donnees PubMed
10 raw_corpus,clean_corpus=parse_pubmed_api(pm_query
       ,n_doc)
11 # 3. Apprentissage des coclusters
12 wg = WordGraph()
13 wg.fit_coclusters(clean_corpus, n_clust)
14 plot_top_terms(wg, nb_terms=n_terms)
15 # 4. Construction du reseau de mots
16 wg.set_cocluster(cocluster=3, nb_terms=n_terms)
17 wg.build_graph(n_shuffles=g_shf,conf_threshold=
       g_thresh)
```

Listing 1 – Utilisation simple de WordGraph

Le Listing 1 présente une utilisation simple de `WordGraph`, depuis la recherche de document via PubMed (l.10) jusqu'au calcul de l'objet graphe `MIIC` (l.17). Nous fournissons dans `WordGraph` un *notebook jupyter* qui intègre cette proposition d'utilisation simple (voir *WordGraph utilisation simple.ipynb*).

## 3.2 Un notebook interactif d'exploration

`WordGraph` dispose d'un module *visualization* qui utilise le package `ipycytoscape`. En particulier, *plot_graph* affiche des réseaux interactifs dans un *notebook jupyter*, sur lequel la disposition des graphes peut être directement ajustée à l'aide de la souris. Il fournit des histogrammes de fréquence pour les principaux termes des co-clusters et permet ainsi de résumer facilement les thématiques (Fig. 3, à gauche). Un clic sur un nœud du modèle graphique de mots permet l'affichage d'un réseau interactif secondaire, dans lequel les sommets sont choisis en fonction de leur similarité avec le nœud précédemment sélectionné (Fig. 3, droite). De plus, `WordGraph` propose un *notebook jupyter* faci-

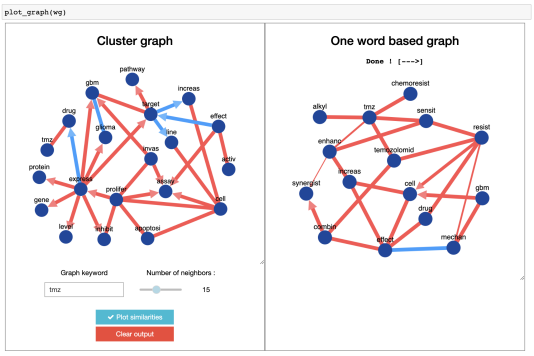



FIGURE 3 – Graphes interactifs dans le notebook jupyter.

lement personnalisable (*WordGraph Custom.ipynb*), dans lequel l'ensemble du pipeline peut être incorporé dans une seule cellule. Le notebook intègre plusieurs éléments `ipywidgets` (par exemple, un curseur, un bouton, des entrées textuelles et numériques). La figure 4 montre comment le pipeline s'organise dans une seule cellule.

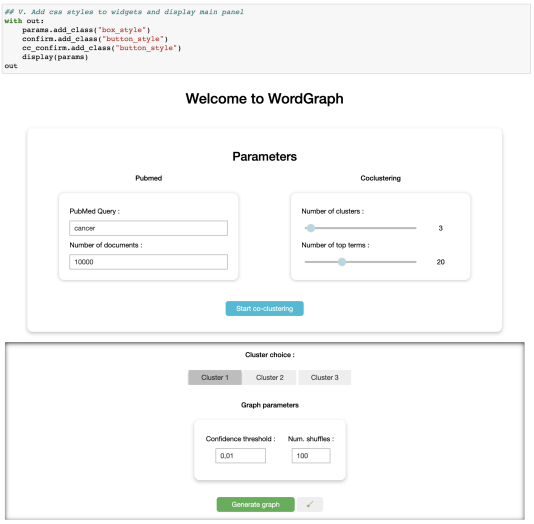



FIGURE 4 – Pipeline interactif dans un jupyter notebook.

La modification ou l'ajout d'un *widget* ne nécessite que quelques lignes de code. Le Listing 2 fournit un exemple d'implémentation pour la mise en place d'un bouton *confirm* et l'appel de sa fonction au moment du clic.

```
# Bouton 'Confirm' (label 'Do graph')
confirm = widgets.Button(description='Do graph',
          disabled=False, button_style='success')
# Fonction callback via clic souris
def on_confirm_button_clicked(button):
    button.disabled=True
    # Par exemple: generation du reseau../..
    button.disabled=False
# Association de la fonction au bouton
confirm.on_click(on_confirm_button_clicked)
```

Listing 2 – Personnalisation du notebook interactif

## 4 Une application web interactive

Le notebook *WordGraph Custom.ipynb* a également été conçu pour être automatiquement transformé en application web, avec le package *voila*[7]. Ainsi, l'utilisateur peut exploiter le *WordGraph Custom.ipynb* tel quel ou le personnaliser, puis obtenir facilement l'application web correspondante (Fig. 5), prête à l'emploi pour les utilisateurs qui ne sont pas familiers avec la programmation.

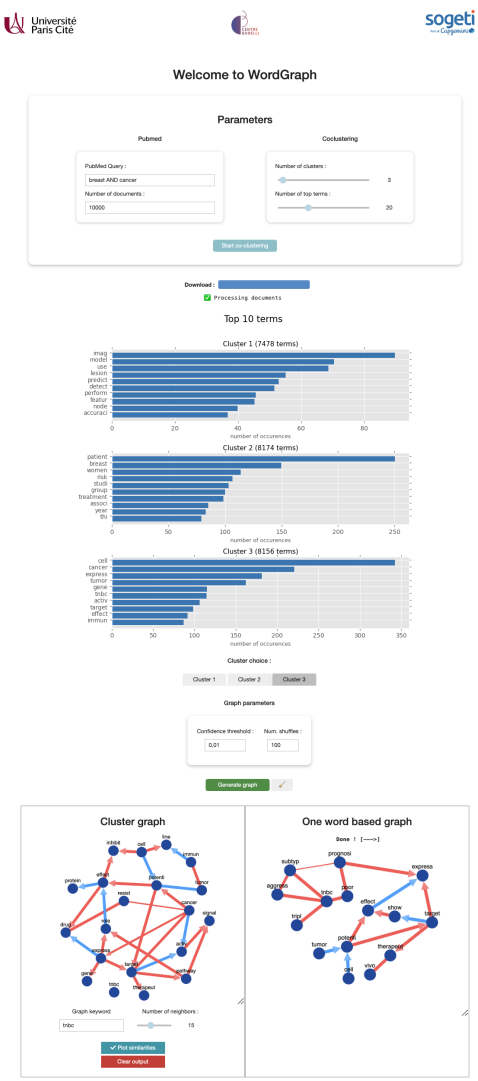



FIGURE 5 – Application web automatiquement générée.

## 5 Application biomédicale

Nous avons généré trois corpus en lien avec le cancer et d'une taille de $10,000$ documents, et avons exploré leurs thématiques avec `WordGraph`.
La Figure 1 se focalise sur le **cancer du poumon**. L'opposition (lien bleu) entre *NSCLC* (Non-Small Cell Lung

7. https ://voila.readthedocs.io/en/stable/using.html

Cancer) et *LUAD* (LUng ADenocarcinoma) suggère que les documents traitant de *LUAD* ignorent le sujet *NSCLC*. En effet, même si les cancers de type *LUAD* appartiennent à la famille *NSCLC*, leur pronostic et leur expression génétique sont très distincts [29] et ils sont considérés comme des entités séparées. Le cancer *EGFR*-positive NSCLC résulte d'une mutation du gène EGFR (récepteur du facteur de croissance épidermique). Notre réseau de mots établit un lien causal entre la présence dans les documents du terme *EGFR* et de l'acronyme *NSCLC*. Le terme *ICI* (Immune Checkpoint Inhibitor) *treatment*[8] a également une relation de cause à effet avec le nœud *immun*. Le réseau secondaire, basé sur *egfr*, se concentre sur le *treatment*, qui vise généralement à prévenir la croissance anormale de la tumeur en ciblant la *Tyrosin* Kinase *Inhibitor* (*TKI*), comme dans le permet le traitement par *osimertinib* [30].

Les modèles graphiques de la Figure 3 correspondent à un corpus centré sur le **glioblastoma cancer**. Le glioblastome multiforme (GBM) est le type de cancer du cerveau le plus aggressif, avec une augmentation de l'*invas*ion, de la *proliferation* et une diminution de l'*apoptosis*, comme le détaille notre graphe. Nous remarquons également le nœud *TMZ*, qui est le médicament (*drug*) chimiothérapeutique prédominant dans le GBM. Le graphe secondaire (Figure 3, droite), qui se concentre sur *TMZ*, apporte d'autres informations. En particulier, nous comprenons que la tumeur *GBM* a tendance à développer une *TMZ* (*temozolomid*) *resist*ance ou *chimioresist*ance [36].

Enfin, la Figure 5 propose une exploration d'un corpus qui concerne le **cancer du sein**. En particulier, le graphe de cluster met l'accent sur la possibilité de *drug resist*ance pour ce cancer. Ceci est particulièrement vrai pour le cancer du sein de type Triple Négatif *TNBC*. Séparé du réseau de cluster (Figure 5, gauche), le *TNBC* s'impose comme une sous-thématique, pour laquelle le graphique secondaire (Figure 5, droite) fournit des informations supplémentaires. Plus précisément, il s'agit d'un *subtype agressive*, dont le *prognosis* est généralement *poor* [21].

## 6 Conclusion

L'exploration de corpus devient de plus en plus difficile car d'énormes quantités de données textuelles sont disponibles dans de nombreux domaines. Nous avons créé `WordGraph` afin de fournir un package Python facile à exploiter pour les utilisateurs souhaitant extraire un maximum d'informations autour d'un sujet biomédical. Nous avons fait des efforts particuliers pour faire de `WordGraph` une application web autonome pour les utilisateurs n'ayant aucune connaissance de Python ou de *jupyter notebook*, tout en permettant la personnalisation de l'interface. Nos applications d'exploration sur divers corpus autour du cancer démontrent l'intérêt de notre package, qui propose la première interface de programmation Python du package R `MIIC`, pour la reconstruction de réseaux causaux.

8. Les documents des corpus étant en anglais pour cette application, nous rendons compte des termes également en anglais, tels qu'il apparaissent dans les réseaux. La procédure de *stemming* explique la troncature de certains mots.

## Remerciements

Ce travail est soutenu par une subvention de l'Agence Nationale de la Recherche (ANR) (ANR-19-CE23-0002). Il a également reçu la labellisation des pôles de compétitivité *Cap Digital* et *EuroBiomed*. Nous remercions le Isambert lab pour son support concernant le package R `MIIC`.

## Références

[1] Séverine Affeldt and Hervé Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *ACI@ UAI*, pages 1–29, 2015.

[2] Séverine Affeldt, Louis Verny, and Hervé Isambert. 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, pages 149–165. BioMed Central, 2016.

[3] Melissa Ailem, François Role, and Mohamed Nadif. Co-clustering document-term matrices by direct maximization of graph modularity. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1807–1810, 2015.

[4] Melissa Ailem, Aghiles Salah, and Mohamed Nadif. Non-negative matrix factorization meets word embedding. In *SIGIR*, pages 1081–1084, 2017.

[5] Simon Barkow, Stefan Bleuler, Amela Prelić, Philip Zimmermann, and Eckart Zitzler. Bicat : a biclustering analysis toolbox. *Bioinformatics*, 22(10) :1282–1283, 2006.

[6] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data : Diffusion maps. *Proceedings of the national academy of sciences*, 102(21) :7426–7431, 2005.

[7] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[8] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, 2001.

[9] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, 2003.

[10] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD in-*

*ternational conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[11] Kemal Eren, Mehmet Deveci, Onur Küçüktunç, and Ümit V Çatalyürek. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics*, 14(3) :279–292, 2013.

[12] Mickael Febrissy, Aghiles Salah, Melissa Ailem, and Mohamed Nadif. Improving nmf clustering by leveraging contextual relationships among words. *Neurocomputing*, 495 :105–117, 2022.

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, 2008.

[14] Thomas George and Srujana Merugu. A scalable collaborative filtering framework based on co-clustering. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

[15] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6) :3233–3245, 2008.

[16] Gérard Govaert and Mohamed Nadif. *Co-clustering : models, algorithms and applications*. John Wiley & Sons, 2013.

[17] Gérard Govaert and Mohamed Nadif. Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in data analysis and classification*, 12 :455–488, 2018.

[18] Blaise Hanczar and Mohamed Nadif. Ensemble methods for biclustering tasks. *Pattern Recognition*, 45(11) :3938–3949, 2012.

[19] John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337) :123–129, 1972.

[20] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks : The combination of knowledge and statistical data. *Machine learning*, 20 :197–243, 1995.

[21] William J Irvin Jr and Lisa A Carey. What is triple-negative breast cancer ? *European journal of cancer*, 44(18) :2799–2805, 2008.

[22] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9) :939–952, 1982.

[23] Sebastian Kaiser and Friedrich Leisch. A toolbox for bicluster analysis in r. 2008.

[24] Lazhar Labiod and Mohamed Nadif. Co-clustering for binary and categorical data with maximum modularity. In *2011 IEEE 11th international conference on data mining*, pages 1140–1145. IEEE, 2011.

[25] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[26] W. McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4) :93–111, 1954.

[27] Mohamed Nadif and Gérard Govaert. Block clustering of contingency table and mixture model. In *Advances in Intelligent Data Analysis*, pages 249–259, 2005.

[28] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

[29] Valeria Relli, Marco Trerotola, Emanuela Guerra, and Saverio Alberti. Abandoning the notion of non-small cell lung cancer. *Trends in Molecular Medicine*, 25(7) :585–594, 2019.

[30] J Remon, CE Steuer, SS Ramalingam, and E Felip. Osimertinib and other third-generation egfr tki in egfr-mutant nsclc patients. *Annals of Oncology*, 29 :i20–i27, 2018.

[31] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4) :962–1030, 2002.

[32] François Role, Stanislas Morbieu, and Mohamed Nadif. Coclust : a python package for co-clustering. *Journal of Statistical Software*, 88 :1–29, 2019.

[33] Aghiles Salah, Melissa Ailem, and Mohamed Nadif. Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[34] Aghiles Salah and Mohamed Nadif. Directional co-clustering. *Advances in Data Analysis and Classification*, 13 :591–620, 2019.

[35] Nadir Sella, Louis Verny, Guido Uguzzoni, Séverine Affeldt, and Hervé Isambert. Miic online : a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics*, 34(13) :2311–2313, 2018.

[36] Neha Singh, Alexandra Miner, Lauren Hennis, and Sandeep Mittal. Mechanisms of temozolomide resistance in glioblastoma-a comprehensive review. *Cancer drug resistance*, 4(1) :17–43, 2021.

[37] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1) :62–72, 1991.

[38] Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS computational biology*, 13(10) :e1005662, 2017.

[39] Raymond W Yeung. A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37(3) :466–474, 1991.

# Modules dans les Espaces de Robinson

Mikhaël Carmona[1,2], Victor Chepoi[1], Guyslain Naves[1], Pascal Préa[1,2]

[1] Aix-Marseille Université, Université de Toulon, CNRS, LIS
[2] École Centrale de Marseille

{mikhael.carmona, victor.chepoi, guyslain.naves, pascal.prea}@lis-lab.fr

## Résumé

*Une dissimilarité $d$ sur $X$ est* Robinson *s'il existe un ordre total $<$ sur X tel que $x < y < z \Rightarrow d(x,z) \geq \max\{d(x,y), d(y,z)\}$. Les dissimilarités de Robinson ont de nombreuses applications en sériation et classification. Un* mmodule *de $(X,d)$ est un sous-ensemble $M$ de $X$ indistingable depuis l'extérieur de $M$. Pour $p \in X$, un $p$-*copoint *est un mmodule maximal ne contenant pas $p$. Nous présentons quelques propriétés des mmodules et copoints, ainsi qu'un algorithme simple et optimal pour reconnaitre les dissimilarités de Robinson, basé sur ces notions.*

## Mots-clés

*Dissimilarités de Robinson, Sériation, Classification, Modules, Copoints, Diviser-pour-Régner.*

## Abstract

*A dissimilarity on $X$ is* Robinson *if there exists a total order $<$ on $X$ such that $x < y < z \Rightarrow d(x,z) \geq \max\{d(x,y), d(y,z)\}$. Robinson dissimilarities have numerous applications in seriation and classification. An* mmodule *of $(X,d)$ is a subset $M$ of $X$ which is not distinguishable from the outside of $M$. If $p \in X$, a maximal by inclusion mmodule not containing $p$ is a $p$-*copoint*. In this paper, we investigate the structure of mmodules and copoints and use it to design a simple and practical algorithm to recognize of Robinson dissimilarities in optimal time.*

## Keywords

*Robinson dissimilarity, Seriation, Classification, Module, Copoint, Divide-and-conquer.*

# 1 Introduction

Étant donné un ensemble (fini) $X$, une *dissimilarité* $d$ sur $X$ est une fonction symétrique $X \times X \mapsto \mathbf{R}^+$ telle que $d(x,x) = 0$ pour tout $x \in X$. Les dissimilarités de Robinson [14] correspondent à des points sur une droite. Ces dissimilarités se définissent ainsi : $d$ est *Robinson* si il existe un ordre (total), dit *compatible* sur $X$ tel que :

$$x < y < z \Longrightarrow d(x,z) \geq \max\{d(x,y), d(y,z)\} \quad (1)$$

$(X,d)$ est alors appelé un *espace de Robinson*. W.S. Robinson a inventé ces dissimilarités pour résoudre le problème de la sériation en archéologie, et elle sont devenues depuis le modèle standard pour la sériation (qui a de nombreuses autres applications), en particulier parce qu'il existe plusieurs caractérisations intéressantes de ces dissimilarités (une dissimilarité est Robinson si et seulement si ses boules/2-boules/clusters forment un hypergraphe d'intervalles [1]). De plus, les dissimilarités de Robinson sont équivalentes aux pyramides [7, 8]. Il faut remarquer que le modèle de Robinson est plus "souple" que les line-distances (définies par, étant donné $x_1, \ldots, x_n \in \mathbf{R}$, $d(x_i, x_j) = |x_i - x_j|$).
En raison de l'importance de ces dissimilarités, il existe déjà plusieurs algorithmes pour reconnaitre les dissimilarités de Robinson, par exemple [12, 6, 15, 13, 11, 3]. Il faut remarquer que l'algorithme de [13], si il est optimal, est complexe et assez délicat à programmer. La complexité des autres algorithmes varie entre $O(n^2 \log n)$ et $O(n^4)$.

Dans un graphe non orienté, un *module* est un ensemble $M$ de sommets qui ont tous le même voisinage à l'extérieur de $M$. Cette notion, très classique en théorie des graphes, a été généralisée aux relations asymétriques [5, 9, 10].

Nous présentons ici une généralisation/adaptation de la notion de modules aux dissimilarités et nous construisons, à partir de là, un algorithme simple et optimal pour reconnaitre les dissimilarités de Robinson.
Dans la suite, $X$ est un ensemble fini ($|X| = n$) et $d$ une dissimilarité (pas forcément Robinson) sur $X$.

# 2 Mmodules et Copoints

Un *mmodule* (ou *module métrique*) est un sous-ensemble $M$ de $X$ tel que $\forall z \in X \setminus M, x, y \in M, d(x,z) = d(y,z)$. Les singletons, $X$ et $\emptyset$ sont des mmodules, dits *triviaux*. On note $\mathcal{M}(X,d)$ l'ensemble des mmodules de $(X,d)$. Un mmodule *maximal* est un mmodule différent de $X$ et maximal pour l'inclusion.

**Proposition 1** *L'ensemble $\mathcal{M} = \mathcal{M}(X,d)$ a les propriétés suivantes :*

1. *$M_1, M_2 \in \mathcal{M} \Longrightarrow M_1 \cap M_2 \in \mathcal{M}$ ;*
2. *si $M \in \mathcal{M}$ et $M' \subset M$, alors $M' \in \mathcal{M}$ si et seulement si $M'$ est un mmodule de $(M,d)$ ;*
3. *si $M_1, M_2 \in \mathcal{M}$ et $M_1 \cap M_2 \neq \emptyset$, alors $M_1 \cup M_2 \in \mathcal{M}$, de plus, si $M_1 \setminus M_2 \neq \emptyset$ et $M_2 \setminus M_1 \neq \emptyset$, alors $M_1 \setminus M_2, M_2 \setminus M_1, M_1 \triangle M_2 \in \mathcal{M}$ ;*

4. *l'union $M_1 \cup M_2$ de deux mmodules maximaux qui s'intersectent est $X$ ;*
5. *si $M_1$ et $M_2$ sont des mmodules maximaux disjoints et $M$ un mmodule non trivial contenu dans $M_1 \cup M_2$, alors soit $M \subset M_1$, soit $M \subset M_2$ ;*
6. *si $M_1, M_2 \in \mathcal{M}$ et $M_1 \cap M_2 = \emptyset$, alors $d(u, v) = d(u', v')$ pour tous (pas forcément distincts) points $u, u' \in M_1$ et $v, v' \in M_2$ ;*

**Proposition 2** *L'ensemble des mmodules maximaux de $X$ est soit une partition, soit une co-partition de $X$.*

Un $\cup\cap$-arbre sur $X$ est un arbre $\mathcal{T}$ dont les feuilles sont indexées par $X$ et les nœuds internes par $\cup$ ou $\cap$ qui représente un sous-ensemble $\mathcal{S}(\mathcal{T})$ de $2^X$ défini par :

1. Si $\alpha$ est un sommet de $\mathcal{T}$, l'ensemble des feuilles de $\alpha$ est dans $\mathcal{S}(\mathcal{T})$.
2. Si $\alpha$ est un $\cap$-sommet de $\mathcal{T}$, l'ensemble des feuilles de n'importe quel ensemble d'enfants de $\alpha$ est dans $\mathcal{S}(\mathcal{T})$.

**Proposition 3** *Il existe un unique $\cup\cap$-arbre $\mathcal{T}_\mathcal{M}$ sur $X$, dit* arbre de mmodules *de $(X, d)$, tel que $\mathcal{M}(X, d) = \mathcal{S}(\mathcal{T}_\mathcal{M})$.*

Étant donné $p \in X$, un *copoint* attaché à $p$, ou *$p$-copoint* est un mmodule $C$, maximal pour l'inclusion, qui ne contient pas $p$. Le point $p$ est le *point d'attachement* de $C$. On note $\mathcal{C}_p$ l'ensemble des $p$-copoints union $\{\{p\}\}$. L'espace quotient $(\mathcal{C}_p, \widehat{d})$ est défini par $\widehat{d}(C, C') := d(u, v)$ pour $u \in C$, $v \in C'$, $C \neq C'$ et $\widehat{d}(C, C) = 0$.

**Proposition 4** *Pour tout $p \in X$, $\mathcal{C}_p$ est une partition de $X$.*

Si $\mathcal{C}_p = \{x : x \in X\}$, on dit que $\mathcal{C}_p$ est *$p$-trivial* ; Si $\mathcal{C}_p = \{\{p\}, X \setminus \{p\}\}$, $\mathcal{C}_p$ est *co-trivial*, dans ce cas, $X$ est conique d'apex $p$.

**Proposition 5** *$(\mathcal{C}_p, \widehat{d})$ est $\{p\}$-trivial.*

# 3 Copoints dans les dissimilarités de Robinson

Étant donné un point $p$ et un copoint $C \in \mathcal{C}_p$, on dit que $C$ est *compact* si $\mathrm{diam}(C) < d(p, C)$, *large* si $\mathrm{diam}(C) > d(p, C)$ et *limite* si $\mathrm{diam}(C) = d(p, C)$.

**Proposition 6** *Si $(X, d)$ est un espace de Robinson et $C$ un copoint de $(X, d)$, alors, pour tout ordre compatible :*
*Si $C$ est compact, $C$ est un interval,*
*Si $C$ est large, $C$ est constitué de deux intervalles,*
*Si $C$ est limite, $C$ est constitué de un ou deux intervalles.*

**Proposition 7** *Soit $(X, d)$ un espace de Robinson, il existe un ordre compatible $<$ tel que, quand $X$ est ordonné selon $<$, tout copoint non large est fait d'un seul intervalle.*

Étant donné espace de Robison $(X, d)$, un ordre compatible $<$ et un point $p$ de $X$, un *ordre de $p$-proximité* est un ordre $\prec$ sur $\mathcal{C}_p$ tel que, si $C_1 \prec C_2$ :

1. $d(p, C_1) \leq d(p, C_2)$
2. $\forall x \in C_1, y \in C_2$, on n'a ni $p < y < x$ ni $x < y < p$ (*ie.* $C_2$ n'est pas entre $p$ et $C_1$).

Si $\prec$ est un ordre de $p$-proximité pour tout ordre compatible, alors $\prec$ est *universel*.

**Proposition 8** *L'algorithme 1 recursiveRefine, lancé avec $(p, \{p\}, X \setminus \{p\}, \emptyset)$ construit $\mathcal{C}_p$ et l'ordonne, si $(X, d)$ est Robinson, selon un ordre de $p$-proximité universel, et ce, sans que l'on connaisse, au départ, un ordre compatible.*

**Algorithme 1 :** recursiveRefine$(p, In, S, Out)$

**Input :** Un espace de Robinson $(X, d)$ (implicite), un point $p \in X$, un ensemble $S \subseteq X$, deux sous-ensembles disjoints $In, Out \subseteq X \setminus S$ (pivots internes et externes).
**Output :** Une partition ordonnée $[S^\star_1, S^\star_2, \ldots, S^\star_{k^\star}]$ of $S \setminus \{p\}$ selon un ordre de $p$-proximité (partiel).

```
begin
    if In ∪ Out = ∅ then
        return [S] ;
    Let q ∈ In ∪ Out ;
    [S_1, ..., S_m] ← refine(q, S) ;
    /* refine partitionne S en [S_1, ..., S_m]
       de telle sorte que les d(q, S_i)
       soient bien définis et
       i < j ⇒ d(q, S_i) < d(q, S_j)                */
    if q ∈ Out then
        α ← min({j ∈ {1, ..., m} : d(S_j, q) >
          d(p, q)} ∪ {m + 1}) ;
        [S'_1, ..., S'_m] ←
          [S_{α−1}, S_{α−2}, ..., S_1, S_α, S_{α+1}, ..., S_m] ;
    else
        [S'_1, ..., S'_m] ← [S_1, ..., S_m] ;
    forall i ∈ {1, ..., m} do
        In_i ← concat(S'_1, ..., S'_{i−1}, In \ {q}) ;
        Out_i ← concat(S'_{i+1}, ..., S'_m, Out \ {q}) ;
        T_i ← recursiveRefine(p, In_i, S'_i, Out_i) ;
    return concat(T_1, ..., T_m)
```

**Proposition 9** *Sans compter le temps pris par les appels à la fonction refine, l'algorithme 1 recursiveRefine, avec $(p, [q], S, [])$ comme entrée et $[S^\star_1, \ldots, S^\star_{k^\star}]$ comme sortie, tourne en temps $O(|S|^2 - \sum_{j=1}^{k^\star} |S^\star_j|^2)$.*

# 4 Ordres compatibles dans un espace de Robinson conique

Un espace $(X, d)$ est *$\delta$-conique* ou *conique* d'*apex* $p$ si $\forall x \neq p$, $d(x, p) = \delta$. L'importance des espaces coniques vient de ce que si $C$ est un $p$-copoint, alors $C \cup \{p\}$ est conique d'apex $p$. La proposition 10 correspond au cas d'un

copoint compact ou limite, la proposition 11 au cas d'un copoint large. Quand un $p$-copoint $C$ est compact ou limite, l'algorithme 2 separateIfSeparable laisse $C$ en un seul morceau et met $p$ avant le copoint ; quand le $p$-copoint est large, il partitionne $C$ en deux sous-ensembles et met $p$ à un "bon endroit" (il peut en exister plusieurs).

**Proposition 10** *Si $X$ est $\delta$-conique d'apex $p$, $X \setminus \{p\}$ est Robinson d'ordre compatible $x_1 < \ldots < x_{n-1}$ et si $d(x_1, x_{n-1}) \leq \delta$, alors $X$ est Robinson d'ordre compatible $p < x_1 < \ldots < x_{n-1}$.*

**Proposition 11** *Si $X$ est $\delta$-conique d'apex $p$, $X \setminus \{p\}$ est Robinson d'ordre compatible $x_1 < \ldots < x_{n-1}$ et si $d(x_1, x_{n-1}) > \delta$, alors $X$ est Robinson si et seulement si il existe $k \in \{1, \ldots, n-2\}$ tel que, pour tout $i, j \in \{1, \ldots, n-1\}$ :*

- $i, j \leq k \vee i, j > k \Longrightarrow d(x_i, x_j) \leq \delta$
- $i \leq k < j \Longrightarrow d(x_i, x_j) \geq \delta$

*Dans ce cas, $x_1 < \ldots < x_k < p < x_{k+1} < \ldots < x_{n-1}$ est un ordre compatible de $X$.*

**Proposition 12** *L'algorithme 2 separateIfSeparable rend, si $(X, d)$ est Robinson, un ordre compatible en $O(n)$.*

**Algorithme 2 :** separateIfSeparable$(p, X')$

**Input :** Un espace de Robinson $(X, d)$ (implicite), conique d'apex $p$, avec $X' = X \setminus \{p\}$ trié selon un ordre compatible $<'$.
**Output :** $X'$ ou une bipartition de $X'$, selon que $\mathrm{diam}(X') \leq d(p, X')$ ou pas.

$x_* \leftarrow$ minimum de $(X', <')$ ;
$x^* \leftarrow$ maximum de $(X', <')$ ;
$\Delta \leftarrow d(x_*, x^*)$ ;
$\delta \leftarrow d(p, x_*)$ ;
**if** $\Delta \leq \delta$ **then**
  **return** $[(x_*, X')]$
**forall** $y \in X' \setminus \{x^*\}$ *(en ordre croissant)* **do**
  $z \leftarrow$ l'element suivant $x$ pour $<'$ ;
  **if** $(d(x_*, y) \leq \delta) \wedge (d(z, x^*) \leq \delta) \wedge (d(y, z) \geq \delta)$ **then**
    **return** $[(x_*, \{u \in X' : u \leq' y\}), (x^*, \{u \in X' : z \leq' u\})]$

À partir de l'espace quotient $(\mathcal{C}_p, \widehat{d})$, on peut définir le *quotient étendu* $(\mathcal{C}_p^*, d^*)$ de la manière suivante :

- Si $C \in \mathcal{C}_p$ n'est pas large, alors $C \in \mathcal{C}_p^*$,
- Si $C \in \mathcal{C}_p$ est large, il est divisé par separateIfSeparable en deux *semi-copoints* $C^1$ et $C^2$ et on a $C^1, C^2 \in \mathcal{C}_p^*$ avec $d^*(C^1, C^2) = \mathrm{diam}(C)$. Pour toutes les autres paires d'éléments de $\mathcal{C}_p^*$, $d^* = \widehat{d}$.

Si $(X, d)$ est Robinson, il en est de même pour $(\mathcal{C}_p^*, d^*)$.

# 5 Un algorithme récursif pour reconnaitre les dissimilarités de Robinson

À partir de ce qui précède, on peut déterminer si un espace $(X, d)$ est Robinson de la manière suivante :

1. Prendre un point $p$ et construire $\mathcal{C}_p$,
2. Ordonner $\mathcal{C}_p$ selon un ordre de $p$-proximité $\prec$,
3. Construire un ordre compatible $<$ à partir de $\prec$,
4. Construire $(\mathcal{C}^*, d^*)$,
5. Traiter récursivement tous les copoints et semi-copoints, on obtient une suite d'ordres $(<_i)$,
6. Concaténer les ordres $<_i$ selon $<$,
7. Vérifier si l'ordre obtenu est compatible.

Le point 3 (construction d'un ordre compatible à partir d'un ordre de $p$-proximité) peut se faire par l'algorithme 3 partitionSort, où :

- $\oplus$ concatène deux listes
- $\bullet$ ajoute un élément en tête de liste

**Algorithme 3 :** partitionSort$(p, X)$

**Input :** Un espace de Robinson $(X, d)$ ($d$ est implicite), un point $p \in X$, $X \setminus \{p\}$ est donné selon un ordre de $p$-proximité.
**Output :** $X$ selon un ordre compatible.

$L \leftarrow []$ ; $R \leftarrow []$ ; $Undecided \leftarrow$ reverse$(X \setminus \{p\})$ ;
**forall** $q \in X \setminus \{p\}$ *in decreasing order* **do**
  **if** $q \in Undecided$ **then**
    **Au choix** : $R \leftarrow q \bullet R$ **ou** $L \leftarrow q \bullet L$ ;
    $Undecided \leftarrow Undecided \setminus \{q\}$ ;
  $Skipped \leftarrow []$ ;
  **forall** $x \in Undecided$ **do**
    **if** $d(x, q) = d(p, q)$ **then**
      $Skipped \leftarrow x \bullet Skipped$
    **else**
      **if** $(d(x, q) < d(p, q) \wedge q \in L) \vee (d(x, q) > d(p, q) \wedge q \in R)$ **then**
        $L \leftarrow x \bullet L$ ;
        $R \leftarrow Skipped \oplus R$
      **else**
        $R \leftarrow x \bullet R$ ;
        $L \leftarrow Skipped \oplus L$
      $Skipped \leftarrow []$
  $Undecided \leftarrow$ reverse$(Skipped)$
**return** reverse$(L) \oplus [p] \oplus R$

**Proposition 13** *Si $(X, d)$ est Robinson et est ordonné selon selon un ordre de p-proximité, alors l'algorithme 3 partitionSort construit un ordre compatible en $O(n^2)$.*

On peut maintenant écrire formellement l'algorithme 4 findCompatibleOrder, qui calcule, si $(X, d)$ est Robinson, un

**Algorithme 4 :** findCompatibleOrder$(X)$

**Input :** Un espace de Robinson $(X, d)$ ($d$ est implicite).
**Output :** Un ordre compatible de $X$.
**if** $X = \emptyset$ **then**
  **return** []
Let $p \in X$ ;
$X' \leftarrow X \setminus \{p\}$ ;
$[C_1, \ldots, C_k] \leftarrow$ recursiveRefine$(p, [p], X', [])$ ;
$repCopoints \leftarrow []$ ;
**forall** $i \in \{1, \ldots, k\}$ *en ordre décroissant* **do**
  $C'_i \leftarrow$ findCompatibleOrder$(C_i)$ ;
  $repCopoints \leftarrow$ separateIfSeparable$(p, C'_i) \oplus repCopoints$
$[(x_1, T_1), \ldots, (x_{k'}, T_{k'})] \leftarrow repCopoints$ ;
$[x_{\sigma(1)}, \ldots, p, \ldots, x_{\sigma(k')}] \leftarrow$ partitionSort$(p, [x_1, \ldots, x_{k'}])$ ;
**return** concatenate$(T_{\sigma(1)}, \ldots, [p], \ldots, T_{\sigma(k')})$

ordre compatible de $X$. La fonction concatenate concatène un suite de listes.
Pour déterminer la complexité de l'algorithme 4 findCompatibleOrder, nous avons besoin du lemme suivant (qui donne la complexité globale des appels à la fonction refine) :

**Lemme 1** *Si $T : \mathbf{N} \to \mathbf{N}$ satisfait la relation de récurrence $T(n) \leq \sum_{i=1}^{k} T(n_i) + n \log k$ pour toute partition $\sum_{i=1}^{k} n_i = n$ de $n$ en $k \geq 2$ entiers positifs, alors $T(n) = O(n^2)$.*

**Théorème 1** *L'algorithme 4 findCompatibleOrder, avec un espace de Robinson $(X, d)$ en entrée, calcule un ordre compatible de $X$ en $O(n^2)$.*

Il suffit, pour déterminer si une dissimilarité est Robinson, de vérifier que l'ordre obtenu est bien compatible, ce qui revient à vérifier que les lignes et colonnes d'une matrice sont croissantes quand on s'éloigne de la diagonale principale.

**Algorithme 5 :** RobinsonRecognition$(X, d)$

**Input :** Une dissimilarité $d$ sur un ensemble $X$.
**Output :** Un booléen qui indique si $(X, d)$ est Robinson.
$< \leftarrow$ findCompatibleOrder$(X)$ ;
/* $X = \{x_1, \ldots, x_n\}$ avec $x_1 < \ldots < x_n$ */
**forall** $k \in \{1, \ldots, n-1\}$ **do**
  **forall** $i \in \{1, \ldots, n-k\}$ **do**
    $j \leftarrow i + k$ ;
    **if** $d(x_i, x_j) < \max\{d(x_i, x_{j-1}), d(x_{i+1}, x_j)\}$ **then**
      **return** ***False***
**return** ***True***

**Corollaire 1** *L'algorithme 5 RobinsonRecognition détermine si une dissimilarité est Robinson en temps optimal $O(n^2)$.*

# 6 Conclusion

Nous avons présenté ici les mmodules et les copoints ainsi que leurs principales propriétés. De ces notions, nous avons de plus pu dériver un algorithme simple et optimal pour la reconnaissance des dissimilarités de Robinson.
Il existait déjà un algorithme optimal pour reconnaitre les dissimilarités de Robinson [13]. Mais par rapport à cet algorithme, celui-ci présente plusieurs avantages :

- Il est beaucoup plus simple à comprendre et à programmer, ce qui est un avantage en soi et permet en plus des implémentations plus rapides (une matrice $1000 \times 1000$ se traite en une demi-seconde et une matrice $10000 \times 10000$ en moins d'une minute).
- Il peut être rendu *certifiant* : il est en effet possible de l'adapter, sans changer la complexité, de telle sorte que, si $(X, d)$ n'est pas Robinson, RobinsonRecognition rende un certificat, vérifiable en $O(n)$ qui assure que $(X, d)$ n'est pas Robinson. Si $(X, d)$ est Robinson, l'ordre compatible constitue un certificat que l'on peut vérifier en $O(n^2)$. Ces certificats sont optimaux : pour les dissimilarités de Robinson, il est impossible de certifier une réponse négative en $o(n)$ (puisqu'il existe, pour tout $n$, des espaces à $n$ éléments qui ne sont pas Robinson mais dont tous les sous-espaces le sont) et une réponse positive en $o(n^2)$ (puisque, en changeant quatre valeurs, on peut transformer une matrice de Robinson de n'importe quelle taille en une matrice qui ne l'est pas).
- L'algorithme de [13] utilise des PQ-trees [2], une structure de données qui, pour des dissimilarités, ne peut être définie que pour les dissimilarités de Robinson. L'algorithme présenté ici utilise les mmodules et les copoints, que l'on peut, eux, définir pour toutes les dissimilarités. Ces outils peuvent donc être utilisés pour autre chose que reconnaitre les dissimilarités de Robinson :
  - Il est possible de caractériser les ultramétriques par leur arbre de mmodules ou leurs copoints.
  - Ils ont déjà été à l'origine d'un algorithme optimal pour la sériation circulaire
  - ...

# Remerciements

Cette recherche a été soutenue par l'ANR via le projet DISTANCIA (ANR-17 CE40-0015) et par une aide du gouvernement français au titre du Programme Investissements d'Avenir Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (Institut Archimède AMX-19-IET-009).

# Références

[1] J-P. Barthélemy & F. Brucker, *Éléments de classification : Aspects combinatoires et algorithmiques*, Hermès Science Publications, 2007.

[2] K.S. Booth & G.S. Lueker, Testing for the Consecutive Ones Property, Interval Graphs and Graph Planarity Using PQ-Tree Algorithm, *Journal of Computer and System Sciences* 13, 335–379, 1976.

[3] M. Carmona, V. Chepoi, G. Naves & P. Préa, Two simple but efficient algorithms to recognize Robinson dissimilarities, *IFCS*, Porto, 2022.

[4] M. Carmona, V. Chepoi, G. Naves & P. Préa, A simple and optimal algorithm for strict circular seriation, *SIAM J. on Mathematics of Data Science*, to appear.

[5] M. Chein, M. Habib & M. Maurer, Partitive hypergraphs, *Discr. Math.* 37, 35–50, 1981.

[6] V. Chepoi & B. Fichet, Recognition of Robisonian dissimilarities, *J. of Classification* 18, 159–183, 1997.

[7] E. Diday, Orders and overlapping clusters by pyramids, in *Multidimensionnal Data Analysis*, J. de Leeuw, W. Heiser, J. Meulman and F. Critchley Eds., 201–234, DSWO, 1986.

[8] C. Durand, & B. Fichet, One-to-one correspondences in pyramidal representation : an unified approach, in *Classification and Related Methods of Data Analysis*, H.H. Bock Ed., 85–90, North-Holland, 1988.

[9] A. Ehrenfeucht & G. Rozenberg, Theory of 2-structures, part I : Clans, basic subclasses and morphisms, *Theor. Comput. Sci.* 70, 277–303, 1990.

[10] A. Ehrenfeucht & G. Rozenberg, Theory of 2-structures, part II : Representations through tree families, *Theor. Comput. Sci.* 70, 305–342, 1990.

[11] M. Laurent & M. Seminaroti, Similarity-first search : a new algorithm with applications to Robinson matrix recognition, *SIAM J. Discr. Math.* 31, 1765–1800, 2017.

[12] B. Mirkin & S. Rodin, *Graphs and genes*, Springer-Verlag, 1984.

[13] P. Préa & D. Fortin, An optimal algorithm to recognize Robinsonian dissimilarities, *J. of Classification* 31, 351–385, 2014.

[14] W.S. Robinson, A method for chronologically ordering archeological deposits, *American Antiquity* 16, 293–301, 1951.

[15] M. Seston, A simple algorithm to recognize Robinsonian dissimilarities, in *Proceedings in Computational Statistics*, P. Brito Ed., 241–248, Physica-Verlag, 2008.

# PAC-Bayesian bornes pour l'adaptation de domaine non supervisé dans un cadre d'apprentissage multi-vue

Mehdi Hennequin[1,2], Khalid Benabdeslem[2], Haytham Elghazel[2]

[1] Galilé Group, 28 Bd de la République, 71100 Chalon-sur-Saône, France
[2] Université Lyon 1, LIRIS, UMR CNRS 5205, F-69622, France

{mehdi.hennequin,khalid.benabdeslem,haytham.elghazel}@univ-lyon1.fr

## Résumé

*Dans cet article, nous explorons le cadre PAC-Bayesien pour proposer des bornes multi-view domain adaptation dans une situation de classification binaire sans étiquettes cibles (adaptation de domaine non supervisée). À la connaissance des auteurs, aucune limite de généralisation n'a été proposée pour adaptation de domaine multi-vues non supervisée. Dans un premier temps, nous proposons une nouvelle distance adaptée à notre contexte pour mesurer la distance entre les distributions. Dans un deuxième temps, nous introduisons un théorème Pac-Bayesien général tel que proposé dans [9, 10, 2] pour estimer la nouvelle distance introduite. Ensuite, nous proposons une spécialisation de notre théorème général pour les trois versions des théorèmes PAC-Bayesian [18, 21, 15, 5] en suivant les mêmes principes que Germain et al. [9, 7, 2, 11]. Enfin, nous proposons une limite d'adaptation de domaine PAC-Bayesian dans le cadre multi-vues.*

## Mots-clés

*PAC-Bayesian, Domain Adaptation, Multi-view Learning.*

## Abstract

*This paper presents a series of new results for domain adaptation in the multi-view learning setting. The incorporation of multiple views in the domain adaptation was paid little attention in the previous studies. In this way, we propose an analysis of generalization bounds with Pac-Bayesian theory to consolidate the two paradigms, which are currently treated separately. Firstly, building on previous work by Germain et al. [6, 7], we adapt the distance between distribution proposed by Germain et al. for domain adaptation with the concept of multi-view learning. Thus, we introduce a novel distance that is tailored for the multi-view domain adaptation setting. Then, we give Pac-Bayesian bounds for estimating the introduced divergence. Finally, we compare the different new bounds with the previous studies.*

## Keywords

*Example, model, template.*

## 1 Introduction

Predictive models must fit data with different distributions. In fact, in the theory of statistical learning, the strong hypothesis that training and test data are to be drawn from the same probability distribution. However, this assumption is often too restrictive to be used in practice or in many real-life applications. Indeed, a hypothesis is learned and deployed in different and significantly changing environments. Due to that, we obtain a shift in the data distributions. A typical solution for addressing this issue is to retrain the models. Nonetheless, this process of retraining can result in both time and financial expenses. Thus, we need to design methods for adapting a model from learning (source) data to test (target) data. In the field of machine learning, this scenario is commonly known as *domain adaptation (DA)* or *covariate shift* [19]. Essentially, DA techniques aim to address the challenge of learning when the learning task is the same but the domains exhibit variations in their feature spaces or marginal conditional probabilities.
On the other hand, data can be expressed through multiple independent feature sets, as stated in [25]. As a result, the data can be partitioned into independent groups, known as views [25]. In domain adaptation, these views are usually merged into a single view to align with the learning objective. Nonetheless, this process of merging can potentially result in negative transfer, as highlighted in [27], whereby the integration of each view's unique statistical characteristics could lead to the introduction of unwanted data or knowledge. We find little research on *multi-view domain adaptation (MVDA)* [26, 13, 20] where considerable attention has been paid to algorithm, while analysis of generalization bound remains largely understudied. In this way, we propose a theoretical analysis by means of Pac-Bayesian theory, with the aim of unifying the two paradigms that have conventionally been treated as separate entities [6, 7, 11].

## 2 Related Works

In this section, we present theoretical studies of multi-view learning and domain adaptation related to the Pac-Bayesian theory. First, we introduce the theoretical concepts needed for the following sections. In a second phase, we recall the general PAC–Bayesian generalization bounds in the

setting of binary classification. Then, we present the work done by on the PAC–Bayesian domain adaptation. Finally, we present the works achieved to Pac-Bayesian theory and multi-view learning.

## 2.1 Notation and Assumptions

This section introduces the definitions and concepts needed for the following sections. Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$, denote respectively input space of dimension $d$ and output space. In the *scenario of unsupervised multi-view domain adaptation (UMVDA)* we consider, the learner receives two samples : a labeled sample from a source domain $\mathbb{D}_S$, defined by a distribution $Q$ over $\mathcal{X} \times \mathcal{Y}$; and unlabelled sample according to the target domain $\mathbb{D}_T$, defined by a distribution $P$ over $\mathcal{X} \times \mathcal{Y}$; $Q_{\mathcal{X}}, P_{\mathcal{X}}$ being the respective marginal distributions over $\mathcal{X}$. We denote by $\mathbb{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ the labeled sample of size $m$ received from the source domain, which is drawn i.i.d. from $Q$. In addition, we consider that the data instances can be represented or partitioned in $V$ different views. More formally, for $v \in \{1, \ldots, V\}$, and $V \geq 2$ is the number of views of not-necessarily the same dimension. Throughout the paper, we will abbreviate $v \in \{1, \ldots, V\}$ with $v \in [\![\mathcal{V}]\!]$. The labeled samples multi-view is defined as follows, $\forall\ v \in [\![\mathcal{V}]\!],\ \mathbb{S} = \{(\boldsymbol{x}_i^v, y_i)\}_{i=1}^m \in (\mathcal{X}^v \times \mathcal{Y})^m$, with $\{\boldsymbol{x}_i^v\}_{i=1}^m$ supposed to be drawn i.i.d. according to distribution $Q$. Note that the multi-view observations $\{\boldsymbol{x}_i\}_{i=1}^m = \{(\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^v)\}_{i=1}^m$ belong to a multi-view input set $\mathcal{X} = \mathcal{X}^1 \times, \ldots, \times \mathcal{X}^v$, with $\mathcal{X}^v \in \mathbb{R}^{d_v}$, and $d_v$ denote the dimension of the $v^{th}$ view, where $d = d_1 \times, \ldots, \times d_v$. In the same way, we define unlabeled samples from the target domain, $\forall\ v \in [\![\mathcal{V}]\!], \mathbb{T}_{\mathcal{X}} = \{\boldsymbol{x}_i'^v\}_{i=1}^n$ of size $n$ drawn i.i.d. according to $P_{\mathcal{X}}$ (note that, $\mathbb{T} = \{(\boldsymbol{x}_i'^v, y_i)\}_{i=1}^n$ drawn i.i.d. according to $P$). In our context, we consider that we have no labels in the target domain, however we have prior knowledge about the views in both domains.

We define a hypothesis class $\mathcal{H}$ of hypotheses $h : \mathcal{X} \to \mathcal{Y}$ . Besides, for the concept of multi-view learning, we consider for each view $v \in [\![\mathcal{V}]\!]$, a set $\mathcal{H}_v$ of hypothesis $h : \mathcal{X}_v \to \mathcal{Y}$. The expected source risk or true source risk of $h \in \mathcal{H}$ over the distribution $Q$, are the probability that h errs on the entire source domain, $\mathcal{R}_Q(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q}\big[\mathcal{L}_{0\text{-}1}\big(h(\boldsymbol{x}), y\big)\big]$, where $\mathcal{L}_{0\text{-}1}(a, b) = \mathbb{I}\big[a \neq b\big]$ is the 0-1-loss function and where $\mathbb{I}[a \neq b]$ is the indicator function which returns 1 if $a \neq b$ and 0 otherwise. The extension of the 0-1-loss to real-valued voters (of the form $f : \mathcal{X} \to [-1, 1]$) is given by the following definition, $\mathcal{L}_{0\text{-}1}\big(f(x), y\big) = \mathbb{I}[y.f(x) \neq 0]$. For any two functions $(h, h') \in \mathcal{H}$, we denote by $\mathcal{R}_{Q_{\mathcal{X}}}(h, h')$ the expected disagreement of $h(x)$ and $h'(x)$, which measures the probability that $h$ and $h'$ do not agree on the entire marginal distributions $Q_{\mathcal{X}}$ over $\mathcal{X}$, $\mathcal{R}_{Q_{\mathcal{X}}}(h, h') = \mathbb{E}_{\boldsymbol{x}\sim Q_{\mathcal{X}}}\big[\mathcal{L}_{0\text{-}1}\big(h(\boldsymbol{x}), h'(\boldsymbol{x})\big)\big]$. The empirical source risk $\mathcal{R}_{\mathbb{S}}(h)$ for a given hypothesis $h \in \mathcal{H}$ and a training sample $\mathbb{S} = \{(x_i, y_i)\}_{i=1}^m$ where each example is drawn i.i.d. from $Q$ is defined as, $\mathcal{R}_{\mathbb{S}}(h) = \frac{1}{m}\sum_{i=1}^m \mathcal{L}_{0\text{-}1}\big(h(\boldsymbol{x}_i), y_i\big)$. In the same way, we define the empirical source disagreement by $\mathcal{R}_{\mathbb{S}_{\mathcal{X}}}(h, h') = \frac{1}{m}\sum_{i=1}^m \mathcal{L}_{0\text{-}1}\big(h(\boldsymbol{x}_i), h'(\boldsymbol{x}_i)\big)$, where $\mathbb{S}_{\mathcal{X}} = \{(x_i)\}_{i=1}^m$ where each example is drawn i.i.d. from $Q_{\mathcal{X}}$ . The expected target risk $R_P(\cdot)$ over $P$, the expected target disagreement $R_{P_{\mathcal{X}}}(\cdot, \cdot)$ over $P_{\mathcal{X}}$, the empirical target risk $R_{\mathbb{T}}(\cdot)$ over $P$, the empirical target disagreement $R_{\mathbb{T}_{\mathcal{X}}}(\cdot, \cdot)$ over $P_{\mathcal{X}}$ are defined in a similar way.

## 2.2 Simple Pac-Bayesian Bounds

The PAC-Bayesian approach abbreviated Pac-Bayes is analysis techniques of generalization in the theory of statistical learning. PAC-Bayes inequalities were introduced by [22], and [17, 18]; and further formalised [5, 3, 4] and other (see [12] for a recent survey and [1] for an introduction to the field). It provides PAC (probably approximately correct, Valiant, 1984) generalization bounds by expressing a trade-off between the empirical risk on the training set and a measure of complexity of the predictors class as a weighted majority vote over a set of functions from the hypothesis space $\mathcal{H}$.

In this section, we recall the general PAC–Bayesian generalization bounds in the setting of binary classification with the 0-1-loss defined in the above section. To derive such a generalization bound, one assumes a prior distribution $\mathcal{P}$ over $\mathcal{H}$, which models an a priori belief on the hypothesis from $\mathcal{H}$ before the observation of the training sample $\mathbb{S} \sim Q^m$. Given the training sample $\mathbb{S}$, the learner aims at finding a posterior distribution $\mathcal{Q}$ over $\mathcal{H}$ that leads to a well-performing $\mathcal{Q}$-weighted majority vote $\mathcal{B}_{\mathcal{Q}}$ (also called the bayes classifier) defined as :

$$\mathcal{B}_{\mathcal{Q}}(\boldsymbol{x}) = \operatorname{sign}\Big[\mathop{\mathbb{E}}_{h\sim\mathcal{Q}} h(x)\Big]. \tag{1}$$

We want to learn $\mathcal{Q}$ over $\mathcal{H}$ such that it minimizes the true risk $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}})$ of the $\mathcal{Q}$-weighted majority vote. However, the risk of $B_{\mathcal{Q}}$ is known to be NP-hard, therefore PAC–Bayesian generalization bounds do not directly focus on the risk of $\mathcal{B}_{\mathcal{Q}}$. Instead, it gives an upper bound over the expectation over $\mathcal{Q}$ of all the individual hypothesis true risk called the expected/true Gibbs risk :

$$\mathcal{R}_Q(G_{\mathcal{Q}}) = \mathop{\mathbb{E}}_{h\sim\mathcal{Q}}\Big[\mathcal{R}_Q(h)\Big]. \tag{2}$$

The expected Gibbs risk is closely related to the deterministic $\mathcal{Q}$-weighted majority vote. Indeed, if $\mathcal{B}_{\mathcal{Q}}(\cdot)$ misclassifies $\boldsymbol{x} \in \mathcal{X}$, then at least half of the classifiers (under measure $\mathcal{Q}$) make a prediction error on $\boldsymbol{x}$. Therefore, we have $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}}) \leq 2\,\mathcal{R}_Q(G_{\mathcal{Q}})$. Another result on the relation between $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}})$ and $\mathcal{R}_Q(G_{\mathcal{Q}})$ know as C-bound ([14]) and defined as :

$$\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}}) \leq 1 - \frac{\big(1 - 2\,\mathcal{R}_Q(G_{\mathcal{Q}})\big)^2}{1 - 2\,d_{Q_{\mathcal{X}}}(\mathcal{Q})}, \tag{3}$$

where $d_{Q_{\mathcal{X}}}(\mathcal{Q})$ corresponds to the expected disagreement between pairs of voters on the marginal distribution $Q_{\mathcal{X}}$ :

$$d_{Q_{\mathcal{X}}}(\mathcal{Q}) = \mathop{\mathbb{E}}_{(h,h')\sim\mathcal{Q}^2}\Big[\mathcal{R}_{Q_{\mathcal{X}}}\big(h, h'\big)\Big]. \tag{4}$$

The authors in [14] observed that in a binary classification context, the expected disagreement $d_{Q_{\mathcal{X}}}(\mathcal{Q})$ is closely related to the notion of expected joint error $e_Q(\mathcal{Q})$ between

pairs of voters, $e_Q(\mathcal{Q})$ :

$$e_Q(\mathcal{Q}) = \mathop{\mathbb{E}}_{(h,h')\sim\mathcal{Q}^2} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim Q}\Big[\mathcal{L}_{0\text{-}1}\big(h(\boldsymbol{x}),y\big)\times\mathcal{L}_{0\text{-}1}\big(h'(\boldsymbol{x}),y\big)\Big]. \tag{5}$$

Indeed, for all distribution $Q$ on $\mathcal{X}\times\mathcal{Y}$ and a distribution $\mathcal{Q}$ on $\mathcal{H}$, we can decompose $\mathcal{R}_Q(G_\mathcal{Q})$ as follows :

$$\mathcal{R}_P(G_\mathcal{Q}) = \frac{1}{2}\, d_{P_\mathcal{X}}(\mathcal{Q}) + e_P(\mathcal{Q}). \tag{6}$$

The PAC-Bayesian theory, suggests that minimizing the expected Gibbs risk $\mathcal{R}_Q(G_\mathcal{Q})$ can be done by minimizing the trade-off between the empirical Gibbs risk $\mathcal{R}_\mathbb{S}(G_\mathcal{Q})$ and Kullback–Leibler divergence minimization $D_{KL}(\mathcal{Q}||\mathcal{P})$. Note that PAC–Bayesian generalization bounds do not directly take into account the complexity of the hypothesis class $\mathcal{H}$, but measure the deviation between the prior distribution $\mathcal{P}$ and the posterior distribution $\mathcal{Q}$ on $\mathcal{H}$ through the Kullback–Leibler divergence. In the literature, we find three main PAC-Bayesian bound proposed by [18] ; [21, 15] ; [5].

## 2.3 Analysis of Domain Adaptation Pac-Bayesian Bounds

In this section, we recall the work done by [6, 7] on how the PAC–Bayesian theory can help to theoretically understand domain adaptation through the weighted majority vote learning point of view.

The first Pac-Bayesian generalization bound for domain adaptation was introduced in [6]. The authors defined a divergence measure that follows the idea of C-bound equation 3. Thus, Germain et al. underlined that the domains $\mathbb{D}_\mathcal{S}$ and $\mathbb{D}_\mathcal{T}$ are close according to $\mathcal{Q}$ if the expected disagreement over the two domains tends to be close. More formally, if $\mathcal{R}_Q(G_\mathcal{Q})$ and $\mathcal{R}_P(G_\mathcal{Q})$ are similar, then and $\mathcal{R}_Q(\mathcal{B}_\mathcal{Q})$ and $\mathcal{R}_P(\mathcal{B}_\mathcal{Q})$ are similar when $d_{Q_\mathcal{X}}(\mathcal{Q})$ and $d_{P_\mathcal{X}}(\mathcal{Q})$ are also similar. In this way, the authors introduced the following domain disagreement pseudometric [1].

**Definition 1** *Let $\mathcal{H}$ be a hypothesis class. For any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any distribution $\mathcal{Q}$ on $\mathcal{H}$, the domain disagreement $dis_\mathcal{Q}(Q_\mathcal{X},P_\mathcal{X})$ between $Q_\mathcal{X}$ and $P_\mathcal{X}$ is defined by :*

$$dis_\mathcal{Q}(Q_\mathcal{X},P_\mathcal{X}) = \Big|d_{P_\mathcal{X}}(\mathcal{Q}) - d_{Q_\mathcal{X}}(\mathcal{Q})\Big|. \tag{7}$$

Note that $dis_\mathcal{Q}(\cdot,\cdot)$ is symmetric and fulfills the triangle inequality [6]. Note that for the sake of simplicity, we suppose that m = n, i.e., the size of $\mathbb{S}/\mathbb{S}_\mathcal{X}$ and $\mathbb{T}/\mathbb{T}_\mathcal{X}$ are equal. The authors showed that $dis_\mathcal{Q}(Q_\mathcal{X},P_\mathcal{X})$ can be bounded in terms of the classical PAC-Bayesian quantities and propose the following theorem (the theorem and it's proof can be found in [6, 7]) :

**Theorem 1** *(Germain [6, 7]) For any distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of hypothesis $\mathcal{H}$, any prior distribution $\mathcal{P}$ over $\mathcal{H}$, any $\delta\in(0,1]$, and any real number $\alpha>0$, with a probability at least $1-\delta$ over the choice of $(\mathbb{S}_\mathcal{X}\times\mathbb{T}_\mathcal{X})\sim(Q_\mathcal{X}\times P_\mathcal{X})^{m=n}$, for every $\mathcal{Q}$ on $\mathcal{H}$, we have,*

$$dis_\mathcal{Q}(Q_\mathcal{X},P_\mathcal{X}) \leq \frac{2\alpha}{1-e^{-2\alpha}}\Bigg[dis_\mathcal{Q}(\mathbb{S}_\mathcal{X},\mathbb{T}_\mathcal{X}) + \frac{2D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{2}{\delta}}{m\times\alpha} + 1\Bigg] - 1.$$

Note that the authors in [6, 7] propose the theorem 1 as "Catoni's type" with $c=2\alpha$. Indeed, as described in section 2.2 and in [6, 7] "Catoni's type" with PAC-Bayesian bound present interesting characteristics. First, its minimization is closely related to the minimization problem associated with the SVM when $\mathcal{Q}$ is an isotropic Gaussian over the space of linear classifiers (Germain et al. [9]). Second, the value $c=2\alpha$ allows to control the trade-off between the empirical risk and the complexity term $D_{\mathrm{KL}}(\cdot\|\cdot)$.

Thereby, from this domain's divergence, the authors proved the following domain adaptation bound (the theorem and its proof can be found in [7]).

**Theorem 2** *(Germain et al. [6, 7]). Let $\mathcal{H}$ be a hypothesis class. We have, $\forall$ $\mathcal{Q}$ on $\mathcal{H}$,*

$$\mathcal{R}_P(G_\mathcal{Q}) \leq \mathcal{R}_Q(G_\mathcal{Q}) + \frac{1}{2}dis_\mathcal{Q}(Q_\mathcal{X},P_\mathcal{X}) + \lambda_\mathcal{Q}, \tag{8}$$

where $\lambda_\mathcal{Q}$ is the deviation between the expected joint errors between pairs for voters ad pairs of views defined in section 2.2 on the target and source domains, which is defined as $\lambda_\mathcal{Q} = \Big|e_P(\mathcal{Q}) - e_Q(\mathcal{Q})\Big|$ and where
$e_P(\mathcal{Q}) = \mathbb{E}_{(\boldsymbol{x},y)\sim P} \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2} \big[\mathcal{L}_{0\text{-}1}\big(h(x),y\big) \times \mathcal{L}_{0\text{-}1}\big(h'(x),y\big)\big]$,
$e_Q(\mathcal{Q}) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q} \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2} \big[\mathcal{L}_{0\text{-}1}\big(h(x),y\big) \times \mathcal{L}_{0\text{-}1}\big(h'(x),y\big)\big]$.

In this section, we presented the principal bounds for Pac-Bayesian domain adaptation, in the next section we will discuss about Pac-Bayesian bounds in the multi-view learning setting.

## 2.4 Analysis of Pac-Bayesian Multi-view Bounds

First, the authors in [24] provided PAC-Bayesian bounds over the concatenation of the views, using priors that reflect how well the views agree on average over all examples, and deduced a SVM-like learning algorithm from this framework. However, this concatenation is designed for two views and kernel method, it is not generalizable to other methods. A more general framework of Pac-Bayesian bounds for multi-views was introduced in [11]. In the paper, the authors inroduced the two-level multiview approach. For each view $v\in[\![\mathcal{V}]\!]$, they consider a view-specific set $\mathcal{H}_v$ of voters $h:\mathcal{X}^v\to Y$, and a prior distribution $\mathcal{P}_v$ on $\mathcal{H}_v$. Given a hyper-prior distribution $\pi$ over the views $[\![\mathcal{V}]\!]$. In the paper, PAC-Bayesian learner objective has two parts. The first part is finding a posterior distribution $\mathcal{Q}_v$ over $\mathcal{H}_v, \forall\ v\in[\![\mathcal{V}]\!]$;

1. A pseudometric $d$ is a metric for wich the property $d(x,y)=0\Leftrightarrow x=y$ is relaxed to $d(x,y)=0\Leftarrow x=y$

The second is finding a hyper-posterior $\rho$ distribution on the set of views $[\![\mathcal{V}]\!]$. Thereby, [11] defined the multi-view weighted majority vote $\mathcal{B}_\rho^{\text{MV}}$ as :

$$\mathcal{B}_\rho^{\text{MV}}(\boldsymbol{x}) = \text{sign}\Big[\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_v} h(\boldsymbol{x}^v)\Big]. \tag{9}$$

Thus, the authors propose to build a learner that intends to construct posterior and hyperposterior distributions that minimize the actual risk $\mathcal{R}_D(\mathcal{B}_\rho^{\text{MV}})$ of the multiview weighted majority vote defined as :

$$\mathcal{R}_Q(\mathcal{B}_\rho^{\text{MV}}) = \mathop{\mathbb{E}}_{(\boldsymbol{x}^v,y)\sim Q}\Big[\mathcal{L}_{0\text{-}1}\big(\mathcal{B}_\rho^{\text{MV}}(\boldsymbol{x}^v), y\big)\Big]. \tag{10}$$

As stated in Section 2.2 the Gibbs risk is related to the expected disagreement $d_{Q_\mathcal{X}}(\mathcal{Q})$ and expected joint error $e_{Q_\mathcal{X}}(\mathcal{Q})$. Goyal et al. [11] note that the stochastic Gibbs classifier $G_\rho^{\text{MV}}$ defined as follows in the multiview setting, follows the same idea and it can be rewritten in terms of expected disagreement $d_Q^{\text{MV}}(\rho)$ and expected joint error $e_Q^{\text{MV}}(\rho)$ :

$$\begin{aligned}
\mathcal{R}_Q(G_\rho^{\text{MV}}) &= \frac{1}{2}d_Q^{\text{MV}}(\rho) + e_Q^{\text{MV}}(\rho),\\
d_{Q_\mathcal{X}}^{MV}(\rho) &= \mathbb{E}_{x^v\sim Q_\mathcal{X}}\ \mathbb{E}_{(v,v')\sim\rho^2}\\
&\qquad \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2}\big[\mathcal{L}_{0\text{-}1}\big(h(x^v), h'(x^v)\big)\big],\\
e_Q^{MV}(\rho) &= \mathbb{E}_{x^v\sim Q}\ \mathbb{E}_{(v,v')\sim\rho^2}\\
&\qquad \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2}\ \big[\mathcal{L}_{0\text{-}1}\big(h(x^v), y\big) \times \mathcal{L}_{0\text{-}1}\big(h'(x^v), y\big)\big].
\end{aligned} \tag{11}$$

As in the single-view PAC-Bayesian setting, [11] note that the multiview weighted majority vote $\mathcal{B}_\rho^{\text{MV}}$ is closely related to the stochastic multiview Gibbs classifier $G_\rho^{\text{MV}}$, and a generalization bound for $G_\rho^{\text{MV}}$ gives rise to a generalization bound for $\mathcal{B}_\rho^{\text{MV}}$. Moreover, the authors extend the C-Bound 3 to multiview setting by Lemma 3 below (the theorem and it's proof can be found in [11]).

**Lemma 3** *(Goyal et al. [11]). Let $V \geq 2$ be the number of views. For all posterior $\{\mathcal{Q}_v\}_{v=1}^V$ and hyper-posterior $\rho$ distribution, if $\mathcal{R}_Q(G_\rho^{\text{MV}}) < \frac{1}{2}$, then we have :*

$$\begin{aligned}
\mathcal{R}_Q(\mathcal{B}_\rho^{\text{MV}}) &\leq 1 - \frac{\big(1 - 2\,\mathcal{R}_Q\big(G_\rho^{\text{MV}}\big)\big)^2}{1 - 2d_{Q_\mathcal{X}}^{MV}(\rho)}\\
&\leq 1 - \frac{\big(1 - 2\,\mathbb{E}_{v\sim\rho}\mathcal{R}_Q\big(G_{\mathcal{Q}_v}\big)\big)^2}{1 - 2\,\mathbb{E}_{v\sim\rho}\,d_{Q_\mathcal{X}}(\mathcal{Q}_v)}.
\end{aligned} \tag{12}$$

The authors in the paper provide general multi-view PAC-Bayesian theorems and derive also a generalization bound with the approaches of [17]; [21]; [15]; and [5] introduce in section 2.2. The main difference between Goyal et al.'s bounds to theorems [17]; [21]; [15]; and [5] relies on the introduction of view-specific prior and posterior distributions, which mainly leads to an additional term $\mathbb{E}_{v\sim\rho}\ D_{\text{KL}}(\mathcal{Q}_v\|\mathcal{P}_v)$, expressed as the expectation of the view-specific Kullback-Leibler divergence term over the views $[\![\mathcal{V}]\!]$ according to the hyper-posterior distribution $\rho$.

# 3 Analysis of Unsupervised Multi-view

In this section we propose to introduce the concept of multi-view learning in the DA with generalization Pac-Bayesian guarantees. Then, we adapt the divergence proposed by Germain et al., [6, 8] with the concept of multi-view weighted majority vote introduced in [11]. In a second phase, we propose a general multi-view domain adaptation PAC-Bayesian theorem to upper-bound our divergence. Then, we present specialization of our theorem to the classical approaches. Finally, we propose a Pac-Bayesian domain adaptation bound in the multi-view setting.

## 3.1 Multi-view Domain Disagreement

Germain et al. [6] and Mansour et al.[16] propose a divergence measure that is based on the expected disagreement over the two domains. In the idea of measure disagreement we propose to adapt the definition 1 proposed by [6] to multi-view learning. Thus, we define the multi-view domain disagreement as follows :

**Definition 2** *(Multi-view domain disagreement) $\forall\, v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_v\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho$ over $[\![\mathcal{V}]\!]$, the multi-view domain disagreement $dis_\rho^{\text{MV}}(Q_\mathcal{X}, P_\mathcal{X})$ between $Q_\mathcal{X}$ and $P_\mathcal{X}$ is defined by :*

$$dis_\rho^{\text{MV}}(Q_\mathcal{X}, P_\mathcal{X}) = \Big|d_{P_\mathcal{X}}^{\text{MV}}(\rho) - d_{Q_\mathcal{X}}^{\text{MV}}(\rho)\Big|, \tag{13}$$

where $d_{Q_\mathcal{X}}^{\text{MV}}(\rho)$, $d_{P_\mathcal{X}}^{\text{MV}}(\rho)$, are expected disagreement defined in 2.2 and defined as follows in our multiview setting [11].

From this domain's divergence, we can find the same domain adaptation bound 2 with an adaptation to multi-view learning :

**Theorem 4** *$\forall\, v \in [\![\mathcal{V}]\!]$, for any distributions $Q$ and $P$ over $\mathcal{X}\times\mathcal{Y}$, for any set of voters $\mathcal{H}_v$, for any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_v\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho$ over $[\![\mathcal{V}]\!]$, we have :*

$$\mathcal{R}_P(G_\rho^{\text{MV}}) \leq \mathcal{R}_Q(G_\rho^{\text{MV}}) + \frac{1}{2}dis_\rho^{\text{MV}}(Q_\mathcal{X}, P_\mathcal{X}) + \lambda_\rho, \tag{14}$$

where $\lambda_\rho$ is the deviation between the expected joint errors between pairs for voters ad pairs of views defined in section 2.4 on the target and source domains, which is defined as $\lambda_\rho = \Big|e_P^{\text{MV}}(\rho) - e_Q^{\text{MV}}(\rho)\Big|$ and where
$e_P^{\text{MV}}(\rho) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\ \mathbb{E}_{(v,v')\sim\rho^2}\ \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2}\ \big[\mathcal{L}_{0\text{-}1}\big(h(x), y\big)\times \mathcal{L}_{0\text{-}1}\big(h'(x), y\big)\big]$,
$e_Q^{\text{MV}}(\rho) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q}\ \mathbb{E}_{(v,v')\sim\rho^2}\ \mathbb{E}_{(h,h')\sim\mathcal{Q}_v^2}\ \big[\mathcal{L}_{0\text{-}1}\big(h(x), y\big)\times \mathcal{L}_{0\text{-}1}\big(h'(x), y\big)\big]$.

**Proof 1** *The proof borrows the straightforward proof technique of Theorem 9 in [7].*

Note that the theorem 4 is very similar to the theorem 2. In fact, we trivially have $dis_{\rho}^{\text{MV}}(Q_{\mathcal{X}}, P_{\mathcal{X}}) \leq \mathbb{E}_{(v,v')\sim\rho^2} dis_{\mathcal{Q}}(Q_{\mathcal{X}}, P_{\mathcal{X}})$, that inequality exhibits the role of diversity among the views thanks to the disagreement's expectation over the views as showed in the equation 3.

## 3.2 General Multi-view Domain Disagreement Pac-Bayesian Theorem

In this section we show that $dis_{\rho}^{\text{MV}}(Q_{\mathcal{X}}, P_{\mathcal{X}})$ can be bounded in terms of the general PAC-Bayesian quantities [9, 10]. Note that, the authors in [11] also propose a variation of the general PAC-Bayesian theorem of [9, 10]. Besides, Goyal et al. [11] bounds $\mathbb{E}_{\mathbb{S}\sim Q^m}\mathcal{R}_Q(G_{\mathcal{Q}_{\mathbb{S}}})$, where $\mathcal{Q}_{\mathbb{S}}$ is the posterior distribution outputted by a given learning algorithm after observing the learning sample $\mathbb{S}$. Whereas PAC-Bayesian bounds from [9, 10] and [2] bounds $\mathcal{R}_Q(G_{\mathcal{Q}})$ uniformly for all distribution $Q$, with high probability over the draw of $\mathbb{S} \sim Q^m$. Thereby, Goyal et al. [11]'s theorem has the advantage to involve an expectation over all the possible learning samples (of a given size) in bounds itself. Moreover, the authors in [11] adapt their general Pac-Bayesian bound to multi-view setting. Then, we propose the following theorem as a generalization of Theorem ; It takes the form of an upper bound on the deviation between the true risk of the Multi-view domain disagreement $2\,\mathbb{E}_{\mathbb{S}\sim Q}\, dis_{\rho_{\mathbb{S}}}^{\text{MV}}(Q_{\mathcal{X}}, Q_{\mathcal{X}})$ and its empirical counterpart $\mathbb{E}_{\mathbb{S}\sim Q}\, dis_{\rho_{\mathbb{S}}}^{\text{MV}}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}})$, according to a convex function $\Delta : [0,1]^2 \to \mathbb{R}$. Note that, as the Goyal et al.'s bounds we incorporate in our general theorem the posterior and hyper-posterior distribution $\mathcal{Q}_{v,\mathbb{S}}/\rho_{\mathbb{S}}$ outputted by a given learning algorithm after observing the learning sample $\mathbb{S}$

**Theorem 5** *$\forall\, v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_{\mathcal{X}}$ and $P_{\mathcal{X}}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_{\mathbb{S}}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, for any $m > 0$, for any convex function $\Delta : [0,1]^2 \to \mathbb{R}$, with probability at least $1-\delta$ over the choice of $(\mathbb{S}_{\mathcal{X}} \times \mathbb{T}_{\mathcal{X}}) \sim (Q_{\mathcal{X}} \times P_{\mathcal{X}})^m$, with probability at least $1-\delta$, we have :*

$$\Delta\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}}), \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(Q_{\mathcal{X}}, P_{\mathcal{X}})\right) \leq \frac{2}{m}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_{\mathbb{S}}} D_{\text{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} D_{\text{KL}}(\rho_{\mathbb{S}}\|\pi) + \ln\sqrt{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{(v,v')\sim\pi^2}\mathop{\mathbb{E}}_{(h,h')\sim\mathcal{P}_v^2} e^{m\Delta(\mathcal{R}_{\hat{d}}(\hat{h}),\mathcal{R}_d(\hat{h}))}}\right], \tag{15}$$

where $\hat{h} = (h, h')$ a pair of hypothesis, $(x^{(v)}, x'^{(v)}) \sim (Q_{\mathcal{X}} \times P_{\mathcal{X}})^m$ a pair of examples, $\mathcal{L}_d(\hat{h}, x^{(v)}, x'^{(v)}) = |\mathcal{L}_{0\text{-}1}(h(x'^{(v)}), h'(x'^{(v)})) - (h(x^{(v)}), h'(x^{(v)}))|$ and $\mathcal{R}_d(\hat{h}) = \mathbb{E}_{(x^{(v)},x'^{(v)})\sim(Q_{\mathcal{X}}\times P_{\mathcal{X}})^m}\mathcal{L}_d(\hat{h}, x^{(v)}, x'^{(v)})$, $\mathcal{R}_{\hat{d}}(\hat{h}) = \mathbb{E}_{(x^{(v)},x'^{(v)})\sim(\mathbb{S}_{\mathcal{X}}\times\mathbb{T}_{\mathcal{X}})^m}\mathcal{L}_d(\hat{h}, x^{(v)}, x'^{(v)})$ the risk of $\hat{h}$ on the joint distribution.

**Proof 2** *The proof uses the ideas of the techniques and tricks of Bégin et al. [2], Theorem 4. A detailed proof is available in additional materials.*

## 3.3 Specialization of multi-view domain disagreement to the Classical Approaches

In this section, we provide specialization of our multiview theorem to the most popular PAC-Bayesian approaches. To do so, we follow the same principles as Germain et al. [9, 10]. Selecting a well-suited deviation function $\Delta$ and by upper-bounding $\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{(v,v')\sim\pi^2}\mathop{\mathbb{E}}_{(h,h')\sim\mathcal{P}_v^2} e^{m\Delta(\mathcal{R}_{\hat{d}}(\hat{h}),\mathcal{R}_d(\hat{h}))}$, we can derive easily the classical PAC-Bayesian theorems of [17], [21], [5] presented in the section 2.2. First, we derive the specialization theorem 5 to the McAllester [18]'s point of view.

**Corollary 1** *$\forall\, v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_{\mathcal{X}}$ and $P_{\mathcal{X}}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_{\mathbb{S}}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, we have :*

$$\left|\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}}) - \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(Q_{\mathcal{X}}, P_{\mathcal{X}})\right| \leq \sqrt{\frac{1}{2m}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_{\mathbb{S}}} 2D_{\text{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} 2D_{\text{KL}}(\rho_{\mathbb{S}}\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}. \tag{16}$$

**Proof 3** *The proof uses the ideas of the techniques and tricks of Germain et al. [7, 10]. A detailed proof is available in additional materials.*

To derive a generalization bound with the Catoni [5]'s point of view—given a convex function $\mathcal{F}$ and a real number $c > 0$ we define the measure of deviation as $\Delta(a,b) = \mathcal{F}(b) - c$ (Germain et al., [9, 7, 10]). We obtain the following generalization bound :

**Corollary 2** *$\forall\, v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_{\mathcal{X}}$ and $P_{\mathcal{X}}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_{\mathbb{S}}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^{V}$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, $\forall\, \alpha > 0$, with probability at least $1-\delta$, we have :*

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(Q_{\mathcal{X}}, P_{\mathcal{X}}) \leq \frac{2\alpha}{1-e^{-2\alpha}}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis_{\rho_{\mathbb{S}}}^{MV}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}}) + \frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_{\mathbb{S}}} D_{\text{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} D_{\text{KL}}(\rho_{\mathbb{S}}\|\pi) + \ln\sqrt{\frac{1}{\delta}}}{m\times\alpha}\right]. \tag{17}$$

**Proof 4** *The proof uses the ideas of the techniques and tricks of Germain et al. [7, 10]. A detailed proof is available in additional materials.*

As stated in [8], to recover a PAC-Bayesian bound similar to that proposed by Seeger [21] ; Langford [15], we use as $\Delta$-function the Kullback-Leibler divergence :

**Corollary 3** *$\forall v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_\mathbb{S}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, we have :*

$$D_{\mathrm{KL}}\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis^{MV}_{\rho_\mathbb{S}}(\mathbb{S}_\mathcal{X},\mathbb{T}_\mathcal{X}), \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis^{MV}_{\rho_\mathbb{S}}(Q_\mathcal{X},P_\mathcal{X})\right) \leq \frac{1}{m}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_\mathbb{S}} 2D_{\mathrm{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} 2D_{\mathrm{KL}}(\rho_\mathbb{S}\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]. \tag{18}$$

**Proof 5** *The proof uses the ideas of the techniques and tricks of Germain et al. [7, 10]. A detailed proof is available in additional materials.*

The bounds 2, 3, 1 are demonstrated for m=n, i.e., the sizes of samples from source domain $\mathbb{S}/\mathbb{S}_\mathcal{X}$ are the same of the samples from target domain $\mathbb{T}/\mathbb{T}_\mathcal{X}$. The last result of this section tackles the situation where we assume $m \neq n$, i.e., the sizes of $\mathbb{S}/\mathbb{S}_\mathcal{X}$ and $\mathbb{T}/\mathbb{T}_\mathcal{X}$ are different.

**Corollary 4** *$\forall v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_\mathbb{S}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, we have :*

$$\left|\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis^{\mathrm{MV}}_{\rho_\mathbb{S}}(\mathbb{S}_\mathcal{X},\mathbb{T}_\mathcal{X}) - \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} dis^{\mathrm{MV}}_{\rho_\mathbb{S}}(Q_\mathcal{X},P_\mathcal{X})\right| \leq \sqrt{\frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_\mathbb{S}} 2D_{\mathrm{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v)}{2m}} + \frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} 2D_{\mathrm{KL}}(\rho_\mathbb{S}\|\pi) + \ln\frac{2\sqrt{m}}{\delta}}{2m} + \sqrt{\frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathop{\mathbb{E}}_{v\sim\rho_\mathbb{S}} 2D_{\mathrm{KL}}(\mathcal{Q}_{v,\mathbb{S}}\|\mathcal{P}_v)}{2n}} + \frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} 2D_{\mathrm{KL}}(\rho_\mathbb{S}\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{2n}. \tag{19}$$

**Proof 6** *The proof uses the ideas of the techniques and tricks of Germain et al. [7, 10]. A detailed proof is available in additional materials.*

### 3.4 The PAC-Bayesian DA-Bound

Finally, the Theorem 4 leads to a PAC-Bayesian bound based on both the empirical source error of the Gibbs classifier and the empirical Multi-view domain disagreement pseudo-metric estimated on a source and target samples. The following bound is based on Catoni's approach 2 :

**Theorem 6** *$\forall v \in [\![\mathcal{V}]\!]$, for any set of voters $\mathcal{H}_v$ for any marginal distributions $Q_\mathcal{X}$ and $P_\mathcal{X}$ over $\mathcal{X}$, any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-posterior distribution $\rho_\mathbb{S}$ over $[\![\mathcal{V}]\!]$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^V$ on $\mathcal{H}_v$, for any hyper-prior distribution $\pi$ over $[\![\mathcal{V}]\!]$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, we have :*

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\mathcal{R}_P(G^{\mathrm{MV}}_{\rho_\mathbb{S}}) \leq \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} c'\mathcal{R}_\mathbb{S}(G^{\mathrm{MV}}_{\rho_\mathbb{S}}) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m}\alpha'\frac{1}{2}dis^{\mathrm{MV}}_{\rho_\mathbb{S}}(\mathbb{S}_\mathcal{X},\mathbb{T}_\mathcal{X}) + \left(\frac{c'}{c}+\frac{\alpha'}{\alpha}\right)\frac{\mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} D_{KL}(\mathcal{Q}_{v,\mathbb{S}}||\mathcal{P}_v) + \mathop{\mathbb{E}}_{\mathbb{S}\sim Q^m} D_{KL}(\rho_{v,\mathbb{S}}||\pi) + \ln\frac{1}{\delta}}{m} + \lambda_\rho + \frac{1}{2}(\alpha'-1), \tag{20}$$

where $c' = \frac{c}{1-e^{-c}}$ and $\alpha' = \frac{2\alpha}{1-e^{-2\alpha}}$.

**Proof 7** *In Theorem 4, replace $\mathbb{E}_{\mathbb{S}\sim Q^m}\ \mathcal{R}_Q(G^{\mathrm{MV}}_{\rho_\mathbb{S}})$ and $\mathbb{E}_{\mathbb{S}\sim Q^m}\ dis^{\mathrm{MV}}_{\rho_\mathbb{S}}(Q_\mathcal{X},P_\mathcal{X})$ by their upper bound, obtained from Corollary 2 in [11] and Corollary 2.*

## 4 Discussions and Conclusion

The primary contrast between our bounds 2 ; 3, 1, 4, 6, and Germain et al.'s bounds [7] lies in the incorporation of view-specific prior and posterior distributions. This results in an extra term, $\mathbb{E}_{v\sim\rho}\ D_{\mathrm{KL}}(\mathcal{Q}_v\|\mathcal{P}_v)$, which represents the expected value of the view-specific Kullback-Leibler divergence term over the views$[\![\mathcal{V}]\!]$, based on the hyper-posterior distribution $\rho$. The second difference comes from the expectation over all the possible learning samples in bounds itself [11]. In this way, the expectation $\mathbb{E}_{\mathbb{S}\sim Q^m}$ is distributed for the all terms in the bounds. Thereby, the $D_{\mathrm{KL}}(\cdot\|\cdot)$ terms take account of the the posterior and hyper-posterior distribution $\mathcal{Q}_{v,\mathbb{S}}/\rho_\mathbb{S}$ outputted by a given learning algorithm after observing the learning sample $\mathbb{S}$.

Finally in this paper, we propose a first PAC-Bayesian analysis of weighted majority vote classifiers for domain adaptation with the concept of multi-view learning. Our works is based on theoretical results and for the future we aim to derive from introduced bounds a new domain adaptation multi-view algorithm. We will build on the work of Germain et al. [9, 6] to propose a specialized algorithm for linear classifiers or to propose a specialized algorithm for neural networks [23].

.

# Références


[1] Pierre Alquier. User-friendly introduction to pac-bayes bounds, 2021.

[2] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. Pac-bayesian bounds based on the rényi divergence. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 435–444, Cadiz, Spain, 09–11 May 2016. PMLR.

[3] O. Catoni. *Statistical learning theory and stochastic optimization. Ecole d'été de probabilités de Saint-Flour XXXI-2001*. Springer, 2004. Collection : Lecture notes in mathematics n°1851.

[4] Olivier Catoni. A pac-bayesian approach to adaptive classification. 2004.

[5] Olivier Catoni. Pac-bayesian supervised classification : The thermodynamics of statistical learning. *Lecture Notes-Monograph Series*, 56 :i–163, 2007.

[6] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 738–746, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[7] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. Pac-bayesian theorems for domain adaptation with specialization to linear classifiers, 2015.

[8] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 859–868, New York, New York, USA, 20–22 Jun 2016. PMLR.

[9] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 353–360, New York, NY, USA, 2009. Association for Computing Machinery.

[10] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario March, and Jean-Francis Roy. Risk bounds for the majority vote : From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26) :787–860, 2015.

[11] Anil Goyal, Emilie Morvant, Pascal Germain, and Massih-Reza Amini. Pac-bayesian analysis for a two-step hierarchical multiview learning approach. In Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens, and Sašo Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 205–221, Cham, 2017. Springer International Publishing.

[12] Benjamin Guedj. A primer on pac-bayesian learning. 2019.

[13] Mehdi Hennequin, Khalid Benabdeslem, and Haytham Elghazel. Adversarial multi - view domain adaptation for regression. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[14] Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[15] John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(10) :273–306, 2005.

[16] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation : Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[17] David A. McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA, 1998. Association for Computing Machinery.

[18] David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, page 164–170, New York, NY, USA, 1999. Association for Computing Machinery.

[19] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1) :521–530, 2012.

[20] Jonathan Munro, Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Domain adaptation in multi-view embedding for cross-modal video retrieval. *CoRR*, abs/2110.12812, 2021.

[21] Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct) :233–269, 2002.

[22] John Shawe-Taylor and Robert C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, page 2–9, New York, NY, USA, 1997. Association for Computing Machinery.

[23] Anthony Sicilia, Katherine Atwell, Malihe Alikhani, and Seong Jae Hwang. Pac-bayesian domain adaptation bounds for multiclass learners, 2022.

[24] Shiliang Sun, John Shawe-Taylor, and Liang Mao. Pac-bayes analysis of multi-view learning. *Information Fusion*, 35 :117–131, 2017.

[25] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.

[26] Pei Yang, Wei Gao, Qi Tan, and Kam-Fai Wong. Information-theoretic multi-view domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 270–274, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[27] Wen Zhang, Lingfei Deng, and Dongrui Wu. Overcoming negative transfer : A survey. *CoRR*, abs/2009.00909, 2020.

# Propriétés émergentes du *multi-clustering* bayésien non paramétrique: Application aux données images multivues

Reda Khoufache[1,2], Mohamed Djallel Dilmi[2], Hanene Azzag[2], Etienne Goffinet[3], Mustapha Lebbah[1,2]

[1] DAVID, Université de Versailles, Université Paris-Saclay
[2] LIPN (CNRS UMR 7030), Université Sorbonne Paris Nord
[3] Technology Innovation Institute, TII

## Résumé

*Ce papier a déjà été publié dans [16]. Dans cet article, nous avons développé un nouveau framework qui traite le problème de clustering d'images multivues collectées dans un magasin intelligent. Le framework combine les transformers pré-entrainés avec un nouveau modèle de multi-clustering bayésien non paramétrique. Dans ce travail, nous proposons une approche MCMC afin d'inférer la partition colonne et les partitions lignes. Notre modèle infère de multiples partitions et permet d'estimer automatiquement le nombre de clusters. Notre méthode fournit des blocs homogènes sur un jeu de données d'images multivues. Elle souligne d'une part la puissance des vision transformers pré-entrainés combinés au multi-clustering et d'autre part l'utilité de la modélisation bayésienne non paramétrique qui automatise l'estimation du nombre de clusters. Le code source du framework complet est disponible dans ce lien github :* `https://github.com/unsupervise/acomi`

## Mots-clés

*Modèles bayésiens non paramétriques, sélection de modèles, multi-clustering, clustering d'images multivues, vision transformers*

## Abstract

*This paper has already been published in [16]. In this article, we have developed a new framework to address the multi-view image clustering of images collected from a smart store. The proposed framework combines a pre-trained Vision Transformer with a new bayesian nonparametric multiple clustering model. In this work, we propose an MCMC-based inference approach to learn the column-partition and the row-partitions. This method infers multiple clustering solutions and allows to find the number of clusters automatically. Our method provides interesting results on a multi-view image dataset and emphasizes, on the one hand, the power of pre-trained Vision Transformers combined with the multiple clustering algorithm, on the other hand, the usefulness of the bayesian nonparametric modeling, which automatically performs a model selection. The complete code of our framework is available at* `https://github.com/unsupervise/acomi`

## Keywords

*Bayesian nonparametric, model selection, multiple clustering, multi-view image clustering, vision transformers*

## 1 Introduction

L'intelligence artificielle a un impact considérable sur le secteur du *retail* ; les entreprises de grande distribution investissent de plus en plus dans des solutions de vision par ordinateur afin d'améliorer la gestion de leurs magasins et mieux satisfaire les attentes de leurs clients.
La classification de produits dans un magasin autonome est une tâche délicate, ceci est dû à la quantité massive de produits ainsi que les variables qui décrivent chaque produit. Un des enjeux majeurs est l'amélioration de la classification de produits dans un magasin intelligent. Le développement d'algorithmes d'apprentissage passant à l'échelle et qui permet de traiter simultanément les données et les variables est indispensable pour construire des modèles prédictifs. L'apprentissage non supervisé multivues où chaque vue représente une variable fournit des résultats supérieurs à ceux à vue unique, et ce, en raison de l'utilisation des informations complémentaires provenant de différents espaces de représentation.
Dans le cas multivarié, le *clustering* infère uniquement une partition ligne, tandis que le *multi-clustering* infère une partition en colonne (partition de variables ou vues), et une partition ligne pour chaque vue.
Afin de traiter le problème du *multi-clustering*, une approche probabiliste a été introduite par [10] ; ce modèle suppose la présence d'une structure avec de multiples blocs où chaque cellule est une observation univariée continue. Les modèles paramétriques supposent que le vrai nombre de blocs est connu a priori. Cette hypothèse n'est généralement pas satisfaite en pratique. Par conséquent, une étape intermédiaire de sélection de modèle est nécessaire afin d'estimer ce nombre. Cette sélection est généralement réalisée soit avec une recherche exhaustive, soit avec recherche gloutonne. Cependant, le nombre de modèles possibles est très grand même avec un petit nombre de blocs [12]. Ainsi, une recherche exhaustive n'est plus envisageable. Des stratégies heuristiques sont alors considérées afin d'explorer une partie de l'espace des solutions au coût d'hypothèses supplémentaires sur la structure du modèle. Dans [13], les

auteurs ont proposé un modèle de *multi-clustering* paramétrique fonctionnel désigné pour traiter des séries temporelles multivariées ; la sélection de modèle est réalisée en utilisant le critère *ICL* (*Integrated Classification Likelihood*).
La modélisation bayésienne non paramétrique permet d'estimer le nombre de composantes durant l'inférence en mettant une distribution a priori sur les paramètres du modèle. Le processus de dirichlet pour les modèles de mélange (*DPMM*) [1] est un exemple de modèle bayésien non paramétrique qui suppose la présence d'un nombre infini de composantes latentes. Deux représentations sont souvent utilisées afin de mieux interpréter le processus de dirichlet (*DP*) : processus du restaurant chinois (*CRP*) [2] et le *Stick-Breaking* (*SB*) [24]. Ces deux représentations son liées, l'une peut découler de l'autre [22]. Dans [21], les auteurs ont introduit un modèle de catégorisation croisée, et [14] ont proposé un modèle de *multi-clustering* bayésien non paramétrique. Ces deux travaux partagent la même définition du modèle, qui met d'abord un a priori sur la partition colonne, qui estime automatiquement le nomble de *clusters* colonnes, ensuite met un a priori indépendant sur les proportions de chaque partition ligne. Par ailleurs, dans [14], les auteurs proposent une inférence variationnelle afin d'estimer les paramètres du modèle, tandis que [21] utilise une méthode *MCMC*.
Dans ce travail, nous proposons une nouvelle approche pour le *clustering* d'images multivues basée sur le *clustering* bayésien non paramétrique. Le *clustering* d'images est une tâche complexe car les images sont présentées sous forme de données de grande dimension, ce qui affecte la performance des algorithmes du *clustering*. Des méthodes d'apprentissage de représentation tels que SIFT [19], HOG [7] sont utilisées afin d'en extraire des caractéristiques. Celles-ci sont ensuite utilisées pour discriminer les images. Les approches récentes combinent les méthodes du *clustering* avec les réseaux de neuronnes profonds (*deep clustering*), ce qui a permis d'atteindre des performances remarquables. Ces méthodes peuvent être divisées en deux catégories : la première catégorie réalise d'abord l'extraction des caractéristiques en utilisant des réseaux de neurones à convolution tels que *deep belief network* [17], et les auto-encodeurs [25]. Ensuite un algorithme de *clustering* est appliqué sur les caractéristiques extraites. La deuxième catégorie prend en considération cette séparation, ainsi des travaux récents tels que [18, 27] proposent des méthodes basées sur un outil unifié qui réalise de façon jointe l'apprentissage profond des représentations et les *clusters* d'images.

# 2 Contexte

## 2.1 *Vision transformers*

Les *transformers* sont des réseaux de neurones qui utilisent des mécanismes d'attention [26], leurs permettant de se focaliser sur certaines régions de l'entrée, et ainsi d'extraire des caractéristiques intrinsèques et significatives. Les *transformers* ont d'abord été introduits pour le traitement naturel du langage (*NLP*) [26], ces modèles ont eu un énorme succès avec leur capacité de représentation remarquable.
Récemment, les *transformers* ont été adaptés afin de traiter des tâches de vision par ordinateur. Le modèle *ViT* (*Vision Transformer*) [9] applique une architecture du *transformer* sur des séquences de petits patchs d'images afin de les classifier. Le modèle a atteint des performances exceptionnelles sur de nombreux jeux de données tel que *Imagenet* [8]. Une étude dédiée à l'apprentissage par transfert de ces modèles a été réalisée dans [20]. En effet, l'apprentissage par transfert [3] est un outil important qui permet de gagner en temps et en ressources nécessaires au réentrainement complet de ces réseaux de neurones. Ainsi, au lieu de réentrainer le *ViT* entièrement, nous considérons un réseau déjà pré-entraîné sur *ImageNet* et l'utilisons pour l'extraction des caractéristiques.

## 2.2 *Multi-clustering*

### 2.2.1 Définition du modèle

Notons $X \in \mathbb{R}^{n\times p\times d}$ l'espace latent obtenu après une certaine transformation du jeu de données d'images multivues, où $n$ est le nombre d'observations, $p$ est le nombre de variables $d$ est la dimension de l'espace de représentation. Soit $H$ le nombre de *clusters* de variables, $v$ la partition de variables, $Z$ une matrice indicatrice $n \times H$, des partitions lignes. Le modèle est défini comme suit :

$$
\begin{aligned}
x_{i,j} \mid \left\{v_j = h, z_i^h = k, \theta_k^h\right\} &\sim \mathcal{N}\left(\theta_k^h\right), \\
\theta_k^h \sim G_0,\, v_j \sim \mathrm{Mult}(\eta),\, z_i^h &\sim \mathrm{Mult}\left(\pi_h\right), \\
\eta_j(\mathbf{r}) = r_j \prod_{j'=1}^{j-1}\left(1 - r_{j'}\right), r_j &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1,\gamma), \\
\pi_j^h\left(\mathbf{t}^h\right) = t_j^h \prod_{j'=1}^{j-1}\left(1 - t_{j'}^h\right), t_j^h &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}\left(1,\alpha_h\right), \\
\gamma \sim \mathrm{Gamma}\left(a_\gamma, b_\gamma\right), \alpha_h &\sim \mathrm{Gamma}\left(a_\alpha, b_\alpha\right).
\end{aligned}
$$

Où $x_{i,j}$ est l'élément de la ligne $i$ et la colonne $j$, $v_j$, la variable d'appartenance de la colonne $j$, $z_i^h$ la variable d'appartenance de la ligne $i$ dans le cluster de variables (vue) $h$. Les proportions de la partition de variables $\eta$ et des partitions lignes $\pi_h$ suivent le processus *Stick-Breaking* [24]. Chaque paramètre de concentration suit une loi a priori Gamma, ce qui permet de les adapter à chaque bloc qui peut contenir un nombre différent de composantes. La densité associée à chaque bloc est une gaussienne multivariée de dimension $d$ et de paramètre $\theta_k^h = (\mu_k^h, \Sigma_k^h)$. $G_0$ est la loi à priori conjuguée Normale-Wishart-Inverse (*NIW*), l'ensemble des hyperparamètres est noté $\chi$.
Le modèle défini dans ce papier est une extension au cas bayésien non paramétrique du *multi-clustering* paramétrique introduit dans [13]. Il est important de noter que dans ce travail, les densités composantes associées à chaque bloc modélisent les cellules et non pas les lignes. En d'autres termes, les cellules appartenant au même bloc suivent indépendamment la même distribution multivariée.

#### 2.2.2 Inférence

L'inférence est réalisée en utilisant une variante de l'algorithme Gibbs et alterne entre trois étapes principales : en partant d'une partition initiale, nous commençons par mettre à jour la partition des variables sachant la partition des lignes. Ceci permets d'estimer les différentes vues. Ensuite, pour chaque vue, nous mettons à jour la partition des lignes sachant la partition de variables. Finalement, nous mettons à jours les paramètres de concentration. Dans ce papier, nous détaillons uniquement l'inférence de la partition de variables, les deux autres étapes peuvent être retrouvées dans [16].

**Inférence de la partition de variables :** La partition des variables $v$ est mise à jour sachant les partitions lignes $Z$. Chaque variable d'appartenance $(v_j)_p$ est échantillonnée selon $\mathrm{p}\left(v_j \mid \mathbf{v}_{-j}, Z, \mathbf{x}_{.,j}, \chi\right) \propto$ :

$$\begin{cases} \dfrac{p_h}{p-1+\gamma}\mathrm{p}(\mathbf{x}_{.,j} \mid \mathbf{z}^h, \chi), & \textit{cluster} \text{ existant } h, \quad (1) \\ \dfrac{\gamma}{p-1+\gamma}\mathrm{p}(\mathbf{x}_{.,j} \mid \chi), & \text{nouveau } \textit{cluster}, \quad (2) \end{cases}$$

Avec $p_h$ le cardinal du *cluster* en colonne $h$ et $\mathbf{v}_{-j} = \{v_i : i \neq j\}$. Dans l'équation 1, en vertu du choix approprié de la loi a priori conjuguée $G_0$, la distribution prédictive jointe a priori $\mathrm{p}(\mathbf{x}_{.,j}|z^h, \chi)$ se réduit à un produit de densités de Student multivariées [12]. Dans l'équation 2, $\mathrm{p}(\mathbf{x}_{.,j}|\chi)$ est estimé en appliquant un *DPM* par variable avant l'inférence du *multi-clustering*, ce qui fournit également la partition ligne optimale $\hat{\mathbf{z}}^j$ associée à la variable $j$.

#### 2.2.3 Implémentation

Nous initialisons l'algorithme comme suit : $\mu_0$ est le vecteur nul de dimension $d$. D'après [23], la matrice de précision $\Psi_0$ obtenue avec l'estimateur du maximum de vraisemblance est une bonne initialisation, $\kappa_0$ et $\nu_0$ sont initialisés avec leur plus faible valeur afin qu'ils soient le moins informatifs possible. La partition initiale est la partition à un seul *cluster*.

#### 2.2.4 Inférer la partition finale

L'algorithme de Gibbs produit une chaîne de Markov de partitions échantillonnées à partir de la distribution a posteriori approximée. Ces partitions échantillonnées doivent être agrégées à partir d'un certain nombre d'itérations. Dans notre cas, nous gardons la dernière partition échantillonnée.

#### 2.2.5 La complexité algorithmique

Avant l'inférence principale, un DPM est lancé sur chaque variable $\mathbf{x}_{.,j}$ pour estimer $\mathrm{p}(\mathbf{x}_{.,j} \mid \chi)$. Chaque DPM a une compléxité de $\mathrm{O}(Mnd^2)$. Comme l'inférence est appliquée sur chaque variable, la complexité est $O(Mnpd^2)$. Durant l'inférence principale, la mise à jour des *clusters* en colonne a une complexité de $O(p\bar{H}(nd^2 + \bar{K}d^3))$ avec $\bar{H}$ et $\bar{K}$ les nombres maximaux des *clusters* en colonne et *clusters* lignes respectivement. La mise à jour des *clusters* lignes a une complexité de $O(npd^2 + (n+p)\bar{K}d^3)$. Finalement, le processus complet d'inférence a une complexité de $O(M(npd^2 + (n+p)\bar{K}d^3))$.

## 3 Le framework proposé

**1. Préparation du jeu de données**

Le jeu de données d'images est réarrangé sous forme d'un tableau bi-dimensionnel où les lignes représentent les observations, les colonnes sont les différentes variables et chaque cellule est une image qui décrit l'observation correspondante.

**2. Extraction des caractéristiques et réduction de dimension**

Dans la première étape, nous proposons d'utiliser un *Vision Transformers* pré-entraîné comme extracteur de caractéristiques. Nous traitons chaque image avec le *ViT* et gardons la dernière couche avant la prédiction des classes comme vecteur de représentation latent. Nous appliquons ensuite une analyse en composantes principales (*PCA*) [15] afin de réduire la dimension de l'espace de représentation en raison du fléau de la dimension et de la complexité algorithmique. L'espace latent obtenu est un *tensor* de dimension $n \times p \times d_r$, où $d_r$ est le nombre de composantes conservées après l'application de la *PCA*.

**3. Multi clustering**

Le jeu de données obtenu après la transformation est ensuite passé en entrée de l'algorithme du *multi-clustering* détaillé dans la section 2.2, ce qui nous fournit de multiples partitions. Le *framework* complet est illustré dans la figure 1.

## 4 Expériences

### 4.1 Jeu de données

Aloi[1] [11] est un jeu de données d'images multi-vues contenant 1000 objets, chacun est décrit par 108 images différentes, celles-ci sont obtenues en variant trois critères : l'angle de prise de l'image, l'intensité lumineuse et la position de la caméra. Ainsi, 72 variables notées $\{\mathrm{r\_0}, \mathrm{r\_5}, \cdots, \mathrm{r\_355}\}$ représentent des prises du même objet pivoté par incrément de $5°$ couvrant les $360°$ ; 12 variables notées $\{\mathrm{i\_110}, \mathrm{i\_120}, \cdots, \mathrm{i\_250}\}$ représentent des prises avec différentes intensités lumineuses et 24 variables $\{\mathrm{l1\_}c1, \cdots, \mathrm{l8\_}c3\}$ qui représentent des prises de 3 différentes caméras avec 8 positions différentes de la source lumineuse. Le jeu de données contient des annotations des caractéristiques de ces produits tels que la matière (*Material*) (vêtement, bois, fruit, etc). Nous avons choisi ce jeu de données, car celui-ci contient des objets similaires à ceux qu'on peut retrouver dans un magasin, de plus, comme dans un magasin intelligent, ces objets sont capturés sous différents angles et sous différentes conditions lumineuses (intensité et angles). En notre connaissance, le jeu de données Aloi, est le seul jeu de données images multi-vues public satisfaisant ces caractéristiques.

1. https://aloi.science.uva.nl/

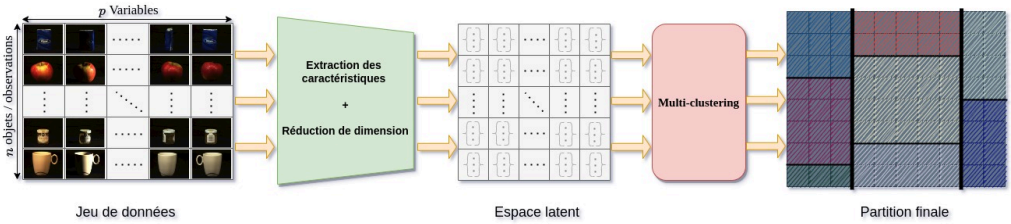



FIGURE 1 – Le *framework* complet

| Méthode | Scores / Vues | *RI* | *Purity* | *NMI* |
|---|---|---|---|---|
| *GMM* | - | 0.890 | 0.299 | 0.367 |
| *LBM* | - | 0.892 | 0.298 | 0.367 |
| Le notre | Vue 2 | **0.899** | 0.669 | 0.592 |
| | Vue 1 | **0.899** | 0.682 | 0.596 |
| | Vue 3 | **0.899** | **0.685** | **0.599** |
| | Vue 10 | 0.898 | 0.579 | 0.560 |

TABLE 1 – Résultats des métriques *RI, Purity* et *NMI*.

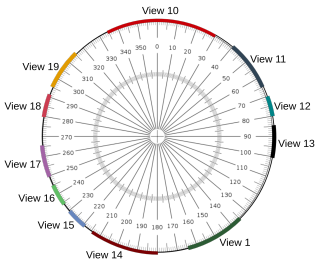



FIGURE 2 – Positions spatiales des différentes vues

| Vues | *Clusters* de variables | Interprétation |
|---|---|---|
| Vue 1 | r_135, ... , r_165 | Agrégation de légères rotations |
| Vue 2 | i110, i120, i130, i150, i160 | Température faible |
| Vue 3 | l3_c1, l3_c2, l3_c3 | Projecteur frontal et caméra 3 |
| Vue 4 | l1_c1, l2_c1, l6_c1 | Projecteur à droite et caméra frontale |
| Vue 5 | l1_c2, l2_c2, l6_c2 | Projecteur à droite et caméra 2 |
| Vue 6 | l1_c3, l2_c3, l6_c3 | Projecteur à droite et caméra 3 |
| Vue 7 | l4_c1, l5_c1, l7_c1 | Projecteur à gauche caméra frontale |
| Vue 8 | l4_c2, l5_c2, l7_c2 | Projecteur à gauche et caméra 2 |
| Vue 10 | i_140, i_170, i_180, i_190, i_210, i_230, i_250, l8_c1, l8_c2, l8_c3, r_0, ..., r_30, r_335, ..., r_355 | Capture frontale |
| Vue 14 | r_180, ... , r_215 | Capture prise de derrière |

TABLE 2 – Les dix-neuf (33 - 14) *clusters* en colonnes obtenus par le *multi-clustering*

## 4.2 Extraction des caractéristiques

Chaque image a été encodée en utilisant le vit_b16[2] pour obtenir une représentation vectorielle de dimension $d = 768$. L'analyse en composantes principales avait permis la réduction de la dimension à $d_r = 53$ conservant variance expliquée à $50\%$.

## 4.3 Évaluation du *multi-clustering*

### 4.3.1 Modèles concurrents

Afin d'évaluer l'approche proposée dans ce papier, le multi-clustering a été comparé aux deux méthodes paramétriques de l'état de l'art : le Modèle de Mélange Gaussien (*GMM*) [4] et le Modèle aux Blocs Latents (*LBM*) [6].

### 4.3.2 Paramétrisation

Le multi-clustering non-paramétrique possède 4 paramètres de concentration $(a_\gamma, b_\gamma)$ et $(a_\alpha, b_\alpha)$. L'expérience réalisée dans ce travail, fixe ces paramètres à $(1, 1)$ et $(5, 1)$ respectivement, le nombre d'itérations utilisé est $M = 20$.
GMM et LBM sont deux modèles paramétriques ; GMM possède un paramètre $k$ (nombre de *clusters* lignes). Une recherche exhaustive en optimisant le critère *ICL* [5] a été réalisée sur les élements de la liste $\{60, 76, 100, 200\}$ pour trouver le $k$ optimal. LBM quant à lui possède deux paramètres $k$ et $l$ (nombre de *clusters* lignes et colonnes). Une recherche similaire dans les listes $\{60, 76, 100, 200\}$ resp. $\{10, 14, 18, 20, 30, 33\}$ a été realisée pour estimer les $k$ et $l$ optimaux. L'étape d'extraction des caractéristiques et la réduction de dimension est communes aux trois méthodes.

### 4.3.3 Critères d'évaluation

La qualité des partitions lignes inférées (l'unique partition ligne du *GMM* et *LBM* et les multiples partitions inférées par le *multi-clustering*) ont été comparées en utilisant la variable *Material* comme annotation externes (labels) : trois métriques avaient été évaluées sur les résultats des trois méthodes sur la base de cette annotation : l'indice de Rand (RI), la pureté (purity), et l'information mutuelle normalisée (NMI).

### 4.3.4 Résultats

Le *multi-clustering* non-paramétrique avait inféré $(33)$ *cluster* en colonne (Vues) : seuls $(19)$ d'entre eux constituent réellement des agglomérations des colonnes originales, les $(14)$ autres *clusters* sont des singletons. Ceci est peut-être dû à une surestimation du nombre de *clusters* en colonne. Le tableau 2 présente certaines *v*ues ainsi que les variables qui y sont affectées ainsi que les interprétations proposées.
Les résultats des évaluations métriques sont présentées dans le tableau 1 et montrent que les solutions obtenues avec le *multi-clustering* enregistrent de meilleurs scores comparées aux approches classiques (*GMM* et *LBM*). Ces dernières

2. https://github.com/faustomorales/vit-keras

FIGURE 3 – Un exemple de partition inférée par le *multi-clustering*. Elle représente la vue 2 et les partitions lignes, de haut en bas (23, 0, 65) respectivement.

présentent des indices de Rand (RI) comparables à celui du *multi-clustering* mais les scores de la pureté et de l'information mutuelle normalisée sont beaucoup plus faibles. A noter que seules les vues (2,1,3 et 10) sont présentées dans le tableau pour *multi-clustering* pour alléger la lecture de ce dernier, les autres vues présentent des métriques similaires. En plus des bons scores enregistrés par les *clusters* lignes du *multi-clustering*, les *clusters* colonnes (vues) préservent la topologie spatiale des prises comme l'illustrent le tableau 2 et la figure 2.
En effet, la figure 2 montre les positions spatiales des différentes vues. elles sont ordonnées et forment des partitions compactes d'un point de vue spatial. La figure 3 montre un exemple de partition inférée par le *multi-clustering*. Elle représente la vue 2, et les quelques partitions lignes correspondantes, respectivement, $(23, 0, 65, 162)$ de haut en bas. Les résultats montrent que les *clusters* lignes regroupent des objets similaires (même nature, formes,...). Ces résultats sont également vérifiés pour les $(33)$ autres vues.

Contrairement au *multi-clustering*, le modèle à bloc latent (*LBM*) a échoué dans la tâche du regroupement des prises similaires dans le même *cluster*. La figure 4 montre un exemple de bloc latent regroupant dans le même *cluster* colonne $(0)$ différentes prises telles que (r_85, r_235, et r_275) qui restent différentes et éloignées du point de vue spatial. De plus, il regroupe dans la même partition ligne des objets différents d'un point de vue humain (formes différentes,...) ; pour certains objets affectés à ce bloc latent, les affectations par *multi-clustering* conservent une cohérence (voir exemple sur la vue $(3)$ de la figure 5).

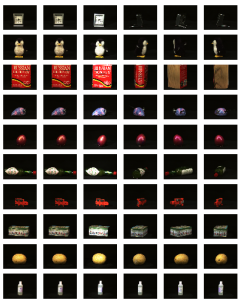


FIGURE 4 – Bloc latent du *cluster* en colonne (0) et *cluster* ligne (75) inféré par le *LBM*

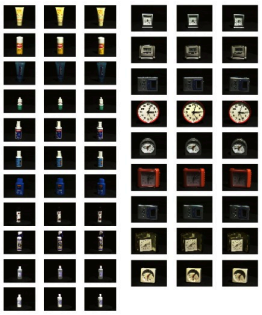


FIGURE 5 – *Cluster* ligne (3) (droite) et (18) (gauche) de la vue (3) inférés par le *multi-clustering*.

## 5 Conclusion et perspectives

Dans ce papier, nous avons proposé un *framework* pour inférer de multiples solutions de *clustering* sur un jeu de données d'images multivues. Notre *framework* combine un *vision transformers* pré-entraîné avec un algorithme du *multi-clustering*. Le problème majeur des algorithmes de *multi-clustering* paramétriques est l'étape de sélection de modèle lorsque le nombre de modèles possibles est grand. Nous avons proposé une approche bayésienne non paramétrique, qui permet d'estimer la structure du modèle durant l'inférence. Notre approche a fournit des blocs homogènes et cohérents. De plus, elle a surpassé les méthodes traditionnelles telles que le *GMM* et *LBM*.
Une de nos perspectives consiste à étudier la possibilité d'unifier notre *framework* afin de joindre l'apprentissage profond des représentations à l'aide des *transformers* et l'étape du *multi-clustering*.

## Références

[1] Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6) :1152–1174, 1974.

[2] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6) :1152 – 1174, 1974.

[3] Timothy Baldwin and J. Ford. Transfer of training : A review and directions for future research. *Personnel Psychology*, 41 :63–105, 03 1988.

[4] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3) :803–821, 1993.

[5] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7) :719–725, jul 2000.

[6] Charles Bouveyron, Laurent Bozzi, Julien Jacques, and François-Xavier Jollois. The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves. *Journal of the Royal Statistical Society : Series C Applied Statistics*, December 2017.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet : A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale, 2020.

[10] Giuliano Galimberti and Gabriele Soffritti. Model-based methods to identify multiple cluster structures in a data set. *Computational Statistics & Data Analysis*, 52(1) :520–536, 2007.

[11] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1) :103–112, 2005.

[12] Etienne Goffinet. *Multi-Block Clustering and Analytical Visualization of Massive Time Series from Autonomous Vehicle Simulation*. Theses, Université Paris 13 Sorbonne Paris Nord, December 2021.

[13] Etienne Goffinet, Anthony Coutant, Mustapha Lebbah, Hanane Azzag, and Loïc Giraldi. CONDITIONAL LATENT BLOCK MODEL : A MULTIVARIATE TIME SERIES CLUSTERING APPROACH FOR AUTONOMOUS DRIVING VALIDATION. working paper or preprint, January 2022.

[14] Yue Guan, Jennifer Dy, Donglin Niu, and Zoubin Ghahramani. Variational inference for nonparametric multiple clustering. 01 2010.

[15] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.

[16] Reda Khoufache, Mohamed Djallel Dilmi, Hanene Azzag, Etienne Gofinnet, and Mustapha Lebbah. Emerging properties from bayesian non-parametric for multiple clustering : Application for multi-view image dataset. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 31–38, 2022.

[17] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. page 77, 01 2009.

[18] Hongfu Liu, Ming Shao, Sheng Li, and Yun Fu. Infinite ensemble for image clustering. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1745–1754, New York, NY, USA, 2016. Association for Computing Machinery.

[19] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[20] Durvesh Malpure, Onkar Litake, and Rajesh Ingle. Investigating transfer learning capabilities of vision transformers and cnns by fine-tuning a single trainable block, 2021.

[21] Vikash Mansinghka, Eric Jonas, Cap Petschulat, Beau Cronin, Patrick Shafto, and Joshua Tenenbaum. Cross-categorization : A method for discovering multiple overlapping clusterings. 01 2009.

[22] Jeffrey W. Miller. An elementary derivation of the chinese restaurant process from sethuraman's stick-breaking process, 2018.

[23] Noémi Schuurman, Raoul Grasman, and Ellen Hamaker. A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate behavioral research*, 51 :1–22, 03 2016.

[24] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2) :639–650, 1994.

[25] F. Tian, B. Gao, Q. Cui, Enhong Chen, and T.-Y Liu. Learning deep representations for graph clustering. *Proceedings of the National Conference on Artificial Intelligence*, 2 :1293–1299, 01 2014.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[27] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters, 2016.

# Réduction de la Dimension et Classification: approche jointe

Z. Tighidet, L. Labiod, M. Nadif

Centre Borelli UMR 9010, Université Paris Cité, France

zineddine.tighidet@etu.u-paris.fr | lazhar.labiod@u-paris.fr | mohamed.nadif@u-paris.fr

## Résumé

*Pour permettre à la fois un clustering et une réduction de la dimensionalité, nous présentons une approche combinant simultanément l'analyse en composantes principales (*ACP*) et l'algorithme (*CEM*) qui maximise la vraisemblance classifiante. Nous dérivons un algorithme que nous évaluons dans une première étape sur des données simulées.*

## Mots-clés

*Classification, Modèles de mélange, CEM algorithme, ACP.*

## Abstract

*To deal with both clustering and dimensionality reduction, we present an approach combining simultaneously principal component analysis (*PCA*) and an algorithm (*CEM*) that maximizes the complete data likelihood. We derive an algorithm that we evaluate in a first step on simulated data.*

## Keywords

*Clustering, Mixture models, CEM algorithm, PCA.*

## 1 Introduction

Dans le domaine de l'apprentissage automatique, la capacité à classifier les données de manière efficace et précise est cruciale. Une méthode populaire pour y parvenir est l'utilisation de l'ACP, qui réduit la dimensionnalité des données tout en conservant les informations les plus pertinentes, suivie d'un algorithme EM [3] qui utilise un Modèle de Mélange Gaussien (GMM) [1, 2] pour classifier les données. Bien que ces approches aient été appliquées avec succès d'une manière séquentielle dans de nombreuses applications, leur combinaison d'une manière simultanée a le potentiel d'améliorer encore davantage la précision et la vitesse d'exécution comme illustré dans Figure 1.

Dans cet article, nous présentons une approche qui utilise simultanément la PCA et l'algorithme CEM [2] pour améliorer les performances. Nous illustrons l'intérêt de cette approche sur plusieurs ensembles de données simulées et comparons ses performances à celles d'autres techniques couramment utilisées. Nos résultats montrent que l'approche proposée surpasse les méthodes traditionnelles et ouvre une voie prometteuse pour de futures recherches dans le domaine de la classification.

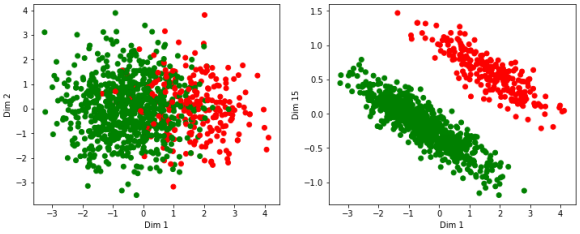



FIGURE 1 – Projection ACP des données de Chang sur la première et deuxième CP (à gauche) et sur la première et quinzième CP (à droite).

## 2 Notre Proposition

Les données sont représentées par une matrice $\mathbf{X} = (x_{ij})$ de taille $n \times d$ , où les $x_{ij}$ sont supposés être échantillonnés à partir d'une distribution paramétrique donnée de densité $\Phi$. La valeur de chaque entrée de la matrice de données dépend de la ligne latente des paramètres du modèle. La partition de l'ensemble de lignes en $g$ classes est représentée par la matrice de classification latente $\mathbf{Z} = (z_{ik})$, avec $\sum_{k=1}^{g} z_{ik} = 1$, où $z_{ik} = 1$ si la ligne $i$ appartient au classe de lignes $k$ et $z_{ik} = 0$ dans le cas contraire. Nous écrivons également $z_i \in \{1, \ldots, g\}$ comme étant l'indice de la classe de $i$. Nous considérons le problème de la réduction de la dimension et du clustering simultanément. L'idée est de combiner l'ACP et le CEM via une régularisation. De cette manière, nous proposons la fonction objective suivante à minimiser :

$$\mathcal{F}(\mathbf{X}; \mathbf{Q}, \mathbf{Z}, \mathbf{\Sigma}, \mathbf{S}) = ||\mathbf{X} - \mathbf{B}\mathbf{Q}^\top||^2 + \delta||\mathbf{B} - \mathbf{M}||^2 - \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \log(\pi_k \varphi(\mathbf{m}_i^p, (\mathbf{s}_k, \mathbf{\Sigma}_k)) \quad (1)$$

avec $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{Q} = \mathbf{X}^T\mathbf{B} \in \mathbb{R}^{d\times p}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$ est une matrice orthonormée en colonnes, $\mathbf{M} \in \mathbb{R}^{n\times p}$, $\mathbf{s}_k = (s_{k1}, \ldots, s_{kp})$ le centroide de la $k$ème classe. $\mathbf{Z} \in \{0, 1\}^{n\times g}$ matrice de classification. $\varphi(\mathbf{m}_i^p, (\mathbf{s}_k, \mathbf{\Sigma}_k))$ est une distribution gaussienne dans $\mathbf{M} \in \mathbb{R}^{n\times p}$. Les $\pi_k$ sont les probabilités à priori du mélange.

# 3 Algorithme

L'estimation des paramètres est réalisée d'une manière alternative, permettant ainsi un renforcement mutuel entre la tâche de la réduction de la dimensionalité et la tâche de clustering réalisée sur une matrice de taille réduite. Ci-après, les principales étapes de l'algorithme proposé.

**Algorithm 1** Classification avec PCA-CEM

**Entrée :** $\mathbf{X}$, $g$ nombre de classes, $p$ le nombre de dimensions latentes, $\delta$ paramètre de régularisation
**Initialisation :** Calculer $\mathbf{B}$ $\mathbf{Q}$ avec une PCA, $\mathbf{Z}$, $\mathbf{\Sigma}$ et $\mathbf{S}$, avec une CEM appliquée sur $\mathbf{B}$
**repeat**
**Etape 1.** Calculer $\mathbf{M}$,
**Etape 2.** Mettre à jour $\mathbf{Z}, \mathbf{S}, \mathbf{\Sigma}$ et des proportions $\pi_k$'s à l'aide de CEM.
**Etape 3.** Calculer $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$ où

$$\mathbf{U}\Delta\mathbf{V}^\top = \text{svd}(\mathbf{X}\mathbf{Q} + \delta\mathbf{M}),$$

**Etape 4.** Calculer $\mathbf{Q} = \mathbf{X}^\top\mathbf{B}$
**until** convergence
**return** $\mathbf{Z}$, $\mathbf{B}$, $\mathbf{M}$, $\Sigma$ et $\pi_k$'s.

# 4 Expériences

L'algorithme `PCA-CEM` est intialisé au hasard. Sur la base de 20 initialisations, on retient le meilleur essai qui minimise (1). Pour la convergence, sur les données synthétiques décrites ci-après, l'algorithme a nécessité moins de 20 itérations. Quant au paramètre $\delta$, dans nos expériences nous l'avons fixé à $10^{-6}$.

## 4.1 Données Synthétiques

Les caractéristiques des données `FCPS` qui ont servi d'évaluation et de point de comparaison sont illustrées dans Table 1.

TABLE 1 – Caractéristiques des jeux de données `FCPS`

| | Caractéristiques `FCPS` | | | |
|---|---|---|---|---|
| Données | Type | #lignes | #colonnes | #classes |
| `Atom` | Numérique | 800 | 3 | 2 |
| `Chainlink` | Numérique | 1000 | 3 | 2 |
| `Hepta` | Numérique | 212 | 3 | 7 |
| `Lsun3D` | Numérique | 404 | 3 | 4 |
| `Tetra` | Numérique | 400 | 3 | 4 |

## 4.2 Evaluation et comparaisons

Les résultats obtenus sur des données synthétiques (`FCPS`) sont illustrées dans Table 1. On remarque que les performances de la méthode proposée (`PCA-CEM`) sont nettement meilleures que les autres méthodes de référence, `K-means` [7], `ClusPCA` [8], `MFA` [5], `Deep-K-means` [4] et `PCA-GMM` [6] sauf pour les données `Chainlink`. Le fait de réaliser les deux étapes d'une manière séquentielle peut donner un objectif qui est très éloigné de la classification qui intervient juste après. En revanche, `PCA-CEM` effectue simultanément les deux étapes, ce qui crée un renforcement mutuel entre la réduction de la dimension et la classification.

TABLE 2 – Performance de Classification (NMI % and ARI % ) sur les données `FCPS`

| Méthode | `FCPS` | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `Atom` | | `Chainlink` | | `Hepta` | | `Lsun3D` | | `Tetra` | |
| | NMI | ARI | NMI | ARI | NMI | ARI | NMI | ARI | NMI | ARI |
| `K-means` | 0.29 | 0.18 | 0.0 | 0.0 | 1.0 | 1.0 | 0.73 | 0.6 | 1.0 | 1.0 |
| `ClusPCA` | 0.56 | 0.17 | 0.0 | 0.0 | 1.0 | 1.0 | 0.65 | 0.59 | 1.0 | 1.0 |
| `MFA` | 0.16 | 0.08 | 0.25 | 0.17 | 1.0 | 1.0 | 0.68 | 0.65 | 1.0 | 1.0 |
| `Deepkmeans` | 0.33 | 0.23 | 0.07 | 0.10 | 0.90 | 0.88 | 0.84 | 0.86 | 0.72 | 0.69 |
| `PCA-GMM` | 0.00 | 0.00 | 0.31 | 0.28 | 0.90 | 0.78 | 0.74 | 0.61 | 0.64 | 0.54 |
| Ours | 0.93 | 0.97 | 0.17 | 0.09 | 1.0 | 1.0 | 0.98 | 0.99 | 1.0 | 1.0 |

# 5 Conclusion

Dans cet article nous avons présenté une nouvelle approche de classification se basant sur une minimisation jointe et régularisée de l'objectif de l'`ACP` et de `CEM`. On a montré que l'algorithme qui en découle donne des résultats encourageants qui mériteront bien entendu d'être confirmés sur des données réelles et de grande taille.

# Références


[1] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3) :803–821, 1993.

[2] Gilles Celeux and Gérard Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793, 1995.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[4] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means : Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138 :185–192, 2020.

[5] Zoubin Ghahramani, Geoffrey E Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Citeseer, 1996.

[6] Johannes Hertrich, Dang-Phuong-Lan Nguyen, Jean-Francois Aujol, Dominique Bernard, Yannick Berthoumieu, Abdellatif Saadaldin, and Gabriele Steidl. Pca reduced gaussian mixture models with applications in superresolution, 2022.

[7] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2) :451–461, 2003. Biometrics.

[8] Angelos Markos. Joint dimension reduction and clustering in R (part I).

# Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering*

Julien Ah-Pine[1,2,3]

[1] Université de Lyon, Lyon 2, ERIC UR 3083
[2] Université Clermont Auvergne, LMBP CNRS, UMR 6620
[3] Université Clermont Auvergne, CERDI CNRS IRD, UMR 6587

julien.ah-pine@univ-lyon2.fr

## Résumé

*Nous nous intéressons à la tâche de clustering du point de vue graphe à l'instar du partionnement spectral (spectral clustering). Dans ce cas, la matrice d'affinité qui mesure l'intensité du lien (arête du graphe) pour chaque paire d'éléments (sommets du graphe) joue un rôle crucial. Plusieurs travaux antécédents ont montré l'intérêt de transformer une matrice d'affinité initiale de sorte à satisfaire certaines propriétés. La bistochasticité est une condition pertinente à cet égard. Dans ce travail, nous mettons en avant une autre condition : l'idempotence. Par la suite, En utilisant les propriétés existantes entre les matrices bistochastiques et idempotentes d'une part, et leurs matrices Laplaciennes associées d'autre part, nous proposons une nouvelle méthode d'apprentissage non-supervisé de matrice d'affinité. Notre procédure d'optimisation repose sur la méthode des multiplicateurs de Lagrange avec directions alternées (ADMM). Des résultats expérimentaux montrent l'intérêt pratique de notre approche.*

## Mots-clés

*Clustering, Matrice d'affinité, Bistochasticité, Idempotence, ADMM.*

## Abstract

*We are interested in graph based clustering such as spectral clustering. In this context, the affinity matrix that provides the strenght of the similarity between each pair of elements plays a crucial role. Several previous works have showed that transforming a given affinity matrix so that it becomes double stochastic was beneficial. In this work, we highlight another property : idempotency. By leveraging the relationships between double stochastic and idempotent matrices on the one hand, and their related Laplacian matrices on the other hand, we introduce a new unsupervised learning method for affinity matrices. Our learning algorithm is based on ADMM. Some experimental results are provided in order to demonstrate the interest of our proposal.*

## Keywords

*Clustering, Affinity matrix, doubly stochasticity, Idempotence, ADMM.*

*Cette communication est issue de l'article suivant : [**?**].

## 1 Contexte et travaux antérieurs

La tâche de clustering consiste à partitionner un ensemble d'éléments en des sous-ensembles homogènes appelés clusters. Soit un ensemble de $n$ vecteurs $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ appartenant à $\mathbb{R}^p$, que nous cherchons à analyser. Nous nous intéressons à la partition en $k$ clusters $C = \{C_1, \ldots, C_k\}$ qui minimise le critère SSE (Sum of Squared Errors) suivant :

$$\sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\phi(\mathbf{x}_i) - \mathbf{c}_j\|^2 \tag{1}$$

où $\phi : \mathbb{R}^p \to \mathbb{F}$ est une projection des $\{\mathbf{x}_i\}$ dans un espace de grande dimension $\mathbb{F}$, $\mathbf{c}_j = \sum_{\mathbf{x}_i \in C_j} \phi(\mathbf{x}_i)/n_j$ est le vecteur moyen du cluster $C_j$ qui est de cardinal $n_j$ et $\|.\|$ est la norme Euclidienne dans $\mathbb{F}$.
Le critère SSE peut être formalisé à l'aide de la matrice de noyau $\mathbf{K}$ de terme général $\mathbf{K}_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle = \kappa(\mathbf{x}_i, \mathbf{x}_{i'})$ où $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ est une fonction noyau. Il s'agit dans ce cas de déterminer $\mathbf{X}$, la matrice de $\mathbb{R}^{n \times n}$ qui minimise la fonction objectif suivante :

$$\mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{2}$$

où Tr est l'application trace dans $\mathbb{R}^{n \times n}$, $\mathbf{I}_n$ est la matrice identité d'ordre $n$, et $\mathbf{X}$ est de terme général :

$$\mathbf{X}_{ii'} = \begin{cases} 1/n_j & \text{si } \mathbf{x}_i \text{ et } \mathbf{x}_{i'} \text{ sont dans } C_j, \\ 0 & \text{sinon.} \end{cases} \tag{3}$$

La matrice $\mathbf{X}$ ainsi définie possède plusieurs propriétés. Plus précisément [**?**] montre que la minimisation du SSE peut s'exprimer de façon équivalente comme suit :

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{4}$$
$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2, \mathrm{Tr}(\mathbf{X}) = k.$$

où $\mathbf{0}_n$ est la matrice nulle d'ordre $n$, $(\mathbf{X} \geq \mathbf{0}_n) \Leftrightarrow (\mathbf{X}_{ii'} \geq 0, \forall i, i' = 1, \ldots, n)$, $\mathbf{X}^\top$ est la matrice transposée de $\mathbf{X}$ et $\mathbf{e}_n$ est le vecteur rempli de 1 de dimension $n$.

La matrice $\mathbf{X}$ recherchée est ainsi non-négative, symétrique, bistochastique (les sommes de chaque ligne et de chaque colonne valent 1), idempotente et de trace égale à $k$ le nombre de clusters désiré. En fait, il existe une bijection entre l'ensemble des partitions d'un ensemble de $n$ éléments et l'ensemble des classes d'équivalence des matrices bistochastiques et idempotentes pour la relation $\mathbf{X} \sim \mathbf{Y}$ si et seulement si il existe une matrice de permutation $\mathbf{P}$ telle que $\mathbf{X} = \mathbf{PYP}^\top$ [**?**].
Par exemple la partition $\{(a, e), (b, c, d)\}$ peut être représentée par les matrices bistochastiques et idempotentes suivantes appartenant à la même classe d'équivalence :

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{pmatrix} \end{array} \sim \begin{array}{c} \\ a \\ e \\ b \\ c \\ d \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} a & e & b & c & d \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

Le problème ainsi formulé, donne un point de vue graphe à la tâche de clustering : $\mathbf{K}$ (à condition d'être non-négative -comme pour le noyau Gaussien par exemple-) peut être vue telle une matrice d'adjacence pondérée d'un graphe sans structure particulière et $\mathbf{X}$ peut être interprétée comme la matrice d'adjacence pondérée d'un graphe représentant une partition des sommets. Il s'agit alors d'approximer $\mathbf{K}$ par $\mathbf{X}$ au sens de la norme de Frobenius. En effet, soit $\mathrm{Tr}(\mathbf{X}^\top\mathbf{Y}) = \langle \mathbf{X}, \mathbf{Y} \rangle_F$ le produit scalaire de Frobenius dans $\mathbb{R}^{n\times n}$. Si $\mathbf{X}$ vérifie les contraintes stipulées dans (4), alors il est facile de montrer que :

$$\min_{\mathbf{X}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \Leftrightarrow \min_{\mathbf{X}} \|\mathbf{K} - \mathbf{X}\|_F^2 \tag{5}$$

Le modèle d'optimisation 4, appelé 0-1 Semi-Definite Program dans [**?**], est NP-difficile en raison de la nature discrète et de l'idempotence de la matrice $\mathbf{X}$. En pratique, une démarche heuristique permettant de résoudre de façon approchée 4 consiste à (i) définir un problème relaxé solutionnable en temps polynomial, (ii) discrétiser la solution optimale du problème relaxé afin d'obtenir une solution réalisable du problème initial.
De nombreuses méthodes d'apprentissage de matrice d'affinité et de clustering découlent de cette démarche. Dans ce travail nous nous penchons plus particulièrement, sur les approches présentées dans [**?**] et [**?**]. Ces travaux reviennent à remplacer $\mathbf{K}$ par une matrice $\mathbf{X}$ non-négative, symétrique et bistochastique (étape (i)). Autrement dit, la contrainte d'idempotence est abandonnée, le nombre de cluster $k$ n'est alors plus associé à la trace et les contraintes restantes sont toutes linéaires. Des procédures efficaces sont proposées pour déterminer $\mathbf{X}$ : dans [**?**] il s'agit d'une version symétrique de l'algorithme de Sinkhorn-Knoop [**?**] dénoté SSK, alors que dans [**?**] est introduit l'algorithme DSN (Double Stochastic Normalization). Plus précisément, l'algorithme de Sinkhorn-Knoop vise à minimiser la divergence de Kullback-Leibler entre $\mathbf{K}$ et $\mathbf{X}$ alors que l'approche DSN résulte de la minimisation de la distance de Frobenius entre $\mathbf{K}$ et $\mathbf{X}$. Une fois $\mathbf{X}$ déterminée une méthode de discrétisation est utilisée. Le spectral clustering qui, en bref, applique l'algorithme des $k$-means sur les $k$ premiers vecteurs propres de $\mathbf{X}$ est une approche classique à cet égard (étape (ii)).

# 2 Approche proposée

Contrairement aux deux approches précédentes, nous cherchons à tenir compte de la contrainte d'idempotence tout en évitant de se ramener à un problème NP-difficile. Néanmoins, nous ne considérons pas le nombre de clusters $k$ comme paramètre de notre modèle et n'imposons donc pas $\mathrm{Tr}(\mathbf{X}) = k$. L'approximation basée sur la distance de Frobenius reste centrale dans notre approche qui vise, en somme, à définir un problème relaxé du modèle suivant :

$$\begin{aligned} &\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 \\ &\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2. \end{aligned} \tag{6}$$

$\mathbf{X}$ étant bistochastique, la matrice des degrés vaut $\mathbf{I}_n$ et la matrice Laplacienne associée à $\mathbf{X}$ est donnée par $\mathbf{L}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}$. Clairement, le problème (6) peut être reformulé de façon équivalente en fonction de $\mathbf{L}_\mathbf{X}$ comme suit :

$$\begin{aligned} &\min_{\mathbf{L}_\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 \\ &\text{s.t. } \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^2. \end{aligned} \tag{7}$$

où $\mathbf{n}_n$ est le vecteur nul de dimension $n$.
Nous exploitons à présent les relations algébriques existantes entre $\mathbf{X}$ et $\mathbf{L}_\mathbf{X}$. En effet, ces deux matrices étant symétriques et idempotentes, elles représentent des projections orthogonales. De façon plus singulière, l'une est l'unique projecteur orthogonale complémentaire de l'autre et *vice-versa* : l'image de $\mathbf{X}$ est le noyau de $\mathbf{L}_\mathbf{X}$, l'image de $\mathbf{L}_\mathbf{X}$ est le noyau de $\mathbf{X}$ et nous avons la relation suivante, centrale dans notre travail :

$$\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}_n \tag{8}$$

Nous proposons de relaxer (6) et (7) en considérant le modèle suivant :

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{L}_\mathbf{X}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2 \\ &\text{s.t. } \begin{cases} \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \\ \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \\ \mathbf{X} + \mathbf{L}_\mathbf{X} = \mathbf{I}_n. \end{cases} \end{aligned} \tag{9}$$

où $\mu > 0$ est un paramètre de pénalité.
Notre modèle intitulé DSNI (Doubly Stochastic and Nearly Idempotent), consiste en un apprentissage joint de $\mathbf{X}$ et de sa matrice Laplacienne associée $\mathbf{L}_\mathbf{X}$. Il est facile de montrer que sous la condition $\mathbf{X} + \mathbf{L}_\mathbf{X} = \mathbf{I}_n$, les trois propriétés qui suivent sont équivalentes : $\mathbf{X} = \mathbf{X}^2$, $\mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^2$, $\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}_n$. Cependant, ces propriétés étant la source première de la complexité des problèmes (6) et (7), nous ne les intégrons pas dans les contraintes de notre modèle. Pour pallier à ce manque, nous ajoutons, en revanche, un terme de pénalité dans la fonction objectif, $\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2$, afin d'encourager les solutions obtenues à être ainsi quasi-idempotentes.

Le problème DSNI (9) étant bi-convexe, nous pouvons utiliser la méthode ADMM (voir par exemple [**?**]) comme procédure d'optimisation. Les différentes étapes sont alors les suivantes :

0. Initialisation : $\mathbf{X}^0 \leftarrow \mathbf{K}$ (en ayant au préalable annuler les valeurs négatives de $\mathbf{K}$ le cas échéant).
1. Déterminer $\mathbf{L}_{\mathbf{X}}^{t+1}$ avec $\mathbf{X}^t$ fixé :

$$\begin{aligned}\mathbf{L}_{\mathbf{X}}^{t+1} \leftarrow \underset{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}}{\arg\min} & \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 \\ & + \frac{\mu}{2}\|\mathbf{X}^t \mathbf{L}_{\mathbf{X}}\|_F^2 \\ & + \frac{\rho}{2}\|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \\ & \text{s.t. } \mathbf{L}_{\mathbf{X}} \leq \mathbf{I}_n, \mathbf{L}_{\mathbf{X}} = \mathbf{L}_{\mathbf{X}}^\top, \mathbf{L}_{\mathbf{X}}\mathbf{e}_n = \mathbf{n}_n.\end{aligned} \tag{10}$$

2. Déterminer $\mathbf{X}^{t+1}$ avec $\mathbf{L}_{\mathbf{X}}^{t+1}$ fixé :

$$\begin{aligned}\mathbf{X}^{t+1} \leftarrow \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\arg\min} & \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 \\ & + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 \\ & + \frac{\rho}{2}\|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \\ & \text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n.\end{aligned} \tag{11}$$

3. Déterminer $\mathbf{U}^{t+1}$ :

$$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{X}^{t+1} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n \tag{12}$$

4. Répéter 1., 2., 3. tant qu'une condition d'arrêt n'est pas satisfaite.

Les sous-problèmes (10) et (11) peuvent être résolus efficacement par projections successives sur des ensembles convexes (voir par exemple [**?**]).

# 3 Validation empirique de l'approche proposée

Nous avons testé notre approche sur plusieurs jeux de données réels classiques disponibles en ligne [1]. Le protocole expérimental est le suivant : calcul de $\mathbf{K}$ en utilisant un noyau Gaussien ; approximation de $\mathbf{K}$ par une matrice d'affinité bistochastique $\mathbf{X}$ obtenue par SSK ou DSN ou DSNI (sauf pour la baseline) ; application du spectral clustering [**?**] sur $\mathbf{X}$ en fixant $k$ au nombre correct de clusters ; comparaison de la partition obtenue et de la vérité terrain en utilisant la mesure NMI (Normalized Mutual Information). Pour le noyau Gaussien, l'hyperparamètre $\sigma^2$ est fixé à $p$ et pour DSNI le paramètre de pénalité $\mu$ est fixé à $\sqrt{n}$.
Les résultats obtenus sont donnés dans la Table 1. La colonne SC représente la baseline et utilise le spectral clustering directement sur $\mathbf{K}$ la matrice de noyau Gaussien. Sur l'ensemble des jeux de données, nous constatons que DSNI donne de meilleurs résultats que SSK et DSN ce qui valide l'intérêt de notre modèle.

1. `https://archive.ics.uci.edu/ml/index.php`

| Dataset | $n$ | $p$ | $k$ | SC | SSK | DSN | DSNI |
|---|---|---|---|---|---|---|---|
| Glass | 214 | 9 | 6 | 0.253 | 0.276 | 0.243 | **0.297** |
| Ionosphere | 351 | 34 | 2 | 0.038 | 0.066 | 0.076 | **0.131** |
| Breast cancer | 569 | 30 | 2 | 0.010 | 0.010 | 0.010 | **0.670** |
| Yeast | 1484 | 8 | 10 | 0.070 | 0.258 | 0.256 | **0.263** |
| Digits | 1797 | 64 | 10 | 0.015 | 0.044 | 0.743 | **0.767** |

TABLE 1 – Statistiques des jeux de données et mesures NMI des différentes méthodes.

# Références


[1] J. Ah-Pine. Learning doubly stochastic and nearly idempotent affinity matrix for graph-based clustering. *European Journal of Operational Research*, 299(3) :1069–1078, 2022.

[2] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3) :367–426, 1996.

[3] S. Boyd, N. Parikh, and E. Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[4] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering : Analysis and an algorithm. *Advances in neural information processing systems*, 2 :849–856, 2002.

[5] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1) :186–205, 2007.

[6] J. Peng and Y. Xia. A new theoretical framework for k-means-type clustering. In *Foundations and advances in data mining*, pages 79–96. Springer, 2005.

[7] R. Sinkhorn. Two results concerning doubly stochastic matrices. *The American Mathematical Monthly*, 75(6) :632–634, 1968.

[8] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4) :395–416, 2007.

[9] F. Wang, P. Li, and A. C. Konig. Learning a bistochastic data similarity matrix. In *2010 IEEE International Conference on Data Mining*, pages 551–560. IEEE, 2010.

[10] R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 294–301. IEEE, 2005.

[11] R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in neural information processing systems*, pages 1569–1576, 2007.

# Un algorithme simple et efficace pour la sériation circulaire

Mikhaël Carmona[1,2], Victor Chepoi[1], Guyslain Naves[1], Pascal Préa[1,2]

[1] Aix-Marseille Université, Université de Toulon, CNRS, LIS
[2] École Centrale de Marseille

{mikhael.carmona, victor.chepoi, guyslain.naves, pascal.prea}@lis-lab.fr

## Résumé

*Suite aux travaux d'Armstrong, Guzmán et Sing Long sur la sériation circulaire stricte nous avions cherché à améliorer leur solution. La continuité naturelle maintenant est de montrer qu'il existe aussi un algorithme simple et efficace pour la sériation circulaire dans le cas général. Dans ce papier nous montrons une puissante propriété qui permet de déterminer si un ordre quasi-circulaire est aussi un ordre circulaire compatible à partir d'un PC-arbre. De cette propriété découle l'algorithme de reconnaissance pour la sériation circulaire.*

## Mots-clés

*Robinson, dissimilaritées, PC-arbres, sériation circulaire.*

## Abstract

*Following previous work from Armstrong, Guzmán and Sing on strict circular seriation we improved their result in a previous paper. What comes next naturally is to show that a simple and efficient algorithm exists for the general case too. In this paper we show a property that decides from a PC-tree representing all the quasi-circular orders if one of them is also a compatible circular order. From this property follows the recognition algorithm for circular seriation.*

## Keywords

*Robinson, dissimilarities, PC-trees, circular seriation.*

## 1 Introduction

Un enjeu majeur en classification et en analyse de données est de visualiser de simples structures relationelles et géométriques entre des objets à partir de leurs distances deux à deux. Le problème de la sériation (linéaire) introduit par Robinson [8] propose une solution à ce problème. Pour ce problème classique il existe 4 définitions équivalentes : la définition métrique, par les boules, par les 2-boules et par les clusters. Qui permettent d'attaquer le problème d'autant de façons différentes.

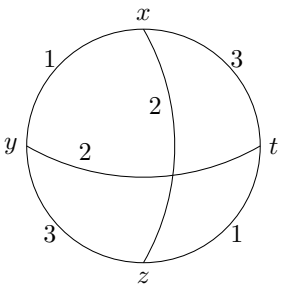
Pour rappel la définition métrique considère un ensemble fini $X$, une dissimilarité sur $X$ qui est une fonction sur $X \times X$ à valeurs positives ou nulles, symétrique et telle que $\forall x \in X, d(x,x) = 0$. Le couple $(X,d)$ est appelé espace. On cherche alors à déterminer si les points de $X$ peuvent être arrangés compatiblement avec $d$. Un tel ordre est dit compatible si :

$\forall x < y < z, d(x,z) \geq max\{d(x,y), d(y,z)\}$

On dit d'une dissimilarité qui admet un tel ordre compatible qu'elle est Robinson.

Pour la sériation circulaire, qui consiste étant donné un espace $(X,d)$ à déterminer si les points de $X$ peuvent être disposés sur un cercle de manière compatible avec $d$, il existe aussi 4 définitions qui ne sont pas équivalentes. Parmis elles on retrouve une définition par les boules, une par les 2-boules et une par les clusters qui sont des définitions par classes et une autre définition qui repose sur la convexité des boules/2-boules/clusters. Un problème étant qu'il n'y a pas dde structure de donnée pour représenter ces ordres, une solution pour ces trois premières, est de représenter l'ensemble des ordres compatibles par un PC-arbre.

Ces définitions présentent des problèmes, la définition par les 2-boules et celle par les clusters peuvent donner modéliser des ensembles non connexes. Inversement la définition par les boules est trop large et donne des ordres potentiellement non circulaires comme dans ce contre-exemple classique :




En effet, si l'on considère la définition de quasi-circulaire : un espace $(X,d)$ est quasi-circulaire si il existe un ordre circulaire $\lessdot$ tel que :

$\forall x \lessdot y \lessdot z \lessdot t, d(x,z) \geq min\{d(y,z), d(t,z)\}$.

Ce qui est respecté dans l'exemple ci-dessus, alors que la définition de cirulaire n'est pas respectée : Un espace $(X,d)$ est circulaire si il existe un ordre circulaire $<$ tel que pour tout $x, y \in X$ au moins un des deux espaces $(X^{\lessdot}_{xy}, d)$ $(X^{\lessdot}_{yx}, d)$ est Robinson. Avec $X^{\lessdot}_{xy} := \{t \in X : x \lessdot t \lessdot y\}$ un arc.

Enfin définissons pré-circulaire que l'on sait équivalent à circulaire [3] : Un espace $(X,d)$ est pré-circulaire si il existe un ordre circulaire $\lessdot$ tel que :

$\forall x \lessdot y \lessdot z \lessdot t$,
$d(x,y) \geq min\{max\{d(x,y), d(y,z)\},$
$max\{d(x,t), d(t,z)\}\}$.
L'idée de l'algortihme que nous présentons est de partir du cas le plus large (le cas quasi-circulaire) et de s'appuyer sur une contrainte supplémentaire pour s'assurer que cet ordre est aussi circulaire. Pour cela nous avons besoin de travailler sur le PC-arbre représentant l'ensemble des ordres quasi-circulaires compatibles avec la dissimilarité. Un algortihme en $O(n^3)$ pour construire un PC-arbre à partir d'une dissimilarité est connu [2].

# 2 Propriétés

Introduisons d'abord quelques définitions qui seront utile dans notre algorithme.

**Définition 1.** *L'*excentricitée *d'un point $x$ est $r_x := \max\{d(x,y) : y \in X\}$. Prenons un point $x \in X$, un point $y \in X$ est appelé* point excentrique *de $x$ if $d(x,y) = r_x$.*

**Définition 2.** *[4] Un PC-arbre, est un arbre $T$ avec deux types de noeuds interne : $P$ et $C$. Deux PC-arbres équivalents $T$ et $T'$ sont notés $T \sim T'$. Il existe deux transformations qui préservent l'équivalence :*

1. *On peut permuter librement tous les fils d'un P-noeud.*
2. *On peut changer l'ordre circulaire des voisins d'un C-noeud du sens horaire au sens anti-horaire.*

**Propriété 1.** *[3, Proposition 4.2] Soit $(X, d, \lessdot)$ un espace quasi-circulaire Robinson et $x \in X$. Si $y, z \in L_x$ et $x \lessdot y \lessdot z$ ou $y, z \in R_x$ et $z \lessdot y \lessdot x$, alors $d(x,y) \leq d(x,z)$.*

**Propriété 2.** *[3, Proposition 4.4] Soit $(X, d, \lessdot)$ un espace circulaire Robinson. Alors pour tout $x, y \in X$, $x' \in F_x$, $y' \in F_y$ avec $|\{x, x', y, y'\}| \geq 3$, une des affirmations suivantes est vérifiée :*
*(a)* $x \lessdot y \lessdot x' \lessdot y'$,
*(b)* $x \lessdot y' \lessdot x' \lessdot y$,
*(c)* $\{y, y'\} \cap F_x \neq \varnothing$ *or* $\{x, x'\} \cap F_y \neq \varnothing$.

Avec ces deux propriétés ci-dessus on peut montrer la propriété centrale de l'algorithme, l'idée étant que si on prend un PC-arbre représentant l'ensemble des quasi-circulaire et que l'on vérifie que les diamètres entre toutes les paires de points "ne se croisent pas" (c'est ce que dit la propriété 2) alors cette permutation est circulaire.

**Propriété 3.** *Soit $(X, d, \lessdot)$ un espace de dissimilarité, $(X, d, \lessdot)$ est circulaire si et seulement si toutes ces affimations sont vérifiées :*

1. *$(X, d, \lessdot)$ est quasi-circulaire*
2. *pour tout $x, y \in X$, $x' \in F_x$, $y' \in F_y$ avec $|\{x, x', y, y'\}| \geq 3$, une des affirmations suivantes est vérifiée :*
   *(a)* $x \lessdot y \lessdot x' \lessdot y'$,
   *(b)* $x \lessdot y' \lessdot x' \lessdot y$,
   *(c)* $\{y, y'\} \cap F_x \neq \varnothing$ *ou* $\{x, x'\} \cap F_y \neq \varnothing$.

# 3 L'algorithme

De cette propriété découle naturellement l'algorithme. On construit d'abord le PC-arbre représentant l'ensemble des ordres quasi-circulaires possibles avec l'algorithme proposé dans [2]. Parmis ces ordres on cherche maintenant à en trouver un qui respecte la propriété 2 pour cela on vérifie deux à deux tous les points du cercle en vérifiant si les diamètres se croisent et si ce n'est pas le cas on modifie la structure du PC-arbre pour croiser les diamètres, c'est à dire que si la structure de l'arbre le permet on interverti soit les deux points qui posent problème soit leurs points excentriques.

# Remerciements

Ce travail a bénéficié d'un financement de l'ANR via le projet DISTANCIA (ANR-17- CE40-0015) et d'une aide du gouvernement français au titre du Programme Investissements d'Avenir Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (Institut Archimède AMX-19-IET-009)

# Références

[1] S. Armstrong, C. Guzmán, and C.A. Sing Long, An optimal algorithm for strict circular seriation, SIAM Journal on Mathematics of Data Science, 3 (2021), 1223–1250.

[2] F. Brucker et C. Osswald, Hypercycles and dissimilarities, *Journal of Classification*, accepté, (2008).

[3] M. Carmona, V. Chepoi, G. Naves et P. Préa, A simple and optimal algorithm for strict circular seriation, submitted to Siam J. of Mathematics of Data Science (2022)

[4] W. Hsu et R. M. McConnell, *PC trees and circular-ones arrangements*, Theor. Comput. Sci., 296 (2003), pp 99-116

[5] L. Hubert, P. Arabie, et J. Meulman, Linear and circular unidimensional scaling for symmetric proximity matrices, *British Journal of Mathematical* Statistics and Psychology, 50 (1997), 253–284.

[6] L. Hubert, P. Arabie, et J. Meulman, Graph-theoretic representations for proximity matrices through strongly-anti-robinson or circular strongly-anti-robinson matrices, *Psychometrika*, 63 (1998), 341—358.

[7] E.V. Huntington, A set of independent postulates for cyclic order, *Proc. Natl. Acad. Sci. U.S.A.*, 2 (1916), 630–631.

[8] W.S. Robinson, A method for chronologically ordering archeological deposits, American Antiquity 16 (1951), 293–301.

# Une comparaison de quelques méthodes de classification de variables mixtes

Ndèye Niang [1], Mory Ouattara[2], Gilbert Saporta[1]

[1] Cédric-CNAM, Paris, France
ndeye.niang_keita@cnam.fr gilbert.saporta@cnam.fr
[2] Université de San Pedro, Côte d'Ivoire
ouattara.mory@usp.edu.ci

**Résumé**

*Nous proposons une nouvelle méthode de classification d'un ensemble de variables qualitatives et quantitatives que nous comparons à celles plus anciennes utilisant des mesures de corrélation entre les variables.*

**Mots-clés**

*Classification de variables, données mixtes, ACP, coefficient RV*

**Abstract**

*We compare several old and recent methods for clustering a set of qualitative and quantitative variables.*

**Keywords**

*clustering of variables, mixed data, PCA, RV coefficient*

## 1 Introduction

Des données à grande échelle et hétérogènes sont de plus en plus récoltées dans de nombreux domaines tels que l'environnement, le domaine médical et clinique, etc... Nous nous intéressons ici à l'hétérogénéité des variables. Le traitement simultané d'un mélange de variables quantitatives et qualitatives, que ce soit en analyse factorielle ou en classification, a fait l'objet d'un grand nombre de travaux Tenenhaus [4],Escofier [7], Saporta [5], Kiers [8] , Vigneau et Qanari [2], Pagès [9], Chavent[3].
Une question essentielle est d'utiliser des mesures de similarité cohérentes et comparables pour les différents couples des variables possibles. Plusieurs mesures ont été proposées. Le coefficient de corrélation linéaire ou les rapport de corrélation sont d'usage courant, tandis que diverses solutions ont été proposées pour le cas d'un couple de variables qualitatives.
Nous nous intéressons à la classification d'un ensemble de variables qualitatives et quantitatives. Nous proposons une nouvelle méthode que nous comparons celles plus anciennes utilisant des mesures de corrélation entre les variables.

## 2 Méthodes de classification

Le traitement simultané d'un mélange de $J$ variables quantitatives $\mathbf{x}_j$ et $Q$ qualitatives $\tilde{\mathbf{x}}_q$, que ce soit en analyse factorielle ou en classification repose souvent sur la détermination d'une ou plusieurs variables synthétiques globales ou locales (ie par classe) optimisant le critère suivant introduit par Tenenhaus [4], réutilisé par Escofier [7], puis Saporta [5], Kiers [8] sous le nom de PCAMIX, et Pagès [9].

$$\max_{\mathbf{c}} \left( \sum_{j=1}^{J} r^2 \left(\mathbf{c}, \mathbf{x}_j\right) + \sum_{q=1}^{Q} \eta^2 \left(\mathbf{c}, \tilde{\mathbf{x}}_q\right) \right) \qquad (1)$$

L'algorithme ClustOfVar [3] utilise ce critère pour effectuer une classification d'un ensemble de variables de nature différentes autour de composantes latentes par groupe, étendant la méthode de Vigneau et Qanari [2] introduite pour des variables exclusivement quantitatives.

Le regroupement de variables autour de composantes est une alternative intéressante aux algorithmes directs qui partent d'un tableau de similarités, de dissimilarités ou de distances entre toutes les variables car il optimise simultanément le regroupement et la représentation des classes par une composante, comme dans une approche clusterwise.

Il est fondamental d'utiliser des mesures de similarité cohérentes et comparables dans les trois cas : un couple de variables quantitatives, un couple de variables qualitatives et un couple formé d'une variable quantitative et d'une variable qualitative.
Les coefficients $r^2$ de corrélation linéaire et $\eta^2$ pour le rapport de corrélation sont d'usage courant, tandis que diverses solutions ont été proposées pour le cas d'un couple de variables qualitatives : le chi-deux et ses dérivés comme le carré du coefficient de Tschuprow $T^2$ [2] ou la plus grande valeur propre de l'AFC du tableau croisant deux variables qualitatives [3].
Les coefficients associés aux variables qualitatives ne sont cependant pas comparables entre eux ni avec un $r^2$ car leurs distributions dépendent de leurs nombres de moda-

lités. Dans (1) une variable qualitative joue un rôle d'autant plus grand que son nombre de modalités $m_q$ est élevé.
Les coefficients $RV$ d'Escoufier [10] entre tableaux engendrés par chaque variable quantitative et par les tableaux d'indicatrices des modalités des variables qualitatives permettent de définir des similarités euclidiennes égales selon les cas à $r^2$ , $\frac{\eta^2}{\sqrt{m_q-1}}$ ou $T^2$ (voir [2]).
On peut alors effectuer des classifications hiérarchiques avec l'algorithme de Ward ou des partitions avec les $k$-means, soit directement sur la matrice des similarités, soit sur les coordonnées obtenues par la formule de Torgerson.

Cette solution élégante mais un peu oubliée souffre quand même d'un défaut : diviser par la racine carrée du degré de liberté ne corrige pas complètement l'effet du nombre de modalités.
Pour cela, il peut être judicieux d'utiliser comme dissimilarité la $p$-value du test d'indépendance dans l'esprit de l'algorithme de la vraisemblance du lien [1]. Mais on perd les propriétés euclidiennes. De plus, lorsque le nombre d'observations est très grand, les $p$-value se rapprochent de zéro (*paradox of large samples*) et ne sont plus utilisables.

Nous proposons de les remplacer par les fractiles correspondants de la loi normale standard dans l'esprit des valeurs-test du logiciel SPAD [11].

## 3 Applications

Ces différentes approches seront comparées, en terme d'indice de qualité externes (indice de Rand) et de distance entre hiérarchies, sur des jeux de données réelles en particulier sur des données relatives à la pollution de l'air intérieur.

## Références


[1] Nicolau, F. Costa and Bacelar-Nicolau, H.,Some trends in the classification of variables. *Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan* Hayashi, C. and Yajima, K. and Bock, H.H. and Ohsumi, N. and Tanaka, Y. and Baba, Y., 1998.

[2] Qannari, E.M. and Vigneau, E. and Courcoux, Ph., Une nouvelle distance entre variables. Application en classification, *Revue de Statistique Appliquée*, vol. 46,pp. 21-32,1998.

[3] Chavent, M. and Kuentz-Simonet, V. and Liquet, B. and Saracco, J., ClustOfVar : An R Package for the Clustering of Variables, *Journal of Statistical Software*, vol. 50, number. 13, pp. 1–16, 2012

[4] Tenenhaus, M., Analyse en composantes principales d'un ensemble de variables nominales ou numériques, *Revue de Statistique Appliquée* ,vol. 25,pp. 39-56,1977

[5] Saporta, G., Simultaneous analysis of qualitative and quantitative data, *Atti della XXXV Riunione Scientifica, Societa Italiana di Statistica, Padova, Italy*,vol. 1, pp. 62-72, 1990

[6] Vigneau, E. and Qannari, E.M., Clustering of variables around latent components, *Communications in Statistics-Simulation and Computation* ,vol. 32,pp. 1131-1150,2003

[7] Escofier, B., Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des Données*, vol.4,pp.137-146,1979

[8] Kiers, H., Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, vol. 56, pp. 197-212, 1991

[9] Pagés, J., Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, vol.52,number. 4, pp. 93-111, 2004.

[10] Robert, P. and Escoufier, Y., A unifying tool for linear multivariate statistical methods : the RV-coefficient, *Journal of the Royal Statistical Society*, Series C : Applied Statistics, vol. 25(3), pp. 257–265, 1976.

[11] Morineau, A., SPAD.N logiciel pour l'analyse statistique des données, *Modulad-Le Monde des Utilisateurs de L'Analyse de Données*, vol. 6, pp. 27-60, 1991