Analyse exploratoire et classification de textes

Florian Barbaro

Université Paris 1 Panthéon Sorbonne
Laboratoire : Statistique, Analyse et Modélisation Multidisciplinaire, EA 4543
École Doctorale Sciences Mathématiques de Paris Centre (ED 386)

Analyse exploratoire et classification de textes

Florian Barbaro

Thèse de doctorat de l'Université Paris 1 Panthéon Sorbonne

Spécialité de doctorat : Mathématiques Appliquées

Date de soutenance : 2 juin 2022

Mathilde Mougeot	Professeure	Rapportrice
Julien Velcin	Professeur	Rapporteur
Eric Brousseau	Professeur	Examinateur
Pierre Latouche	Professeur	Examinateur
Joseph Rynkiewicz	Professeur	Examinateur
Fabrice Rossi	Professeur	Directeur de thèse

Labor omnia vincit improbus.

Remerciements

Je tiens ici à remercier le Professeur Fabrice ROSSI de m'avoir dirigé lors de la thèse, toutes les personnes du SAMM dont son directeur, le Professeur Jean-Marc BARDET, et la Chaire Gouvernance et Régulation de l'Université Paris Dauphine-PSL.

Je remercie le directeur de l'école doctorale, le Professeur Elisha FALBEL, de m'avoir écouté et accordé sa confiance.

Mes remerciements vont aussi au Professeur Gilles NOTTON pour m'avoir soutenu et conseillé tout au long de la thèse. Ton aide m'a été si précieuse.

Je souhaite remercier la personne qui m'a soutenu dans l'entreprise qui sponsorisait ma thèse à savoir Monsieur Benjamin COHEN. Ton aide est arrivée à point nommé.

Mes remerciements vont aux Professeurs Mary et Mathieu MATTEI pour leurs nombreux conseils.

Je souhaite maintenant remercier les personnes qui ont su m'accompagner lors de ce périlleux chemin, mes parents, Dominique et Jean-François BARBARO, ma mina, Rose TOMASI, ma marraine, Marie-Christine TOMASI, mes cousins, Julien POZZO DI BORGO et Eric BERAUD, ma tante, Anita BERAUD ainsi que toutes les autres personnes de ma famille. Même si tu n'as pu m'accompagner lors de cette épreuve, tu m'as appris à rêver, à croire que rien n'est impossible et atteindre tous mes objectifs même les plus fous, merci pépé, Dume TOMASI.

Cum'è dicemu in casa : VINCEREMU.

Enfin, une pensée à tous mes amis qui m'ont soutenu lors de ce périple.

Vi Ringraziu - Śukriyā - Дуже дякую - Большое спасибо - Çok teşekkürler - Muchas gracias - Nagyon szépen köszönjük - Muito obrigado - Thank you - Ευχαριστώ πάρα πολύ

Florian Dominique Barbaro Tomasi Pozzo Di Borgo, Paris, 2021

Résumé

Le traitement automatique des langues (NLP) a connu ces dernières années un grand engouement auprès de la communauté scientifique mais aussi des industriels pour les nombreuses opportunités offertes. En effet, nous sommes, de nos jours, submergés par les informations et par les différentes façons de les analyser. C'est dans cette voie que nous avons orienté notre thèse, à savoir comment rendre le résultat d'une classification facilement analysable et compréhensible.

Le **Chapitre 1** permettra une contextualisation de la thèse et de son intérêt tout en y présentant sa structure. De plus, un résumé synthétique des travaux de la thèse y est présenté.

Le **Chapitre 2** présentera l'état de l'art des modèles de représentation de textes utilisés, quelques méthodes de modélisations adaptées ainsi que des modèles pour les données directionnelles.

Puis, le **Chapitre 3** présentera les données qui nous ont accompagné tout au long de la thèse. Ce jeu de données, que nous avons constitué et qui est disponible librement, contient les rapports 8-K des entreprises du $S\&P500$ pour les années 2015 à 2019.

Dans le **Chapitre 4**, en nous inspirant de certains articles traitant des 8-K, nous essayerons de prédire le mouvement d'un actif financier selon la publication d'un rapport 8-K. Différentes techniques sont testées, tant pour la représentation de textes que pour la classification, et nous démontrerons que la complexification des modèles n'améliore que très légèrement les résultats de ladite classification.

Ensuite, dans le **Chapitre 5**, nous nous concentrerons sur la comparaison de représentations de textes en vue d'une analyse exploratoire à l'aide d'un algorithme de classification. Notre but est d'analyser plusieurs représentations de textes et de proposer de nouvelles manières de visualiser pour mieux appréhender la classification.

De même, dans le **Chapitre 6**, nous proposerons un modèle de mélange de distribution de von Mises-Fisher pénalisé par la norme l_1. Ceci permet d'améliorer l'interprétabilité des clusters obtenus grâce notamment à la parcimonie des moyennes directionnelles. Nous dériverons un algorithme EM pour ce modèle et nous illustrerons l'intérêt de notre approche sur plusieurs jeux de données réelles. De plus, nous proposerons une méthode de suivi de chemin qui permet en adéquation de critères de sélection de modèles, de sélectionner automatiquement le paramètre de pénalisation.

Keywords : NLP, sélection de modèles, analyse exploratoire de textes, finance, mélanges de lois de von Mises-Fisher, pénalisation l_1, données de grande dimension, classification, regroupement.

Abstract

In recent years, Natural Language Processing (NLP) has become very popular with the scientific community and also with industry because of the numerous opportunities it offers. Indeed, nowadays we are submerged by information and this raises the question of its analysis. This is the direction we have taken in our thesis, namely how to make the result of a classification easily analysable and understandable.

Chapter 1 provides a contextualisation of the thesis and its interest. In addition, we present its structure and a synthetic summary of the thesis' work.

Chapter 2 is here to recall the state of the art of text representation methods, some adapted modelling methods, as well as models for directional data.

Then comes **Chapter 3**, which presents the data that has followed us throughout the thesis. This dataset, which we built, is freely available, contains the 8-K reports of the $S\&P500$ companies for 2015 to 2019.

In **Chapter 4**, inspired by some papers that worked on 8-K, we try to predict the movement of a financial asset according to the publication of an 8-K report. Different techniques are tried, both for text representation and classification, and we show in our case that increasing the complexity of the models does not necessarily improve the results.

Chapter 5 follows on from the previous chapter, in which we focus on the comparison of text representations for exploratory analysis using a classification algorithm. Our aim here is to analyse several text representations and to propose new ways of visualisation for an easier understanding of classification.

To continue in this direction, in **Chapter 6**, we propose a mixture model of von Mises-Fisher distribution penalized by the l_1 norm. This leads to sparse prototypes that improve clustering interpretability. We derive an EM algorithm for this model and illustrate the interest of our approach on a real data set. Moreover, we propose a path-following method that allows, in adequacy with model selection criteria, to automatically select the penalization parameter.

Keywords : NLP, model selection, exploratory text analysis, finance, mixtures distribution of von Mises-Fisher, l_1 penalization, high dimensional data, classification, clustering.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte de la thèse

Depuis plus d'une décennie et l'explosion des moyens de communication, nous sommes submergés par les informations, certaines d'une importance capitale, d'autres beaucoup plus futiles. Se pose alors la question du traitement de cette information, en particulier textuelle. Pour cela, les chercheurs ont développé des techniques de traitement automatique des langues, en anglais *Natural language processing* (NLP), dans le but d'aider à la compréhension de l'information textuelle. Ce champ de recherche a connu ces dernières années de prodigieuses avancées dans de nombreuses applications, notamment la classification ou le regroupement de textes, la traduction automatique et bien d'autres.

Dans le cadre de notre thèse, nous nous sommes plus particulièrement intéressés à des textes impactant la vie économique et politico-économique des entreprises et des États. Dans un premier temps, nous nous sommes focalisés sur des rapports publiés par les entreprises américaines et transmis à la *Securities and Exchange Commission* (SEC) puis, dans un deuxième temps, aux consultations européennes.

Une entreprise vient de publier ses résultats trimestriels, une autre société a des nouvelles d'auditeurs qui pourraient déclencher un *red flag* [1], et une troisième société dépose le bilan. Les informations relatives à ces événements, et à bien d'autres, se trouvent dans un document appelé "rapport actuel sur formulaire 8-K". Le formulaire 8-K fournit aux investisseurs des informations actualisées qui leur permettent de prendre des décisions en connaissance de cause. Les types d'informations qui doivent être divulgués sur le formulaire 8-K sont généralement considérés comme "importants". Cela signifie que, en général, il y a une forte probabilité qu'un investisseur raisonnable considère l'information comme cruciale pour prendre une décision d'investissement. Les sociétés fournissent généralement un certain nombre de formulaires 8-K tout au long de l'année, chaque fois

1. Un *red flag* est un avertissement ou un indicateur, suggérant qu'il existe un problème ou une menace potentielle concernant les actions, les états financiers ou les rapports d'information d'une société. Il peut s'agir de toute caractéristique indésirable qui attire l'attention d'un analyste ou d'un investisseur.

que se produisent des événements importants pour la société qui déclenchent une divulgation. Les sociétés doivent déposer les 8-K rapidement, plutôt que d'attendre leur prochain rapport périodique, tel que le rapport trimestriel (sur le formulaire 10-Q) ou le rapport annuel (sur le formulaire 10-K). Les entreprises sont tenues de procéder à la plupart des divulgations 8-K dans les quatre jours ouvrables suivant l'événement déclencheur et, dans certains cas, même avant. Ces rapports sont disponibles sur le site Web EDGAR de la SEC [2].

L'un des principaux éléments du programme "Mieux légiférer", *Better Regulation Agenda*, consiste à rendre le processus d'élaboration des politiques de l'Union Européenne (UE) plus ouvert en consultant régulièrement les parties prenantes (entreprises, États, ONG [3], citoyens,...). L'objectif est de "consulter le plus tôt et le plus largement possible" afin d'inclure toutes les parties intéressées dans les nouvelles initiatives. Des consultations publiques sont organisées par la Commission européenne à plusieurs étapes du processus d'élaboration des politiques de l'UE. Ces consultations peuvent différer dans leur approche et leur format. En général, les consultations ouvertes peuvent être structurées de manière à recueillir des réactions générales ou spécifiques.

Lorsque la Commission européenne a l'intention de lancer une nouvelle action, par exemple l'évaluation d'un acte européen existant ou la préparation d'une proposition de nouvel acte, elle fait part de ses intentions au public en publiant une feuille de route. Les feuilles de route sont publiées sur le site web *Have your Say* et sont ouvertes à la consultation publique pendant une période de quatre semaines. Au cours de cette période de consultation, les parties prenantes et les citoyens peuvent donner leur avis très tôt dans le processus d'élaboration des politiques de l'UE.

La question posée est alors de définir des algorithmes qui permettent de comprendre et classer toutes ces informations.

C'est à ces questions qu'essaie de répondre notre travail de thèse.

Tout d'abord, nous devions collecter des données. Pour ce faire, nous avons compilé les rapports 8-K des entreprises présentes dans l'indice $S\&P$ 500 [4] durant les années 2015 à 2019. De plus, pour pouvoir travailler sur les consultations, nous avons agrégé différentes bases de données, à savoir les consultations provenant du site *Have your Say* mais aussi les discours des parlementaires à partir du site *LinkedEP* [5] et de la base de données *ParlGov project* [6] qui contient des informations sur les orientations politiques des différents partis de l'UE mais aussi de tous les pays de l'OCDE [7].

Puis, nous avons cherché à comprendre si la complexification des algorithmes permettait une compréhension plus aisée des textes à travers des regroupements mais aussi s'ils amélioraient la prédiction des effets engendrés par cette information. Nous avons

2. https://www.sec.gov/edgar/searchedgar/companysearch.html
3. Organisation non-gouvernementale
4. Le $S\&P$ 500 est un indice boursier basé sur 500 grandes sociétés cotées sur les bourses aux États-Unis.
5. https://linkedpolitics.project.cwi.nl/web/html/home.html
6. https://parlgov.org/
7. Organisation de coopération et de développement économiques.

analysé différentes représentations de textes de la plus simple, les sacs de mots, à celle plus complexe, les plongements de mots, mais aussi de nombreux algorithmes de prédiction, du *Random Forest* aux Réseaux de Neurones (RN), et de classification, du k-moyennes aux mélanges de distributions.

Enfin, nous avons exploré très largement les méthodes de visualisation des résultats, ce qui nous a mené à développer une technique de regroupement basée sur les mélanges de distribution de von Mises-Fisher. Cette méthode est adaptée à l'analyse exploratoire et permet d'améliorer l'interprétabilité des résultats par la sélection des variables importantes pour chaque classe. De plus, à la manière des techniques de *co-clustering*, nous réorganisons les moyennes de représentation de chaque classe par blocs permettant de mettre en évidence les variables communes, celles qui discriminent, et aussi celles qui sont inutiles pour définir le clustering.

1.2 Organisation de la thèse

Cette partie aborde les travaux effectués de manière synthétique et servent de plan pour la suite de la thèse.

État de l'art Ce chapitre a pour but de présenter tour à tour les modèles de représentation de textes utilisés, quelques méthodes de modélisations adaptées ainsi que des modèles pour les données directionnelles.

Constitution d'une base de données Nous étudierons dans cette partie, le jeu de données qui nous a servi tout au long de la thèse. Nous présentons en détails les rapports 8-K, l'indice $S\&P500$ ainsi que le processus d'obtention des données. Ce jeu de données s'intéresse aux entreprises du $S\&P500$ pour les années 2015 à 2019. Puis, nous effectuerons une analyse exploratoire qui se concentrera notamment sur la distribution des événements de ces rapports ainsi que sur la saisonnalité des publications. Enfin, nous réaliserons une analyse détaillée sur la période 2015-2016 pour l'entreprise Wells Fargo, de symbole WFC, qui nous permet d'avoir une meilleure vision de la difficulté d'analyse de ces rapports.

Prédiction des marchés financiers Suite à notre analyse exploratoire de la base de données, nous nous intéresserons dans ce chapitre à la possibilité de prédire la réaction du marché sur le cours d'un actif suite à la publication d'un rapport 8-K comme dans [72]. Pour ce faire, nous explorerons plusieurs stratégies de représentations de textes ainsi que différents algorithmes d'apprentissage automatique. Nous montrerons que la complexification des représentations et des algorithmes de classification n'améliore que très légèrement les résultats.

Analyse exploratoire de textes Nous étudierons dans ce chapitre l'impact de différentes approches de représentations vectorielles de textes sur l'analyse exploratoire d'un corpus. Nous comparerons une représentation élémentaire par sac de mots (unigrammes) à celle obtenue par *topic model* ainsi qu'à une plus complexe, construite à partir de la distance Sinkhorn entre les textes calculée sur une représentation vectorielle des mots. Nous construirons une classification des textes ainsi représentés à l'aide du modèle *high-dimensional data clustering*. Nous illustrerons les différences entre les représentations grâce à un corpus de textes constitués à partir des rapports 8-K des entreprises du Standard & Poor's 500 (pour les années 2015 et 2016). Nous analyserons la cohérence des classes ainsi obtenues et chercherons à les caractériser en termes de vocabulaire et de sujets spécifiques.

Mélange de lois de von Mises-Fisher Les mélanges de distributions de von Mises-Fisher peuvent être utilisés pour regrouper des données sur l'hypersphère unitaire. Ceci est particulièrement adapté aux données directionnelles de haute dimension telles que les textes. Nous estimerons dans cette partie un mélange de von Mises en utilisant une vraisemblance pénalisée par la norme l_1. Cela conduira à des prototypes parcimonieux qui améliorent l'interprétabilité des résultats. Nous introduirons un algorithme EM pour cette estimation et montrerons les avantages de l'approche sur des données réelles de référence. Nous proposerons d'explorer le compromis entre le terme de parcimonie et celui de vraisemblance avec un algorithme simple de suivi de chemin.

Chapitre 2

État de l'art

Résumé : Ce chapitre a pour but de présenter tour à tour les modèles de représentation de textes utilisés, quelques méthodes de modélisations adaptées ainsi que des modèles pour les données directionnelles.

2.1 Représentation de texte

Le traitement du langage naturel (NLP) est un domaine de recherche et d'application qui explore comment les ordinateurs peuvent être utilisés pour comprendre et manipuler des textes ou des discours en langage naturel afin d'en extraire l'essence. Les chercheurs en NLP ont pour objectif de rassembler des connaissances sur la façon dont les êtres humains comprennent et utilisent le langage afin de développer des outils et des techniques appropriées pour que ces systèmes puissent comprendre et manipuler les langues naturelles pour effectuer les tâches souhaitées. Les fondements de la NLP reposent sur de nombreuses disciplines, à savoir l'informatique et les sciences de l'information, la linguistique, les mathématiques, l'intelligence artificielle, la robotique, la psychologie, ... Les applications NLP comprennent un certain nombre de domaines d'études [86, 109], tels que la traduction automatique, le traitement de texte en langage naturel, les interfaces utilisateurs, la recherche d'information multilingue, la reconnaissance vocale, l'intelligence artificielle et les systèmes experts, ainsi de suite.

Dans cette partie, nous nous concentrerons sur différentes représentations de textes étudiées dans notre travail de thèse ainsi que les différents modèles adaptés.

Notations Un corpus C est un ensemble de N textes notés $(t_i)_{1\leq i\leq N}$. Un texte est une séquence de M unités sémantiques - paragraphes, phrases, mots - notées $(w_i)_{1\leq i\leq M}$ qui forment le vocabulaire V. Les matrices sont indiquées en gras et majuscules, les vecteurs en gras et en minuscules. La norme l_1 est notée par $\|.\|_1$ et la norme l_2 par $\|.\|_2$. La sphère unité de dimension $(d-1)$ intégrée dans $\mathbb{R}^d$ est noté $\mathbb{S}^{d-1}$. Les données sont représentées par une matrice $\boldsymbol{X} = (x_{ij})$ de dimension $n \times d$ avec $x_{ij} \in \mathbb{R}$ et la i^{eme} ligne de cette

matrice est représentée par un vecteur $\boldsymbol{x_i} = (x_{i1}, ..., x_{id})^T$, où T dénote la transposée. La partition de l'ensemble des lignes I en K classes peut être représentée par une matrice de classification $\boldsymbol{Z}$ d'éléments z_{ih} dans $\{0, 1\}$ satisfaisant $\sum_{h=1}^{K} z_{ih} = 1$. Nous notons $\mathbb{I}$ la fonction caractéristique.

2.1.1 Modèles vectoriels

Les modèles vectoriels ont été parmi les premières méthodes développées pour transformer une donnée non-structurée, un texte, en une donnée structurée, un vecteur. Nous considérons ici qu'un corpus peut être représenté par un dictionnaire et qu'un texte en son sein est donc représenté par un vecteur de la même taille que ce dictionnaire.

2.1.1.1 Sac de mots

La représentation par sac de mots, proposée dans [53], est une méthode élémentaire encore très utilisée dans l'analyse de texte [72]. Elle consiste à représenter un texte par un vecteur indiquant les occurrences de chaque terme dans le texte. Pour ce faire, nous décidons d'un ensemble de M mots, qui forment le vocabulaire, et nous attribuons à chaque mot du vocabulaire un index unique. Ensuite, chaque document est représenté par un vecteur de longueur M, dans lequel la i-ème entrée contient le nombre d'occurrences du mot i dans le document. En prenant le corpus contenant les deux documents (phrases) suivants *The cat is sitting on the table* et *The car is red*, la représentation par sac de mots serait la suivante dans la table 2.1.

Table 2.1 – Exemple sac de mots

Doc/Termes	the	cat	is	sitting	on	table	car	red
1	2	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	1

Un corpus de documents est donc défini par la matrice $\boldsymbol{C} \in \mathbb{N}^{N \times M}$. Nous avons $c_{ij} = o_{ij}$ où o_{ij} représente l'occurrence du terme j dans le document i.

Le sac de mots a été utilisé presque exclusivement et avec grand succès pendant des décennies sur diverses tâches de NLP. Et ceci, en dépit de la perte de toutes les informations autres que l'occurrence (par exemple la place des mots) et de la taille grandissante de la représentation, si le vocabulaire est riche. Même avec les progrès significatifs de la représentation vectorielle pour le texte au cours des dernières années, de légères variations de cette méthode sont encore utilisées aujourd'hui avec succès.

2.1.1.2 Pondération *Term Frequency-Inverse Document Frequency*

Une autre façon classique de traiter une tâche NLP est de construire un modèle de sac de mots avec une pondération donnée par la méthode *Term Frequency-Inverse Document Frequency* (TF-IDF). Supposons qu'il y ait N documents dans le corpus, et que le terme w_j apparaisse dans n_j de ces documents. Alors, la fréquence inverse des documents relative au terme w_j peut être calculée comme suit :

$$idf_j = \log \frac{N}{n_j}. \tag{2.1}$$

En fait, la mesure originale était une approximation entière de cette formule, et le logarithme était en base 2. Cependant, (2.1) est la forme la plus couramment citée de l'IDF. Pour plus d'informations, le lecteur peut se tourner vers l'article original [95].

On désigne par tf_{ij} la fréquence du terme j dans le document i considéré [95]. tf_{ij} s'exprime comme suit :

$$tf_{ij} = \frac{o_{ij}}{\sum_{j=1}^{M} o_{ij}}. \tag{2.2}$$

Enfin, le TF-IDF est défini pour un terme t_j et un document i donnés comme suit :

$$tfidf_{ij} = tf_{ij} \times idf_j. \tag{2.3}$$

Ainsi, un corpus de documents est donc défini par la matrice $\boldsymbol{C} \in \mathbb{R}^{N \times M}$. Nous avons $d_{ij} = tfidf_{ij}$ où $tfidf_{ij}$ représente le TF-IDF du terme j dans le document i.

2.1.1.3 N-grammes

Pour récupérer une partie de l'information sur l'ordre des mots perdue par l'approche du sac de mots (unigramme), la fréquence des courtes séquences de mots (de longueur deux, trois, etc.) peut être utilisée (en complément ou en remplacement) pour construire des vecteurs de mots. Naturellement, les unigrammes sont un cas particulier de cette méthode.

Pour la phrase *The car is red*, les paires de mots sont *the car*, *car is* et *is red*. Le vocabulaire est constitué (ou enrichi) de toutes les paires de mots successives du corpus d'entrée.

En conséquence, un corpus de documents est défini par la matrice $\boldsymbol{C} \in \mathbb{N}^{N \times M}$. Nous avons $c_{ij} = o_{ij}$ où o_{ij} représente l'occurrence du N-gramme j dans le document i.

L'un des principaux inconvénients de cette approche est la dépendance non linéaire de la taille du vocabulaire par rapport au nombre de mots uniques, qui peut être très

importante pour les grands corpus. Les techniques de filtrage sont couramment utilisées pour réduire la taille du vocabulaire (cf. **section 4.5.1.1**, p. 47).

2.1.2 Plongement de mots

Le plongement de mots ou *Word Embedding*, est une méthode qui consiste à représenter chaque mot sous forme de vecteurs denses.

Il s'agit d'une amélioration par rapport aux schémas d'encodage plus traditionnels que sont les modèles vectoriels, où de grands vecteurs épars sont utilisés pour noter chaque mot dans un vecteur afin de représenter un vocabulaire entier. Ces représentations sont éparses car le vocabulaire est vaste et un document donné est représenté par un grand vecteur composé principalement de valeurs nulles.

Au contraire, dans un plongement, les mots sont représentés par des vecteurs denses où un vecteur représente la projection du mot dans un espace vectoriel continu. La position d'un mot dans l'espace vectoriel est apprise à partir du texte et est basée sur les mots qui entourent le mot lorsqu'il est utilisé. Ces méthodes ont connu de grandes avancées depuis les articles présentant les méthodes word2vec [82] et GloVe [89].

2.1.2.1 Word2vec

Word2vec, est le nom sous lequel sont connus deux modèles de langage basés sur des réseaux neuronaux et proposés par Mikolov et al. [82] pour générer des représentations vectorielles denses des mots. Plus précisément, deux architectures de modèles sont proposées : Le modèle Continuous Bag of Words (CBOW), et le modèle Continuous Skip Gram (SG). D'une part, le modèle CBOW vise à prédire l'occurrence d'un mot en fonction des autres mots qui constituent son contexte. Le contexte d'un mot w_i est compris comme le voisinage composé par les k mots à la gauche de w_i , et les k mots à la droite de w_i . D'autre part, le SG tente de prédire un contexte donné par le mot w_i . La taille k de la fenêtre de contexte local est un hyperparamètre du modèle. Dans les deux cas, les modèles fournissent des représentations vectorielles denses pour les mots, qui se sont avérées efficaces pour préserver les caractéristiques sémantiques de ceux-ci, même en réduisant la taille de la fenêtre de contexte. De plus, l'apprentissage est effectué à l'aide d'un réseau de neurones peu profond composé de trois couches (une d'entrée, une cachée, une de sortie), ce qui accélère également le processus .

2.1.2.2 GloVe

Le modèle *Global Vectors for Word Representation* (GloVe) a été proposé par Pennington *et al.* [89]. Contrairement à word2vec, qui est un modèle prédictif, GloVe est plus proche d'un modèle qui réduit la dimensionnalité d'une matrice de co-occurrence de type mot-mot, générée par une fenêtre de contexte local de dimension fixe. GloVe tire son nom du fait que les statistiques du corpus entier (à un niveau global) sont capturées

directement par le modèle. En outre, il s'est avéré performant et a donné de meilleurs résultats que d'autres méthodes de pointe au moment de sa sortie, comme word2vec, dans des tâches telles que l'analogie et la similarité des mots, ainsi que la reconnaissance des entités nommées.

2.1.2.3 Défaut de ces modèles

Le défaut principal de ces méthodes est de ne pas prendre en compte le contexte des mots dans leur représentation. Pour pallier cela, de nombreuses méthodes sont apparues comme FastText[62], ELMO[90], Infersent [35] et surtout BERT [40]. Ce dernier, et notamment ses dérivées comme RoBERTa [75], FlauBERT [71] pour sa version en français ou bien FinBERT [4] spécialement entraîné pour la finance, constitue l'état de l'art pour de nombreux sujets en NLP.

2.1.2.4 BERT

Le modèle BERT comporte deux étapes : le pré-entraînement (*pre-training*) et le réglage fin (*fine-tuning*) [40]. Pendant le pré-entraînement, le modèle est formé sur un grand corpus non étiqueté. Pour le réglage fin, le modèle est initialisé avec les paramètres pré-entraînés, et tous les paramètres sont affinés en utilisant des données étiquetées pour des tâches spécifiques. L'architecture du modèle de BERT est un Transformer multi-couche encoder [40] basé sur l'implémentation originale décrite dans [112].

Ce type d'encodeur est composé d'une pile de $N = 6$ couches identiques. Chacune de ces couches possède deux sous-couches. La première sous-couche est un mécanisme d'auto-attention multi-têtes alors que la seconde est un simple réseau *feedforward* entièrement connecté. Il utilise une connexion résiduelle [54] autour des deux sous-couches, suivie d'une normalisation des couches [7]. Il faut donc comprendre que la sortie de chaque sous-couche est $LayerNorm(x + Sublayer(x))$, où $Sublayer(x)$ est la fonction implémentée par la sous-couche [112].

En ce qui concerne l'auto-attention à têtes multiples, nous devons d'abord définir l'attention par produit scalaire. Elle se définit comme suit :

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}, \tag{2.4}$$

Où $\boldsymbol{Q}$ est la matrice des requêtes, $\boldsymbol{K}$ la matrice des clés, $\boldsymbol{V}$ celle des valeurs et d_k la dimension des matrices $\boldsymbol{Q}$ et $\boldsymbol{K}$. Maintenant, l'attention multi-têtes peut être défini comme suit :

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, ..., head_h)\boldsymbol{W}^O, \tag{2.5}$$

ou $head_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}})$ et les $\boldsymbol{W}^O$, $\boldsymbol{W}^{\boldsymbol{Q}}$, $\boldsymbol{W}^{\boldsymbol{K}}$ et $\boldsymbol{W}^{\boldsymbol{V}}$

représentent les paramètres du modèle.

L'attention multi-têtes consiste à projeter les requêtes, les clés et les valeurs h fois avec différentes projections linéaires apprises sur d_k, d_k et d_v (dimension de la matrice des valeurs), respectivement. Ensuite, sur chacune de ces versions projetées des requêtes, clés et valeurs, la fonction d'attention est exécutée en parallèle, produisant des valeurs de sortie à d_v dimensions. Enfin, celles-ci sont concaténées et projetées, ce qui donne les valeurs finales [112]. L'auto-attention signifie que toutes les clés, valeurs et requêtes proviennent du même endroit.

BERT représente une phrase unique ou une paire de phrases (par exemple, la paire ⟨question, réponse⟩) comme une séquence de *tokens*, représentation des termes, selon les caractéristiques suivantes. Le premier jeton de la séquence est "[CLS]". Lorsqu'il y a une paire de phrases dans la séquence, elles sont séparées par le jeton "[SEP]". Et, un encastrement est ajouté à chaque *token* indiquant s'il appartient à la première ou à la deuxième phrase. Pour un *token* donné, sa représentation d'entrée est construite en additionnant le *token*, la position, et le segment correspondant [40]. Notons que BERT utilise des encastrements WordPiece [114].

2.1.3 *Bags of vectors*

Une représentation populaire des textes consiste à s'appuyer sur un plongement vectoriel des mots eux mêmes, dans la lignée de *word2vec* [82] ou bien de *GloVe* [89] mais aussi sur les différents *topics* présents dans le texte.

Ainsi chaque texte est représenté par un sac de vecteurs, *Bags of vectors* en anglais, l'ensemble des vecteurs associés aux mots ou *topics* du texte.

Formellement, un texte est représenté par $\boldsymbol{T} \in \mathbb{R}^{M \times D}$. M indique le nombre d'unités sémantiques qui constituent un texte. D désigne la dimension de l'espace qui représente ces éléments. Par exemple, si nous travaillons avec des *topics*, $D = 1$. Si maintenant, nous exploitons les plongements de mots, D est égal à la dimension de ce plongement.

Cependant, il est difficile de comparer directement de tels ensembles et de nombreuses solutions ont été développées.

Une des premières manières les plus simples, notamment utilisée dans le cas des *topics*, est de représenter un texte sous forme d'histogramme. Dans ce cas, chaque index correspond à un *topic* et le nombre d'occurrence des mots du topic dans le texte est répertorié dans l'index en question. Cela permet de représenter les différents textes d'un corpus avec des vecteurs de même longueur. Nous présentons en détails cette méthode dans la **section 4.4.2** (p. 45).

Une deuxième méthode est appelé l'apprentissage multi-instance [42] qui est utilisée pour la classification de textes [117]. C'est un type d'algorithme d'apprentissage faiblement supervisé dans lequel les données d'apprentissage sont organisées en sacs (les textes), chaque sac contenant un ensemble d'instances $\boldsymbol{T} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$ (les mots ou leurs représentations vectorielles). Il existe une seule étiquette Y par sac, $Y \in \{0, 1\}$

dans le cas d'un problème de classification binaire. Nous supposons que les étiquettes individuelles $y_1, y_2, \ldots, y_M$ existent pour les instances d'un sac, mais qu'elles sont inconnues pendant l'apprentissage. Dans l'hypothèse standard, un sac est considéré comme négatif si toutes ses instances sont négatives. D'autre part, un sac est positif si au moins une instance du sac est positive.

En outre, pour représenter un texte à partir des méthodes de *Word Embedding* notamment, nous pouvons utiliser la moyenne [37], la moyenne pondérée selon l'information mutuelle, la moyenne pondérée selon la probabilité d'apparition du mot [5], le Min Max, la concaténation du Min, du Max et de la moyenne [96] ou bien une représentation Gaussienne Multivariée des textes [85]. Ces méthodes sont détaillées dans la **section 4.4.1** (p. 45). L'article [5] montre d'ailleurs que ces techniques peuvent obtenir de meilleurs résultats que des réseaux de neurones récurrents RNN [97] ou bien les Long Short Term Memory (LSTM) [55]

2.1.4 Topic models

Dans les tâches qui incombent à la NLP, il existe une série de loupes à travers lesquelles nous pouvons extraire du sens, des mots aux phrases, en passant par les paragraphes et les documents. Au niveau du document, l'une des façons les plus utiles de comprendre un texte est d'analyser ses *topics*. Le processus d'apprentissage, de reconnaissance et d'extraction de ces *topics* dans une collection de documents est appelé *topics modeling*.

Les *topic models* sont construits autour de l'idée que la sémantique de notre document est en fait régie par des variables cachées, ou *latentes*, que nous n'observons pas. Par conséquent, le but de la modélisation thématique est de découvrir ces variables latentes, en d'autres termes *topics*, qui façonnent la signification du document et du corpus.

Deux des méthodes les plus utilisées sont détaillées ci-dessous, à savoir l'Analyse Sémantique Latente et l'Allocation de Dirichlet.

2.1.4.1 Analyse sémantique latente

L'Analyse sémantique latente [70], ou *Latent semantic analysis* en anglais, est une méthode statistique qui met en évidence les relations entre les mots dans les textes. Il s'agit d'une méthode basée sur un corpus qui n'utilise pas de dictionnaires, de réseaux sémantiques, de grammaires, d'analyseurs syntaxiques ou morphologiques, et dont l'entrée est représentée uniquement par du texte brut divisé en unité sémantique.

Formellement, $\boldsymbol{A}$ est la matrice terme-document $M \times N$ d'un corpus. Si le terme i apparaît a fois dans le document j, alors $a_{i,j} = a$.

Observons que $\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}$ est la matrice document-document. Si les documents i et j ont b mots en commun, alors $b_{i,j} = b$. D'autre part, $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^T$ est la matrice terme-terme. Si les termes i et j apparaissent ensemble dans c documents, alors $c_{i,j} = c$. Il est

clair que $\boldsymbol{B}$ et $\boldsymbol{C}$ sont toutes deux carrées et symétriques : $\boldsymbol{B}$ est une matrice $N \times N$, tandis que $\boldsymbol{C}$ est une matrice $M \times M$.

Nous effectuons une décomposition en valeurs singulières sur $\boldsymbol{A}$ en utilisant les matrices $\boldsymbol{B}$ et $\boldsymbol{C}$:

$$\boldsymbol{A} = \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{U}^T, \tag{2.6}$$

où $\boldsymbol{S}$ est la matrice des vecteurs propres de $\boldsymbol{C}$, $\boldsymbol{U}$ étant la matrice des vecteurs propres de $\boldsymbol{B}$, et $\boldsymbol{\Sigma}$, la matrice diagonale des valeurs singulières obtenues comme racines carrées des valeurs propres de $\boldsymbol{B}$.

Certaines des valeurs singulières sont négligeables. Ainsi, de manière empirique, nous choisissons dans ce modèle d'ignorer ces petites valeurs singulières et les remplaçons par 0. Nous ne gardons alors que k valeurs singulières dans $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ est nulle à l'exception des k premières valeurs sur sa diagonale. Ainsi, nous pouvons réduire la matrice $\boldsymbol{\Sigma}$ en $\boldsymbol{\Sigma}_k$ qui est une matrice $k \times k$ contenant uniquement les k valeurs singulières que nous conservons. Nous réduisons également $\boldsymbol{S}$ et $\boldsymbol{U}^T$, en $\boldsymbol{S}_k$ et $\boldsymbol{U}_k^T$, pour avoir respectivement k colonnes et k lignes.

La matrice $\boldsymbol{A}$ est maintenant approximée par

$$\boldsymbol{A}_k = \boldsymbol{S}_k\boldsymbol{\Sigma}_k\boldsymbol{U}_k^T. \tag{2.7}$$

Ainsi, les k composantes restantes des vecteurs propres dans $\boldsymbol{S}$ et $\boldsymbol{U}$ correspondent aux k concepts latents auxquels participent les termes et les documents. Les termes et les documents ont maintenant une nouvelle représentation dans cet espace latent.

À savoir, les termes sont représentés par les vecteurs de ligne de la matrice $M \times k$:

$$\boldsymbol{S}_k\boldsymbol{\Sigma}_k, \tag{2.8}$$

tandis que les documents par les vecteurs colonnes de la matrice $k \times N$:

$$\boldsymbol{\Sigma}_k\boldsymbol{U}_k^T. \tag{2.9}$$

2.1.4.2 Allocation de Dirichlet

L'Allocation de Dirichlet latente [17], ou *Latent Dirichlet Allocation* (LDA) en anglais, est un modèle probabiliste génératif conçu pour extraire des sujets de textes.

L'idée de base de cette méthode est que les documents sont représentés comme des mélanges aléatoires de k sujets latents. Chaque sujet est représenté comme une distribution multinomiale sur les M mots du vocabulaire. Un document est généré en échantillonnant un mélange de ces sujets, puis en échantillonnant des mots de ce mélange.

Plus précisément, un document de M termes $\boldsymbol{w} = (w_1, \ldots, w_M)$ est généré par un processus génératif. Tout d'abord, $\boldsymbol{\theta}$ est échantillonné grâce à une distribution de $Dirichlet(\alpha_1, \ldots, \alpha_k)$. Cela signifie que $\boldsymbol{\theta}$ se trouve dans le simplexe de dimension $(k-1) : \theta_i \geq 0, \sum_i \theta_i = 1$. Puis, pour chacun des M mots, un *topic* $z_n \in \{1, \ldots, k\}$ est échantillonné à partir d'une distribution multinomiale de paramètres $\boldsymbol{\theta}$, $p(z_n = i|\boldsymbol{\theta}) = \theta_i$. Enfin, chaque mot w_m est échantillonné, conditionné par le z_n-ième sujet, à partir de la distribution multinomiale $p(w_m|z_n)$. Ainsi, θ_i peut être compris comme la proportion du *topic* i dans le document.

La probabilité d'un document est donc le mélange suivant :

$$p(\boldsymbol{w}) = \int_{\boldsymbol{\theta}} \left(\prod_{m=1}^{M} \sum_{z_n=1}^{k} p(w_m|z_n; \boldsymbol{\beta}) p(z_n|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}; \boldsymbol{\alpha}) d\boldsymbol{\theta}, \tag{2.10}$$

où $p(\boldsymbol{\theta}; \boldsymbol{\alpha})$ est une distribution de Dirichlet, $p(z_n|\boldsymbol{\theta})$ est une distribution multinomiale de paramètres $\boldsymbol{\theta}$, $p(w_m|z_n; \boldsymbol{\beta})$ est une distribution multinomiale sur les mots. Ce modèle est paramétré par le paramètre de Dirichlet $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ et une matrice $\boldsymbol{\beta}$ de dimension $k \times M$, qui contrôlent les k distributions multinomiales sur les mots.

Plusieurs procédures d'inférence ont été proposées pour ce modèle. Par exemple, une approche *Variational Expectation-Maximization* (VEM) a été adoptée dans l'article original de Blei *et al.* [17] ou bien l'algorithme d'échantillonnage de Gibbs [46, 16] que nous utiliserons dans cette thèse.

Enfin, de nombreux critères permettent de sélectionner le nombre optimal de sujets pour un corpus donné [6, 29, 49, 39].

2.2 Modèles

2.2.1 Modèles classiques

Dans le cas des textes considérés comme des ensembles de données de grande dimension, le choix d'un modèle adapté dépend de la problématique posée. Le plus souvent, nous cherchons soit à classifier les textes, soit à les regrouper pour permettre une analyse exploratoire des résultats en utilisant des méthodes de visualisation.

Pour le premier objectif, à savoir la classification, les modèles de type Random Forest [72] ou de *boosting* [91] se sont révélés bien adaptés aux données de grande dimension. Cependant, comme il s'agit de textes, les notions de séquence et de contexte y sont très importantes. Pour tenter de les prendre en compte, une architecture de réseau neuronal récurrent (RNN) [56] est particulièrement intéressante, à savoir les *Long short-term memory* (LSTM) [55]. Ils possèdent des liens de rétroaction, contrairement aux réseaux neuronaux à action directe normaux, et sont appliqués avec succès à cette problématique [121]. Plus récemment, l'introduction du modèle *BERT* a bouleversé le monde de la NLP. Une fois entraîné dans un contexte général, il est possible d'affiner les para-

mètres du modèle de manière supervisée pour obtenir des résultats qui nécessiteraient un long apprentissage spécifique sans cela. RoBERTa [75], FlauBERT [71] pour sa version en français ou bien FinBERT [4] spécialement entraîné pour la finance, en sont trois exemples concrets.

Pour la deuxième tâche, à savoir le regroupement, c'est un problème tout autant difficile. La difficulté est due au fait que les données de haute dimension vivent généralement dans différentes sous-dimensions, de plus faible dimension, cachées dans la structure d'origine. Pour répondre à ce problème, une des méthodes est d'utiliser des modèles adaptés à partir des modèles de mélange gaussien. Ces derniers modèles n'étant pas adaptés à ce type de données, de nombreuses variantes sont apparues [23, 87, 119] et notamment l'algorithme *High-dimensional data clustering (HDDC)* [23].

HDDC est un modèle de mélange gaussien avec comme spécificité principale la sélection automatique d'une projection spécifique en basse dimension pour chaque classe. En ce sens, HDDC peut être vu comme une généralisation du principe du mélange d'analyse en composantes principales [108].

Comme dans le cadre du modèle de mélange gaussien classique [79], nous supposons que les densités conditionnelles de classe sont gaussiennes $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, ayant pour moyennes $\boldsymbol{\mu}_k$ et comme matrices de covariance $\boldsymbol{\Sigma}_k$, pour $k = 1,\ldots,K$. K étant le nombre de classes. Dans HDDC, nous contraignons le spectre de $\boldsymbol{\Sigma}_k$ à s'écrire $(a_{k1}, \ldots, a_{kd_k}, b_k, \ldots, b_k)$. Les d_k premiers termes sont libres et correspondent à la covariance conservée pour la classe k (qui a donc une dimension intrinsèque de d_k). Le reste du spectre caractérisé par b_k est une composante de bruit isotrope (sur les $p - d_k$ dimensions restantes).

HDDC propose plusieurs variantes régularisées du modèle le plus général qui imposent un partage de paramètres entre les classes (ou au sein d'une même classe). Par exemple, le bruit b_k est spécifique ou non à chaque classe. La sélection automatique, selon le critère BIC, opère non seulement sur le nombre de classes K, mais aussi sur les variantes : HDDC sélectionne le meilleur compromis entre la complexité du modèle et l'adéquation aux données.

2.2.2 Modèles pour données directionnelles

Un cas particulier intéressant des données de grande dimension, consiste en ce que l'on appelle les données directionnelles [78]. Pour ce cas, la corrélation entre deux vecteurs est plus informative que la norme de leur différence (c'est-à-dire la distance euclidienne). Ce type de données apparaît naturellement dans la représentation vectorielle de texte, ainsi que dans de nombreux autres domaines comme l'analyse des gènes. Pour ces données, de nombreux modèles, et notamment les modèles de type gaussien, sont doublement inadaptés : ils souffrent des effets négatifs de la grande dimension des données et sont basés sur une métrique sous-jacente non adaptée.

Une façon naturelle de traiter ce type de données est de procéder à une normalisation qui les place sur la sphère unitaire. Ensuite, des techniques de *clustering* qui tiennent compte de cette spécificité peuvent être utilisées.

Une première technique qui peut être utilisée est celle des k-means sphériques [41]. Elle diffère de la méthode k-means conventionnelle dans la projection des centroïdes estimés des clusters sur la sphère unitaire, à la fin de chaque étape de maximisation. L'article [64] montre la pertinence de l'utilisation de cette méthodologie sur les textes.

Une deuxième technique est d'utiliser une distribution sphérique comme les distributions de Fisher-Bingham introduite dans [77] ou bien de Kent [63] mais surtout la distribution de von Mises-Fisher qui se révèle très bien adaptée dans le cas des textes et leurs classifications (non supervisées), cf [9, 48]. Nous explorerons plus en détails cette distribution dans le **chapitre 6** (p. 78).

2.2.3 Réseaux de neurones

L'architecture de base de l'apprentissage profond, également connue sous le nom de réseau neuronal artificiel à action directe, est appelée perceptron multicouche (MLP) [88]. Un MLP typique est un réseau entièrement connecté composé d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie. Chaque noeud d'une couche est connecté à chaque noeud de la couche suivante avec une pondération. Il utilise la technique de *Back propagation* [51]. Il s'agit du bloc de construction le plus fondamental d'un réseau neuronal, pour ajuster les valeurs de poids en interne tout en construisant le modèle. Le MLP typique est sensible aux caractéristiques d'échelle et permet de régler une variété d'hyperparamètres, tels que le nombre de couches cachées, de neurones et d'itérations, ce qui peut entraîner un modèle coûteux en termes de calcul.

Un autre type de réseau de neurones adapté aux textes est le *Long short-term memory* (LSTM) [55]. Le LSTM est une architecture de réseau neuronal récurrent (RNN) [56] utilisée dans le domaine de l'apprentissage profond [47]. Le LSTM possède des liens de rétroaction, contrairement aux réseaux neuronaux à action directe. Les réseaux LSTM sont bien adaptés à l'analyse et à l'apprentissage de données séquentielles, comme la classification, le traitement et la prédiction de données basées sur des séries chronologiques, ce qui les différencie des autres réseaux conventionnels. Ainsi, les réseaux LSTM peuvent être utilisés lorsque les données sont dans un format séquentiel, comme le temps, la phrase, etc., et sont couramment appliqués dans le domaine de l'analyse des séries temporelles, du traitement du langage naturel, de la reconnaissance vocale ...

2.2.4 Transport optimal

Les solutions simples, comme la représentation par la moyenne des vecteurs, induisent une perte considérable d'information [5, 37]. Pour pallier ce problème, il est possible de s'appuyer sur le transport optimal [113], qui permet de définir une distance entre des lois de probabilité, distance qui tient compte de l'espace ambiant. Chaque texte est vu comme une distribution discrète dans l'espace vectoriel de plongement des mots : nous comparons deux textes au sens de la distance du transport optimal entre leur distribution [116, 107]. Le coût du calcul du transport optimal étant élevé, au moins $O(m^3 log(m))$

lors du calcul de la distance entre une paire de distributions de dimension m, il est possible d'utiliser une version régularisée du problème, proposée dans [36] avec un coût de $O(m^2)$. La distance correspondante est dite distance Sinkhorn.

Formellement, le polytope du transport optimal peut être interprété en tant qu'un ensemble de probabilités jointes. Dans ce qui suit, $\langle \cdot, \cdot \rangle$ est le produit scalaire de Frobenius. Pour deux vecteurs de probabilité $\boldsymbol{r}$ et $\boldsymbol{c}$ dans le simplexe : $\boldsymbol{\Sigma}_d := \{\boldsymbol{x} \in \mathbb{R}_+^d : \boldsymbol{x}^T \mathbf{1}_d = 1\}$, où $\mathbf{1}_d$ est un vecteur de uns de dimension d, nous notons $U(\boldsymbol{r}, \boldsymbol{c})$ le polytope du transport de $\boldsymbol{r}$ et $\boldsymbol{c}$, à savoir l'ensemble polyédrique de matrices $d \times d$,

$$U(\boldsymbol{r}, c) := \{\boldsymbol{P} \in \mathbb{R}_+^{d \times d} | \boldsymbol{P}\mathbf{1}_d = \boldsymbol{r}, \boldsymbol{P}^T \mathbf{1}_d = \boldsymbol{c}\}. \tag{2.11}$$

$U(\boldsymbol{r}, \boldsymbol{c})$ contient toutes les matrices $d \times d$ non négatives dont les sommes des lignes et des colonnes sont respectivement $\boldsymbol{r}$ et $\boldsymbol{c}$. $U(\boldsymbol{r}, \boldsymbol{c})$ peut être interprété de manière probabiliste : pour $\boldsymbol{X}$ et $\boldsymbol{Y}$ deux variables aléatoires multinomiales prenant valeurs dans $\{1, \ldots, d\}$, chacune avec pour distribution $\boldsymbol{r}$ et $\boldsymbol{c}$ respectivement, l'ensemble $U(\boldsymbol{r}, \boldsymbol{c})$ contient toutes les probabilités jointes possibles de $(\boldsymbol{X}, \boldsymbol{Y})$. En effet, chaque matrice $\boldsymbol{P} \in U(\boldsymbol{r}, \boldsymbol{c})$ peut être identifiée avec une probabilité jointe de $(\boldsymbol{X}, \boldsymbol{Y})$ telle que $p(\boldsymbol{X} = i, \boldsymbol{Y} = j) = p_{ij}$. Nous définissons l'entropie h et la divergence de Kullback-Leibler de $\boldsymbol{Q}, \boldsymbol{P} \in U(\boldsymbol{r}, \boldsymbol{c})$ et d'une marginale $\boldsymbol{r} \in \boldsymbol{\Sigma}_d$ comme

$$h(\boldsymbol{r}) = -\sum_{i=1}^{d} r_i \log(r_i),\ h(\boldsymbol{P}) = -\sum_{i,j=1}^{d} p_{ij} \log(p_{ij}),\ \boldsymbol{KL}(\boldsymbol{P}||\boldsymbol{Q}) = \sum_{ij} p_{ij} \log(\frac{p_{ij}}{q_{ij}}). \tag{2.12}$$

Étant donné une matrice $d \times d$ de coûts $\boldsymbol{M}$, le coût du transport de $\boldsymbol{r}$ à $\boldsymbol{c}$ en utilisant une matrice de transport (ou une probabilité jointe) $\boldsymbol{P}$ peut être quantifié comme $\langle \boldsymbol{P}, \boldsymbol{M} \rangle$.

Le problème du transport optimal entre $\boldsymbol{r}$ et $\boldsymbol{c}$ étant donné une matrice de coûts M se définit alors par :

$$d_{\boldsymbol{M}}(\boldsymbol{r}, \boldsymbol{c}) := \min_{\boldsymbol{P} \in U(\boldsymbol{r}, \boldsymbol{c})} \langle \boldsymbol{P}, \boldsymbol{M} \rangle. \tag{2.13}$$

Nous pouvons maintenant définir la distance Sinkhorn,

$$d_{\boldsymbol{M},\alpha}(\boldsymbol{r}, \boldsymbol{c}) := \min_{\boldsymbol{P} \in U_\alpha(\boldsymbol{r}, \boldsymbol{c})} \langle \boldsymbol{P}, \boldsymbol{M} \rangle, \tag{2.14}$$

où $U_\alpha(\boldsymbol{r}, \boldsymbol{c})$ est défini comme suit,

$$\begin{aligned} U_\alpha(\boldsymbol{r}, \boldsymbol{c}) &:= \{\boldsymbol{P} \in U(\boldsymbol{r}, \boldsymbol{c}) | \boldsymbol{KL}(\boldsymbol{P}||\boldsymbol{r}\boldsymbol{c}^T) \leq \alpha\}, \\ &= \{\boldsymbol{P} \in U(\boldsymbol{r}, \boldsymbol{c}) | h(\boldsymbol{P}) \geq h(\boldsymbol{r}) + h(\boldsymbol{c}) - \alpha\} \subset U(\boldsymbol{r}, \boldsymbol{c}). \end{aligned} \tag{2.15}$$

Quand α est suffisamment grand, la distance Sinkhorn coïncide avec la distance définie à l'équation (2.13). Quand $\alpha = 0$, la distance Sinkhorn a une forme fermée (cf. [36] pour plus de détails).

En pratique, la représentation par sac de vecteurs suivie par le calcul des distances Sinkhorn ne donne pas directement une représentation vectorielle. Pour obtenir celle-ci, un plongement euclidien de la matrice des distances peut être effectué en appliquant simplement un *multidimensional scaling* métrique [110].

Ainsi, si $\boldsymbol{D}$ est une matrice de distance, pas nécessairement euclidienne, $\boldsymbol{B}$ la matrice de produit scalaire associée, alors, pour une dimension q fixée, la configuration issue du *multidimensional scaling* a une matrice de distance $\hat{\boldsymbol{D}}$ qui rend $\sum_{i,j=1}(d_{ij} - \hat{d_{ij}})^2$ minimum et, c'est équivalent, une matrice de produit scalaire $\hat{\boldsymbol{B}}$ qui minimise $\left\|\boldsymbol{B} - \hat{\boldsymbol{B}}\right\|^2$.

Chapitre 3

Constitution d'un jeu de données de rapports 8-K

Résumé : Dans ce chapitre, nous constituons une base de données de rapports 8-K pour les années 2015 à 2019 des entreprises du Standard & Poor's 500. Ces rapports, fournissant des détails exhaustifs et de qualité sur la santé économique ainsi que le contexte dans lequel l'entreprise évolue, pourraient nous aider à prédire les variations des cours sur le marché. De plus, nous effectuons une analyse exploratoire sur l'ensemble du jeu de données et une analyse détaillée pour l'entreprise Wells Fargo pour les années 2015 - 2016.

3.1 Introduction

Une des grandes difficultés pour tenter de comprendre l'évolution des mouvements des marchés financiers est l'obtention de données qualitatives concernant les entreprises et les événements qui les impactent. La plupart du temps, ces données appartiennent à de grands agrégateurs de données comme Bloomberg [1], Thomson Reuters [2] ou encore SIX Financial Information [3] qui ne les mettent pas à disposition gratuitement. En ce sens, de nombreuses études se sont concentrées sur l'analyse des tweets [92, 115] et donc une information déjà disponible par les différents acteurs du marché.

De ce fait, les formulaires 8-K constituent un ensemble d'informations précieuses pour comprendre, par exemple, l'influence de divers événements sur le cours des actions. Étant donné que les rapports 8-K sont légalement obligatoires et mis à disposition publiquement, ils fournissent une base complète et empêchent le biais de sélection de l'échantillon.

1. https://www.bloomberg.com/
2. https://www.thomsonreuters.com/en.html
3. https://www.six-group.com/fr/products-services/financial-information.html

Ces rapports, émis par l'entreprise, représentent donc une information de meilleure qualité que les tweets en fournissant des détails intrinsèques à sa santé économique et financière, mais aussi au contexte dans lequel elle évolue, ce qui pourrait nous aider à mieux prédire les variations du cours de son action.

3.2 Description des rapports 8-K et de l'indice $S\&P500$

Cette section présente les rapports 8-K en résumant leur contenu et les contraintes légales associées. Nous rappelons ensuite ce qu'est le $S\&P500$.

3.2.1 Les rapports 8-K

Un rapport ou formulaire 8-K de l'anglais *Form 8-K*, abrégé en 8-K dans ce texte, est un rapport d'événements importants non prévus ou de changements au sein d'une entreprise qui pourraient être importants pour les actionnaires ou la Securities and Exchange Commission (SEC) [4], organisme fédéral américain de réglementation et de contrôle des marchés financiers. Ce rapport informe le public d'événements tels que des acquisitions, une faillite, la démission d'administrateurs ou des changements dans l'exercice financier, dont la liste officielle peut se trouver dans la table 3.1. L'entreprise dispose généralement de quatre jours ouvrables pour déposer un 8-K pour la plupart des éléments spécifiés. Il est généralement composé de deux parties, une première, décrivant les événements et une deuxième, permettant à l'entreprise d'ajouter des pièces jointes qu'elle juge utiles comme des communiqués de presse ou des états financiers.

Les documents répondant aux exigences de la réglementation de divugaltion équitable, *Regulation Fair Disclosure* (Reg FD) [5], peuvent être exigés pour une période plus courte que quatre jours ouvrables pour améliorer la confiance des investisseurs dans l'intégrité des marchés financiers. L'objectif de la Reg FD est d'accroître la transparence, la responsabilité et, fondamentalement, d'uniformiser les règles du jeu entre les investisseurs individuels et les investisseurs institutionnels. Par exemple, par le passé, de nombreux émetteurs divulguaient des informations importantes non publiques, telles que des avertissements préalables sur les résultats, à des analystes financiers ou à des investisseurs institutionnels sélectionnés, ou aux deux, avant de divulguer intégralement ces mêmes informations au grand public. Lorsque cela se produisait, ceux qui ont eu connaissance de l'information à l'avance ont pu réaliser un bénéfice ou éviter une perte au détriment de ceux qui étaient tenus dans l'ignorance.

Une organisation doit donc déterminer si l'information est importante et soumettre le rapport à la SEC. Cette dernière met les rapports à disposition via la plateforme EDGAR (Electronic Data Gathering, Analysis, and Retrieval) [6]. Dans le cas où l'entreprise ne remplirait pas ses obligations en soumettant son rapport à temps, elle s'exposerait à

4. https://www.sec.gov/.
5. https://www.sec.gov/rules/final/33-7881.htm.
6. https://www.sec.gov/edgar/searchedgar/companysearch.html.

des sanctions administratives de la part de la SEC dont la pénalité dépendra des raisons et de la date à laquelle le 8-K a finalement été déposé. De plus, les bourses, sur lesquelles l'action est cotée, exigeraient que l'entreprise publie un communiqué de presse annonçant son incapacité à publier un rapport dans les délais, ce qui pourrait nuire au cours de son action et à sa réputation auprès des investisseurs.

La SEC décrit les différentes situations qui nécessitent un formulaire 8-K. Il y a neuf sections, chacune composée de une à huit sous-sections, dans le Bulletin des investisseurs qui sont organisées ainsi :

— Section 1 : *Registrant's Business and Operations* (conclusion ou résiliation d'une entente importante ou bien faillite) ;
— Section 2 : *Financial Information* (acquisition ou cession d'actifs, résultats d'exploitation et situation financière, …) ;
— Section 3 : *Securities and Trading Markets* (avis de radiation de la cote, ventes non enregistrées de titres de participation, …) ;
— Section 4 : *Matters Related to Accountants and Financial Statements* (changements de l'expert-comptable, Non-respect des états financiers publiés, …) ;
— Section 5 : *Corporate Governance and Management* (départ d'administrateurs ou de certains dirigeants, soumission de questions au vote d'actionnaires, …) ;
— Section 6 : *Asset-Backed Securities* (changement de prestataire de services ou de fiduciaire,…) ;
— Section 7 : *Regulation FD* (divulgation de la réglementation FD) ;
— Section 8 : *Other Events* (l'entreprise peut utiliser cette rubrique pour signaler des événements qui ne sont pas spécifiquement prévus par le formulaire 8-K, mais que le déclarant considère comme importants) ;
— Section 9 : *Financial Statements and Exhibits* (états financiers et annexes).

La liste des événements est présentée dans la table 3.1. Un exemple de rapport 8-K de l'entreprise Amazon est proposé dans la figure 3.1 ainsi que le formulaire vierge en annexe A.

En plus des 8-K, l'entreprise doit fournir des 10-Q ainsi que des 10-K à la SEC. Les 10-Q sont des rapports trimestriels qui sont obligatoires pour les trois premiers trimestres de l'année. Ils présentent les états financiers non audités et la situation financière de l'entreprise sur le trimestre. Pour le dernier trimestre, l'entreprise doit fournir un 10-K. Il s'agit d'un rapport annuel résumant la performance financière de l'organisation et incluant son histoire, sa structure organisationnelle, ses états financiers audités, son bénéfice par action, ses filiales, la rémunération de ses dirigeants et toute autre donnée pertinente.

Code	Type	Occurrences
1	*Regulation FD Disclosure*	8606
2	*Financial Statements and Exhibits*	28264
3	*Other Events*	9001
4	*Material Modifications to Rights of Security Holders*	319
5	*Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers*	6914
6	*Completion of Acquisition or Disposition of Assets*	538
7	*Entry into a Material Definitive Agreement*	3743
8	*Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant*	2090
9	*Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year*	1449
10	*Results of Operations and Financial Condition*	10637
11	*Submission of Matters to a Vote of Security Holders*	2686
12	*Cost Associated with Exit or Disposal Activities*	230
13	*Unregistered Sales of Equity Securities*	161
14	*Termination of a Material Definitive Agreement*	475
15	*Material Impairment*	105
16	*Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing*	99
17	*Changes in Control of Registrant*	74
18	*Mine Safety - Reporting of Shutdowns and Patterns of Violations*	22
19	*Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement*	40
20	*Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics*	38
21	*Temporary Suspension of Trading Under Registrant's Employee Benefit Plans*	98
22	*Changes in Registrant's Certifying Accountant*	26
23	*Shareholder Nominations Pursuant to Exchange Act Rule a-*	9
24	*Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review*	11
25	*Bankruptcy or Receivership*	5

TABLE 3.1 – Liste complète officielle avec le nombre total d'occurrences des événements du corpus pour les années 2015 à 2019 contenant 37238 rapports (un rapport peut contenir plusieurs événements).

Evénement : *DEPARTURE OF DIRECTORS OR CERTAIN OFFICERS; ELECTION OF DIRECTORS; APPOINTMENT OF CERTAIN OFFICERS; COMPENSATORY ARRANGEMENTS OF CERTAIN OFFICERS.*;

Texte : "On September 20, 2021, the Board of Directors of Amazon.com, Inc. (the "Company") elected Edith W. Cooper as a director of the Company, and also appointed her to the Leadership Development and Compensation Committee of the Board. Ms. Cooper is a co-founder of Medley Living, Inc., a membership-based community for personal and professional growth that launched in September 2020. In addition, Ms. Cooper served as Executive Vice President, Global Head of Human Capital Management of Goldman Sachs Group, Inc. from March 2008 to December 2017. Previously at Goldman Sachs, Ms. Cooper led various client franchise businesses for the firm. Ms. Cooper has served as a director of PepsiCo, Inc. since September 2021, a director of MSD Acquisition Corp. since March 2021, a director of EQT AB since October 2018, a director of Etsy, Inc. from April 2018 to September 2021, and a director of Slack Technologies, Inc. from January 2018 to July 2021. In connection with her election, Ms. Cooper was granted a restricted stock unit award under the Company's 1997 Stock Incentive Plan for 285 shares of common stock of the Company, to vest in three equal annual installments beginning on November 15, 2022, assuming continued service as a director. Ms. Cooper also entered into an indemnification agreement with the Company in the same form as its other directors have entered, which is filed as an exhibit to Amendment No. 1, filed April 21, 1997, to the Company's Registration Statement on Form S-1 (Registration No. 333-23795).".

Figure 3.1 – Exemple de rapport 8-K de l'entreprise Amazon publié le 20 septembre 2021

3.2.2 L'indice $S\&P500$

L'indice $S\&P500$, ou indice Standard $\&$ Poor's 500, est un indice pondéré en fonction de la capitalisation boursière des 500 principales sociétés cotées en bourse aux États-Unis (NYSE [7], NASDAQ [8], Cboe BZX Exchange [9]). Il a a été officiellement introduit le 4 mars 1957 par la société américaine du même nom, Standard $\&$ Poor's, qui est une société de notation financière. Il est considéré comme l'un des meilleurs indicateurs pour les actions américaines à grande capitalisation. Il ne s'agit pas d'une liste exacte des 500 premières sociétés américaines en fonction de leur capitalisation boursière, car il existe d'autres critères pour être inclus dans l'indice. Un comité sélectionne chacune des 500 sociétés de l'indice en fonction de leur liquidité, de leur taille et de leur secteur d'activité. Il rééquilibre l'indice tous les trimestres, en mars, juin, septembre et décembre.

Pour faire partie de l'indice, une société doit être située aux États-Unis et avoir une capitalisation boursière d'au moins 11,8 milliards de dollars. Au moins 50% des actions de la société doivent être accessibles au public. Le prix unitaire des actions doit être d'au moins un dollar. Elle doit déposer un rapport annuel 10-K. Au moins 50% de ses actifs fixes et de ses revenus doivent se trouver aux États-Unis. Enfin, elle doit avoir au moins quatre trimestres consécutifs de bénéfices positifs.

Au 30 septembre 2021, les 9 plus grandes entreprises de la liste des sociétés du $S\&P500$ représentaient 28.1% de la capitalisation boursière de l'indice et étaient, par ordre de pondération, Apple Inc, Microsoft, Alphabet Inc. (y compris les actions de classe A et C [10]), Amazon.com, Facebook, Tesla Inc. et Nvidia, Berkshire Hathaway et JPMorgan Chase $\&$ Co.

3.3 Constitution du jeu de données

Nous décrivons dans cette section les sources et la méthodologie utilisées pour constituer notre jeu de données.

Comme vu précédemment, suivant les évènements auxquels une entreprise est confrontée, sa situation sur les marchés peut évoluer et son action peut donc être listée ou, au contraire, délistée de l'indice $S\&P500$. Il était donc nécessaire avant de s'intéresser aux rapports 8-K, d'obtenir la liste la plus exhaustive possible des entreprises faisant, ou ayant fait, partie de l'indice $S\&P500$. Nous nous sommes basés sur la liste proposée par *Wikipedia* [11], en considérant les entreprises pour les années 2015 à 2019, pour mettre en place un algorithme de *scraping* permettant de récupérer les rapports disponibles sur la plateforme EDGAR de la SEC.

Enfin, pour être en mesure d'étudier l'impact d'un rapport sur le cours de l'action,

7. The New York Stock Exchange.
8. National Association of Securities Dealers Automated Quotations.
9. Chicago Board Options Exchange.
10. Les actions de classe A sont associées à un droit de vote au contraire des actions de classe C.
11. https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

nous avons cherché à obtenir les données boursières des entreprises présentes dans notre liste $S\&P500$. Pour faire cela, trois sources de données ont été agrégées :

— la bibliothèque R BatchGetSymbols [12] pour récupérer celles disponibles sur Yahoo Finance ;
— celles mises à disposition par Cameron Nugent [13] ;
— celles disponibles sur IEX Stocks [14].

L'utilisation de plusieurs sources a été nécessaire car, la plupart du temps, ces sources ne stockent pas les cours boursiers d'entreprises ayant subi une faillite ou ayant été absorbées par une autre entreprise.

3.4 Analyse exploratoire

Suite à la collecte des différentes données pour les années 2015 à 2019, nous avons obtenu 37238 rapports pour 592 entreprises. La figure 3.4 représente pour chaque mois de la période, le nombre d'entreprises ayant publié au moins un formulaire. Nous constatons que la variation mensuelle est relativement importante avec une modification moyenne d'environ 20% entre deux mois consécutifs.

3.4.1 Pré-traitement des textes

Une des difficultés, avant d'analyser les rapports, est d'en extraire les textes. Suite au téléchargement, nous avons obtenu des fichiers *HTML* qui n'étaient pas toujours encodés de manière similaire et pouvaient contenir des pièces jointes dans un autre format. Dans ce cadre, pour effectuer les différentes analyses qui suivent dans ce chapitre et dans le reste de la thèse, nous avons mis en place une seule méthodologie de base pour le traitement de textes bruts :

1. extraction des informations importantes du texte contenu dans le *HTML* (évènements et des dates de publication pour les rapports 8-K par exemple) [15] ;
2. mise en casse minuscule du texte ;
3. suppression des caractères non-alphanumériques excepté la ponctuation ;
4. lemmatisation de chaque mot [16] ;
5. suppression des mots vides et des nombres ;
6. suppression des unigrammes très rares apparaissant moins de dix fois [17].

12. https://cran.r-project.org/web/packages/BatchGetSymbols/index.html
13. https://camnugent.wordpress.com/
14. https://iextrading.com/apps/stocks/
15. Nous ne travaillons pas sur les pièces jointes éventuelles.
16. Rappelons que la lemmatisation consiste à donner à un mot sa forme neutre canonique.
17. Pour des raisons de temps de calcul dans la **section 6.5.2** (p. 123), nous avons supprimé les unigrammes apparaissant moins de cent fois.

Suite à ce pré-traitement général, nous avons représenté les textes, dans la plupart des cas, sous une forme vectorielle faisant abstraction de l'ordre des mots. Chaque texte est ainsi représenté par un vecteur contenant les occurrences des mots du corpus qui le composent. Cette occurrence a pu être pondérée, nous le préciserons le cas échéant, à l'aide de la méthode *TF-IDF* ou par l'information mutuelle par exemple.

Cependant, dans des cas spécifiques, il était nécessaire de garder l'ordre des mots notamment pour ajouter des marques de négativité ou, comme nous le détaillerons dans la **section 4.5.2** (p. 52), lors de la représentation des textes sous forme de phrase.

Les figures 3.2 et 3.3 permettent de se rendre compte de la grande variabilité des textes et des mots qui les composent. La figure 3.2 représente la distribution du nombre de mots uniques par texte, sur laquelle nous voyons qu'une partie des textes est composée de moins de 2500 mots. La figure 3.3 illustre la distribution de la longueur des textes, ici aussi, nous notons une grande variabilité, avec un léger pic pour les textes contenants moins de 5000 termes correspondant dans leur grande majorité à l'événement *Financial Statements and Exhibits*.

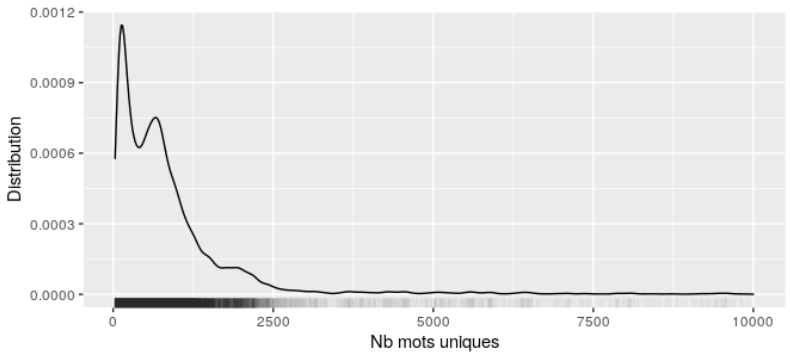


FIGURE 3.2 – Distribution des mots uniques par rapport 8-K pour les années 2015 à 2019.

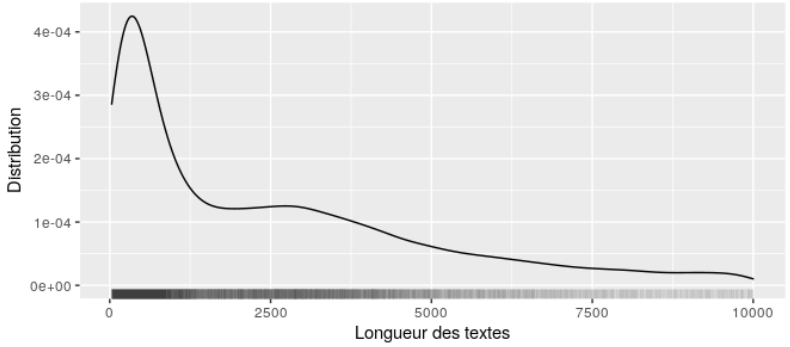


FIGURE 3.3 – Distribution de la longueur des rapports 8-K pour les années 2015 à 2019.

3.4.2 Distribution des événements

Dans la table 3.1, les 25 évènements possibles pour un rapport 8-K sont présents avec leurs occurrences. Les évènements *Financial Statements and Exhibits* et *Results of Operations and Financial Condition* sont les plus présents au contraire des évènements *Shareholder Nominations Pursuant to Exchange Act Rule a-* et *Bankruptcy or Receivership* qui apparaissent moins de dix fois. Certains évènements, apparaissant plus de 1000 fois, sont liés à la vie courante d'une entreprise comme la présentation des résultats financiers. Les autres, bien plus rares, sont liés à des évènements impactant plus fortement l'entreprise tels que des faillites.

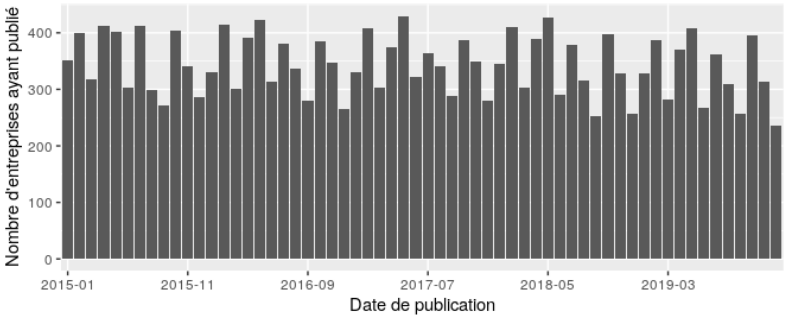


Figure 3.4 – Nombre d'entreprises ayant publié par mois pour les années 2015 à 2019.

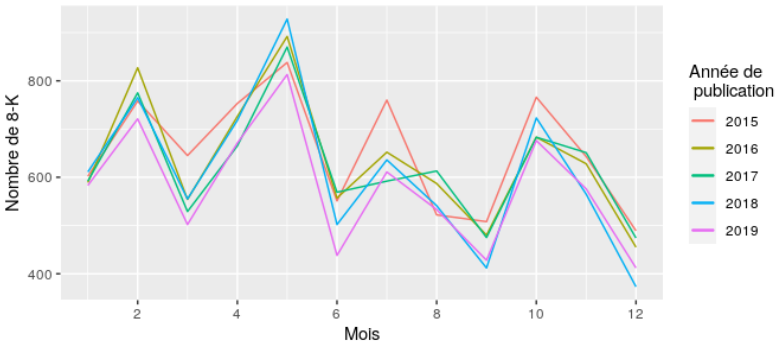


Figure 3.5 – Nombre de rapports publiés par mois pour les années 2015 à 2019.

3.4.3 Saisonnalité

Sur les figures 3.5, 3.6 et 3.7, nous pouvons apercevoir la répartition des rapports respectivement par mois, jour et heure. La figure 3.5 fait apparaître un pic de publication le mois précédent la fin de chaque trimestre. Ce pic s'explique par la volonté des entreprises de publier des informations financières qui résument généralement les états financiers complets, lesquels figureront ultérieurement dans son rapport trimestriel (formulaire 10-Q) ou annuel (formulaire 10-K).

La figure 3.6 montre que les entreprises privilégient les mardis et jeudis pour la publication des rapports. Le jeudi s'explique par le fait que tous les évènements ayant lieu le vendredi et pendant le week-end doivent être annoncés pendant cette journée.

La figure 3.7 montre que les entreprises publient les rapports juste avant l'ouverture, à 9h30, et après la fermeture des marchés, 16h, jusqu'à 17h30. Il est préférable de publier à ces heures-ci pour ne pas surprendre les investisseurs et permettre aux services de relation des entreprises de ne pas être pris au dépourvu par tous les appels de ces derniers. De plus, pour respecter les dates limites de publication, un rapport doit être soumis avant 17h30 pour être comptabilisé le jour même. Les soumissions qui sont faites après 17h30 sont considérées comme étant déposées le jour ouvrable suivant [18].

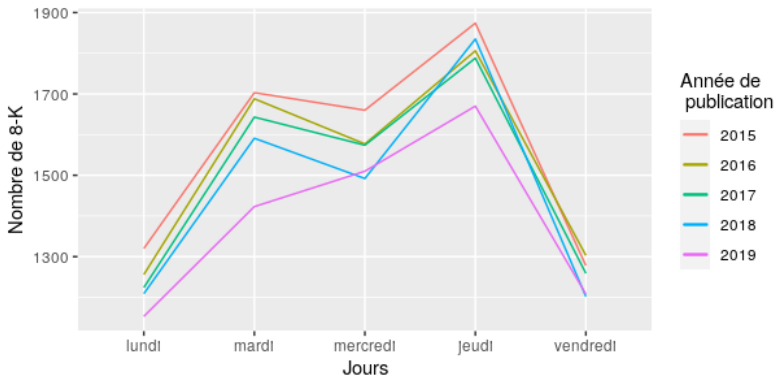


FIGURE 3.6 – Nombre de rapports publiés par jour pour les années 2015 à 2019.

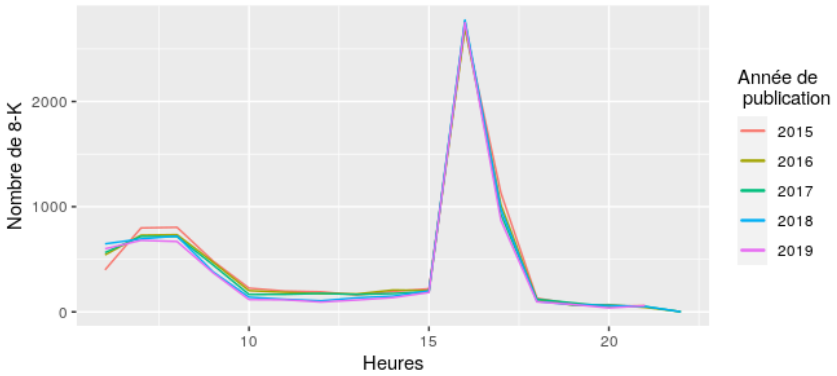


FIGURE 3.7 – Nombre de rapports publiés par heure pour les années 2015 à 2019.

Nous regardons maintenant si une temporalité des événements peut être constatée à l'image des observations faites sur la figure 3.5. Sur les figures 3.8, le nombre d'évènements par mois est représenté. Pour une meilleure visualisation, nous avons scindé en trois les évènements selon leurs occurrences :

18. https://www.wilmerhale.com/en/insights/publications/20201019-keeping-current-with-form-8-k-a-practical-guide.

— Évènements communs : > 1000 ;

— Évènements moins communs : < 1000 et > 97 ;

— Évènements rares : < 97.

Comme nous l'avons précisé plus tôt, les évènements des deux premières figures sont liés au cycle de vie normal d'une entreprise au contraire de ceux de la troisième figure qui sont beaucoup plus rares et impliquent un changement profond comme une banqueroute.

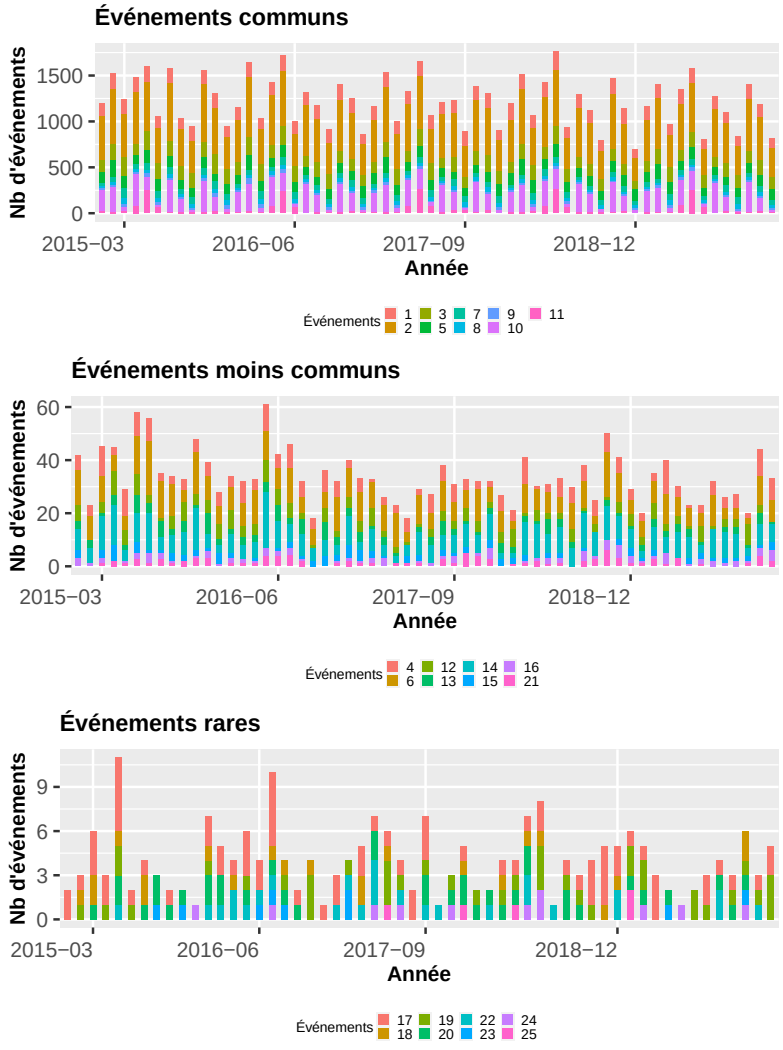


Figure 3.8 – Type d'événements par mois pour la période 2015 - 2019. Les événements sont arrangés par occurrence. Les événements communs apparaissent plus de 1000 fois, les événements moins communs se produisent entre 1000 et 97 fois et les événements rares présents moins de 97 fois.

Cet évènement *Bankruptcy or Receivership* apparaît 5 fois, entre juin 2017 et janvier 2019, en lien avec trois entreprises *NRG* (NRG Energy), *FE* (First Energy) et *PCG* (Pacific

Gas and Electric Company) du secteur de l'énergie. Ces entreprises ont été victimes d'une réorganisation de secteur [19] ou de catastrophes naturelles [20].

La figure 3.8 ne permet donc pas de trouver de temporalité dans les événements qui apparaissent selon le contexte qui entoure une entreprise.

3.4.4 Analyse exploratoire sur un échantillon

Du fait du nombre important de rapports, nous nous concentrons ici, pour une analyse plus fine, sur une période de temps restreinte, 2015 - 2016, et sur l'entreprise ayant le plus publié durant celle-ci à savoir Wells Fargo (WFC) [10].

Cette entreprise a publié 248 rapports pour les années 2015 et 2016. Sur les 25 évènements possibles, seuls 7 sont représentés, avec une domination de l'évènement *"financial statements and exhibits"*, ce qui tend à montrer que ces rapports ont pour sujet principal l'état financier de l'entreprise (cf la table 3.2 pour les intitulés des évènements et leur fréquences).

Dans le graphique 3.9, nous pouvons observer la répartition des publications selon les mois au cours des années 2015 - 2016 et l'augmentation de ces publications pour l'année 2016.

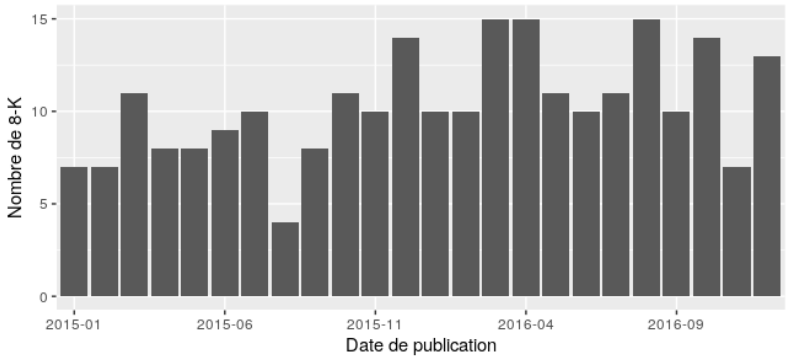


FIGURE 3.9 – Nombre de 8K par mois pour l'entreprise Wells Fargo pour les années 2015 à 2016.

3.4.4.1 Temporalité et saisonnalité

Dans la figure 3.10, une certaine similarité des courbes peut être observée notamment à l'approche de chaque fin de trimestre.

19. https://static1.squarespace.com/static/5b64a999a2772cef1fe10e54/t/5bf59239562fa7445421483d/1542820412197/GenOnEnergy_Inc_Interim_Financial_Report_September_30_2018.pdf.
20. https://finance.yahoo.com/news/pg-e-corp-files-chapter-11-bankruptcy-protection-082001142--finance.html.

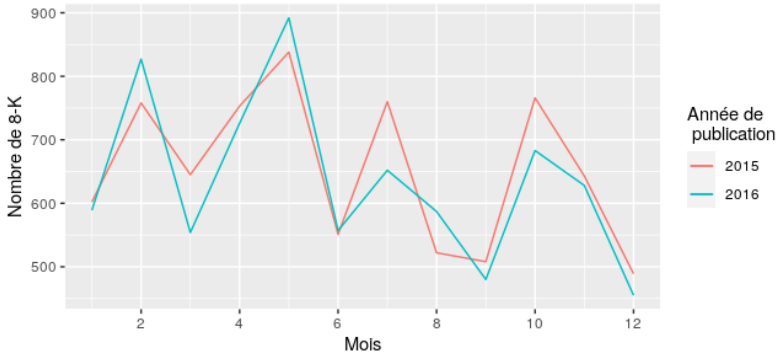


Figure 3.10 – Nombre de 8K par mois selon les années pour l'entreprise Wells Fargo pour les années 2015 à 2016.

Code	Type	Occurrences
1	*amendments to articles of incorporation or bylaws; change in fiscal year*	11
2	*amendments to the registrant's code of ethics, or waiver of a provision of the code of ethics*	1
3	*departure of directors or certain officers; election of directors; appointment of certain officers : compensatory arrangements of certain officers*	9
4	*financial statements and exhibits*	243
5	*other events*	8
6	*results of operations and financial condition*	9
7	*submission of matters to a vote of security holders*	2

Table 3.2 – Évènements pour Wells Fargo pour les années 2015 - 2016.

La figure 3.11 montre le peu de similarité entre les deux années, nous notons tout de même de nombreuses publications le vendredi pour l'année 2016.

Suite aux diverses condamnations lors de l'année 2016 par rapport à l'année 2015, nous notons sur le graphique 3.12 une grande différence dans l'heure de publication des rapports. En effet, lors de l'année 2016, l'entreprise a publié beaucoup de plus de rapport en lien avec l'actualité et notamment :

— son exposition, par rapport à la chute des cours de pétrole [21] ;

— diverses condamnations [22] ;

— démission du directeur de la banque et licenciement d'employés en lien avec les condamnations [23].

En 2015, l'entreprise publiait l'essentiel de ces rapports à 15 heures soit 1 heure avant la fermeture des marchés. Au contraire, une autre stratégie émerge pour l'année 2016 avec l'essentiel des publications ayant lieu vers 12 heures. Sans doute pour mieux appréhender les variations de marché en ayant la possibilité de donner des explications avant la fermeture des cotations ou suite au changement de direction et à la mise en place d'une nouvelle stratégie de communication.

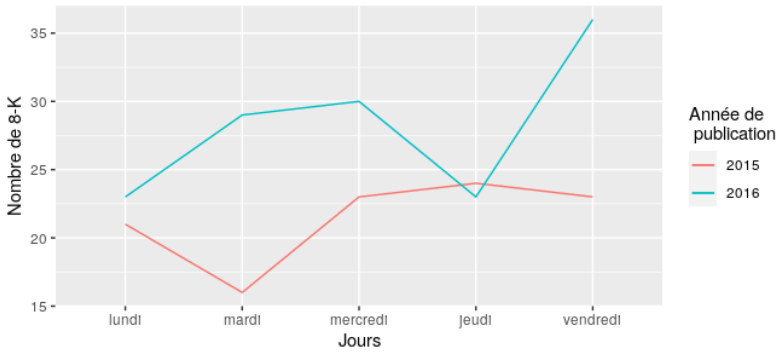


FIGURE 3.11 – Nombre de 8K par jour pour l'entreprise Wells Fargo pour les années 2015 à 2016.

3.4.4.2 Comparaisons de deux représentations

Dans cette partie, nous souhaitons observer si une structure apparaît entre les textes de ce corpus. Pour ce faire, nous représentons les textes de différentes manières et effectuons une analyse basée sur un regroupement hiérarchique à l'aide de la méthode de Ward.

21. https://www.bloomberg.com/news/articles/2016-01-15/banks-brace-for-bigger-losses-as-oil-drops-below-30-a-barrel
22. https://www.justice.gov/opa/pr/wells-fargo-bank-agrees-pay-12-billion-improper-mortgage-lending-practices
23. https://www.usatoday.com/story/money/2016/09/08/wells-fargo-fined-185m-over-unauthorized-accounts/90003212/

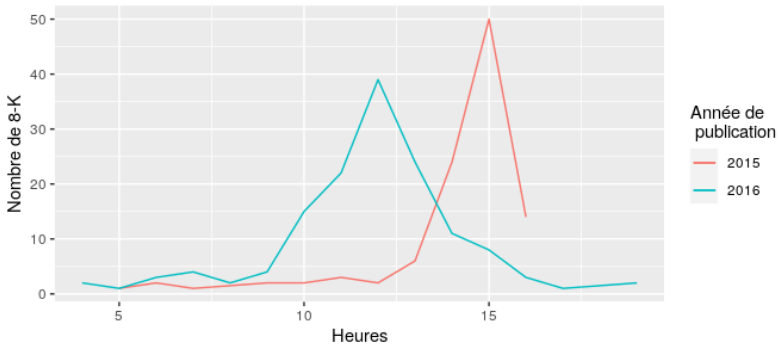


Figure 3.12 – Nombre de 8K par heure pour l'entreprise Wells Fargo pour les années 2015 à 2016.

Différentes représentations de texte ont été abordées dans l'état de l'art, ici nous explorons un modèle vectoriel, les unigrammes, et un plongement de mots, GloVe [89] pour montrer leurs influences sur les résultats. Cette influence sera plus longuement étudiée dans le **chapitre 5** (p. 59).

Dans ce qui suit et suivant le pré-traitement énoncé précédemment, seuls 3 778 mots (racines) sont utilisés dans les rapports et nous faisons abstraction des données temporelles telles que les dates, les années ou les heures ainsi que des événements.

Unigramme Les rapports sont ici représentés sous forme d'unigrammes. Une normalisation par la norme 1 ainsi que le calcul de la distance euclidienne entre ces textes ont été effectués. Dans ce cas, comme le vecteur du texte A et celui du texte B deviennent des vecteurs unitaires, l'égalité suivante entre la distance euclidienne et la similarité cosinus [24] apparaît :

$$\|\boldsymbol{w}_A - \boldsymbol{w}_B\|_2 = 2(1 - \cos(\boldsymbol{w}_A, \boldsymbol{w}_B)) \tag{3.1}$$

où $\boldsymbol{w}_A$ et $\boldsymbol{w}_B$ sont les représentations vectorielles des textes A et B.

Pour sélectionner le nombre de classes, la partition ayant la plus grande perte relative d'inertie a été choisie. En prenant ce critère, la meilleure partition est celle à deux classes suivie par celle à trois classes, cf. figure 3.13.

Pour une analyse plus fine des rapports, nous privilégions un découpage en trois classes. La figure 3.14 montre ce partitionnement sur le dendrogramme.

La répartition des rapports selon leur classe se trouve dans la table 3.3. Nous constatons que les classes 1 et celle contenant 9 rapports sont assez stables, quelque soit le nombre de classes.

24. Mesure souvent employée pour comparer la similitude entre deux textes comme cela est abordé dans le **chapitre 6** (p. 78).

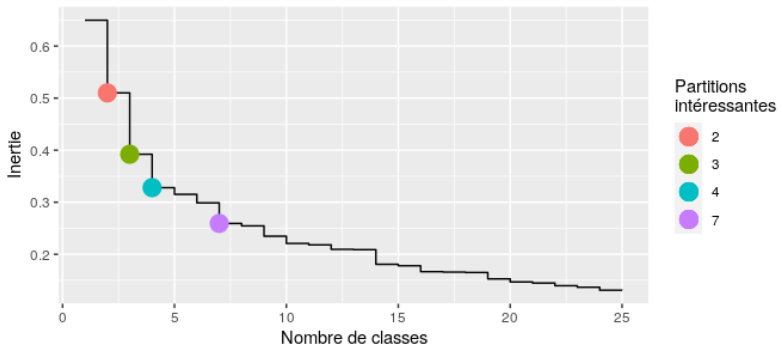


FIGURE 3.13 – Représentation de l'inertie en fonction du nombre de classes pour la représentation des textes par unigramme.

TABLE 3.3 – Distribution des rapports par nombre de classes pour la représentation des textes par unigramme.

	Nombre de rapports par classe						
Nombre de classes	1	2	3	4	5	6	7
2	239	9	-	-	-	-	-
3	144	95	9	-	-	-	-
4	144	57	38	9	-	-	-
7	144	8	11	31	7	38	9

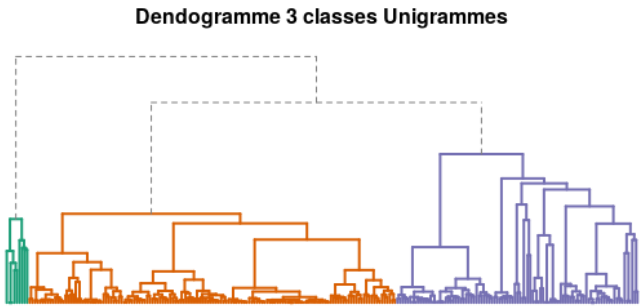


FIGURE 3.14 – Dendrogramme 3 classes pour la représentation des textes par unigramme.

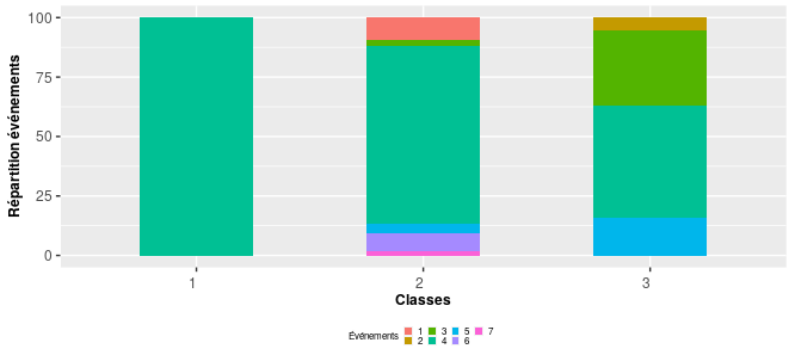


Figure 3.15 – Répartition des évènements selon les classes pour la représentation des textes par unigramme.

La figure 3.15 montre que la classe 1 contient principalement l'événement majoritaire *Financial Statements and Exhibits* au contraire des classes 2 et 3 qui sont composées d'événements hétérogènes. Pour la classe 2, il s'agit des événements *amendments to articles of incorporation or bylaws; change in fiscal year* et *results of operations and financial condition* alors que pour la 3 se sont les évènements *amendments to the registrant's code of ethics, or waiver of aprovision of the code of ethics, departure of directors or certain officers; election of directors; appointment of certain officers : compensatory arrangements of certain officers* et *other events*.

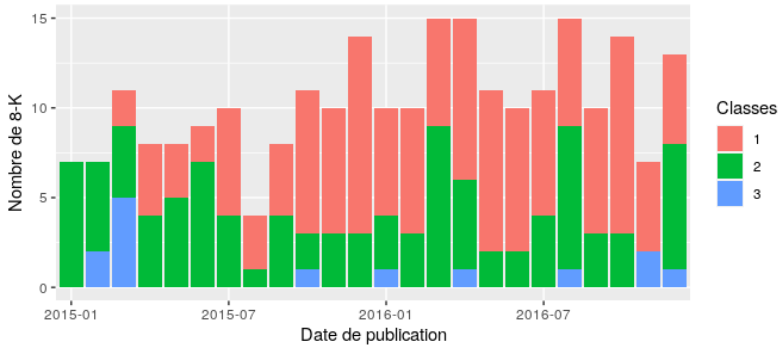


Figure 3.16 – Répartition des classes par mois pour la représentation des textes par unigramme.

La figure 3.16 montre la répartition des classes selon les mois. Il est intéressant de noter que la classe 1 apparaît suite au rapport, appartenant à la classe 2, publié le 25 février 2015 avec l'événement *Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers*. Suite à cet événement, la classe 1 devient plus prépondérante, ce qui nous indique une nouvelle méthode de communication imprégnée par la nouvelle direction et s'inscrit parfaitement avec l'historique de l'entreprise. La classe 3, composée en grande partie des événements *Financial Statements and Exhibits* et *Other Events*, apparaît le plus souvent suite aux rapports trimestriels (février, août par exemple).

Plongement de mots Dans cette partie, les rapports sont représentés à l'aide du plongement de mots GloVe [89]. Pour chaque texte, la moyenne de la représentation des mots qui les composent est calculée. De la même manière que précédemment, les moyennes sont normalisées et nous avons calculé la distance euclidienne entre les représentations des textes.

Pareillement aux unigrammes, la partition ayant la plus grande perte relative d'inertie a été choisie. Selon ce critère, la meilleure partition est celle avec 3 classes. La figure 3.18 montre ce partitionnement sur le dendrogramme.

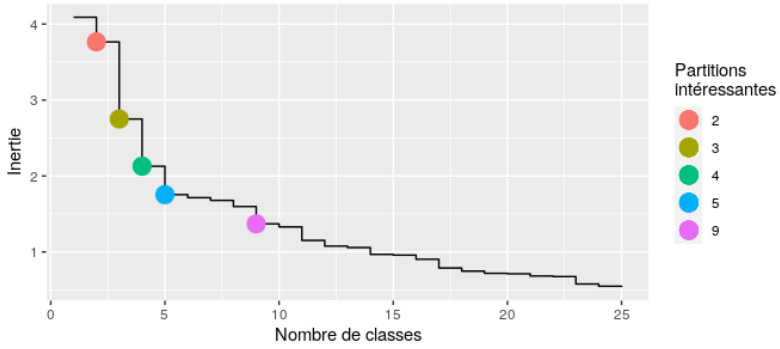


FIGURE 3.17 – Représentation de l'inertie en fonction du nombre de classes pour la représentation des textes par plongement de mots.

TABLE 3.4 – Distribution des rapports par nombre de classes pour la représentation des textes par plongement de mots.

	Nombre de rapports par classe								
Nombre de classes	1	2	3	4	5	6	7	8	9
2	229	19	-	-	-	-	-	-	-
3	221	8	19	-	-	-	-	-	-
4	143	8	78	19	-	-	-	-	-
5	81	8	78	62	19	-	-	-	-
9	81	8	75	24	3	7	38	3	9

Des remarques similaires à celles qui précédent peuvent être établies. Dans la table 3.4, nous observons une certaine stabilité pour la classe 1 et deux autres contenants 8 et 19 rapports.

De plus, la classe 1 contient majoritairement l'évènement *Financial Statements and Exhibits*, visible sur les figures 3.15 et3.20. Sur cette dernière, nous pouvons observer que les deux autres classes sont composées d'événements hétérogènes avec des événements sous-représentés en comparaison de la classe 1. Notamment, la classe 3 où l'évènement *Financial Statements and Exhibits* est sous-représenté.

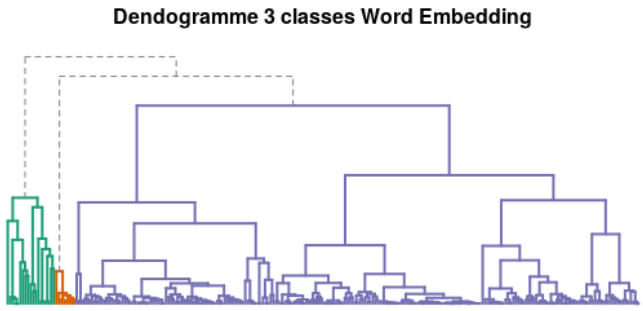


Figure 3.18 – Dendrogramme 3 classes pour la représentation des textes par plongement de mots.

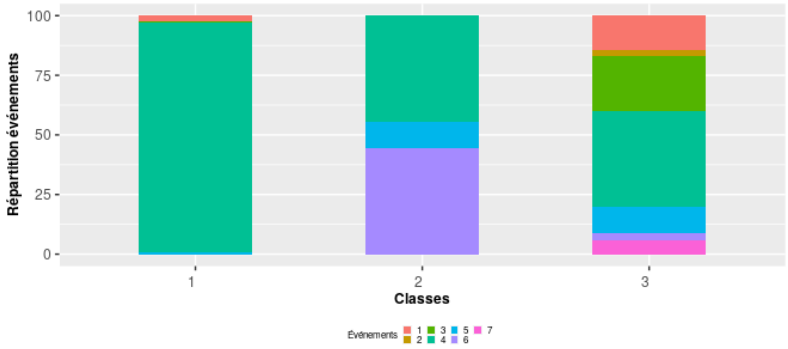


Figure 3.19 – Répartition des évènements selon les classes pour la représentation des textes par plongement de mots.

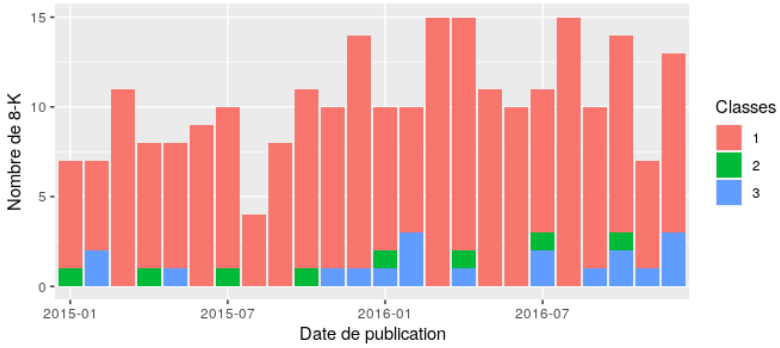


Figure 3.20 – Répartition des classes par mois pour la représentation des textes par plongement de mots.

La figure 3.20 montre la répartition des classes selon les mois de publication. Il est intéressant de noter la saisonnalité de la classe 2 qui apparaît à tous les débuts de trimestre (janvier, avril, juillet et octobre) car composée en majorité de l'évènement *Results of Operations and Financial Condition*. Pour la classe 3 composée en partie par les événements *Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers*, *Submission of Matters to a Vote of Security Holders* et *Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics*, sa part croît, suite aux déboires subis par l'entreprise à partir de la fin de l'année 2015.

Analyse comparative Comparons maintenant les résultats obtenus avec les représentations unigramme et plongement de mots.

A première vue, notamment sur les figures 3.15 et 3.19, une certaine similarité entre les deux regroupements apparaît. Les classes 1 sont composées majoritairement de l'évènement *Financial Statements and Exhibits*. Les classes 2 et 3 sont dans les deux cas, les classes les plus intéressantes avec des évènements hétérogènes. Notamment pour l'évènement *departure of directors or certain officers; election of directors; appointment of certain officers : compensatory arrangements of certain officers* en sur-représentation dans les classes 3.

Pour confirmer cette impression, nous utilisons une méthode visuelle pour comparer les dendrogrammes, appelée *Tanglegram* [104]. Cette méthode permet d'aligner les dendrogrammes de sorte que les paires de feuilles correspondantes, les textes dans notre cas, soient reliées entre elles en minimisant les croisements entre les arêtes. La figure 3.21 montre cet alignement entre les deux dendrogrammes précédents, dans laquelle les lignes en pointillé représentent une combinaison de noeuds/rapports qui n'est pas présente dans l'autre dendrogramme. Le nombre important de lignes en pointillé met en évidence une grande différence pour les partitions de petites tailles, cependant les arêtes permettent de nous faire observer une certaine similitude au niveau des sous-classes.

La table 3.5 vient confirmer la figure 3.21, dans laquelle les Adjusted Rand Index (ARI) [60] entre les différents regroupements sont comparés. Pour les unigrammes, le regroupement à deux classes est très différent des trois autres avec un ARI d'au plus 0.13. Ces trois autres regroupements sont quant à eux beaucoup plus similaires entre eux avec un ARI minimum de 0.79. Nous notons la même différence pour le plongement de mots avec seulement les regroupements à 2 et 3 classes ainsi que ceux à 5 et 9 classes qui obtiennent un ARI de plus de 0.80.

De plus, entre les représentations unigramme et plongement de mots, le regroupement hiérarchique aboutit à des résultats très différents avec un maximum de 0.52 de ARI entre les regroupements à deux classes.

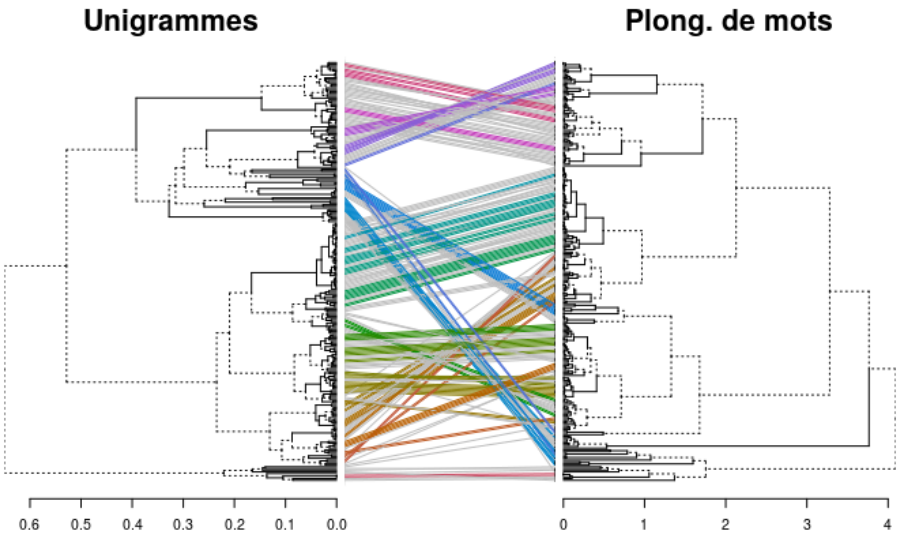


Figure 3.21 – Alignement des dendrogrammes des représentations par unigramme et plongement de mots.

Table 3.5 – Similarité des classes pour les regroupements hiérarchiques selon les ARI.

Classe	U 2	U 3	U 4	U 7	WE 2	WE 3	WE 4	WE 5	WE 9
Uni 2	1	0.13	0.10	0.08	0.52	0.38	0.09	0.04	0.04
Uni 3	-	1	0.86	0.79	0.15	0.18	0.10	0.43	0.38
Uni 4	-	-	1	0.93	0.15	0.20	0.12	0.42	0.49
Uni 7	-	-	-	1	0.17	0.25	0.17	0.46	0.54
WE 2	-	-	-	-	1	0.80	0.23	0.12	0.10
WE 3	-	-	-	-	-	1	0.32	0.17	0.14
WE 4	-	-	-	-	-	-	1	0.65	0.56
WE 5	-	-	-	-	-	-	-	1	0.90
WE 9	-	-	-	-	-	-	-	-	1

Regardons plus spécifiquement deux textes regroupés dans la classe 3 pour le plongement de mots avec comme événements respectifs *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* pour le premier et *Other Events* ainsi que *Financial Statements and Exhibits* pour le deuxième. Le premier texte signifie à la SEC que Madame Elaine L. Chao quitte ses fonctions de membre du Conseil d'administration pour le poste de Secrétaire aux Transports des États-Unis sous l'administration du président Trump. Pour le deuxième, il s'agit d'un rapport [25] faisant suite aux poursuites contre Wells Fargo et dans lequel l'entreprise s'engage à respecter les directives du régulateur. Il s'agit donc de deux textes très différents et notons qu'ils ont été classés différemment pour la représentation unigramme.

La même remarque peut être faite pour deux textes de cette dernière représentation classés dans la classe 3 et ayant les mêmes événements, *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* et *Financial Statements and Exhibits*. La classification les a regroupés au sein de la classe 3, à savoir l'évènement *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* alors que le regroupement faisant suite à la représentation par plongement de mots a préféré inclure le premier texte dans la classe majoritaire 1 comportant principalement l'événement *Financial Statements and Exhibits*.

Nous pouvons voir à travers ces exemples l'importance de la représentation de textes et les différences qu'elle induit dans les résultats obtenus.

3.5 Comparaison à l'existant

De nombreuses études se sont concentrées sur des données de type tweets mais très peu sur les conséquences des rapports 8-K. Seul l'article de Lee *et al.* [72] propose de s'intéresser aux variations engendrées par la parution d'un rapport 8-K d'une entreprise du $S\&P500$ et mettent en avant l'intérêt de l'utilisation de l'analyse sémantique couplée à des données financières pour la prédiction de la variation de l'actif de l'entreprise concernée, cf. l'**annexe B** pour plus de détails.

3.6 Conclusion

Dans cette partie, nous avons présenté le jeu de données que nous avons constitué, basé sur les rapports 8-K des entreprises du $S\&P500$ pour les années 2015 - 2019. Notre base est disponible de manière gratuite et a fait l'objet d'une première analyse dans [10].

Cette partie a permis, tout d'abord, de présenter la procédure d'obtention des données, puis d'y effectuer une analyse globale. Ensuite, nous avons présenté une analyse

25. https://www.sec.gov/Archives/edgar/data/72971/000007297116001369/ex991dec132016.htm.

plus détaillée pour l'entreprise Wells Fargo (WFC).

Enfin, nous avons mis en évidence l'importance jouée par les différentes représentations dans les résultats obtenus avec des algorithmes de regroupement, cf. la figure 3.21 ainsi que la table 3.5, et donc de leur intérêt dans l'obtention de meilleurs résultats.

Dans le prochain chapitre, nous étudierons la possibilité de prédire les variations d'un actif financier, puis dans le chapitre suivant, de mettre en place une méthodologie d'analyse exploratoire grâce à différentes représentations de textes.

Chapitre 4

Prédiction des Marchés financiers

Résumé : Suite à notre analyse exploratoire de la base de données, nous nous intéressons dans ce chapitre à la possibilité de prédire la réaction du marché sur le cours d'un actif, suite à la publication d'un rapport 8-K comme dans [72]. Pour ce faire, nous explorons plusieurs stratégies de représentations de textes ainsi que différents algorithmes d'apprentissage automatique. Nous montrons que la complexification des représentations et des algorithmes de classification n'améliore que très légèrement les résultats.

4.1 Introduction

La prévision du prix des actions est réputée difficile en raison des multiples facteurs qui affectent les fluctuations des cours [102, 18]. Les effets de divers facteurs, tels que les politiques nationales, les relations diplomatiques ou encore l'impact de la psychologie des investisseurs, doivent être pris en compte lors de la prévision des mouvements boursiers.

En ce qui concerne la psychologie des investisseurs, il est logique que ces derniers analysent divers types d'informations avant d'investir. Nous pouvons donc affirmer qu'il existe une forte corrélation entre les informations disponibles et le comportement desdits investisseurs. Donc, l'information joue un rôle très important pour estimer la valeur des performances futures d'une action [28, 76].

Ce phénomène a pris de l'ampleur, d'autant plus qu'il est devenu possible pour tout un chacun de s'informer sur Internet où circule une information abondante, accessible à tous. Ainsi, pour estimer au mieux le prix d'une action, il devient vital d'analyser tout ce flot d'informations, qu'elles proviennent des médias, des réseaux sociaux, des rapports institutionnels ou bien des marchés.

Avec l'avènement de l'ère des Big-Data, une quantité gigantesque de nouvelles financières est publiée chaque jour, et il est devenu difficile pour un individu de sélectionner les informations pertinentes qui affectent le prix d'une action à travers toutes ces sources.

C'est pourquoi, dans le monde universitaire comme dans celui de l'entreprise, des techniques d'explorations de texte dans le but d'analyser et de prédire l'actualité financière ont été largement étudiées et développées, comme par exemple les sacs de mots [38, 83, 72], les N-grammes [50], les *topics models* [84] et les plongements de mots [67].

Dans ce chapitre, nous chercherons la représentation de texte la plus adaptée pour la prédiction du mouvement du cours d'une action. Pour ce faire, nous utiliserons la base de données des rapports 8-K pour les années 2015 à 2017 sur laquelle nous étudierons différentes représentations de texte en adéquation avec plusieurs modèles de classification.

4.2 Prédiction des tendances boursières

L'ensemble de données utilisé a été présenté en détail dans le **chapitre 3** (p. 18). Dans le présent chapitre, nous nous concentrerons sur un corpus contenant les rapports 8-K des entreprises du $S\&P500$, pour les années 2015 à 2017.

Durant cette période, 580 entreprises ont fait partie de l'indice $S\&P500$. Pour ce nombre d'entreprises, 22953 rapports ont été collectés, mais suite à la difficulté d'obtenir toutes les données financières, seuls 21676 ont été conservés.

Dans ce corpus, tous les types d'événements de la table 3.1 sont présents. Notons qu'une entreprise publie 1.83 rapports par mois en moyenne et le maximum de 15 parutions par mois revient à Wells Fargo.

L'objectif est de prédire les effets de la publication d'un rapport 8-K sur le cours des actions d'une société. Pour chaque publication enregistrée dans l'ensemble de données, nous calculons la différence en pourcentage entre les prix des actions de la société déclarante avant et après la publication du rapport. Cette différence est ensuite normalisée en soustrayant la même différence calculée pour la même période sur l'indice S&P 500. Enfin, de manière semblable à la procédure de [72], l'évolution normalisée du cours de l'action est transformée en une variable catégorielle qui donne sa tendance : les augmentations supérieures à 0.5% sont codées comme *UP*, les diminutions inférieures à -0.5% comme *DOWN* et les mouvements avec moins de 0.5% d'amplitude sont codés comme *STAY*. Les figures 4.1 et 4.2 permettent de constater respectivement la densité des fluctuations et les fluctuations par heure. Sur cette dernière, les lignes vertes représentent l'ouverture et la fermeture du marché et les rouges les seuils pour la labellisation (-0.5% et 0.5%).

La tâche de prédiction consiste à classer un communiqué dans l'une de ces trois classes compte tenu des informations disponibles dans le contenu textuel du rapport 8-K. Les années 2015 et 2016 sont utilisées pour l'estimation du modèle et 2017 pour l'évaluation de la qualité (ensemble de test) avec une distribution des classes visible dans la table 4.1. Comme dans [72], nous rapportons la précision des modèles, c'est-à-dire le pourcentage de rapports correctement classés. Remarquons que la structure à trois classes conduit à une précision de $55,5\%$ sur l'ensemble de test pour un classifieur naïf basé sur la classe majoritaire (qui prédit toujours *STAY*).

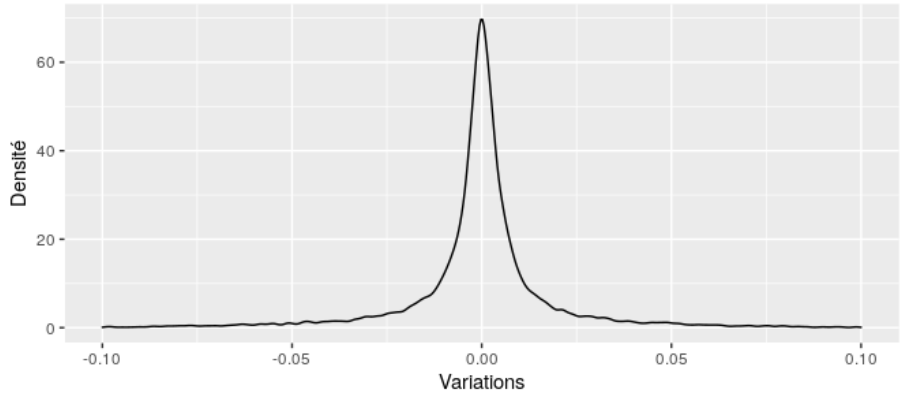


FIGURE 4.1 – Densité des fluctuations normalisées après parution des 8-K des entreprises du $S\&P500$ pour les années 2015 à 2017.

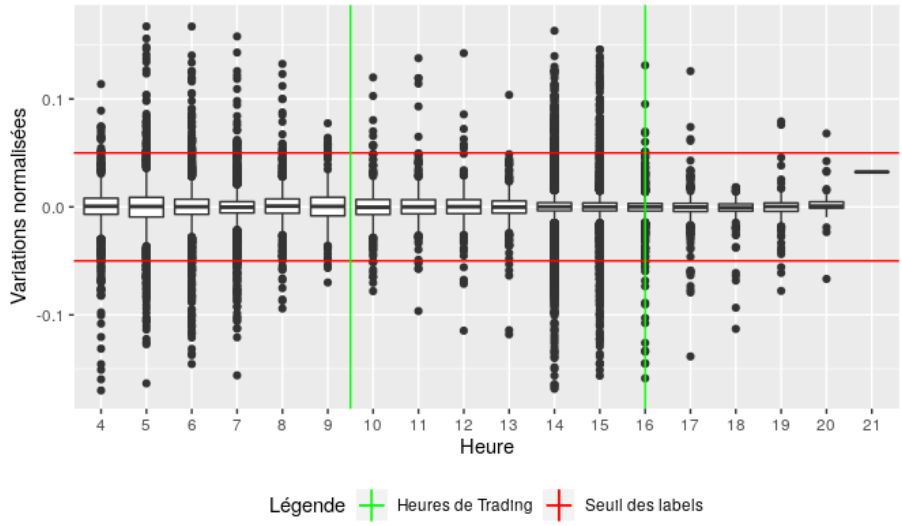


FIGURE 4.2 – Boxplot des fluctuations normalisées par heure avec les seuils de labellisation. $DOWN$ si inférieure à -0.5%, UP si supérieure à 0.5%, sinon $STAY$.

TABLE 4.1 – Distribution des classes.

Dataset	*DOWN*	*STAY*	*UP*
Entraînement	3754 (26%)	6910 (48%)	3755 (26%)
Test	1577 (22%)	4025 (55%)	1655 (23%)

La figure 4.3 montre la répartition des labels (*DOWN*, *STAY* et *UP*) selon les événements. Cette figure met en avant le fait qu'aucun événement n'est à l'origine d'une seule variation.

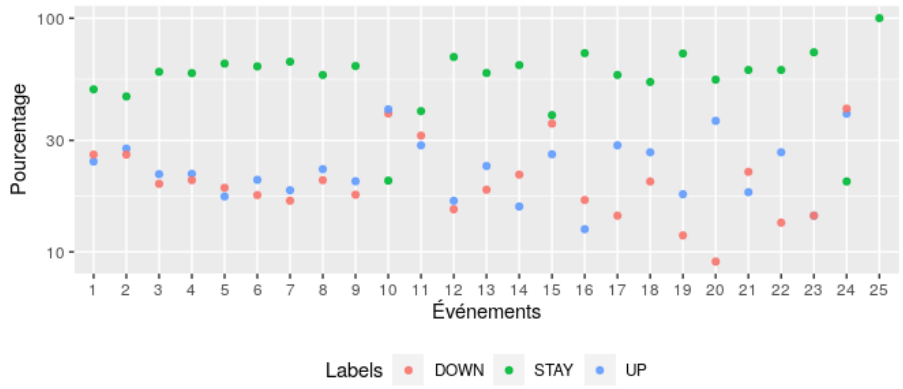


Figure 4.3 – Répartition en pourcentage des évènements selon les labels $DOWN$, $STAY$ et UP.

4.3 Approche proposée

Afin de prédire les variations engendrées par la publication des rapports 8-K, nous avons étudié les alternatives suivantes :

1. Représentation par unigramme avec utilisation de deux techniques de réduction de dimension :
   (a) Sélection par Information Mutuelle puis élimination récursive de variables ;
   (b) Sélection à l'aide de dictionnaires ;
2. Représentation par le plongement de mots GloVe :
   (a) Moyenne, Moyenne pondérée, Min Max …
   (b) Représentation par histogramme des centroïdes ;
   (c) Résumé extractif [45] et représentation par la moyenne des textes ;
   (d) Distance *Earth Mover's* [26, 69] ;
   (e) Représentation vectorielle par phrase.

4.4 Traitement des sacs de vecteurs

Nous présenterons en détails dans cette section les méthodes de traitement des sacs de vecteurs qui seront utilisées dans la **section 4.5**. Dans ce qui suit, un texte est représenté par une séquence de n vecteurs $\boldsymbol{w_1}, \ldots, \boldsymbol{w_n} \in \mathbb{R}^d$.

4.4.1 Agrégation et statistiques

Plusieurs méthodes sont à notre disposition [37] pour agréger ces sacs de vecteurs.

Une première manière de les agréger est d'utiliser :

1. la moyenne ($\bar{\boldsymbol{w}} \in \mathbb{R}^d$) ;
2. la concaténation du minimum et du maximum ($\bar{\boldsymbol{w}} \in \mathbb{R}^{2d}$) ;
3. la concaténation de la moyenne, du minimum et du maximum ($\bar{\boldsymbol{w}} \in \mathbb{R}^{3d}$) [96].

Pour cela, nous prenons, pour chaque dimension $j \in \{1, \ldots, d\}$, la moyenne généralisée. Ainsi pour un texte représenté par une séquence $\boldsymbol{w_1}, \ldots, \boldsymbol{w_n} \in \mathbb{R}^d$ et en reprenant la notation de [52, 98] avec $p \in \bar{\mathbb{Z}}$, la moyenne généralisée a pour formule :

$$\bar{w_j} = \left(\frac{w_{1j}^p + \ldots + w_{nj}^p}{n} \right)^{\frac{1}{p}}, \forall j \in \{1, \ldots, d\} \tag{4.1}$$

avec,

— $p = 1$, pour la moyenne arithmétique ;
— $p = -\infty$, pour le minimum ;
— $p = +\infty$, pour le maximum.

Une deuxième manière est d'utiliser la moyenne pondérée selon :

1. l'information mutuelle ;
2. la probabilité d'apparition du mot[5].

Pour le calcul de ces moyennes, la nouvelle valeur de la représentation du mot doit être calculée de cette manière :

$$\boldsymbol{w_i'} = \boldsymbol{w_i} \times \frac{a}{a + B} \tag{4.2}$$

avec,

— $\boldsymbol{w_i'}$, la nouvelle représentation du mot i ;
— $\boldsymbol{w_i}$, la représentation du mot i ;
— $a = 10^{-3}$ obtenu empiriquement [5] ;
— B est égal respectivement à la probabilité d'apparition ou l'information mutuelle du mot qui ont été calculées précédemment.

4.4.2 Histogrammes

Après avoir représenté chaque mot d'un corpus de textes sous forme d'un vecteur à l'aide d'un plongement de mots, nous appliquons un algorithme des k-moyennes aux vecteurs afin d'obtenir un ensemble de classes de taille fixe k.

Le centroïde de chaque classe peut être interprété comme un *champ lexical* qui englobe tous les vecteurs de mots sémantiquement liés dans une certaine région de l'espace du plongement de mots [27]. Chaque vecteur de mots du corpus est ensuite assigné au centroïde de la classe la plus proche. Ainsi, chaque texte est représenté comme un sac de *champs lexicaux* en calculant l'occurrence de chaque *champ lexical* présent dans le texte en question.

Cela revient à appliquer la méthodologie suivante :

1. représentation vectorielle à l'aide du plongement de mots choisi des mots du corpus (44362) ;
2. utilisation des k-moyennes pour déterminer les centroïdes de chaque classe (*champ lexical*) ;
3. les mots sont assignés au centroïde le plus proche ;
4. calcul de l'occurrence de chaque *champ lexical* dans un texte ;
5. le vecteur $\boldsymbol{w_h} \in \mathbb{R}^k$ représente le texte.

4.4.3 Transport optimal

Nous aborderons dans cette section la voie du transport optimal [113]. Pour ce faire, notre intérêt s'est porté sur la distance Earth mover's (ou distance de Wasserstein) et plus particulièrement à celle développée par M. Cuturi [36] qui considère le problème du transport optimal d'un point de vue de l'entropie maximale. En ajoutant un terme de régularisation entropique à ce problème, Cuturi *et al.* vont lisser le problème et imposer une structure plus simple au transport optimal régularisé et ainsi réduire les temps de calcul de manière considérable, passant d'une complexité de $O(m^3 \log(m))$ à $O(m^2)$. Pour plus de détails, nous invitons le lecteur à la **section 2.2.4** (p. 15).

Toujours dans des considérations de temps de calcul, nous avons appliqué la méthodologie suivante :

1. les textes sont représentés sous forme de sac de mots de dimension 1500 (cf. **section 4.5.1.1** pour plus de précisions) ;
2. sélection de 10% des rapports d'entraînement (1441), en gardant la distribution des labels des données d'origine ;
3. la distance euclidienne entre les différentes représentations normalisées du plongement de mots choisi sert comme matrice de poids ;
4. calcul de la distance Sinkhorn entre les textes du corpus (d'entraînement et de test) et les textes sélectionnés ;
5. un texte est représenté par un vecteur $\boldsymbol{w_{TO}} \in \mathbb{R}^{\mathbf{1441}}$.

La matrice résultante est de dimension 14419×1441 pour les données d'entraînement et de 7257×1441 pour celles de test. Ainsi, chaque texte est représenté sous forme d'histogramme normalisé, dont les index sont les textes sélectionnés et les valeurs, la distance de chacun de ces textes en fonction de la distance Sinkhorn.

4.5 Résultats

Dans cette section, nous présenterons les résultats obtenus avec les différentes stratégies proposées. Nous commencerons par le modèle qui sert de référence, puis nous passerons à des modèles plus complexes utilisant les plongements de mots, les résumés extractifs ou bien le transport optimal. La plupart des résultats ont été obtenus en s'appuyant sur des forêts aléatoires et sont rassemblés dans la **section 4.5.1**. Les résultats obtenus avec un réseau de neurones sont présentés dans la **section 4.5.2**.

4.5.1 Résultats obtenus avec des forêts aléatoires

Pour toutes les expériences rapportées ici, une forêt aléatoire [24] composée de 2000 arbres a été utilisée pour tester ces modèles avec une optimisation du taux de sélection, c'est-à-dire le nombre de variables testées à chaque division. Nous effectuons cette optimisation par *Out of Bag*. Rappelons qu'une forêt aléatoire travaille sur des données vectorielles classiques, ainsi chaque texte doit être représenté par un vecteur de taille identique.

4.5.1.1 Unigrammes

Le pré-traitement des textes décrit dans la **section 3.4.1** (p. 24) a été utilisé pour obtenir 44362 unigrammes. Pour établir notre modèle de référence, nous sélectionnons successivement ces unigrammes par information mutuelle et élimination récursive. Nous comparons le résultat de cette sélection à celui d'une sélection opérée par dictionnaire.

Information Mutuelle Nous avons estimé l'information mutuelle entre chaque unigramme et la variable cible sur l'ensemble d'apprentissage. Les résultats sont visibles dans la table 4.2.

TABLE 4.2 – Résultats de l'Information Mutuelle par Quantile.

Quantile	Information Mutuelle
10%	0.002293047
30%	0.004097382
50%	0.006142110
70%	0.009102122
90%	0.038102557

Compromis entre quantité d'information et temps de calcul, le seuil du quantile à 90% de l'information mutuelle a été choisi pour ne sélectionner que 4395 unigrammes.

Élimination récursive de variables Suite à cette première sélection, la méthode de l'élimination récursive de variables [73] , *Recursive Feature Elimination* (RFE) en anglais, en utilisant une forêt aléatoire a été appliquée sur l'ensemble d'apprentissage.

Pour ce faire, nous avons utilisé l'implémentation proposée par la bibliothèque *caret* et sa fonction *rfe* avec validation croisée (10 *folds*) permettant de ne sélectionner que 1500 unigrammes. Puis, nous entraînons une forêt aléatoire pour obtenir le résultat présenté dans la table 4.3.

Table 4.3 – Résultat avec les unigrammes sélectionnés par information mutuelle (MI) et élimination récursive de variables (RFE) - modèle de référence.

Modèle	Précision en %
MI et RFE	59.10

Sélection des unigrammes à l'aide de dictionnaires Afin d'améliorer les résultats obtenus avec les unigrammes, nous les avons sélectionné à l'aide de dictionnaires, à l'image de ce qui se fait pour l'analyse des textes légaux ou des brevets [61]. Nous avons utilisé les dictionnaires *Financial*, *Accounting* et *Legal* du projet *LexPredict*[1] et conservé les termes du corpus qui y sont représentés. Sur les 44362 termes, seuls 1346 étaient présents. Puis, nous avons entraîné une forêt aléatoire pour obtenir une précision de 58.5% qui est inférieure au résultat précédent.

4.5.1.2 Plongement de mots

Suite aux unigrammes, nous étudions l'apport en capacité de prédiction de l'utilisation de représentations plus sophistiqués des mots et des textes. Pour ce faire, les mots sont ici représentés sous forme vectorielle, autrement dit en utilisant les plongements de mots.

Dans cette partie, le plongement de mots GloVe [89] a été utilisé. Le dictionnaire pré-entraîné de cette méthode, dans notre cas le glove.6B.zip[2] de dimension 300, est composé de 400000 mots et couvre 54% des mots présents dans tous les textes de notre corpus. La figure 4.4 offre une vue du pourcentage pondéré des mots présents dans GloVe pour chaque texte. Ajoutons que les termes manquants sont le plus souvent des noms propres ayant une faible occurrence dans le corpus, comme *Wintersburg* (ville d'Arizona), *Trayport* (entreprise de *trading* d'énergie) ou bien *Parmeswar* (Vice Président de Johnson & Johnson) avec une occurrence respective de 22, 26 et 17 dans le corpus.

Agrégation de représentation vectorielle par texte Suite à l'utilisation du plongement des mots GloVe, les textes sont représentés par des sacs de vecteurs qu'il est

1. https://github.com/LexPredict/lexpredict-legal-dictionary/tree/master/en
2. https://nlp.stanford.edu/projects/glove/

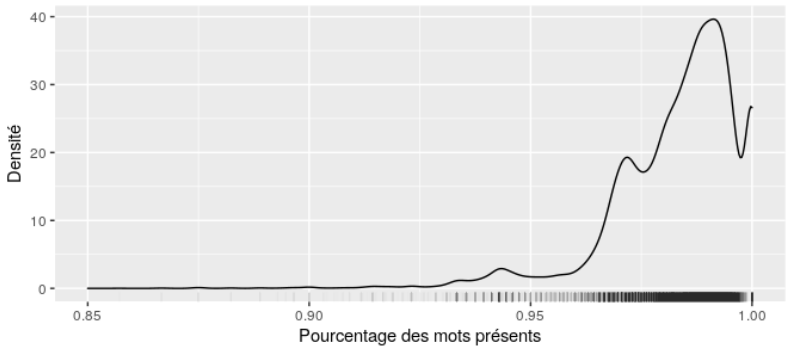


FIGURE 4.4 – La distribution du pourcentage pondéré de mots d'un 8-K présents dans GloVe.

nécessaire d'agréger pour pouvoir travailler avec une forêt aléatoire. Pour ce faire, nous utilisons les méthodes présentées dans la **section 4.4.1**.

La table 4.4 présente quelques-uns des meilleurs résultats. Il est visible que ces différentes représentations n'améliorent que légèrement la précision, au mieux de 0.38% par rapport à notre modèle de référence.

TABLE 4.4 – Les résultats obtenus avec le plongement de mots GloVe.

Modèles	Précision en %
Moyenne	59.38
Moyenne Pondérée par MI	59.23
Min Max	58.81
Moyenne et Min Max concaténés	59.48

Entraînement du dictionnaire GloVe Dans la continuité des recherches, nous avons entraîné le dictionnaire GloVe sur notre corpus selon plusieurs dimensions. Pour ce faire, nous avons utilisé le corpus d'entraînement comme corpus d'apprentissage pour calculer la matrice de co-occurrence. Suite à cela, nous avons appliqué l'algorithme GloVe avec les paramètres présentés dans la table 4.5.

TABLE 4.5 – Paramètres de l'algorithme GloVe.

Fenêtre	Dimensions
5	50 - 250

En représentant chaque texte par la moyenne comme précédemment, nous avons obtenu les résultats présentés dans la table 4.6. La représentation en dimension 100 améliore un peu les résultats précédents de plus de 0.5% au regard de la représentation de référence.

Table 4.6 – Les résultats obtenus avec le dictionnaire GloVe entraîné sur notre corpus.

Dimensions Glove	Précision en %
50	59.10
100	59.64
150	59.05
200	59.04
250	57.17

Nous conservons maintenant la représentation en dimension 100 et les termes sélectionnés dans la **section 4.5.1.1**. Nous représentons les textes par la moyenne et le résultat obtenu n'améliore pas celui de référence, comme indiqué dans la table 4.7.

Table 4.7 – Résultats en conservant la représentation entraînée de dimension 100 et les termes sélectionnés dans la **section 4.5.1.1**.

Nb. termes	Dimension	Précision
1500	100	58.94%

Ajout d'une marque de négation À l'image de celle utilisée dans [72], nous avons ajouté une marque de négation.

Deux méthodes furent testées pour la négation :

1. après chaque marque de négation jusqu'à la prochaine ponctuation, nous avons ajouté pour chaque mot la représentation du mot "no" ;
2. avec la même méthodologie, nous avons ici simplement multiplié la représentation du mot par -1.

Les deux méthodes ont obtenu respectivement une précision de 58.32% et 58.51%.

Représentation sous forme d'histogramme Suivant la méthodologie présentée dans la **section 4.4.2**, plusieurs valeurs de centroïdes ont été testées dont les résultats se trouvent dans la table 4.8.

Le meilleur résultat est ici obtenu pour 175 centroïdes avec 58.99% qui est inférieur à notre modèle de référence.

TABLE 4.8 – Résultats obtenus avec les histogrammes.

Centroïdes	Précision en %
125	58.47
135	58.83
145	58.67
155	58.63
165	58.84
175	58.99
185	58.33

4.5.1.3 Résumé extractif

Il existe plusieurs méthodes de résumé dont les méthodes abstractive et extractive [3]. La première méthode a pour but de reformuler l'information essentielle d'une nouvelle manière alors que la deuxième, identifie les parties les plus importantes et les agrège. La deuxième étant pour le moment la plus utilisée, nous l'avons mise en place à travers deux méthodes. La première, LexRank [43] avec la bibliothèque lexRankr [3] et la deuxième en suivant l'article de [20], basé sur une étude de la similarité cosinus et de la morphologie des phrases pour sélectionner les phrases les plus importantes.

Pour la première méthode, nous appliquons l'algorithme LexRank en faisant varier le nombre de phrases de 5, dont un exemple [4] est visible dans la figure 4.5, jusqu'à 20. Puis, nous utilisons la représentation moyenne du plongement des mots GloVe pour chaque texte et nous avons noté que plus l'algorithme conservait les phrases plus la prédiction était précise, commençant à environ 52% pour aller jusqu'à 56%.

La deuxième méthode est basée quant à elle sur la similarité cosinus et morphologique. Il s'agit d'une approche basée sur les graphes qui, en plus d'utiliser la mesure de similarité du cosinus [81, 43] ajoute une mesure de similarité morphologique calculée à l'aide d'une version modifiée de la mesure de la plus longue chaîne commune (LCS - *Longest Common Substring*). Le facteur d'amortissement a été fixé à 0.9, comme dans [20], et nous avions un paramètre supplémentaire fixé à 0.45, obtenu de manière empirique, pour ne garder que les phrases les plus liées. Dans ce cas, un texte est représenté en moyenne par 8.35 phrases. Enfin, nous avons représenté les textes avec la moyenne des représentations GloVe pour obtenir une prédiction de 55.3%.

3. https://github.com/AdamSpannbauer/lexRankr/
4. Le rapport pris en exemple est composé de 21 phrases.

Evénements : *Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers | Financial Statements and Exhibits*;

Texte : "bank nyse usb large bank unite state announce today andrew cecere currently vice chairman chief financial officer cfo promote vice chairman chief operate officer coo will responsible. bank believe comprehensive executive development process ensure leadership team breadth depth experience require execute plan create value customer shareholder say davis. bank confident ability manage business effectively execute customer - focus growth strategy. bank s core strength continuity executive leadership establish carefully manage time - test process. bank create value customer shareholder develop leverage internal executive talent highlight tenure stability commitment preserve"

Figure 4.5 – Exemple de rapport 8-K résumé en 5 phrases de l'entreprise US Bancorp (USB) publié le 20 janvier 2015

4.5.1.4 Transport Optimal

La méthodologie présentée dans la **section 4.4.3** est appliquée en utilisant le plongement de mots GloVe. Tout comme dans l'article [106], nous appliquons une forêt aléatoire et le résultat obtenu est comparé aux précédents résultats dans la table 4.9. Cette table montre que cette méthodologie, malgré sa complexité, n'améliore pas la précision.

Table 4.9 – Résultats obtenus pour le transport optimal.

Modèles	Précision en %
Classification naïve	55.5
Unigram 1500	59.10
WE Mean	59.38
Sinkhorn RF	57.79

4.5.2 Résultats obtenus avec des réseaux de neurones de type *long short term memory*

Pour les expériences rapportées ici, et pour prendre en compte la séquentialité des données, un LSTM [55] a été utilisé avec les paramètres présentés dans la table 4.10.

Précédemment, un texte indépendamment de sa longueur était représenté par la même quantité d'informations. Ici, nous considérons un texte comme une série temporelle de phrases, c'est à dire qu'un texte n'est plus représenté par un seul et même vecteur, mais par autant de vecteurs qu'il contient de phrases.

TABLE 4.10 – Paramètres LSTM.

Couches cachées	*Dropout*	Perte	Optimisation
100	0.2	*categorical crossentropy*	*Adam* [65]

Pour mettre en place cette méthodologie, un simple calcul de la moyenne de la représentation GloVe pour chaque phrase de chaque texte a été effectué.

Chaque texte ayant un nombre de phrases différent, et dans l'idée que par l'organisation d'un texte même, une phrase était impactée par celle d'avant, l'utilisation d'un LSTM est le plus approprié pour prendre en compte la séquentialité.

En utilisant l'architecture présentée du LSTM, nous avons obtenu un résultat de $57,07\%$ de précision ce qui est inférieur à notre modèle de référence.

4.6 Discussion des résultats

Les résultats obtenus dans la section précédente étant décevants, en comparaison de la classification naïve, il est naturel de se demander si l'information contenue dans les textes est suffisante pour faire des prévisions correctes. Pour ce faire, nous nous pencherons sur trois points à savoir, la richesse du vocabulaire, les termes ayant la plus forte information mutuelle et l'étude des résultats à travers une comparaison des textes.

4.6.1 Richesse du vocabulaire

La diversité lexicale est la mesure du nombre de mots différents employés dans un texte ou dans une production orale. Pour calculer cette diversité, ou richesse, il existe de nombreuses méthodes [111] comme les indices de MASS, TTR pour *type-token ratio* et surtout l'indice de Guiraud [25]. Cet indice est intéressant, car il équilibre les inégalités liées à la longueur des textes et se calcule selon la formule :

$$IG = \frac{N}{\sqrt{T}} \tag{4.3}$$

avec,

— N, nombre total de mots différents ;

— T, nombre total de mots énoncés.

Notons que plus sa valeur est élevée, plus le vocabulaire utilisé par l'entreprise est riche.

La figure 4.6 permet de visualiser la distribution de cet indice pour les cinq entreprises avec le plus de rapports publiés sur la période, à savoir Wells Fargo (WFC), Labcorp (LH), Alliance Data (ADS), Citigroup (C) et JPMorgan Chase (JPM). Nous nous apercevons que

la richesse lexicale varie beaucoup selon les entreprises, Alliance Data a un vocabulaire restreint au contraire de Wells Fargo et Citigroup.

De plus, l'indice de Jaccard est disponible dans la table 4.11. Il compare la similarité entre deux ensembles, dans notre cas les mots utilisés par les entreprises, et met en lumière la différence de vocabulaire utilisé. Nous constatons dans la table 4.11 qu'au minimum 47% des termes sont uniques pour chaque entreprise.

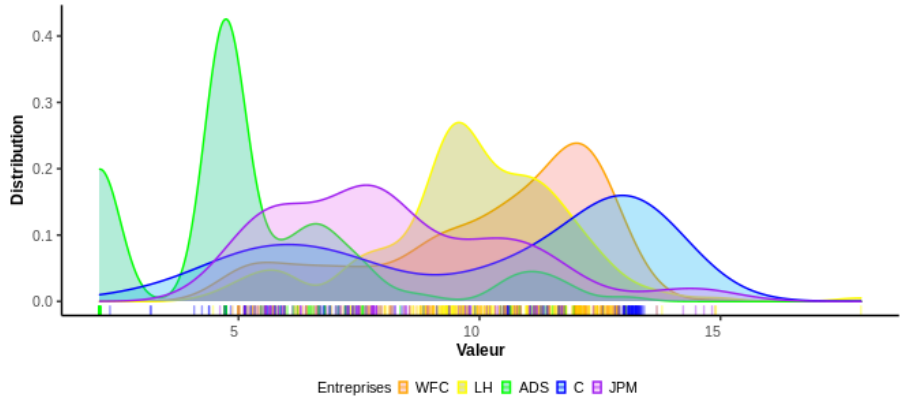


Figure 4.6 – Richesse lexicale pour les 5 entreprises ayant le plus publié.

	WFC	LH	ADS	C	JPM
WFC	1.00	0.42	0.46	0.53	0.46
LH	-	1.00	0.39	0.39	0.41
ADS	-	-	1.00	0.53	0.40
C	-	-	-	1.00	0.43
JPM	-	-	-	-	1.00

Table 4.11 – Indice de Jaccard entre le vocabulaire des entreprises.

4.6.2 Information mutuelle

Notre intérêt se porte maintenant sur l'information mutuelle pour ces cinq entreprises. Ainsi, dans la table 4.12, nous nous intéressons aux termes ayant la plus grande information mutuelle selon la variation observée sur les marchés suite à la publication d'un rapport. Cette table met en évidence la grande variabilité dans l'importance des mots pour chacune des entreprises, même pour des entreprises du même secteur d'activité, le secteur bancaire, comme Wells Fargo, Citigroup ou JPMorgan Chase.

WFC	LH	ADS	C	JPM
portfolio	brca	purchaser	award	defendant
gain	cancer	indenture	guarantor	plea
real	phase	initial	ytd	pleer
growth	coupon	offer	meet	nine
preference	zero	participant	committee	sib
offs	record	board	australian	growth
percent	compare	paragraph	six	stress
esop	tax	default	meaningful	six
card	increase	holder	hereunder	projection
yoy	earnings	group	noncontrolling	industry

TABLE 4.12 – Mots ayant la plus grande Information mutuelle selon la variation engendrée pour chacune des entreprises.

4.6.3 Étude des résultats

Analysons maintenant les résultats précédemment obtenus avec les algorithmes de classification. Pour ce faire, nous nous attachons à observer la similarité des rapports mal classés pour la représentation moyenne GloVe dont les résultats ont été présentés dans la **section 4.4.1.2**. La précision pour ce modèle est de 59.38% et la matrice de confusion est présentée dans la table 4.13 avec les résultats, qui subiront une analyse plus approfondie, notés en rouge.

	DOWN	STAY	UP
DOWN	365	291	337
STAY	807	3455	825
UP	402	286	489

TABLE 4.13 – Matrice de confusion pour la représentation GloVe moyenne.

4.6.3.1 Classé *DOWN* - réel *UP*

La figure 4.7 permet de visualiser les variations normalisées des rapports qui sont dans la classe *UP* mais qui ont été classés *DOWN*. La grande majorité de ces variations se situent entre 0.5% et 10%.

Prenons un rapport se situant dans cet intervalle avec le rapport de l'entreprise FedEx (FDX) publié le 19 décembre 2017 dont le texte [5] est :

> **Évènements :** Results of Operations and Financial Condition | Financial Statements and Exhibits

5. Les textes ou extraits de textes sont pré-traités.

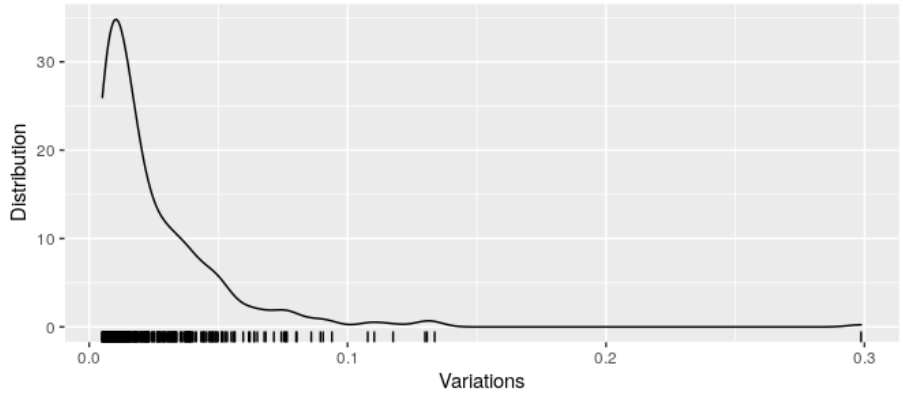


Figure 4.7 – Variations normalisées des rapports mal classés en *DOWN* à la place de *UP*

> **Texte :** follow exhibit furnish part report. exhibit numb description. press release fedex corporation date december. signature pursuant requirement security exchange act registrant duly cause report sign behalf undersigned hereunto duly authorize. fedex corporation date december s john l. merino john l. merino corporate vice president principal account officer.

En calculant le texte le plus proche au sens de la similarité cosinus, nous nous apercevons qu'il s'agit d'un autre rapport de l'entreprise FedEx publié le 20 décembre 2016, dont voici le texte :

> **Évènements :** Results of Operations and Financial Condition | Financial Statements and Exhibits
> **Texte :** follow exhibit furnish part report. exhibit numb description. press release fedex corporation date december. signature pursuant requirement security exchange act registrant duly cause report sign behalf undersigned hereunto duly authorize. fedex corporation date december s john l. merino john l. merino corporate vice president principal account officer.

Cependant, ce dernier texte a eu un impact négatif sur le cours de l'action au contraire du rapport mal classé. Ces deux rapports sont similaires mais, les communiqués de presse dont ils font référence et qui leur sont attachés [6] [7], ne sont pas disponibles dans leurs textes. Ces communiqués diffèrent dans les résultats et le contexte, ce qui explique la différence de variation. Notamment, rappelons que FedEx a subi une cyberattaque [8] en 2017 lui coûtant plus de 300 millions de dollars et qui a impacté la confiance des investisseurs.

6. Pour le premier : https://www.sec.gov/Archives/edgar/data/1048911/000119312516798888/d310707dex991.htm.
7. Pour le deuxième : https://www.sec.gov/Archives/edgar/data/1048911/000119312517373624/d450375dex991.htm.
8. https://www.bbc.com/news/technology-41336086.

4.6.3.2 Classé *UP* - réel *DOWN*

De la même manière que précédemment, la figure 4.8 permet de visualiser les variations normalisées des rapports qui sont dans la classe *DOWN* mais qui ont été classés *UP*. La grande majorité de ces variations se situent entre -0.5% et -10%.

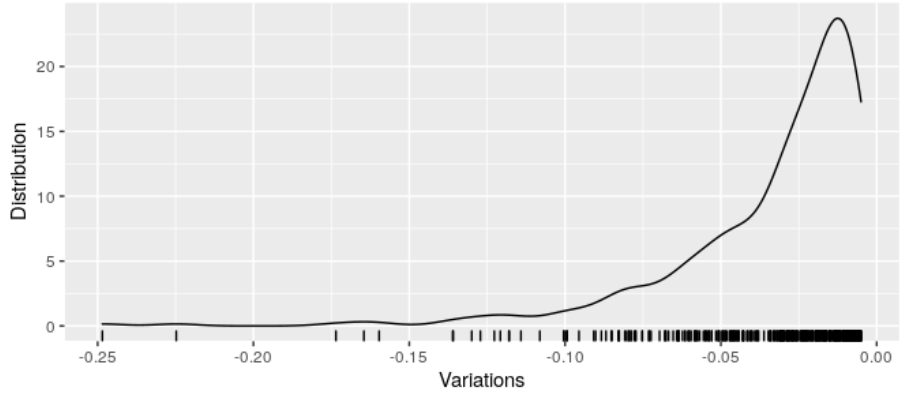


FIGURE 4.8 – Variations normalisées des rapports mal classés en *UP* à la place de *DOWN*

Un rapport se situant dans cet intervalle est celui de l'entreprise Nike (NKE) publié le 21 décembre 2017 dont voici un extrait [9] :

> **Évènements :** Results of Operations and Financial Condition | Financial Statements and Exhibits
> **Extrait du texte :** [...]quarter revenue growth drive international geography continue strength nike direct partly offset expect decline north america wholesale revenue. dilute earnings per share. percent compare period last year due primarily decline gross margin high sell administrative expense offset solid revenue growth low tax rate low average share count. quarter lead consumer direct offense accelerate international growth build underlie momentum [...]

Le texte le plus proche au sens de la similarité cosinus est ici aussi un texte de la même entreprise publié le 20 décembre 2020, dont voici un extrait :

> **Évènements :** Results of Operations and Financial Condition | Financial Statements and Exhibits
> **Extrait du texte :** [...] today report financial result fiscal quarter end november . global consumer demand drive revenue growth across nike brand portfolio. dilute earnings per share percent grow fast revenue primarily due sell administrative expense leverage low average share count. nike s ability attack opportunity consistently drive growth near long term set us [...]

Cependant, ce dernier rapport a eu un impact positif sur le cours de l'action au contraire de celui mal classifié.

9. Les textes ou extraits de textes sont pré-traités.

Ce constat de similarité, au sens de la similarité cosinus, entre les rapports peut être fait dans un cadre général. Un rapport mal classé sera, dans 90% des cas, le plus similaire à un autre rapport de la même entreprise quelle que soit la variation engendrée.

Les **sections 4.6.3.1** et **4.6.3.2** mettent donc en évidence la difficulté de créer un modèle unique de prédiction pour toutes les entreprises du $S\&P500$.

4.7 Conclusion

Suite aux différentes expérimentations menées, nous pouvons conclure que la complexification des représentations des textes n'améliore que très légèrement les résultats des prédictions. Seul l'apprentissage du dictionnaire GloVe nous a permis d'améliorer nos prédictions de plus de $0,5\%$ par rapport à notre modèle de base.

Comme nous l'avons vu dans la **section 4.6**, chaque entreprise utilise un lexique très spécifique qui ne permet pas de trouver de similitude entre des textes ayant impliqué une variation semblable, mais des textes d'une même entreprise. De plus, deux textes semblables d'une entreprise peuvent impliquer une variation différente ce qui est une piste pour expliquer les faibles améliorations obtenues dans ce chapitre.

Dans le chapitre suivant, comme les méthodes prédictives ne nous permettent pas d'extraire de l'information dans les 8-K, nous nous orientons vers des méthodes d'exploration. Pour ce faire, nous nous concentrons sur une seule entreprise, Wells Fargo, et nous étudions l'impact de différentes représentations de textes pour l'analyse exploratoire d'un corpus.

Chapitre 5

Comparaison de représentations de textes en vue d'une analyse exploratoire

Résumé : Nous étudions dans ce chapitre l'impact de différentes approches de représentations vectorielles de textes sur l'analyse exploratoire d'un corpus. Nous comparons une représentation élémentaire par sac de mots (unigrammes) à celle obtenue par *topic model* ainsi qu'à une plus complexe, construite à partir de la distance Sinkhorn entre les textes calculée sur une représentation vectorielle des mots. Nous construisons une classification des textes ainsi représentés à l'aide du modèle *high-dimensional data clustering*. Nous illustrons les différences entre les représentations grâce à un corpus de textes constitués à partir des rapports 8-K des entreprises du Standard & Poor's 500 (pour les années 2015 et 2016). Nous analysons la cohérence des classes ainsi obtenues et cherchons à les caractériser en termes de vocabulaire et de sujets spécifiques.

5.1 Introduction

Dans le chapitre précédent, nous avons montré qu'il était difficile d'extraire de l'information des rapports 8-K pour obtenir de bons résultats pour une classification supervisée. De plus, nous avons mis en avant qu'une représentation basique par sac de mots donnait un modèle de référence très difficile à améliorer.

Dans le présent chapitre , nous nous tournons vers l'analyse exploratoire afin de mieux comprendre les faibles résultats de prédiction, mais aussi pour révéler l'information qui se cache dans les rapports 8-K. À cette fin, nous étudions différentes représentations de texte. Nous comparons la représentation basique par sac de mots à deux solutions concurrentes : la première basée sur les *topic models* [17], la deuxième sur le transport optimal [113].

Pour ce faire, nous nous concentrons sur le jeu de données présenté dans la **section 3.4.4** (p. 29) concernant l'entreprise Wells Fargo (WFC) pour les années 2015 et 2016. Nous présentons les méthodes retenues dans la section 5.2, les différents résultats dans

la section 5.3 et la section 5.4 compare les résultats avec ceux obtenus précédemment.

5.2 Méthodes

Nous présentons brièvement dans cette partie les stratégies de représentations étudiées ainsi que le modèle de classification retenu.

5.2.1 Le modèle High-Dimensional Data Clustering

Les représentations vectorielles de texte ont généralement une grande dimension, ce qui dégrade fréquemment les performances des méthodes de classification. Nous avons donc retenu le modèle *High-dimensional data clustering (HDDC)* [23] présenté dans la **section 2.2.1** (p. 13). Il a été spécifiquement conçu pour lutter contre le fléau de la dimension. En outre, HDDC est entièrement automatique car il s'appuie sur le critère BIC [103] pour sélectionner automatiquement ses paramètres, notamment le nombre de classes.

5.2.2 Représentations des textes

Nous nous proposons de comparer plusieurs représentations vectorielles des textes.

Unigrammes Chaque texte est représenté par un vecteur indiquant les occurrences de chaque terme en son sein (cf. **section 2.1.1** p. 6 pour plus de détails).

Topic models Chaque texte est caractérisé par un vecteur θ (notation de [17]) qui donne les proportions de chaque *topic* en son sein. Nous pouvons donc représenter un texte par sa position θ dans le simplexe des probabilités sur les *topics*. Notons que la dimension de cette représentation est naturellement beaucoup plus faible que celle des unigrammes. Le critère de [39] permet de sélectionner le nombre optimal de *topics*, qui est ici de 35 (cf. **section 2.1.4** p. 11 pour plus de détails).

Transport optimal Chaque texte est vu comme une distribution discrète dans l'espace vectoriel de plongement des mots *GloVe* [89] : nous comparons deux textes au sens de la distance du transport optimal entre leur distributions [116, 107] à l'aide de la distance Sinkhorn [36]. Pour obtenir une représentation vectorielle, nous réalisons un plongement euclidien de la matrice des distances, en appliquant un *multidimensional scaling* métrique [110] (cf. **section 2.2.4** p. 15 pour plus de détails).

5.3 Expérimentations et résultats

5.3.1 Représentation par sac de mots (unigrammes)

L'utilisation d'HDDC sur les 248 textes de Wells Fargo (en dimension 3778) conduit à la sélection d'une solution à 3 classes, dont la table 5.1 résume les caractéristiques. Chaque classe possède sa propre matrice de rotation, mais les spectres des covariances sont tous de la forme $(a, \ldots, a, b, \ldots, b)$.

TABLE 5.1 – Caractéristiques des 3 classes HDDC pour la représentation unigramme.

Classe	1	2	3
Dimension intrinsèque	5	1	2
Nombre de rapports	51	49	148

La classification distingue les textes selon leur longueur comme nous pouvons le voir dans la figure 5.1. Ainsi, les textes de la classe 2 sont plus courts que ceux de la classe 3 mais de longueur homogène dans chacune de ces deux classes, au contraire de ceux de la classe 1.

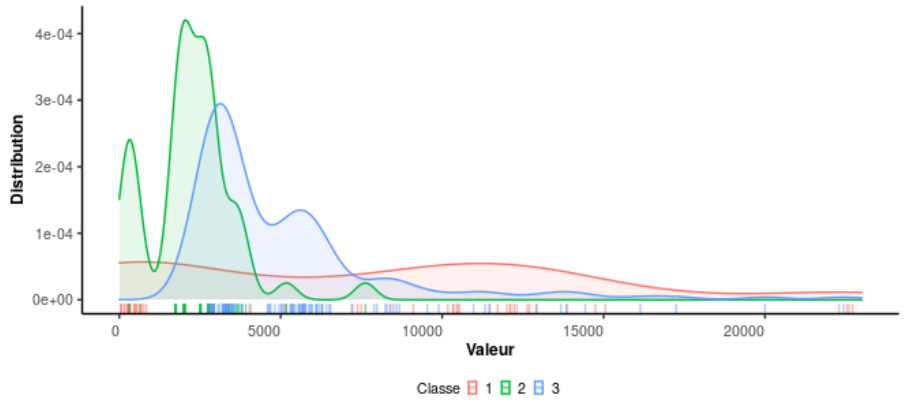


FIGURE 5.1 – Longueur des textes de Wells Fargo selon la classification HDDC pour la représentation unigramme.

La table 5.1 montre que le modèle HDDC permet ici une forte réduction de dimension, passant de 3778 à au plus 5 dimensions. Les figures 5.2, 5.3 et 5.4 montrent la distance euclidienne entre la représentation du texte et sa projection dans chaque dimension. Ainsi malgré la diminution importante des dimensions, ces figures permettent de constater la qualité et la richesse de la représentation dans chaque représentation. Elles montrent une bonne représentation dans leurs espaces de projection respectifs pour les classes 2 et 3 pour lesquelles chaque texte est proche de son projeté. Pour la classe 2, 95.9% des textes sont plus proches de leurs projetés que le premier d'une autre classe et pour le cluster 3, cela atteint 96.6%. À l'inverse, le cluster 1 est constitué de textes plus hétérogènes,

seuls 62.7% des textes sont plus proches que le premier d'une autre classe ce qui tend à indiquer que cette projection est discutable et aurait peut être nécessité la création d'une autre classe.

La table 5.2 montre le résultat du modèle HDDC pour 4 classes. Un Adjusted Rand Index (ARI) de 98% entre cette classification et celle à 3 classes est trouvé. Cela nous indique une forte similarité entre ces deux classifications. Nous constatons que les classes 2 et 3 sont identiques avec les mêmes nombres de rapports ainsi que les mêmes dimensions intrinsèques. La seule différence vient de la classe 1 de la classification à 3 classes qui est scindée ici en deux classes à savoir les classes 1 et 4. La classe 1 est de dimension intrinsèque 1 et contient 5 rapports. La classe 4 est de dimension intrinsèque 4 et contient 46 rapports. De plus, il apparaît pour la classe 1, que 100% des textes sont plus proches de leurs projetés que le premier d'une autre classe. A l'inverse, pour la classe 4, seuls 65.2% des textes sont plus proches que le premier d'une autre classe. Ainsi, la classification à quatre classes a vu l'apparition d'une classe spécialisée mais ne nous permet pas d'obtenir une meilleure représentation pour les autres textes de la classe 4. Nous conservons donc pour la suite de notre analyse la classification à trois classes.

Table 5.2 – Caractéristiques des 4 classes HDDC pour la représentation unigramme.

Classe	1	2	3	4
Dimension intrinsèque	1	1	2	4
Nombre de rapports	5	49	148	46

Par la suite, pour illustrer notre propos, deux types de représentations seront privilégiées notamment à travers l'analyse des classes 2 et 3. La première est basée sur les représentations générales issues du modèle HDDC, c'est à dire les projections des classes ainsi que des matrices de rotation. La deuxième sera quand à elle basée sur une approche plus fine d'analyse des mots impactant le plus la classification.

La figure 5.5 permet une représentation plus détaillée de la classe 2 de dimension une que la figure 5.3. Nous représentons la classe 2 par la distance euclidienne entre la représentation des textes et leurs projections selon cette classe. Nous constatons que les textes de cette classe sont très proches de leurs projetés mais, qu'en même temps, ils sont projetés de deux manières différentes amenant à la création de deux sous-groupes notés *1* et *2*.

La figure 5.6 étaye nos propos précédents en rappelant une ACP, dans le sens où seuls les textes projetés de la classe 3 sont décrits par cette projection. Sur cette figure apparaît un sous groupe à l'écart des autres textes de cette classe que nous notons *1*.

Pour étudier ces sous-groupes, nous nous inspirons de [34]. Ainsi, chaque dimension d'une classe est représentée par les cinq variables qui ont la plus grande valeur en valeur absolue de rotation. Chaque variable est différenciée selon sa fréquence et la valeur de la rotation qu'elle entraîne. Comme nous travaillons ici avec les unigrammes, cela nous permet d'analyser finement les termes impliqués dans chaque dimension de chaque classe.

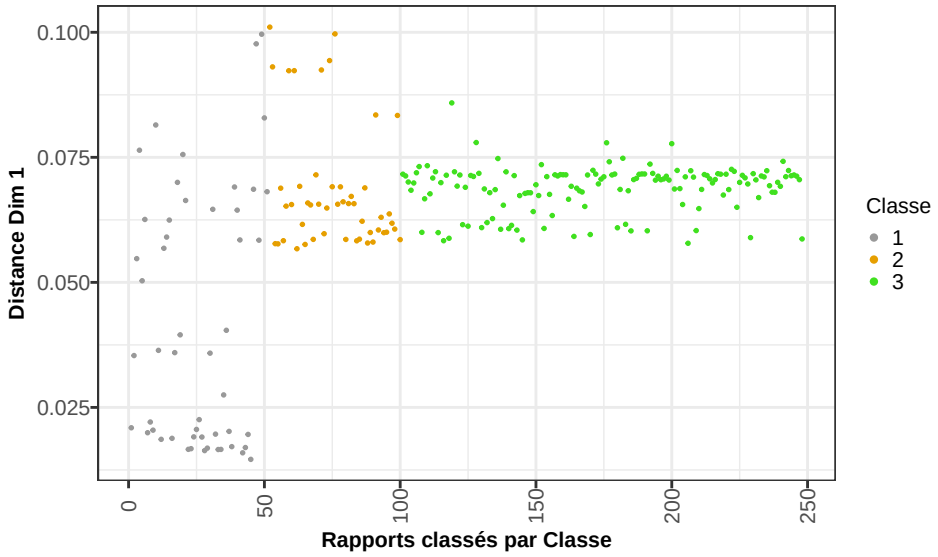


FIGURE 5.2 – Distance entre les représentations des textes et leurs projections pour la classe 1 HDDC pour la représentation unigramme.

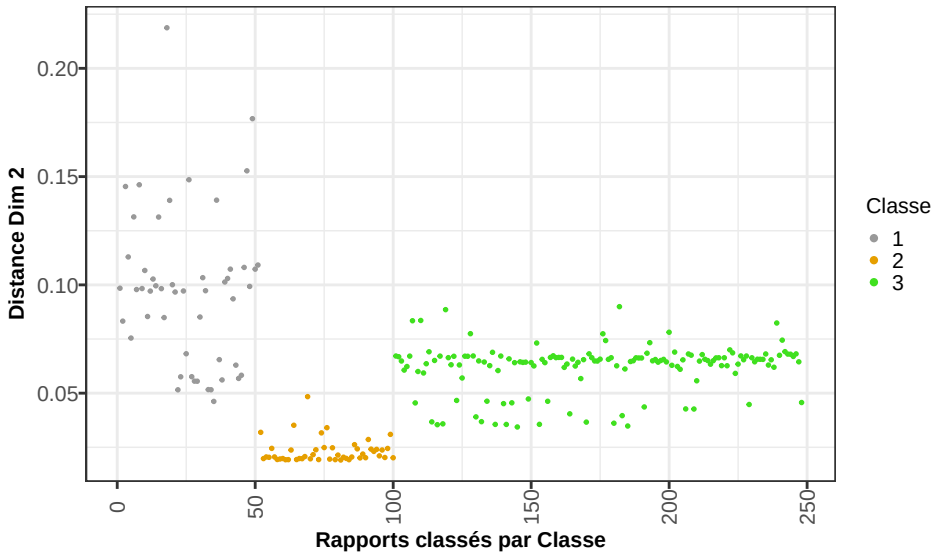


FIGURE 5.3 – Distance entre les représentations des textes et leurs projections pour la classe 2 HDDC pour la représentation unigramme.

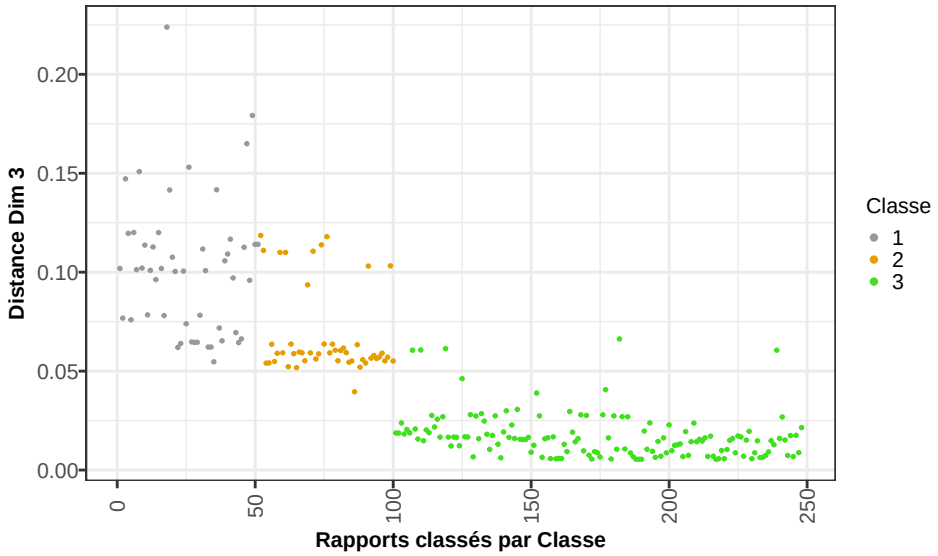


Figure 5.4 – Distance entre les représentations des textes et leurs projections pour la classe 3 HDDC pour la représentation unigramme.

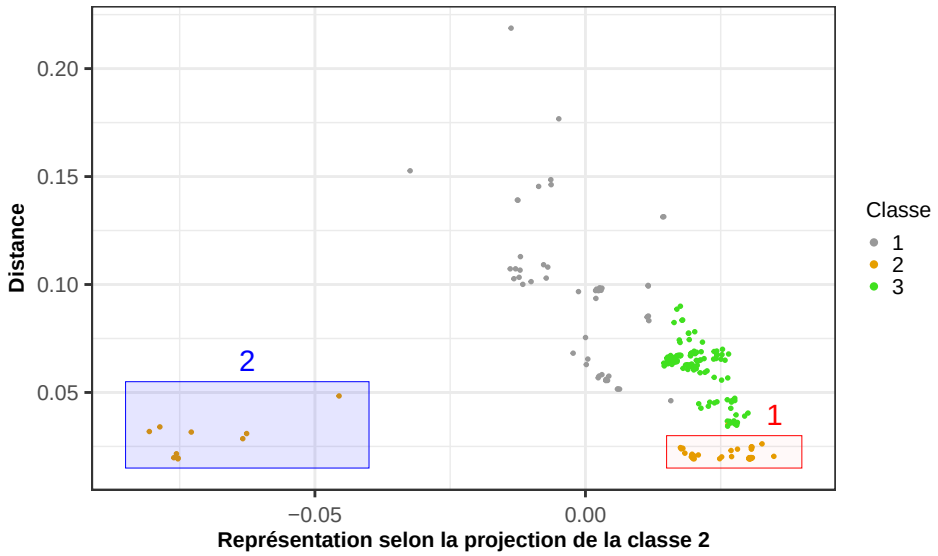


Figure 5.5 – Distance euclidienne par rapport à la projection selon la classe 2 HDDC pour la représentation unigramme avec les sous groupes *1* et *2*.

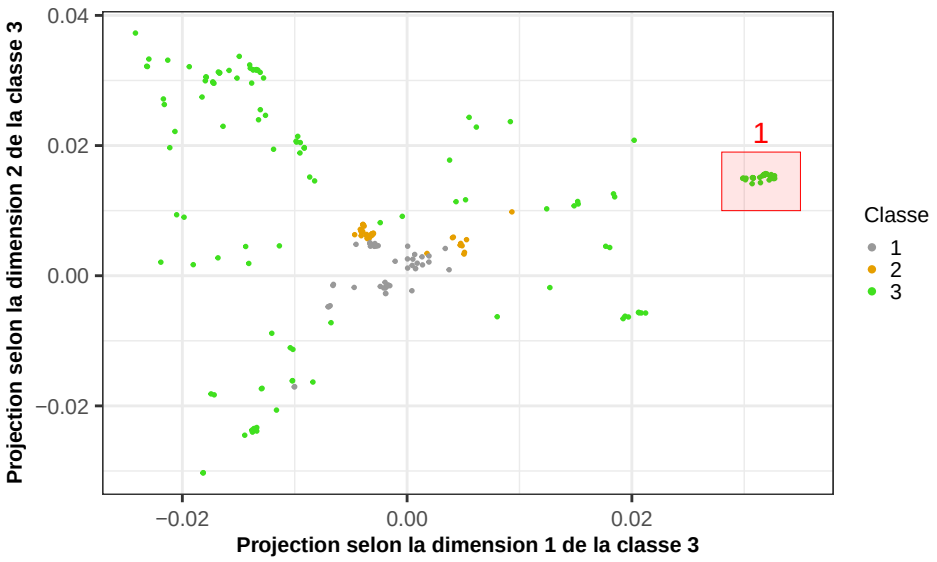


FIGURE 5.6 – Projection des textes selon la classe 3 HDDC pour la représentation unigramme avec un sous groupe noté *1*.

Sur la figure 5.7, nous constatons que la classe 2 est subdivisée par deux types de rotation. Les mots les plus fréquents comme *security* ont une rotation positive alors que des mots plus rares comme *note* en ont une négative. Cela explique les deux sous groupes *1* et *2* de la figure 5.5. Prenons l'exemple des textes du sous-groupe *2* qui sont projetés négativement. La figure 5.8 présente les termes qui entraînent les plus grandes rotations positives et négatives. Ainsi, nous constatons que les textes de ce sous-groupe sont fortement liés au cabinet d'avocats *Faegre Baker Daniels llp* pour l'émission de billets à terme. Voici un extrait [1] pré-traité de texte appartenant à ce sous-groupe qui confirme notre analyse :

> [...]signature pursuant requirement *security* exchange act registrant duly cause report sign behalf undersigned hereunto duly authorize well fargo barbara brett barbara brett senior vice president assistant treasurer form index *exhibit exhibit* description method *file opinion faegre baker daniels llp* electronic transmission consent [...]

Nous pouvons analyser les textes de la classe 3 de manière semblable. Grâce à la figure 5.7, nous pouvons constater que la dimension 1 de cette classe est décrite par les termes *index* et *underlier* alors que la dimension 2 est décrite par les termes *fund* et *index*.

Nous constatons sur la figure 5.7 que le terme *underlier* entraîne une projection positive sur cette dimension. Ainsi, l'occurrence de ce terme doit être forte dans les textes du sous-groupe *1* de la figure 5.6.

1. Le rapport complet est disponible à cette adresse : https://www.sec.gov/Archives/edgar/data/72971/000119312515029789/d861786d8k.htm.

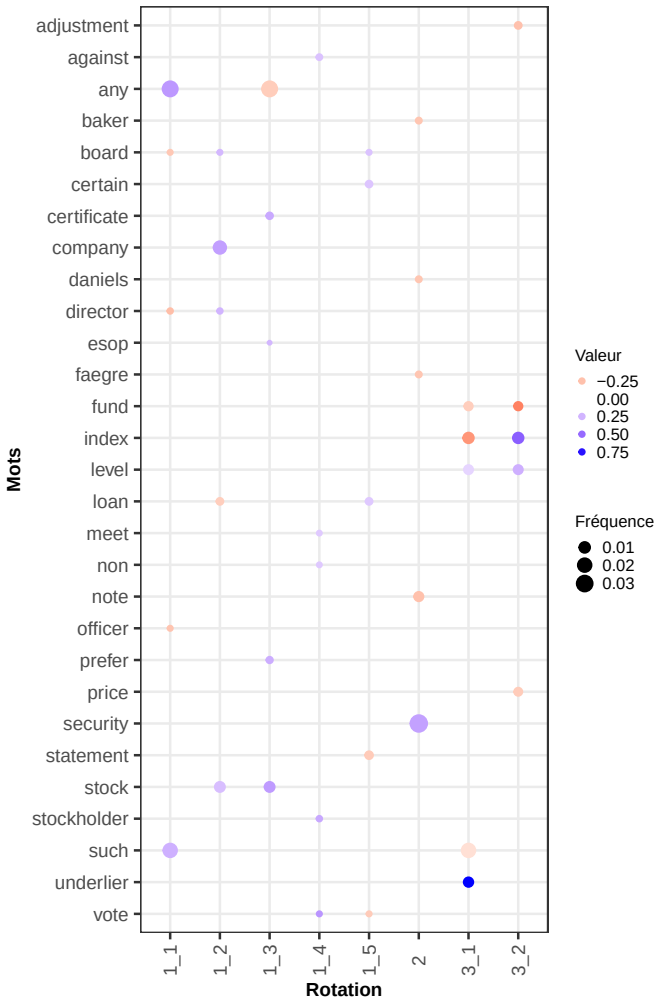


Figure 5.7 – Les 5 termes avec la plus grande rotation en valeur absolue par dimension de chaque classe HDDC pour la représentation unigramme. La couleur indique la valeur de la rotation. La taille du point exprime la fréquence du terme dans le corpus.

L'extrait de texte [2] suivant appartenant à ce sous-groupe confirme notre analyse :

> [...] relate future option exchange ii submission deadline order enter relate future option exchange system execution close *trade such* relate future option exchange relevant *underlier* sponsor fail publish *level underlier* any successor *underlier* result relevant *underlier* sponsor discontinue publication *underlier* successor *underlier* successor *underlier* available any relate future option exchange fail open *trade* [...]

2. Le rapport complet est disponible à cette adresse : https://www.sec.gov/Archives/edgar/data/72971/000119312516487253/d149760d424b2.htm.

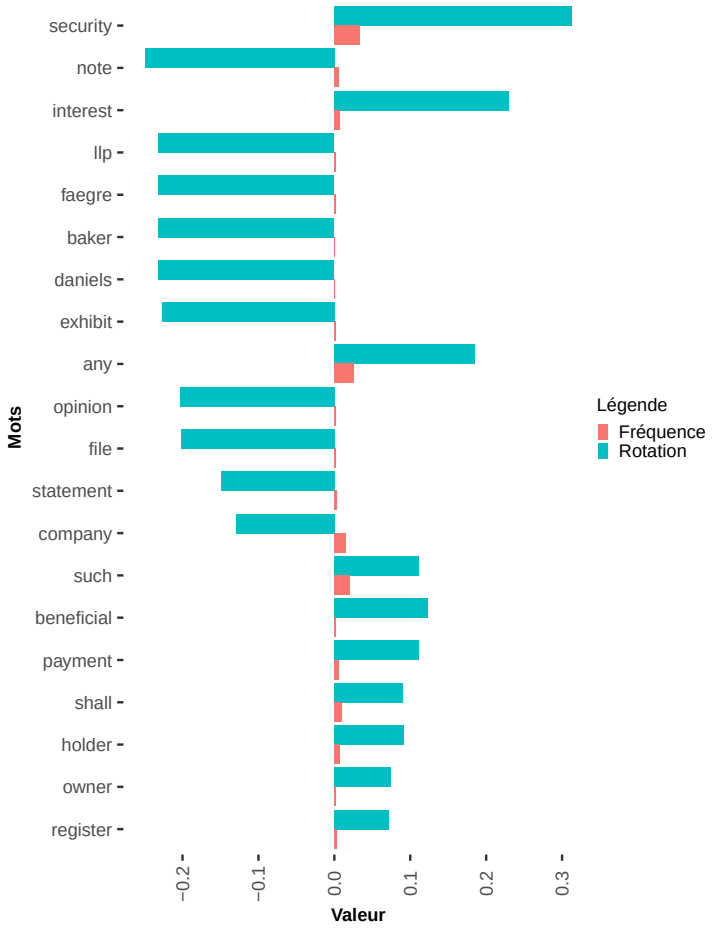


FIGURE 5.8 – Graphique en barres des mots ayant les plus grandes valeurs de rotation en valeur absolue pour la classe 2 HDDC pour la représentation unigramme. Les barres rouges expriment la fréquence des termes. Les barres bleues indiquent la valeur des rotations.

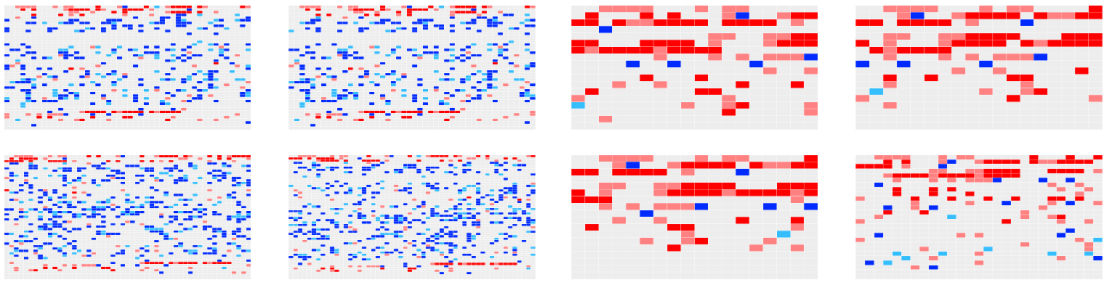
Pour une analyse plus fine des textes, nous nous inspirons de [31, 68]. Chaque texte est représenté sous forme d'image ayant pour pixel des mots, en *bleu* pour une rotation positive, en *rouge* si négative, la teinte variant selon la valeur de la rotation.

Nous appliquons cette méthodologie aux deux sous-groupes de la classe 2.

En premier lieu, les textes de ces deux sous-groupes apparaissent différents tant sur le fond que sur la forme. La figure 5.9a expose les textes du premier sous-groupe de la classe 2. Nous constatons que ces textes sont longs et décomposés en trois parties. Les parties d'introduction et de conclusion de chaque texte sont composées de mots d'orientation négative comme *opinion*, *statement* ou *faegre*. Il s'agit en effet d'une déclaration de la société d'avocats *Faegre Baker Daniels llp* concernant les produits financiers de Wells Fargo. Le corps du texte, d'autre part, expose toutes les dispositions de ces produits et est composé de mots d'orientation positive comme *interest*, *payment*, *security*.

La figure 5.9b montre les textes du deuxième groupe de la classe 2. Ils sont presque entièrement composés de mots d'orientation négative et sont plus courts que le premier groupe. Ceci explique la différence dans l'espace de projection de la classe 2.

En deuxième lieu, par une analyse attentive des textes, une grande similarité du vocabulaire apparaît entre ces deux groupes. Ils ont en commun 90% des termes. Un groupe introduit des produits à émettre, le deuxième groupe exprime des opinions sur les produits émis, justifiant ainsi la classification HDDC et leurs différences dans la projection.



(a) Aperçu des textes du groupe 1.

(b) Aperçu des textes du groupe 2.

Figure 5.9 – Aperçu des textes de la classe 2 HDDC pour la représentation unigramme.

5.3.2 Transport optimal

Nous appliquons le positionnement multidimensionnel, ou *multidimensional scaling* en anglais, sur les distances Sinkhorn qui aboutit à une représentation vectorielle des textes de dimension 171. Pour obtenir cette dimension, nous conservons seulement les vecteurs ayant une valeur propre associée strictement positive. Nous appliquons le modèle HDDC sur cette représentation pour retenir un modèle à 7 classes dont les caractéristiques sont présentées dans la table 5.3. L'indice de Rand [93] ajusté entre ce résultat et celui obtenu avec la représentation unigramme est calculé. Nous obtenons 31.5% qui nous signale une faible similarité entre ces deux résultats.

Table 5.3 – Caractéristiques des 7 classes HDDC pour la représentation issue du transport optimal. La dimension intrinsèque de chaque classe est de 5.

Classe	1	2	3	4	5	6	7
Nombre de rapports	38	36	29	42	42	41	20

La figure 5.11 est le graphique t-SNE de la représentation obtenue avec le positionnement multidimensionnel qui permet d'observer la bonne classification des textes. Cette même observation peut être faite avec la représentation des distances Sinkhorn sur la figure 5.10.

Sur cette dernière figure, les droites horizontales en pointillé permettent de séparer les classes. Les cinq premières classes sont homogènes entre elles ce qui n'est pas le cas

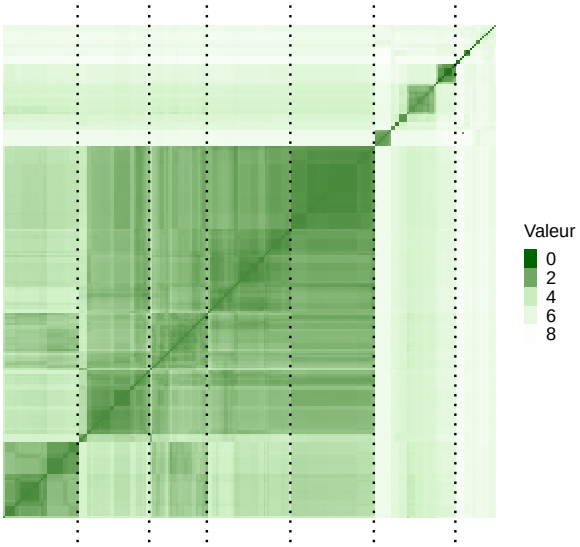


FIGURE 5.10 – Distance Sinkhorn regroupée selon les classes HDDC pour la représentation issue du transport optimal. Plus la valeur de la distance est faible, plus le vert est prononcé.

des deux dernières. En revenant aux textes, nous nous apercevons que les textes des cinq premières classes ont, dans plus de 95% des cas, l'événement *Financial Statements and Exhibits*. Au contraire des deux dernières classes qui sont plus hétérogènes comme nous pouvons le voir sur la figure 5.12. Dans celles-ci, nous retrouvons de nombreux autres évènements tels que *Amendments to Articles of Incorporation or Bylaws ; Change in Fiscal Year* , *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* ou *Submission of Matters to a Vote of Security Holders* qui semblent bien plus rares dans la vie d'une entreprise.

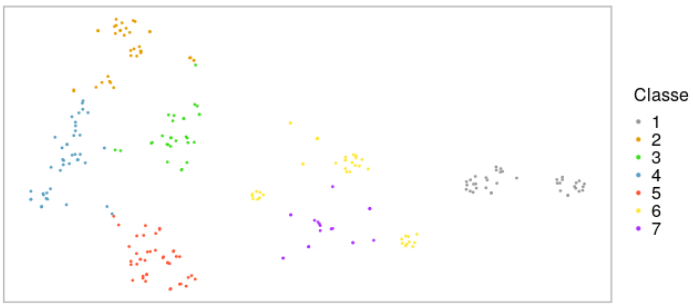


FIGURE 5.11 – t-SNE des classes HDDC pour la représentation *multidimensional scaling*.

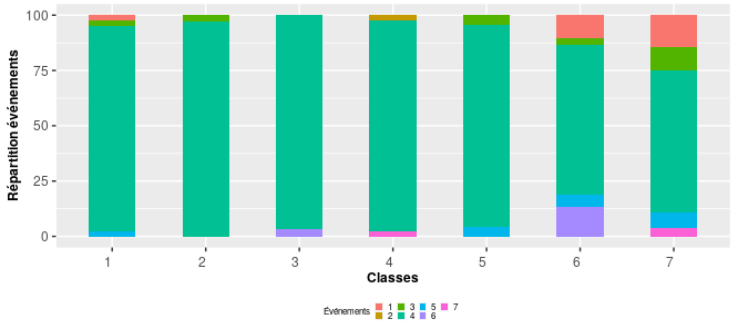


Figure 5.12 – Diagramme en bâtons des événements par classe HDDC pour la représentation basée sur le transport optimal.

Nous nous intéressons maintenant aux textes centraux de chaque classe. Pour ce faire, nous calculons, pour chaque classe, la moyenne des représentations obtenues avec le *multidimensional scaling*. Nous conservons seulement le texte dont la représentation a la plus faible distance euclidienne à la représentation moyenne de sa classe. La figure 5.13 présente la matrice de distance des textes centraux respectifs. Les classes 6 et 7 sont toujours représentées de manière bien distinctes ainsi que la classe 1 qui se démarque des autres.

Comparons le texte central 1 [3] et celui de la classe 3 [4] dont les extraits sont ci-dessous. Ces deux textes ont l'évènement *financial statements and exhibits* en commun mais diffèrent par les termes financiers employés notamment au niveau du sous-jacent auxquels les bons à moyen terme [5] sont adossés. Pour le premier, il s'agit du LIBOR [6] à 3 mois et pour le deuxième, du taux à 10 ans publié par la Réserve Fédérale américaine.

> Texte central 1 : item 9.01. financial statements and exhibits [...] linked to 3 month libor [...] indices exchange *trade*d funds securities [...] limits delays or *prohibit*s the making of payments [...]

> Texte central 3 : item 9.01. financial statements and exhibits [...] 10 year constant maturity swap rate [...] to establish the validity of the *beneficial ownership* in this security [...] successor equity index are *trade*d [...]

L'analyse du plan de transport entre ces deux textes permet de voir que les mots communs aux deux se transportent principalement sur eux-mêmes, dans plus de 96% des cas. Si nous regardons maintenant les mots qui ne sont présents que dans le texte 1, certains sont transportés sur des mots du texte 3 qui sont sémantiquement au plus

3. Le rapport complet est disponible à cette adresse https://www.sec.gov/Archives/edgar/data/72971/000119312516484926/d145535d8k.htm.

4. Le rapport complet est disponible à cette adresse https://www.sec.gov/Archives/edgar/data/72971/000119312515311961/d45070d8k.htm.

5. Medium Term Notes en anglais, qui sont des titres de créance.

6. London Interbank Offered Rate, taux d'intérêt moyen auquel les grandes banques internationales prêtent et empruntent de l'argent entre elles.

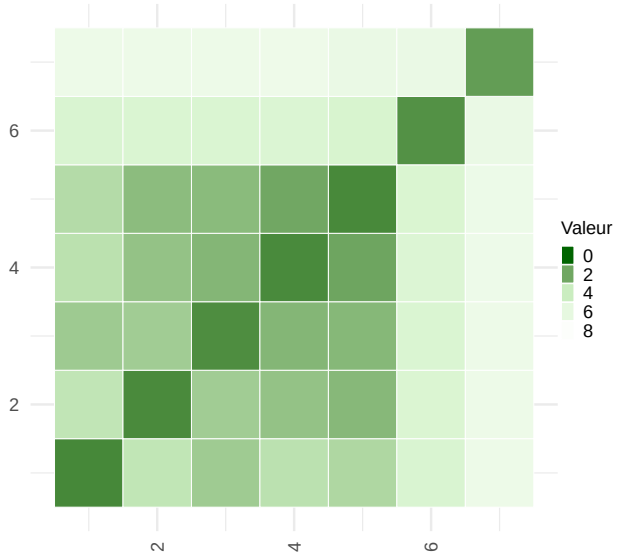


FIGURE 5.13 – Distance Sinkhorn entre les textes centraux de chaque classe. Plus la valeur de la distance est faible, plus le vert est prononcé.

proche dans la représentation GloVe. Prenons l'exemple des mots *february* et *four* qui sont transportés respectivement sur *september*, avec une similarité cosinus de 0.95 dans l'espace de représentation GloVe, et *three*, avec une similarité cosinus de 0.97.

L'analyse entropique du plan de transport permet une analyse plus fine encore des similitudes, mais aussi des différences lexicales entre ces deux textes. Une entropie forte sur ce plan de transport (entre les textes 1 et 3), permet d'affirmer qu'un mot n'est que peu ou pas présent dans le texte sur lequel il est projeté et où ne figurent pas de mots similaires, au sens de la représentation GloVe. Cela est visible dans les extraits précédents où le mot *prohibit* (texte 1), qui a une entropie de 8.16, est projeté sur les mots *beneficial* et *ownership* (texte 3). Au contraire, les mots comme *trade* ayant une entropie faible, dans ce cas une entropie de 0.08, sont présents dans les deux textes.

5.3.3 *Topic Models*

Une fois les *Topic models* obtenus, chaque texte est représenté par son vecteur θ de dimension 35. Nous appliquons la méthode HDDC qui retient une solution à 10 classes dont la table 5.4 résume les caractéristiques.

La figure 5.14 montre la représentation t-SNE selon les classes obtenues. Nous constatons que de nombreuses classes se retrouvent dispersées. De même, la figure 5.15, qui représente la distance euclidienne entre les vecteurs θ de chaque texte regroupés selon le regroupement obtenu, montre que les classes sont difficilement identifiables les unes des autres.

Table 5.4 – Caractéristiques des 10 classes HDDC pour la représentation issue des *Topic Models*.

Classe	1	2	3	4	5	6	7	8	9	10
Dimension intrinsèque	2	5	5	4	3	3	3	4	3	3
Nombre de rapports	43	20	15	15	24	28	15	12	42	34

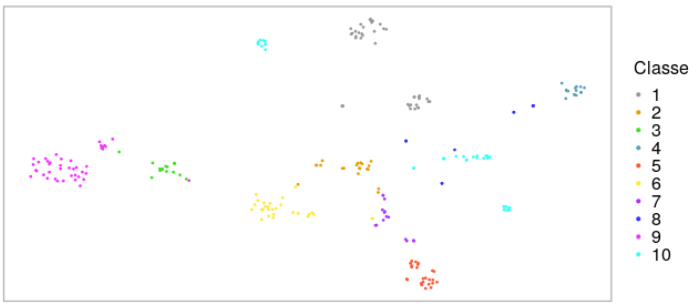


Figure 5.14 – t-SNE des classes HDDC pour la représentation *topic model* .

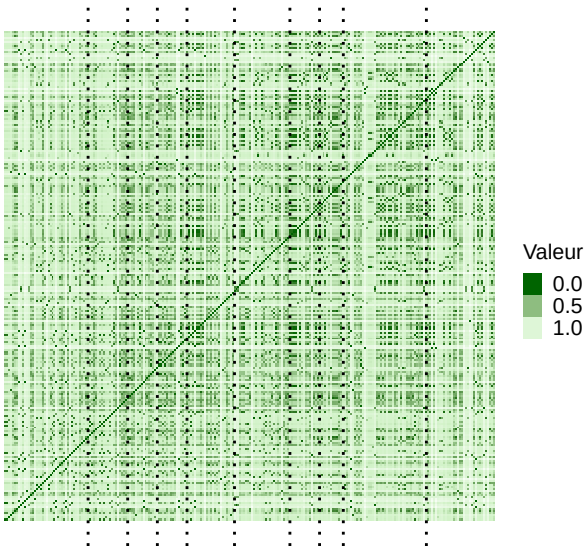


Figure 5.15 – Distance Euclidienne entre les vecteurs θ regroupés selon les classes HDDC pour la représentation *topic model*. Plus la valeur de la distance est faible, plus le vert est prononcé.

Nous analysons maintenant l'importance jouée par chaque *topic* dans la représentation des classes. Ainsi, la figure 5.16 présente les cinq *topics* ayant la plus grande rotation en valeur absolue par dimension de chaque classe. Plus la valeur de la rotation en valeur absolue est grande plus un *topic* est important pour la dimension de la classe en question. Nous observons que seuls les *topics* 14 et 25 sont uniques à la dimension d'une classe. Au contraire, de nombreux *topics* sont communs à de nombreuses dimensions comme les *topics* 2 et 24. De même, en prenant exemple sur les premières dimensions des classes 5 et 6, nous remarquons que plusieurs *topics* ont des comportements semblables au sein des dimensions de classes différentes, en l'occurrence les *topics* 22 et 24.

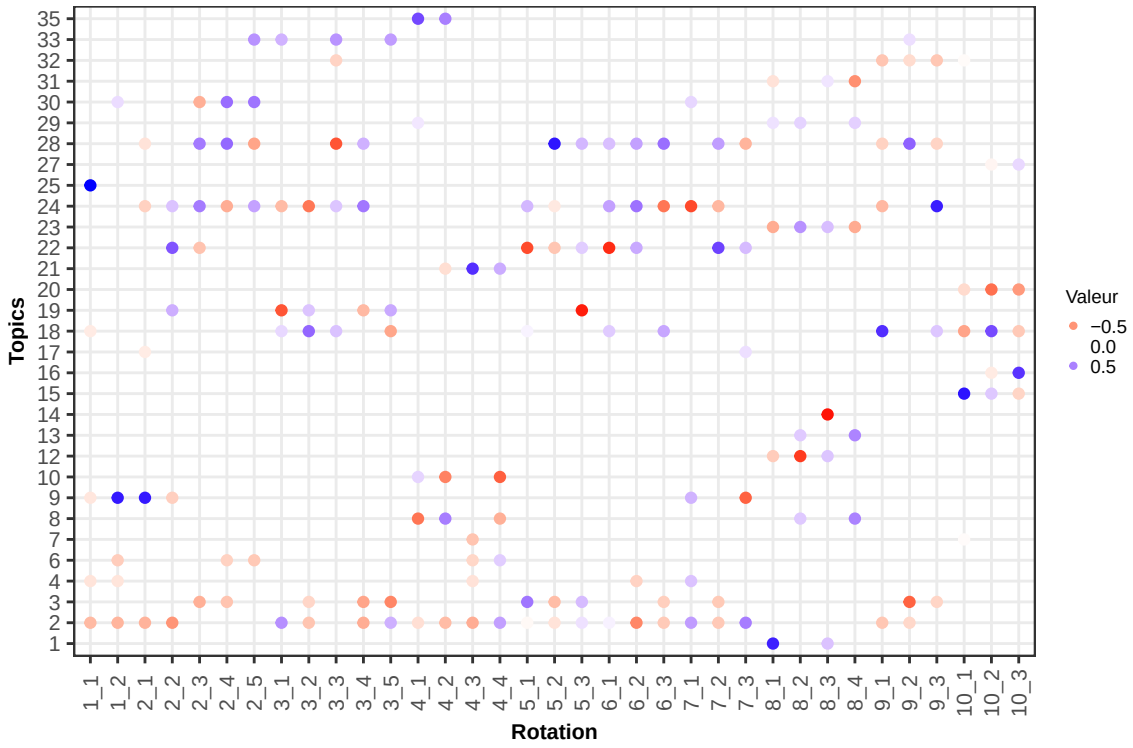


FIGURE 5.16 – Les 5 topics avec la plus grande rotation en valeur absolue par dimension de chaque classe. La couleur indique la valeur de la rotation.

L'analyse de la répartition des *topics* et leurs implications dans les dimensions de chaque classe nous fait constater que la réduction de dimension opérée par les *Topic Models* fait perdre en richesse la représentation des textes et ne permet pas une analyse fine de ceux-ci.

5.4 Comparaison avec les résultats obtenus précédemment

5.4.1 Comparaison avec les classifications hiérarchiques

La table 5.5 montre les Adjusted Rand Index (ARI) entre les regroupements ci-dessus et ceux du **chapitre 3** (p. 18). Elle met en avant le peu de similarité entre les regroupements des deux méthodes avec un maximum de 0.45 pour les deux regroupements unigrammes.

Table 5.5 – ARI entre les résultats obtenus avec HDDC et les classifications hiérarchiques.

Classe	Uni Chap 3	WE Chap 3	Uni	TO	LDA
Uni Chap 3	1	0.18	0.45	0.16	0.12
WE Chap 3	-	1	0.12	0.01	0.01
Uni	-	-	1	0.31	0.23
TO	-	-	-	1	0.68
LDA	-	-	-	-	1

Ces deux regroupements unigrammes contiennent une classe majoritaire ayant peu ou prou le même nombre de rapports, cf. les tables 3.3 (p. 33) et 5.1. La différence vient des deux autres classes qui étaient auparavant déséquilibrées et qui sont maintenant d'une taille homogène, passant respectivement de 95 et 9 à 51 et 49.

La composition des classes diffère aussi au niveau des événements, comme nous pouvons le voir dans la figure 5.17 (versus figure 3.15 (p. 34)). Notamment, nous observons que l'événement *Financial Statements and Exhibits* est majoritaire dans les trois classes. Nous constatons toutefois que la classe 1 a une composition plus hétérogène. Les autres événements y représentent plus de 40% des événements.

La figure 5.18 montre la répartition des classes selon les mois. Il est intéressant de noter que la classe 2 est sur-représentée entre les mois de février à juillet 2015. Cela s'explique avec la présence des événements *Other Events* et *Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year*, en plus de l'événement majoritaire *Financial Statements and Exhibits*, qui correspondent aux premiers ennuis judiciaires de l'entreprise sur cette période à savoir la violation des lois new-yorkaises sur les cartes de crédit en février 2015 (*Violation of New York credit card laws*[7]) ainsi qu'une affaire de délit d'initié (*SEC settlement for insider trading case*[8]). Au contraire, la classe 1 devient prépondérante à partir de janvier 2016 avec en plus de l'événement *Financial Statements*

7. https://www.reuters.com/article/us-wells-credit-settlement-idUSKBN0L92C720150205.

8. https://www.reuters.com/article/usa-insidertrading-wellsfargo-idUSL1N11K1MU20150914.

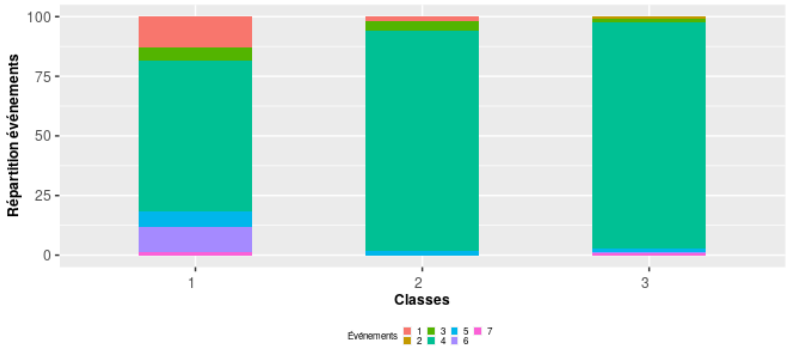


FIGURE 5.17 – Répartition des évènements selon les classes HDDC pour la représentation unigramme.

and Exhibits de nombreux événements *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* et *Results of Operations and Financial Condition* qui correspondent aux règlements d'une partie des poursuites judiciaires, dont notamment le règlement de 1,2 milliards de dollars pour l'affaire *Lawsuit by FHA over loan underwriting*[9], le départ du CEO John Stumpf et la mise en place d'une nouvelle équipe de direction.

Rappelons que l'événement *Results of Operations and Financial Condition* permet à l'entreprise d'expliquer le contexte de ses états financiers et donc d'expliquer les événements qui ont grevé ses comptes, comme le paiement des amendes. Cette analyse n'était pas visible dans la figure 3.16 (p. 34) qui ne mettait pas en relief l'apparition des problématiques de l'entreprise et sa tentative de les résoudre.

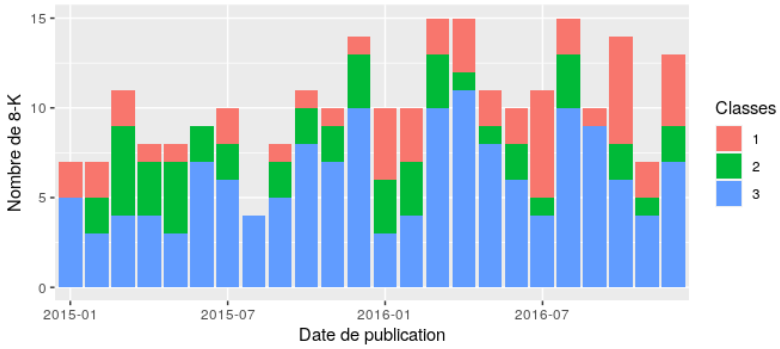


FIGURE 5.18 – Répartition par mois des classes HDDC pour la représentation unigramme.

9. https://www.justice.gov/opa/pr/wells-fargo-bank-agrees-pay-12-billion-improper-mortgage-lending-practices.

5.4.2 Comparaison avec les résultats obtenus dans le chapitre 4

La table 5.6 montre les Adjusted Rand Index (ARI) entre les regroupements ci-dessus et les classifications du **chapitre 4** (p. 41). Nous observons peu de similitudes entre les résultats obtenus avec les différentes méthodes.

Table 5.6 – Similarité des résultats *adj.rand.index* entre les résultats obtenus avec HDDC et ceux du **chapitre 4**

Classe	Uni Chap 3	WE Chap 3	Uni	TO	LDA
Uni Chap 4	1	0.55	0.05	0.00	0.01
WE Chap 4	-	1	0.05	0.00	0.01
Uni	-	-	1	0.31	0.23
TO	-	-	-	1	0.68
LDA	-	-	-	-	1

La figure 5.19 montre la répartition des variations réelles ($DOWN$, $STAY$, UP) selon les classes obtenues avec HDDC pour les unigrammes. Nous observons que chaque classe a sensiblement la même répartition des variations, ce qui tend à expliquer les résultats obtenus dans le **chapitre 4** (p. 41).

Figure 5.19 – Répartition des variations réelles par classe HDDC pour la représentation unigramme.

Pour valider ce constat, nous allons reprendre la méthodologie de la **section 4.6.3** (p. 55) qui consiste à étudier les rapports au sens de la similarité cosinus.

Prenons un texte publié le 23 janvier 2015, appartenant à la classe 1 (HDDC) avec les événements *Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year* et *Financial Statements and Exhibits*. La variation observée sur les marchés et celle prédite par le modèle avec la représentation unigramme sont les mêmes, à savoir $STAY$. Le texte le plus similaire, au sens de la similarité cosinus, a été publié le 29 janvier 2016, avec les mêmes événements et classé dans 1 (HDDC), cependant la variation engendrée sur les marchés était $DOWN$ et celle prédite était $STAY$.

La même remarque peut être faite pour la classe 2, dans laquelle deux textes similaires ayant le même événement *Financial Statements and Exhibits*, publiés respectivement le 22 avril et le 30 juin 2015 ont été prédits avec la même variation $STAY$ ce qui est vrai pour le premier, mais erroné pour le second car une variation UP a été constatée sur les marchés.

5.5 Conclusion

Dans ce chapitre, nous cherchons à sélectionner la meilleure représentation textuelle, dans le cadre d'une approche exploratoire, permettant une meilleure compréhension de la classification non supervisée et de la visualisation du texte.

Nous montrons, tout d'abord, sur un corpus composé des rapports 8-K de l'entreprise Wells Fargo publiés entre 2015 et 2016, qu'une représentation des textes par unigrammes ainsi que celle du transport optimal permettent une compréhension plus aisée de la classification. Puis, nous montrons également que ces dernières favorisent une visualisation optimale des textes classés et offrir ainsi une meilleure appréhension du texte, plus performante que des techniques usuelles comme le modèle de sujet [17]. Ainsi, pour la représentation unigramme, l'analyse des paramètres intrinsèques de la méthode HDDC permet de connaître l'impact de chaque terme dans la classification. Pour celle issue du transport optimal, c'est l'étude entropique des plans de transport qui met en avant les similitudes, mais aussi les différences lexicales entre les textes.

Enfin, la **section 5.4** met en évidence les difficultés de prédictions apparues dans le chapitre précédent. Nous constatons que deux rapports de la même entreprise avec les mêmes événements et contenus peuvent générer des variations différentes.

Chapitre 6

Pénalisation l_1 pour un mélange de lois de von Mises-Fisher

Résumé : Les mélanges de distributions de von Mises-Fisher peuvent être utilisés pour regrouper des données sur l'hypersphère unitaire. Ceci est particulièrement adapté aux données directionnelles de haute dimension telles que les textes. Nous proposons dans ce chapitre d'estimer un mélange de von Mises-Fisher en utilisant une vraisemblance pénalisée l_1. Cela conduit à des prototypes parcimonieux qui améliorent l'interprétabilité du clustering. Nous introduisons un algorithme EM pour cette estimation et explorons le compromis entre le terme de parcimonie et celui de la vraisemblance avec un algorithme de suivi de chemin. Le comportement du modèle est étudié sur des données simulées et nous montrons les avantages de l'approche sur des données réelles de référence. Nous utilisons également l'ensemble de données sur les rapports financiers de l'entreprise Wells Fargo pour les années 2015 à 2019 et montrons les avantages de notre méthode pour l'analyse exploratoire.

6.1 Introduction

Beaucoup de modèles de mélanges classiques sont peu adaptés aux données de grande dimension, comme celles qui sont issues de la représentation vectorielle de textes. Quand les données sont directionnelles [78], c'est-à-dire quand ce sont plutôt leurs corrélations que leurs distances euclidiennes qui importent, les modèles de type Gaussien sont encore moins adaptés. Pour de telles données, il est naturel d'opérer une normalisation qui les place sur la sphère unité. Nous montrons alors que les mélanges de lois de von Mises-Fisher (vMF) sur cette sphère sont bien adaptés pour la classification (non supervisée), cf [120, 9, 48].

La distribution de von Mises-Fisher (vMF) est une distribution de probabilité sur une hypersphère unitaire qui appartient au domaine des statistiques directionnelles [78]. Elle se concentre principalement sur les directions des objets et mesure la distance entre eux en utilisant la similarité en cosinus. Les premiers travaux utilisant la distribution de von

Mises-Fisher ont porté sur des données de faible dimension en raison de la difficulté à estimer le paramètre κ, qui implique l'inversion des rapports de fonctions de Bessel, cf [80]. Puis, dans le contexte du clustering et des données de grande dimension, les auteurs de l'article [9] ont proposé un modèle dérivé d'un mélange de distributions vMF utilisant une solution basée sur l'algorithme EM pour estimer les paramètres, ainsi qu'une approximation pour estimer le paramètre de concentration κ.

Cette contribution a conduit au développement de nombreuses applications de la distribution de von Mises-Fisher pour les données de grande dimension comme un *topic model* sphérique, [94], inspiré par l'Allocation de Dirichlet latente, ou pour le regroupement de textes [48].

Plus récemment, [101] a introduit un nouveau modèle basé sur les distributions de von Mises-Fisher dans le contexte du co-clustering en proposant un co-clustering diagonal. Les lignes et les colonnes possèdent alors le même nombre de classes. Après une réorganisation appropriée des lignes et des colonnes, la méthode permet d'obtenir une structure diagonale en blocs. Puisque le co-clustering a tendance à générer des solutions très asymétriques, avec des classes très déséquilibrées ou même vides, l'article [100] a ensuite introduit un *conscience mechanism*.

Entre-temps, de nombreuses recherches ont été menées sur le regroupement de données de grande dimension. Dans [23], les auteurs développent un modèle de mélange gaussien dont la principale caractéristique est la sélection automatique d'une projection spécifique de basse dimension pour chaque classe. En ce sens, il peut être considéré comme une généralisation du principe de mélange d'analyse en composantes principales [108]. En gardant la même idée, les auteurs de l'article [22] présentent un modèle de mélange latent discriminant, qui ajuste les données dans un sous-espace discriminant orthonormal latent avec une dimension intrinsèque inférieure à la dimension de l'espace original. Ils utilisent comme modèle d'estimation, le modèle dit de Fisher-EM. Puis, dans [21], ils introduisent une pénalité de type l_1 pour apporter de la parcimonie à la matrice d'orientation du sous-espace discriminant. En gardant cette idée de la pénalité de type l_1, [64] propose un k-means sphérique parcimonieux et qui est bien adapté au regroupement de textes.

Dans ce chapitre, en s'inspirant de [87], nous proposons une pénalisation l_1 pour un mélange de distributions vMF dans le but d'augmenter la parcimonie des moyennes directionnelles et ainsi améliorer la compréhension des résultats de classification des données de grande dimension. Notre solution s'appuie sur une modification de l'algorithme espérance - maximisation proposé par [9]. De plus, nous proposons une méthodologie pour estimer ce terme de pénalité, appelé le chemin de β, en ligne avec les critères de sélection de modèles. En outre, comme dans [101, 100], nous réorganisons les colonnes permettant une analyse fine des variables communes entre les classes mais surtout des variables qui permettent de discriminer ces classes (cf. figure 6.26). Cela n'était pas possible avec les méthodes de co-clustering telles que celles mentionnées précédemment.

6.2 Mélange de lois de von Mises-Fisher

Nous présentons rapidement dans cette section le modèle de mélange de lois de von Mises-Fisher issu de l'article [9]. Ce modèle génératif fournit une distribution sur $\mathbb{S}^{d-1}$, la sphère unité de dimension $(d-1)$ intégrée dans $\mathbb{R}^d$, c'est à dire

$$\mathbb{S}^{d-1} = \left\{ \boldsymbol{x} \in \mathbb{R}^d | \, \|\boldsymbol{x}\|_2 = 1 \right\},$$

où $\|.\|_2$ désigne la norme (euclidienne) l_2 dans $\mathbb{R}^d$.

6.2.1 Loi de von Mises-Fisher (VMF)

Dans $\mathbb{S}^{d-1}$ ($d \geq 2$), la densité de probabilité d'une loi de von Mises Fisher s'écrit :

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}}. \tag{6.1}$$

où $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ est la moyenne directionnelle de la distribution et $\kappa \geq 0$, son paramètre de concentration. La constante de normalisation $c_d(\kappa)$ est donnée par :

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}. \tag{6.2}$$

où I_r est la fonction de Bessel modifiée du premier type et d'ordre r.

6.2.2 Estimateur du maximum de vraisemblance

Comme le montre par exemple [9], l'estimation du maximum de vraisemblance (MLE) de la moyenne directionnelle d'une loi de von Mises-Fisher à partir d'un échantillon de données indépendantes identiquement distribuées $\boldsymbol{X} = (\boldsymbol{x}_i)_{1 \leq i \leq N}$ est direct, car nous avons

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n \boldsymbol{x}_i}{\|\sum_{i=1}^n \boldsymbol{x}_i\|_2}. \tag{6.3}$$

Cependant, l'estimation de κ est seulement indirect. En effet, nous pouvons montrer que $\widehat{\kappa}$ est la solution de l'équation suivante

$$\frac{I_{d/2}(\widehat{\kappa})}{I_{d/2-1}(\widehat{\kappa})} = \frac{1}{n} \left\| \sum_{i=1}^n \boldsymbol{x}_i \right\|_2, \tag{6.4}$$

qui n'admet pas de solution de forme fermée. Nous suivons la stratégie de [9] qui estime κ par

$$\widetilde{\kappa} = \frac{\bar{r}d - \bar{r}^3}{1 - \bar{r}^2}, \tag{6.5}$$

avec

$$\bar{r} = \frac{1}{n} \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \right\|_2 . \tag{6.6}$$

Il est à noter que nous utilisons cette approche, car elle offre un bon compromis entre complexité et précision, mais des méthodes numériques plus avancées peuvent être utilisées, voir par exemple [57] pour une discussion à ce sujet.

6.2.3 Mélange de lois de von Mises-Fisher

Nous considérons maintenant un mélange de K lois de von Mises-Fisher. Notons $f_k(\boldsymbol{x}|\theta_k)$ une loi de von Mises Fisher de paramètre $\theta_k = (\boldsymbol{\mu_k}, \kappa_k)$ pour $1 \leq k \leq K$. Alors un mélange de ces K lois de von Mises Fisher a pour densité :

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{x}|\theta_k). \tag{6.7}$$

où $\boldsymbol{\Theta} = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$ rassemble les moyennes directionnelles $(\boldsymbol{\mu}_k)_{1\leq k\leq K}$, les paramètres de concentration $(\kappa_k)_{1\leq k\leq K}$ et les proportions du mélange $(\alpha_k)_{1\leq k\leq K}$ avec $\alpha_k \geq 0$ et $\sum_{k=1}^{k} \alpha_k = 1$.

Les paramètres $\boldsymbol{\Theta}$ peuvent être estimés à partir d'un ensemble de données par maximum de vraisemblance à l'aide de l'algorithme EM, comme le montre [9]. Dans la section suivante, nous détaillons une variation de cet algorithme, adaptée à l'estimation régularisée que nous proposons.

6.3 Mélange de lois de von Mises-Fisher parcimonieuses

En s'inspirant de [87], nous proposons de remplacer l'estimateur du maximum de vraisemblance de $\boldsymbol{\Theta}$ par une version régularisée selon la norme l_1. Ceci permet d'augmenter la parcimonie des moyennes directionnelles et, ainsi, faciliter l'interprétation des résultats.

6.3.1 Vraisemblance pénalisée pour des moyennes directionnelles parcimonieuses

6.3.1.1 Représentation du mélange

Nous utilisons la représentation classique d'un mélange via des variables latentes. Nous supposons que l'ensemble complet de données consiste en N paires indépendantes et identiquement distribuées $(\boldsymbol{x}_i, z_i)_{1\leq i\leq N} = (\boldsymbol{X}, \boldsymbol{Z})$. Les $(z_i)_{i\leq i\leq N}$ sont les variables latentes non observées tandis que les $(\boldsymbol{x}_i)_{1\leq i\leq N}$ le sont. Chaque z_i suit une distribution catégorielle sur $\{1, \ldots, K\}$ avec le paramètre $\boldsymbol{\alpha} = (\alpha_k)_{1\leq k\leq K}$, c'est à dire $\mathbb{P}(z_i = k|\boldsymbol{\alpha}) = \alpha_k$.

Alors la densité conditionnelle de $\boldsymbol{x}_i$ étant donné $z_i = k$ est f_k, la k-ième composante du mélange vMF, c'est à dire

$$p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta}) = f_k(\boldsymbol{x}_i|\theta_k) = c_d(\kappa_k) \exp^{\kappa \boldsymbol{\mu}_k^T \boldsymbol{x}_i} . \tag{6.8}$$

Cela conduit à la distribution marginale de $p(\boldsymbol{x}_i|\boldsymbol{\Theta})$ donnée par l'équation (6.7) et la log-vraisemblance des données observées est donc

$$L(\boldsymbol{\Theta}|\boldsymbol{X}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{x}_i|\theta_k) \right) . \tag{6.9}$$

Pour faciliter la dérivation de l'algorithme EM, nous introduisons la représentation classique d'encodage 1 parmi n, ou *one-hot encoding* en anglais, des variables latentes : z_i est représenté par le vecteur binaire $\boldsymbol{z}_i$ avec $\sum_{k=1}^{K} z_{ik} = 1$ et tel que $z_i = k \Leftrightarrow z_{ij} = 0$ pour $j \neq k$ et $z_{ik} = 1$. Alors la log-vraisemblance complétée est donnée par

$$L(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\ln \alpha_k + \ln f_k(\boldsymbol{x}_i|\theta_k) \right) . \tag{6.10}$$

6.3.1.2 Vraisemblance pénalisée

Nous proposons de pénaliser la vraisemblance par la norme l_1 permettant ainsi d'augmenter la parcimonie de la représentation des moyennes directionnelles. Plus précisément, nous cherchons à estimer $\boldsymbol{\Theta}$ en maximisant la log-vraisemblance pénalisée :

$$L_p(\boldsymbol{\Theta}|\boldsymbol{X}) = L(\boldsymbol{\Theta}|\boldsymbol{X}) - \beta \sum_{k=1}^{K} \|\boldsymbol{\mu_k}\|_1 , \tag{6.11}$$

où β règle le compromis entre la vraisemblance et la parcimonie. Comme nous utilisons la log-vraisemblance complétée dans l'algorithme EM, nous introduisons maintenant sa

version pénalisée

$$L_p(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z}) = L(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z}) - \beta \sum_{k=1}^{K} \|\boldsymbol{\mu_k}\|_1 . \tag{6.12}$$

6.3.2 Algorithme EM

6.3.2.1 Étape E

Nous suivons les articles [87] et [9] pour obtenir un algorithme EM pour notre estimateur pénalisé. Dans l'étape espérance de l'algorithme EM, nous calculons l'espérance de $\ln L_p(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})$ par rapport à une distribution sur les variables latentes $\boldsymbol{Z}$. De toute évidence,

$$\mathbb{E}_{\boldsymbol{Z}\sim q}\left(L_p(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})\right) = \mathbb{E}_{\boldsymbol{Z}\sim q}\left(L(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})\right) - \beta \sum_{k=1}^{K} \|\boldsymbol{\mu_k}\|_1 , \tag{6.13}$$

pour toute distribution q car le terme de pénalité ne dépend pas de $\boldsymbol{Z}$. Ainsi l'étape E n'est pas modifiée malgré la pénalisation de la vraisemblance.

Notons

$$\tau_{ik}^{(m)} = \mathbb{P}(z_i = k|\boldsymbol{x}_i, \boldsymbol{\Theta^{(m)}}) = \frac{\alpha_k^{(m)} f_k(\boldsymbol{x}_i, \theta_k^{(m)})}{\sum_{l=1}^{K} \alpha_l^{(m)} f_l(\boldsymbol{x}_i, \theta_l^{(m)})}. \tag{6.14}$$

Nous savons de [9] que

$$\begin{aligned} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(m)}) &= \mathbb{E}_{\mathcal{Z}\sim q(\mathcal{Z}|\mathcal{X},\boldsymbol{\Theta}^{(m)})}\left(L(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})\right), \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(m)} \left(\ln \alpha_k + \ln f_k(\boldsymbol{x}_i|\theta_k)\right), \end{aligned} \tag{6.15}$$

et par conséquent

$$Q_p(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(m)} \left(\ln \alpha_k + \ln f_k(\boldsymbol{x}_i|\theta_k)\right) - \beta \sum_{k=1}^{K} \|\boldsymbol{\mu_k}\|_1 . \tag{6.16}$$

Comme il est classique dans les modèles de mélange, le $\tau_{ik}^{(m)}$ peut être utilisé pour définir un partitionnement fort des observations en K clusters. L'indice de cluster de l'observation $\boldsymbol{x}_i$, $c_i^{(m)}$, est donné par

$$c_i^{(m)} = \arg \max_{1\leq k\leq K} \tau_{ik}^{(m)}. \tag{6.17}$$

6.3.2.2 Étape M

L'étape M maximise le terme de l'équation (6.16) pour mettre à jour les estimations des paramètres avec pour contrainte $\sum_{k=1}^{K} \alpha_k = 1$, $\|\boldsymbol{\mu}_k\|_2 = 1$ et $\kappa_k \geq 0$ pour $1 \leq k \leq K$. Pour respecter ces contraintes sur α_k et $\boldsymbol{\mu_k}$, nous introduisons $K+1$ multiplicateurs, respectivement λ_k et ζ :

$$\mathcal{L}(\boldsymbol{\Theta}, \zeta, \boldsymbol{\lambda}|\boldsymbol{\Theta}^{(m)}) = Q_P(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(m)}) + \zeta(\sum_k \alpha_k - 1) + \sum_k \lambda_k(1 - \boldsymbol{\mu_k}^T \boldsymbol{\mu_k}), \tag{6.18}$$

Une dérivation simple montre que les dérivées partielles de $\mathcal{L}$ par rapport aux α_k sont égales à zéro si et seulement si

$$\forall k,\ \alpha_k = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(m)}. \tag{6.19}$$

Il s'agit de la mise à jour standard de la phase M obtenue dans [9], un fait évident considérant que le terme de pénalisation ne s'applique pas aux α_k.

Le cas des autres paramètres est plus compliqué. En effet, pour les paramètres de concentration, nous avons

$$\frac{\partial}{\partial_{\kappa_k}} \mathcal{L}(\boldsymbol{\Theta}, \zeta, \boldsymbol{\lambda}|\boldsymbol{\Theta}^{(m)}) = \sum_{i=1}^{N} \tau_{ik}^{(m)} \left(\frac{c_d'(\kappa_k)}{c_d(\kappa_k)} + \boldsymbol{\mu}_k^T \boldsymbol{x}_i \right). \tag{6.20}$$

En mettant la dérivée à zéro et en suivant [9], cela nous amène à

$$\frac{I_{d/2}(\kappa_k)}{I_{d/2-1}(\kappa_k)} = \boldsymbol{\mu}_k^T \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{ik}^{(m)}}. \tag{6.21}$$

Si le côté droit de cette équation est connu, nous pouvons alors utiliser l'estimateur rappelé dans la **section 6.2.2** pour calculer κ_k. Cependant, le découplage entre κ_k et et $\boldsymbol{\mu}_k$ qui apparaît en l'absence de régularisation ne s'applique pas dans le cas présent. Cela conduit à des mises à jour plus compliquées, comme détaillé ci-dessous.

Pour les moyennes directionnelles, nous devons considérer le sous gradient de $\mathcal{L}$. Nous avons

$$\partial_{\mu_{kj}} \mathcal{L}(\boldsymbol{\Theta}, \zeta, \boldsymbol{\lambda}|\Theta^{(m)}) = \kappa_k \sum_{i=1}^{n} \tau_{ik}^{(m)} x_{ij} - 2\lambda_k \mu_{kj} - \beta \partial_{\mu_{kj}} |\mu_{kj}|. \tag{6.22}$$

En utilisant la propriété du sous-gradient de la valeur absolue, nous obtenons

$$\partial_{\mu_{kj}} \mathcal{L}(\boldsymbol{\Theta}, \zeta, \boldsymbol{\lambda} | \Theta^{(m)}) = \begin{cases} \{\kappa_k r_{kj}^{(m)} - 2\lambda_k \mu_{kj} + \beta\} & \text{when } \mu_{kj} < 0, \\ \{\kappa_k r_{kj}^{(m)} - \epsilon\beta | \epsilon \in [-1; 1]\} & \text{when } \mu_{kj} = 0, \\ \{\kappa_k r_{kj}^{(m)} - 2\lambda_k \mu_{kj} - \beta\} & \text{when } \mu_{kj} > 0, \end{cases} \tag{6.23}$$

où

$$\boldsymbol{r}_k^{(m)} = \sum_i \tau_{ik}^{(m)} \boldsymbol{x}_i. \tag{6.24}$$

La condition d'optimalité du premier ordre est $0 \in \partial_{\mu_{kj}} \mathcal{L}(\boldsymbol{\Theta}, \zeta, \boldsymbol{\lambda} | \Theta^{(m)})$, ce qui emmène à l'analyse suivante.

Si nous recherchons une solution positive $\mu_{kj} > 0$, la condition d'optimalité est remplie lorsque

$$\mu_{kj} = \frac{\kappa_k r_{kj}^{(m)} - \beta}{2\lambda_k}. \tag{6.25}$$

Cette solution est compatible avec $\mu_{kj} > 0$ si $\kappa_k r_{kj}^{(m)} - \beta > 0$, c'est-à-dire quand $r_{kj}^{(m)} > \frac{\beta}{\kappa_k}$. Dans ce cas, nous avons aussi

$$\mu_{kj} = \text{sign}\left(r_{kj}^{(m)}\right) \frac{\kappa_k |r_{kj}^{(m)}| - \beta}{2\lambda_k}. \tag{6.26}$$

Si nous recherchons une solution négative $\mu_{kj} < 0$, la condition d'optimalité est remplie lorsque

$$\mu_{kj} = \frac{\kappa_k r_{kj}^{(m)} + \beta}{2\lambda_k}. \tag{6.27}$$

Ceci est compatible avec l'hypothèse $\mu_{kj} < 0$ si $\kappa_k r_{kj}^{(m)} + \beta < 0$, c'est à dire $r_{kj}^{(m)} < -\frac{\beta}{\kappa_k}$. Dans ce cas, nous avons également

$$\mu_{kj} = \text{sign}\left(r_{kj}^{(m)}\right) \frac{\kappa_k |r_{kj}^{(m)}| - \beta}{2\lambda_k}. \tag{6.28}$$

Enfin, une valeur nulle, $\mu_{kj} = 0$, remplit la condition d'optimalité si

$$0 \in \left[\kappa_k r_{kj}^{(m)} + \beta; \kappa_k r_{kj}^{(m)} - \beta\right].$$

C'est le cas quand $-\frac{\beta}{\kappa_k} \leq r_{kj}^{(m)} \leq \frac{\beta}{\kappa_k}$, c'est-à-dire quand $\kappa_k |r_{kj}^{(m)}| - \beta \leq 0$.

En résumé, la condition d'optimalité du premier ordre est remplie lorsque

$$\mu_{kj} = \frac{\operatorname{sign}\left(r_{kj}^{(m)}\right)}{2\lambda_k} \max(\kappa_k |r_{kj}^{(m)}| - \beta, 0). \tag{6.29}$$

Rappelons que le multiplicateur de Lagrange est calculé en utilisant les contraintes d'égalité $\|\boldsymbol{\mu_k}\|_2^2 = 1$. Cela donne

$$\left\| \sum_{j=1}^{d} \frac{\operatorname{sign}\left(r_{kj}^{(m)}\right)}{2\lambda_k} \max(\kappa_k |r_{kj}^{(m)}| - \beta, 0) \right\|_2^2 = 1,$$

$$\frac{1}{4\lambda_k^2} \sum_{j=1}^{d} (\max(\kappa_k |r_{kj}^{(m)}| - \beta, 0))^2 = 1,$$

et donc

$$\lambda_k = \frac{1}{2} \sqrt{\sum_{j=1}^{d} (\max(\kappa_k |r_{kj}^{(m)}| - \beta, 0))^2}. \tag{6.30}$$

Notons que l'équation (6.21), les équations (6.29) pour tous les j et l'équation (6.30) sont couplées.

Lorsque $\beta = 0$, μ_{kj} se simplifie en $\frac{\kappa_k r_{kj}^{(m)}}{2\lambda_k}$, ce qui implique $\lambda_k = \frac{1}{2}\sqrt{\sum_{j=1}^{d} \kappa_k^2 (r_{kj}^{(m)})^2}$. En retour, cela se simplifie en

$$\boldsymbol{\mu}_k = \frac{\sum_i \tau_{ik}^{(m)} \boldsymbol{x}_i}{\left\| \sum_i \tau_{ik}^{(m)} \boldsymbol{x}_i \right\|_2}.$$

Cette dernière est utilisée dans [9] pour obtenir des équations de forme fermée pour la phase M.

Cependant, dans notre cas où $\beta > 0$, nous ne pouvons pas tirer parti d'un tel découplage des équations. Nous proposons donc de résoudre la phase M de manière approximative, en utilisant une stratégie de point fixe. En utilisant l'estimation actuelle de κ_k, nous calculons une estimation mise à jour de $\boldsymbol{\mu}_k$ en utilisant les équations (6.29) et (6.30). Ensuite, nous mettons à jour κ_k en utilisant l'estimateur rappelé dans la **section 6.2.2**, c'est à dire

$$\kappa_k = \frac{d\rho_k - \rho_k^3}{1 - \rho_k^2}, \tag{6.31}$$

avec

$$\rho_k = \frac{\boldsymbol{\mu}_k^T \boldsymbol{r}_k^{(m)}}{\sum_i \tau_{ik}^{(m)}}. \tag{6.32}$$

Comme indiqué dans la **section 6.2.2**, des méthodes numériques plus avancées peuvent être utilisées pour estimer κ_k. Elles peuvent être intégrées dans l'algorithme EM sans aucune difficulté puisqu'elles résolvent simplement l'équation (6.21).

Nous itérons ces deux mises à jour jusqu'à convergence. Remarquons que pour assurer la consistance de cette stratégie avec les équations de forme fermée de [9] dans le cas où $\beta = 0$, nous devons mettre à jour $\boldsymbol{\mu}_k$ puis κ_k. La séquence inverse ne génère pas de mises à jour consistantes.

L'algorithme EM final est résumé dans l'algorithme 1. Nous discutons de l'implémentation dans la **section C.1** (p. 153).

6.3.2.3 κ commun

Comme le montre par exemple l'article [57], dans des contextes de haute dimension, les composantes des mélanges de von Mises-Fisher ont tendance à se spécialiser de manière excessive dans des sous-ensembles de données lorsque leurs paramètres de concentration deviennent très grands. Le problème peut être réduit en utilisant un seul paramètre κ commun entre toutes les composantes. Dans ce cas, les K équations (6.21) sont remplacées par la seule équation

$$\frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} = \frac{1}{N}\sum_{k=1}^{K}\boldsymbol{\mu}_k^T\left(\sum_{i=1}^{n}\tau_{ik}^{(m)}\boldsymbol{x}_i\right). \tag{6.33}$$

Alors, l'équation (6.29) est remplacée par

$$\mu_{kj} = \frac{\operatorname{sign}\left(r_{kj}^{(m)}\right)}{2\lambda_k}\max(\kappa|r_{kj}^{(m)}| - \beta, 0), \tag{6.34}$$

et l'équation (6.30) par

$$\lambda_k = \frac{1}{2}\sqrt{\sum_{j=1}^{d}(\max(\kappa|r_{kj}^{(m)}| - \beta, 0))^2}.. \tag{6.35}$$

Le reste de l'algorithme 1 reste inchangé.

6.3.3 Suivi de chemin de β

L'algorithme 1 peut être appliqué pour toute valeur fixe de β. Une stratégie possible, pour explorer l'effet de β, serait d'appliquer l'algorithme à partir de zéro pour différentes valeurs, par exemple régulièrement espacées sur une grille. Pour réduire la charge de calcul, améliorer la convergence et assurer la cohérence entre les modèles, nous proposons au contraire d'adopter une stratégie de suivi de chemin.

L'idée principale est de commencer par une solution dense pour $\beta = 0$ et d'augmenter progressivement la valeur de β, en recommençant chaque fois l'algorithme 1 à partir de la solution précédente. En outre, des incréments significatifs de β peuvent être calculés à partir de l'équation (6.29) : nous pouvons en effet rechercher une augmentation minimale de β qui garantit une augmentation de la parcimonie des moyennes directionnelles (au moins pendant la première itération de l'algorithme 1).

Désignons par $\boldsymbol{\Theta}\{\beta\}$ les paramètres estimés en appliquant l'algorithme 1 jusqu'à convergence pour une valeur donnée de β. Par exemple, $\mu_{kj}\{0\}$ est la coordonnée j de la moyenne directionnelle de la composante k lorsque $\beta = 0$. Naturellement, $\boldsymbol{r}_k\{\beta\}$ est le résultat de l'application de l'équation (6.24) à $\boldsymbol{\Theta}\{\beta\}$ (en utilisant l'équation (6.14)).

Pour illustrer le calcul des incréments de β, considérons d'abord la solution initiale obtenue avec $\beta_0 = 0$ et définissons β_1 comme suit

$$\beta_1 = \min_{1 \leq k \leq K, 1 \leq j \leq d, \kappa_k\{0\}|r\{0\}_{kj}|>0} \kappa_k\{0\}\,|r_{kj}\{0\}|. \tag{6.36}$$

Considérons $0 < \beta < \beta_1$ et la première itération de l'algorithme 1 initialisée avec $\boldsymbol{\Theta}^{(0)} = \boldsymbol{\Theta}\{0\}$. La phase E ne dépend pas de β et aucune des quantités calculées dans cette phase ne change par rapport à $\boldsymbol{\Theta}\{0\}$ (au fur et à mesure que l'algorithme 1 converge). C'est également le cas pour la première partie de la phase M, les κ_k et le $\boldsymbol{\mu}_k$ restent inchangés (par exemple, $\kappa_k^{(1)} = \kappa_k\{0\}$). Considérons ensuite la mise à jour de $\mu_{kj}^{(1)}$. Selon l'équation (6.29), $\mu_{kj}\{0\} = 0$ ne peut être qu'une conséquence de $r_{kj}\{0\} = 0$. Alors pour toute valeur de $\beta > 0$, $\mu_{kj}^{(1)} = 0$. Au contraire, si $|r_{kj}\{0\}| > 0$, alors $|\mu_{kj}^{(1)}| > 0$ pour tout $\beta < \beta_1$ comme conséquence de la définition de β_1 et de l'équation (6.29). Évidemment, $|\mu_{kj}^{(1)}| < |\mu_{kj}\{0\}|$ à cause de l'effet de rétrécissement induit par $\beta > 0$ dans l'équation (6.29), mais à moins que $\beta \geq \beta_1$, la parcimonie des moyennes directionnelles n'augmentera pas pendant cette première étape de l'algorithme. Cette simple analyse ne permet pas de prédire tous les effets d'une valeur non nulle de β, et la parcimonie pourrait augmenter en raison de la modification du κ_k et du τ_{ij} induite par le rétrécissement. Néanmoins, fixer β à β_1 est la plus petite augmentation par rapport à β_0 qui *garantit* l'augmentation de la parcimonie de la solution pendant la première étape de l'algorithme.

Un raisonnement similaire montre que nous pouvons garantir une augmentation de la parcimonie (dans la première étape de l'algorithme) en commençant avec $\boldsymbol{\Theta}\{\beta_{p-1}\}$ en choisissant β_p donné par

$$\beta_p = \beta_{p-1} + \min_{h,j,\kappa_k\{\beta_{p-1}\}|r_{kj}\{\beta_{p-1}\}|-\beta_{p-1}>0} \kappa_k\{\beta_{p-1}\}\,|r_{kj}\{\beta_{p-1}\}| - \beta_{p-1}. \tag{6.37}$$

En pratique, nous proposons de commencer avec $\beta_0 = 0$ et d'itérer les mises à jour basées sur l'équation (6.37) pour générer une série de solutions. Pour éviter d'effectuer trop d'étapes, nous mettons à zéro les valeurs inférieures au seuil de précision numérique choisi après avoir mis à jour β. L'algorithme final de suivi de chemin est donné dans l'algorithme 2. Dans ce résumé, $EM(\beta)$ est un appel à l'algorithme 1 avec une ini-

tialisation aléatoire pour $\boldsymbol{\Theta}^{(0)}$, tandis que $EM(\beta, \boldsymbol{\Theta})$ utilise $\boldsymbol{\Theta}$ comme valeur initiale de $\boldsymbol{\Theta}^{(0)}$.

De même, le nombre de pas effectués par l'algorithme peut être aussi élevé que le nombre de dimensions multiplié par K, lorsque les coordonnées sont mises à zéro presque une par une. Afin de réduire la charge de calcul, nous pouvons imposer une augmentation minimale (relative) de β entre deux étapes. Il est également possible de limiter le chemin à P étapes (comme dans l'algorithme 2) ou de continuer à l'explorer jusqu'à ce que la parcimonie maximale soit atteinte (un seul paramètre non nul par moyenne directionnelle). Ces heuristiques seront utilisées dans les expériences.

6.3.4 Sélection de modèles

Suivant [19, 99], nous proposons d'utiliser des critères d'information pour la sélection des modèles, notamment pour fixer la valeur de β. Des études précédentes [19, 99] n'ont pas été très concluantes quant à la capacité du critère d'information d'Akaike [2] (AIC), du critère d'information bayésien [103] (BIC) et de leurs variantes à sélectionner systématiquement un nombre approprié de composantes. Pour les mélanges de von Mises-Fisher, l'AIC a tendance à surajuster en sélectionnant trop de composantes, tandis que le BIC a tendance à sous-ajuster à moins que le nombre d'observations soit suffisamment grand (plusieurs fois le nombre de dimensions). Pour la variante de co-clustering des mélanges de von Mises-Fisher proposée dans [101], l'AIC semble être la solution la plus appropriée étant donné le petit nombre de paramètres libres de ce modèle (cf. [99]).

Les limites de l'AIC et du BIC dans des contextes de haute dimension sont bien connues, et plusieurs variantes ont été proposées pour résoudre le problème dans le contexte de l'apprentissage supervisé (principalement la régression linéaire). Les variantes comprennent le *Risk Inflation Criterion* (RIC, [44]) et son extension spécifique à des contextes de grande dimension à savoir le RICc [118], ainsi que le BIC étendu (EBIC [32, 33]). D'autres variantes peuvent être trouvées, par exemple, dans [19].

La formule générale de ces critères est donnée par

$$IC(\boldsymbol{\Theta}\{\beta\}) = \phi(n, d) \times C(\boldsymbol{\Theta}\{\beta\}) - 2 \times \log L(\boldsymbol{\Theta}\{\beta\} | \boldsymbol{X}), \tag{6.38}$$

où $C(\boldsymbol{\Theta}\{\beta\})$ est le nombre de paramètres libres dans le modèle et $\phi(n, d)$ est un coefficient propre au critère qui peut dépendre du nombre d'observations n et de leur dimension d. La table 6.1 donne la définition de la fonction du coefficient pour une sélection de critères considérés dans la suite de ce chapitre.

Lorsque $\beta = 0, C(\boldsymbol{\Theta}\{0\})$ est facile à calculer. Comme les $\boldsymbol{\kappa}$ ne sont pas contraints, ils contribuent à K paramètres libres (et un seul paramètre lorsqu'un κ commun est utilisé). La somme des $\boldsymbol{\alpha}$ est égale à un, et contribue donc à $K - 1$ paramètres libres. Lorsque $\beta = 0$, les moyennes directionnelles sont simplement contraintes par leur norme unitaire et contribuent donc à $K(d - 1)$ paramètres libres [1].

1. Notons que [19, 99] négligent la contrainte de la norme unitaire et considèrent Kd paramètres.

Critère	$\phi(n, d)$
AIC [2]	2
BIC [103]	$\log n$
RIC [44]	$2 \log d$
RICc [118]	$2(\log d + \log \log d)$
EBIC [32]	$\log n + 2\gamma \log d$

Table 6.1 – Coefficients pour les différents critères : n est le nombre d'observations et d leur dimension. Le paramètre γ de l'EBIC est fixé à $0,5$ comme recommandé dans [32].

Malheureusement, l'estimation du nombre de paramètres libres pour les moyennes directionnelles sous régularisation n'est pas évidente. Il a été montré dans [122] que dans le cas de la régression lasso, un estimateur cohérent du degré de liberté du modèle est donné en comptant le nombre de termes non nuls dans la régression. Cependant, les auteurs soulignent que ce résultat ne s'applique pas à d'autres contextes, comme par exemple dans le cadre *Elasticnet*. En conséquence, nous proposons d'utiliser comme nombre de paramètres libres pour une moyenne directionnelle donnée $\boldsymbol{\mu}$

$$C_{dm}(\boldsymbol{\mu}_k) = \max\left(1, \sum_{j=1}^{d} \mathbb{I}_{\mu_{kj} \neq 0} - 1\right), \tag{6.39}$$

où $\mathbb{I}$ est la fonction caractéristique. Dans le cas particulier où une seule coordonnée est non nulle en raison d'une forte régularisation, la contrainte de la norme unitaire réduit l'ensemble des valeurs possibles pour cette coordonnée à $\{-1, 1\}$. Nous le considérons toujours comme un paramètre libre et nous fixons donc $C(\boldsymbol{\mu})$ à 1 dans ce cas particulier (d'où l'opérateur $\max$ dans la définition). Alors le nombre de paramètres libres est donné par

$$C(\boldsymbol{\Theta}\,\{\beta\}) = (2K - 1) + \sum_{k=1}^{K} C_{dm}(\boldsymbol{\mu}_k\,\{\beta\}). \tag{6.40}$$

Dans la pratique, nous proposons d'utiliser le BIC ou l'AIC pour sélectionner le β optimal sur le chemin de régularisation. Nous proposons d'utiliser les autres critères comme guides pour sélectionner des configurations intéressantes en termes de nombre de composantes dans le mélange. En raison de la difficulté inhérente à l'estimation d'un modèle dans le cas de la haute dimension et du faible nombre d'observations, nous ne pouvons pas recommander de se concentrer sur un seul critère.

6.4 Expériences sur des données simulées

Nous présentons dans cette section des expériences qui illustrent le comportement de notre algorithme sur des données simulées. Banerjee et al. ont déjà démontré dans

[9] l'intérêt du mélange de la distribution de von Mises-Fisher par rapport aux autres solutions de partitionnement pour les données directionnelles. Par conséquent, les principaux axes de notre évaluation sont les effets de la régularisation, la pertinence des critères d'information pour la sélection de modèles et le comportement de l'algorithme de suivi de chemin.

6.4.1 Génération de données simulées

Pour étudier le comportement du modèle, nous utilisons des ensembles de données simulées qui sont générés par des mélanges de distributions de von Mises-Fisher. Nous générons les paramètres des distributions d'une manière semi-aléatoire qui nous permet de contrôler la séparation entre les composantes. La procédure générale pour un mélange de K composantes en dimension d est la suivante :

- nous échantillonnons $20 \times K$ vecteurs aléatoires uniformément sur l'hypersphère unitaire $\mathbb{S}^{d-1}$;
- nous extrayons de ces vecteurs, K vecteurs séparés de manière maximale, en minimisant leurs produits internes par paire de manière gloutonne : ce sont les moyennes directionnelles du mélange $(\boldsymbol{\mu}_k)_{1 \leq k \leq K}$;
- dans la plupart des simulations, nous ajoutons de la parcimonie aux moyennes directionnelles en mettant à zéro un sous-ensemble de leurs coordonnées choisi au hasard. Nous nous assurons de conserver des moyennes directionnelles non nulles et de les avoir toutes distinctes ;
- nous avons choisi $\boldsymbol{\kappa}$ de manière à assurer un degré donné de chevauchement entre les composantes. Le chevauchement est mesuré comme le taux d'erreur des affectations fortes obtenues par le modèle en utilisant les vrais paramètres par rapport à la vérité terrain. Pour une dimension $d = 100$, nous utilisons de base $\kappa = 17,34$ pour obtenir $2,5\%$ de chevauchement, et $\kappa = 15,09$ pour obtenir 5% de chevauchement ;
- pour chaque composante, κ_k est échantillonné à partir de la distribution gaussienne $\mathcal{N}(\mu = \kappa, \sigma = 0,025\kappa)$;
- la concentration finale de chaque composante du mélange est ajustée pour la séparabilité intrinsèque. Cela consiste à utiliser κ'_k défini par

$$\kappa'_k = \frac{2\kappa_k}{1 - \max_{l \neq k} \boldsymbol{\mu}_k^T \boldsymbol{\mu}_l}. \tag{6.41}$$

L'**annexe C.2** (p. 154) explicite la procédure de calibration. La table 6.2 montre les valeurs de κ utilisées pour une sélection de dimensions ainsi que de paramètres de chevauchement. Pour simplifier, nous utilisons systématiquement un mélange équilibré avec $\alpha_k = \frac{1}{K}$. En utilisant ces données, nous pouvons étudier la pertinence des critères d'information pour la sélection des modèles ainsi que le comportement de l'algorithme de suivi de chemin. Sauf indication contraire, nous présentons des statistiques obtenues en générant 100 ensembles de données pour chacune des configurations considérées.

Dans chaque exécution, le modèle est obtenu en exécutant l'algorithme EM à partir de dix configurations initiales aléatoires (cf. l'**annexe C.1** , p. 153) et en conservant les meilleurs résultats selon la vraisemblance (pénalisée).

Chevauchement Dimension	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
2	9.04	6.52	5.66	4.30	3.31	2.36	1.87	1.51
10	11.34	8.86	7.90	6.48	5.37	4.27	3.66	3.14
100	26.09	21.98	20.10	17.34	15.09	12.61	10.95	9.66
1000	76.52	65.78	60.70	52.74	46.35	39.08	34.05	30.08

Table 6.2 – Valeurs de chevauchement et κ associés : chaque colonne correspond à une valeur de chevauchement et chaque ligne à une dimension.

6.4.2 Sélection de modèles pour des données denses

Dans cette partie, nous nous concentrons sur les critères de l'AIC et du BIC. Pour étudier leur pertinence, nous générons des ensembles de données avec des séparations variables entre les composantes et nous testons si le nombre réel de composantes est retrouvé en minimisant l'AIC ou le BIC. Nos expériences complètent celles rapportées dans [19] car nous considérons plus de cas, des dimensions plus élevées et un plus grand nombre de réplications de chaque paramètre (100 ensembles de données contre 20).

6.4.2.1 Cas en faible dimension

Comme dans [9], nous étudions d'abord un cas de faible dimension en dimension $d = 2$ et avec $K^* = 2$ composantes. Nous considérons plusieurs tailles d'échantillons et huit valeurs de chevauchement (cf. table 6.2). Nous testons $K \in \{1, 2, 3, 4, 5, 6\}$ et rapportons le taux de simulations pour lesquelles chaque valeur est considérée comme optimale, à la fois pour le BIC et l'AIC. Les résultats sont présentés dans les tableaux 6.3 et 6.4.

Pour un échantillon de très petite taille (50 observations), le BIC ne sélectionne le nombre correct de composantes que pour un faible chevauchement (5 %) et seulement environ 80% du temps. Pour des valeurs de chevauchement plus importantes, il se rabat prudemment sur une seule composante. Comme prévu, l'AIC surestime le nombre de composantes même dans les cas fortement séparés. La figure 6.1 montre une représentation graphique des résultats.

Un ensemble de données plus important, avec 200 observations, améliore considérablement les résultats du BIC, qui peut maintenant gérer davantage de chevauchement entre les composantes. Les performances de l'AIC s'améliorent également, mais il continue à surestimer le nombre de composantes. La figure 6.2 détaille ces résultats.

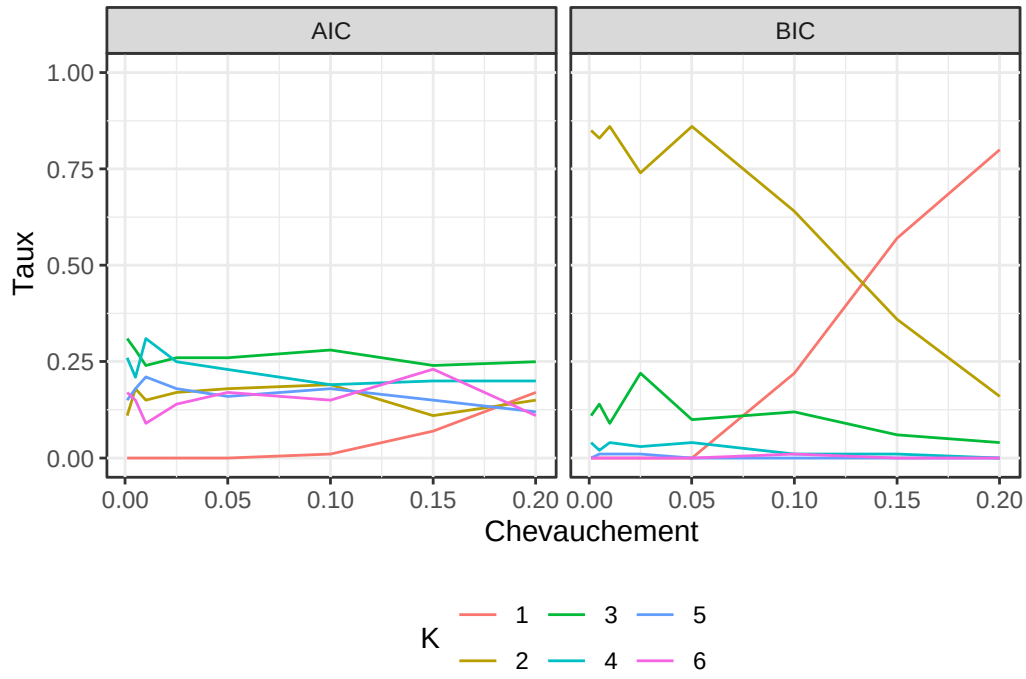


FIGURE 6.1 – Résultats de la sélection du nombre de composantes basée sur l'AIC et sur le BIC sur l'ensemble de données en dimension $d = 2$ pour 50 observations avec $K^* = 2$. Pour chaque K et chaque valeur de chevauchement, la figure affiche le pourcentage des jeux de données pour lesquels cette valeur de K a été considérée comme optimale.

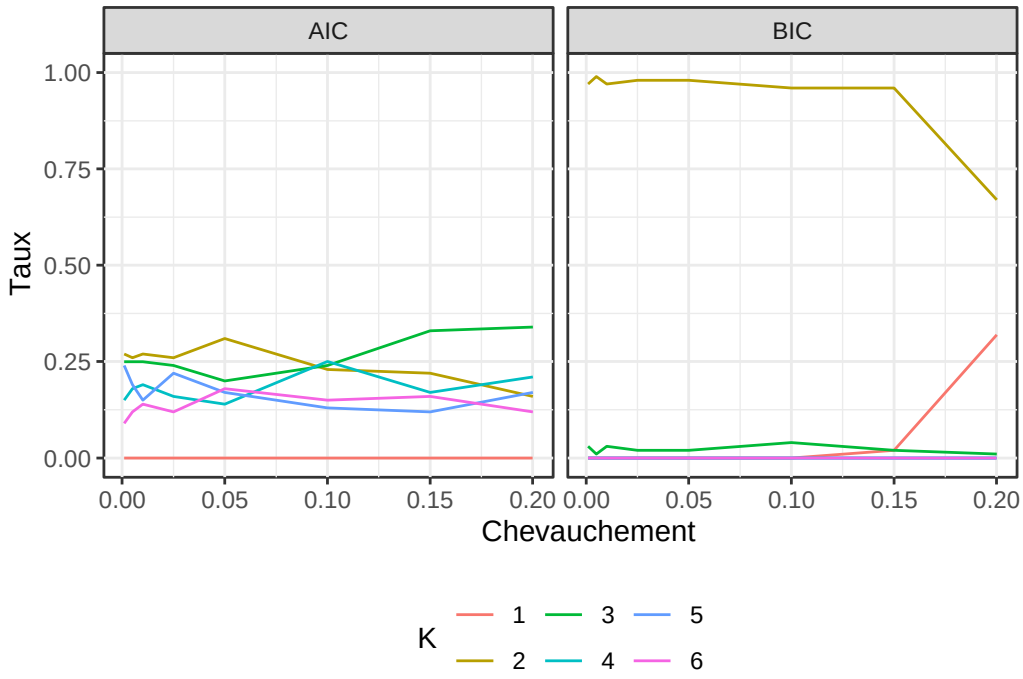


Figure 6.2 – Résultats de la sélection du nombre de composantes basée sur l'AIC et sur le BIC sur l'ensemble de données en dimension $d = 2$ pour 200 observations avec $K^* = 2$. Pour chaque K et chaque valeur de chevauchement, la figure affiche le pourcentage des jeux de données pour lesquels cette valeur de K a été considérée comme optimale.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
50	1	0.00	0.00	0.00	0.00	0.00	0.22	0.57	0.80
50	2	0.85	0.83	0.86	0.74	0.86	0.64	0.36	0.16
50	3	0.11	0.14	0.09	0.22	0.10	0.12	0.06	0.04
50	4	0.04	0.02	0.04	0.03	0.04	0.01	0.01	0.00
50	5	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
50	6	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
100	1	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.63
100	2	0.96	0.97	0.97	0.94	0.96	0.95	0.70	0.32
100	3	0.03	0.03	0.03	0.06	0.04	0.05	0.04	0.05
100	4	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	1	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.32
200	2	0.97	0.99	0.97	0.98	0.98	0.96	0.96	0.67
200	3	0.03	0.01	0.03	0.02	0.02	0.04	0.02	0.01
200	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
500	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
500	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
500	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 6.3 – Résultats complets du **BIC** sur le cas à 2 dimensions avec $K^* = 2$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
50	1	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.17
50	2	0.11	0.18	0.15	0.17	0.18	0.19	0.11	0.15
50	3	0.31	0.28	0.24	0.26	0.26	0.28	0.24	0.25
50	4	0.26	0.21	0.31	0.25	0.23	0.19	0.20	0.20
50	5	0.15	0.18	0.21	0.18	0.16	0.18	0.15	0.12
50	6	0.17	0.15	0.09	0.14	0.17	0.15	0.23	0.11
100	1	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05
100	2	0.19	0.20	0.22	0.25	0.25	0.31	0.21	0.20
100	3	0.33	0.30	0.33	0.32	0.31	0.23	0.20	0.32
100	4	0.22	0.19	0.21	0.10	0.20	0.23	0.25	0.20
100	5	0.12	0.20	0.21	0.17	0.15	0.13	0.21	0.15
100	6	0.14	0.11	0.03	0.16	0.09	0.10	0.11	0.08
200	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	2	0.27	0.26	0.27	0.26	0.31	0.23	0.22	0.16
200	3	0.25	0.25	0.25	0.24	0.20	0.24	0.33	0.34
200	4	0.15	0.18	0.19	0.16	0.14	0.25	0.17	0.21
200	5	0.24	0.19	0.15	0.22	0.17	0.13	0.12	0.17
200	6	0.09	0.12	0.14	0.12	0.18	0.15	0.16	0.12
500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	2	0.34	0.29	0.30	0.37	0.31	0.35	0.29	0.19
500	3	0.25	0.31	0.32	0.27	0.29	0.29	0.36	0.32
500	4	0.19	0.15	0.16	0.18	0.21	0.14	0.13	0.20
500	5	0.14	0.18	0.12	0.09	0.09	0.16	0.13	0.17
500	6	0.08	0.07	0.10	0.09	0.10	0.06	0.09	0.12

Table 6.4 – Résultats complets du **AIC** sur le cas à 2 dimensions avec $K^* = 2$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

Ces résultats sont confirmés par une autre série d'expériences pour $d = 10$, comme indiqué dans l'**annexe C.3.1** (p. 157). Globalement, l'AIC n'est pas adapté à la sélection de modèles en basse dimension.

6.4.2.2 Comportement en dimension supérieure

Comme nous l'avons souligné dans l'introduction, un des objectifs du mélange de distributions von Mises-Fisher est de permettre le clustering pour des données de haute dimension. Nous devons donc considérer le comportement des critères d'information dans une dimension supérieure, en particulier dans les situations où le nombre d'observations est relativement faible par rapport à la dimension. Nous étudions ici les dimensions moyennes à grandes avec des données denses en considérant $d = 100$ et $d = 1000$, avec $K^* = 4$ composantes. Nous discutons ici des résultats pour $d = 1000$ tandis que les résultats pour $d = 100$ sont rapportés dans l'**annexe C.3.2** (p. 161).

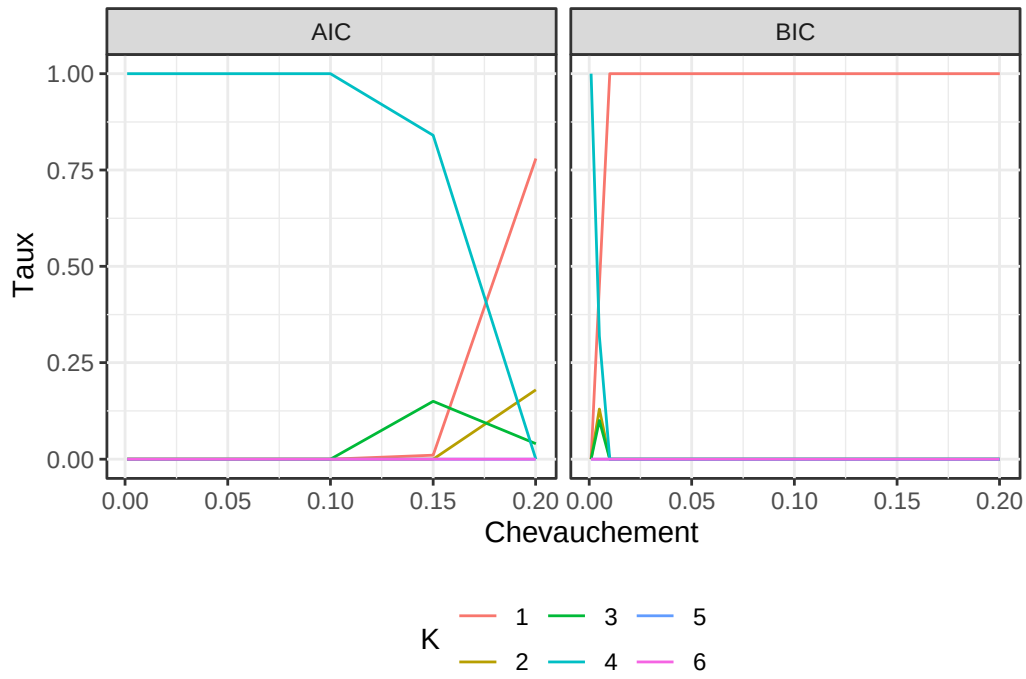


FIGURE 6.3 – Résultats de la sélection du nombre de composantes basée sur l'AIC et sur le BIC sur l'ensemble de données en dimension $d = 1000$ pour 2000 observations avec $K^* = 4$. Pour chaque K et chaque valeur de chevauchement, la figure affiche le pourcentage des jeux de données pour lesquels cette valeur de K a été considérée comme optimale.

Les résultats sont résumés dans les tables 6.5 et 6.6. Pour le BIC, ils montrent un comportement similaire à celui en basse dimension mais d'une manière plus extrême. Remarquons d'abord que dans ces expériences, le nombre d'observations est faible par rapport à la dimension, afin de se rapprocher du cas d'utilisation des données textuelles. Ceci a un effet très négatif sur le BIC qui ne parvient pas à retrouver le vrai nombre

de composantes dans des contextes sous-spécifiés tels que $n = 500$ observations en dimension $d = 1000$, même pour des composantes fortement séparées. Le BIC ne fonctionne correctement que lorsque n est supérieur à plusieurs fois la dimension et qu'en même temps les composantes sont bien séparées. La tendance au surajustement de l'AIC conduit ici à de meilleurs résultats, même pour les petits ensembles de données et un certain chevauchement. La différence de comportement entre les deux critères est bien illustrée par la figure 6.3 pour $n = 2000$ observations.

Cette analyse est confirmée par les résultats pour $d = 100$ (cf. l'**annexe C.3.2**, p. 161). En résumé, lorsqu'il y a suffisamment de données par rapport à la dimension, le BIC sélectionne correctement le vrai nombre de composantes tant qu'elles sont suffisamment séparées (où suffisamment dépend du nombre d'observations). Au contraire, dans le régime de faible taille des données, la tendance au surajustement de l'AIC fournit une meilleure estimation du nombre de composantes.

6.4.3 Illustration de la stratégie de suivi du chemin

Nous illustrons dans cette section le comportement de l'algorithme 2 de suivi de chemin proposé dans la **section 6.3.3** sur un exemple simple. Nous utilisons $K^* = 4$ composantes en dimension $d = 10$, avec une séparation de 5% ($\kappa = 5.37$). Nous générons $n = 500$ observations, ce qui rend l'estimation relativement facile compte tenu de la faible dimension des données (nous n'introduisons pas de parcimonie dans les moyennes directionnelles). Nous exécutons notre algorithme de suivi de chemin à partir de la meilleure configuration (en termes de vraisemblance) parmi dix configurations initiales aléatoires.

Les figures 6.4 et 6.5 montrent le comportement de l'algorithme. Dans cet exemple particulier, le chemin contient treize étapes. Au cours de la dernière étape, l'algorithme EM n'a pas convergé vers une configuration à 4 composantes, comme prévu lorsque la parcimonie devient trop importante. Bien qu'aucune parcimonie n'ait été imposée lors de la génération de l'ensemble de données, il était néanmoins utile de fixer certaines des composantes à zéro, car cela a conduit à une légère diminution du BIC (autour de l'étape 5).

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
500	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2000	1	0.00	0.45	1.00	1.00	1.00	1.00	1.00	1.00
2000	2	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00
2000	3	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00
2000	4	1.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00
2000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	1	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
5000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	4	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
5000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 6.5 – Résultats complets du **BIC** sur le cas à 1000 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
500	1	0.00	0.00	0.00	0.72	1.00	1.00	1.00	1.00
500	2	0.00	0.00	0.02	0.26	0.00	0.00	0.00	0.00
500	3	0.00	0.00	0.25	0.02	0.00	0.00	0.00	0.00
500	4	1.00	1.00	0.73	0.00	0.00	0.00	0.00	0.00
500	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	1	0.00	0.00	0.00	0.00	0.00	0.82	1.00	1.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00
1000	3	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00
1000	4	1.00	1.00	1.00	1.00	0.88	0.00	0.00	0.00
1000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.78
2000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18
2000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.04
2000	4	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.00
2000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64
5000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34
5000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02

Table 6.6 – Résultats complets du **AIC** sur le cas à 1000 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

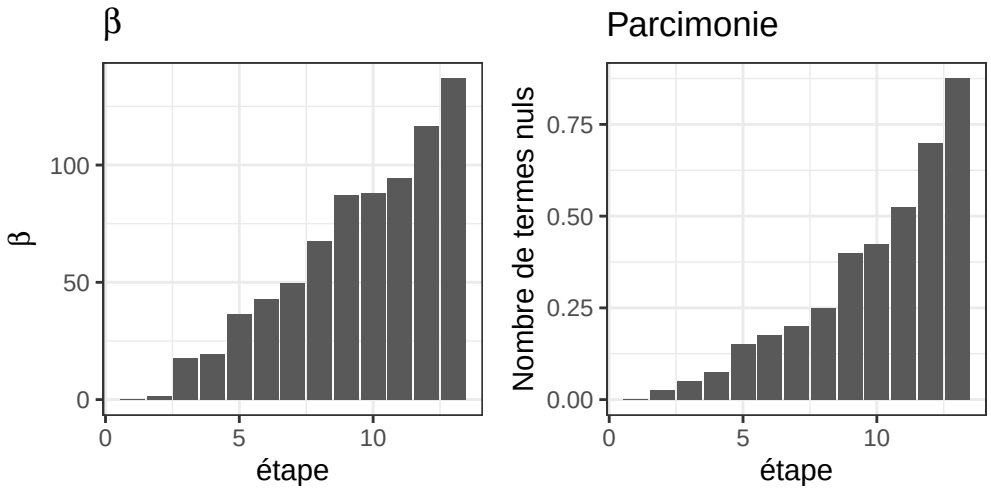


FIGURE 6.4 – Évolution de β et de la parcimonie de la solution au cours de l'algorithme de suivi du chemin.

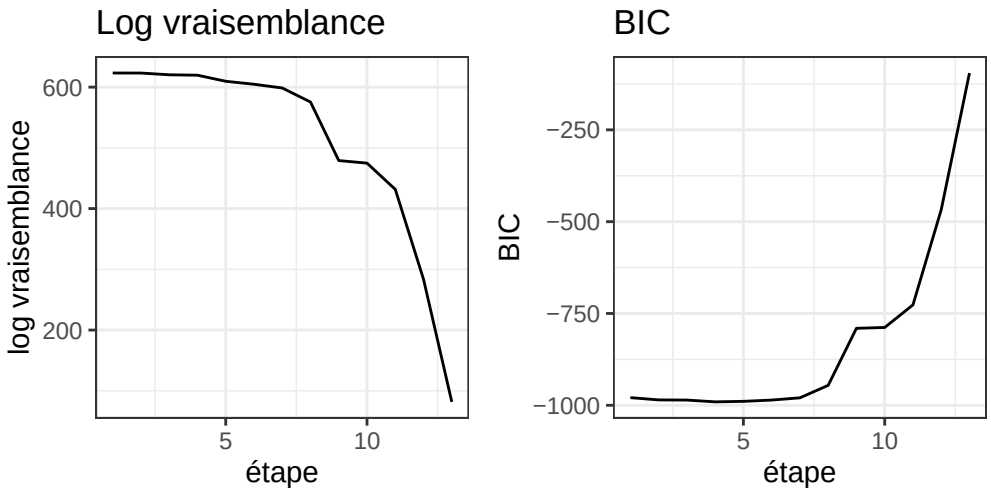


FIGURE 6.5 – Évolution de la log vraisemblance et du BIC de la solution au cours de l'algorithme de suivi de chemin.

Pour évaluer l'intérêt de l'algorithme de suivi de chemin sur cet exemple simple, nous avons d'abord utilisé comme configuration de départ la même solution que celle utilisée par le suivi de chemin et estimé directement les modèles régularisés à partir de cette configuration, en utilisant les β calculés par le chemin. Cela a généré les mêmes modèles estimés en un temps d'exécution plus long (25% d'itérations supplémentaires de l'algorithme EM).

Puis, nous avons utilisé les β calculés par le chemin mais nous avons démarré l'algorithme EM à partir de dix configurations initiales aléatoires pour chaque β. Une fois encore, nous avons obtenu des résultats identiques à ceux obtenus par l'algorithme de suivi de chemin. Cependant, nous avons évidemment utilisé environ dix fois plus de ressources de calcul et, en outre, un grand nombre de configurations initiales n'ont pas permis à l'algorithme EM de converger pour des valeurs plus importantes de β, comme le montre la figure 6.6.

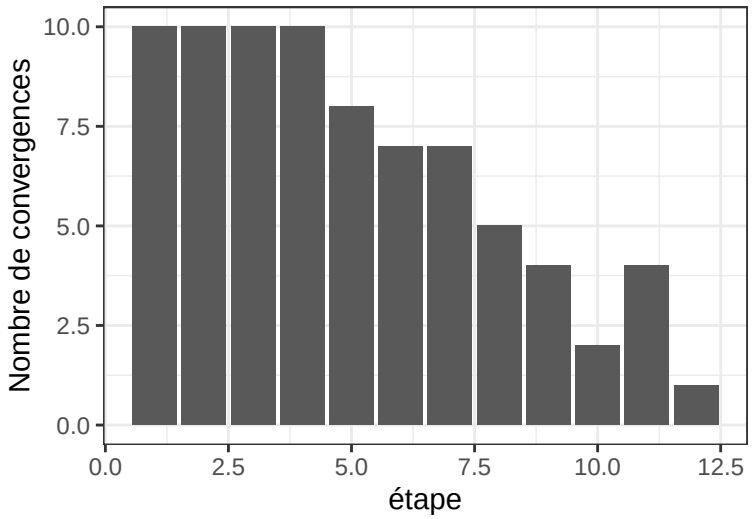


Figure 6.6 – Nombre d'exécutions EM convergentes (parmi 10) en fonction de β (représenté ici par l'étape dans l'algorithme de suivi du chemin).

Notons enfin que les valeurs de β sont assez imprévisibles. Sans la stratégie de suivi du chemin, nous aurions dû étudier l'effet de β échantillonné à partir d'une grille arbitraire de valeurs. Nous avons testé cette stratégie avec une grille de 50 valeurs non nulles régulièrement espacées entre 0 et la valeur maximale obtenue par l'algorithme de suivi de chemin. Sur la base des résultats des expériences précédentes, nous avons utilisé la solution obtenue avec $\beta = 0$ comme configuration initiale pour chaque valeur de β. Les résultats sont présentés dans les figures 6.7 et 6.8. Ils montrent un comportement identique de la recherche basée sur une grille et de l'algorithme de suivi de chemin en termes de vraisemblance et de BIC. Certains niveaux de parcimonie peuvent être manqués pendant le suivi du chemin (en comparant la figure 6.7 et la figure 6.4), mais cela est facilement corrigeable en testant quelques valeurs supplémentaires pour β à l'intérieur des intervalles où le saut de parcimonie est important.

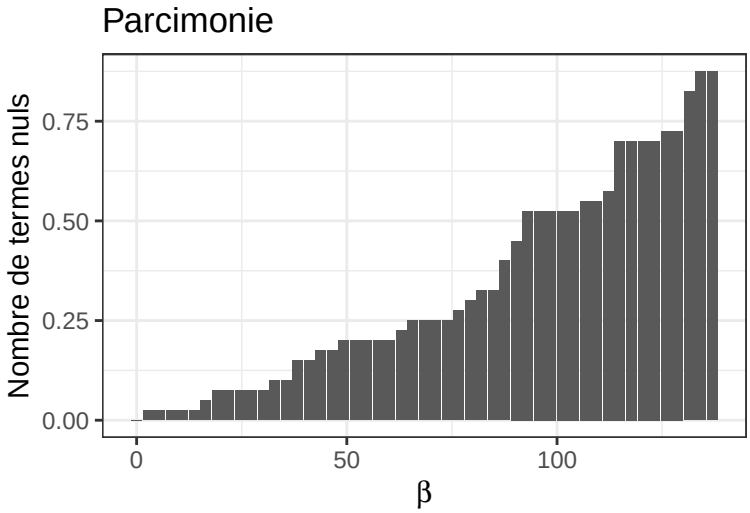


FIGURE 6.7 – Parcimonie de la solution en fonction de β.

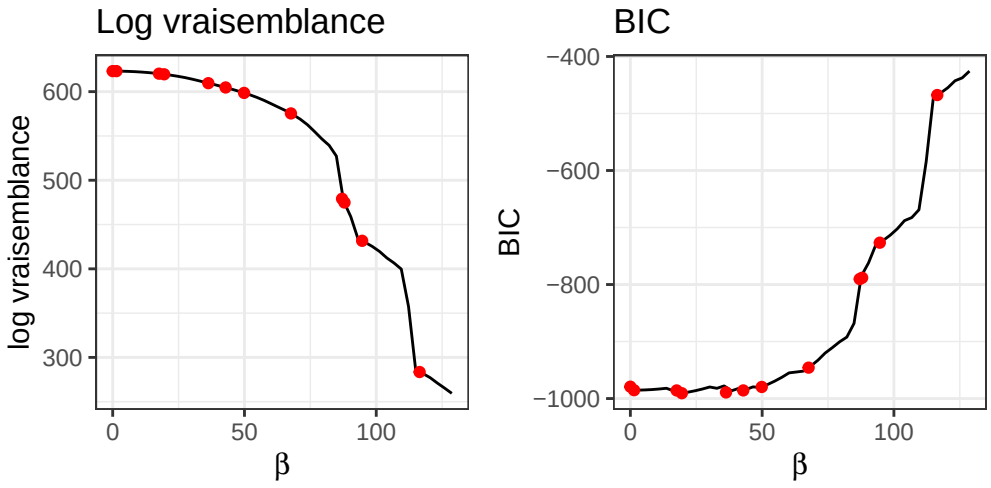


FIGURE 6.8 – Log vraisemblance et BIC de la solution en fonction de β. Les points rouges sont les configurations obtenues par l'algorithme de suivi de chemin.

En résumé, l'algorithme de suivi de chemin fournit efficacement un bon échantillonnage des valeurs de β qui ont un effet significatif sur la parcimonie de la solution. Si une analyse plus fine est nécessaire, nous pouvons échantillonner les intervalles entre les valeurs sur le chemin qui montrent une grande modification de la parcimonie de la solution.

6.4.4 Étude sur des simulations

Nous étudions le comportement de l'algorithme de suivi de chemin dans le cas de la dimension $d = 100$, avec $K^* = 4$ composantes et pour différents degrés de séparation entre les composantes, différents niveaux de parcimonie dans les moyennes directionnelles et les nombres d'observations. Remarquons que si les moyennes directionnelles sont parcimonieuses, ce n'est pas le cas des observations elles-mêmes, à moins que les $\boldsymbol{\kappa}$ ne soient fixés à des valeurs significativement plus grandes que celles que nous utilisons. Nous ne présentons ici que les résultats obtenus pour les valeurs spécifiques de $\boldsymbol{\kappa}$ des composantes car ceux obtenus avec un κ commun ne s'en écartent pas de manière significative.

Dans ces simulations, l'algorithme de suivi du chemin a été paramétré pour assurer une augmentation relative minimale de 10^{-3} entre deux valeurs consécutives de β.

6.4.4.1 Caractéristiques du chemin

Le comportement de l'algorithme de suivi de chemin est résumé par la figure 6.10 qui montre la distribution du nombre d'étapes effectuées sur le chemin ainsi que la distribution du nombre total d'itérations de l'algorithme EM. Par rapport au cas dense (c'est-à-dire à l'initialisation de l'algorithme) représenté sur la figure 6.9, suivre le chemin augmente considérablement la charge de calcul. Cependant, l'augmentation est bien moins importante que ce que l'on pourrait attendre du nombre de valeurs différentes de β considérées pendant l'exploration du chemin. En effet, le nombre médian d'itérations EM nécessaires pour obtenir une configuration initiale dense est supérieur à 500 (pour $K \geq 2$), alors qu'il est inférieur à 10000 pour l'exploration ultérieure du chemin. Ce ratio de 20 fois, est significativement plus petit que le nombre médian d'étapes (au moins 150 pour $K \geq 2$).

Les résultats présentés ici pour $n = 200$ observations sont représentatifs des résultats obtenus avec plus d'observations. Le nombre d'itérations a tendance à augmenter pour les grands K lorsque n augmente, mais cela ne change pas de manière significative le nombre d'étapes sur le chemin ou le rapport entre le nombre d'itérations EM dans le cas dense et sur le chemin.

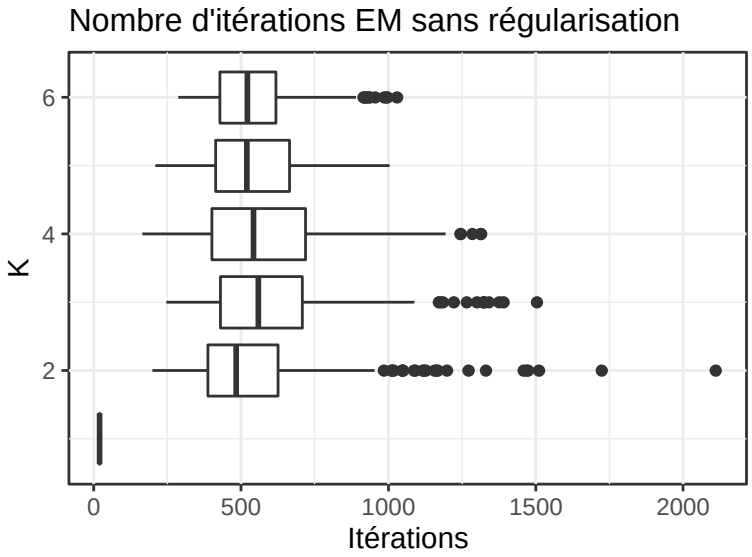


FIGURE 6.9 – Distributions du nombre d'itérations EM nécessaires pour obtenir le premier modèle avec $\beta = 0$ sur 600 ensembles de données avec $d = 100$ et $n = 200$, en fonction de K, le nombre de composantes. La figure regroupe les résultats pour toutes les valeurs de séparation et de parcimonie.

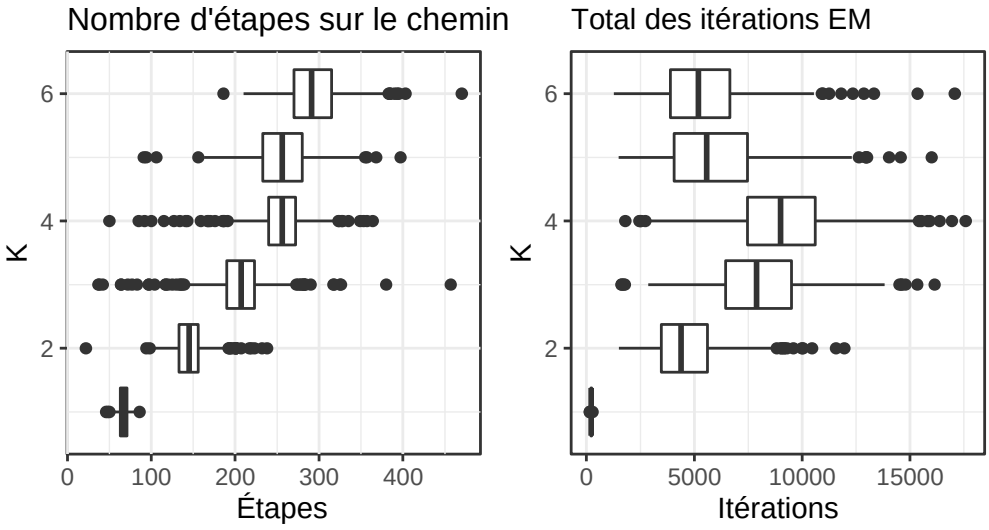


FIGURE 6.10 – Distributions du nombre d'étapes (c'est-à-dire des valeurs de β) et du nombre total d'itérations EM sur 600 ensembles de données avec $d = 100$ et $n = 200$, en fonction de K, le nombre de composantes. Les figures regroupent les résultats pour toutes les valeurs de séparation et de parcimonie.

6.4.4.2 Sélection du modèle

Pour chacune des 100 réplications, nous gardons le modèle dense original comme référence. Puis, nous sélectionnons le long du chemin β le meilleur modèle selon chacun des critères d'information présentés dans la **section 6.3.4**. Enfin, nous indiquons le nombre de composantes sélectionnées dans ces deux cas (dense et parcimonieux) en minimisant les critères d'information. Remarquons que dans le cas dense, nous avons un seul modèle évalué par plusieurs critères, alors que dans le cas parcimonieux, chaque critère sélectionne un modèle différent sur le chemin.

Les figures 6.11 et 6.12 montrent les résultats de cette approche dans le cas dense et celui parcimonieux (pour AIC et BIC), avec $n = 200$ observations. Comme la régularisation réduit le nombre de paramètres effectifs sans trop réduire la vraisemblance, elle favorise les modèles avec plus de composantes. Dans ce contexte, cela s'avère bénéfique pour le BIC mais conduit déjà l'AIC dans son régime de surajustement.

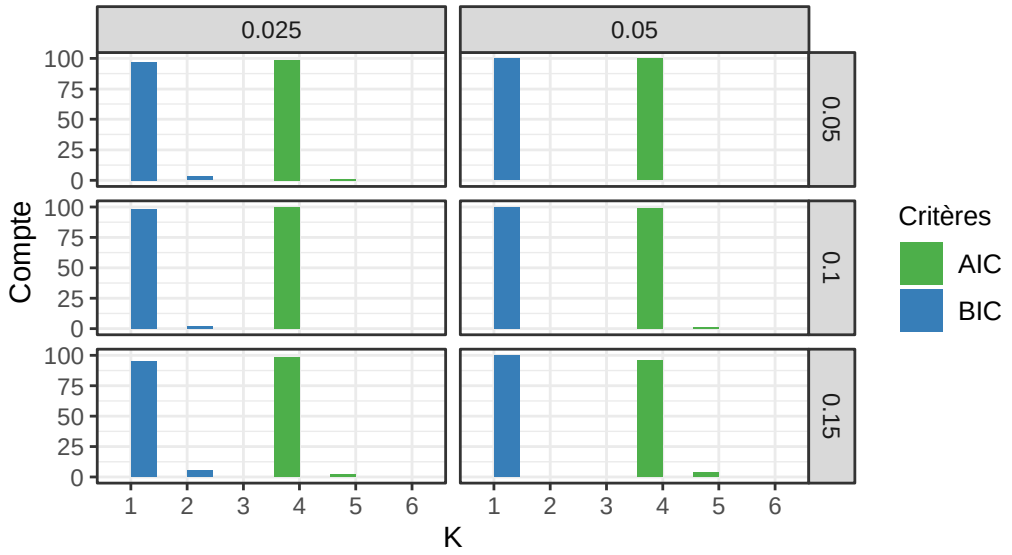


Figure 6.11 – **Cas dense** : nombre de fois où chaque K est sélectionné comme la meilleure configuration par l'AIC ou le BIC pour $\mathbf{n = 200}$ observations et $\beta = 0$, à travers les valeurs de chevauchement (en colonne) et la parcimonie des moyennes directionnelles (en ligne).

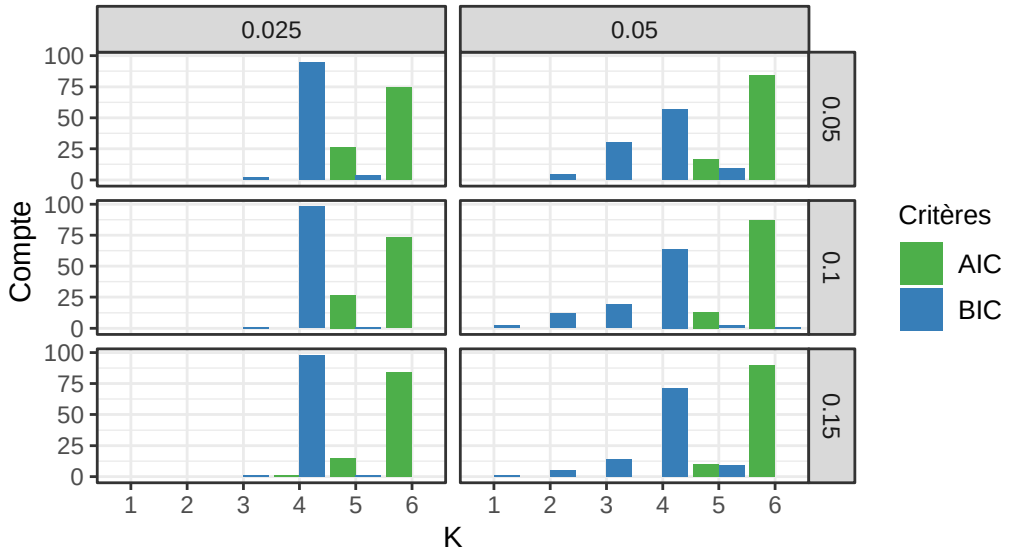


FIGURE 6.12 – **Cas parcimonieux** : nombre de fois où chaque K est sélectionné comme la meilleure configuration par l'AIC ou le BIC pour $\mathbf{n = 200}$ observations et pour le β optimal sélectionné par chaque critère, à travers les valeurs de chevauchement (en colonne) et la parcimonie des moyennes directionnelles (en ligne).

Malheureusement, ce comportement de surajustement de l'AIC se manifeste d'autant plus dans le cas plus simple où les observations sont de l'ordre de $n = 1000$ (cf. les figures 6.13 et 6.14), alors que le BIC, au contraire, est capable de retrouver le vrai nombre de composantes, avec ou sans régularisation.

6.4.5 Sélection sur le chemin et parcimonie

Afin d'étudier l'effet de la régularisation, nous calculons l'*Adjusted Rand Index* (ARI) entre la vérité terrain et les assignations fortes produites par les différents modèles. La figure 6.15 montre les résultats pour $n = 200$ observations. Dans ce cas, le BIC a tendance à rendre les moyennes directionnelles trop parcimonieuses par rapport à l'ARI, en particulier lorsque le $K = 4$, le nombre réel de composantes.

Ce phénomène est lié à la difficulté de l'estimation, comme le montre la figure 6.16 avec $n = 1000$ observations. Lorsque nous avons plus d'observations, lorsque les vraies moyennes directionnelles sont plus parcimonieuses ou lorsque les composantes se chevauchent moins, la chute de l'ARI entre le BIC et l'AIC est moins prononcée.

Il est également lié à la parcimonie réalisable compte tenu du nombre d'observations, comme l'illustrent les figures 6.17 et 6.18. En effet, avec plus d'observations, les estimations des composantes moyennes directionnelles sont plus compactes et les composantes non nulles nécessitent une plus grande valeur de β pour être éliminées. Le compromis entre parcimonie et vraisemblance est plus prononcé pour les modèles denses.

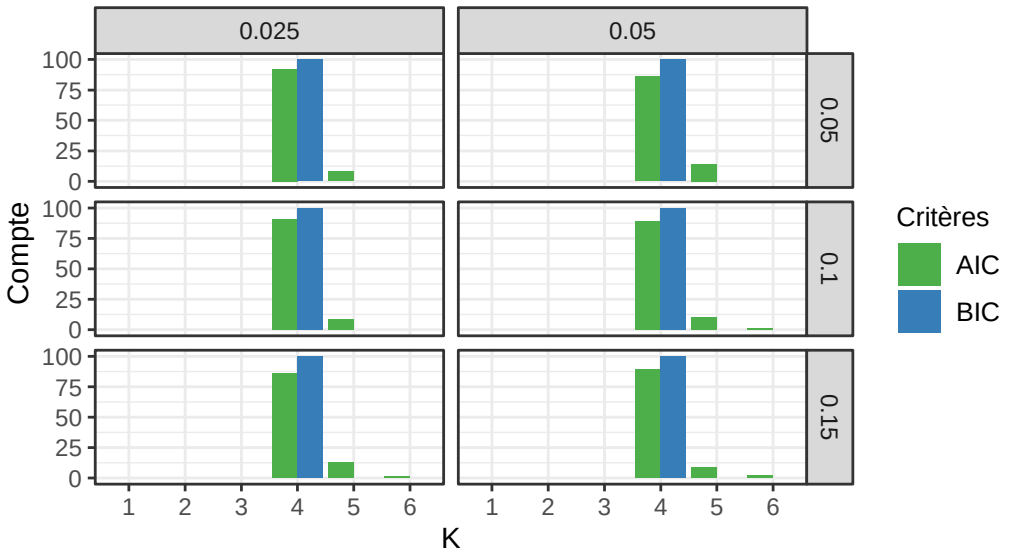


Figure 6.13 – **Cas dense** : nombre de fois où chaque K est sélectionné comme la meilleure configuration par l'AIC ou le BIC pour $\mathbf{n = 1000}$ observations et $\beta = 0$, à travers les valeurs de chevauchement (en colonne) et la parcimonie des moyennes directionnelles (en ligne).

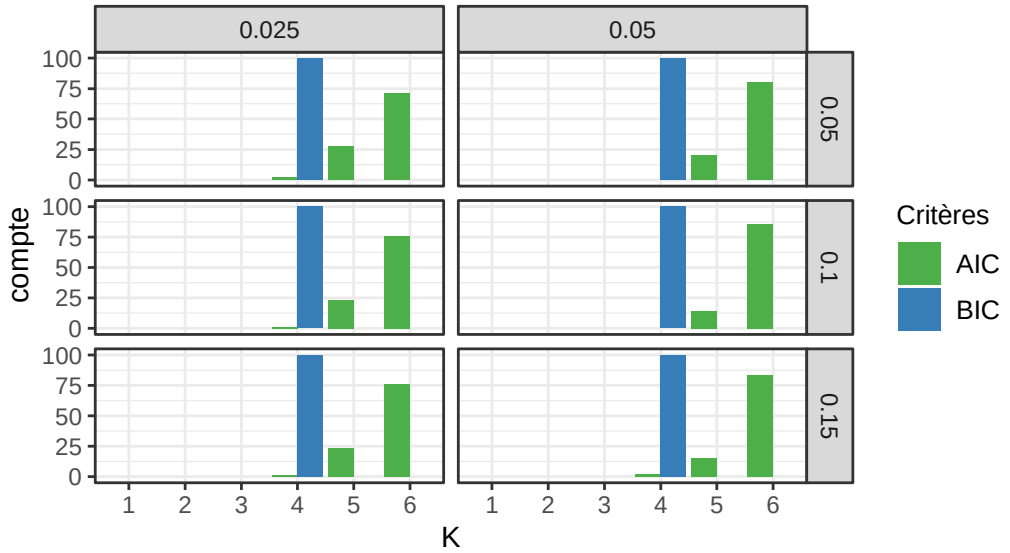


Figure 6.14 – **Cas parcimonieux** : nombre de fois où chaque K est sélectionné comme la meilleure configuration par l'AIC ou le BIC pour $\mathbf{n = 1000}$ observations et pour le β optimal sélectionné par chaque critère, à travers les valeurs de chevauchement (en colonne) et la parcimonie des moyennes directionnelles (en ligne).

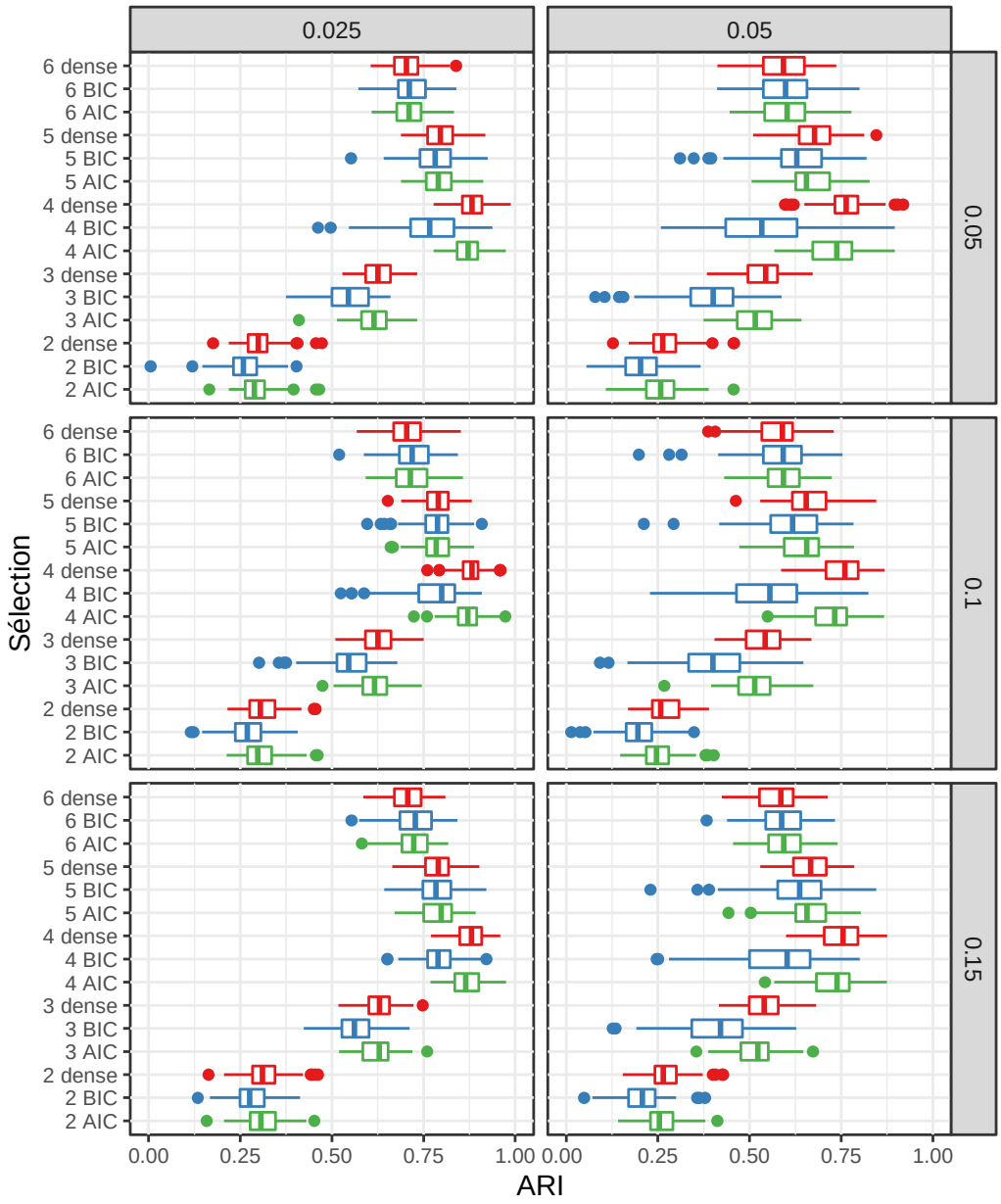


FIGURE 6.15 – Distribution de l'*Adjusted Rand Index* pour le modèle dense optimal (en rouge) et pour les modèles parcimonieux optimaux selon l'AIC (en vert) et le BIC (en bleu), en fonction de K, le nombre de composantes, pour $\mathbf{n = 200}$. Les panneaux sont organisés en fonction du chevauchement (verticalement) et de la parcimonie (horizontalement).

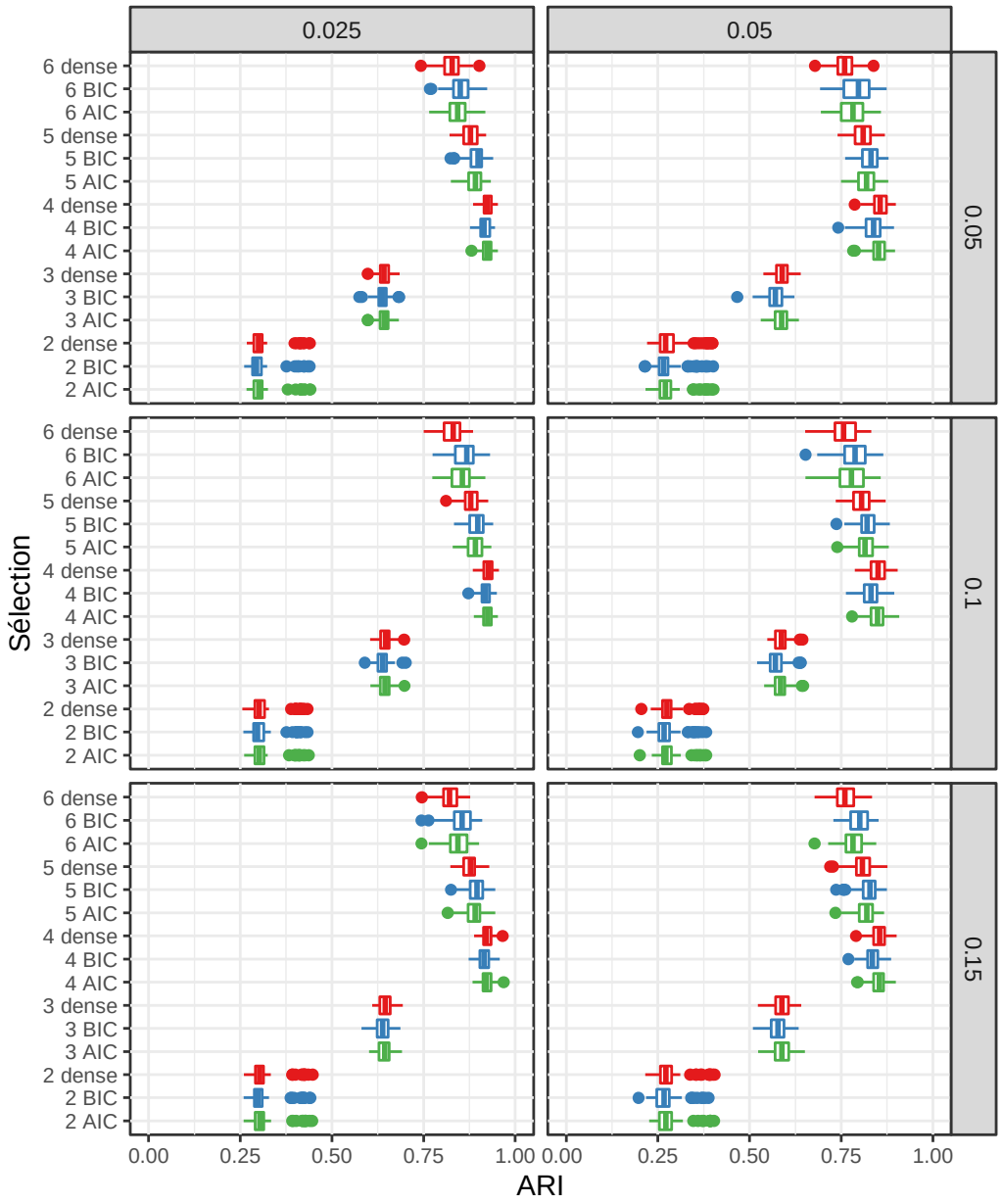


Figure 6.16 – Distribution de l'*Adjusted Rand Index* pour le modèle dense optimal (en rouge) et pour les modèles parcimonieux optimaux selon l'AIC (en vert) et le BIC (en bleu), en fonction de K, le nombre de composantes, pour $\mathbf{n = 1000}$. Les panneaux sont organisés en fonction du chevauchement (verticalement) et de la parcimonie (horizontalement).

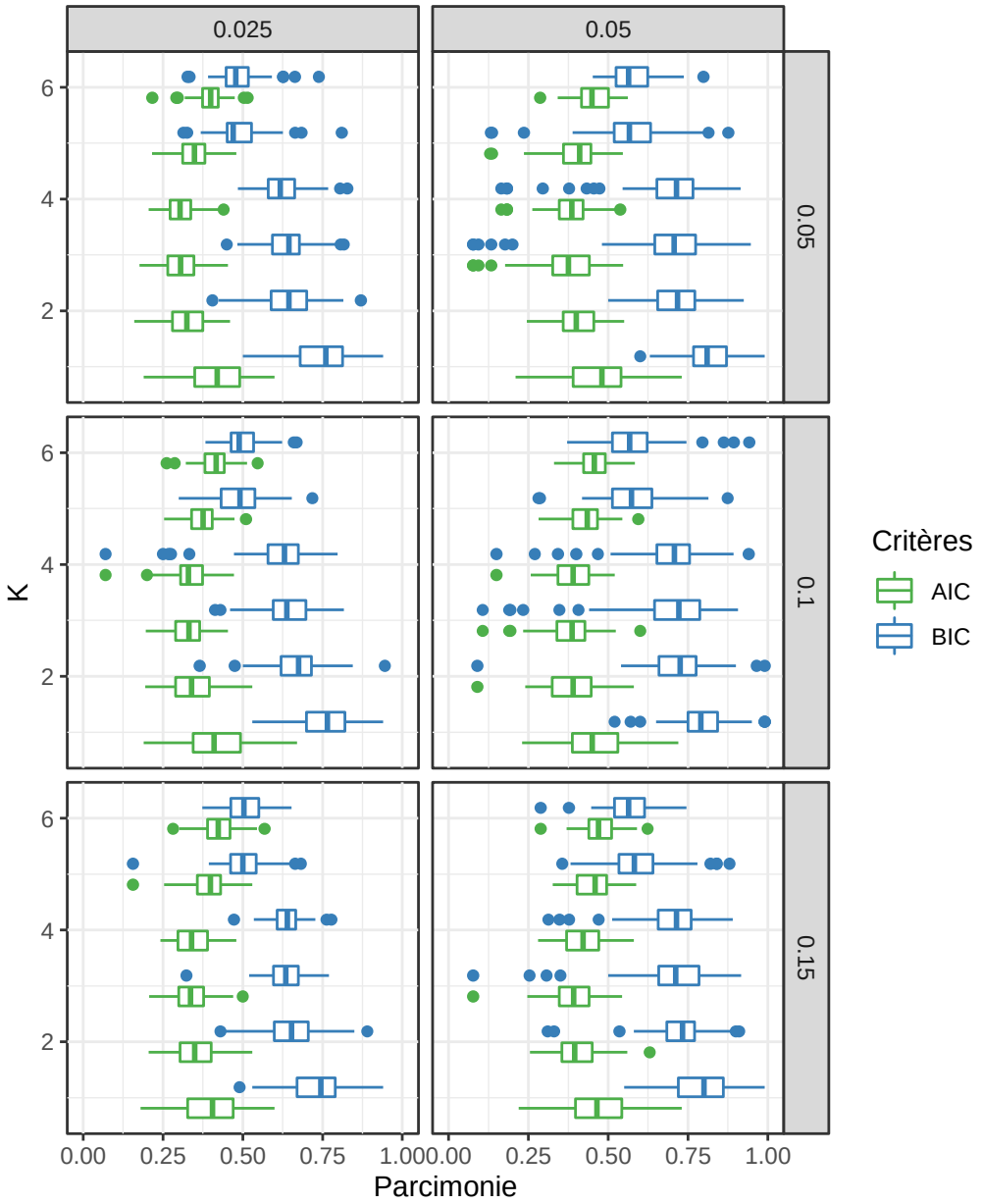


FIGURE 6.17 – Parcimonie obtenue par les modèles sélectionnés par l'AIC et le BIC, en fonction de K, le nombre de composantes, pour $\mathbf{n = 200}$. Les panneaux sont organisés en fonction du chevauchement (verticalement) et de la parcimonie (horizontalement).

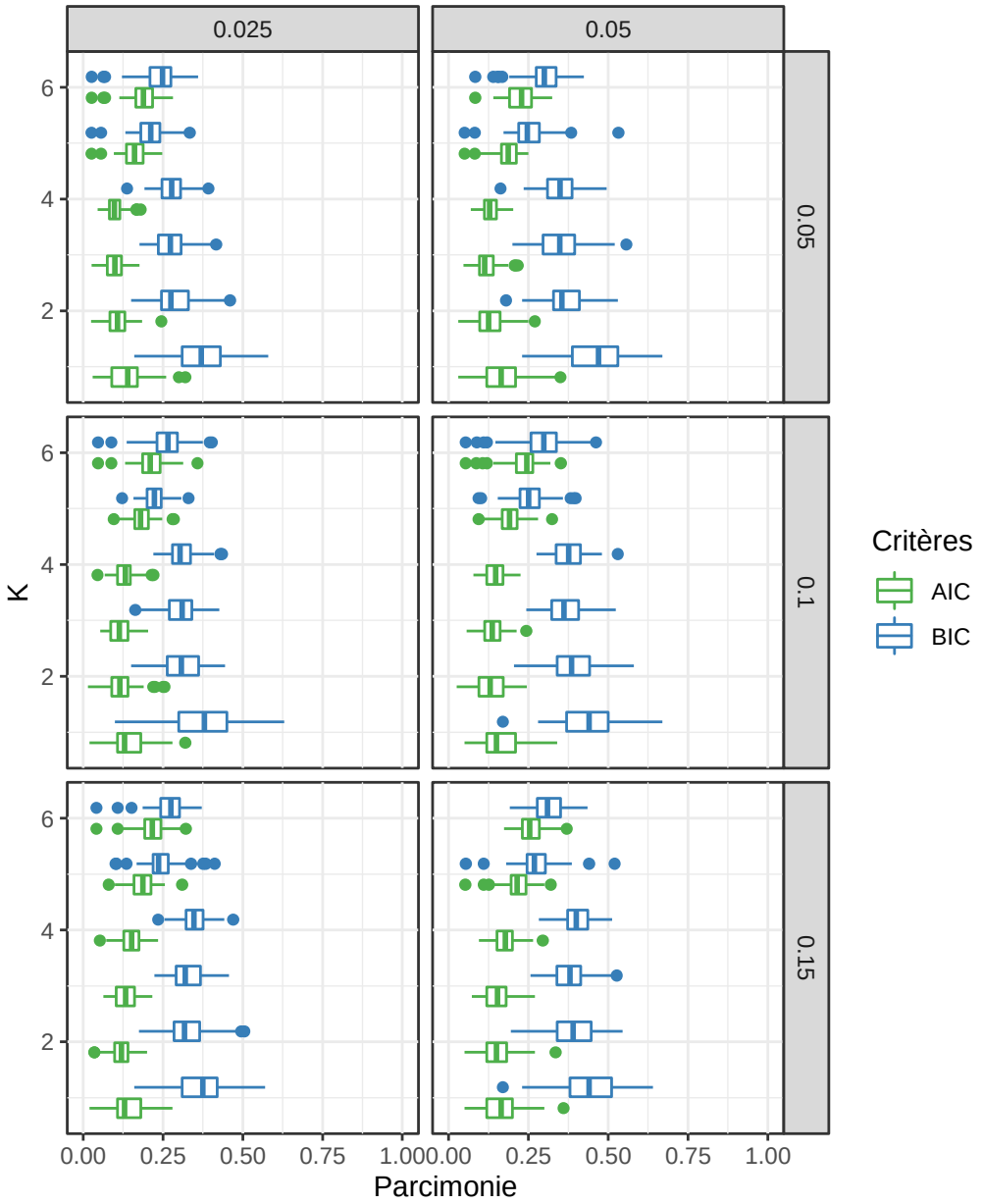


Figure 6.18 – Parcimonie obtenue par les modèles sélectionnés par l'AIC et le BIC, en fonction de K, le nombre de composantes, pour $\mathbf{n = 1000}$. Les panneaux sont organisés en fonction du chevauchement (verticalement) et de la parcimonie (horizontalement).

Enfin, la figure 6.19 montre la précision et le rappel des modèles AIC et BIC optimaux pour 100 ensembles de données avec $n = 1000$ et $d = 100$. Ces valeurs sont mesurées en comparant la classification des coordonnées des moyennes directionnelles en deux classes (composantes nulles et non nulles) avec la classification réelle induite par la régularisation des composantes directionnelles pendant la génération des données artificielles (remarquons que cela n'a de sens que lorsque $K = K^*$). La faible valeur de la précision confirme la tendance des deux critères à sélectionner des représentations trop parcimonieuses. Sur un ensemble de données suffisamment grand, le BIC a un rappel significativement meilleur que l'AIC, mais avec une perte significative en précision. Les résultats pour des ensembles de données plus petits ont tendance à être moins bons en précision et plus ou moins équivalents en rappel.

6.4.5.1 Conclusion

En résumé, la stratégie de suivi de chemin est un moyen efficace d'explorer la régularisation des solutions. Dans le cas d'un faible nombre d'observations, l'utilisation de la régularisation permet de sélectionner un nombre optimal de composantes en utilisant le BIC. Cependant, dans ce régime, elle tend également à sélectionner des moyennes directionnelles trop parcimonieuses par rapport aux vrais paramètres. En raison du grand nombre de paramètres et de la dimension élevée des données considérées, cela n'est pas surprenant, mais les expériences montrent qu'il faut être prudent lors de l'utilisation de ce type de modèle (régularisé ou non).

Nous n'avons pas inclus dans cette section les résultats obtenus pour les autres critères d'information rappelés dans la **section 6.3.4**. Sur des données simulées, ils sont uniformément moins performants que l'AIC et le BIC dans le cas d'un petit nombre d'observations ($n = 200$ pour $d = 100$ par exemple) et à peu près identiques au BIC dans le cas d'un grand nombre d'observations ($n = 1000$). Nous étudions leur pertinence pratique sur des données du monde réel dans la section suivante.

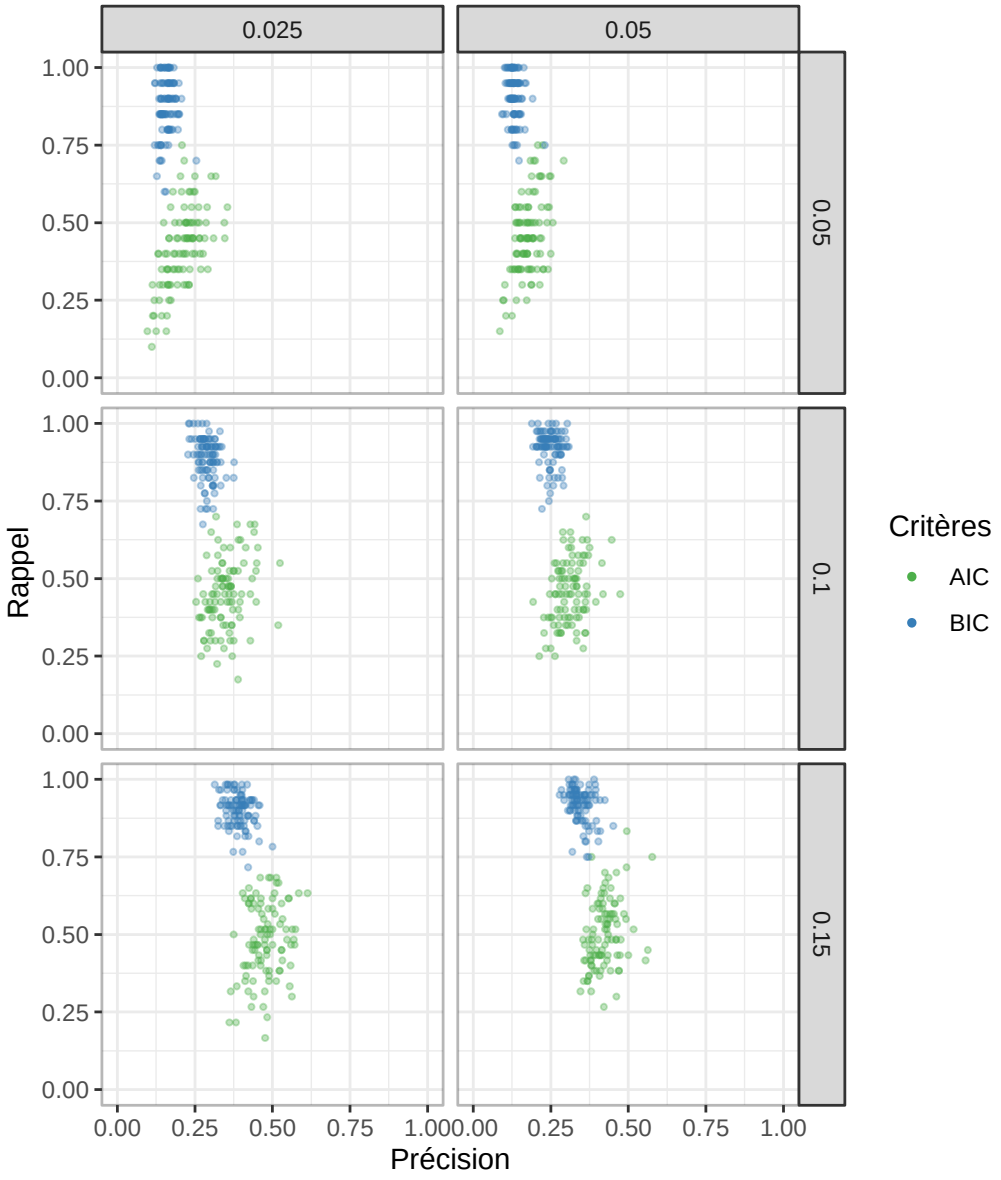


Figure 6.19 – Précision et rappel des composantes mis à zéro des moyennes directionnelles dans les modèles spartiates optimaux selon l'AIC et le BIC pour $K = K^* = 4$ pour $n = \mathbf{1000}$ et $d = 100$. Les panneaux sont organisés en fonction du chevauchement (verticalement) et de la parcimonie (horizontalement).

6.5 Expériences sur des données du monde réel

Nous testons notre algorithme sur deux ensembles de données du monde réel, le jeu de données populaire CSTR et un nouvel ensemble de données basé sur les rapports 8-K de Wells Fargo pour les années 2015-2019.

6.5.1 Computer Science Technical Reports (CSTR)

Le jeu de données CSTR, proposé dans [74] [2], est un bon exemple d'un ensemble de données de dimension assez élevée mais de petite taille avec $n = 475$ échantillons en dimension $d = 1000$. Il a été produit à partir d'une sélection de 475 résumés de rapports techniques [3] publiés par le département d'informatique de l'université de Rochester entre 1991 et 2002. Les rapports sont représentés sur un dictionnaire non divulgué de 1000 mots, avec un encodage binaire (un mot est présent ou non dans un résumé). Sur la base des domaines de recherche développés par le département d'informatique au moment de la collecte, les résumés sont regroupés en $K = 4$ classes (Traitement du langage naturel, Robotique/Vision, Systèmes, et Théorie).

Les classes de l'ensemble de données CSTR ne sont pas des clusters comme le montre une expérience simple : en utilisant comme partition initiale les vraies classes, une application de l'algorithme standard des k-means sphériques [58] conduit à une partition différente après convergence. L'*Adjusted Rand Index* (ARI) entre les deux partitions est de $0,835$. Comme le montre la matrice de confusion entre les deux partitions, disponible dans la table 6.7, deux des classes sont quelque peu difficiles à retrouver du point de vue du partitionnement.

	1	2	3	4
1	71	26	3	1
2	0	70	1	0
3	0	1	176	1
4	0	2	5	118

TABLE 6.7 – Matrice de confusion entre les classes de l'ensemble de données CSTR (en ligne) et les classes obtenues par les k-means sphériques (en colonne).

Le comportement du mélange de distributions von Mises-Fisher sur CSTR est similaire à celui du k-means sphérique. En utilisant la même initialisation, nous obtenons après convergence un ARI de $0,818$ avec des composants κ spécifiques et de $0,837$ avec un κ commun. Les matrices de confusions (omises) sont presque identiques à celles des k-means sphériques.

2. Disponible par exemple ici à cette URL https://github.com/dbmovMFs/DirecCoclus/tree/master/Data

3. Les rapports peuvent être téléchargés sur le site Web du département https://www.cs.rochester.edu/research/technical_reports.html.

En conséquence de la petite taille de l'ensemble de données par rapport à sa dimension, le regroupement final obtenu par différentes méthodes à partir d'une initialisation aléatoire tend à être beaucoup plus dépendant de cette configuration initiale que dans le cas d'un ensemble de données plus simple (comme par exemple dans les expériences de données artificielles rapportées ci-dessus). Pour fournir des résultats significatifs, nous procédons comme suit. Pour chaque algorithme, nous utilisons un ensemble commun de 50 configurations initiales aléatoires. Après convergence d'un algorithme donné, nous conservons la meilleure configuration en termes de critère de qualité de cet algorithme (par exemple, la vraisemblance pour les modèles de mélange) et rapportons l'ARI du clustering correspondant. Nous répétons cette procédure 50 fois (en considérant donc 250 configurations initiales aléatoires) pour évaluer la variabilité des résultats.

La figure 6.20 et la table 6.8 résument les résultats obtenus par le k-means sphérique et les deux variantes des mélanges de vMF. Le mélange avec des paramètres spécifiques de concentration pour chaque composante présente de loin la plus grande variabilité et les plus mauvais résultats. Les effets négatifs d'une valeur trop élevée du paramètre de concentration sur des données réelles ont déjà été établis, par exemple, dans [57, 101]. Pour autant que nous le sachions, la très forte sensibilité des résultats à la configuration initiale, dans ce cas, ne l'était pas. Ces deux problèmes sont résolus par l'utilisation d'un paramètre de concentration partagé. La variabilité des résultats est alors plus faible que celle observée pour le k-means sphérique et sur la configuration optimale avec $K = 4$, les résultats sont à peu près identiques. En particulier, un t-test apparié ne montre pas de différences significatives à un niveau de 1% entre le k-means sphérique et le mélange de vMF avec un κ commun pour $K \in \{3, 4, 5, 6\}$.

	SK-means		Shared κ		Free κ		Co-clustering	
K	moyenne	sd	moyenne	sd	moyenne	sd	moyenne	sd
2	0.471	$4.52\,10^{-3}$	0.344	$2.95\,10^{-3}$	0.395	$3.28\,10^{-4}$	0.442	$6.28\,10^{-2}$
3	0.757	$1.31\,10^{-2}$	0.756	$3.83\,10^{-3}$	0.567	$1.89\,10^{-2}$	0.772	$7.35\,10^{-3}$
4	0.802	$1.77\,10^{-2}$	0.804	$1.22\,10^{-2}$	0.519	$4.48\,10^{-2}$	0.803	$1.72\,10^{-2}$
5	0.659	$4.05\,10^{-2}$	0.650	$2.11\,10^{-2}$	0.497	$8.78\,10^{-2}$	0.716	$4.06\,10^{-2}$
6	0.572	$4.59\,10^{-2}$	0.569	$2.24\,10^{-2}$	0.520	$9.51\,10^{-2}$	0.663	$4.15\,10^{-2}$
7	0.535	$5.21\,10^{-2}$	0.493	$2.82\,10^{-2}$	0.463	$8.28\,10^{-2}$	0.625	$5.18\,10^{-2}$
8	0.481	$4.99\,10^{-2}$	0.448	$3.13\,10^{-2}$	0.441	$7.37\,10^{-2}$	0.588	$6.16\,10^{-2}$

Table 6.8 – *Adjusted Rand Index* entre les classes de CSTR et les clusters obtenus par les modèles étudiés.

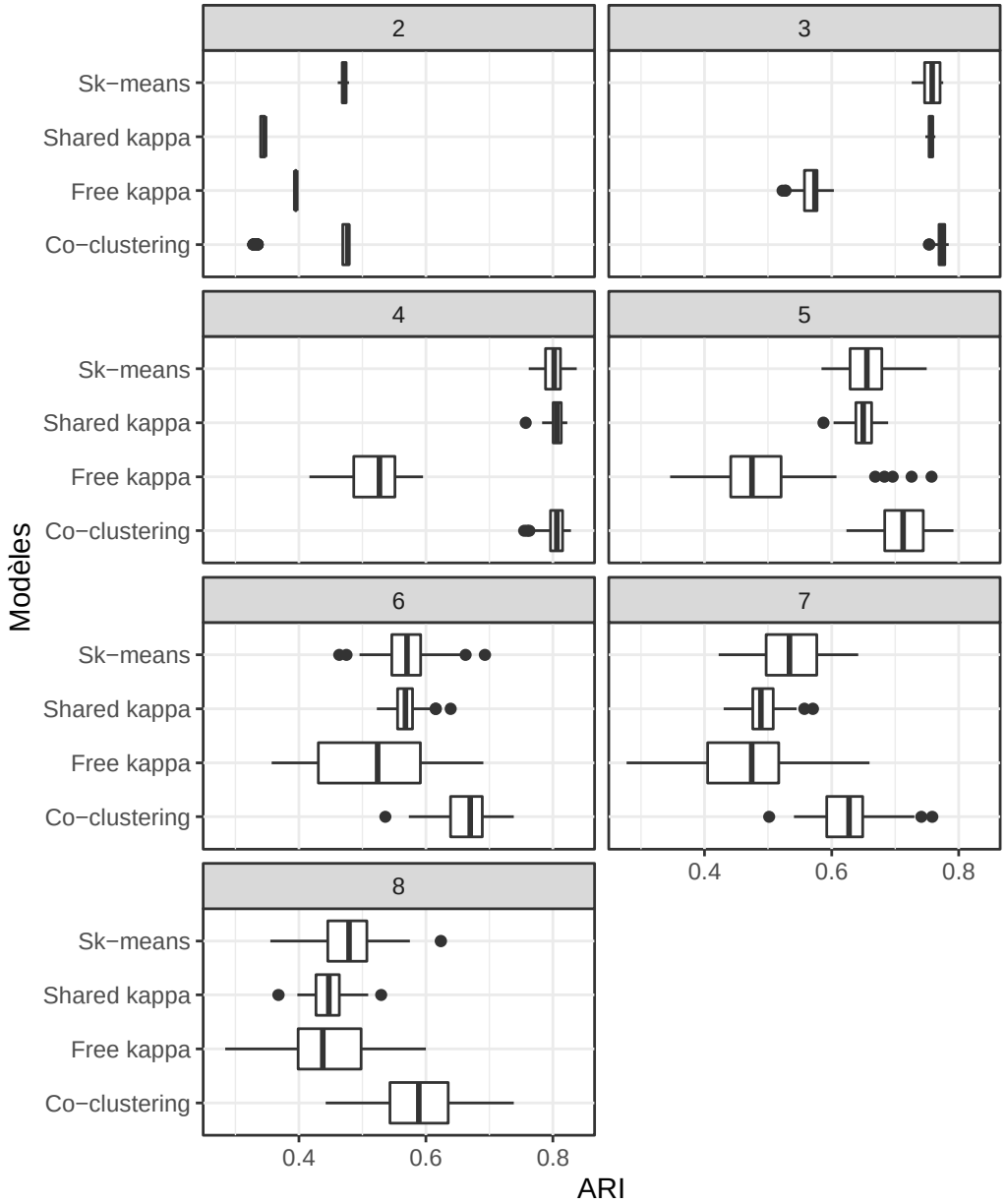


FIGURE 6.20 – *Adjusted Rand Index* entre les classes de CSTR et les clusters obtenus par k-means sphérique (Sk-means), mélange de distributions vMF avec un paramètre κ commun (shared kappa) et mélange de distributions vMF avec des κ spécifiques aux composants (free kappa), et co-clustering pour différentes valeurs de K.

Comme le montre [101], une approche de co-clustering du mélange de von Mises-Fisher, qui impose une structure diagonale sur les moyennes directionnelles, est également un moyen efficace de contrôler les effets négatifs des paramètres de concentration. Alors que les résultats pour $K = 4$ sont identiques à ceux obtenus par les autres méthodes, le modèle de co-clustering est beaucoup plus robuste face à une mauvaise spécification du nombre de composantes. A part pour $K = 2$ où les k-means sphériques fournissent le meilleur ARI (différence significative au niveau 1%), dans toutes les autres configurations avec $K \neq 4$, l'ARI obtenu par le co-clustering est significativement plus grand que ceux obtenus par les autres méthodes.

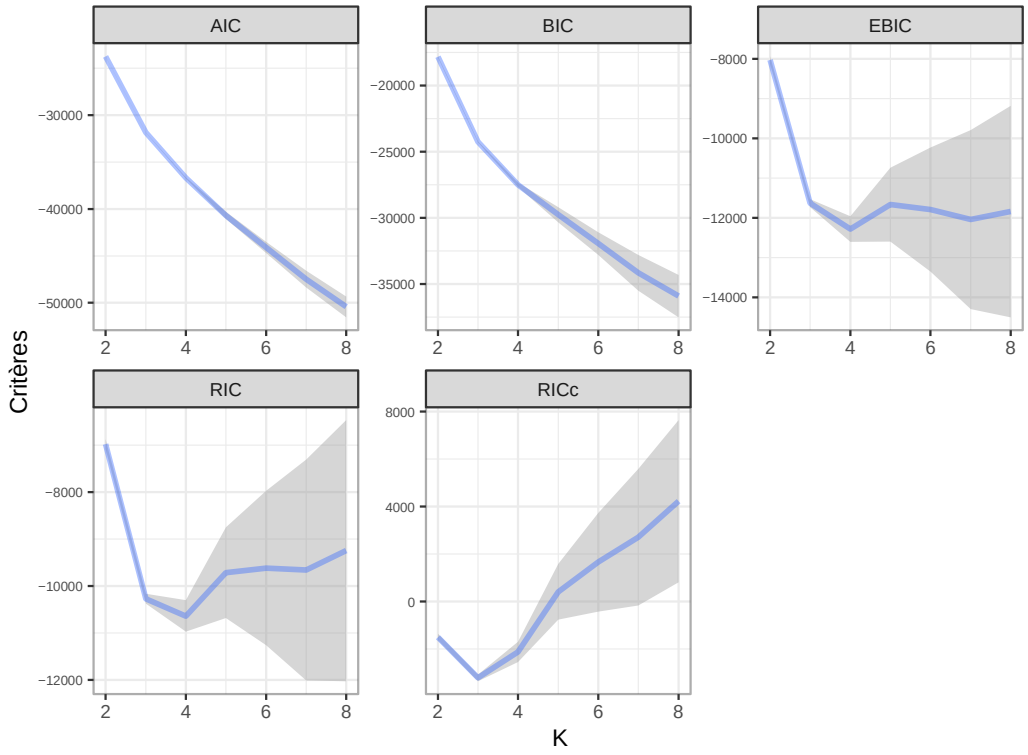


Figure 6.21 – Critères de sélection de modèles pour le mélange de distributions vMF avec un paramètre κ commun : la courbe bleue représente la valeur moyenne, tandis que l'enveloppe grise affiche un intervalle de 2 écarts types autour de celle-ci.

Les figures 6.21 et 6.22 montrent le comportement des critères de sélection du modèle pour le mélange de von Mises-Fisher avec un κ commun et pour l'approche de co-clustering. Ils montrent des comportements assez différents. Pour la distribution de von Mises-Fisher, des critères fortement pénalisés doivent être utilisés pour récupérer les meilleurs modèles, alors qu'au contraire, le petit nombre de paramètres de l'approche de co-clustering conduit à un meilleur comportement de l'AIC.

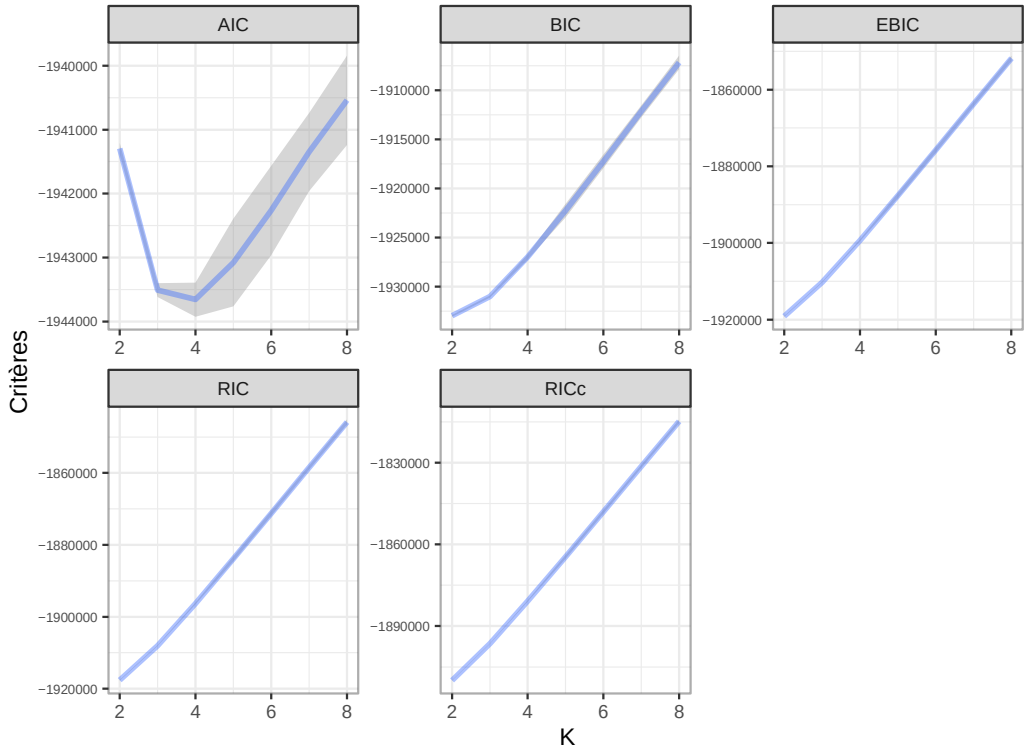


FIGURE 6.22 – Critères de sélection pour le modèle de co-clustering : la courbe bleue représente la valeur moyenne, tandis que l'enveloppe grise affiche un intervalle de 2 écarts types autour de celle-ci.

Nous calculons maintenant le chemin de β pour chacune des 50 réplications de notre procédure, en partant à chaque fois de la meilleure initialisation obtenue à partir des 50 configurations initiales aléatoires. Nous nous limitons au modèle κ partagé. La figure 6.23 résume les résultats. En termes de sélection de modèles parcimonieux, l'AIC, le BIC et l'EBIC fournissent de bons compromis entre l'ARI et la parcimonie. Le RIC et les RICs sélectionnent un modèle trop parcimonieux.

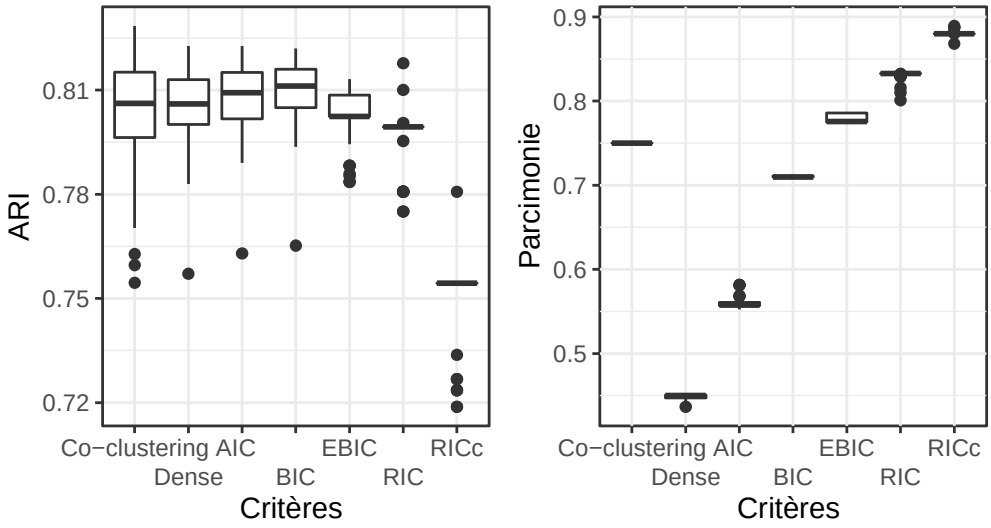


Figure 6.23 – *Adjusted rand index* et la parcimonie pour les modèles sélectionnés sur le chemin β en utilisant les différents critères de sélection des modèles. La configuration *dense* correspond à la solution obtenue sans régularisation. Les résultats du co-clustering sont donnés à titre de référence.

Un point très important est qu'aucun des critères n'est en mesure de fournir une sélection tout-en-un. En effet, comme le montre la figure 6.21, le nombre de composantes devrait être sélectionné avec l'EBIC ou le RIC (et éventuellement avec le RICc), car l'AIC et le BIC diminuent de façon monotone avec le nombre de composantes. Cependant, si l'on calcule le chemin de β pour différents nombres de composantes et que l'on garde comme modèle sélectionné ceux qui minimisent chaque critère, ce comportement s'applique à tous les critères.

En d'autres termes, la régularisation compense l'augmentation du nombre de composantes. Ainsi, il faut d'abord sélectionner le nombre de composantes en fonction de l'EBIC ou du RIC, puis sélectionner le niveau de parcimonie avec le BIC ou l'EBIC, en gardant le nombre de composantes fixe.

Finalement, nous illustrons l'intérêt d'obtenir des moyennes directionnelles parcimonieuses (nous restreignons l'illustration à $K = 4$). Pour faciliter l'interprétation des résultats du clustering, nous suivons [101, 100] et réordonnons les dimensions des données. Le réordonnancement est basé sur les coordonnées non nulles partagées entre les moyennes directionnelles : la partie gauche de la représentation rassemble les coordonnées non nulles pour toutes les composantes/clusters tandis que la partie droite cor-

respond aux coordonnées isolées. Notons qu'en analyse de texte, cela correspond à un vocabulaire partagé par opposition à un vocabulaire spécifique. Chaque ligne de la représentation correspond à une composante et a une hauteur proportionnelle à la taille du cluster correspondant. La valeur réelle de chaque coordonnée est représentée par la valeur grise du rectangle correspondant.

La figure 6.24 représente la structure en blocs obtenue par l'algorithme de co-clustering de [101]. Il s'agit d'un modèle très brut qui favorise la parcimonie au détriment de la révélation de coordonnées partagées et d'une structure plus fine.

FIGURE 6.24 – Représentation des moyennes directionnelles obtenues par l'algorithme de co-clustering sur le jeu de données CSTR.

La figure 6.25 montre la structure des moyennes directionnelles pour le mélange de vMF obtenu sans régularisation. Elle confirme qu'il existe bien des coordonnées spécifiques, mais elle montre que les clusters partagent des dimensions dans une grande proportion, ce qui confirme que la solution de co-clustering cache la majeure partie de la structure.

FIGURE 6.25 – Représentation des moyennes directionnelles obtenues par le mélange de vMF avec un κ commun sur l'ensemble de données CSTR.

Enfin, la figure 6.26 représente les moyennes directionnelles obtenues en sélectionnant le meilleur modèle parcimonieux avec le BIC le long du chemin de β. Le résultat

est un compromis entre la structure strictement diagonale obtenue par l'algorithme de co-clustering et la solution plus dense obtenue sans régularisation. Il isole mieux les dimensions/vocabulaires spécifiques tout en conservant un sous-ensemble plus petit de dimensions partagées. Ceci est confirmé par les figures 6.27 et 6.28 qui montrent le réordonnancement de l'ensemble des données de la même manière que les moyennes directionnelles. Le réordonnancement induit par le mélange parcimonieux de von Mises-Fisher révèle de manière plus claire la structure sous-jacente des données. Nous verrons dans la **section 6.5.2** comment tirer parti de cette représentation pour explorer un ensemble de données.

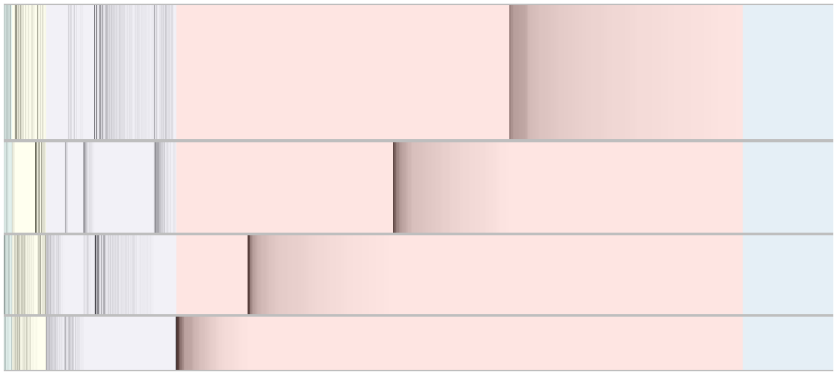


Figure 6.26 – Représentation des moyennes directionnelles obtenues par le mélange de vMF parcimonieux avec un κ commun sur l'ensemble de données CSTR.

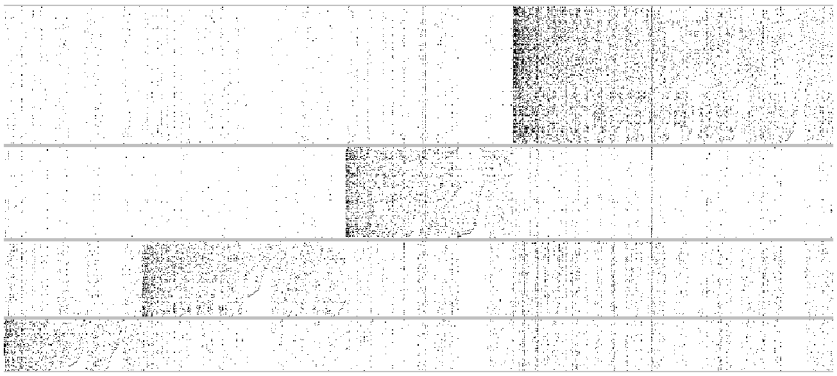


Figure 6.27 – Représentation de l'ensemble de données CSTR réorganisé comme les moyennes directionnelles obtenues par l'algorithme de co-clustering.

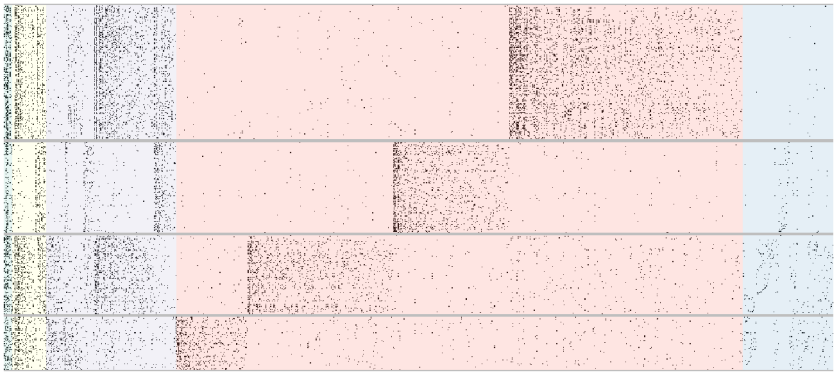


FIGURE 6.28 – Représentation de l'ensemble de données CSTR réorganisé comme les moyennes directionnelles obtenues par le mélange de vMF parcimonieux avec un κ partagé.

6.5.2 Analyse exploratoire des rapports 8-K de l'entreprise Wells Fargo pour les années 2015 - 2019

Le corpus, présenté dans la **section 3.4** (p. 24), de toutes les entreprises du $S\&P500$ pour les années 2015 - 2019 contient 37238 rapports émis par 592 entreprises. Les textes ont été pré-traités en appliquant le pipeline présenté dans la **section 3.4.1** (p. 24). Nous nous concentrons dans cette partie à l'entreprise Wells Fargo (WFC) car c'est elle qui a publié le plus au cours de cette période.

Cette entreprise a publié 672 rapports pour les années 2015 à 2019 et sur 25 événements possibles, seuls 7 sont représentés, avec une domination de l'événement *financial statements and exhibits*, ce qui tend à montrer que ces rapports concernent principalement l'état financier de l'entreprise (voir le tableau 6.9 pour les titres des événements et leurs fréquences). Notons que les rapports peuvent partager plusieurs événements. Seuls 4377 mots (racines) sont utilisés dans les rapports et cet ensemble de données est le suivant : $n = 672$ en dimension $d = 4377$.

Code	Type	Frequencies
1	*Financial Statements and Exhibits*	658
2	*Results of Operations and Financial Condition*	24
3	*Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year*	19
4	*Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers*	27
5	*Submission of Matters to a Vote of Security Holders*	5
6	*Other Events*	36
7	*Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics*	2

TABLE 6.9 – Événements Wells Fargo pour les années 2015 à 2019.

Comme le nombre de clusters K est inconnu dans ce cas, nous procédons comme exposé précédemment en utilisant le mélange de von Mises-Fisher avec un paramètre κ commun. Dans un premier temps, nous sélectionnons le nombre de composantes grâce au RICc pour obtenir $K = 14$ comme le montre la figure 6.29.

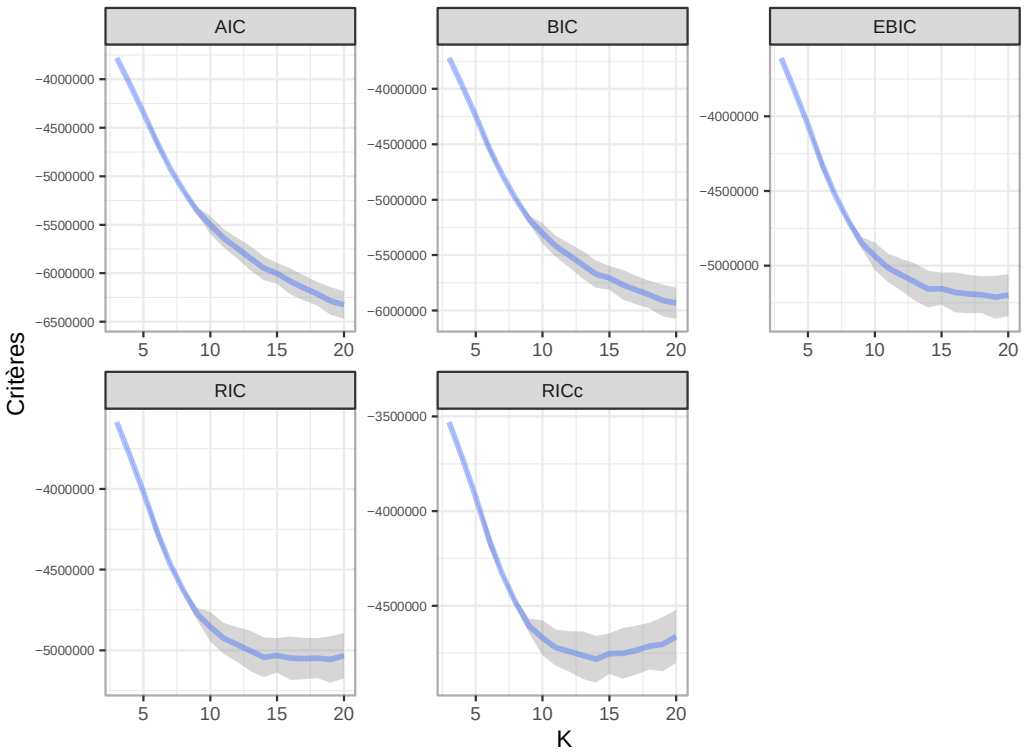


Figure 6.29 – Critères de sélection de modèles pour le mélange de distributions vMF avec un paramètre κ commun concernant l'analyse de Wells Fargo : la courbe bleue représente la valeur moyenne, tandis que l'enveloppe grise affiche un intervalle de 2 écarts types autour de celle-ci.

Dans un deuxième temps, le niveau de parcimonie a été sélectionné en utilisant la stratégie de suivi de chemin avec un maximum de 1000 étapes et l'augmentation relative minimale entre deux valeurs de β fixée à $0,01$. Nous avons obtenu une valeur de β de $1072,253$ et une parcimonie de $82,16\%$. La figure 6.30 représente les moyennes directionnelles. La figure 6.31 présente l'ensemble des données réorganisé comme les moyennes directionnelles. Cette réorganisation révèle de manière claire la structure sous-jacente des données.

La table 6.10 montre la distribution des rapports par cluster obtenue. Nous pouvons noter que le cluster 3 est le plus grand avec 156 rapports tandis que les clusters 1 et 4 sont composés de très peu d'entre eux et doivent être concentrés sur un seul sujet.

Pour sa part, la table 6.11 montre les mots uniques par cluster et ceux qui sont communs. Ces derniers sont logiques dans la mesure où ils contiennent des termes génériques des rapports de l'entreprise, comme sa dénomination ou le nom d'un instrument financier par exemple.

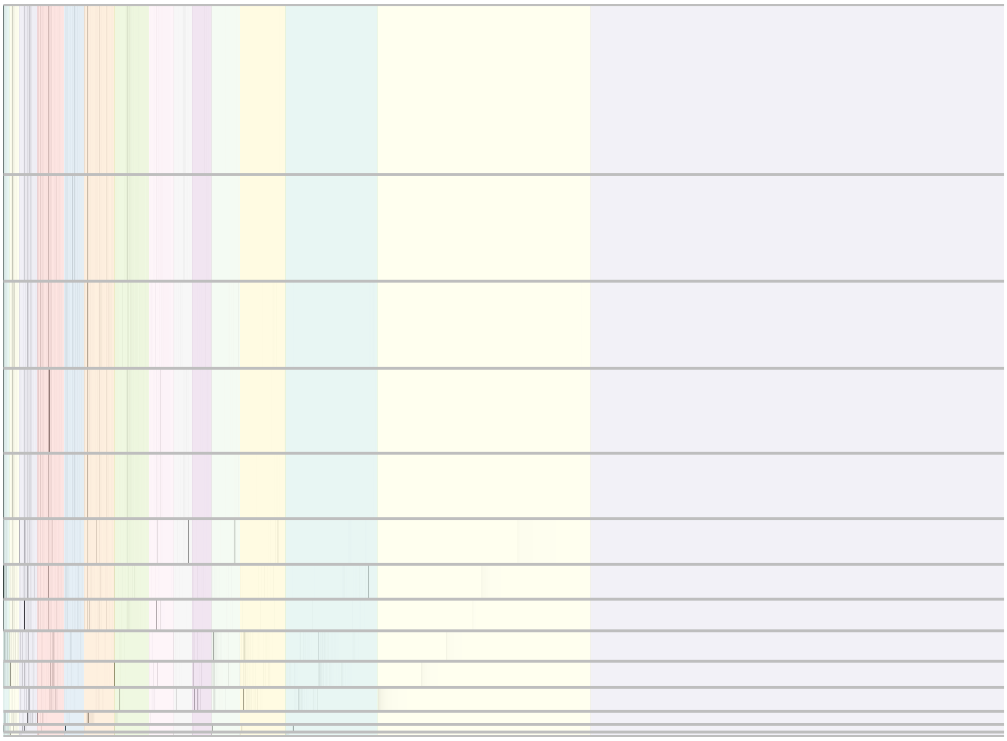


FIGURE 6.30 – Représentation des moyennes directionnelles obtenues par le mélange de vMF parcimonieux avec un κ partagé sur l'ensemble de données Wells Fargo pour les années 2015 à 2019.

	Clusters													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Nb. 8-K	4	80	156	7	12	78	32	60	42	98	24	28	22	29

TABLE 6.10 – Distribution des rapports par cluster obtenue par le modèle parcimonieux sélectionné avec le RICc durant le chemin de β.

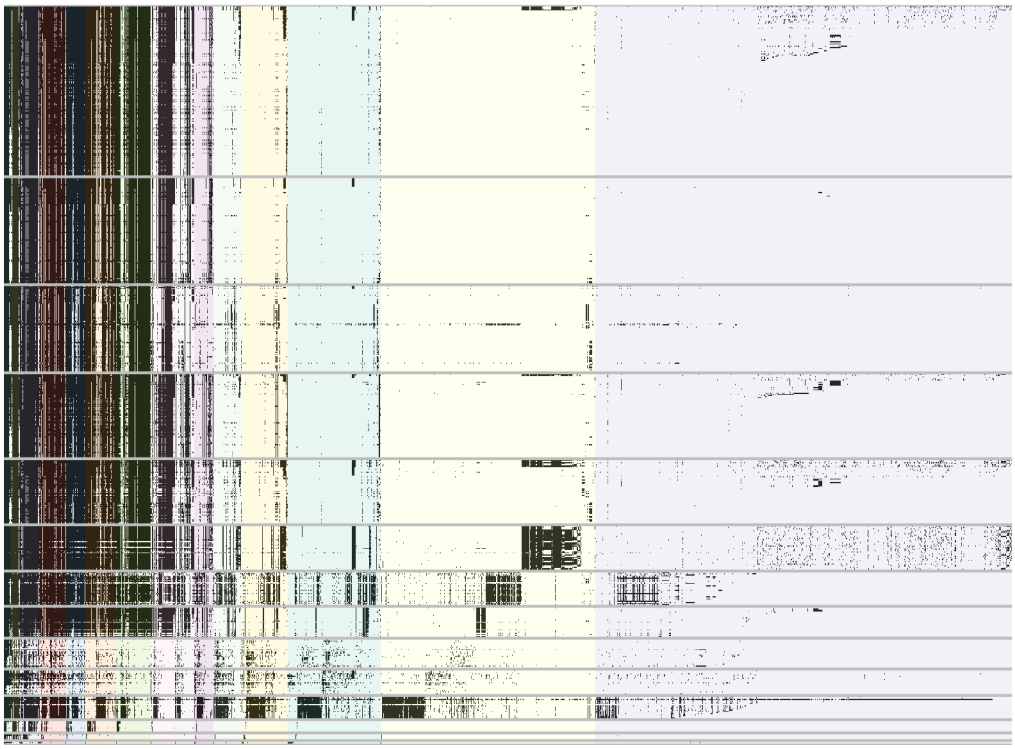


Figure 6.31 – Représentation de l'ensemble de données Wells Fargo pour les années 2015 à 2019 réorganisé comme les moyennes directionnelles obtenues par le mélange de vMF parcimonieux avec un κ partagé.

Plus intéressants sont les mots uniques pour chaque cluster car ils forment des sujets cohérents. Notons l'exception pour les clusters 5 et 10, qui partagent tous les mots de leurs représentants avec au moins un autre cluster.

Par exemple, les mots uniques du cluster 1- *abstention*, *cast*, *ratify*, *shareowner* - sont issus du lexique de l'assemblée générale annuelle. La figure 6.34 montre que ce cluster est entièrement composé de l'événement *Submission of Matters to a Vote of Security Holders* qui a lieu chaque année comme le montre la figure 6.35. La figure 6.32 montre un extrait d'un rapport du Cluster 1 publié par Wells Fargo le 1er mai 2015 [4]. Les mots en *bleu* représentent les mots communs entre tous les clusters et en *rouge*, ceux spécifiques à ce cluster.

Cluster	1	2	3	4	5
1	abstention	cast	ratify	shareowner	-
2	continuance	bankrupt	insolvent	receiver	annually
3	vme	monthly	shewchuk	sonia	cqr
4	advisable	convene	nonassessable	-	-
5	-	-	-	-	
6	sector	bad	homebuilders	gold	miner
7	untrue	omission	canadian	directive	representation
8	adr	absent	determinable	fluctuation	bloomberg
9	domainitemtype	false	thinterestinshareof	shr	text
10	-	-	-	-	-
11	defendant	chair	bonus	rsrs	hear
12	mack	banker	unauthorized	parent	controller
13	portfolio	revenue	offs	sep	jun
14	gics	spin	otc	bulletin	antidilution
commun	security	company	any	well	fargo

TABLE 6.11 – Mots uniques pour chaque cluster obtenus par le modèle parcimonieux sélectionné à l'aide du RICc avec lde chemin de β. La ligne *commun* montre les mots partagés par tous les clusters.

Event : *Submission of Matters to a Vote of Security Holders.*;

Text : [...] *well*s *fargo company* held its annual meeting of stockholders on april 28, 2015. at the meeting, stockholders elected all 16 of the directors nominated by the board of directors as each director received a greater number of votes *cast* for his or her election than votes *cast* [...] *ratify* the appointment of kpmg llp as independent registered public accounting firm for 2015 [...].

FIGURE 6.32 – Exemple d'un rapport 8-K du cluster 1 publié le 1er mai 2015. *Mots* montrent les mots en commun entre tous les clusters et *mots*, ceux spécifiques au cluster 1.

4. Le texte complet est disponible sur https://www.sec.gov/Archives/edgar/data/0000072971/000119312515166149/d920037d8k.htm

Si nous examinons maintenant le cluster 11, qui apparaît de manière aléatoire dans le temps dans la figure 6.35, il est composé des événements *Financial Statements and Exhibits*, *Other Events* et surtout *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers*. Ce groupe se concentre sur les changements au sein du conseil d'administration et leurs conséquences éventuelles sur les résultats de l'entreprise. La figure 6.33 montre un extrait d'un rapport du cluster 11 publié par Wells Fargo le 12 octobre 2016 [5] notifiant le départ du PDG John Stumpf à la suite de nombreux scandales [6]. Il est intéressant de noter que les mots uniques de ce cluster expriment ce contexte. Premièrement, le mot *chair* fait référence à une personne qui siège au conseil d'administration. Ensuite, le terme *defendant* implique une procédure judiciaire. Enfin, les termes *bonus* et *rsrs* [7] mentionnent la compensation due à la rotation des membres du conseil d'administration.

Event : *Departure of Directors or Certain Officers ; Election of Directors ; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers & Financial Statements and Exhibits.* ;

Text : [...] on october 12, 2016, john g. stumpf notified *wells fargo company*) of his decision to retire as chairman and chief executive officer and a director of the *company*, effective immediately. [...] elected director elizabeth a. duke as the *company* s non-executive vice *chair*. [...].

Figure 6.33 – Exemple d'un rapport 8-K du Cluster 11 publié le 12 octobre 2016. *Mots* montrent les mots en commun entre tous les clusters et *mots*, ceux spécifiques au cluster 11.

Concentrons-nous maintenant sur les clusters qui sont constitués du même type d'événement unique et qui n'ont pas de termes uniques, tels que les clusters 5 et 10, comme le montre la figure 6.34.

La figure 6.35 montre que ces clusters apparaissent différemment dans le temps. Le cluster 5 se concentre principalement sur la période précédant la démission du PDG, c'est-à-dire avant octobre 2016, tandis que le cluster 10 se retrouve de manière significative dans deux périodes, c'est-à-dire entre juillet 2015 et mars 2017 mais aussi entre avril 2018 et juillet 2019. Ces deux périodes correspondent à de nombreuses affaires judiciaires mais aussi à des revers commerciaux pour Wells Fargo. Il s'agit notamment d'une forte exposition à la chute des prix du pétrole en janvier 2016 et de nombreux règlements d'amendes pour des pratiques commerciales frauduleuses en avril 2018 et concernant la crise des subprimes en août 2018.

5. Le texte intégral est disponible sur https://www.sec.gov/Archives/edgar/data/0000072971/000119312516736870/d271369d8k.htm

6. Exemple de scandale auquel Wells Fargo a été confronté https://www.cnbc.com/2016/10/20/wells-fargo-just-lost-its-accreditation-with-the-better-business-bureau.html.

7. RSRs est l'acronyme de Restricted Share Rights.

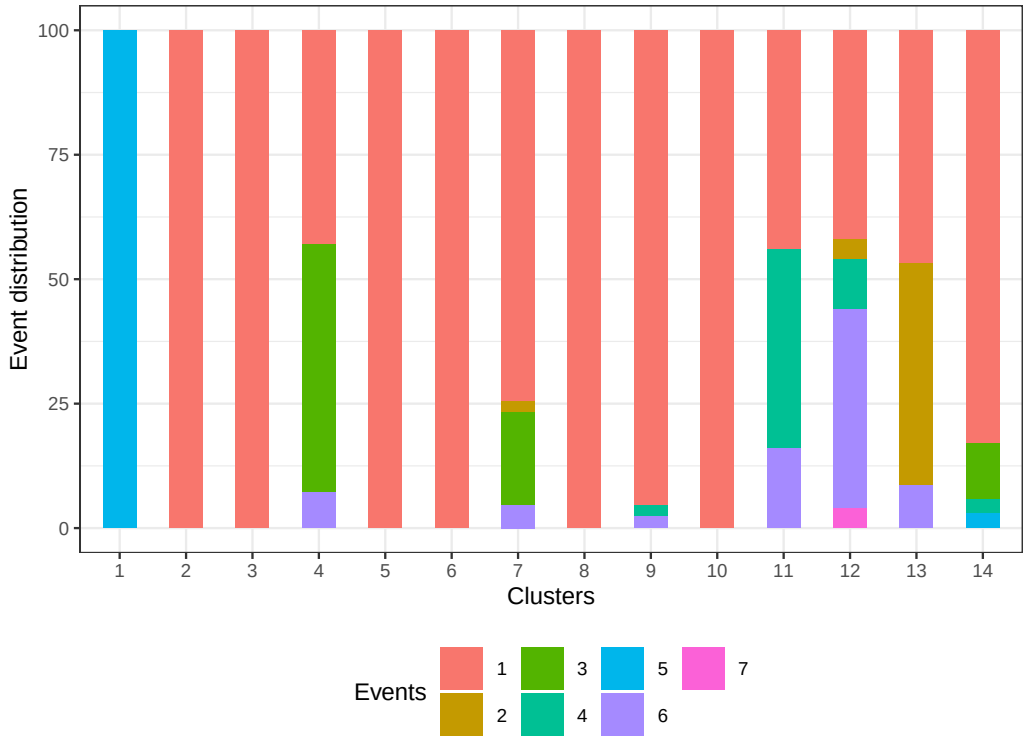


FIGURE 6.34 – Distribution des événements par cluster dans l'ensemble de données Wells Fargo pour les années 2015 à 2019 avec le modèle obtenu par l'approche suivi de chemin.

Une lecture approfondie des textes de ces groupes révèle un sujet commun entre eux, à savoir des billets à moyen terme (*medium-term note* en anglais), mais de séries différentes et de sous-jacents différents. Le groupe 10 est lié aux billets à moyen terme, série K, liés à des indices basés sur les marchés émergents comme le *iShares MSCI Emerging Markets ETF*[8] ou les marchés développés comme le *MSCI EAFE Index*[9]. Le cluster 5 est associé à des notes à moyen terme, série N, liées à des taux de référence [10]. Ces clusters montrent donc que l'entreprise a émis différents types de dette pour faire face à son contexte et assurer ses besoins de financement.

8. L'ETF iShares MSCI Emerging Markets cherche à répliquer les résultats d'investissement d'un indice composé d'actions des marchés émergents à grande et moyenne capitalisation.

9. L'indice MSCI EAEO est conçu pour représenter le rendement de titres de grande et moyenne capitalisation dans 21 marchés développés, y compris des pays d'Europe, d'Australasie et d'Extrême-Orient, à l'exclusion des États-Unis et du Canada.

10. Plus de détails disponibles sur : https://saf.wellsfargoadvisors.com/emx/dctm/Marketing/Marketing_Materials/Fixed_Income_Bonds/e7434.pdf

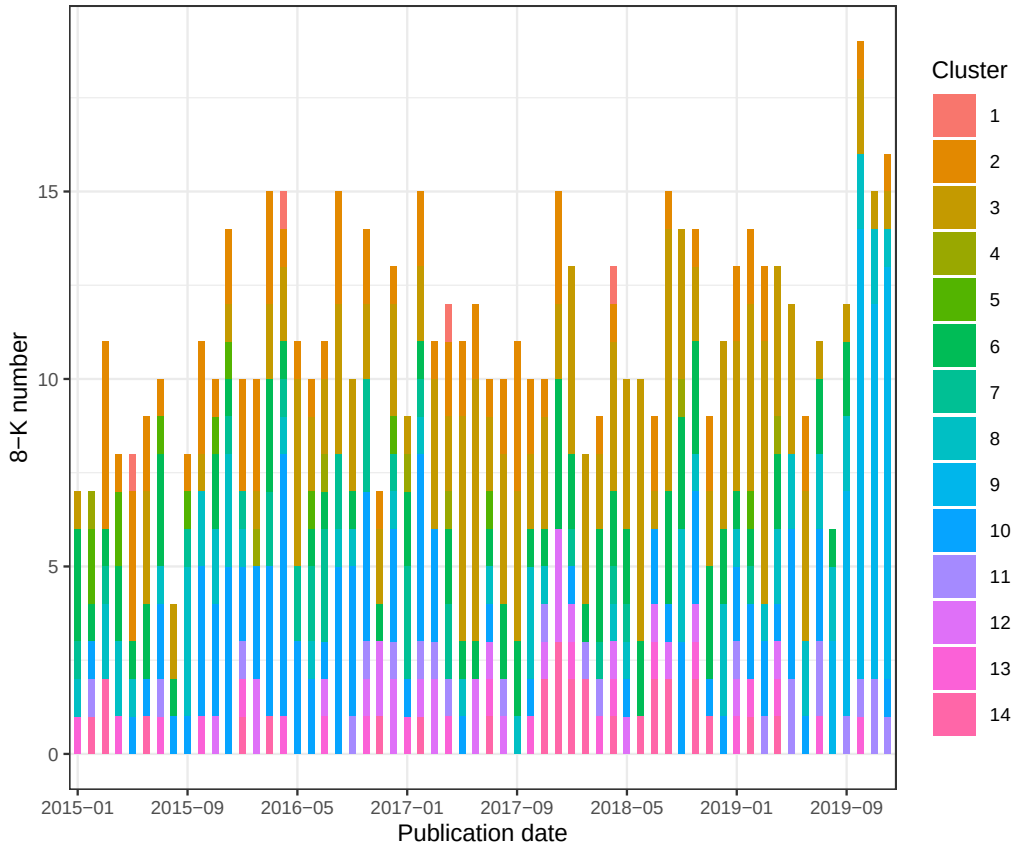


Figure 6.35 – Distribution des clusters par mois dans l'ensemble de données Wells Fargo pour les années 2015 à 2019 avec le modèle obtenu par l'approche suivi de chemin.

Nous comparons maintenant nos résultats avec ceux obtenus par l'approche de co-clustering. Dans ce cas, le nombre de composantes sélectionnées par l'AIC est de $K = 3$, comme le montre la figure 6.36.

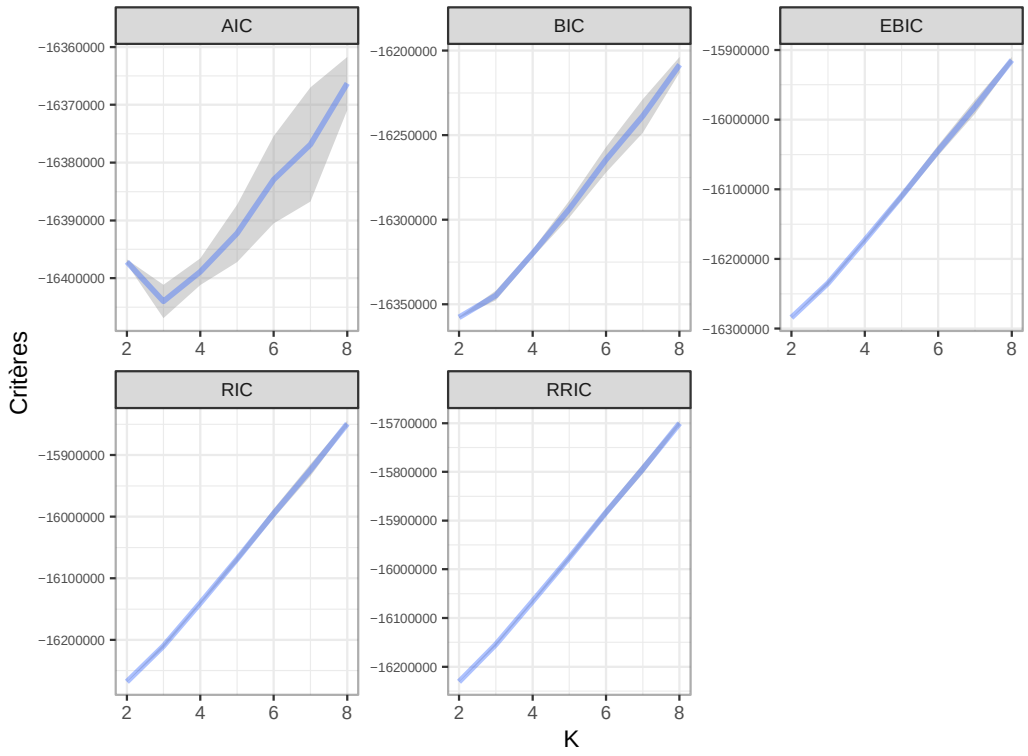


FIGURE 6.36 – Critères de sélection du modèle de co-clustering concernant l'analyse de Wells Fargo : la courbe bleue représente la valeur moyenne, tandis que l'enveloppe grise affiche un intervalle de 2 écarts types autour de celle-ci.

La table 6.12 montre la distribution des rapports par cluster obtenue par le modèle de co-clustering. Nous pouvons noter que les classes sont déséquilibrées et un ARI de $6,62\%$ montre la différence avec le clustering obtenu par notre méthode. La classe 3 contient la grande majorité des rapports.

	Clusters		
	1	2	3
Nb. 8-K	22	53	597

TABLE 6.12 – Distribution des rapports par cluster obtenue par le modèle de co-clustering sélectionné par l'AIC.

La figure 6.37 représente la structure en blocs obtenue par l'algorithme de co-clustering. Comme observé précédemment, la solution de co-clustering cache la plupart de la structure des données par rapport aux résultats obtenus avec notre méthode, comme le montre la figure 6.30.

Figure 6.37 – Représentation des moyennes directionnelles obtenues par l'algorithme de co-clustering sur le jeu de données Wells Fargo.

La figure 6.38 montre la distribution des événements par cluster. Il apparaît que la classe 1 est identique à la classe 13 trouvée par le mélange de von Mises-Fisher. La classe 2 est principalement concernée par des événements spécifiques tels que : *Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers : Compensatory Arrangements of Certain Officers* ou *Submission of Matters to a Vote of Security Holders*. La figure 6.39 montre que cette classe apparaît lorsque l'entreprise a dû faire face à un contexte négatif et a voulu se réorganiser. La classe 3, constituée principalement de l'événement *Financial Statements and Exhibits*, concerne les différentes communications financières de l'entreprise. Cependant, contrairement à l'analyse détaillée possible avec le mélange de von Mises-Fisher, il est très difficile ici de voir les différents aspects de sa communication financière et les produits financiers qu'elle émet.

Enfin, l'analyse précédente montre les avantages de notre méthode par rapport au modèle de co-clustering pour une analyse exploratoire. Elle met en évidence la spécialisation de chacun des clusters qui permet de comprendre facilement les différents événements qui impactent une entreprise dans le temps. De plus, lorsqu'ils existent, les mots uniques à chaque cluster donnent une idée précise du sujet principal de ce cluster. De leur côté, les termes communs à tous les clusters donnent une vue d'ensemble du sujet du corpus.

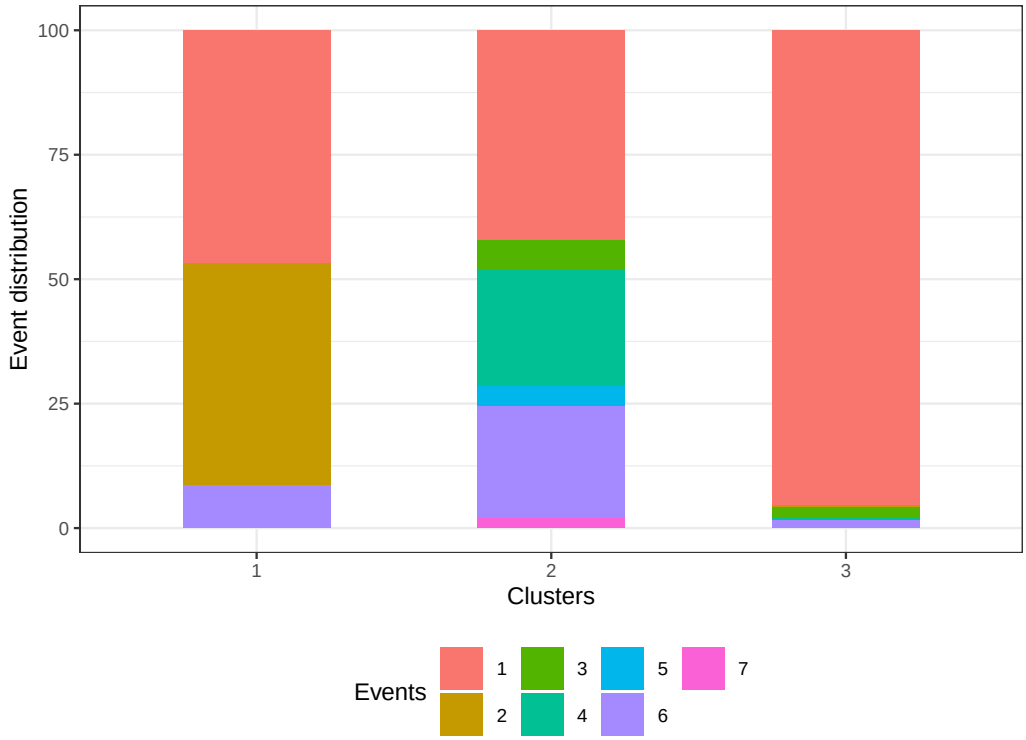


FIGURE 6.38 – Distribution des événements par cluster dans le jeu de données Wells Fargo pour les années 2015 à 2019 obtenu avec le modèle de co-clustering.

6.6 Conclusion

Dans ce chapitre, nous cherchons à pénaliser la vraisemblance d'un mélange de distributions de von Mises-Fisher pour augmenter la parcimonie des moyennes directionnelles. Nous montrons que la pénalisation au moyen de la norme l_1 nous permet d'atteindre notre but en utilisant une adaptation de l'algorithme d'espérance-maximisation ainsi que la combinaison entre l'approche de suivi de chemin et les critères de sélection de modèle pour sélectionner automatiquement le paramètre de pénalisation.

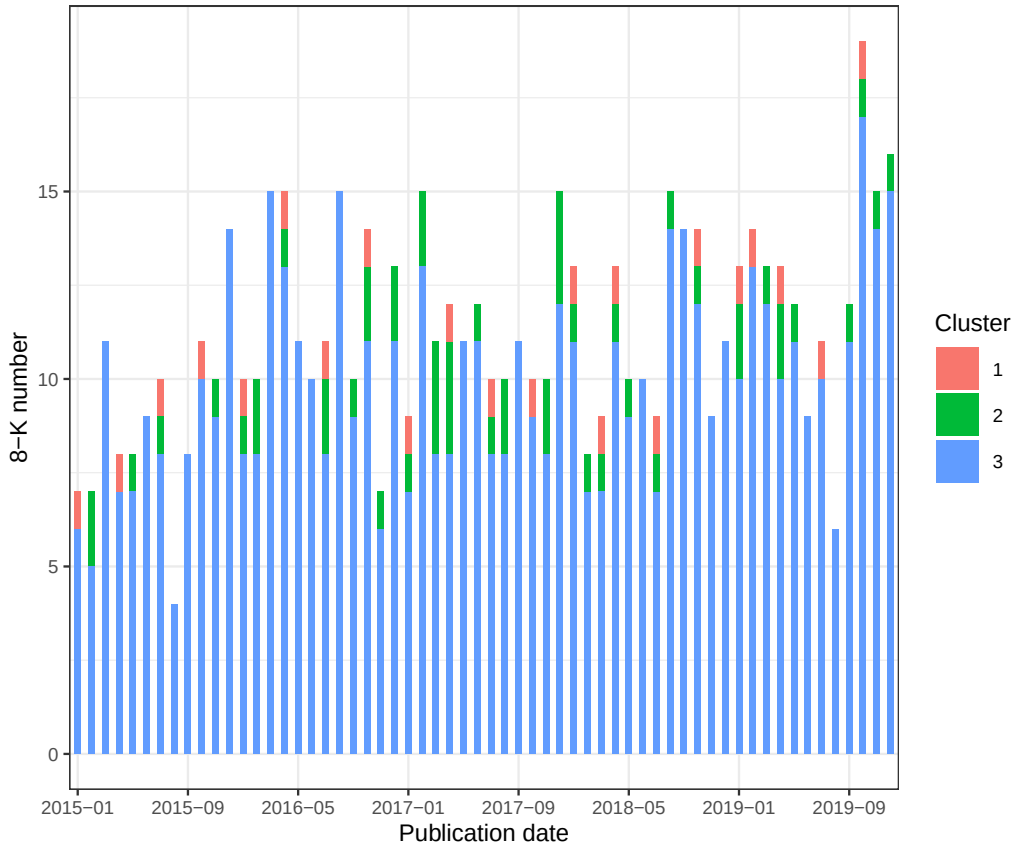


Figure 6.39 – Distribution des clusters par mois dans le jeu de données Wells Fargo pour les années 2015 à 2019 obtenu avec le modèle de co-clustering.

Algorithm 1 EM pour l'estimation de la vraisemblance pénalisée

Require: $\beta \geq 0$ (le paramètre de régularisation)

Require: $\boldsymbol{\Theta}_{init}$ (une valeur d'initialisation facultative pour $\boldsymbol{\Theta}^{(0)}$)

Initialisation de $\boldsymbol{\Theta}^{(0)}$ à $\boldsymbol{\Theta}_{init}$ ou aléatoirement (voir l'algorithme 3 (p. 154))

$m \leftarrow 0$

repeat

$$\tau_{ik}^{(m)} \leftarrow \frac{\alpha_k^{(m)} f_k(\boldsymbol{x}_i, \theta_k^{(m)})}{\sum_{l=1}^{K} \alpha_l^{(m)} f_l(\boldsymbol{x}_i, \theta_l^{(m)})} \qquad \boldsymbol{r}_k^{(m)} \leftarrow \sum_{i=1}^{n} \tau_{ik}^{(m)} \boldsymbol{x}_i$$

$$\alpha_k^{(m+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(m)}$$

$$\kappa_k^{(m+1)} \leftarrow \kappa_k^{(m)}$$

repeat

$$\mu_{kj}^{(m+1)} \leftarrow \frac{\operatorname{sign}\left(r_{kj}^{(m)}\right)}{\sqrt{\sum_{j=1}^{d} (\max(\kappa_k^{(m+1)} |r_{kj}^{(m)}| - \beta, 0))^2}} \max(\kappa_k^{(m+1)} |r_{kj}^{(m)}| - \beta, 0)$$

$$\rho_k \leftarrow \frac{{\boldsymbol{\mu}_k^{(m+1)}}^T \boldsymbol{r}_k^{(m)}}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \qquad \kappa_k^{(m+1)} \leftarrow \frac{d\rho_k - \rho_k^3}{1 - \rho_k^2}$$

until convergence de $\kappa_k^{(m+1)}$ et ${\boldsymbol{\mu_k}}^{(m+1)}$

$m \leftarrow m + 1$

until convergence de $\boldsymbol{\Theta}^{(m+1)}$

Algorithm 2 Suivi de chemin

Require: $P > 0$ (le nombre de β à explorer sur le chemin)

Require: $\epsilon > 0$ (la précision numérique en dessous de laquelle les coordonnées des moyennes directionnelles sont mises à 0)

$\beta_0 \leftarrow 0$

$\mathbf{\Theta}\{0\} \leftarrow EM(\beta_0)$

for $p = 1$ **to** $P - 1$ **do**

$$\beta_p \leftarrow \beta_{p-1} + \min_{h,j,\kappa_k\{\beta_{p-1}\}|r_{kj}\{\beta_{p-1}\}|-\beta_{p-1}>0} \kappa_k\{\beta_{p-1}\} \left| r_{kj}\{\beta_{p-1}\} \right| - \beta_{p-1}$$

$$\mathbf{\Theta}\{\beta_p\} \leftarrow EM(\beta_p, \mathbf{\Theta}\{\beta_{p-1}\})$$

if $|\mu_{kj}| < \epsilon$ **then**

$\mu_{kj} \leftarrow 0$

end if

end for

Chapitre 7

Conclusion

Tout au long de cette thèse, notre fil conducteur a été de rechercher la représentation ainsi que la visualisation les mieux adaptées de textes, pour une meilleure compréhension des résultats obtenus par divers modèles de classification.

Dans le **chapitre 3**, nous avons abordé la construction de la base de données de rapports 8-K pour les entreprises du $S\&P500$ des années 2015 à 2019. Puis, une présentation générale des données a été effectuée. Cette dernière met en évidence la difficulté pour une classification future avec des textes de différentes tailles et des vocabulaires spécifiques à chaque entreprise. Enfin, nous avons effectué une analyse préliminaire des années 2015 à 2016 pour l'entreprise Wells Fargo (WFC) [10], dans laquelle nous avons montré que la représentation de textes pouvait jouer un rôle vital pour une classification.

Dans la continuité de notre recherche et celle de l'article [72], dans le **chapitre 4**, nous souhaitions montrer que des techniques plus complexes de représentations de textes permettraient de battre la représentation unigramme. Cependant, au contraire des résultats obtenus par [72], l'utilisation de ces méthodes, que cela soit pour la représentation ou la classification, ne permet pas une amélioration substantielle. Nous en concluons que la construction syntaxique des textes et l'utilisation de vocabulaire trop spécifique à chaque entreprise en est la cause principale. Une autre explication à ceci pourrait être l'anticipation du marché ou la disponibilité de l'information antérieurement à la publication du rapport.

Nous avons donc voulu comprendre dans le **chapitre 5** d'où pouvait provenir cette disparité avec les résultats de [72] et avons proposé une méthode de *Comparaison de représentations de textes en vue d'une analyse exploratoire* [10]. Dans celle-ci, nous analysons trois représentations, les unigrammes, les *topic models* ainsi qu'une représentation basée sur le transport optimal, en utilisant un algorithme de classification non-supervisée HDDC [21]. De plus, nous proposons une méthode de visualisation adaptée à chaque représentation pour une analyse aisée des résultats et nous montrons que ces méthodes sont une bonne alternative au *topic model*.

En suivant cette logique, dans le **chapitre 6** nous présentons un modèle de mélange de von Mises-Fisher pénalisé à l'aide de la norme l_1. Nous montrons, dans [11, 12], que cet

algorithme est idéal pour les données directionnelles [78], c'est-à-dire lorsque leur corrélation importe plus que leur distance euclidienne. De plus, la pénalisation permet d'obtenir une représentation parcimonieuse de la moyenne directionnelle, et donc à l'image de la figure 6.26 (p. 122), d'obtenir une vision claire des variables ayant un fort pouvoir discriminant ou non. Enfin, la méthode de suivi de chemin, alliée aux critères de sélection de modèles permettent une sélection automatique des paramètres du modèle.

Dans la continuité de ce travail de thèse, de nombreuses pistes s'ouvrent à nous.

Tout d'abord, à la suite du **chapitre 5**, une solution de visualisation interactive pourrait être développée, à l'image de l'outil *LDAvis* [105] qui simplifie d'autant plus l'analyse des résultats.

Puis, nous pourrions améliorer notre méthode de recherche de chemin présentée dans le **chapitre 6** pour notre modèle de mélange en le rendant plus rapide, notamment avec la mise en place de *mini-batch* par exemple. De plus, des recherches préliminaires indiquent que notre méthode de suivi de chemin s'appliquent avec succès au modèle de mélange gaussien pénalisé, proposé dans [87], ce qui laisse à penser que celle-ci pourrait facilement être applicable à d'autres types de modèles.

Enfin, nous présentons dans l'**annexe D** (p. 163) un travail, toujours en cours de réalisation, en collaboration avec la Chaire Gouvernance et Régulation de l'Université Paris Dauphine-PSL : il s'agit, en se basant sur nos recherches précédentes, d'évaluer l'impact des Lobbys dans le processus parlementaire au sein de l'Union Européenne. Pour ce faire, nous proposons une méthode simple pour créer un dictionnaire permettant de cliver le discours politique. Puis, nous entraînons un modèle permettant de prédire l'orientation politique d'un répondant. Ainsi, nous pourrions envisager plusieurs actions. La première serait de mettre en place des algorithmes de classification multi-classes pour avoir une vision complète de l'orientation politique d'un répondant. Dans la deuxième, nous pourrions utiliser le modèle BERT [40] pour tenter d'améliorer nos résultats même si cela nécessiterait l'élargissement de notre base de données. Enfin, la troisième action serait d'implémenter un algorithme d'apprentissage incrémental (*Online machine learning*) à l'aide d'un site internet où nous demanderions à des participants de juger les résultats de l'algorithme.

Ces pistes ne sont bien sûr, pas exhaustives et pourraient peut-être servir de base à d'autres exploitations ou répondre à d'autres interrogations ?

Bibliographie

[1] Astrid Aggelen et al. « The debates of the European Parliament as Linked Open Data ». In : *Semantic Web* 8 (déc. 2016), p. 271-281. doi : 10.3233/SW-160227.

[2] Hirotogu Akaike. « Information Theory and an Extension of the Maximum Likelihood Principle ». In : *Selected Papers of Hirotugu Akaike*. Sous la dir. d'Emanuel Parzen, Kunio Tanabe et Genshiro Kitagawa. New York, NY : Springer New York, 1998. Chap. 4, p. 199-213. doi : 10.1007/978-1-4612-1694-0_15.

[3] Mehdi Allahyari et al. « Text Summarization Techniques: A Brief Survey ». In : *CoRR* abs/1707.02268 (2017). arXiv : 1707.02268.

[4] Dogu Araci. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. 2019. arXiv : 1908.10063 [cs.CL].

[5] Sanjeev Arora, Yingyu Liang et Tengyu Ma. « A Simple but Tough-to-Beat Baseline for Sentence Embeddings ». In : *5th International Conf. on Learning Representations, ICLR 2017, Toulon, France*. 2017. url : https://openreview.net/forum?id=SyK00v5xx.

[6] R. Arun et al. « On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. » In : *PAKDD (1)*. Sous la dir. de Mohammed Javeed Zaki et al. T. 6118. Lecture Notes in Computer Science. Springer, 2010, p. 391-402. doi : 10.1007/978-3-642-13657-3.

[7] Jimmy Lei Ba, Jamie Ryan Kiros et Geoffrey E. Hinton. *Layer Normalization*. 2016. arXiv : 1607.06450 [stat.ML].

[8] Ryan Bakker et al. « Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010 ». In : *Party Politics* 21.1 (2015), p. 143-152. doi : 10.1177/1354068812462931.

[9] Arindam Banerjee et al. « Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions ». In : *J. Mach. Learn. Res.* 6 (déc. 2005), 1345–1382. issn : 1532-4435. url : http://jmlr.org/papers/v6/banerjee05a.html.

[10] Florian Barbaro et Fabrice Rossi. « Comparaison de représentations de textes en vue d'une analyse exploratoire ». In : *Revue des Nouvelles Technologies de l'Information* Extraction et Gestion des Connaissances, RNTI-E-37 (2021), p. 505-506. url : https://editions-rnti.fr/?inprocid=1002690.

[11] Florian Barbaro et Fabrice Rossi. « Pénalisation l 1 pour un mélange de lois de von Mises-Fisher ». In : *JDS 2021*. Nice, France, juin 2021. url : https://hal.archives-ouvertes.fr/hal-03285717.

[12] Florian Barbaro et Fabrice Rossi. « Sparse mixture of von Mises-Fisher distribution ». In : *ESANN 2021 proceedings*. Bruges, Belgium : European Symposium On Artificial Neural Networks, Computational Intelligence et Machine Learning, oct. 2021. doi : 10.14428/esann/2021.ES2021-115.

[13] David P. Baron. « Integrated Market and Nonmarket Strategies in Client and Interest Group Politics ». In : *Business and Politics* 1.1 (1999), 7–34. doi : 10.1515/bap.1999.1.1.7.

[14] David P. Baron. « Private Politics ». In : *Journal of Economics & Management Strategy* 12.1 (2003), p. 31-66. doi : 10.1111/j.1430-9134.2003.00031.x.

[15] Kenneth Benoit et Michael Laver. *Party Policy in Modern Democracies*. T. 19. Oct. 2006. doi : 10.4324/9780203028179.

[16] David M. Blei et Michael I. Jordan. « Variational Methods for the Dirichlet Process ». In : *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada : Association for Computing Machinery, 2004, p. 12. doi : 10.1145/1015330.1015439.

[17] David M. Blei, Andrew Y. Ng et Michael I. Jordan. « Latent Dirichlet Allocation ». In : *J. Mach. Learn. Res.* 3.null (2003), 993–1022. url : https://jmlr.org/papers/v3/blei03a.html.

[18] Oliver Boguth et al. « Horizon effects in average returns: The role of slow information diffusion ». In : *The Review of Financial Studies* 29.8 (2015), p. 2241-2281. doi : 10.2139/ssrn.1787215.

[19] Wafia Parr Bouberima, Mohamed Nadif et Yamina Khemal Bencheikh. « Assessing the Number of Clusters From a Mixture of Von Mises-Fisher ». In : *Proceedings of the World Congress on Engineering (WCE 2010)*. Sous la dir. de S. I. Ao et al. T. III. London (U.K.) : Newswood Limited, juin 2010, p. 2006-2011. url : http://www.iaeng.org/publication/WCE2010/WCE2010_pp2006-2011.pdf.

[20] Florian Boudin et Juan-Manuel Torres-Moreno. « Résumé automatique multi-document et indépendance de la langue: une première évaluation en français ». In : (2009). url : http://talnarchives.atala.org/TALN/TALN-2009/taln-2009-court-035.html.

[21] Charles Bouveyron et Camille Brunet. *Discriminative variable selection for clustering with the sparse Fisher-EM algorithm*. 2012. arXiv : 1204.2067 [stat.ME].

[22] Charles Bouveyron et Camille Brunet. « Simultaneous model-based clustering and visualization in the Fisher discriminative subspace ». In : *Statistics and Computing* 22.1 (jan. 2012), p. 301-324. doi : 10.1007/s11222-011-9249-9.

[23] Charles Bouveyron, Stéphane Girard et Cordelia Schmid. « High-dimensional data clustering ». In : *Computational Statistics & Data Analysis* 52.1 (2007), p. 502-519. doi : 10.1016/j.csda.2007.02.009.

[24] Leo Breiman. « Random Forests ». In : *Machine Learning* 45.1 (2001), p. 5-32. doi : 10.1023/A:1010933404324.

[25] Peter Broeder, G. Extra et R. van Hout. « Richness and variety in the developing lexicon ». English. In : *Adult language acquisition: Cross-linguistic perspectives. Vol. I: Field methods*. Sous la dir. de C. Perdue. United Kingdom : Cambridge University Press, 1993, p. 145-163. url : https://research.tilburguniversity.edu/en/publications/richness-and-variety-in-the-developing-lexicon.

[26] Georgios-Ioannis Brokos, Prodromos Malakasiotis et Ion Androutsopoulos. « Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering ». In : *CoRR* abs/1608.03905 (2016). arXiv : 1608.03905.

[27] Andrei M. Butnaru et Radu Tudor Ionescu. « From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings ». In : *Procedia Computer Science* 112 (2017). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France, p. 1783-1792. doi : 10.1016/j.procs.2017.08.211.

[28] John Y Campbell et Robert J Shiller. « Cointegration and tests of present value models ». In : *Journal of political economy* 95.5 (1987), p. 1062-1088. doi : 10.3386/w1885.

[29] Juan Cao et al. « A density-based method for adaptive LDA model selection. » In : *Neurocomputing* 72.7-9 (2009), p. 1775-1781. doi : 10.1016/j.neucom.2008.06.011.

[30] Francis Castles et PETER MAIR. « Left-Right Political Scales: Some 'Expert' Judgments ». In : *European Journal of Political Research* 12 (mai 2006), p. 73 -88. doi : 10.1111/j.1475-6765.1984.tb00080.x.

[31] Senthil Chandrasegaran et al. « Integrating Visual Analytics Support for Grounded Theory Practice in Qualitative Text Analysis ». In : *Computer Graphics Forum* 36.3 (2017), p. 201-212. doi : 10.1111/cgf.13180.

[32] Jiahua Chen et Zehua Chen. « Extended Bayesian information criteria for model selection with large model spaces ». In : *Biometrika* 95.3 (sept. 2008), p. 759-771. issn : 0006-3444. doi : 10.1093/biomet/asn034.

[33] Jiahua Chen et Zehua Chen. « EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM ». In : *Statistica Sinica* 22.2 (2012), p. 555-574. doi : 10.5705/ss.2010.216.

[34] Jason Chuang, Christopher D. Manning et Jeffrey Heer. « Termite: Visualization Techniques for Assessing Textual Topic Models ». In : *Proc. of the Int. Working Conf. on Adv. Visual Interfaces*. AVI '12. Capri Island, Italy : Association for Computing Machinery, 2012, 74–77. doi : 10.1145/2254556.2254572.

[35] Alexis Conneau et al. « Supervised Learning of Universal Sentence Representations from Natural Language Inference Data ». In : *CoRR* abs/1705.02364 (2017). arXiv : 1705.02364.

[36] Marco Cuturi. « Sinkhorn Distances: Lightspeed Computation of Optimal Transport ». In : *Advances in Neural Information Processing Systems 26*. Curran Ass., Inc., 2013, p. 2292-2300. url : https://papers.nips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

[37] Cedric De Boom et al. « Representation Learning for Very Short Texts Using Weighted Word Embedding Aggregation ». In : *Pattern Recogn. Lett.* 80.C (sept. 2016), p. 150-156. doi : 10.1016/j.patrec.2016.06.012.

[38] Enric Junqué De Fortuny et al. « Evaluating and understanding text-based stock price prediction models ». In : *Information Processing & Management* 50.2 (2014), p. 426-441. doi : 10.1016/j.ipm.2013.12.002.

[39] Romain Deveaud, Eric Sanjuan et Patrice Bellot. « Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval ». In : *Document Numérique* (juin 2014), p. 61-84. doi : 10.3166/DN.17.1.61-84.

[40] Jacob Devlin et al. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *Proc. of the 2019 Conf. of the North American Chapter of the Ass. for Comp. Ling.: Hum. Lang. Technologies*. Juin 2019, p. 4171-4186. doi : 10.18653/v1/N19-1423.

[41] Inderjit S. Dhillon et Dharmendra S. Modha. « Concept Decompositions for Large Sparse Text Data Using Clustering ». In : *Machine Learning* 42.1 (2001), p. 143-175. issn : 1573-0565. doi : 10.1023/A:1007612920971.

[42] Thomas G. Dietterich, Richard H. Lathrop et Tomás Lozano-Pérez. « Solving the Multiple Instance Problem with Axis-Parallel Rectangles ». In : *Artif. Intell.* 89.1–2 (jan. 1997), 31–71. doi : 10.1016/S0004-3702(96)00034-3.

[43] Günes Erkan et Dragomir R. Radev. « LexRank: Graph-based Lexical Centrality As Salience in Text Summarization ». In : *J. Artif. Int. Res.* 22.1 (déc. 2004), p. 457-479. doi : 10.1613/jair.1523.

[44] Dean P. Foster et Edward I. George. « The Risk Inflation Criterion for Multiple Regression ». In : *The Annals of Statistics* 22.4 (1994), p. 1947 -1975. doi : 10.1214/aos/1176325766.

[45] Mahak Gambhir et Vishal Gupta. « Recent Automatic Text Summarization Techniques: A Survey ». In : *Artif. Intell. Rev.* 47.1 (jan. 2017), p. 1-66. doi : 10.1007/s10462-016-9475-9.

[46] S. Geman et D. Geman. « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images ». In : *IEEE Trans. on Pattern Anal. and M. Intel.* PAMI-6.6 (1984), p. 721-741. doi : 10.1109/TPAMI.1984.4767596.

[47] Ian J. Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA : MIT Press, 2016.

[48] Siddharth Gopal et Yiming Yang. « Von Mises-Fisher Clustering Models ». In : *Proceedings of the 31st International Conference on Machine Learning*. Sous la dir. d'Eric P. Xing et Tony Jebara. T. 32. Proceedings of Machine Learning Research. Bejing, China : PMLR, 2014, p. 154-162. url : http://proceedings.mlr.press/v32/gopal14.html.

[49] Thomas L. Griffiths et Mark Steyvers. « Finding scientific topics ». In : *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), p. 5228-5235. doi : 10.1073/pnas.0307752101.

[50] Michael Hagenau, Michael Liebmann et Dirk Neumann. « Automated news reading: Stock price prediction based on financial news using context-capturing features ». In : *Decision Support Systems* 55.3 (2013), p. 685-697. doi : 10.1016/j.dss.2013.02.006.

[51] Jiawei Han, Micheline Kamber et Jian Pei. *Data mining concepts and techniques, third edition*. Waltham, Mass. : Morgan Kaufmann Publishers, 2012. doi : 10.1016/C2009-0-61819-5.

[52] G. H. Hardy, John E. Littlewood et George Polya. *Inequalities*. Cambridge : Cambridge University Press, 1988. isbn : 0521358809.

[53] Zellig S. Harris. « Distributional Structure ». In : *WORD* 10.2-3 (1954), p. 146-162. doi : 10.1080/00437956.1954.11659520.

[54] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv : 1512.03385 [cs.CV].

[55] Sepp Hochreiter et Jürgen Schmidhuber. « Long Short-Term Memory ». In : *Neural Computation* 9.8 (1997), p. 1735-1780. doi : 10.1162/neco.1997.9.8.1735.

[56] J J Hopfield. « Neural networks and physical systems with emergent collective computational abilities ». In : *Proceedings of the National Academy of Sciences* 79.8 (1982), p. 2554-2558. doi : 10.1073/pnas.79.8.2554.

[57] Kurt Hornik et Bettina Grün. « movMF: an R package for fitting mixtures of von Mises-Fisher distributions ». In : *Journal of Statistical Software* 58.10 (2014), p. 1-31. doi : 10.18637/jss.v058.i10.

[58] Kurt Hornik et al. « Spherical k-Means Clustering ». In : *Journal of Statistical Software, Articles* 50.10 (2012), p. 1-22. DOI : 10.18637/jss.v050.i10.

[59] John Huber et Ronald Inglehart. « Expert Interpretations of Party Space and Party Locations in 42 Societies ». In : *Party Politics* 1.1 (1995), p. 73-111. DOI : 10.1177/1354068895001001004.

[60] L. Hubert et P. Arabie. « Comparing partitions ». In : *Journal of classification* 2.1 (1985), p. 193-218. DOI : 10.1007/BF01908075.

[61] Michael J. Bommarito II, Daniel Martin Katz et Eric M. Detterman. « LexNLP: Natural language processing and information extraction for legal and regulatory texts ». In : *CoRR* abs/1806.03688 (2018). DOI : 10.2139/ssrn.3192101.

[62] Armand Joulin et al. *FastText.zip: Compressing text classification models*. 2016. arXiv : 1612.03651 [cs.CL].

[63] John T. Kent. « The Fisher-Bingham Distribution on the Sphere ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 44.1 (1982), p. 71-80. DOI : 10.1111/j.2517-6161.1982.tb01189.x.

[64] Hyunjoong Kim, Han Kyul Kim et Sungzoon Cho. « Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling ». In : *Expert Systems with Applications* 150 (2020), p. 113288. DOI : 10.1016/j.eswa.2020.113288.

[65] Diederik P Kingma et Jimmy Ba. « Adam: A method for stochastic optimization ». In : (2014). arXiv : 1412.6980.

[66] Heike Klüver. « The contextual nature of lobbying: Explaining lobbying success in the European Union ». In : *European Union Politics* 12.4 (2011), p. 483-506. DOI : 10.1177/1465116511413163.

[67] Mathias Kraus et Stefan Feuerriegel. « Decision support from financial disclosures with deep neural networks and transfer learning ». In : *Decision Support Systems* 104 (2017), p. 38-48. DOI : 10.1016/j.dss.2017.10.001.

[68] Kostiantyn Kucher., Carita Paradis. et Andreas Kerren. « DoSVis: Document Stance Visualization ». In : *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: IVAPP,* 2018, p. 168-175. DOI : 10.5220/0006539101680175.

[69] Matt J. Kusner et al. « From Word Embeddings to Document Distances ». In : *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France : JMLR.org, 2015, p. 957-966. URL : https://proceedings.mlr.press/v37/kusnerb15.html.

[70] Thomas K. Landauer et Susan T. Dumais. « A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge ». In : *Psychological Review* 104 (1997), p. 211-240. doi : 10.1037/0033-295X.104.2.211.

[71] Hang Le et al. *FlauBERT: Unsupervised Language Model Pre-training for French*. 2020. arXiv : 1912.05372 [cs.CL].

[72] Heeyoung Lee et al. « On the Importance of Text Analysis for Stock Price Prediction ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland : European Language Resources Association (ELRA), 2014, p. 1170-1175. isbn : 978-2-9517408-8-4. url : http://www.lrec-conf.org/proceedings/lrec2014/pdf/1065_Paper.pdf.

[73] Fan Li et Yiming Yang. « Analysis of recursive feature elimination methods ». In : *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 2005, p. 633-634. doi : 10.1145/1076034.1076164.

[74] Tao Li. « A General Model for Clustering Binary Data ». In : *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. KDD '05. New York, NY, USA : Association for Computing Machinery, 2005, 188–197. doi : 10.1145/1081870.1081894.

[75] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv : 1907.11692 [cs.CL].

[76] A Craig MacKinlay. « Event studies in economics and finance ». In : *Journal of economic literature* 35.1 (1997), p. 13-39. url : http://www.jstor.org/stable/2729691.

[77] K. V. Mardia. « Statistics of Directional Data ». In : *Journal of the Royal Statistical Society: Series B (Methodological)* 37.3 (1975), p. 349-371. doi : 10.1111/j.2517-6161.1975.tb01550.x.

[78] K.V. Mardia et P.E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009. isbn : 9780470317815. doi : 10.1002/9780470316979.

[79] G. McLachlan et John Wiley & Sons. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics. Wiley, 1992. doi : 10.1002/0471725293.

[80] G.J. McLachlan et D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, 2004. doi : 10.1002/0471721182.

[81] Rada Mihalcea. « Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization ». In : *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. ACLdemo '04. Barcelona, Spain : Association for Computational Linguistics, 2004. doi : 10.3115/1219044.1219064.

[82] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv : 1301.3781 [cs.CL].

[83] KiHwan Nam et NohYoon Seong. « Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market ». In : *Decision Support Systems* 117 (2019), p. 100-112. doi : 10.1016/j.dss.2018.11.004.

[84] Thien Hai Nguyen, Kiyoaki Shirai et Julien Velcin. « Sentiment analysis on social media for stock movement prediction ». In : *Expert Systems with Applications* 42.24 (2015), p. 9603-9611. doi : 10.1016/j.eswa.2015.07.052.

[85] Giannis Nikolentzos et al. « Multivariate Gaussian Document Representation from Word Embeddings for Text Categorization ». In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain : Association for Computational Linguistics, 2017, p. 450-455. doi : 10.18653/v1/e17-2072.

[86] Daniel W. Otter, Julian R. Medina et Jugal K. Kalita. *A Survey of the Usages of Deep Learning in Natural Language Processing*. 2019. arXiv : 1807.10854 [cs.CL].

[87] Wei Pan et Xiaotong Shen. « Penalized Model-Based Clustering with Application to Variable Selection ». In : *Journal of Machine Learning Research* 8.41 (2007), p. 1145-1164. url : http://jmlr.org/papers/v8/pan07a.html.

[88] Fabian Pedregosa et al. « Scikit-learn: Machine learning in Python ». In : *Journal of Machine Learning Research* 12.Oct (2011), p. 2825-2830. url : http://jmlr.org/papers/v12/pedregosa11a.html.

[89] Jeffrey Pennington, Richard Socher et Christopher D. Manning. « GloVe: Global Vectors for Word Representation ». In : *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1532-1543. doi : 10.3115/v1/D14-1162.

[90] Matthew E. Peters et al. « Deep Contextualized Word Representations ». In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, juin 2018, p. 2227-2237. doi : 10.18653/v1/N18-1202.

[91] Zhang Qi. « The text classification of theft crime based on TF-IDF and XGBoost model ». In : *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE. 2020, p. 1241-1246. DOI : 10.1109/ICAICA50127.2020.9182555.

[92] Gabriele Ranco et al. « The Effects of Twitter Sentiment on Stock Price Returns ». In : *PLOS ONE* 10.9 (sept. 2015), p. 1-21. DOI : 10.1371/journal.pone.0138441.

[93] William M. Rand. « Objective Criteria for the Evaluation of Clustering Methods ». In : *Journal of the American Statistical Association* 66.336 (1971), p. 846-850. DOI : 10.1080/01621459.1971.10482356.

[94] Joseph Reisinger et al. « Spherical Topic Models ». In : *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel : Omnipress, 2010, 903–910. URL : https://icml.cc/Conferences/2010/papers/45.pdf.

[95] Stephen Robertson. « Understanding inverse document frequency: On theoretical arguments for IDF ». In : *Journal of Documentation* 60 (2004), p. 2004. DOI : 10.1108/00220410410560582.

[96] Andreas Rücklé et al. « Concatenated p-mean Word Embeddings as Universal Cross-Lingual Sentence Representations ». In : *CoRR* abs/1803.01400 (2018). arXiv : 1803.01400.

[97] David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams. « Learning Representations by Back-Propagating Errors ». In : *Neurocomputing: Foundations of Research*. Cambridge, MA, USA : MIT Press, 1988, 696–699. DOI : 10.1038/323533a0.

[98] Andreas Rücklé et al. *Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations*. 2018. arXiv : 1803.01400 [cs.CL].

[99] Aghiles Salah. « Von Mises-Fisher based (co-)clustering for high-dimensional sparse data : application to text and collaborative filtering data ». PhD thesis. Université Sorbonne Paris Cité, nov. 2016. URL : https://tel.archives-ouvertes.fr/tel-01835699.

[100] Aghiles Salah et Mohamed Nadif. « Model-based von Mises-Fisher Co-clustering with a Conscience ». In : *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM'17)*. SIAM. Houston, TX, United States, 2017, p. 246-254. DOI : 10.1137/1.9781611974973.28.

[101] Aghiles Salah, Nicoleta Rogovschi et Mohamed Nadif. « Model-based Co-clustering for High Dimensional Sparse Data ». In : *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Arthur Gretton et Christian C. Robert. T. 51. Proceedings of Machine Learning Research. Cadiz,

Spain : PMLR, 2016, p. 866-874. URL : http://proceedings.mlr.press/v51/salah16.html.

[102] Robert P SCHUMAKER et Hsinchun CHEN. « A quantitative stock prediction system based on financial news ». In : *Information Processing & Management* 45.5 (2009), p. 571-583. DOI : 10.1016/j.ipm.2009.05.001.

[103] Gideon SCHWARZ et al. « Estimating the dimension of a model ». In : *The annals of statistics* 6.2 (1978), p. 461-464. DOI : 10.1214/aos/1176344136.

[104] Celine SCORNAVACCA, Franziska ZICKMANN et Daniel H. HUSON. « Tanglegrams for rooted phylogenetic trees and networks ». In : *Bioinformatics* 27.13 (juin 2011), p. i248-i256. DOI : 10.1093/bioinformatics/btr210.

[105] Carson SIEVERT et Kenneth SHIRLEY. « LDAvis: A method for visualizing and interpreting topics ». In : *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA : Association for Computational Linguistics, juin 2014, p. 63-70. DOI : 10.3115/v1/W14-3110.

[106] Aaron SIM, Dimosthenis TSAGKRASOULIS et Giovanni MONTANA. « Random forests on distance matrices for imaging genetics studies ». In : *Statistical applications in genetics and molecular biology* 12.6 (2013), p. 757-786. DOI : 10.1515/sagmb-2013-0040.

[107] Sidak Pal SINGH et al. *Context Mover's Distance & Barycenters: Optimal Transport of Contexts for Building Representations*. 2020. arXiv : 1808.09663 [cs.CL].

[108] Michael E. TIPPING et Christopher M. BISHOP. « Mixtures of Probabilistic Principal Component Analyzers ». In : *Neural Computation* 11.2 (1999), p. 443-482. DOI : 10.1162/089976699300016728.

[109] Amirsina TORFI et al. « Natural Language Processing Advancements By Deep Learning: A Survey ». In : *CoRR* abs/2003.01200 (2020). arXiv : 2003.01200.

[110] Warren S TORGERSON. « Multidimensional scaling: I. Theory and method ». In : *Psychometrika* 17.4 (1952), p. 401-419. DOI : 10.1007/BF02288916.

[111] Joan TORRUELLA et Ramon CAPSADA. « Lexical Statistics and Tipological Structures: A Measure of Lexical Richness ». In : *Procedia - Social and Behavioral Sciences* 95 (oct. 2013), p. 447-454. DOI : 10.1016/j.sbspro.2013.10.668.

[112] Ashish VASWANI et al. « Attention is All you Need ». In : *Advances in Neural Information Processing Systems*. Sous la dir. d'I. GUYON et al. T. 30. Curran Associates, Inc., 2017. URL : https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[113] Cédric VILLANI. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Berlin : Springer, 2009. DOI : 10.1007/978-3-540-71050-9.

[114] Yonghui Wu et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv : 1609.08144 [cs.CL].

[115] Yumo Xu et Shay B. Cohen. « Stock Movement Prediction from Tweets and Historical Prices ». In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia : Association for Computational Linguistics, juil. 2018, p. 1970-1979. doi : 10.18653/v1/P18-1183.

[116] Mikhail Yurochkin et al. « Hierarchical Optimal Transport for Document Representation ». In : *NIPS 32*. Curran Ass., 2019, p. 1601-1611. url : http://papers.nips.cc/paper/8438-hierarchical-optimal-transport-for-document-representation.pdf.

[117] Jian-Bing Zhang, Yi-Xin Sun et De-Chuan Zhan. « Multiple-instance learning for text categorization based on semantic representation ». In : *Big Data and Information Analytics* 2 (sept. 2017), p. 69-75. doi : 10.3934/bdia.2017009.

[118] Yongli Zhang et Xiaotong Shen. « Model selection procedure for high-dimensional data ». In : *Statistical Analysis and Data Mining: The ASA Data Science Journal* 3.5 (2010), p. 350-358. doi : 10.1002/sam.10088.

[119] Yang Zhao, Abhishek K. Shrivastava et Kwok Leung Tsui. « Regularized Gaussian Mixture Model for High-Dimensional Clustering ». In : *IEEE Transactions on Cybernetics* 49.10 (2019), p. 3677-3688. doi : 10.1109/TCYB.2018.2846404.

[120] Shi Zhong et Joydeep Ghosh. « Generative Model-Based Document Clustering: A Comparative Study ». In : *Knowl. Inf. Syst.* 8.3 (sept. 2005), 374–384. issn : 0219-1377. doi : 10.1007/s10115-004-0194-1.

[121] Chunting Zhou et al. *A C-LSTM Neural Network for Text Classification*. 2015. arXiv : 1511.08630 [cs.CL].

[122] Hui Zou, Trevor Hastie et Robert Tibshirani. « On the degrees of freedom of the lasso ». In : *The Annals of Statistics* 35.5 (2007). issn : 0090-5364. doi : 10.1214/009053607000000127.

Annexe A

Exemple de rapport 8-K vierge

OMB APPROVAL
OMB Number: 3235-0060
Expires: October 31, 2024
Estimated average burden hours per response.....9.21

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 8-K

CURRENT REPORT
Pursuant to Section 13 OR 15(d) of The Securities Exchange Act of 1934

Date of Report (Date of earliest event reported) _____

_____

(Exact name of registrant as specified in its charter)

_____

(State or other jurisdiction of incorporation)	(Commission File Number)	(IRS Employer Identification No.)

_____

(Address of principal executive offices) (Zip Code)

Registrant's telephone number, including area code _____

_____

(Former name or former address, if changed since last report.)

Check the appropriate box below if the Form 8-K filing is intended to simultaneously satisfy the filing obligation of the registrant under any of the following provisions (see General Instruction A.2. below):

[ ] Written communications pursuant to Rule 425 under the Securities Act (17 CFR 230.425)

[ ] Soliciting material pursuant to Rule 14a-12 under the Exchange Act (17 CFR 240.14a-12)

[ ] Pre-commencement communications pursuant to Rule 14d-2(b) under the Exchange Act (17 CFR 240.14d-2(b))

[ ] Pre-commencement communications pursuant to Rule 13e-4(c) under the Exchange Act (17 CFR 240.13e-4(c))

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading Symbol(s)	Name of each exchange on which registered

Indicate by check mark whether the registrant is an emerging growth company as defined in Rule 405 of the Securities Act of 1933 (§230.405 of this chapter) or Rule 12b-2 of the Securities Exchange Act of 1934 (§240.12b-2 of this chapter).

Emerging growth company ☐

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. ☐

SEC 873 (02-21) Potential persons who are to respond to the collection of information contained in this form are not required to respond unless the form displays a currently valid OMB control number. 1 of 22

Annexe B

Article de Lee et al.

La période d'étude de l'article [72] couvre les années 2002 à 2012. Les caractéristiques de la base collectée sont résumées dans la table B.1 qui est mise à disposition sous forme de données brutes [1].

Table B.1 – La base de données de l'article de Lee et al.

Dataset	# de 8-Ks	# de words	# d'entreprises
Train	6652	13M	453
Dev	3433	7.1M	461
Test	3586	7.8M	478

La base de données a été complétée avec des données financières :

- les *earnings surprise* [2]. Les rapports qui ne comportent pas cette information sont supprimés.
- les variations du cours de l'action. Les variations de ces dernières se calculent entre chaque ouverture/fermeture de marchés et des mouvements récents à l'aide d'une moyenne mobile à une semaine, un mois, un trimestre et une année. Chaque type de variation est normalisée par la variation respective du $S\&P500$;
- la valeur de la volatilité de l'index $S\&P500$.

Cependant, la reproduction de cette base de données s'est avérée impossible, car les auteurs :

- différencient des événements identiques qui possèdent une typographie différente, c'est-à-dire qu'ils distinguent *Other event* et *other event* ;
- ne se souviennent plus de la méthode de sélection des textes et des termes. La procédure n'est pas explicitée dans l'article ;

1. https://nlp.stanford.edu/pubs/stock-event.html
2. Différence entre les gains par action reportés et attendus par les analystes.

— utilisent la méthode de factorisation matricielle non négative, *non-negative matrix factorization* en anglais, sur des données *out-of-sample* qui aboutissent à des facteurs négatifs.

Annexe C

Annexes du chapitre 6

C.1 Détails de l'implémentation

Nous abordons dans cette section les détails techniques importants concernant l'implémentation concrète de l'algorithme 1 (p. 135).

Premièrement, il est bien connu que l'initialisation joue un rôle important dans les algorithmes EM. Dans notre cas, une stratégie simple a été suffisante pour obtenir des résultats satisfaisants. Nous sélectionnons, sans remplacement, de manière aléatoire et uniforme K observations dans l'ensemble de données $\boldsymbol{X}$ qui servent de valeurs initiales pour le $(\boldsymbol{\mu}_k)_{1\leq k\leq K}$. Puis, nous effectuons des affectations fortes de toutes les observations à leur moyenne directionnelle la plus proche (par rapport au produit scalaire, c'est-à-dire la similarité en cosinus). Cela nous permet de calculer les valeurs initiales de $\boldsymbol{\alpha}$ comme le rapport des observations assignées à chaque prototype. Enfin, nous calculons les valeurs initiales de $\boldsymbol{\kappa}$ en utilisant l'estimateur EM, c'est-à-dire en résolvant l'équation (6.21, p. 84) en utilisant pour le τ_{ik} la matrice d'affectations fortes. L'algorithme 3 résume le processus. Notons que l'estimation finale peut échouer et que le processus complet peut devoir être répété plusieurs fois afin de produire une configuration initiale correcte (voir ci-dessous pour plus de détails).

Deuxièmement, les modèles de mélange peuvent tomber dans des configurations locales problématiques. Comme indiqué dans [9], κ_k peut devenir sans limite si la composante correspondante se concentre sur une seule observation, dans un comportement similaire à celui observé pour le mélange de distributions gaussiennes, lorsque l'écart-type de la composante disparaît. Comme dans [9], nous évitons ce problème en plafonnant κ_k à une valeur limite (10^6 dans nos expériences).

Au contraire, une composante du mélange peut également devenir inutile lorsque $\kappa_k \to 0$. Cela correspond à la composante qui converge vers une distribution uniforme. Ce comportement est facilement détecté car il se manifeste par le fait que le côté droit de l'équation (6.21, p. 84) prend une valeur supérieure ou égale à 1. Nous surveillons cette quantité et interrompons l'algorithme lorsqu'une telle situation est rencontrée. Nous si-

Algorithm 3 Initialisation de l'EM

Sélection de manière aléatoire et uniforme $(\boldsymbol{\mu}_k)_{1\leq k\leq K}$ parmi les lignes de $\boldsymbol{X}$ sans remplacement
$c_i \leftarrow \arg\max_{1\leq k\leq K} \boldsymbol{\mu}_k^T \boldsymbol{x}_i$
$\tau_{ik} \leftarrow \mathbb{I}_{k=c_i}$
$\alpha_k = \frac{1}{n}\sum_{i=1}^n \tau_{ik}$
définir κ_k comme la solution de

$$\frac{I_{d/2}(\kappa_k)}{I_{d/2-1}(\kappa_k)} = \boldsymbol{\mu}_k^T \frac{\sum_{i=1}^n \tau_{ik}\boldsymbol{x}_i}{\sum_{i=1}^n \tau_{ik}}.$$

gnalons dans ce cas un problème de convergence. Notons que le processus d'initialisation décrit ci-dessus peut également échouer pour cette raison.

Enfin, lorsque $\beta > 0$, l'équation (6.29) (p. 86) peut produire une moyenne directionnelle nulle : cela signifie en pratique que l'étape M a échoué. Lorsque nous détectons ce problème, nous arrêtons l'algorithme et signalons un problème de convergence.

C.2 Génération des données simulées

La principale difficulté concernant le mélange de distributions de von Mises-Fisher est de régler $\boldsymbol{\kappa}$ pour contrôler le chevauchement entre les composantes. En effet, le problème de l'estimation est évidemment plus facile lorsque les composantes sont bien séparées. Malheureusement, le degré de chevauchement dépend de $\boldsymbol{\kappa}$ mais aussi fortement de la dimension : les données de haute dimension sont séparées pour des valeurs beaucoup plus grandes que pour les données de basse dimension. Pour permettre une analyse équitable du comportement du modèle, nous nous appuyons sur une procédure de calibration.

Nous définissons d'abord le chevauchement d'un mélange comme suit. Supposons un vecteur de paramètres $\boldsymbol{\Theta}$. Nous générons un ensemble complet de données $(\boldsymbol{Z}, \boldsymbol{X})$, incluant en particulier la vérité terrain $\boldsymbol{Z}$. Nous définissons alors le vecteur d'affectation forte $\boldsymbol{C}$ par

$$C_i = \arg\max_{1\leq k\leq K} \frac{\alpha_k f_k(\boldsymbol{x}_i, \theta_k)}{\sum_{l=1}^K \alpha_l f_l(\boldsymbol{x}_i, \theta_l)}. \tag{C.1}$$

Le chevauchement dans $\boldsymbol{X}$ est le taux d'erreur de $\boldsymbol{C}$ par rapport à $\boldsymbol{Z}$, soit

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{C_i\neq Z_i}.$$

Pour calibrer $\boldsymbol{\kappa}$, nous utilisons une grille unidimensionnelle régulière de nombres

réels positifs, $\mathcal{G}$, et pour chaque $g \in \mathcal{G}$, nous définissons $\kappa_k = g$. Nous échantillonnons $20 \times K$ vecteurs aléatoires uniformément sur l'hypersphère unitaire $\mathbb{S}^{d-1}$ et sélectionnons les K vecteurs les plus séparés (en minimisant leurs produits scalaires par paires). On ajuste ensuite $\boldsymbol{\kappa}$ pour tenir compte de la séparabilité intrinsèque. Cela consiste à utiliser $\boldsymbol{\kappa}$ défini par l'équation (6.41) (p. 91) que nous rappelons ici

$$\kappa'_k = \frac{2\kappa_k}{1 - \max_{l \neq k} \boldsymbol{\mu}_k^T \boldsymbol{\mu}_l}. \tag{C.2}$$

Enfin, nous générons un ensemble de données équilibré ($\alpha_k = \frac{1}{K}$) avec les paramètres choisis et mesurons son chevauchement. Pour une dimension d donnée, la procédure est répétée 100 fois pour générer une correspondance entre la grille $\mathcal{G}$ et le chevauchement moyen. En utilisant une interpolation linéaire de base, nous pouvons alors calculer un bon candidat pour $\boldsymbol{\kappa}$ afin d'obtenir un chevauchement souhaité dans l'ensemble de données résultant.

Nous avons mené cette procédure de calibration pour quatre dimensions différentes (2, 10, 100, 1000), en utilisant $K = 4$ et $n = 1000$, et avec une grille de taille 100. Les résultats sont présentés sur la figure C.1. Remarquons que nous avons utilisé une grille différente pour chaque dimension. Il est intéressant de noter que les paramètres utilisés dans [9] pour l'ensemble de données *big-mix* correspondent à une séparation parfaite et donc à un problème d'estimation très facile, malgré la dimensionnalité de l'ensemble de données.

Pour introduire une variabilité supplémentaire dans l'ensemble des données générées à l'aide de cette procédure de calibration, avant d'appliquer la mise à l'échelle définie dans l'équation (6.41) (p. 91), nous ajoutons un bruit gaussien à chaque κ_k, avec un écart type de $0{,}025\kappa_k$.

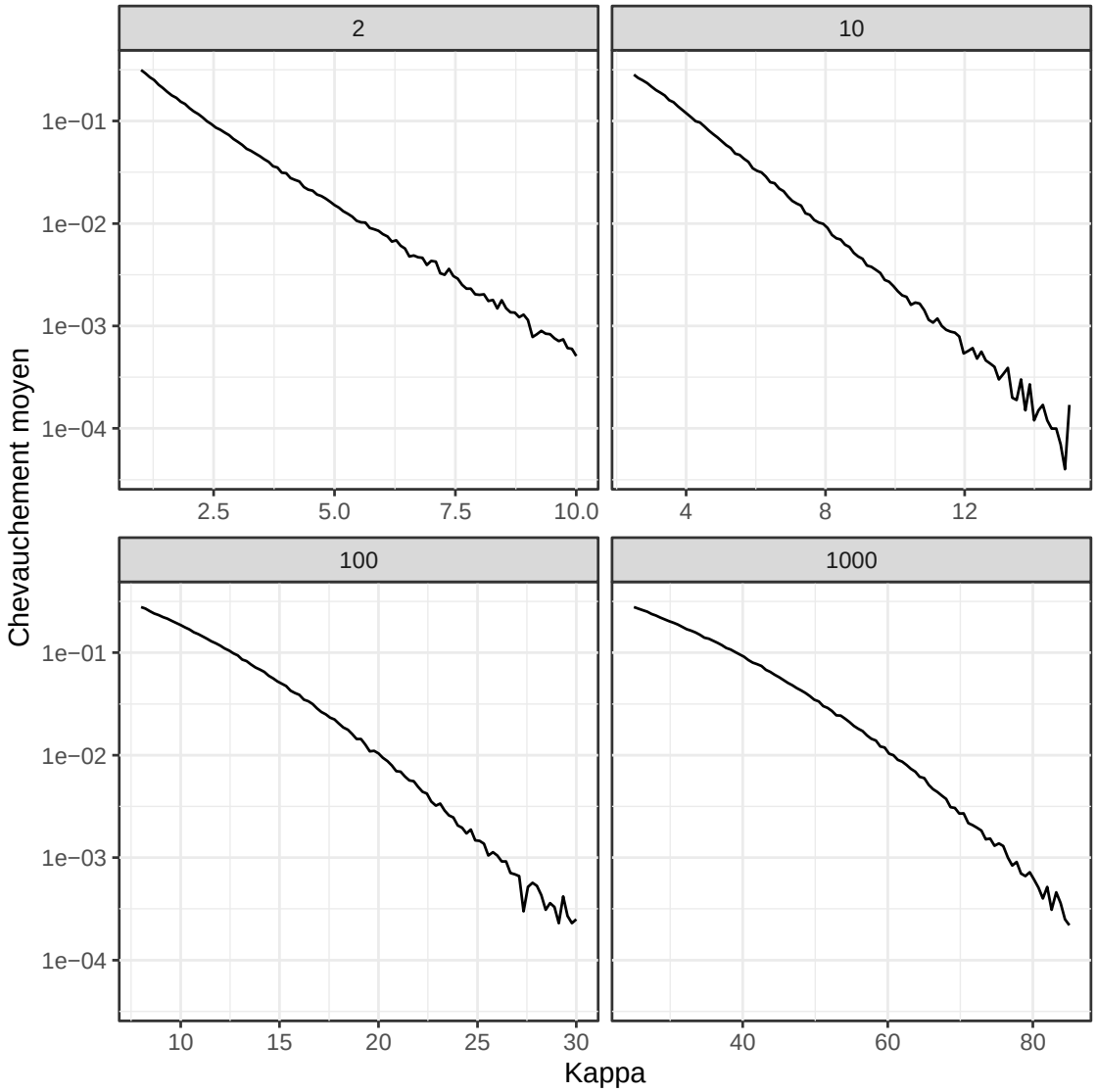


FIGURE C.1 – Résultats de calibration pour quatre dimensions différentes (2, 10, 100, 1000). Le chevauchement est en échelle logarithmique. Comme certaines valeurs de κ n'entraînent aucun chevauchement, les valeurs nulles ont été remplacées par 10^{-5} uniquement à des fins de visualisation.

C.3 Résultats supplémentaires sur des données simulées denses

C.3.1 Sélection de modèles en basse dimension

Nous rapportons ici les résultats obtenus en dimension $d = 10$ avec $K^* = 4$ composantes. Comme dans la **section 6.4.2.1** (p. 92), nous comparons différents nombres d'observations. Les résultats sont présentés dans les tables C.1 et C.2, ainsi que dans la figure C.2 pour 50 observations et la figure C.3 pour 200 observations. Le comportement général est similaire à celui rapporté dans la **section 6.4.2.1** (p. 92).

Le BIC sélectionne raisonnablement bien le nombre réel de composantes lorsque le chevauchement n'est pas important et que le nombre d'observations est suffisant. Un petit ensemble de données peut conduire à un surajustement dans le cas d'un faible chevauchement, tandis qu'un chevauchement important conduit à un sous-ajustement.

L'AIC surajuste dans toutes les situations, avec une nette tendance à surajuster d'autant plus, pour les grands ensembles de données.

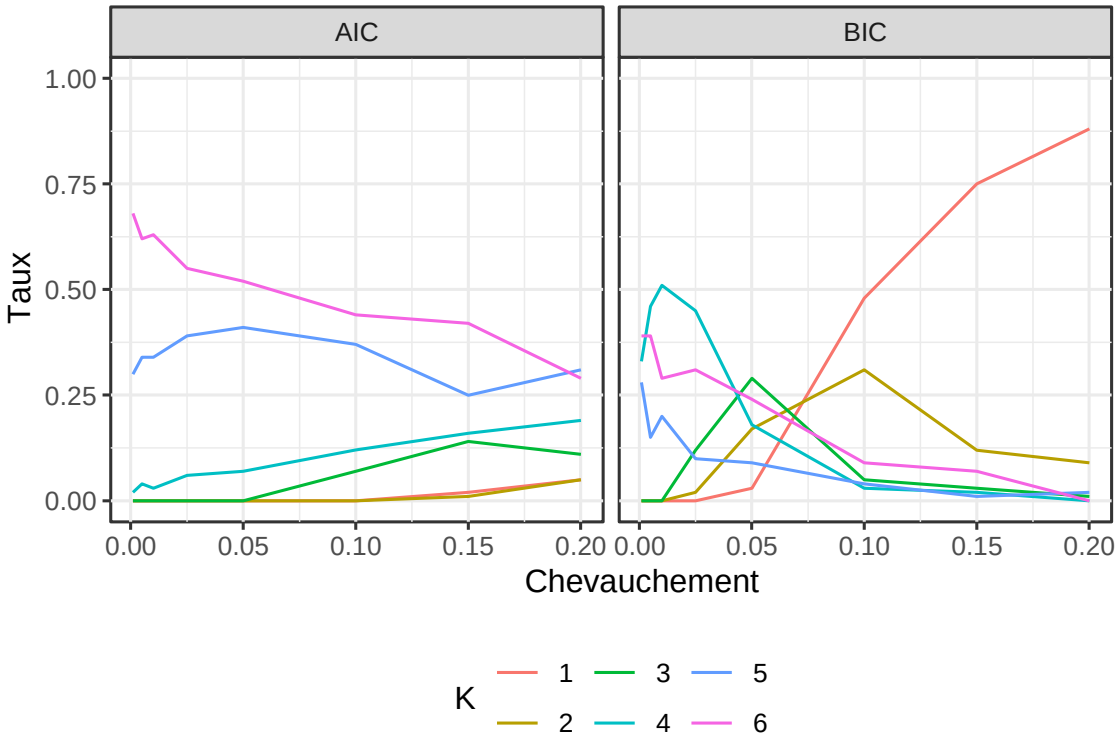


FIGURE C.2 – Résultats de la sélection du nombre de composantes basée sur l'AIC et sur le BIC sur l'ensemble de données en dimension $d = 10$ pour 50 observations avec $K^* = 4$. Pour chaque K et chaque valeur de chevauchement, la figure affiche le pourcentage des jeux de données pour lesquels cette valeur de K a été considérée comme optimale.

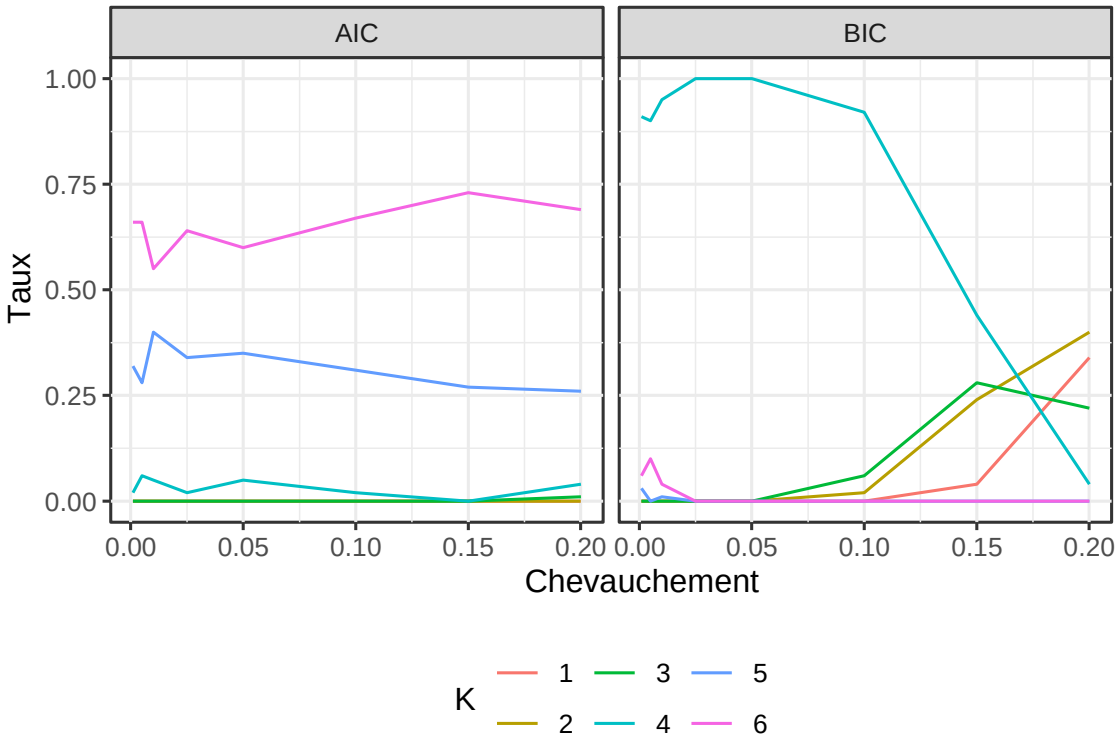


FIGURE C.3 – Résultats de la sélection du nombre de composantes basée sur l'AIC et sur le BIC sur l'ensemble de données en dimension $d = 10$ pour 200 observations avec $K^* = 4$. Pour chaque K et chaque valeur de chevauchement, la figure affiche le pourcentage des jeux de données pour lesquels cette valeur de K a été considérée comme optimale.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
50	1	0.00	0.00	0.00	0.00	0.03	0.48	0.75	0.88
50	2	0.00	0.00	0.00	0.02	0.17	0.31	0.12	0.09
50	3	0.00	0.00	0.00	0.12	0.29	0.05	0.03	0.01
50	4	0.33	0.46	0.51	0.45	0.18	0.03	0.02	0.00
50	5	0.28	0.15	0.20	0.10	0.09	0.04	0.01	0.02
50	6	0.39	0.39	0.29	0.31	0.24	0.09	0.07	0.00
200	1	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.34
200	2	0.00	0.00	0.00	0.00	0.00	0.02	0.24	0.40
200	3	0.00	0.00	0.00	0.00	0.00	0.06	0.28	0.22
200	4	0.91	0.90	0.95	1.00	1.00	0.92	0.44	0.04
200	5	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00
200	6	0.06	0.10	0.04	0.00	0.00	0.00	0.00	0.00
500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
500	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
500	4	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.84
500	5	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	5	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE C.1 – Résultats complets du **BIC** sur le cas à 10 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
50	1	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05
50	2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05
50	3	0.00	0.00	0.00	0.00	0.00	0.07	0.14	0.11
50	4	0.02	0.04	0.03	0.06	0.07	0.12	0.16	0.19
50	5	0.30	0.34	0.34	0.39	0.41	0.37	0.25	0.31
50	6	0.68	0.62	0.63	0.55	0.52	0.44	0.42	0.29
200	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
200	4	0.02	0.06	0.05	0.02	0.05	0.02	0.00	0.04
200	5	0.32	0.28	0.40	0.34	0.35	0.31	0.27	0.26
200	6	0.66	0.66	0.55	0.64	0.60	0.67	0.73	0.69
500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	4	0.03	0.04	0.00	0.01	0.00	0.01	0.00	0.02
500	5	0.28	0.36	0.36	0.31	0.33	0.22	0.25	0.22
500	6	0.69	0.60	0.64	0.68	0.67	0.77	0.75	0.76
1000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	4	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.03
1000	5	0.24	0.20	0.24	0.32	0.27	0.20	0.20	0.22
1000	6	0.74	0.80	0.75	0.67	0.72	0.80	0.80	0.75

TABLE C.2 – Résultats complets du **AIC** sur le cas à 10 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

C.3.2 Sélection de modèles en grande dimension

Nous rapportons ici les résultats obtenus en dimension $d = 100$ avec $K^* = 4$ pour les composantes. Comme dans la **section 6.4.2.1** (p. 92), nous comparons différents nombres d'observations. Les résultats sont présentés dans les tableaux C.3 et C.4. Nous commentons les résultats dans la **section 6.4.2.2** (p. 97).

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
200	1	0.00	0.00	0.01	0.91	1.00	1.00	1.00	1.00
200	2	0.00	0.00	0.09	0.09	0.00	0.00	0.00	0.00
200	3	0.00	0.01	0.20	0.00	0.00	0.00	0.00	0.00
200	4	1.00	0.99	0.70	0.00	0.00	0.00	0.00	0.00
200	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	1	0.00	0.00	0.00	0.00	0.02	0.99	1.00	1.00
500	2	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.00
500	3	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00
500	4	1.00	1.00	1.00	1.00	0.83	0.00	0.00	0.00
500	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	1	0.00	0.00	0.00	0.00	0.00	0.01	0.97	1.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00
1000	3	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
1000	4	1.00	1.00	1.00	1.00	1.00	0.89	0.00	0.00
1000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE C.3 – Résultats complets du **BIC** sur le cas à 100 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

n	K	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2
200	1	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.28
200	2	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.36
200	3	0.00	0.00	0.00	0.00	0.00	0.06	0.46	0.28
200	4	1.00	0.99	1.00	1.00	1.00	0.92	0.38	0.08
200	5	0.00	0.01	0.00	0.00	0.00	0.02	0.04	0.00
200	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	4	0.98	0.94	0.94	0.92	0.90	0.66	0.32	0.12
500	5	0.01	0.06	0.06	0.08	0.10	0.27	0.53	0.60
500	6	0.01	0.00	0.00	0.00	0.00	0.07	0.15	0.28
1000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	4	0.79	0.77	0.76	0.55	0.45	0.26	0.03	0.02
1000	5	0.20	0.22	0.19	0.40	0.51	0.53	0.46	0.32
1000	6	0.01	0.01	0.05	0.05	0.04	0.21	0.51	0.66
5000	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	4	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
5000	5	0.40	0.30	0.34	0.22	0.22	0.10	0.05	0.01
5000	6	0.60	0.68	0.66	0.78	0.78	0.90	0.95	0.99

TABLE C.4 – Résultats complets du **AIC** sur le cas à 100 dimensions avec $K^* = 4$. Chaque ligne donne le taux des ensembles de données d'une taille donnée (n) pour lesquels le critère a considéré que le K donné était optimal, à travers les différents chevauchements.

Annexe D

Analyse de la contribution des lobbys dans le processus parlementaire de l'Union Européenne

Résumé : Cette annexe concerne un travail, toujours en cours de réalisation, qui est effectué en collaboration avec la Chaire Gouvernance et Régulation de l'Université Paris Dauphine-PSL. Il porte sur l'analyse du comportement des Lobbys lors des consultations de l'Union Européenne. Pour ce faire, à l'aide des discours parlementaires, nous créons des dictionnaires que nous appliquons aux consultations. Ceci nous permet d'avoir un bon aperçu des positions des différents acteurs selon les consultations et d'analyser leurs changements de positionnement en fonction des problématiques.

D.1 Introduction

Les stratégies politiques d'entreprises ont un potentiel énorme, offrant des avantages à celles qui les déploient. Une augmentation des prix réglementés, une décision favorable ou une exonération fiscale sont autant d'avantages qu'elles peuvent rechercher par leur biais. Ces stratégies comprennent toutes les initiatives adressées aux institutions politiques qui tentent d'aligner leur environnement commercial sur leurs préférences. Dans ce contexte, le lobbying est une stratégie fréquemment employée. Selon la définition utilisée par [14], le lobbying est la fourniture stratégique d'informations politiquement pertinentes aux représentants du gouvernement. Selon [13], les entreprises sont également en concurrence dans l'arène politique. Ainsi, pour celles qui décident de participer au processus politique, cela inclut la compétition pour obtenir plus d'espace et un meilleur accès aux politiciens. L'accès aux titulaires de fonctions ciblées est la condition première pour mettre en œuvre des stratégies de lobbying. Par conséquent, une compréhension large de la dynamique du lobbying doit prendre en compte les déterminants de l'accès. Même si l'accès ne signifie pas l'influence, le premier est une condition nécessaire pour tenter d'exercer la seconde.

Dans ce contexte, l'Union européenne apparaît comme un environnement pertinent pour étudier ces dynamiques. Ses caractéristiques supranationales apportent des contraintes supplémentaires au processus d'élaboration des politiques. Avec ses 27 pays, l'UE est la plus grande économie du monde. La Commission européenne, qui peut être considérée comme l'organe exécutif de l'UE, publie chaque année des centaines d'actes juridiques parmi les directives, les règlements et les décisions.

Un bref coup d'oeil à quelques chiffres peut l'illustrer : nous comptons plus de 2300 associations d'entreprises et de commerce enregistrées pour la représentation d'intérêts auprès de la Commission européenne et plus de 500 entreprises ayant des bureaux à Bruxelles. En outre, des données récentes sur les réunions entre les représentants de la Commission et les groupes d'intérêt indiquent un accès inégal parmi les entreprises : si peu d'entre elles ont régulièrement accès aux représentants de la Commission, la majorité n'a que très peu de réunions avec la Commission. Les faits présentés renforcent l'idée de concurrence dans l'arène politique. En outre, des recherches antérieures suggèrent que l'accès aux représentants politiques peut réduire l'incertitude due aux questions politiques.

Nous présentons ci-dessous les données utilisées pour la création de dictionnaires clivants selon les orientations politiques. Puis, nous mettons en place un algorithme de classification qui nous permet d'exposer le comportement des acteurs selon les consultations.

D.2 Données

Dans cette étude, nous utilisons plusieurs ensembles de données :

1. Les consultations mises à disposition par l'article [66] qui s'étendent du 1 janvier 2000 au 31 décembre 2008 ;
2. Les discours ayant eu lieu au Parlement européen entre juillet 1999 et décembre 2017 [1] [1] ;
3. La base de données *ParlGov* [2] met à disposition une représentation multidimensionnelle des partis politiques nationaux de l'Union européenne, notamment selon quatre axes :
   (a) gauche - droite [30, 59, 15, 8] ;
   (b) état - marché [15, 8] ;
   (c) liberté - autorité [15, 8] ;
   (d) anti - pro EU [15, 8].

1. Disponible à cette adresse https://linkedpolitics.project.cwi.nl/web/html/home.html.
2. Disponible à cette adresse http://www.parlgov.org/.

D.3 Méthodes

Nous présentons brièvement dans cette partie les stratégies de représentations étudiées ainsi que la méthode de création des dictionnaires.

D.3.1 Dictionnaires

Grâce aux discours parlementaires de l'article [1], nous mettons en place des dictionnaires clivants selon qu'un parti soit europhile ou eurosceptique. Pour ce faire, en utilisant la terminologie de la **section 2.1.1.2** (p. 7), nous proposons la méthodologie suivante :

1. calcul de tf_j, la fréquence du terme j dans le corpus des discours parlementaires ;
2. création de deux sous-corpus :
   (a) europhile ;
   (b) eurosceptique ;
3. calcul de tf_{ji}, la fréquence du terme j dans un des sous-corpus i.

Pour sélectionner les mots pouvant discriminer les deux idéologies, nous appliquons la formule suivante :

$$\frac{tf_{ji}}{tf_j} > 1. \tag{D.1}$$

Cela aboutit à un dictionnaire europhile de 2326 mots et un dictionnaire eurosceptique de 1155 mots.

De plus, nous cherchons à contextualiser chaque mot des dictionnaires. Pour cela, nous analysons chaque phrase dans laquelle un mot apparaît et nous calculons la moyenne des sentiments exprimés dans son contexte. Ainsi, un mot appartenant à un dictionnaire mais utilisé dans un contexte différent, peut nous signaler qu'il a été utilisé par une vision politique différente.

D.3.2 Représentations des textes

Nous nous proposons d'utiliser plusieurs représentations vectorielles des textes.

Histogramme Chaque texte est représenté par l'occurrence des mots des dictionnaires selon leurs contextes. Un texte est donc représenté par un vecteur $\boldsymbol{w} \in \mathbb{R}^4$.

Unigramme signé Nous ajoutons un suffixe *cont* à un terme d'un dictionnaire s'il est utilisé dans un contexte différent. Ainsi, les dictionnaires europhile et eurosceptique comportent respectivement 4652 et 2310 termes. Nous comptons les occurrences de chaque terme dans le texte. Il est donc décrit par un vecteur $\boldsymbol{w} \in \mathbb{R}^{6962}$.

Plongement de mots Nous utilisons les unigrammes signés obtenus précédemment. Nous représentons chaque terme par son équivalent dans l'espace de représentation *GloVe* de dimension 300. Si ce terme à un suffixe *cont*, nous multiplions sa représentation par -1. Nous calculons la moyenne des représentations obtenues par dictionnaire. Un texte est donc représenté par un vecteur $\boldsymbol{w} \in \mathbb{R}^{600}$.

D.4 Résultats

Nous cherchons à classifier des textes en fonction de la variable *état - marché*. Pour ce faire, nous utilisons les discours parlementaires sur le thème de l'énergie pour entraîner notre modèle. Chaque discours est catégorisé par *état* ou *marché* selon le parti politique qui s'exprime en fonction de la base de données *ParlGov*. Ainsi, notre jeu de données est composé de 413 textes dont 85% servent comme ensemble d'apprentissage et le reste comme ensemble de test. Nous appliquons les mêmes forêts aléatoires que dans la **section 4.5.1** (p. 47) pour obtenir les résultats présentés dans la table D.1.

TABLE D.1 – Résultats de classification *état - marché* sur l'ensemble de test.

Représentation	Précision
Histo	0.716%
Moy. GloVe	0.73%
Uni. signé	0.754%
Histo + Uni. signé + Moy. GloVe	**0.80%**

Grâce à ce modèle, nous pouvons maintenant analyser le comportement des acteurs selon les consultations de l'article [66]. Il apparaît que les acteurs, entreprises ou organismes publics, adaptent leurs réponses selon les consultations.

Prenons, l'exemple des Pays-bas. Cet État souhaite une réglementation imposée par l'État en ce qui concerne les thérapies géniques. Au contraire, il est favorable à une régulation par les marchés concernant les résidus de pesticides dans les denrées alimentaires.

D.5 Conclusion

Dans ce travail, nous nous sommes intéressés à la contribution des lobbys dans le processus parlementaire de l'Union Européenne. Nous nous sommes concentrés sur l'étude des discours parlementaires pour établir plusieurs dictionnaires prenant en compte les

sentiments. Cela nous permet d'avoir une contextualisation de l'utilisation des termes selon les différentes parties. Puis, nous utilisons ces dictionnaires pour prédire l'orientation politique de la réponse à une consultation. Nous notons alors les différences de comportement des acteurs selon le contexte des consultations.