

# Évaluation hors-ligne d'un modèle prédictif : application aux algorithmes de recommandation et à la minimisation de l'erreur relative moyenne

Arnaud de Myttenaere

▶ **To cite this version:**

Arnaud de Myttenaere. Évaluation hors-ligne d'un modèle prédictif : application aux algorithmes de recommandation et à la minimisation de l'erreur relative moyenne. Machine Learning [stat.ML]. Université paris 1 Panthéon-La Sorbonne, 2016. Français. NNT : . tel-01395290

**HAL Id: tel-01395290**

**https://theses.hal.science/tel-01395290**

Submitted on 16 Nov 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paris 1 Panthéon Sorbonne
Laboratoires SAMM - CRI
Viadeo

# THÈSE

Par Arnaud De MYTTENAERE

POUR OBTENIR LE GRADE DE DOCTEUR

SPÉCIALITÉ : Mathématiques Appliquées

ÉCOLE DOCTORALE : Sciences Mathématiques de Paris Centre (ED 386)

## Évaluation hors-ligne d'un modèle prédictif : application aux algorithmes de recommandation et à la minimisation de l'erreur relative moyenne.

Directeur de thèse : Fabrice ROSSI
Co-directrice de thèse : Bénédicte LE GRAND

Soutenue le 04 novembre 2016 devant le jury d'examen formé de :

| | | | |
|---|---|---|---|
| Nicolas Vayatis | Professeur | École Normale Supérieure de Cachan | Rapporteur |
| Patrick Gallinari | Professeur | Université Pierre et Marie Curie, Paris | Rapporteur |
| Gérard Biau | Professeur | Université Pierre et Marie Curie, Paris | Examinateur |
| Mathilde Mougeot | Maître de conférences | Université Diderot, Paris | Examinatrice |
| Boris Golden | Docteur | Partech Ventures, Paris | Examinateur |
| Fabrice Rossi | Professeur | Université Paris 1 Panthéon Sorbonne | Directeur |
| Bénédicte Le Grand | Professeure | Université Paris 1 Panthéon Sorbonne | Co-directrice |

# Résumé

L'évaluation hors-ligne permet d'estimer la qualité d'un modèle prédictif à partir de données historiques. En pratique, cette approche estime la qualité d'un modèle avant sa mise en production, sans interagir avec les clients ou utilisateurs. Pour qu'une évaluation hors-ligne soit pertinente, il est nécessaire que les données utilisées soient sans biais, c'est-à-dire représentatives des comportements observés une fois le modèle en production.

Dans cette thèse, nous traitons le cas où les données à disposition sont biaisées. A partir d'expériences réalisées au sein de Viadeo nous proposons une nouvelle procédure d'évaluation hors-ligne d'un algorithme de recommandation. Cette nouvelle approche réduit l'influence du biais sur les résultats de l'évaluation hors-ligne. Nous introduisons ensuite le contexte d'*explanatory shift*, qui correspond à une situation dans laquelle le biais réside dans la distribution de la variable cible. Des expériences menées sur les données du site de e-commerce Cdiscount et la base de données Newsgroup montrent alors que, sous certaines hypothèses, il est possible d'inférer la distribution de la variable cible afin de corriger la non-représentativité de l'échantillon d'apprentissage à disposition.

De façon plus théorique, nous nous intéressons ensuite au rôle de la fonction de perte utilisée pour la sélection d'un modèle à partir de la méthode de minimisation du risque empirique. Plus précisément, nous détaillons le cas particulier de la minimisation de l'erreur relative moyenne et nous introduisons le concept de régression MAPE (*Mean Absolute Percentage Error*). Les travaux réalisés dans ce cadre portent alors sur la consistance de l'estimateur de minimisation du risque empirique pour la régression MAPE, et sur la régression MAPE régularisée en pratique. Les expériences menées sur des données simulées ou extraites du réseau social professionnel Viadeo montrent les avantages de la régression MAPE et permettent d'illustrer des propriétés théoriques de l'estimateur obtenu.

# Remerciements

Cette thèse est le résultat d'un travail réalisé au sein de l'université Paris 1 et de Viadeo. Je tiens avant tout à remercier ces deux organismes pour avoir rendu ce travail possible, ainsi que Bénédicte Le Grand et Fabrice Rossi, mes directeurs de thèse, pour leurs conseils au cours de ces trois années et tous les enseignements qu'ils ont pu m'apporter.

Je remercie également Boris Golden pour nos échanges constructifs sur les travaux réalisés à Viadeo et tous ses précieux conseils professionnels et personnels. Merci également à Julie Séguéla pour avoir été à l'initiative de cette thèse CIFRE et m'avoir fait confiance en me recrutant chez Viadeo, et à tous les autres managers avec qui j'ai eu la chance de travailler (François, Najia, Pierre-Emmanuel et Claire).

Merci à Nicolas Vayatis et Patrick Gallinari pour avoir accepté d'être rapporteurs de cette thèse, ainsi que Gérard Biau et Mathilde Mougeot pour m'avoir fait l'honneur de participer au jury de thèse.

Plus personnellement, je tiens à remercier mes professeurs de mathématiques, dont certains m'ont particulièrement marqué, pour m'avoir introduit au monde scientifique et pour tout ce qu'ils m'ont apporté : Mme Romejon, M. Delahaye, M. Audran et M. Eiden. Merci également à mes amis de prépa (Adrien, Gautier, Lacagne, Meunier, Monville, Moreau, Palou, Pélu et Z.) pour avoir partagé nos premières réflexions scientifiques, qui auront parfois nourri des débats jusqu'à Prague, Dublin et Bruxelles des soirs de nouvel an. Puis mes amis de l'ENSAE (Anne-Elisabeth, Clotilde, Julia, Kévin, Linda, Marc, Meriem, Mustapha, Walter) et Charles de l'ENS, pour tous les bons moments passés ensemble, qui permettent de prendre du recul sur les travaux réalisés pendant la thèse. Et enfin les Kagglers férus de machine learning (Christophe, Eric, Matthieu, Nicolas, Pierre, Romain) pour l'émulation collective et les soirées et week-ends passés à coder afin d'obtenir des modèles toujours plus performants, ainsi qu'Arnaud L. et Nicolas M., à l'origine de nombreux challenges dont certains ont abouti à des travaux développés dans cette thèse.

Je remercie aussi mes collègues et amis de Viadeo (Anastasiia, Aude, Imen, Luc, Marion et Nithya) et du SAMM (Aichetou, Clara, Cécile, Cynthia, Diem, Ibrahima, Jean-Marc, Julien, Marie, Marco, Pierre, Rawya, Tsirizo) pour leur bonne humeur et

tous les moments partagés, qui ont favorisé le bon déroulement de cette thèse.

J'adresse par ailleurs des remerciements particuliers à mes trois relectrices : ma mère, Ester et Magali, avec une distinction particulière pour Ester, qui a trouvé le courage de relire le manuscrit pendant ses vacances, qui m'a adressé son soutien en toute circonstance et sans qui le déroulement de la thèse n'aurait pas été le même.

Enfin, je remercie très chaleureusement mes frères Erwan et Yann, et surtout mes parents pour leur soutien inconditionnel durant toutes ces années. À mes parents : merci !

# Table des matières

**Résumé** **i**

**Remerciements** **ii**

**Table des matières** **iv**

**Introduction** **1**

**1 Évaluation hors-ligne d'un algorithme de recommandation** **8**
1.1 Introduction . . . . . 8
1.1.1 Les systèmes évolutifs . . . . . 8
1.1.2 Problème général des prophéties auto-réalisatrices . . . . . 9
1.1.3 Plan du chapitre . . . . . 10
1.2 Motivations à Viadeo . . . . . 10
1.2.1 Systèmes de recommandation . . . . . 10
1.2.2 Reprise du problème général dans le contexte . . . . . 11
1.2.3 Recommandation de compétences . . . . . 12
1.3 Notations et cadre mathématique . . . . . 13
1.3.1 Cadre classique de l'apprentissage statistique . . . . . 13
1.3.2 Évaluation hors-ligne . . . . . 15
1.4 État de l'art . . . . . 17
1.4.1 Bandits multi-bras . . . . . 17
1.4.2 *Covariate Shift* . . . . . 18
1.4.3 Lien avec le *covariate shift* . . . . . 20
1.5 Solution proposée . . . . . 20
1.5.1 Cadre général . . . . . 20
1.5.2 Choix des lois de tirage . . . . . 21
1.5.3 Pondération des items . . . . . 23
1.5.4 Détermination des poids . . . . . 24
1.5.5 Étude de la complexité . . . . . 27
1.6 Applications . . . . . 28

1.6.1 Description des données . . . . . 28
1.6.2 Mise en évidence du biais . . . . . 29
1.6.3 Correction du biais . . . . . 30
1.7 Conclusion . . . . . 34

**2 Explanatory Shift** **36**
2.1 Introduction . . . . . 36
2.1.1 Contexte . . . . . 36
2.1.2 Plan du chapitre . . . . . 39
2.2 État de l'art . . . . . 39
2.2.1 Approche générative en classification supervisée . . . . . 39
2.2.2 Adaptation de domaine . . . . . 41
2.3 Solution proposée : une approche itérative . . . . . 43
2.3.1 Théorie . . . . . 43
2.3.2 Algorithme proposé . . . . . 45
2.4 Simulations . . . . . 46
2.4.1 Génération des données . . . . . 46
2.4.2 Les différentes approches en pratique . . . . . 48
2.4.3 Analyse des résultats . . . . . 55
2.5 Applications . . . . . 56
2.5.1 Challenge Cdiscount . . . . . 56
2.5.2 Classification de documents : cas des données Newsgroup . . . . 58
2.5.3 Résultats . . . . . 60
2.6 Conclusion . . . . . 61

**3 Minimisation de l'erreur relative moyenne** **64**
3.1 Introduction . . . . . 64
3.1.1 Contexte . . . . . 64
3.1.2 Notations et cadre général . . . . . 67
3.1.3 Problèmes théoriques . . . . . 68
3.1.4 Plan du chapitre . . . . . 71
3.2 Existence de la fonction de régression MAPE . . . . . 72
3.2.1 Cas discret . . . . . 74
3.2.2 Cas continu . . . . . 75
3.2.3 Cas général . . . . . 80
3.2.4 Non-unicité de l'optimum . . . . . 84
3.2.5 Comparaison des estimateurs MAE et MAPE . . . . . 85
3.3 Contrôle de complexité : nombres de couverture . . . . . 87
3.3.1 Introduction générale . . . . . 87
3.3.2 Lien avec les nombres de couverture . . . . . 89

3.3.3 Lien entre les nombres de couverture MAE et MAPE . . . . . 93
3.3.4 Nombres de couverture $L_p$ . . . . . 94
3.4 Contrôle de complexité : dimension de Vapnik-Chervonenkis . . . . . 95
3.4.1 Introduction générale . . . . . 95
3.4.2 Contrôle des nombres de couverture $L_p$ par la VC-dimension . . 96
3.4.3 Lien entre les dimensions de Vapnik-Chervonenkis MAE et MAPE 97
3.5 Consistance . . . . . 98
3.6 Applications . . . . . 103
3.6.1 La MAPE en pratique . . . . . 103
3.6.2 La base de données publique *cars* . . . . . 104
3.6.3 Cas pratique : modélisation de l'âge des membres du réseau social Viadeo . . . . . 106
3.7 Conclusion . . . . . 108

**4 Régression MAPE régularisée 110**
4.1 Introduction . . . . . 110
4.2 Régression MAPE régularisée en pratique . . . . . 111
4.2.1 Contexte . . . . . 111
4.2.2 MAPE : problème primal . . . . . 112
4.2.3 MAPE : problème dual . . . . . 113
4.2.4 Comparaison des problèmes d'optimisation MAE et MAPE . . . 116
4.3 Applications . . . . . 116
4.3.1 Génération des données . . . . . 116
4.3.2 Résultats . . . . . 117
4.3.3 Illustration graphique . . . . . 117
4.4 Conclusion . . . . . 118

**Conclusion et perspectives 120**

**Publications et communications 122**

**Références bibliographiques 122**

# Table des figures


1 Exemple d'un profil Viadeo . . . . . 4

1.1 Représentation schématique du cercle vicieux induit par le cercle vertueux 10
1.2 Exemple de recommandations affichées sur le site internet. . . . . 13
1.3 Exemple de recommandations envoyées par email. . . . . 14
1.4 Illustration du *covariate shift*, image extraite de Shimodaira (2000). . . 19
1.5 Distribution du nombre de compétences par membre (échelles logarithmiques) . . . . . 28
1.6 Impact d'une campagne de recommandation sur les probabilités de tirage des items . . . . . 29
1.7 Evolution des scores dans le temps pour deux algorithmes constants. . . 31
1.8 Représentation de l'évolution des scores en fonction de $p$ (le nombre de paramètres ajustés) sur deux algorithmes constants. . . . . 33
1.9 Représentation de l'évolution des scores en fonction du nombre $p$ de paramètres ajustés, pour un algorithme de filtrage collaboratif. . . . . 34

2.1 Distribution des tailles des hommes et des femmes. . . . . 37
2.2 Adaptation de domaines, une stratégie en trois étapes. . . . . 41
2.3 Représentation des bases d'apprentissage et de test en fonction de $a$ . . 47
2.4 Représentation des poids $\omega$ en fonction de $a$. . . . . 50
2.5 Représentation des densités estimées en fonction de $a$. . . . . 51
2.6 Représentation des densités d'apprentissage originales et transportées en fonction de $a$ (approche par adaptation de domaine) . . . . . 53
2.7 Convergence du modèle itératif en fonction de $a$. . . . . 54
2.8 Structure des fichiers issus de la base Newsgroup. . . . . 60
2.9 Évolution de l'estimation de la répartition des classes sur la base de test en fonction des itérations. . . . . 61
2.10 Distance de Kullback-Leibler entre la densité réelle des classes sur la base de test, et la densité obtenue par application du modèle, en fonction du nombre d'itérations. . . . . 62
2.11 Evolution de l'erreur du modèle en fonction du nombre d'itérations. . . 62

3.1 Comparaison des comportements lorsque $y \to 0$ pour la MAE et la MAPE. 70
3.2 Représentation graphique des fonctions $g^-$ et $g^+$. . . . . 82
3.3 Illustration d'une variable aléatoire avec une infinité de minima globaux pour le risque MAPE. . . . . 85
3.4 Représentation graphique d'une $\epsilon-$couverture. . . . . 90
3.5 Illustration de la dimension de Vapnik-Chervonenkis dans $\mathbb{R}^2$ dans le cas où le séparateur est une droite. . . . . 96
3.6 Représentation graphique des différents modèles obtenus sur la base de données *cars* selon la fonction de perte utilisée. . . . . 105
3.7 Distribution de l'année de naissance des membres présents sur Viadeo (parmi ceux ayant renseigné l'information). . . . . 107

4.1 Représentation de la *check-function*, aussi appelée *pinball loss*. . . . . 111
4.2 Représentation graphique des modèles $\widehat{f}_{MAE,a}$ (en pointillés bleu) et $\widehat{f}_{MAPE,a}$ (en trait plein rouge). . . . . 119

# Liste des tableaux


1.1 Ordre de grandeur des ensembles de données traités pour les algorithmes de recommandation chez Viadeo FR. . . . . 12

2.1 Synthèse des hypothèses sous-jacentes à chaque approche. . . . . 38
2.2 Résultats obtenus par application de la stratégie de *covariate shift*. . . . 49
2.3 Résultats des simulations issues de l'approche proposée . . . . . 55
2.4 Synthèse des résultats obtenus par les différentes approches. . . . . 55
2.5 Exemple de catégories issues du référentiel de produits fourni par Cdiscount.com. . . . . 57
2.6 Les catégories Newsgroup regroupées en 6 classes. . . . . 59
2.7 Effectif de chaque classe. . . . . 59
2.8 Effectif des classes sur les bases d'apprentissage et de test. . . . . 60

3.1 Comparaison de l'intérêt des différents critères selon le type d'évaluation. 66
3.2 Extrait des premières lignes de la base *cars*, disponible dans la librairie *datasets* du logiciel *R*. . . . . 104
3.3 Valeurs numériques des coefficients $a_\ell, b_\ell$ associés à chaque modèle (MSE, MAE, MAPE). . . . . 106
3.4 Résultats numériques associés aux différentes fonctions de perte sur la base de données *cars*. . . . . 106
3.5 Extrait de la base de données utilisée pour modéliser l'âge des membres du réseau social Viadeo. . . . . 108
3.6 Comparaison des différents modèles obtenus pour l'inférence de l'âge des membres, selon la fonction de perte utilisée. . . . . 108

# Introduction

*"En moulinant des millions de données [...], nous étions parvenus à décrypter peu à peu les intentions d'achat des consommateurs. Plus précisément, à prédire quel produit spécifique était susceptible d'intéresser tel internaute en particulier."* [1].

Avec l'explosion des volumes de données collectés et l'émergence de nouveaux outils d'analyse des données massives, tout internaute a déjà été confronté à la recommandation de produits ou d'articles publicitaires. Ces recommandations, souvent présentées sous formes de bannières, pop-up ou courriels, sont le résultat de longs travaux de recherche et développement de différents acteurs d'internet afin de mettre au point des modèles prédictifs aussi performants que possible. Ces modèles prédictifs, également appelés algorithmes de recommandation, permettent d'identifier le contenu le plus susceptible d'intéresser un utilisateur donné à partir des informations disponibles sur ce dernier (Shapira, 2011; Park *et al.*, 2012; Su and Khoshgoftaar, 2009). Par exemple, grâce à l'analyse de l'historique de navigation d'un internaute, il est possible d'estimer si celui-ci sera plus réactif à une publicité pour un article de mode ou pour un voyage.

Les enjeux industriels de tels algorithmes sont multiples : ils permettent d'augmenter la satisfaction et l'expérience des utilisateurs en leur proposant un article qui pourrait répondre à leur besoin, et sont également une source de revenus importante.

Aujourd'hui, l'utilisation de ce type de modèle prédictif est de plus en plus fréquente et les champs d'applications bien plus larges que la recommandation de produit ou le ciblage publicitaire.

D'autres modèles prédictifs sont également utilisés dans des secteurs d'activité très différents. Par exemple, en finance, il existe des modèles pour prédire l'évolution des cours boursiers (Fama and Miller, 1972; Taylor, 2007), ou en météorologie pour prédire les déplacements des zones pluvieuses ou l'évolution des températures (Glahn and Lowry, 1972; Michalakes *et al.*, 2001). Par ailleurs, l'apparition des nouvelles technologies est également à l'origine de nouveaux types de modèles prédictifs, comme ceux nécessitant d'anticiper les trajectoires ou des obstacles routiers pour le développement des voitures autonomes (Althoff *et al.*, 2009).

Qu'il s'agisse de la modélisation de comportements (achats, appétence pour un film

1. "On m'avait dit que c'était impossible, manifeste du fondateur de Criteo", page 97, Rudelle (2015)

ou un produit), de la prévision d'événements (météorologiques par exemple), ou de la prédiction de variables temporelles (revenus futurs d'une fonctionnalité, cours financiers, etc), les techniques de modélisation et de prévision sont désormais au centre de la stratégie de nombreuses sociétés. En entreprise, l'apparition de ces nouveaux enjeux et la volonté de modéliser les comportements des utilisateurs pour mieux les comprendre et les anticiper a favorisé l'apparition d'un nouveau métier : le *data scientist*, en charge notamment d'élaborer ce type de modèles. Si certains algorithmes sont spécifiques à un domaine, l'analyste ou le data scientist qui les élabore suit souvent la méthodologie CRISP-DM (*Cross Industry Standard Process for Data Mining*, voir Wirth and Hipp (2000)), qui décrit un processus en six étapes :

1. Identification d'un problème
2. Collecte de données pour répondre au problème
3. Analyse et préparation des données en vue de la modélisation
4. Élaboration d'un modèle
5. Évaluation du modèle
6. Mise en production

Bien que ce schéma soit très classique, plusieurs approches peuvent être choisies à chaque étape. Par exemple, la constitution de la base de données peut être effectuée à partir d'une collecte manuelle, d'un sondage d'individus, ou en considérant des données publiques ou historiques (par exemple les données de navigation ou d'achats pour un utilisateur). Pour la recommandation de produits pour le e-commerce, le data scientist peut ainsi utiliser des données sur les comportements des utilisateurs, comme les mises au panier ou les historiques d'achats. Ces données seront ensuite utilisées pour réaliser des études statistiques et élaborer un modèle prédictif permettant de répondre à une problématique particulière. Nous dirons dans ce cas que ces données constituent la base d'apprentissage du modèle prédictif.

La phase d'analyse des données est également une étape clé dans l'élaboration d'un modèle, au cours de laquelle se font l'identification des variables importantes et le prétraitement des données. Ces prétraitements auront une influence sur la performance du modèle final et peuvent être nécessaires pour rendre les données exploitables, par exemple en traitant les valeurs manquantes ou aberrantes.

Lors de l'élaboration d'un modèle, plusieurs stratégies peuvent aussi être adoptées. Selon des contraintes métier ou techniques, certains modèles peuvent être plus ou moins adaptés. En fonction de leur facilité d'interprétation et leur rapidité d'exécution, d'autres modèles peuvent être préférés. Par exemple, dans le cadre d'une segmentation où l'objectif est d'obtenir une représentation synthétique de l'ensemble des utilisateurs afin de mieux interagir avec eux, nous avons été amenés à choisir des segments ayant un sens métier très fort. A l'inverse, dans le cadre d'un modèle prédictif pour anticiper les

mouvements boursiers, il sera préférable d'avoir un modèle très performant et capable de réagir en temps réel.

Enfin, une fois le modèle choisi, la dernière étape consiste à évaluer sa performance afin de s'assurer de sa qualité avant sa mise en production. Souvent, la phase d'évaluation permet aussi de comparer différents modèles entre eux pour déterminer le plus performant. Dans cette thèse, nous verrons comment choisir, avant la mise en production, une méthode d'évaluation qui reflète la performance réelle de l'algorithme, c'est-à-dire celle mesurée une fois l'algorithme mis en production.

## Motivations

Cette thèse a été réalisée dans le cadre d'un contrat CIFRE au sein de l'entreprise Viadeo [2], site internet français créé en 2007. Viadeo est un réseau social professionnel ayant pour objectif de réunir recruteurs et chercheurs d'emploi sur une même plateforme, de permettre à chacun d'informer son réseau de son activité professionnelle, d'être visible auprès des recruteurs, et de faciliter la recherche d'emploi. Leader en France sur le marché des réseaux sociaux professionnels, Viadeo disposait à l'été 2015 d'une base de 10 millions d'utilisateurs en France, possédant chacun un profil sur lequel il est possible de renseigner ses expériences professionnelles, sa formation, ses diplômes, ainsi que ses compétences ou centres d'intérêt. Un exemple de profil est représenté à la figure 1.

En plus de ces informations, Viadeo propose de nombreuses fonctionnalités aux utilisateurs en vue de les accompagner dans leur vie professionnelle et dans le développement de leur réseau. Par exemple, il est possible d'entrer en contact avec des anciens camarades de promotion, des collègues ou des recruteurs. Il est également possible de candidater à des offres d'emploi directement à partir de son profil en ligne, d'échanger avec des recruteurs grâce à la messagerie interne, ou de suivre des entreprises afin de se tenir au courant de leur actualité. Enfin, différents services sont proposés aux utilisateurs en fonction de leur secteur d'activité, notamment pour aider les travailleurs indépendants à trouver des missions ou les recruteurs à trouver des candidats correspondant à leurs besoins.

Chez Viadeo, les modèles prédictifs sont utilisés comme outil en vue d'améliorer l'expérience des utilisateurs, en anticipant leur comportements et en leur présentant du contenu personnalisé par exemple. Ils permettent également une meilleure compréhension et une bonne anticipation des indicateurs clés de la société, par l'utilisation de tableaux de bord notamment. Les modèles prédictifs ou analyses statistiques constituent donc un élément important dans les processus d'aide à la décision, par exemple pour identifier rapidement des points d'amélioration lors du lancement de nouvelles fonctionnalités, ou pour détecter des changements de comportement des utilisateurs pouvant avoir un

2. `http://www.viadeo.fr`

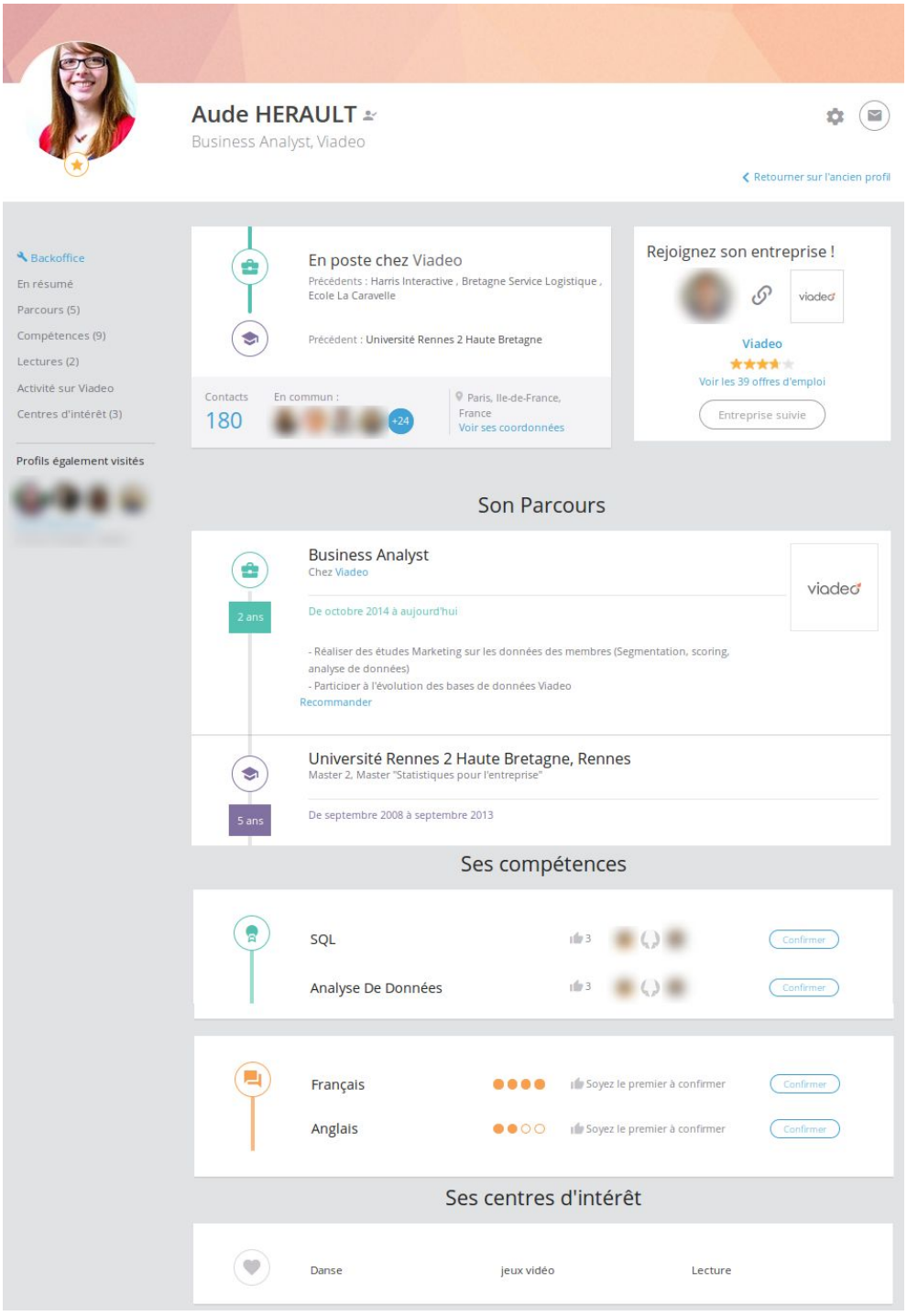



FIGURE 1 – Exemple d'un profil Viadeo

impact significatif sur l'usage ou la santé de l'entreprise (par exemple s'ils sont liés à une panne d'une fonctionnalité existante). Les enjeux liés à l'élaboration et à l'évaluation d'un modèle prédictif sont donc très importants en pratique.

Pour évaluer la qualité d'un modèle, il est courant de distinguer deux approches. D'une part, les méthodes dites hors-ligne, ou *offline*, dont l'objectif est de simuler des comportements d'utilisateurs à partir de données existantes afin d'évaluer la capacité d'un modèle à prédire ce type de comportement. D'autre part, les méthodes dites en-ligne, ou *online*, qui consistent à mettre en production un modèle prédictif afin qu'il soit testé en situation réelle sur l'ensemble ou sur un échantillon d'utilisateurs test, et à évaluer la performance de ce modèle en analysant des indicateurs clés (par exemple : taux de clics, chiffre d'affaires généré, nombre d'achats, etc.). En pratique, il est courant d'effectuer une évaluation hors-ligne (Pradel, 2013; Saxena *et al.*, 2010) afin de s'assurer de la pertinence d'un modèle prédictif avant sa mise en production, puis de réaliser une évaluation en ligne une fois le modèle déployé. Outre sa facilité de mise en œuvre, une évaluation hors-ligne possède de nombreux avantages : rapidité, coûts de développement très faibles, aucune interaction avec les utilisateurs (donc aucun impact sur l'expérience utilisateur), possibilité de comparer plusieurs modèles entre eux ou d'améliorer un modèle de façon itérative avant sa mise en production, etc.

Cependant, il est courant d'observer une différence entre la performance estimée d'un modèle avant sa mise en production et celle observée une fois le modèle déployé. Cette différence peut s'expliquer de nombreuses façons. Une cause très fréquente réside dans la non-représentativité de l'échantillon de données utilisé pour calibrer ou évaluer le modèle, qui peut généralement être détectée par une analyse des données à disposition, et par une comparaison statistique à l'ensemble des données sur lequel est appliqué le modèle (échantillon test).

Pour illustrer les problèmes de non-représentativité, prenons l'exemple d'une étude menée au sein d'un lycée, dont l'objectif est de déterminer si un élève est un garçon ou une fille à partir de son poids et sa taille. Dans le cas idéal, toutes les classes du lycée vérifient les mêmes propriétés : proportion de filles, répartition des tailles et des poids identiques entre les classes. Dans cette situation, à partir de n'importe quelle classe observée il sera possible d'établir un modèle qui permettra de bien distinguer les élèves selon leur sexe en fonction uniquement de leur poids ou de leur taille. Dans le cas contraire, plusieurs hypothèses sont possibles.

Considérons par exemple le cas d'une classe, notée classe A, qui comporte une majorité de très grands élèves (par exemple la classe avec option basket-ball). Alors la classe A possède des propriétés statistiques (répartition des variables observées : poids et taille) différentes de l'ensemble des autres classes. Un modèle calibré sur une classe sélectionnée au hasard se généralisera mal à cette classe particulière et le modèle ne sera pas capable de distinguer de façon pertinente les élèves de la classe A selon leur sexe à partir des données observées (poids et taille). Pour pallier ce problème, nous

verrons que la théorie du *covariate shift* (voir Shimodaira (2000)) propose de pondérer les individus d'une classe de façon à modéliser la classe A avant la calibration du modèle. Cette pondération permettra d'avoir un modèle spécifique à la classe A, plus performant que le modèle générique.

De même, si une classe notée classe B comporte la même répartition des observations (poids et taille) mais une différence au niveau de la variable cible (par exemple une proportion de filles très différente des autres classes), un modèle calibré sur une classe choisie aléatoirement ne se généralisera pas bien à la classe B et ses prédictions sur cette classe seront susceptibles de comporter de nombreuses erreurs (par exemple la surestimation du nombre de filles). Ce cas de biais sur la variable cible n'est à notre connaissance pas encore traité dans la littérature et sort du contexte du *covariate shift* qui pouvait s'appliquer au cas précédent.

## Plan du manuscrit

L'objectif de cette thèse est de détailler différents problèmes liés à l'évaluation de modèles et d'illustrer les solutions proposées sur des jeux de données publics, ou sur des cas d'applications concrets rencontrés à Viadeo au cours de cette thèse CIFRE.

Dans le premier chapitre, nous verrons quel rôle peut jouer la qualité des données présentes dans la base d'apprentissage, et comment bien évaluer un modèle lorsque la base d'apprentissage possède des propriétés différentes de la base réelle. Plus précisément, nous verrons les difficultés liées à l'évaluation hors-ligne d'un algorithme de recommandation et le biais introduit par les algorithmes existants sur le comportement des utilisateurs et donc sur les données collectées. Nous proposerons une procédure d'évaluation hors-ligne sans biais en nous inspirant de la théorie du *covariate shift* et nous illustrerons les résultats à partir de données réelles issues du réseau social professionnel Viadeo.

Dans le deuxième chapitre nous proposerons une adaptation de la stratégie issue du *covariate shift* afin de diminuer le biais entre les bases d'apprentissage et de test, dans le cas ou les hypothèses classiques du *covariate shift* (qui seront vues dans le chapitre 1) ne sont pas respectées. Nous illustrerons d'abord notre approche sur des données simulées avant de réaliser des expériences sur des données réelles. Comme dans le cas du *covariate shift*, nous verrons que l'approche proposée consiste à pondérer les observations de la base d'apprentissage en fonction de poids dépendant des observations elles-mêmes.

Dans le troisième chapitre, nous détaillerons pour les modèles de régression le cas de l'erreur relative moyenne, qui est fréquemment utilisée pour mesurer la performance d'un modèle dans un cadre industriel, en raison de sa facilité d'interprétation. Nous introduirons alors la notion de régression MAPE (*Mean Absolute Percentage Error*). De façon immédiate, ce critère peut être vu comme une erreur absolue pondérée par l'inverse de la quantité à estimer. L'objectif de ce chapitre sera alors de montrer l'existence et la consistance de l'estimateur de minimisation du risque empirique dans ce cas particulier.

Nous illustrerons les résultats par un cas pratique sur les données de Viadeo.

Enfin, le dernier chapitre sera une extension du chapitre 3 dans un cadre plus général. Nous nous intéresserons au problème de la minimisation de l'erreur relative moyenne pour les régressions non-paramétriques régularisées. Nous aborderons la résolution du problème d'optimisation en pratique et illustrerons l'intérêt des régressions MAPE non-paramétriques sur des données simulées.

# Chapitre 1

# Évaluation hors-ligne d'un algorithme de recommandation

## 1.1 Introduction

Dans ce chapitre, nous discuterons des problèmes d'évaluation d'un modèle prédictif lié à un système évolutif, c'est-à-dire évoluant au cours du temps. En particulier, l'objectif de ce chapitre est de proposer une nouvelle procédure d'évaluation hors-ligne adaptée à ce type de système, et d'illustrer les résultats à partir de données réelles issues du réseau social Viadeo.

### 1.1.1 Les systèmes évolutifs

Nous avons vu en introduction que l'élaboration d'un modèle prédictif s'effectue généralement selon une procédure en six étapes, décrites par la méthodologie CRISP-DM (Wirth and Hipp, 2000). En pratique, une fois qu'un modèle est mis en production, il est fréquemment amélioré selon une technique classique en trois étapes :

1. Collecte de nouvelles données : une partie des utilisateurs peut faire des retours sur le modèle en production de façon directe (par sondage par exemple), ou indirecte (par l'observation de leur comportement). Cette phase permet d'observer les premiers résultats du modèle déployé.
2. Amélioration du modèle : les premiers travaux consistent à analyser les données collectées pour en extraire de l'information pertinente et trouver des pistes d'amélioration du modèle en production. Après cette phase d'analyse, il s'agit de développer un nouveau modèle tenant compte des points d'amélioration identifiés. Une fois le modèle développé, il pourra être mis en production et testé sur une partie ou l'ensemble des utilisateurs.

3. Mise en production du nouveau modèle : cette étape consiste à mettre un modèle prédictif en production pour le rendre disponible auprès de l'ensemble ou d'une partie des utilisateurs, et comparé au modèle précédent selon la technique de l'A/B testing (Kohavi, 2015).

Les itérations successives permettent alors de comparer différents modèles et de valider les hypothèses faites par les analystes ou experts métier. Ce processus est particulièrement bien adapté aux systèmes évoluant dans le temps, aussi appelés systèmes évolutifs, pour lesquels la collecte et l'analyse régulière des retours d'utilisateurs permettent d'adapter rapidement le modèle en cas de changement de comportement des utilisateurs.

### 1.1.2 Problème général des prophéties auto-réalisatrices

Dans le cadre des systèmes évolutifs, et en particulier sur internet, les comportements des utilisateurs peuvent être influencés par de multiples facteurs : formulation des informations présentées ou recommandées, choix des images, couleur et taille des polices, etc. Chacun de ces paramètres peut avoir une influence sur l'utilisateur final et donc sur les données collectées.

Par exemple, l'ordre des résultats présentés dans le cadre des moteurs de recherche sur internet influence fortement les comportements des utilisateurs : à pertinence égale un utilisateur est plus susceptible de cliquer sur les articles proposés en haut de page du moteur de recherche Google (Caphyon, 2014), en témoignent également quelques articles de blog, comme Hong (2014), ou Dataiku (2014). Suite à une compétition organisée par Yandex, moteur de recherche russe, sur la plateforme de data-science Kaggle [1], Dataiku (2014) évoque en effet ce biais par la remarque suivante : « *The problem is that the scoring metric was done with regard to the clicks obtained using Yandex initial ranking... And users tend to click on URLs appearing at the top of the page* ». En d'autres termes, le comportement des utilisateurs ayant été influencé par l'algorithme existant, les résultats d'une évaluation hors-ligne basée sur ces comportements dépendent de ce même algorithme.

Un phénomène similaire est observé au sein des sites de vente en ligne, où, de façon évidente, les produits recommandés à un utilisateur ont plus de chance d'être achetés que ceux qui ne le sont pas.

Cependant, l'analyse des données réelles fournies par Viadeo montre que ce processus vertueux induit un processus parallèle non désiré, aussi appelé cercle vicieux, dû à l'algorithme en production et qui biaise les résultats de l'évaluation hors-ligne, donc l'amélioration de la fonctionnalité en cours. Finalement, le processus complet peut-être représenté schématiquement par la figure 1.1.

1. voir `https://www.kaggle.com/c/yandex-personalized-web-search-challenge` pour plus de détails.

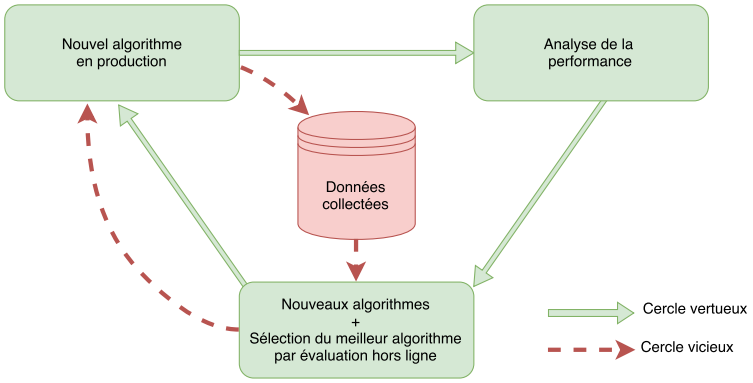



FIGURE 1.1 – Représentation schématique du cercle vicieux induit par le cercle vertueux

En conséquence, en se basant sur les comportements des utilisateurs observés sur une période donnée, un nouveau modèle sera mieux valorisé s'il est capable de recommander les produits proposés durant la période d'observation. Ce résultat est particulièrement vrai lorsque le coût d'acceptation (par exemple le prix du produit recommandé) est faible. Par exemple, en nous appuyant sur des données réelles fournies par Viadeo, nous verrons dans ce chapitre que le biais est très important dans le cas de la recommandation de compétences, cas pour lequel le coût pour l'utilisateur est nul.

### 1.1.3 Plan du chapitre

La suite de ce chapitre est organisée de la façon suivante. Dans la section 1.2 nous introduisons les algorithmes de recommandation dans le cadre de Viadeo et mettant en évidence le biais associé aux algorithmes en production dans le cas particulier de la recommandation de compétences.

Dans la section 1.3 nous définissons les notations utilisées et le contexte mathématique associé au biais présenté.

Dans la section 1.4 nous détaillons les solutions proposées dans l'état de l'art.

Enfin, dans la section 1.5 nous développons la solution proposée et présentons les résultats dans la section 1.6, sur des données réelles issues de Viadeo.

## 1.2 Motivations à Viadeo

### 1.2.1 Systèmes de recommandation

Les algorithmes de recommandation ont pour objectif de recommander un ensemble d'items, ou produits, à un utilisateur particulier en fonction de ses données personnelles (âge, région géographique, etc.) ou comportementales (historique des sites visités, produits achetés, etc.). Ces algorithmes peuvent être utilisés pour répondre à de multiples

besoins industriels (ciblage publicitaire, segmentation de la clientèle, diffusion d'offre d'emploi, ...), et ont beaucoup évolué ces dernières années, en témoignent les nombreuses publications scientifiques sur le sujet (Adomavicius and Tuzhilin, 2005; Park *et al.*, 2012; Shapira, 2011). Aujourd'hui, ils sont très présents et facilitent l'expérience utilisateur de la plupart des sites internet.

Les systèmes de recommandation constituent en effet un axe de recherche important, sur internet notamment, car ils contribuent directement aux revenus générés (recommandations de produits pour les sites commerciaux, ciblage publicitaire), à l'activité (recommandation de contacts dans les réseaux sociaux), et à la satisfaction des utilisateurs (interfaces personnalisées, ...). Dans un contexte industriel, l'élaboration d'un algorithme de recommandation s'effectue en plusieurs étapes selon la méthodologie CRISP-DM (Wirth and Hipp, 2000) vue en introduction.

Afin d'étudier le biais lié à la qualité des données ayant été utilisées durant la phase d'apprentissage, nous proposons de traiter le cas particulier du biais introduit lors de l'élaboration et l'évaluation d'un algorithme de recommandation. Plus précisément, l'objectif de ce chapitre est d'analyser comment l'évaluation hors-ligne d'un algorithme de recommandation peut être biaisée par le processus de collecte des données, et de proposer une nouvelle procédure d'évaluation permettant de réduire ce biais.

### 1.2.2 Reprise du problème général dans le contexte

En pratique, il est courant d'évaluer la qualité d'un algorithme de recommandation en simulant des comportements d'utilisateurs, selon une procédure hors-ligne, ne nécessitant aucune interaction avec les utilisateurs et qui peut donc être réalisée très rapidement et sans risque sur l'expérience utilisateur. Nous détaillerons cette procédure à la section 1.3.2. Cependant, comme nous l'avons évoqué à la section 1.1.2, ce type de procédure peut être biaisé par les précédents algorithmes en production qui ont influencé les comportements des utilisateurs.

Chez Viadeo, de nombreux modèles de recommandation sont utilisés. Par exemple, un algorithme est utilisé pour suggérer des offres d'emplois pertinentes aux utilisateurs en fonction de leur profil et de leurs précédentes expériences, tandis qu'un autre algorithme suggère des contacts aux utilisateurs afin de les aider à retrouver des connaissances et augmenter la taille de leur réseau. Pour améliorer la qualité du profil des utilisateurs, un algorithme propose également des compétences pertinentes aux utilisateurs en fonction de leur domaine d'activité ou des compétences déjà renseignées. Ces recommandations peuvent prendre deux formes différentes : soit elles sont présentées dans l'interface du site internet (voir figure 1.2), sur la page d'accueil de l'utilisateur ou sur son profil, soit elles sont envoyées par courriel à l'utilisateur qui n'a plus qu'à les valider (voir figure 1.3).

La table 1.1 donne quelques ordres de grandeur des nombres d'utilisateurs et d'items

sur la plateforme française du site Viadeo, en fonction des algorithmes à étudier.

| Recommandation visée | Nombre d'utilisateurs ($\|\mathcal{U}\|$) | Nombre d'items ($\|\mathcal{I}\|$) |
|---|---|---|
| Offres d'emplois | $10 \cdot 10^6$ | $2 \cdot 10^3$ |
| Compétences | $10 \cdot 10^6$ | $3 \cdot 10^4$ |
| Contacts | $10 \cdot 10^6$ | $10 \cdot 10^6$ |

TABLE 1.1 – Ordre de grandeur des ensembles de données traités pour les algorithmes de recommandation chez Viadeo FR.

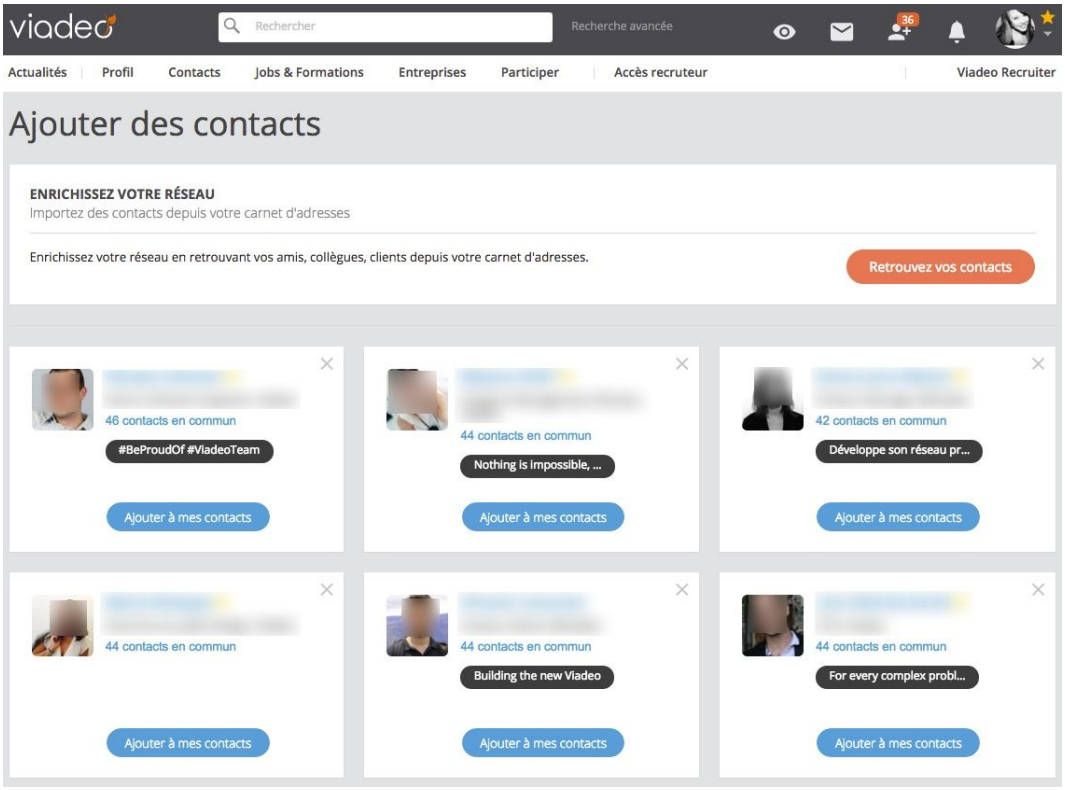
Ces algorithmes étant en production depuis plusieurs années, une grande partie des données collectées sur l'activité des utilisateurs liée à la consultation d'offres d'emplois ou de profils d'utilisateurs a été influencée par ces algorithmes. Pour évaluer un nouvel algorithme à partir des comportements historiques des utilisateurs, il est donc nécessaire de tenir compte de ce biais.

### 1.2.3 Recommandation de compétences

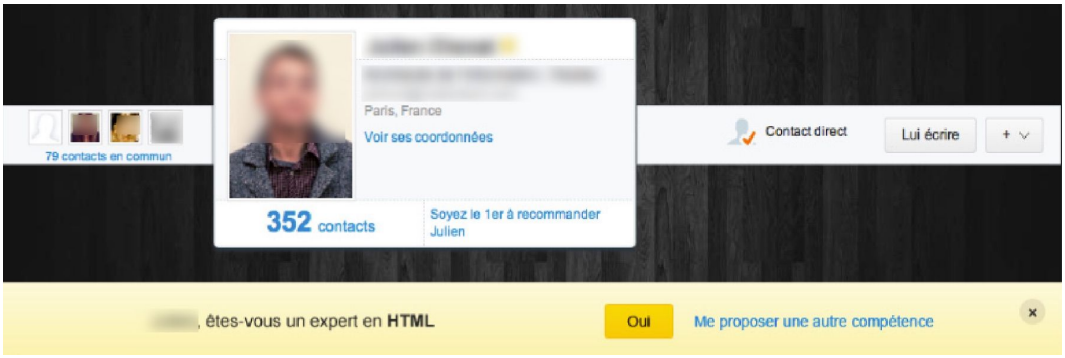
Pour illustrer le biais nous proposons dans ce chapitre d'étudier le cas particulier de la recommandation de compétences. La recommandation de compétences est un élément important dans la stratégie de Viadeo pour améliorer la qualité des profils des utilisateurs du réseau social professionnel. Plus précisément, une recommandation de compétences consistera ici en un courriel envoyé à un membre, afin de lui proposer d'ajouter une compétence à son profil. L'ensemble des compétences pouvant être recommandées est très large, allant des compétences les plus fréquentes sur l'ensemble des utilisateurs de la plateforme (marketing, gestion de projet, communication), aux plus spécifiques (C++, chirurgie, traduction, etc.).

Afin de comprendre le biais lié aux précédents algorithmes en production, prenons par exemple le cas extrême d'un algorithme proposant la compétence *marketing* à tous les utilisateurs. Alors, peu de temps après la mise en production de cet algorithme, la compétence *marketing* sera sur-représentée. En conséquence, en se basant sur les données collectées, la qualité d'un algorithme proposant la compétence marketing sera sur-estimée par rapport à sa performance réelle.

Par ailleurs, le biais dû aux précédentes campagnes de recommandation est d'autant plus important que le coût pour les utilisateurs à accepter les recommandations est faible. Le coût pour un utilisateur à accepter une compétence étant nul, le taux d'acceptation des items recommandés est en effet bien plus important que dans le cas de la recommandation de produits ayant un coût élevé (produits pour le e-commerce par exemple).




(a) Recommandation de contacts



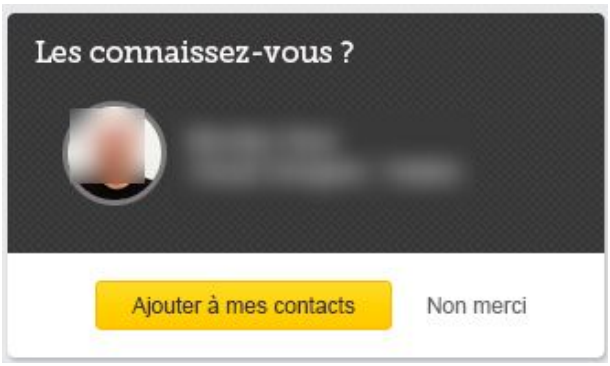

(b) Recommandation de compétence

FIGURE 1.2 – Exemple de recommandations affichées sur le site internet.

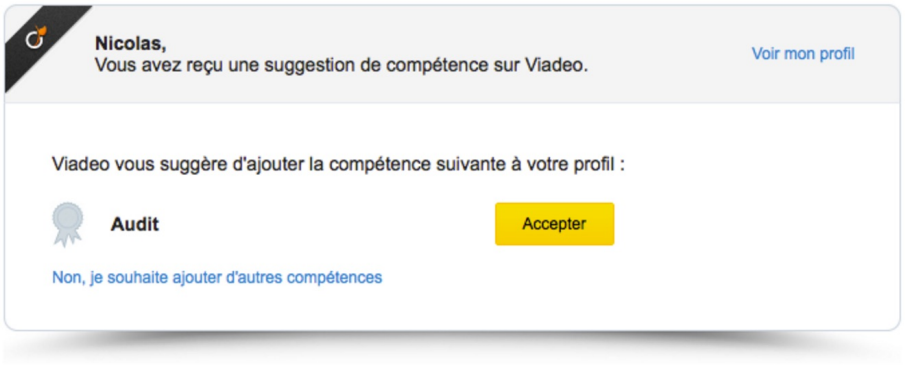
## 1.3 Notations et cadre mathématique

### 1.3.1 Cadre classique de l'apprentissage statistique

Dans la suite de ce chapitre, nous utilisons les notations classiques du contexte de régression (voir Neter *et al.* (1996) par exemple), où les données sont entièrement




(a) Recommandation de contacts




(b) Recommandation de compétences

FIGURE 1.3 – Exemple de recommandations envoyées par email.

décrites par un couple de variables aléatoires $Z = (X, Y)$ à valeurs dans $\mathcal{X} \times \mathcal{Y}$, ou $\mathcal{X}$ et $\mathcal{Y}$ sont deux sous-espaces de $\mathbb{R}^d$ et $\mathbb{R}$, avec $d$ est un entier naturel strictement positif. Notre objectif sera de trouver un modèle approchant les observations, qui soit une fonction mesurable $g$ de $\mathcal{X}$ dans $\mathcal{Y}$ telle que $g(X)$ est "proche" de $Y$.

La qualité d'un modèle est mesurée via une **fonction de perte** $l$ (généralement appelée *loss function* dans la littérature), de $\mathbb{R}^2$ dans $\mathbb{R}^+$. Alors l'erreur ponctuelle d'un modèle $m$ en $X$ est donnée par $l(m(X), Y)$ et le **risque** d'un modèle $m$ vis-à-vis d'une fonction de perte $l$ est défini par

$$L_l(m) = \mathbb{E}(l(m(X), Y)). \tag{1.1}$$

Par exemple, dans le cas des moindres carrés, la fonction de perte, $l_2 = l_{MSE}$, est définie par $l_2(p, y) = (p - y)^2$. Dans le cas où l'on dispose d'observations $(x_i, y_i)$ indépendantes et identiquement distribuées selon la loi de $(X, Y)$, on appelle **risque empirique** la quantité définie par

$$\widehat{L}_l(m)_n = \frac{1}{n} \sum_{i=1}^{n} l(m(X_i), Y_i).$$

Nous noterons $s_n(m)$ le **score empirique** de l'algorithme $m$ défini par

$$\widehat{s_n}(m) = -\widehat{L}_l(m)_n. \tag{1.2}$$

En pratique, le modèle $m^\star$ retenu est généralement celui minimisant le risque empirique ou, de façon symétrique, maximise le score empirique. Le modèle $m^\star$ vérifie donc

$$m^\star = \arg\min_m \sum_{i=1}^{n} \ell(m(X_i), Y_i). \tag{1.3}$$

Dans ce chapitre, nous supposons que nous observons deux jeux de données : d'une part l'ensemble d'apprentissage, et d'autre part l'ensemble de test, aussi appelées bases d'apprentissage et de test. Nous notons alors $p_{train}(E)$ (resp. $p_{test}(E)$) la probabilité de l'événement $E$ sur la base d'apprentissage (resp. de test). Dans le cas particulier de l'évaluation hors-ligne d'un algorithme de recommandation, nous notons $\mathcal{U}$ l'ensemble des utilisateurs et $\mathcal{I}$ l'ensemble des items. Un utilisateur $U$ est alors caractérisé par $\mathcal{I}_U$, l'ensemble des items qui lui sont associés, et nous noterons $U_{-I}$ l'utilisateur $U$ associé à tous ses items excepté $I$. En d'autres termes, nous avons la relation $\mathcal{I}_{U_{-I}} = \mathcal{I}_U \setminus \{I\}$.

### 1.3.2 Évaluation hors-ligne

En pratique, l'évaluation de la qualité d'un algorithme constitue une étape essentielle à sa validation. Cette évaluation a pour but d'estimer la performance d'un algorithme selon des critères statistiques ou métiers (taux de clics, proportion ou nombre de recommandations acceptées, ...), et éventuellement de quantifier la marge d'amélioration du modèle déployé. Pour cela, trois approches d'évaluation peuvent être envisagées (voir Beel *et al.* (2013)) : *user study*, en ligne (*online*) et hors-ligne (*offline*).

La première approche, dite ***user study***, consiste à envoyer des recommandations à un ensemble d'utilisateurs, pour qu'ils en évaluent manuellement la pertinence. Lors de ce type d'expérimentation, les utilisateurs ont conscience de faire partie d'un processus de test et font directement des retours critiques pour améliorer la pertinence des recommandations.

La seconde approche, dite ***en ligne***, consiste à envoyer des recommandations à un ensemble d'utilisateurs et à mesurer la qualité d'un algorithme selon un critère de performance déterminé en amont (taux de clics, pourcentage de recommandations validées, ...). Ici, les utilisateurs ne sont pas conscients qu'ils appartiennent à une population de test, et plusieurs algorithmes peuvent être comparés sur des populations distinctes afin de sélectionner l'algorithme le plus performant avant la mise en production selon la technique de l'A/B testing (Kohavi, 2015).

Enfin, une troisième méthode d'évaluation, dite ***hors-ligne*** (Pradel, 2013; Shani and Gunawardana, 2011), s'appuie sur l'historique des comportements des utilisateurs jusqu'à un instant donné, et cherche à inférer leur comportement futur à partir des

comportements passés. Cette approche consiste à utiliser les données connues sur une population test pour évaluer les algorithmes selon le protocole suivant :

i- sélectionner un utilisateur $u \in \mathcal{U}$ ;

ii- isoler un (ou plusieurs) item(s) de l'utilisateur sélectionné ;

iii- simuler des recommandations à cet utilisateur, à partir d'un algorithme $g$, sans prendre en considération le(s) item(s) isolé(s) ;

iv- mesurer la qualité des recommandations effectuées à l'utilisateur $u$ par l'algorithme $g$.

Dans la suite de ce chapitre, nous notons $P(U = u)$ la probabilité d'isoler l'utilisateur $u$ et $P(I = i|U = u)$ la probabilité d'isoler l'item $i$ en ayant isolé l'utilisateur $u$.

La procédure d'évaluation hors-ligne réalisée à l'instant $t$ est alors notée $\mathcal{P} = \mathcal{P}(U, I, \ell, t)$. Cette procédure dépend d'une fonction de perte $\ell$ (ou d'une fonction de score $s$ vérifiant $s = -\ell$).

Un algorithme de recommandation, noté $g$, est une fonction définie sur l'ensemble des utilisateurs et à valeurs dans l'ensemble des items, $g : \mathcal{U} \to \mathcal{I}$. Nous notons $g_k(u_{-i})$ les $k$ recommandations soumises à l'utilisateur $u$ en ayant isolé $i$ de la base d'apprentissage, un des items associés à $u$ (par exemple $i$ peut être un produit acheté par $u$, un film noté par l'utilisateur $u$, une offre d'emploi à laquelle $u$ a candidaté, etc.).

La précédure d'évaluation d'un algorithme est alors définie par l'algorithme 1.

**Algorithme 1 :** Evaluation hors-ligne

**Paramètres :** $k$, le nombre de recommandations
score $= 0$
**for** n $= 0$, $n < N_{simu}$, n++ **do**
  tirer un utilisateur $u$, avec probabilité $P(U = u)$
  isoler aléatoirement un de ses items $i$, selon $P(I = i|U = u)$
  calculer $g_k(u_{-i})$
  score+ $= -l(g_k(u_{-i}), i)$
**end for**
**return** score$/N_{simu}$

En comparaison avec les procédures *user study* et en ligne, les tests hors-ligne ont l'avantage de fournir des résultats très rapidement et de pouvoir tester des algorithmes sans risque d'impacter négativement le produit (revenus, activité) et l'expérience utilisateur. Par ailleurs, cette approche permet de confronter plusieurs algorithmes entre eux avant de tester en ligne le plus pertinent, ou d'itérer sur les résultats hors-ligne pour proposer un algorithme pertinent dès le premier test réel, bien que l'optimisation d'un critère de performance théorique ne conduise pas toujours au meilleur algorithme en pratique (McNee *et al.*, 2006; Said *et al.*, 2013).

## 1.4 État de l'art

### 1.4.1 Bandits multi-bras

A notre connaissance, le problème du biais lors de l'évaluation hors-ligne d'un algorithme de recommandation a été peu traité dans la littérature, et les articles à ce sujet proposent principalement une approche par bandits multi-bras (voir Li *et al.* (2011); Mary *et al.* (2014) par exemple). Nous avons vu en introduction que la réalisation d'un modèle prédictif s'effectue généralement de façon itérative, en alternant des phases de collecte de données et d'analyse des résultats afin d'améliorer le modèle en productif.

L'utilisation de bandits multi-bras consiste à optimiser les phases de collecte des données, dites d'exploration, utilisées pour en déduire des liens entre les profils des utilisateurs et les items les plus pertinents à recommander. Ainsi, il s'agit de minimiser le nombre de recommandations aléatoires envoyées pour avoir des retours des utilisateurs. Cette approche est particulièrement adaptée dans le cas où les items sont renouvelés régulièrement, ou plus généralement lorsqu'il est fréquent d'être en présence d'items pour lesquels le nombre de retours d'utilisateurs est trop faible pour être exploitable. Il s'agit par exemple du cas de la recommandation d'articles publicitaires (Tang *et al.*, 2013) ou d'articles de presse (Li *et al.*, 2010).

L'approche par bandits multi-bras (Li *et al.*, 2011; Mary *et al.*, 2014) trouve cependant ses limites dans sa mise en pratique, qui nécessite une phase d'exploration pour constituer une base d'apprentissage aléatoire, permettant de s'assurer que, sur le long terme, l'ensemble des actions sont explorées. Dans le cas des algorithmes de recommandation, cela se traduit par la mise en place d'une phase préliminaire pour constituer une base d'apprentissage non-biaisée, durant laquelle des produits sont recommandés aléatoirement à un ensemble d'utilisateurs. Cette phase d'exploration, bien que nécessaire, n'est souvent pas réalisée afin de ne pas détériorer la qualité du service aux yeux de l'utilisateur (les recommandations devant être aléatoires elles sont peu pertinentes lors de la phase d'exploration).

De plus, en pratique il est courant de disposer d'une base de données collectées sur les utilisateurs à partir de la fonctionnalité en cours. Dans ce cas il apparaît pertinent d'utiliser les informations à disposition plutôt que de procéder à une nouvelle phase exploratoire. Cette approche a été évoquée par Li *et al.* dans Li *et al.* (2011). Li *et al.* proposent alors une nouvelle approche d'évaluation de modèles par bandits multi-bras, basée sur la simulation de comportements à partir de données collectées, limitant ainsi les phases d'exploration. Li *et al.* montrent alors la pertinence de l'approche sous l'hypothèse que les données à disposition ont été engendrées de façon aléatoire, où chaque couple $(u, i)$ a été exploré de façon équiprobable. Dans le cadre des données disponibles chez Viadeo, les comportements des utilisateurs ont été influencés par un algorithme en production et les données à disposition ne vérifient pas les hypothèses de

Li *et al.* (2011).

Dans ce chapitre, nous proposons donc une nouvelle méthode d'évaluation hors-ligne permettant de diminuer ce biais. Cette méthode sera inspirée de la théorie du *covariate shift* et validée par des expériences menées sur les données de Viadeo.

### 1.4.2 *Covariate Shift*

Pour être exploitable, un modèle prédictif doit pouvoir être appliqué à des données nouvelles, par exemple pour prédire avec une bonne performance les comportements de nouveaux clients. Pour cela, les nouvelles données de test doivent avoir des propriétés similaires à celles présentes dans la base d'apprentissage. Par exemple, dans le cas d'un modèle estimant la taille des individus à partir de leur poids, si la base d'apprentissage est constituée à 90% de femmes, alors le modèle ne sera pas bien adapté à une population comportant 90% d'hommes car hommes et femmes possèdent des propriétés physiologiques différentes. Dans une telle situation, les bases d'apprentissage et de tests possèdent des propriétés différentes et on parle de *covariate shift* (Shimodaira, 2000).

En reprenant les notations introduites à la section 1.3.1, le modèle $m^\star$ retenu est généralement celui minimisant le risque empirique, défini à l'équation 1.3.

Toutefois, dans le cas où les bases de test et d'apprentissage possèdent des propriétés différentes, l'estimateur défini ci-dessus est sous-optimal, dans le sens où ce modèle aura une bonne capacité prédictive sur un jeu de données ayant des propriétés similaires à la base d'apprentissage, mais une performance moindre sur un jeu de données ayant des propriétés similaires à la base de test.

Dans ses travaux (voir Shimodaira (2000)), Shimodaira montre que la bonne stratégie pour traiter ce type de problème est de pondérer les observations de l'échantillon d'apprentissage afin de *mimer* l'échantillon de test. En notant $P_B(X = x)$ la valeur de la densité de $X$ en $x$ sur la base $B$, cette approche nécessite les deux hypothèses suivantes :

- $H1$) la relation entre la variable explicative et la variable cible est constante entre les bases d'apprentissage et de test : $P_{test}(Y|X) = P_{train}(Y|X)$,
- $H2$) le support de $X$ sur la base de test est inclus dans le support de $X$, de sorte que pour tout $x$, $\frac{p_{test}(X=x)}{p_{train}(X=x)} < \infty$.

Dans ce cas, le modèle optimal, $m^\star_\omega$ est donné par

$$m^\star_\omega = \arg\min_m \sum_{i=1}^{n} \omega_i \ell(X_i, Y_i) \quad \text{où} \quad \omega_i = \frac{p_{test}(X = x_i)}{p_{train}(X = x_i)}.$$

Ce résultat découle de la proposition 1.

**Proposition 1** (Covariate Shift)**.** *Soit $m$ un modèle, $\ell$ une fonction de perte, $(X_{train}, Y_{train})$ et $(X_{test}, Y_{test})$ des couples de variables aléatoires telles que*

$$p(Y_{train}|X_{train}) = p(Y_{test}|X_{test}).$$

*Alors, en définissant les poids* $\omega(x)$ *par* $\omega(x) = p(X_{test} = x)/p(X_{train} = x)$ *on a*

$$\mathbb{E}_{train}\left[\omega(X)\ell(Y, m(X))\right] = \mathbb{E}_{test}\left[\ell(Y, m(X))\right].$$

*Démonstration.* Ce résultat est immédiat par le calcul :

$$\begin{aligned}
\mathbb{E}_{train}\left[\omega(X)\ell(Y, m(X))\right] &= \int \frac{p_{test}(x)}{p_{train}(x)}\ell(y, m(x))p_{train}(x, y)dxdy, \\
&= \int \frac{p_{test}(x)}{p_{train}(x)}\ell(y, m(x))p_{train}(y|x)p_{train}(x)dxdy, \\
&= \int p_{test}(x)p_{train}(y|x)\ell(y, m(x))dxdy.
\end{aligned}$$

Or $p_{train}(y|x) = p_{test}(y|x)$ par hypothèse, donc

$$\mathbb{E}_{train}\left[\omega(X)\ell(Y, m(X))\right] = \int p_{test}(x)p_{test}(y|x)\ell(y, m(x))dxdy,$$

ce qui permet de conclure. □

Une illustration de cette stratégie de pondération est donnée par la figure 1.4, qui représente les simulations issues de l'article original sur le *covariate shift* (Shimodaira, 2000). Sur cette figure, on remarque que l'espace des observations est plus important sur la base d'apprentissage (à gauche) que sur la base de test (à droite), et que la relation entre les données sur la base d'apprentissage ne satisfait pas une relation linéaire. En revanche, une relation linéaire apparaît plus adaptée sur la base de test, qui comprend des observations uniquement sur un sous-intervalle de l'ensemble d'apprentissage ($[-0.5, 0.5]$).

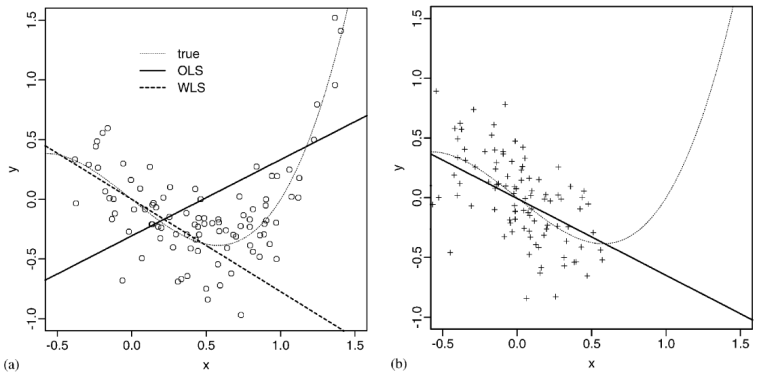



FIGURE 1.4 – Illustration du *covariate shift*, image extraite de Shimodaira (2000).

Dans ce cas particulier, l'utilisation des pondérations permet d'associer un poids nul à toutes les observations de la base d'apprentissage qui ne sont pas présentes sur l'ensemble de définition de la base de test, et ainsi de mimer la base de test. Le modèle calibré de cette façon (modèle WLS pour *Weighted Least Square*) coincide bien mieux avec la base de test que celui calibré sur la base d'apprentissage non pondérée (modèle OLS pour *Ordinary Least Square*).

### 1.4.3 Lien avec le *covariate shift*

Comme présenté dans la section 1.4.2, une hypothèse nécessaire à la stratégie de pondération est la conservation de la relation entre les observations et la valeur de la variable cible. Plus formellement, cela correspond à la constance de $P(Y|X)$ entre les bases d'apprentissage et de test, pour tout $X$ donné.

En reprenant les notations introduites à la section 1.3.1, dans le contexte de la recommandation de compétences l'espace $\mathcal{X}$ désigne l'ensemble de combinaisons sur l'ensemble des compétences $\mathcal{I}$, les observations $X$ sont représentées par l'information disponible sur chaque utilisateur, $\mathcal{I}_U$, et la variable cible $Y$ est l'ensemble de compétences acquises par un utilisateur $U$. L'hypothèse de constance de $P(Y|X)$ signifie alors que, pour toute compétence $i \in \mathcal{I}$ et pour tout utilisateur $u \in \mathcal{U}$ étant associé aux items $\mathcal{I}_u$, la probabilité que $u$ dispose de la compétence $i$ est constante entre les bases de test et d'apprentissage. En revanche, comme les comportements des utilisateurs ont été influencés par les algorithmes en production, la répartition de l'information à disposition sur l'ensemble des utilisateurs diffère entre les bases d'apprentissage et de test.

Le problème identifié à Viadeo s'inscrit dans le cadre de la théorie du *covariate shift*. Dans la section suivante nous proposons donc une approche inspirée du *covariate shift* pour réaliser une évaluation hors-ligne sans biais d'un algorithme de recommandation. Les expériences menées sur les données réelles issues du réseau social Viadeo permettent en effet de mettre en évidence le biais et l'intérêt de l'utilisation de l'approche proposée pour diminuer ce biais d'évaluation.

## 1.5 Solution proposée

### 1.5.1 Cadre général

En reprenant les notations introduites à la section 1.3.2, l'algorithme de recommandation $g_k^\star$ optimal est celui défini par

$$g_k^\star = \arg\min_{g_k} \mathbb{E}[\ell(g_k(U_{-I}), I)].$$

En pratique, la quantité à minimiser est estimée par $\max_g s(g)$ où $s(g)$ représente le score empirique de l'algorithme $g$ défini par :

$$s(g) = -\sum_{(u,i)\in\mathcal{U}\times\mathcal{I}} P(U=u, I=i)\cdot\ell(g(u_{-i}), i).$$

Pour qu'une procédure d'évaluation hors-ligne soit exploitable, il est nécessaire que l'ordre de pertinence obtenu en confrontant $n$ algorithmes par cette procédure soit identique à celui obtenu en évaluant en ligne ces mêmes algorithmes. On dira dans ce cas que la procédure est *cohérente*.

**Définition 1** (Procédure cohérente)**.** *Soit $c$ un critère métier (taux de clics, taux d'acceptations, etc.) utilisé pour une évaluation online, $\mathcal{P}$ une procédure d'évaluation hors-ligne et $s_{\mathcal{P}}$ le score hors-ligne associé à cette procédure. Soit $s_c$ le score online associé au critère $c$.*

*La procédure $\mathcal{P}$ est dite cohérente par rapport à $c$ si, et seulement si, pour tous algorithmes de recommandation $g_1$ et $g_2$ on a*

$$s_{\mathcal{P}}(g_1) > s_{\mathcal{P}}(g_2) \iff s_c(g_1) > s_c(g_2).$$

Afin d'étudier la cohérence d'une procédure d'évaluation hors-ligne, nous proposons dans un premier temps de l'étudier sur la classe des algorithmes constants, définis de la façon suivante :

**Définition 2** (Algorithme constant)**.** *Soit $g$ un algorithme de recommandation, et $u$ un utilisateur. On dit que $g$ est un algorithme constant si les recommandations faites à l'utilisateur $u$ sont indépendantes de cet utilisateur et des informations connues sur ce dernier.*

*Ainsi, pour tous utilisateurs $u_1$ et $u_2$, on a $g(u_1) = g(u_2)$.*

### 1.5.2 Choix des lois de tirage

Les lois de tirage des utilisateurs et items permettent de définir une procédure d'évaluation et ont une influence directe sur le score des algorithmes obtenu par la procédure d'évaluation offline. Ces lois sont en pratique choisies à partir de considérations métier. Par exemple, s'il est nécessaire d'associer le même poids à tous les utilisateurs, il est préférable d'isoler les utilisateurs selon une loi uniforme. Si au contraire il apparaît pertinent d'associer d'autant plus de poids à un utilisateur qu'il possède d'items (par exemple afin de privilégier les utilisateurs les plus actifs), il est plus adapté de choisir une loi permettant de tirer avec plus forte probabilité les utilisateurs avec un nombre d'items élevé.

Il en est de même pour $P(I|U)$ : si tous les items sont d'importance égale, alors le choix le plus naturel est une loi uniforme sur $\mathcal{I}_U$, les items possédés par l'utilisateur $U$. Dans un cadre dynamique, par exemple pour la recommandation de films ou d'articles de presse, où l'aspect temporel a une grande importance, il pourra être préférable

d'accorder plus d'importance aux items les plus récents. Dans le cadre le plus classique (qui est aussi celui que nous utiliserons pour les expériences), en notant $j$ le nombre d'items isolés pour chaque utilisateur, le choix des paramètres est le suivant :

— $j = 1$ : tirage d'un seul item,
— $U \sim unif(\mathcal{U})$ loi uniforme sur l'ensemble des utilisateurs : tous les utilisateurs ont la même probabilité d'être isolés,
— $I|U \sim unif(\mathcal{I}_U)$ loi uniforme sur $\mathcal{I}_U$ : ayant choisi un utilisateur, tous ses items ont la même probabilité d'être isolés.

Intuitivement, plus un item est recommandé dans le passé, plus le nombre d'utilisateurs associés à cet item est élevé et plus sa probabilité d'être isolé lors de la procédure d'évaluation hors-ligne augmente. Une stratégie naturelle pour diminuer ce biais consiste à isoler l'item $i$ avec une probabilité plus faible s'il a été recommandé dans le passé.

Par ailleurs, en s'intéressant au cas particulier des algorithmes constants, on peut montrer que le score obtenu par évaluation hors-ligne dépend uniquement des probabilités de tirage des items. Il est alors évident qu'une procédure cohérente à un instant $t_0$ sur l'ensemble des algorithmes constants sera également cohérente à un instant $t_1$ si les probabilités de tirage des items sont constantes au cours du temps. Ce résultat fait l'objet de la proposition 2.

**Proposition 2.** *Soit $\mathcal{P}_{t_0}$ une procédure cohérente à l'instant $t_0$, et $t_1$ un instant ultérieur à $t_0$. Alors, si les probabilités de tirage des items sont identiques en $t_0$ et $t_1$ la procédure $\mathcal{P}_{t_0}$ est cohérente sur l'ensemble des algorithmes constants à l'instant $t_1$.*

*Démonstration.* Soit $g$ un algorithme de recommandation constant, $t_0$ et $t_1$ des instants tels que $t_0 < t_1$, et $\mathcal{P}$ une procédure d'évaluation hors-ligne cohérente à l'instant $t_0$.

Le score de l'algorithme $g$ obtenu par la procédure $\mathcal{P}$ à l'instant $t_1$ est obtenu par

$$\begin{aligned} s_{t_1}(g) &= -\sum_{i,u \in \mathcal{I}\times\mathcal{U}} P(i,u)_{t_1} \ell(g(u_{-i}), i), \\ &= -\sum_{i\in\mathcal{I}} \sum_{u\in\mathcal{U}} P(u)_{t_1} P(i|u)_{t_1} \ell(g(.), i), \\ &= -\sum_{i\in\mathcal{I}} \ell(g(.), i) \sum_{u\in\mathcal{U}} P(u) P(i|u)_{t_1}, \\ &= -\sum_{i\in\mathcal{I}} P(i)_{t_1} \ell(g(.), i). \end{aligned}$$

En supposant les probabilités de tirage conservées entre les instants $t_0$ et $t_1$, on a $p(i)_{t_1} = p(i)_{t_0}$, et donc $\ell(g(.), i)_{t_1} = \ell(g(.), i)_{t_0}$.

Alors

$$s_{t_1}(g) = -\sum_i p(i)_{t_0} \ell(g(.), i)_{t_0} = s_{t_0}(g)$$

et la procédure $\mathcal{P}$ est donc cohérente à l'instant $t_1$ pour tout algorithme constant. □

Nous utiliserons une procédure qui permet de contrôler les probabilités de tirage des items, tout en conservant le tirage des utilisateurs selon une loi uniforme. Nous pourrons ainsi prendre en compte une évolution des données et l'apparition de nouveaux couples $(u, i)$ indépendants d'une campagne de recommandation. Ce résultat fait l'objet de la proposition 2.

**Remarque** La preuve de la proposition 2 montre que le meilleur algorithme constant est celui qui maximise $-\sum_{i\in\mathcal{I}} P(i)\ell(g(.), i)$. En conséquence, l'algorithme constant optimal parmi ceux proposant $k$ items est celui pour lequel les valeurs de $P(i)\ell(g(.), i)$ sont les $k$ plus élevées. Dans le cas où le score d'évaluation est binaire, et vaut 1 si l'item isolé apparaît dans les recommandations obtenues pour l'utilisateur $u$, 0 sinon, l'algorithme constant optimal est donc celui proposant l'item avec la plus forte probabilité d'être isolé.

### 1.5.3 Pondération des items

La proposition 2 montre que si une procédure est cohérente à un instant $t_0$, il est suffisant que les probabilités de tirage des items soient conservées pour que la procédure reste cohérente sur l'ensemble des algorithmes constants à un instant $t > t_0$. Pour déterminer une procédure cohérente, une approche naturelle consiste donc à rechercher une procédure qui conserve les lois de tirage de items.

Pour maîtriser les probabilités de tirage des items, une stratégie consiste à inverser la procédure d'évaluation, de façon à isoler d'abord un item $I$, puis à isoler un utilisateur parmi $\mathcal{U}_I$, l'ensemble des utilisateurs associés à l'item $I$. Toutefois, cette approche n'est pas satisfaisante en pratique car elle ne permet pas de maîtriser la loi de tirage des utilisateurs, qui sont généralement isolés selon une probabilité uniforme pour des raisons métier. Il est donc nécessaire de conserver la procédure d'évaluation présentée à la section 1.3.2. En utilisant cette procédure, l'unique façon d'agir sur les probabilités de tirage des items sans modifier la loi de tirage des utilisateurs réside dans le choix de la variable aléatoire $I|U$.

Pour tout instant $t \in \{t_0, t_1\}$, l'objectif est de conserver les probabilités de tirage des items définies pour tout $i \in \mathcal{I}$ par

$$P(i|t) = \sum_{u\in\mathcal{U}} P(I = i|U = u, t) \cdot P(U = u).$$

Alors, les conditions $P(i|t_0) = P(i|t_1)$ fournissent un système à $|\mathcal{I}|$ équations, dont les inconnues sont $P(I = i|U = u, t_1)$ pour tout $i \in \mathcal{I}$ et $u \in \mathcal{U}$. En pratique, il est très fréquent d'observer $|\mathcal{U}| > |\mathcal{I}|$, et le problème possède donc plus d'inconnues que d'équations. Afin de contrôler les probabilité de tirage des items et réduire le nombre de

paramètres à identifier, nous proposons d'introduire, pour chaque item $i$, un poids $\omega_i$ de la façon suivante :

$$P(i|u,\omega)_t = \frac{\omega_i P(i|u)_t}{\sum_{j\in I_u} \omega_j P(j|u)_t}$$

avec la contrainte $\sum_i \omega_i = 1$. La probabilité de sélection des items dépend alors de $\omega$ par la relation

$$P(i|\omega) = \sum_{u\in U} P(i|u,\omega)P(u).$$

La section suivante propose une méthode pour déterminer la valeur des poids permettant de diminuer le biais d'évaluation.

### 1.5.4 Détermination des poids

Nous avons choisi d'approcher les coefficients $\omega_i(t)$ à l'aide d'un algorithme d'optimisation, de façon à minimiser la distance entre les distributions de tirage des items aux instants $t_0$ et $t_1$.Comme il s'agit de distributions de probabilités, nous avons choisi d'utiliser la divergence de Kullback-Leibler définie par

$$\begin{aligned} D(\omega) &= D_{KL}(P(.)_{t_0}||P(.|\omega)_{t_1}) \\ &= \sum_{i\in I_{t_0}} P(i)_{t_0} \log \frac{P(i)_{t_0}}{P(i|\omega)_{t_1}}. \end{aligned}$$

La dissymétrie de la divergence de Kullback-Leibler est particulièrement pertinente ici dans la mesure où elle réduit l'importance accordée aux items rares à l'instant $t_0$. Par ailleurs, cette fonction de divergence a l'avantage d'accorder une importance nulle aux nouveaux items (c'est-à-dire tels que $P(i)_{t_1} > 0$ mais $P(i)_{t_0} > 0$), ce qui permet bien d'optimiser uniquement les poids des items présents à l'instant $t_0$.

#### Cadre général

Nous proposons de minimiser la distance de Kullback-Leibler entre les distributions des probabilités de tirage aux instants $t_0$ et $t_1$ par un algorithme classique de descente de gradient, décrit en pseudo code à l'algorithme 2.

La proposition 3 donne une expression du gradient de la distance de Kullback-Leibler.

**Proposition 3.** *Soit $\mathcal{P}$ une procédure d'évaluation hors-ligne, $t_0$ et $t_1$ deux instants tels que $t_0 > t_1$, et $D(\omega)$ la distance de Kullback-Leibler entre les lois de tirage des items aux instants $t_0$ et $t_1$, en fonction du vecteur des poids $\omega$, définie par*

$$D(\omega) = \sum_{i\in I_{t_0}} P(i)_{t_0} \log \frac{P(i)_{t_0}}{P(i|\omega)_{t_1}}.$$

**Algorithme 2 :** Optimisation par descente de gradient

```
f, x0, x1, ε, N_max, p
n = 1
while max(|(x0 − x1)/x0|) > ε  &  n < N_max do
  x0 = x1
  x1 = x0 − p/(n+1) × ∇f(x1)
  n = n + 1
end while
return x0
```

*Alors, le gradient de la distance en fonction de $\omega$ est donné par*

$$\nabla_k D(\omega) = \frac{\partial D(\omega)}{\partial \omega_k} = \sum_{i \in I_{t_0}} \frac{P(i)_{t_0}}{\omega_k P(i|\omega)_{t_1}} \times \big(P(i,k|\omega)_{t_1} - \delta_{ik} P(k|\omega)\big),$$

*où* $P(i,k|\omega)_{t_1} = \sum_u P(i|u,\omega)_{t_1} P(k|u,\omega)_{t_1} P(u)$.

*Démonstration.* Pour tout $k$

$$\frac{\partial D(\omega)}{\partial \omega_k} = -\sum_{i \in I_{t_0}} P(i)_{t_0} \frac{1}{P(i|\omega)_{t_1}} \frac{\partial P(i|\omega)_{t_1}}{\partial \omega_k}.$$

Or $P(i|\omega)_{t_1} = \sum_u P(u) \times P(i|u,\omega)_{t_1}$, donc

$$\frac{\partial P(i|\omega)_{t_1}}{\partial \omega_k} = \sum_u P(u) \frac{\partial P(i|u,\omega)_{t_1}}{\partial \omega_k}.$$

Et nous avons vu que $P(i|u,\omega) = \frac{\omega_i P(i|u)}{\sum_{j \in I} \omega_j P(j|u)}$. Une distinction de cas est donc nécessaire pour exprimer $\frac{\partial P(i|u,\omega)}{\partial \omega_k}$ :

— Si $i \neq k$, alors

$$\frac{\partial P(i|u,\omega)}{\partial \omega_k} = -\frac{\omega_i P(i|u) P(k|u)}{\Big(\sum_{j \in I} \omega_j P(j|u)\Big)^2} = -\frac{P(i|u,\omega) P(k|u,\omega)}{\omega_k}.$$

— Et si $i = k$, alors

$$\begin{aligned}
\frac{\partial P(i|u,\omega)}{\partial \omega_i} &= \frac{P(i|u) \cdot \sum_j \omega_j P(j|u) - \omega_i P(i|u) \cdot P(i|u)}{\Big(\sum_{j \in I} \omega_j P(j|u)\Big)^2}, \\
&= \frac{P(i|u)}{\sum_j \omega_j P(j|u)} - \frac{\omega_i P(i|u)}{\sum_j \omega_j P(j|u)} \cdot \frac{P(i|u)}{\sum_{j \in I} \omega_j P(j|u)}, \\
&= \frac{P(i|u,\omega)}{\omega_i} - P(i|u,\omega) \cdot \frac{P(i|u,\omega)}{\omega_i},
\end{aligned}$$

$$= \frac{P(i|u,\omega)}{\omega_i}(1 - P(i|u,\omega)).$$

Finalement, nous avons pour tout $k$

$$\frac{\partial P(i|u,\omega)_{t_1}}{\partial \omega_i} = \frac{P(k|u,\omega)_{t_1}}{\omega_k}\left(\delta_{ik} - P(i|u,\omega)_{t_1}\right).$$

Par conséquent,

$$\frac{\partial P(i|\omega)_{t_1}}{\partial \omega_k} = \sum_u P(u)\frac{P(k|u,\omega)_{t_1}}{\omega_k}\left(\delta_{ik} - P(i|u,\omega)_{t_1}\right).$$

En notant $P(i,k|\omega)_{t_1} = \sum_u P(i|u,\omega)_{t_1}P(k|u,\omega)_{t_1}P(u)$ on obtient

$$\frac{\partial P(i|\omega)_{t_1}}{\partial \omega_k} = \frac{\delta_{ik}P(k|\omega)_{t_1} - P(i,k|\omega)_{t_1}}{\omega_k},$$

d'où

$$\frac{\partial D(\omega)}{\partial \omega_k} = \sum_{i \in I_{t_0}} \frac{P(i)_{t_0}}{\omega_k P(i|\omega)_{t_1}} \times \big(P(i,k|\omega) - \delta_{ik}P(k|\omega)\big).$$

□

**Application**

Dans le cas le plus simple et le plus fréquent, les lois de tirage des utilisateurs et des items suivent une loi uniforme. Dans ce cas, on a alors :

$$\begin{aligned}
P(U=u) &= \frac{1}{|\mathcal{U}|} \\
P(I=i|U=u) &= \frac{1}{|\mathcal{I}_u|}\cdot \mathbb{1}_{i\in\mathcal{I}_u} \\
P(I=i|U=u,\omega) &= \frac{\omega_i}{\sum_{j\in\mathcal{I}_u}\omega_j}\mathbb{1}_{i\in\mathcal{I}_u} \\
P(I=i|\omega) &= \sum_{u\in\mathcal{U}} P(I=i|U=u,\omega)P(u) \\
&= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}_i} P(i|u,\omega) \\
&= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}_i}\frac{\omega_i}{\sum_{i\in\mathcal{I}_u}\omega_j} \\
P(i,k|\omega) &= \sum_{u\in\mathcal{U}} P(i|u,\omega)P(k|u,\omega)P(u)
\end{aligned}$$

$$= \frac{1}{|\mathcal{U}|} \cdot \sum_{u \in \mathcal{U}_i \cap \mathcal{U}_k} \frac{\omega_i \omega_k}{(\sum_{j \in \mathcal{I}_u} \omega_j)^2}$$

En pratique, dans le cadre de données en grande dimension (nombre élevé d'items), l'optimisation de tous les paramètres est trop coûteuse en temps de calcul pour être réalisée de façon exhaustive. Nous proposons alors d'optimiser les $p$ paramètres ayant le plus dévié dans le temps, où la déviation entre deux instants $t_0$ et $t_1$ est donnée par :

$$\delta_i = |P_i(\omega(t_0)) - P_i(\omega(t_1))| .$$

Dans la section 1.6, nous effectuerons plusieurs expériences en fonction de $p$ pour mesurer l'influence de la valeur de $p$ sur les résultats. Par ailleurs, comme l'ensemble des $\omega_i$ est lié par la relation $\sum_j \omega_j = 1$, l'optimisation d'un paramètre fait varier légèrement les autres. Nous préférerons donc commencer par le $p$-ième paramètre ayant le plus dévié puis itérer jusqu'au paramètre associé à l'item ayant connu la plus grande déviation. Ainsi, nous terminerons l'optimisation des paramètres par le poids associé au produit ayant le plus dévié, ce qui permet de s'assurer que celui-ci ne sera pas modifié par des itérations ultérieures.

### 1.5.5 Étude de la complexité

La proposition 4 donne une expression de la complexité de l'approche proposée.

**Proposition 4.** *Soit $\mathcal{P}$ une procédure d'évaluation hors-ligne, $t_0$ et $t_1$ deux instants tels que $t_0 < t_1$, et $D(\omega)$ la distance de Kullback-Leibler entre les lois de tirage des items aux instants $t_0$ et $t_1$.*

*Le calcul de $p$ coordonnées du gradient de $D$ peut être effectué avec une complexité en $\mathcal{O}(p \times n_{U\times I})$, où $n_{U\times I}$ est le nombre de couples $(u, i)$ avec $u \in U$ et $i \in I_u$ ($n_{U\times I} = \sum_{i\in I} \#U_i$).*

*Démonstration.* Soit $A_u$ (resp. $A_i$) une matrice creuse dans $\mathcal{M}_{n_U,n_I}(\mathbb{R})$, indexée par les colonnes (resp. lignes), telle que $(A_u)_{u,i} = \mathbb{1}_{i\in I_u}$ (resp. $(A_i)_{u,i} = \mathbb{1}_{i\in I_u}$).

De telles matrices peuvent être obtenues en $\mathcal{O}(n_{U\times I})$ et permettent d'accéder à tous les éléments en $\mathcal{O}(\eta_u)$ (resp. $\mathcal{O}(\eta_u)$), où $\mathcal{O}(\eta_u)$ (resp. $\mathcal{O}(\eta_i)$) est le nombre maximum d'éléments non nuls sur une ligne de la matrice $A_u$ (resp. $A_i$).

Alors, comme dans le cas particulier des lois uniformes on a

$$P(i|\omega) = \sum_{u\in U_i} P(i|u,\omega)P(u) = \sum_{u\in U_i} \frac{\omega_i}{\sum_{j\in I_u} \omega_j} P(u)$$

et

$$P(i,k|\omega) = \sum_{u\in U_i} P(i|u,\omega)P(k|u,\omega)P(u) = \sum_{u\in U_i} \frac{\omega_i\omega_k}{(\sum_{j\in I_u} \omega_j)^2} P(u),$$

il est possible de calculer $P(i|\omega)$ pour tout $i \in I$ en $\mathcal{O}(n_{U\times I})$. Et en ayant précalculé $\sum_{u\in U_i} \frac{\omega_i}{(\sum_{j\in I_u}\omega_j)^2} P(u)$ pour tout $i$ (coût en $\mathcal{O}(n_{U\times I})$), on a ensuite $P(i,k|\omega)$ en $\mathcal{O}(\#U_i)$ pour tout $k$.

Alors, après avoir précalculé $P_{t_0}(i)$ et $P_{t_1}(i)$ pour tout $i$, la quantité $\frac{\partial D(\omega)}{\partial \omega_k}$ est une somme de $\#I$ termes calculés en $\mathcal{O}(\#U_i)$ et chaque coordonnée du gradient s'obtient donc en une complexité de $\mathcal{O}(\sum_{i\in I} U_i)$, ce qui correspond bien à $|A_u|$ par définition de $A_u$. □

## 1.6 Applications

### 1.6.1 Description des données

Pour illustrer les résultats de l'approche proposée dans la section 1.5, nous avons réalisé des expériences à partir de données issues du réseau social professionnel Viadeo France.

Le jeu de données utilisé pour ces expériences est constitué de 34 448 utilisateurs, 35 741 compétences distinctes (correspondant aux items), avec en moyenne 5.33 compétences par utilisateur. La distribution du nombre d'items par membre est représentée à la figure 1.5. Les échelles étant logarithmiques, on reconnaît ici une distribution des degrés en loi puissance, caractéristique des données issues des réseaux sociaux.

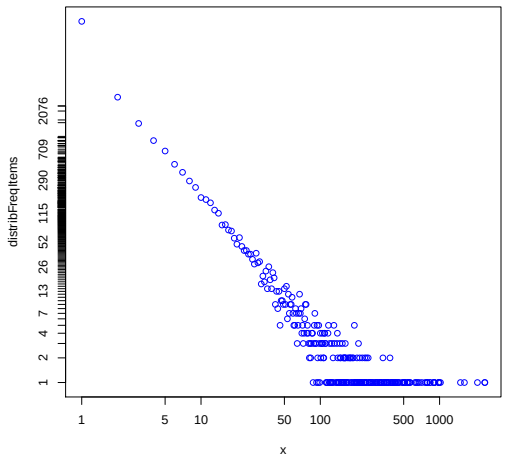



FIGURE 1.5 – Distribution du nombre de compétences par membre (échelles logarithmiques)

Connaissant les dates des campagnes de recommandation et les compétences recommandées dans le passé, nous avons mesuré par évaluation hors-ligne le score de

quelques algorithmes de recommandation à plusieurs instants : avant, pendant et après la campagne de recommandation.

**Protocole d'évaluation** Les évaluations ont été réalisées à partir de 20 000 simulations selon l'algorithme 1.

### 1.6.2 Mise en évidence du biais

Au cours d'une campagne de recommandation, l'impact sur les fréquences des compétences (et donc leur probabilité de tirage) est très rapide, notamment pour celles recommandées auprès du plus grand nombre d'utilisateurs. La figure 1.6 illustre ce phénomène sur un intervalle de temps compris entre $t = 300$ jours et $t = 500$ jours (l'instant $t = 0$ correspond au jour du lancement de la fonctionnalité permettant aux utilisateurs d'ajouter des compétences à leur profil).

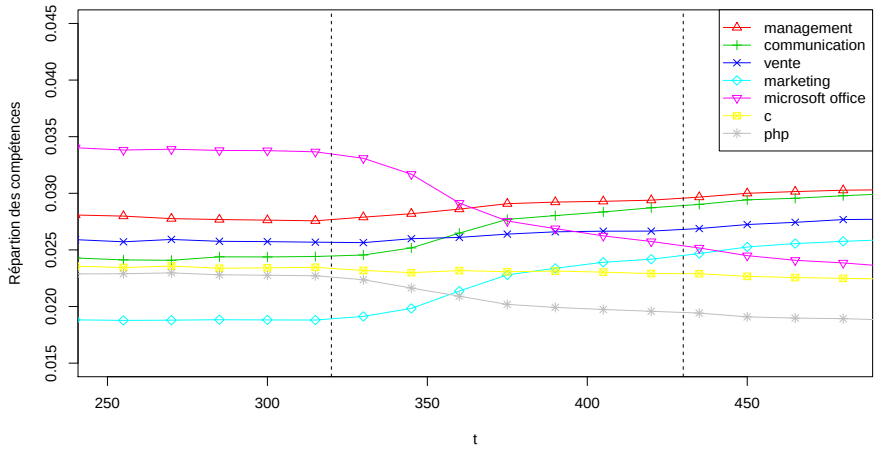



FIGURE 1.6 – Impact d'une campagne de recommandation sur les probabilités de tirage des items

A l'instant $t = 300$, de nombreux utilisateurs ont déjà renseigné leurs compétences et les fréquences des items sont stables. Deux campagnes de recommandation peuvent être mises en évidence sur cette figure, aux instants $t = 320$ et $t = 430$ respectivement. Les probabilités de sélection des items changent alors brusquement, en faveur des items qui ont été recommandés. L'allure de la courbe associée à la compétence marketing indique par exemple que cette compétence a été largement recommandée durant les campagnes de recommandation, contrairement à la compétence vente pour laquelle la courbe est décroissante.

Une procédure d'évaluation hors-ligne classique en $t = 400$ isolerait alors plus souvent la compétence marketing que vente tandis que l'inverse serait observé en $t = 300$. En conséquence, la sélection des items sera biaisée en faveur de ceux déjà recommandés lors d'une procédure d'évaluation hors-ligne à l'instant $t > 320$, ce qui conduira à la surestimation de la qualité des algorithmes proposant des items déjà suggérés. Ce phénomène est illustré aux figures 1.7(a) et 1.7(b), qui représentent respectivement les scores obtenus par un algorithme proche de celui utilisé pour les recommandations, et un algorithme orthogonal, c'est-à-dire proposant des items n'ayant jamais été recommandés précédemment.

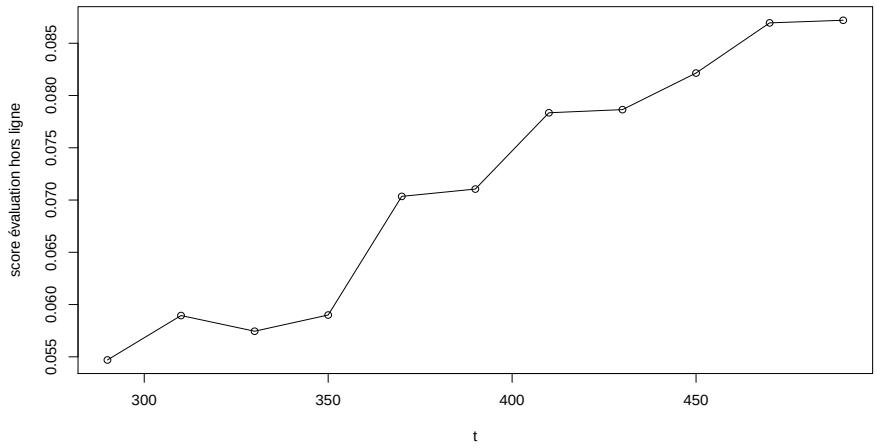
— Figure 1.7(a) : l'algorithme utilisé est constant et consiste à proposer à chaque utilisateur les cinq items les plus souvent recommandés sur la période $t \in [320; 480]$. Cet algorithme permet d'illustrer le biais de l'évaluation hors-ligne d'un algorithme proche de celui utilisé pour les précédentes campagnes de recommandation. Les items recommandés sont de plus en plus souvent isolés lors de la procédure d'évaluation offline, et le score de cet algorithme augmente en conséquence au cours du temps. Ce résultat illustre bien la surestimation de la qualité d'un algorithme allant dans le sens de celui existant.

— Figure 1.7(b) : ce graphique illustre l'évolution du score obtenu par évaluation hors-ligne pour un algorithme orthogonal à celui utilisé pour les recommandations. L'algorithme consiste ici à proposer à chaque utilisateur, parmi les items qui n'ont jamais été recommandés entre les instants t = 300 et t = 480, les cinq items qui étaient les plus fréquents à l'instant t = 300. Il s'agit donc d'un algorithme très différent de ceux utilisés durant les précédentes campagnes de recommandations. Cette figure illustre bien le fait que la procédure d'évaluation hors-ligne sous-estime la qualité d'un algorithme différent de celui existant.

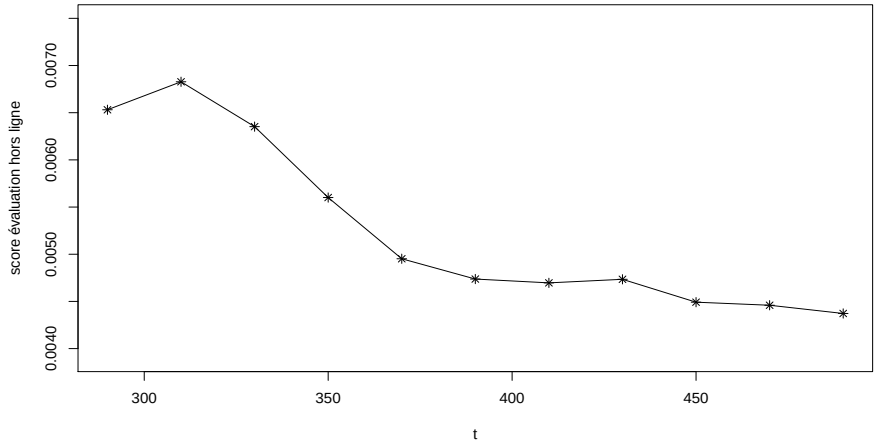
Ces figures témoignent du biais lié à une campagne de recommandation sur les scores obtenus par la procédure d'évaluation hors-ligne classique. La section suivante détaille les résultats obtenus par l'approche proposée dans la section 1.5 pour diminuer le biais de l'évaluation hors-ligne.

### 1.6.3 Correction du biais

Afin d'évaluer la performance de l'approche proposée, nous l'avons testée sur deux types d'algorithmes. Dans un premier temps, nous détaillerons les résultats obtenus sur les algorithmes constants précédemment évoqués. Dans un second temps, nous présenterons le cas plus complexe des algorithmes de filtrage collaboratif (Sarwar *et al.*, 2001; Su and Khoshgoftaar, 2009).




(a) Algorithme similaire




(b) Algorithme orthogonal

FIGURE 1.7 – Evolution des scores dans le temps pour deux algorithmes constants.

### Algorithmes constants

Les figures 1.8(a) et 1.8(b) illustrent les résultats de la procédure hors-ligne proposée dans le cas des deux algorithmes constants présentés à la section 1.6.2. Les expériences montrent que les résultats sont très sensibles à la valeur de $p$, le nombre de poids optimisés :

— dans le cas de l'algorithme similaire à celui utilisé pour les recommandations passées, $p = 20$ suffit à corriger le biais introduit par la campagne et à retrouver

la constance du score. Cela s'explique par le fait que les items recommandés sont ceux dont la proportion a le plus dévié ;

— la correction du biais est beaucoup plus lente dans le cas de l'algorithme orthogonal : aucune amélioration n'est perceptible pour $p \leq 20$. On peut l'expliquer par le fait que les items recommandés par cet algorithme n'ont pas été recommandés dans le passé, leur proportion n'a donc que très peu dévié. En conséquence, il faudra une valeur de $p$ élevée pour que ces items apparaissent parmi les $p$ items recalibrés.

L'analyse de ces résultats montre que la valeur minimale de $p$ dépend de l'algorithme utilisé. Pour comparer la pertinence de deux algorithmes, il est donc nécessaire d'optimiser les paramètres sur un grand nombre d'items, même si cela peut être très long en pratique.

### Filtrage collaboratif

Les méthodes de filtrage collaboratif (Su and Khoshgoftaar, 2009; Sarwar *et al.*, 2001) constituent une classe d'algorithmes de recommandation classiques, qui sont très couramment utilisés en pratique en raison de leur simplicité d'implémentation et de leur efficacité. Pour un utilisateur $u$ donné, le principe des algorithmes de filtrage collaboratif peut être résumé en deux étapes :

1. Utiliser les données comportementales disponibles sur l'ensemble des utilisateurs pour en extraire un groupe d'utilisateurs similaires à l'utilisateur $u$.
2. Recommander à l'utilisateur $u$ des produits populaires au sein de ce groupe d'utilisateurs similaires.

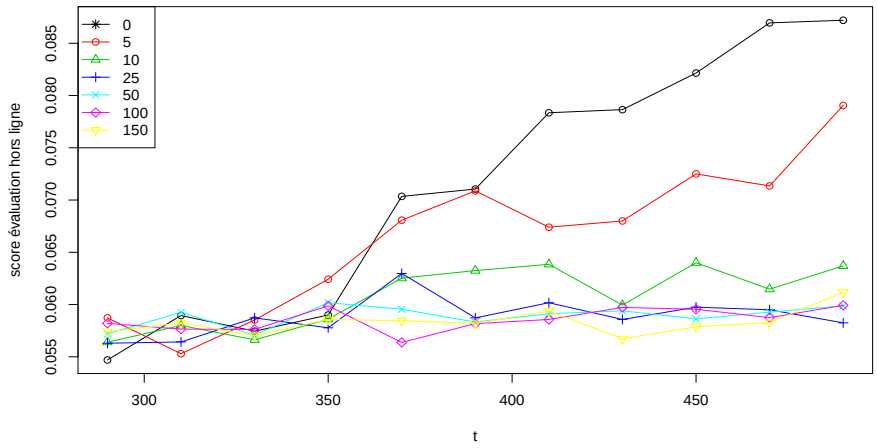
Notons $I_u(t)$ le vecteur des items associés à un utilisateur $u$ à l'instant $t$, *i.e.* $I_u(t)[i] = 1$ si l'item $i$ est associé à l'utilisateur $u$ à l'instant $t$, et 0 sinon.

Les algorithmes de filtrage collaboratif (Su and Khoshgoftaar, 2009; Sarwar *et al.*, 2001) consistent à estimer le vecteur des items associés à un utilisateur $u$ à un instant $t' > t$ par une relation du type

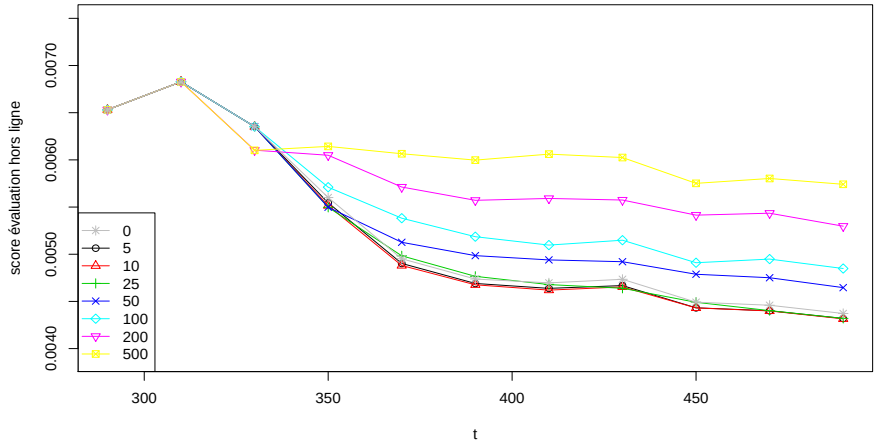
$$\mathcal{I}_u(t') = \sum_{v \neq u} sim(u, v) \cdot \mathcal{I}_v(t)$$

où $sim(u, v)$ représente la similarité entre deux utilisateurs $u$ et $v$. Le choix de la fonction de similarité définit alors le type de filtrage collaboratif utilisé. Un algorithme intuitif, pour lequel la similarité correspond au nombre d'items en commun entre les utilisateurs $u$ et $v$, se traduit par la fonction de similarité $sim(u, v) = \langle I_u(t), I_v(t) \rangle$, où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire entre deux vecteurs.

Bien que cette fonction de similarité soit très intuitive, elle n'est que rarement utilisée en pratique car elle associe trop d'importance aux utilisateurs possédant un grand nombre d'items. Il est en général plus pertinent de pondérer la fonction de




(a) Algorithme similaire




(b) Algorithme orthogonal

FIGURE 1.8 – Représentation de l'évolution des scores en fonction de $p$ (le nombre de paramètres ajustés) sur deux algorithmes constants.

similarité par le nombre d'items associés à chaque utilisateur, ce qui conduit à la similarité cosinus, en anglais *cosine similarity*, donnée par (voir Su and Khoshgoftaar (2009) par exemple)

$$sim(u,v) = \frac{\langle \mathcal{I}_u(t), \mathcal{I}_v(t) \rangle}{\sqrt{|\mathcal{I}_u(t)| \cdot |\mathcal{I}_v(t)|}}.$$

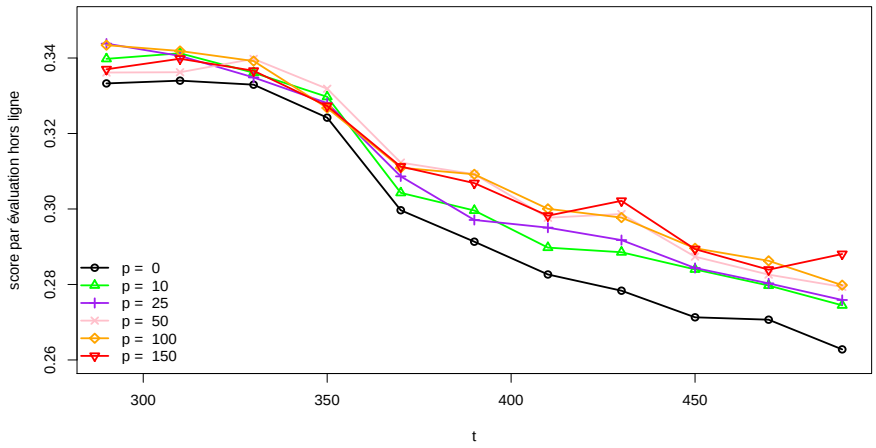



FIGURE 1.9 – Représentation de l'évolution des scores en fonction du nombre $p$ de paramètres ajustés, pour un algorithme de filtrage collaboratif.

Les résultats présentés à la figure 1.9 montrent que la stabilisation des scores s'améliore avec le nombre d'items dont les poids sont optimisés. Comme attendu, la méthode proposée permet donc de diminuer le biais lié aux précédentes recommandations lors de l'évaluation hors-ligne d'un nouvel algorithme.

Cependant, il est difficile de tirer des conclusions très précises de ces résultats, dans la mesure où il n'y a pas de raison de supposer la stabilité des scores dans le cas d'un filtrage collaboratif. En effet, les recommandations proposées par le ou les algorithmes précédemment en production ont biaisé la fréquence des items présents en base, mais également la structure du graphe bipartite *utilisateurs-items*. Or un algorithme de filtrage collaboratif repose principalement sur la structure de ce graphe, et l'optimisation des poids ne permet que de diminuer le biais lié à la modification des fréquences des items. Le cas du biais introduit dans la structure du réseau constitue alors un nouvel axe de travail spécifique, que nous n'aborderons pas dans le cadre de cette thèse.

## 1.7 Conclusion

Nous avons vu que de nombreux facteurs (campagnes marketing, campagnes de recommandation, changement de design du site, ...) influencent les comportements des utilisateurs sur internet. Comme l'évaluation hors-ligne d'un algorithme de recommandation repose sur les comportements des utilisateurs observés sur une période donnée, les résultats obtenus sont également biaisés par des facteurs indépendants des utilisateurs, par exemple les algorithmes précédemment mis en production. Dans le cadre d'une

campagne de recommandation, nous avons par exemple vu qu'une évaluation classique tend à sur-estimer un algorithme proche de celui existant et à pénaliser un algorithme orthogonal.

A l'aide de la théorie du *covariate shift*, nous avons proposé une approche permettant de diminuer le biais introduit par une campagne de recommandation, en mimant une situation observée précédemment et pour laquelle le biais peut être considéré comme négligeable. Les expériences montrent que cette nouvelle procédure d'évaluation hors-ligne permet de réduire le biais lié à des effets extérieurs sur les scores obtenus et d'avoir une évaluation plus représentative de la réalité.

Cette approche, inspirée du *covariate shift*, suppose que les compétences des utilisateurs sont constantes dans le temps. Cette hypothèse, acceptable sur un court intervalle de temps, permet de s'assurer que la probabilité d'observer une compétence, connaissant le profil d'un utilisateur, est constante entre deux instants. Dans le chapitre suivant, nous discutons cette hypothèse et proposons une nouvelle approche dans le cas où la constance de la relation explicative entre les observations et la variable cible n'est pas respectée.

## Contributions scientifiques

Ce chapitre a fait l'objet des publications et des communications suivantes :

1. "Reducing offline evaluation bias of collaborative filtering algorithms.", Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi, Apr 2015, Bruges, Belgium. Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015)
2. "Study of a bias in the offline evaluation of a recommendation algorithm.", Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi. 11th Industrial Conference on Data Mining, ICDM 2015, Jul 2015, Hamburg, Germany. Ibai Publishing, pp.57-70, 2015, Advances in Data Mining.
3. "Reducing offline evaluation bias in recommendation systems", A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, 23rd annual Belgium-Dutch Conference on Machine Learning (Benelearn 2014), Bruxelles, juin 2014
4. "Reducing offline evaluation bias in recommendation systems", A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, séminaire du SAMM, université Paris 1 Panthéon Sorbonne, décembre 2014

# Chapitre 2

# Explanatory Shift

## 2.1 Introduction

### 2.1.1 Contexte

Dans le chapitre précédent, nous avons vu que les bases d'apprentissage et de test peuvent présenter des propriétés différentes. On dit alors qu'on observe un biais, aussi appelé shift, qui peut s'expliquer par : le processus de génération ou de collecte des données, la modification ou la mise en place dans le passé d'une nouvelle fonctionnalité ayant une influence significative sur l'expérience utilisateur, comme une campagne de recommandation par exemple. Dans le cas de données observées sur un intervalle de temps donné, les effets de saisonnalité peuvent également être à l'origine d'un biais. Ce biais, ou shift, peut conduire à une base d'apprentissage non-représentative de la base de test, ce qui aura un impact sur la performance des modèles prédictifs calibrés à partir de cette base d'apprentissage biaisée. Nous avons alors vu que la théorie du *covariate shift* permet d'ajuster la base d'apprentissage et d'obtenir un modèle pouvant être appliqué à des fins prédictives avec une bonne performance.

Toutefois, l'application du *covariate shift* nécessite des hypothèses fortes sur les relations entre les bases d'apprentissage et de test. En particulier, nous avons vu que la stratégie développée par Shimodaira (2000) est pertinente dans le cas où la distribution des données a été modifiée, mais nécessite la conservation de la relation explicative de la variable cible Y sachant les observations X entre les deux ensembles de données. Dans ce cas, Shimodaira montre qu'il est préférable de pondérer les observations de la base d'apprentissage de façon à mimer la base de test, et ainsi avoir un modèle plus performant sur la base de test. En reprenant les notations introduites à la section 1.3.1, le modèle optimal est alors choisi de façon à minimiser le risque empirique pondéré, donné par

$$g^\star = \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \omega_i(x_i)\ell(y_i, g(x_i)) \quad \text{avec} \quad \omega_i(x) = \frac{p_{test}(x)}{p_{train}(x)}.$$

Cependant, bien que cette approche soit pertinente sous les hypothèses mentionnées, elle n'est pas toujours exploitable. En pratique il est en effet courant d'observer une modification du lien entre la variable cible $Y$ et les observations $X$ lorsque la distribution de $X$ est modifiée. Deux cas sont souvent distingués :

1) $P(Y|X)$ est conservé entre les bases d'apprentissage et de test, bien que $P(X)$ évolue. Comme vu précédemment, cette situation s'inscrit dans le cadre du *covariate shift*.
2) $P(Y|X)$ diffère entre les bases d'apprentissage et de test. Ce cas s'inscrit dans le cadre de la théorie de l'adaptation de domaine (Pan and Yang, 2010; Villani, 2008), que nous présenterons à la section 2.2.2.

Dans ce chapitre, nous proposons d'étudier un cas s'inscrivant dans le cadre de l'adaptation de domaine, pour lequel les approches existantes ne fournissent pas de résultats satisfaisants à notre connaissance. Plus précisément, nous traiterons le cas où $P(Y|X)$ et $P(Y)$ évoluent de manière différente entre les bases d'apprentissage et de test, bien que $P(X|Y)$ soit conservé entre ces deux bases. Nous parlerons alors de biais explicatif, ou *explanatory shift*. Pour cela, nous nous placerons dans le cadre où $X$ est une variable aléatoire réelle absolument continue et $Y$ une variable aléatoire discrète.

A titre d'illustration, prenons l'exemple d'une étude menée dans un lycée, entre deux groupes d'élèves, notés A et B, où l'objectif est de modéliser le genre des élèves en fonction de leur taille. En reprenant les notations du chapitre précédent, la variable cible $Y$ est alors le genre (garçon ou fille), et la variable explicative $X$ la taille. En supposant que la distribution de la taille des filles suit une loi normale de moyenne 165 cm et d'écart-type 10 cm, et que la distribution de la taille des garçons suit une loi normale de moyenne 175 cm et d'écart type 10 cm, les distributions obtenues sont celles représentées à la figure 2.1.1.

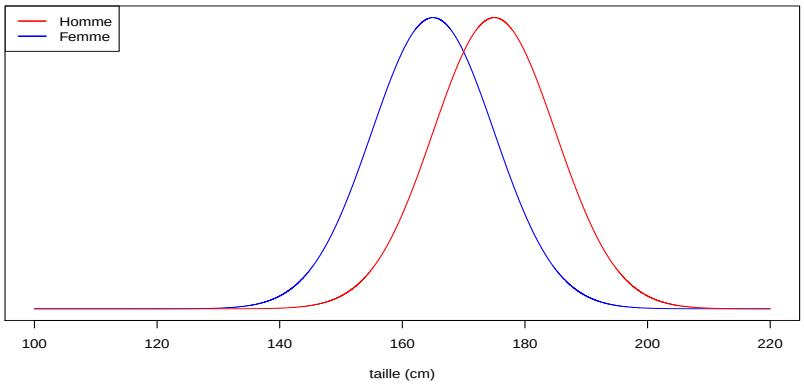



FIGURE 2.1 – Distribution des tailles des hommes et des femmes.

Alors, dans le cas où les deux groupes A et B comportent autant de filles que de garçons, on peut montrer que, pour un individu tiré aléatoirement, le choix optimal pour estimer son genre est de supposer qu'il s'agit d'une femme si la taille de l'individu est inférieure à 170cm, et un homme sinon.

On peut également montrer que dans le cas où la proportion de filles est de 40%, le seuil définissant la règle de décision est de 166cm. A l'inverse, dans le cas où la proportion de filles est de 60%, la règle de décision optimale se situe à 174cm.

Par conséquent, dans l'hypothèse où le groupe A comporte une proportion de filles de $\pi_A = 60\%$, et le groupe B une proportion de filles $\pi_B = 40\%$ (*i.e.*, $P(Y|A) \neq P(Y|B)$), un modèle calibré sur le groupe A ne se généralisera pas bien aux individus du groupe B. Toutefois, au sein de la classe la distribution des tailles conditionnellement au genre est identique : il s'agit d'une loi normale de moyenne 165 (resp. 175) et d'écart-type 10 pour les filles (resp. garçons), *i.e.* $p(X|Y, A) = p(X|Y, B)$. Cet exemple illustre les hypothèses que nous effectuerons dans ce chapitre.

Plus précisément, nous nous intéressons au cas où l'hypothèse fondamentale du *covariate shift* $P(Y_{train}|X_{train}) = P(Y_{test}|X_{test})$ n'est pas respectée, et nous proposons de la remplacer par l'hypothèse de stabilité des variables explicatives conditionnellement à la variable cible :

$$p(X_{train}|Y_{train}) = p(X_{test}|Y_{test}).$$

Pour exploiter cette hypothèse, nous aurons besoin de connaître ou d'estimer la distribution de $Y_{test}$, ce que nous ferons grâce à l'approche itérative présentée à la section 2.3.

Dans ce chapitre, nous détaillerons une approche inspirée de la théorie du *covariate shift* pour avoir un modèle pertinent sous ces hypothèses. Aussi, nous comparerons notre approche à d'autres modélisations évoquées dans la littérature, permettant de traiter le problème de biais entre les bases d'apprentissage et de test sous des hypothèses parfois différentes, résumées à la table 2.1.

| Hypothèse | $P_{Y_s} = P_{Y_t}$ | $P_{X_s} = P_{X_t}$ | $P_{Y_s\|X_s} = P_{Y_t\|X_t}$ | $P_{X_s\|Y_s} = P_{X_t\|Y_t}$ |
|---|---|---|---|---|
| Cadre classique | **oui** | **oui** | **oui** | **oui** |
| *Covariate shift* | non | non | **oui** | non |
| Mélange gaussien | non | non | non | **oui** |
| Adaptation de domaine | **oui** | non | non | non |
| Approche proposée | non | non | non | **oui** |

TABLE 2.1 – Synthèse des hypothèses sous-jacentes à chaque approche.

### 2.1.2 Plan du chapitre

La suite de ce chapitre sera organisée de la façon suivante. Tout d'abord, dans la section 2.2, nous évoquerons l'état de l'art sur le sujet. Nous détaillerons alors les limites du *covariate shift* et des modèles d'adaptation de domaine, et présenterons un exemple simple pour lequel les hypothèses du *covariate shift* ne sont pas respectées. Nous montrerons également que la théorie des mélanges gaussiens peut fournir des résultats pertinents dans ce type de situation. Enfin, nous détaillerons l'approche que nous proposons à la section 2.3.

Dans la section 2.4, nous appliquerons les approches évoquées dans les sections 2.2 et 2.3 (*covariate shift*, adaptation de domaine, mélange gaussien et l'approche proposée) sur des données simulées. Enfin, la section 2.5 sera consacrée à des expériences sur des données réelles, issues du site de e-commerce Cdiscount.com et du jeu de données public Newsgroup.

## 2.2 État de l'art

### 2.2.1 Approche générative en classification supervisée

Dans le cadre de la classification supervisée, l'objectif est de déterminer la relation entre une variable explicative $X$ et une variable $Y$ à partir d'exemples $(X_i, Y_i)_{i=1,\dots,n}$. En notant $K$ le nombre de catégories, ou classes, auxquelles peut appartenir la variable $Y$, l'objectif est donc d'estimer $\mathbb{P}(Y = k|X)$ pour $k \in \{1, \dots, K\}$.

Pour cela, une approche classique consiste à estimer la loi de la variable aléatoire $X$ conditionnellement à $Y$ à partir des exemples à disposition. Une fois cette loi estimée, la relation entre les variables aléatoires $X$ et $Y$ est alors obtenue en appliquant la formule de Bayes, qui indique que pour tout événement $A$ et $B$ tel que $\mathbb{P}(B) = 0$ on a

$$\mathbb{P}(A) = \frac{P(B|A)P(A)}{P(B)}.$$

Alors, en notant $f_k$ la densité de la variable aléatoire $X$ conditionnellement à $Y = k$, la formule de Bayes donne pour tout $k \in \{1, \dots, K\}$

$$\mathbb{P}(Y = k|X) = \frac{\mathbb{P}(Y = k)f_k(X)}{\sum_{i=1}^{K} \mathbb{P}(Y = i) \cdot f_i(X)}.$$

Et donc, en notant $\widehat{f_k}$ un estimateur de la densité de la variable aléatoire $X$ conditionnellement à $Y = k$ et $\pi_k$ un estimateur de $\mathbb{P}(Y = k)$, un estimateur de la relation entre les variables aléatoires $X$ et $Y$ est obtenu par

$$\mathbb{P}(Y = k|X) = \frac{\pi_k \widehat{f_k}(X)}{\sum_{i=1}^{K} \pi_i \cdot \widehat{f}_i(X)}.$$

Pour obtenir cet estimateur, il est donc nécessaire d'avoir des estimateurs de $\mathbb{P}(Y = k)$ et $f_k$. En pratique, il est courant de procéder par modèles de mélange (McLachlan and Peel, 2004; McLachlan and Basford, 1988; Figueiredo and Jain, 2002). La théorie des modèles de mélange (voir par exemple McLachlan and Peel (2004)) permet de séparer une population, ou ensemble d'observations, en plusieurs sous-ensembles issus de lois différentes. En pratique, il est très fréquent d'utiliser des lois gaussiennes (on parle alors de mélange gaussien), mais d'autres distributions peuvent être utilisées selon les hypothèses.

Plus formellement, on dit que la variable aléatoire $X$ de densité $f_X$ est issue d'un modèle de mélange de $p$ densités $f_1, \ldots, f_p$ si la densité de $X$ s'écrit sous la forme

$$f_X(x) = \sum_{i=1}^{p} \pi_i f_i(x).$$

Une variable aléatoire $X$ est dite issue d'un mélange gaussien dans le cas particulier où, pour tout $i$, chaque fonction $f_i$ est la densité d'une loi normale de moyenne $\mu_i$ et de variance $\sigma_i$, et les proportions vérifient la condition suivante :

$$\forall i, \quad \pi_i \in [0; 1] \quad \text{et} \quad \sum_i \pi_i = 1.$$

Les paramètres du mélange $(\mu_i, \sigma_i, \pi_i, i = 1, \ldots, p)$, caractérisant la variable aléatoire $X$ peuvent alors être estimés par un algorithme EM, pour *Expectation-Maximisation* (Dempster *et al.*, 1977; Bishop, 2006).

L'algorithme EM, pour *Expectation-Maximisation* en anglais (Dempster *et al.*, 1977; Bishop, 2006), est une façon d'estimer les paramètres d'un modèle de mélange en maximisant la vraisemblance (ou log-vraisemblance) des données observées. En supposant qu'on dispose de $N$ observations $(X_n)_{n=1,\ldots,N}$ indépendantes et identiquement distribuées selon un mélange gaussien, l'objectif est de trouver les moyennes $\mu$, covariances $\Sigma$ et proportions $\pi$ maximisant la fonction de log-vraisemblance, donnée par

$$\ln\left(P(X|\mu, \Sigma, \pi)\right) = \sum_{n=1}^{N} \ln\left\{\sum_{i=1}^{p} \pi_i \mathcal{N}\left(X_n|\mu_i, \Sigma_i\right)\right\},$$

où $\mathcal{N}\left(x|\mu_i, \Sigma_i\right)$ désigne la densité de la loi normale de moyenne $\mu_i$ et de variance $\Sigma_i$, estimée au point $x$. L'algorithme EM propose une approche itérative pour estimer les paramètres optimaux.

Une spécificité de cette approche est de ne pas faire d'hypothèse de constance des distributions entre les bases d'apprentissage et de test. En particulier, le fait d'estimer les proportions du mélange sur la base de test autorise la présence d'un biais entre les bases d'apprentissage et de test.

Dans la partie 2.4 consacrée aux simulations, nous verrons une application de cette méthode et ses avantages par rapport à la théorie du *covariate shift*. Toutefois, nous

verrons également que l'estimation des paramètres est difficile en pratique et que les limites de l'approche par mélange gaussien apparaissent rapidement lorsque les volumes de données augmentent ou en présence de données hétérogènes ou textuelles.

### 2.2.2 Adaptation de domaine

#### Contexte

L'adaptation de domaine, aussi appelée apprentissage par transfert (Pan and Yang, 2010; Villani, 2008), consiste à élaborer un modèle prédictif performant dans le cas où les ensembles de test et d'apprentissage vérifient des propriétés différentes. Contrairement au cadre du *covariate shift*, il s'agit ici de traiter des cas où la relation $P(Y|X)$ n'est pas forcément conservée entre les deux bases. Dans le cadre de l'adaptation de domaine, il est courant de noter $\Omega_s$ (resp. $\Omega_t$) le domaine d'apprentissage (resp. test).

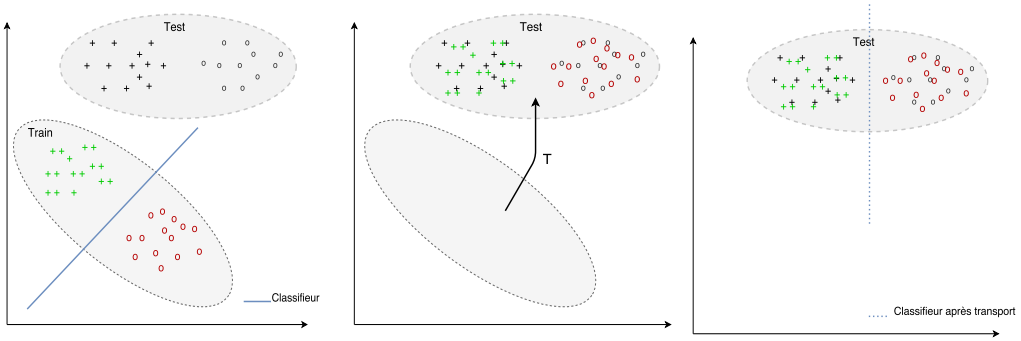
Plus formellement, les problèmes d'adaptation de domaine supposent l'existence d'une fonction de transport $T$ permettant de transporter les observations d'apprentissage et de test dans un même espace (Pan *et al.*, 2011; Gopalan *et al.*, 2011; Courty *et al.*, 2014), de sorte que

$$P_{train}(Y|X) = P_{test}(Y|T(X)).$$

La stratégie d'apprentissage consiste alors en une procédure en trois étapes (Courty *et al.*, 2014) :

1. estimer la fonction de transport optimal entre les domaines d'apprentissage et de test ;
2. transporter les données d'apprentissage sur le domaine de l'ensemble de test ;
3. apprendre le modèle sur les données transportées.

Ces trois étapes sont résumées sur la figure 2.2.




(a) Avant transport (b) Transport des données d'apprentissage (c) Modèle après transport

FIGURE 2.2 – Adaptation de domaines, une stratégie en trois étapes.

**Fonction de transport optimal**

Étant données $\mu_s$ et $\mu_t$ deux mesures de probabilité définies sur $\Omega_s \times \Omega_t$, et $c$ une fonction de coût, définie sur $\Omega_s \times \Omega_t$ et à valeur dans $\mathbb{R}^+$, le problème de transport optimal défini par Kantorovitch (1958) consiste alors à trouver une mesure de probabilité couplée $\gamma$ entre $\Omega_s$ et $\Omega_t$ telle que

$$\begin{aligned} \gamma_0 = \quad & \arg\min_{\gamma \in \mathcal{P}(\Omega_s \times \Omega_t)} \quad \int_{\Omega_s \times \Omega_t} c(x,y)\gamma(x,y)dxdy, \\ & \text{tels que} \qquad \int_{\Omega_t} \gamma(x,y)dy = \mu_s, \\ & \qquad\qquad\quad \int_{\Omega_s} \gamma(x,y)dx = \mu_t. \end{aligned}$$

Dans le cas où $\mu_s$ et $\mu_t$ correspondent à des mesures de probabilités discrètes, ce problème est alors équivalent à

$$\gamma_0 = \arg\min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle$$

où $\mathcal{P} = \{\gamma \in (\mathbb{R}^+)^{n_s \times n_t} \ / \ \gamma \cdot 1_{n_t} = \mu_s, \gamma^T \cdot 1_{n_s} = \mu_t\}$ et $C$ est la matrice de coût induite de la fonction $c$, et $\langle \gamma, C \rangle$ désigne le produit scalaire entre les matrices $\gamma$ et $C$, obtenu par $\langle \gamma, C \rangle = \text{tr}(\gamma \cdot C^T)$.

En pratique, il est courant d'observer, pour chaque espace $\Omega_s$ et $\Omega_t$, des échantillons de données de tailles $n_s$ et $n_t$ respectivement et notés $X_s = (x_1^s, \ldots, x_{n_s}^s)$ et $X_s = (x_1^t, \ldots, x_{n_t}^t)$. Un choix usuel pour $\mu_s$ et $\mu_t$ est alors donné par

$$\mu_s = \sum_{i=1}^{n_s} \frac{1}{n_s} \delta_{x_i^s} \quad \text{et} \quad \mu_t = \sum_{i=1}^{n_t} \frac{1}{n_t} \delta_{x_i^t}$$

où $\delta_{x_i}$ correspond à la fonction de Dirac au point $x_i$. Un choix courant pour la matrice de coût est la matrice de distance euclidienne, de sorte que $C_{i,j} = (x_i^s - x_j^t)^2$

Toutefois, cette approche s'avère vite limitée lorsque le nombre d'observations et la dimension des observations augmentent. Il devient alors difficile de déterminer la fonction de transport optimal, et la résolution du problème d'optimisation peut conduire à une relation de transport entre $\Omega_s$ et $\Omega_t$ possédant de nombreuses irrégularités (cas similaire à la situation de sur-apprentissage). Pour se prémunir de ce phénomène, Cuturi propose dans Cuturi (2013) d'introduire un terme de régularisation, dépendant d'un paramètre $\lambda$, et de minimiser la fonction suivante :

$$\gamma_0^\lambda = \arg\min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle - \frac{1}{\lambda} h(\gamma), \text{ où } h(\gamma) = -\sum_{i,j} \gamma(i,j) \log \gamma(i,j).$$

Cuturi montre alors que, pour $\lambda > 0$ donné, la solution s'écrit sous la forme $\gamma_0^\lambda = \text{diag}(u) \exp(-\lambda C) \text{diag}(v)$, où $u$ et $v$ sont deux vecteurs de réels positifs définis à un facteur multiplicatif près.

Dans Cuturi (2013), Cuturi montre également que la détermination des vecteurs $u$ et $v$ peut être faite de façon itérative par la méthode de Sinkhorn-Knopp (Knight, 2008).

Bien que cette approche soit pertinente dans de nombreux cas, nous verrons ses limites dans la partie 2.4 en réalisant des expériences sur données simulées. Plus précisément, nous verrons que la fonction de transport obtenue ne s'avère pas toujours pertinente en présence d'un fort biais dans les proportions de chaque catégorie associée à la variable cible entre les espaces $\Omega_s$ et $\Omega_t$. Bien qu'il soit possible de trouver une fonction de coût pertinente dans ce cas, nous proposons une nouvelle approche, développée dans la section suivante.

## 2.3 Solution proposée : une approche itérative

### 2.3.1 Théorie

Nous proposons ici de formaliser l'approche utilisée dans le cadre d'un challenge organisé par cDiscount sur la plateforme datascience.net [1], et pour lequel un biais important était observé entre les bases d'apprentissage et de test. Ce biais lié au processus utilisé pour la constitution de la base de test (voir détails à la section 2.5.1), et la spécificité des données textuelles, nous ont poussé à utiliser une approche particulière, inspirée de la théorie du *covariate shift* (présentée dans la chapitre 1), que nous détaillons dans cette section.

Plus précisément, dans le cadre du *covariate shift*, l'utilisation de la distribution des variables explicatives permet de pallier le biais, tandis que la spécificité de l'approche proposée réside dans l'utilisation de la distribution de la variable cible pour l'ajustement du modèle prédictif. Intuitivement, le principe de cette approche consiste à estimer la distribution de la variable cible sur la base de test pour mettre à jour les poids associés aux observations de l'ensemble d'apprentissage et mimer l'ensemble de test.

La proposition 5 adapte les résultats obtenus par Shimodaira dans Shimodaira (2000) sous les hypothèses présentées à la section 2.1.1.

**Proposition 5.** *Pour toute variable aléatoire $Z$, on note $p_Z$ la fonction de densité de la variable aléatoire $Z$.*

*Soit $d$, $n$ des entiers naturels strictement positifs, $\mathcal{X}$ un sous-ensemble de $\mathbb{R}^d$ et $\mathcal{Y} = \{1, \dots, n\}$. Soit $\ell$ une fonction de perte, $(X_s, Y_s)$ et $(X_t, Y_t)$ des couples de variables aléatoires à valeurs dans $\mathcal{X} \times \mathcal{Y}$ tels que*

$$\exists M > 0 \text{ tel que } \ell(Y_s, m(X_s)) \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad p_{Y_s}(y) > 0 \text{ et } \ell(y, m(x)) < M,$$

*et $m$ une fonction définie sur $\mathcal{X}$ et à valeurs dans $\mathcal{Y}$.*

1. `https://www.datascience.net/fr/challenge/20/details`

*Si les couples de variables aléatoires $(X_s, Y_s)$ et $(X_t, Y_t)$ ainsi définis vérifient*

$$p_{X_s|Y_s} = p_{X_t|Y_t},$$

*alors, en définissant pour tout $y \in \mathcal{Y}$ les poids $\omega(y)$ par $\omega(y) = p_{Y_t}(y)/p_{Y_s}(y)$ on a*

$$\mathbb{E}\left[\omega(Y_s)\ell(Y_s, m(X_s))\right] = \mathbb{E}\left[\ell(Y_t, m(X_t))\right].$$

*Démonstration.* Supposons qu'il existe $M > 0$ tel que pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$ on a $\ell(y, m(x)) < M$, et que $p_{Y_s}(y) > 0$ pour tout $y \in \mathcal{Y}$.

Alors, comme $\mathcal{Y}$ est un ensemble de cardinal fini il existe $m > 0$ tel que $p_{Y_s}(y) \geq m$ pour tout $y \in \mathcal{Y}$. Par ailleurs pour tout $y \in \mathcal{Y}$ on a $p_{Y_s}(y) \leq 1$. Donc

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad 0 \leq \frac{p_{Y_t}(y)}{p_{Y_s}(y)}\ell(y, m(x))p_{X_s,Y_s}(x, y) \leq \frac{1}{m}M p_{X_s,Y_s}(x, y).$$

Par conséquent, la fonction $(x, y) \mapsto \frac{p_{Y_t}(y)}{p_{Y_s}(y)}\ell(y, m(x))p_{X_s,Y_s}(x, y)$ est intégrable sur $\mathcal{X} \times \mathcal{Y}$ et

$$\begin{aligned}
\mathbb{E}\left[\omega(Y_s)\ell(Y_s, m(X_s))\right] &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{Y_t}(y)}{p_{Y_s}(y)}\ell(y, m(x))p_{X_s,Y_s}(x, y)dxdy, \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{Y_t}(y)}{p_{Y_s}(y)}\ell(y, m(x))p_{X_s|Y_s}(x|y)p_{Y_s}(y)dxdy, \\
&= \int_{\mathcal{X}\times\mathcal{Y}} p_{Y_t}(y)p_{X_s|Y_s}(x|y)\ell(y, m(x))dxdy.
\end{aligned}$$

Or pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$ on a $p_{X_s|Y_s}(x|y) = p_{X_t|Y_t}(x|y)$ par hypothèse, donc

$$\mathbb{E}\left[\omega(Y_s)\ell(Y_s, m(X_s))\right] = \int_{\mathcal{X}\times\mathcal{Y}} p_{X_t}(x)p_{Y_t|X_t}(y|x)\ell(y, m(x))dxdy.$$

ce qui permet de conclure. □

**Remarque :** Dans le cas où les données d'apprentissage (resp. de test) sont des copies indépendantes et identiquement distribuées du couple de variables aléatoires $(X_s, Y_s)$ (resp. $(X_t, Y_t)$), la proposition 5 montre qu'il est possible de mimer la base de test à partir de poids dépendants de la distribution de la variable cible sur les bases d'apprentissage et de test. Cependant, par définition de la base de test, la variable cible (et donc sa distribution) est inconnue. La section suivante propose une approche itérative pour estimer cette distribution.

### 2.3.2 Algorithme proposé

En l'absence de shift, la distribution de la variable cible $Y$ est identique sur les bases d'apprentissage et de test. Dans ce cas, tous les poids $\omega$ définis à la proposition 5 sont constants (et unitaires). Afin d'établir une approche capable de converger en une itération en l'absence de shift, nous supposons dans un premier temps que les poids sont unitaires. Nous proposons ensuite d'utiliser les prédictions fournies par le modèle sur la base de test pour obtenir une estimation plus précise de la loi de la variable cible sur la base de test, et procéder ainsi de façon itérative. Cette approche est décrite à l'algorithme 3, où $\pi_{\chi^2}(Z_1, Z_2)$ représente la *p-value* associée au test du Chi 2 de Pearson entre les variables aléatoires discrètes $Z_1$ et $Z_2$. Ce test permet, pour un échantillon aléatoire $z_1, \ldots, z_n$, de quantifier la possibilité de distinguer si cet échantillon a été engendré selon la loi de la variable aléatoire $Z_1$ ou $Z_2$. En d'autres termes, si la *p-value* associée à ce test est inférieure à 0.01 par exemple, il n'est statistiquement pas pertinent de distinguer les lois des variables aléatoires $Z_1$ et $Z_2$.

**Algorithme 3 :** Estimation de la distribution de la variable cible

**Input** $\ell$, $(X_{s,1}, Y_{s,1}), \ldots, (X_{s,n_s}, Y_{s,n_s})$, et $X_{t,1}, \ldots, X_{t,n_t}$
**Initialisation** $\omega = 1$, $p^0_{Y_t} = p_{Y_s}$, $p^1_{Y_t} = p_{g_\omega(X^t)}$
**while** $\pi_{\chi^2}(p^0_{Y_t}, p^1_{Y_t}) < 0.01$ **do**
  Calculer les poids $\omega(y) = p^1_{Y_t}(y)/p_{Y_s}(y)$
  Remplacer $p^0_{Y_t}$ par $p^1_{Y_t}$
  Estimer $g_\omega$ qui minimise l'erreur empirique pondérée.
  Remplacer $p^1_{Y_t}$ par $p_{g_\omega(X^t)}$
**end while**
**return** $p^1_{Y_t}$

L'objectif des itérations est d'estimer la distribution de la variable cible sur la base de test, pour mettre à jour les poids associés à chaque observation de la base d'apprentissage. Les itérations sont donc profitables dès lors que la différence entre la distribution estimée de la variable cible évolue significativement entre deux itérations.

Dans cette section, nous avons vu qu'il est possible de mimer la base de test à partir de poids dépendant de la distribution de la variable cible sur les bases d'apprentissage et de test, et que la distribution de la variable cible peut être estimée selon une approche itérative. Dans la section suivante, nous proposons de valider cette approche sur des données simulées, avant de procéder à des expérimentations sur données réelles (section 2.5)

## 2.4 Simulations

### 2.4.1 Génération des données

Afin d'étudier la pertinence de l'approche proposée dans la section 2.3, nous avons procédé à des expériences sur des données simulées. Soit $X_0$ et $X_1$ deux variables aléatoires définies sur $\mathbb{R}^2$ selon

$$X_0 \sim \mathcal{N}\left(\begin{pmatrix}-a\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right) \quad \text{et} \quad X_1 \sim \mathcal{N}\left(\begin{pmatrix}a\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right),$$

où $a$ est un nombre réel strictement positif. Nous noterons $f_0$ et $f_1$ les densités de $X_0$ et $X_1$ respectivement. A partir de ces densités, nous avons engendré deux populations représentant respectivement les échantillons d'apprentissage et de test, constituées chacune de 5 000 observations ($n_{train} = n_{test} = 5\ 000$).

Nous nous placerons ici dans le cadre de la classification binaire. Plus précisément, à partir de l'ensemble de données présent dans l'échantillon d'apprentissage, l'objectif sera d'élaborer un modèle prédictif capable de déterminer si une observation $x_{test,i} \in \mathbb{R}^2$ de l'ensemble de test, avec $i \in \{1, \dots, n_{test}\}$, a été engendrée à partir de la variable aléatoire $X_0$ ou $X_1$.

Dans ce cas, pour tout $i \in \{1, \dots, n_{train}\}$, la variable cible $y_{train,i}$ vaut 1 si l'observation $x_{train,i}$ a été engendrée selon la loi de la variable aléatoire $X_1$, 0 sinon.

Afin d'introduire un biais entre les bases d'apprentissage et de test, nous avons utilisé les modèles de mélange suivants, où pour tout $x \in \mathbb{R}$, $f_{train}$ (resp. $f_{test}$) désigne la densité des observations présentes dans la base d'apprentissage (resp. de test) évaluée au point $x \in \mathbb{R}^2$ :

$$f_{train}(x) = \pi_{train} \cdot f_0(x) + (1 - \pi_{train}) \cdot f_1(x) \quad \text{avec} \quad \pi_{train} = 0.1 \tag{2.1}$$

et

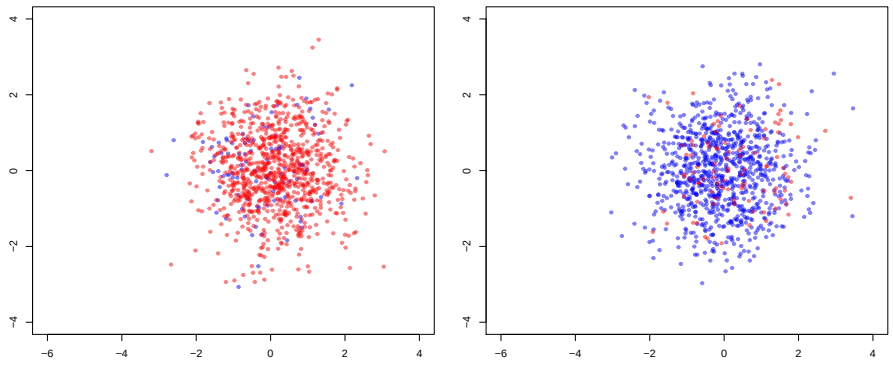
$$f_{test}(x) = \pi_{test} \cdot f_0(x) + (1 - \pi_{test}) \cdot f_1(x) \quad \text{avec} \quad \pi_{test} = 0.9 \tag{2.2}$$

Chaque échantillon est donc issu d'un modèle de mélange de gaussiennes. Ainsi, la base d'apprentissage est composée de 90% de données issues de la variable $X_1$, que nous appellerons *cas positifs*, tandis que la base de test comporte 90% de données issues de la variable aléatoire $X_0$, que nous appellerons *cas négatifs*.

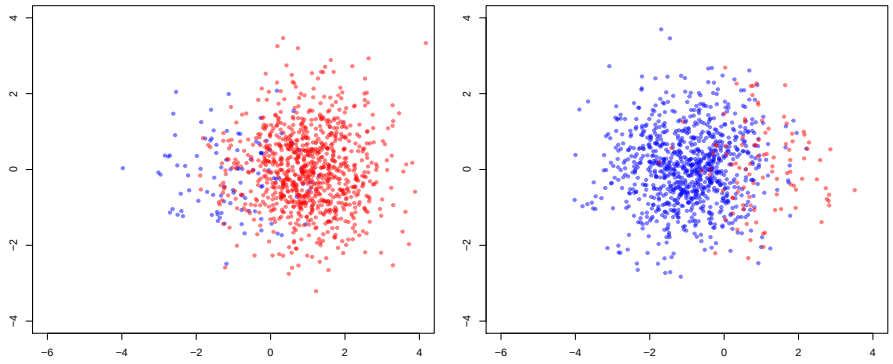
Les données ainsi engendrées sont représentées pour quelques valeurs du paramètre $a$ aux figures 2.3(a) à 2.3(c). On distingue trois cas :

- $a = 0.1$ : les deux populations sont très rapprochées. Nous verrons qu'il est très difficile de les identifier, indépendamment de l'approche choisie ;
- $a = 1$ : une partie des populations peut être distinguée, mais les deux ensembles possèdent un fort recouvrement ;

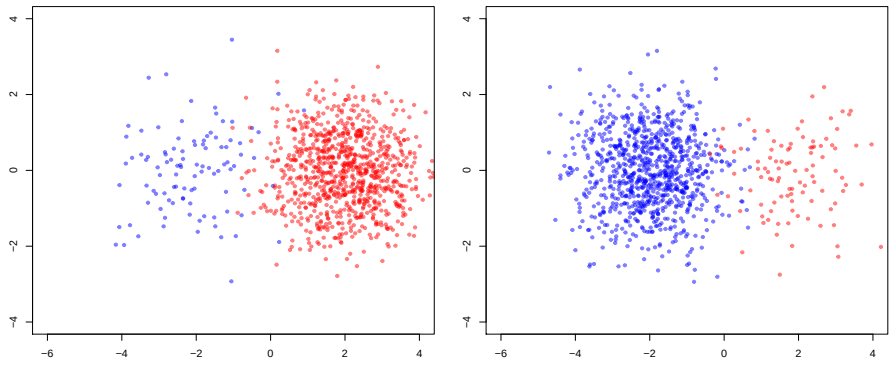
— $a = 2$ : les deux populations sont quasi parfaitement distinctes. Dans ce cas nous verrons qu'il est facile de distinguer les deux groupes d'observations, et le shift n'aura que très peu d'influence sur la performance des modèles.



(a) $a = 0.1$. Gauche : échantillon d'apprentissage, droite : échantillon de test.



(b) $a = 1$. Gauche : échantillon d'apprentissage, droite : échantillon de test.



(c) $a = 2$. Gauche : échantillon d'apprentissage, droite : échantillon de test.

FIGURE 2.3 – Représentation des bases d'apprentissage et de test en fonction de $a$

Cette situation de fort déséquilibre est en pratique très fréquemment observée. Par exemple, dans le cadre d'analyse de données issues de détection de fraude bancaire, l'ensemble des données d'apprentissage ne contient souvent que l'ensemble des cas suspectés de fraudes. Sur cet ensemble d'apprentissage les cas positifs sont alors les cas de fraude avérée, tandis que les cas négatifs représentent ceux pour lesquels, après analyse, aucune fraude n'est détectée. Dans le cas où la base de test comporte l'ensemble des transactions sur une période donnée ou l'ensemble des clients, il est alors évident que la proportion de fraudeurs (et donc de cas positifs) sera largement inférieure à celle observée sur l'ensemble des clients suspectés de fraude.

### 2.4.2 Les différentes approches en pratique

Nous proposons ici d'analyser les résultats, sur les données simulées, issues des approches évoquées dans les sections précédentes et de discuter de leurs performances respectives : *covariate shift* (section 1.4.2), mélanges gaussiens (section 2.2), adaptation de domaine (section 2.2.2) et approche itérative (section 2.3).

Afin d'estimer, pour chaque observation, la probabilité d'appartenir à l'une ou l'autre des classes ($Y$ est binaire), nous avons choisi un modèle logistique. Comme chaque observation est entièrement caractérisée par ses coordonnées ($x = (x_1, x_2)$), cela signifie que nous cherchons un vecteur réel $\alpha = (\alpha_1, \alpha_2)$ de sorte que

$$p(y = 1|x, \alpha) = \frac{1}{1 + \exp(-\alpha^T \cdot x)}.$$

Le vecteur optimal $\alpha$ est celui maximisant la log-vraisemblance du modèle sur l'ensemble d'apprentissage, donnée par

$$\mathcal{L}(Y|X, \alpha) = \sum_{i=1}^{n_{train}} y_i \ln\left(p(y_i = 1|x_i, \alpha)\right) + (1 - y_i) \ln\left(1 - p(y_i = 1|x_i, \alpha)\right).$$

En introduisant les poids $\omega$ dans la régression logistique, le vecteur $\alpha$ est celui maximisant la log-vraisemblance pondérée, définie par

$$\mathcal{L}_\omega(Y|X, \alpha) = \sum_{i=1}^{n_{train}} \omega_i \{ y_i \ln\left(p(y_i = 1|x_i, \alpha)\right) + (1 - y_i) \ln\left(1 - p(y_i = 1|x_i, \alpha)\right) \}.$$

#### *Covariate Shift*

Afin d'appliquer la théorie du *covariate shift*, il convient d'estimer la densité des données sur les bases d'apprentissage et de test. Ces deux densités, $f_{train}$ et $f_{test}$, sont fournies par les équations (2.1) et (2.2), et les poids sont donnés par

$$\omega(x) = \frac{f_{test}(x)}{f_{train}(x)}, \quad \forall x \in \mathbb{R}^2.$$

Ici, les données ayant été engendrées selon des lois normales, les distributions $f_{test}$ et $f_{train}$ sont connues, et nous sommes donc dans une situation plus favorable que la majorité des situations observées en pratique.

Comme nous sommes dans le cas où les variables explicatives appartiennent à $\mathbb{R}^2$ (engendrées selon des lois gaussiennes en dimension 2), chaque observation $x$ de la base d'apprentissage s'écrit sous la forme $x = (x_1, x_2)$. Les figures 2.4(a) à 2.4(c) représentent la valeur des poids en fonction des coordonnées de chaque point (abscisse $x_1$, ordonnée $x_2$). Ces figures montrent que les poids sont décroissants avec l'abscisse des points, ce qui témoigne d'un biais entre la base d'apprentissage et la base de test. Ce biais est par ailleurs uniquement lié à l'abscisse des points. En l'absence de biais, les distributions sur les bases d'apprentissage et de test seraient identiques et tous les poids seraient constants et égaux à 1.

Les performances obtenues sont décrites et analysées à la section 2.4.3.

**Mélange Gaussien**

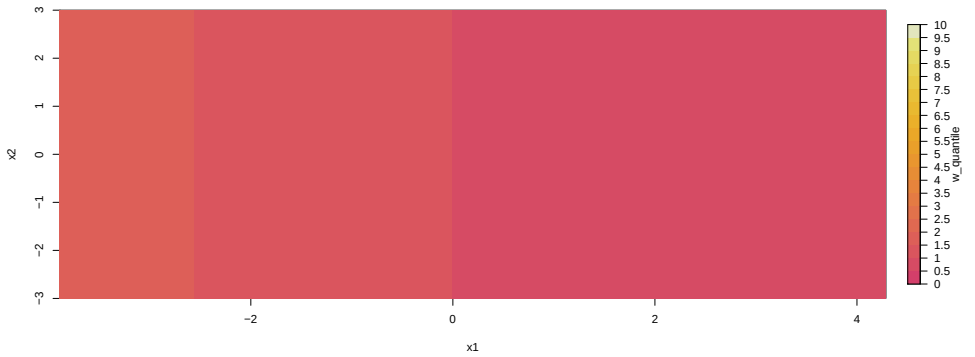
Pour estimer les paramètres du modèle gaussien, nous avons utilisé la bibliothèque *mclust* du logiciel R. La table 2.2 détaille les paramètres obtenus sur la base d'apprentissage, où $\widehat{\mu_0}$ (resp. $\widehat{\mu_1}$) et $\widehat{\sigma_0}$ (resp. $\widehat{\sigma_1}$) désignent les paramètres estimés de la loi de la variable aléatoire $X_0$ (resp. $X_1$), comme introduits à la section 2.4.1, tandis que $\widehat{\pi}_{train}$ désigne la valeur estimée de la proportion de cas positifs dans le mélange. Les lois ainsi estimées sont représentées graphiquement aux figures 2.5(a) à 2.5(c).

| | $\widehat{\mu_0}$ | $\widehat{\sigma_0}$ | $\widehat{\mu_1}$ | $\widehat{\sigma_1}$ | $\widehat{\pi}_{train}$ |
|---|---|---|---|---|---|
| $a = 0.1$ | - | - | $\begin{pmatrix} 0.09545 \\ 0.0097 \end{pmatrix}$ | $\begin{pmatrix} 0.9722 & 0 \\ 0 & 0.9722 \end{pmatrix}$ | 0 |
| $a = 1$ | $\begin{pmatrix} -0.8425 \\ 0.0904 \end{pmatrix}$ | $\begin{pmatrix} 0.9793 & 0 \\ 0 & 0.9793 \end{pmatrix}$ | $\begin{pmatrix} 1.0429 \\ -0.0538 \end{pmatrix}$ | $\begin{pmatrix} 0.9793 & 0 \\ 0 & 0.9793 \end{pmatrix}$ | 0.1200 |
| $a = 2$ | $\begin{pmatrix} -2.009 \\ -0.047 \end{pmatrix}$ | $\begin{pmatrix} 1.005 & 0 \\ 0 & 1.005 \end{pmatrix}$ | $\begin{pmatrix} 2.016 \\ -0.008 \end{pmatrix}$ | $\begin{pmatrix} 1.005 & 0 \\ 0 & 1.005 \end{pmatrix}$ | 0.0985 |

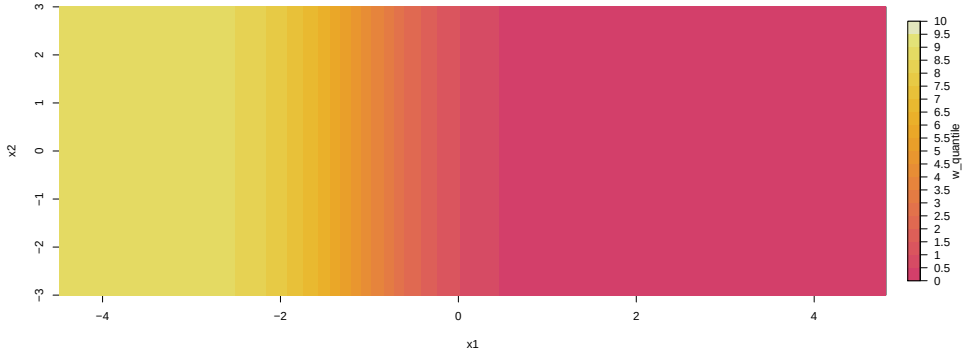
TABLE 2.2 – Résultats obtenus par application de la stratégie de *covariate shift*.

En faisant l'hypothèse d'un mélange gaussien, ces résultats nous permettent d'avoir une estimation des lois de $X$ conditionnellement à $Y = 0$ et $Y = 1$ sur la population de test, notées respectivement $\widehat{f_0}$ et $\widehat{f_1}$. Alors, en notant $\widehat{\pi}_t$ l'estimateur de la proportion de données issues de la variables aléatoire $X_1$ sur la base de test, la formule de Bayes fournit un estimateur de la variable $Y$ par

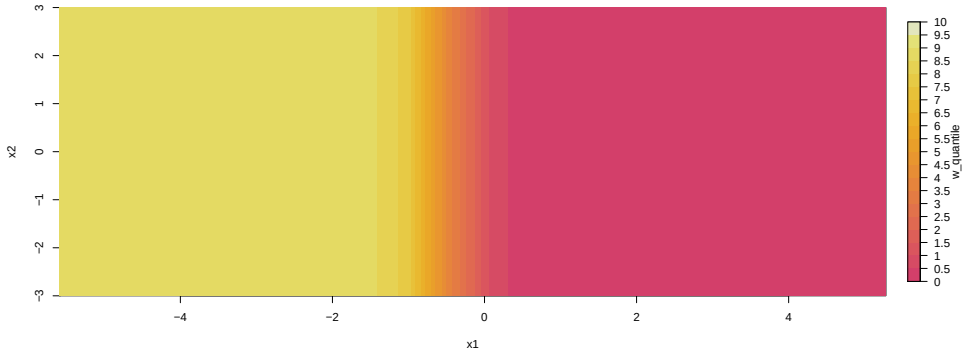
$$\mathbb{P}(Y = 1|X) = \frac{\widehat{\pi}_t \widehat{f_0}(X)}{\widehat{\pi}_t \cdot \widehat{f_1}(X) + (1 - \widehat{\pi}_t) \cdot \widehat{f_0}(X)}.$$




(a) $a = 0.1$

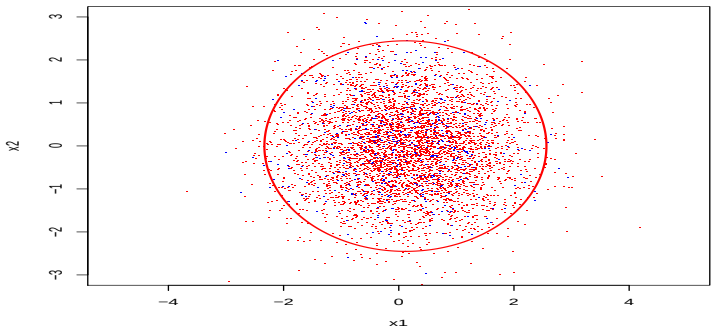



(b) $a = 1$

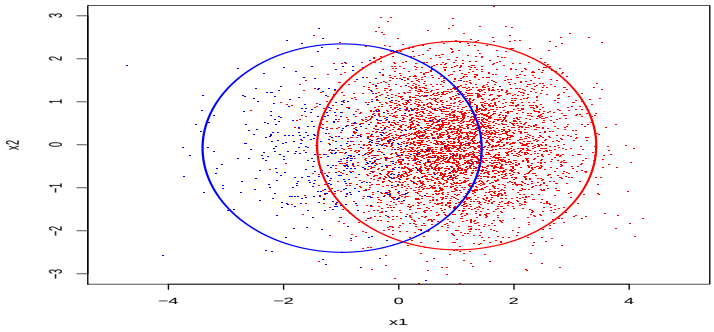



(c) $a = 2$

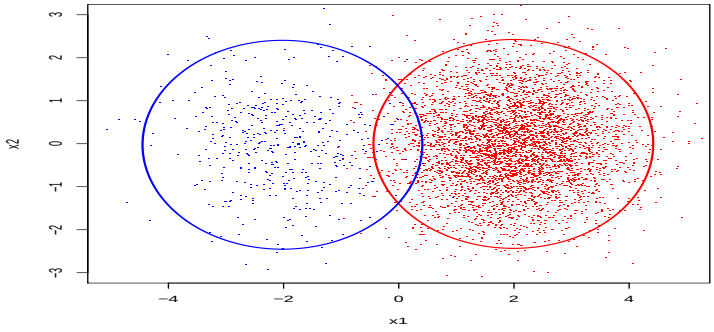
FIGURE 2.4 – Représentation des poids $\omega$ en fonction de $a$.




(a) $a = 0.1$.




(b) $a = 1$.




(c) $a = 0.1$.

FIGURE 2.5 – Représentation des densités estimées en fonction de $a$.

En pratique, il est courant de considérer que les proportions des classes sont conservées entre les bases d'apprentissage et de test. Ici, cette hypothèse est fausse et nous proposons d'estimer $\widehat{\pi}_t$ par maximum de vraisemblance. Étant donné $\widehat{\pi}_t$, la vraisemblance des

données sur la base de test s'écrit

$$\mathcal{L}(X|\widehat{\pi}^t) = \prod_{i=1}^{n_{test}} \left\{(1-\widehat{\pi}_t)\cdot f_0(x_{test,i}) + \widehat{\pi}_t \cdot f_1(x_{test,i})\right\}.$$

La maximisation de cette quantité, à l'aide de la fonction *optim* de $R$, conduit à l'estimation de la valeur de $\pi_t$ dans chacun des cas. Une fois cette quantité estimée, nous pouvons inférer $\mathbb{P}(Y=1|X=x)$ pour toute observation $x$ de la base de test, et en déduire la prédiction des valeurs de la variable cible $Y$ sur la base de test. Les résultats associés à ces prévisions sont présentés et analysés à la section 2.4.3.

### Adaptation de domaine

Afin d'illustrer l'approche par adaptation de domaine, nous avons utilisé la distance euclidienne comme fonction de coût et l'algorithme de Sinkhorn-Knopp (Cuturi, 2013), comme proposé dans les codes mis à disposition par Rémi Flamary [2]. Nous avons par ailleurs adapté ces codes à notre jeu de données test.

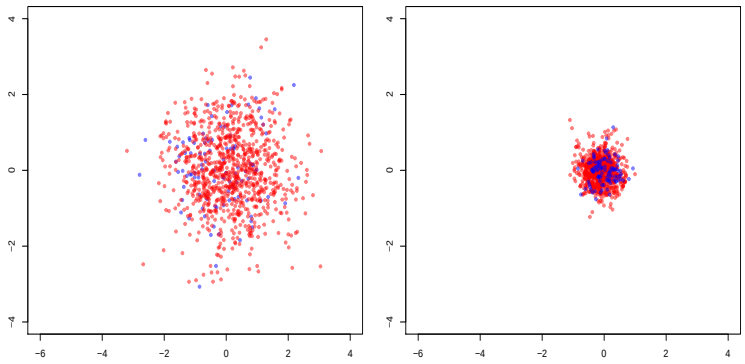
Les données transportées sont représentées sur les figures 2.6(a) à 2.6(c). Celles-ci illustrent les limites de l'approche par adaptation de domaine dans le cas où la distribution de la variable cible varie très fortement entre les ensembles d'apprentissage et de test. Plus précisément, on constate que la fonction de transport optimal trouvée correspond à celle obtenue dans le cas où les données ont subi une rotation d'un angle de 180 degrés, autrement dit, au cas où les observations positives de la base d'apprentissage ont pris la place des observations négatives dans la base de test (et réciproquement).

La fonction de transport optimal obtenue ne correspond donc pas aux modifications observées en pratique. En conséquence, nous verrons à la section 2.4.3 que les résultats numériques issus de la modélisation par adaptation de domaine sont moins bons que ceux obtenus par les autres approches.

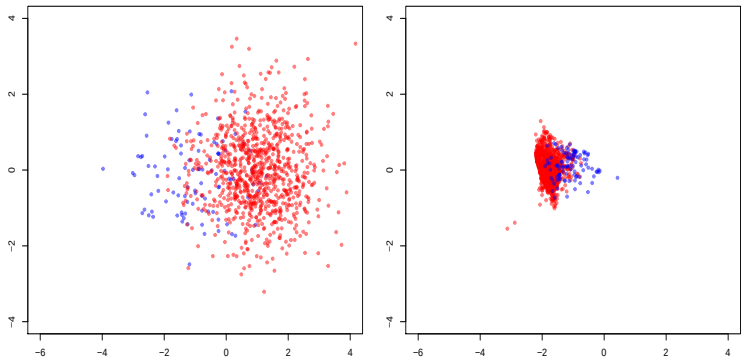
### Approche itérative

Les résultats issus de l'approche itérative proposée sont présentés graphiquement sur les figures 2.7(a) à 2.7(c). Ces dernières illustrent bien la convergence de la frontière estimée entre les données négatives et positives. Lors de la première itération, celle-ci coïncide avec la frontière optimale trouvée sur la base d'apprentissage. Puis la frontière est modifiée jusqu'à converger vers la frontière optimale sur la base de test. Cette frontière, qui définit plus formellement l'ensemble des observations pour lesquelles $P(Y=1|X) = P(Y=0|X) = 0.5$, illustre très clairement le fait que l'hypothèse de constance de $P(Y|X)$ (la relation explicative de $Y$ sachant $X$) n'est pas respectée entre les bases d'apprentissage et de test.

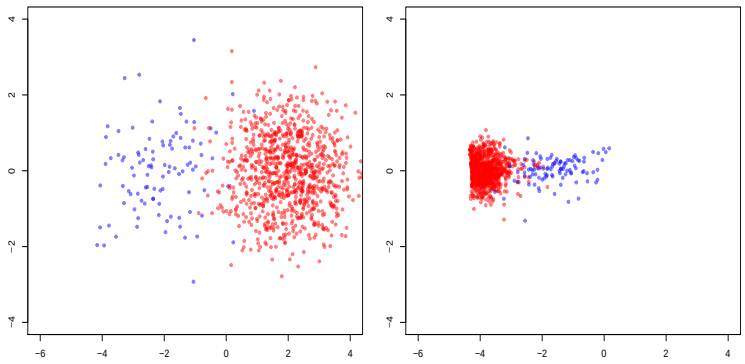
2. `http://remi.flamary.com/soft/soft-transp.html`



(a) $a = 0.1$. Gauche : échantillon d'apprentissage, droite : échantillon d'apprentissage transporté.

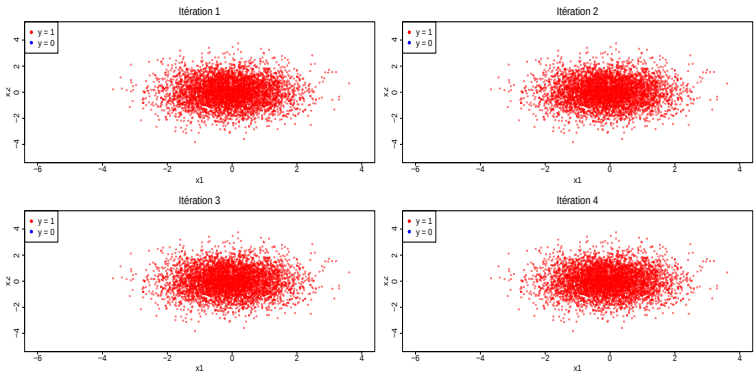


(b) $a = 1$. Gauche : échantillon d'apprentissage, droite : échantillon d'apprentissage transporté.

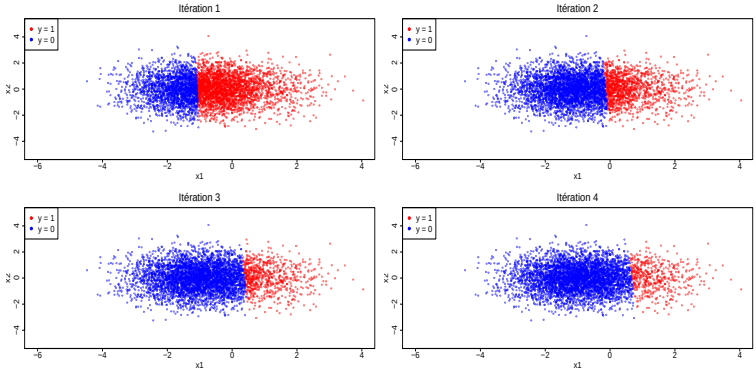


(c) $a = 2$. Gauche : échantillon d'apprentissage, droite : échantillon d'apprentissage transporté.

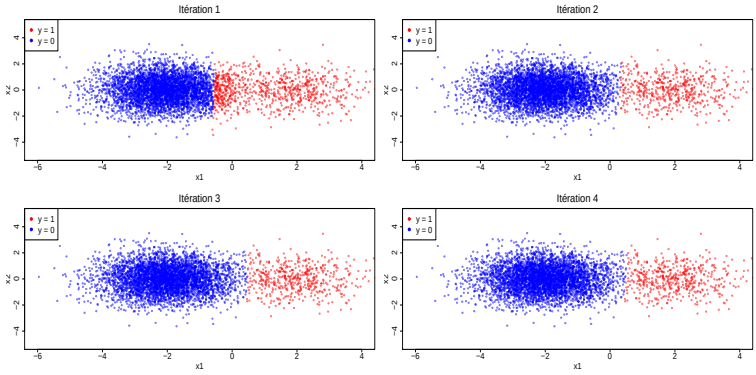
FIGURE 2.6 – Représentation des densités d'apprentissage originales et transportées en fonction de $a$ (approche par adaptation de domaine)




(a) $a = 0.1$




(b) $a = 1$




(c) $a = 2$

FIGURE 2.7 – Convergence du modèle itératif en fonction de $a$.

La table 2.4.2 présente les valeurs numériques associées à l'approche itérative pour les cinq premières itérations, où $n_i$ désigne le nombre d'itérations, $\widehat{\pi}_t$ la proportion de cas positifs estimée sur la base de test, et $\pi_{\chi^2}$ la p-value associée au test de Chi 2 de

Pearson et utilisée pour définir le critère d'arrêt de l'approche itérative (voir section 2.3.2).

| | $a = 0.1$ | | | $a = 1$ | | | $a = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | Erreur | $\widehat{\pi}_t$ | $\pi_{\chi^2}$ | Erreur | $\widehat{\pi}_t$ | $\pi_{\chi^2}$ | Erreur | $\widehat{\pi}_t$ | $\pi_{\chi^2}$ |
| 1 | 89.96% | 0.89 | - | 46.68% | 0.56 | 0.6259 | 6.86% | 0.18 | 0.2632 |
| 2 | 89.96% | 0.89 | - | 19.48% | 0.31 | $< 0.001$ | 1.44% | 0.10 | $< 0.001$ |
| 3 | 89.96% | 0.89 | - | 10.86% | 0.21 | $< 0.001$ | 1.22% | 0.10 | $< 0.001$ |
| 4 | 89.96% | 0.89 | - | 8.56% | 0.16 | $< 0.001$ | 1.22% | 0.09 | $< 0.001$ |
| 5 | 89.96% | 0.89 | - | 8.14% | 0.14 | $< 0.001$ | 1.22% | 0.09 | $< 0.001$ |

TABLE 2.3 – Résultats des simulations issues de l'approche proposée. Dans le cas $a = 0.1$, le test du Chi2 n'est pas réalisable car les prédictions sont constantes.

L'analyse des p-values montre que la convergence est très rapide. Dès la deuxième itération, la distribution de la variable cible obtenue sur la base de test n'est plus significativement différente entre deux itérations successives. En respectant le critère d'arrêt défini dans la section 2.3.2, les prédictions finales sont donc obtenues à l'issue de la deuxième étape, même si la performance des modèles continue de s'améliorer jusqu'à la cinquième itération. Ces résultats semblent indiquer des limites liées à notre critère d'arrêt. En pratique, il peut être plus pertinent de déterminer le nombre d'itérations à réaliser en procédant par validation croisée.

### 2.4.3 Analyse des résultats

Une synthèse des résultats des différentes approches décrites dans la section 2.4.2 est présentée à la table 2.4. Ces résultats montrent clairement que la performance des différentes approches dépend fortement des propriétés des bases de données d'apprentissage et de test. Une synthèse des hypothèses de chaque approche est proposée à la table 2.1. En particulier, lorsque les hypothèses propres à chaque modèle ne sont pas respectées, comme c'est le cas ici pour la régression classique, le *covariate shift*, et l'adaptation de domaine, de fortes erreurs sont observées en analysant les prédictions des modèles sur la base de test.

| Modèle | $a = 0.1$ | $a = 1$ | $a = 2$ |
|---|---|---|---|
| Régression logistique classique | 89.96% | 46.68% | 6.86% |
| *Covariate shift* | 89.96% | 41.46% | 7.64% |
| Mélange gaussien | 89.96% | 8.08% | 1.18% |
| Adaptation de domaine | 88.1% | 51.2% | 39.8% |
| Approche proposée | 89.96% | 8.14% | 1.06% |

TABLE 2.4 – Synthèse des résultats obtenus par les différentes approches.

Par ailleurs, ces expériences sur données simulées mettent en avant les limites de ces approches dans le cas où les deux populations sont quasi confondues ($a = 0.1$). Dans ce cas, aucun des modèles testés n'est capable d'identifier les deux populations et tous ont une erreur très importante, due au biais entre les données d'apprentissage et de test.

Dans le cas où les deux populations sont identifiables, on constate que les approches par mélange gaussien et l'approche itérative ont une performance largement supérieure aux autres modèles dans le cas $a = 1$, voire également dans le cas $a = 2$ pour l'approche itérative.

Outre ces résultats numériques, l'approche proposée montre des avantages en termes de champ d'application (les modèles de mélange gaussien sont peu adaptés aux données textuelles par exemple) et de rapidité d'exécution. En effet, les temps de calculs observés sur nos jeux de données montrent une forte différence selon le type de modèle. Plus précisément, les ordres de grandeurs sont de 0.2 seconde pour l'approche proposée (en effectuant 10 itérations) et de 45 secondes pour le modèle de mélange gaussien réalisé avec la librairie *mclust* du logiciel *R*.

Dans la section suivante, nous proposons de développer une application à partir de données réelles, pour laquelle l'approche par mélange gaussien n'est pas adaptée, tandis que l'approche itérative est pertinente.

## 2.5 Applications

### 2.5.1 Challenge Cdiscount

#### Contexte

Dans le cadre d'une compétition organisée sur la plateforme datascience.net, l'entreprise Cdiscount.com, spécialiste du e-commerce, a mis à disposition un jeu de données dans le but d'améliorer l'algorithme de catégorisation de produits présents sur son site. Cette compétition, qui s'est déroulée de mai à août 2015, a réuni plus de 800 participants. Nous proposons ici de détailler la stratégie que nous avons adoptée, et qui nous a permis d'obtenir le troisième modèle le plus performant de la compétition.

#### Description des données

Pour cette compétition, Cdiscount.com a fourni une base de données composée de 15.8 millions de produits avec pour chacun : la marque, le prix, l'intitulé (ou titre), la description et la catégorie (variable cible $Y$). Par ailleurs, les images de certains produits étaient également disponibles, mais ces dernières n'ont pas été utilisées dans le cadre de notre modèle. L'ensemble de ces données produits (hors catégorie) constituent les observations, $X$, à disposition.

Les catégories, représentant la variable cible, sont structurées selon une hiérarchie propre à cDiscount et sur trois niveaux : les catégories de niveau 1 étant les plus génériques, et les catégories de niveau 3 les plus précises. La table 2.5 donne un exemple de catégories, l'objectif étant de prédire la catégorie de niveau 3. Le référentiel complet comporte 5 790 catégories de niveau 3 (ensemble des valeurs prises par la variable cible).

| Catégorie 1 | Catégorie 2 | Catégorie 3 |
|---|---|---|
| Bateau Moteur | Greement - Voile | Voile Légère |
| Bateau Moteur | Greement - Voile | SPI - Gennaker |
| Bateau Moteur | Electricité | Cablage |
| Tatouage - Piercing | Encre à Tatouage | Encre à Tatouage |
| Téléphonie - GPS | Accessoire Téléphone | Coque - Bumper - Façade Téléphone |
| Téléphonie - GPS | Accessoire Téléphone | Découpe carte SIM |
| Sport | Textile | Kimono |
| Sport | Textile | Short - Bermuda |
| Sport | Textile | Survêtement - Jogging |

TABLE 2.5 – Exemple de catégories issues du référentiel de produits fourni par Cdiscount.com.

L'objectif de cette compétition était de prédire la catégorie de niveau 3 associée à chaque produit présent dans la base test, composée de 35 065 produits.

### Biais

Au sein de la base d'apprentissage, une analyse descriptive de la variable cible montre une forte hétérogénéité de la répartition des produits par catégorie. Par exemple, la catégorie « coque - bumper - facade téléphone » y représente 13.9% des produits présents, soit près de 2.2 millions de produits, tandis que certaines catégories, comme « voile légère » (produit pour bateau) ou « encre à tatouage » ne possèdent qu'un article dans la base d'apprentissage. En revanche, les échanges sur le forum de la compétition indiquent que la base de test a été construite de façon à avoir autant de produits dans chaque catégorie, pour des raisons propres aux organisateurs. Ce constat met clairement en évidence la présence d'un shift entre les bases d'apprentissage et de test.

### Modélisation

Face au volume important d'observations à disposition et pour des raisons pratiques, nous avons choisi d'utiliser des techniques d'apprentissage en ligne, aussi appelées *Online Learning* en anglais (Anderson, 2008), qui permettent de calibrer un modèle en considérant les données les unes après les autres, sans avoir à charger l'ensemble des observations à disposition en mémoire. Pour cela, nous avons utilisé Vowpal Wabbit

(Langford, 2015), une bibliothèque d'apprentissage en ligne, et avec l'option OAA (*One Against All*, ou un contre tous en français).

Dans le cas où la variable cible possède $n$ modalités $y_1, \ldots, y_n$ avec $n > 2$, le principe de la modélisation OAA est d'établir un modèle par classe et d'estimer $P(Y = y_i|X)$ grâce à une régression logistique, pour $i \in \{1, \ldots, n\}$, ce qui conduit à $n$ modèles : $m_1, \ldots, m_n$. Étant donnée une nouvelle observation $X$, une fois ces modèles obtenus la prédiction de la variable cible est donnée par

$$\widehat{Y} = y_{k^\star} \quad \text{avec} \quad k^\star = \arg \max_{i=1,\ldots,n} m_i(X).$$

Avant la modélisation, deux traitements des données ont été effectués. D'abord, nous avons uniformisé le titre et la description de l'ensemble des produits à disposition. Plus précisément, lors de cette étape nous avons converti les champs textuels en minuscules, supprimé les signes de ponctuation et accents, et procédé à une radicalisation (*stemming*) pour conserver uniquement la racine des mots présents dans chaque descriptif. Ensuite, nous avons converti les données uniformisées au format supporté par Vowpal Wabbit, et dont la structure est la suivante :

```
y w |D description |L libellé |M marque |P prix:valeur
```

Où $w$ représente le poids associé à chaque observation. Compte tenu de la durée nécessaire à l'apprentissage du modèle sur les données complètes (quelques heures) nous nous sommes limités à une seule itération pour l'estimation des poids, en les initialisant de façon à avoir une répartition uniforme des classes pondérées au sein de la base d'apprentissage. Les performances sur la base de test du challenge ont été les suivantes : 65.8% de bonnes réponses à la première itération, 66.3% à la seconde itération.

Des techniques d'agrégation de modèles nous ont ensuite permis d'obtenir un score de 66.7%, qui nous a amené à la troisième place à l'issue de la compétition.

Les données cDiscount n'étant plus disponibles publiquement pour des raisons de confidentialité, nous proposons de reproduire cette approche dans la section suivante sur des données publiques.

### 2.5.2 Classification de documents : cas des données Newsgroup

#### Description des données

Les expériences sur données simulées de la section 2.4 ont mis en avant la pertinence de l'approche itérative dans un cas simple, et ont permis de comparer les performances de cette approche avec celles des modèles de mélange. Dans cette section, nous proposons d'illustrer l'approche itérative dans un cas plus complexe, sur des données textuelles, à partir d'un jeu de données réelles issues de la base *20Newsgroups*[3].

3. Accès aux données : `https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups`

Les données *20Newsgroups* (Lang, 1995) représentent une base de 19 997 documents, classés dans 20 catégories distinctes. Pour l'expérience, nous avons regroupé ces catégories en 6 classes, comme suggéré par J. Rennie (2008). Chacune des classes est représentée dans la table 2.6, et les effectifs de chaque classe sont présentés dans la table 2.7.

| **Classe 1** | **Classe 2** | **Classe 3** |
|---|---|---|
| comp.graphics | rec.autos | sci.crypt |
| comp.os.ms-windows.misc | rec.motorcycles | sci.electronics |
| comp.sys.ibm.pc.hardware | rec.sport.baseball | sci.med |
| comp.sys.mac.hardware | rec.sport.hockey | sci.space |
| comp.windows.x | | |
| **Classe 4** | **Classe 5** | **Classe 6** |
| misc.forsale | talk.politics.misc | talk.religion.misc |
| | talk.politics.guns | alt.atheism |
| | talk.politics.mideast | soc.religion.christian |

TABLE 2.6 – Les catégories Newsgroup regroupées en 6 classes.

| Classe | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Effectif | 4 984 | 3 983 | 3 983 | 998 | 2 991 | 2 997 |

TABLE 2.7 – Effectif de chaque classe.

Pour réaliser des expériences sur ces données, nous avons été amenés à extraire le contenu de chaque document pour constituer l'ensemble des variables explicatives. Pour cela, nous avons d'abord analysé la structure de chaque fichier, représentée sur la figure 2.8, afin d'extraire la catégorie et le contenu de chaque document.

Une fois ce prétraitement réalisé, nous avons constitué les bases d'apprentissage et de test. La base d'apprentissage a été utilisée pour apprendre un modèle de correspondance entre le contenu des documents et l'ensemble des catégories, tandis que la base de test a été utilisée pour évaluer la performance du modèle.

La base de test a été constituée en sélectionnant de façon aléatoire 900 documents au sein de chaque catégorie, afin de simuler un biais similaire à celui observé dans le cadre du challenge cDiscount entre la distribution de la variable cible sur les bases de test et d'apprentissage. La répartition des classes sur la base de test est donc uniforme. Une telle base peut par exemple être obtenue lors d'une extraction au cours de laquelle $n$ documents sont tirés au sein de chaque catégorie avant de vérifier manuellement leur classe pour constituer la base d'évaluation.

***Explanatory Shift*** La base d'apprentissage a ensuite été constituée en excluant de la base initiale tous les documents ayant été placés dans la base de test. La base

Path : XXX
From : XXX
Newsgroups : alt.atheism
Subject : Re : Political Atheists ?
Message-ID : XXX
Date : 2 Apr 93 19 :05 :57 GMT
References : XXX
Organization : California Institute of Technology, Pasadena
Lines : 11
*Document*

FIGURE 2.8 – Structure des fichiers issus de la base Newsgroup.

| Classe | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Base d'apprentissage | 28.09% | 21.21% | 21.21% | 0.67% | 14.36% | 14.43% |
| Base test | 16.67% | 16.67% | 16.67% | 16.67% | 16.67% | 16.67% |

TABLE 2.8 – Effectif des classes sur les bases d'apprentissage et de test.

d'apprentissage n'a donc pas été construite de sorte que la variable cible sur cette base respecte les propriétés d'une loi uniforme. La table 2.8 décrit la répartition de chaque classe au sein des différents ensembles. Elle illustre clairement la différence de proportion des classes entre les deux bases, qui témoigne d'un shift.

Nous verrons dans la section suivante la pertinence de l'approche itérative pour traiter ce problème de shift ainsi que le gain de performance associé à cette approche par rapport à une approche directe (sans pondération).

### 2.5.3 Résultats

La pertinence de l'approche itérative a été évaluée en réalisant 20 itérations, et l'ensemble des sources utilisées est disponible publiquement [4]. La figure 2.9 illustre l'évolution des effectifs estimés pour chaque classe dans la base de test, telle qu'estimée à chaque itération. Cette figure montre que les itérations permettent d'estimer avec une meilleure précision l'effectif des différentes classes sur la base de test. En particulier, l'estimation la plus pertinente de la distribution de la variable cible sur la base de test est obtenue dès les premières itérations, comme en témoigne la figure 2.10, qui donne l'évolution de la distance de Kullback-Leibler entre la distribution estimée et la vraie distribution des classes sur la base de test. Cette figure illustre par ailleurs la rapidité de convergence de l'approche itérative. La meilleure estimation est en effet obtenue dès

4. `https://github.com/demytt/Explanatory-Shift`

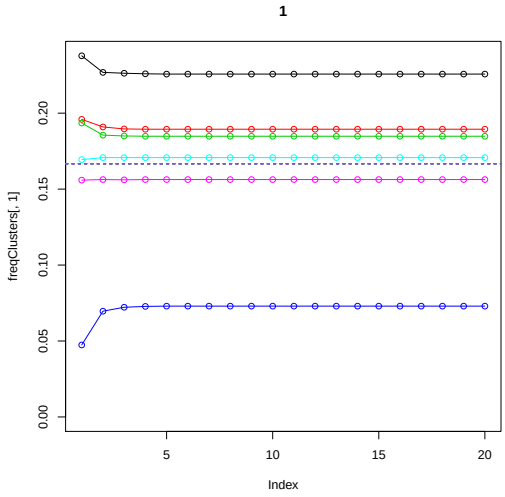



FIGURE 2.9 – Évolution de l'estimation de la répartition des classes sur la base de test en fonction des itérations.

la cinquième itération. Ce résultat est également observé à la figure 2.11, qui représente l'évolution de la performance du modèle obtenu en fonction du nombre d'itérations. Une seule itération est presque suffisante pour avoir le meilleur modèle.

Ces figures montrent que le modèle optimal est obtenu très rapidement et permettent de valider l'approche itérative proposée dans ce chapitre.

## 2.6 Conclusion

Dans ce chapitre, nous avons vu qu'en pratique les hypothèses d'application du *covariate shift* ne sont pas toujours respectées, et nous avons présenté une stratégie pour établir un modèle performant dans le cas où la relation entre les observations et la variable cible diffère entre les bases d'apprentissage et de test.

L'approche proposée peut être vue comme une adaptation de la théorie du *covariate shift* proposée par Shimodaira (2000), sous des hypothèses différentes. Comme dans le cas du *covariate shift*, nous proposons alors d'introduire des poids dans la fonction de perte utilisée afin de mimer la base d'apprentissage. Cependant, nous avons vu que les poids optimaux, sous les hypothèses vues dans le chapitre, s'expriment en fonction de la distribution de la variable cible sur les bases d'apprentissage et de test, contrairement au cas du *covariate shift* où les poids s'expriment en fonction de la distribution des variables explicatives. Pour calculer ces poids, nous avons alors proposé une approche permettant d'estimer la distribution de la variable cible de façon itérative.

Les expériences sur données simulées nous ont permis de mettre en avant les avantages

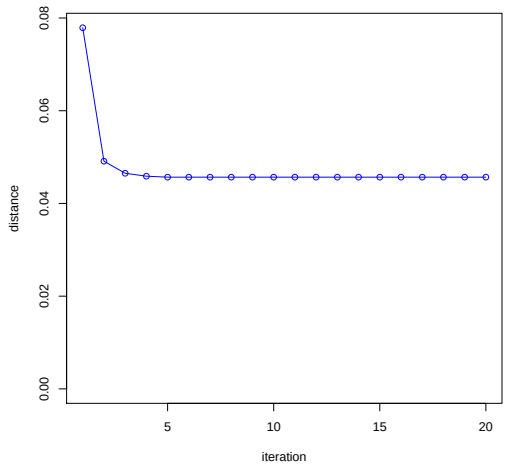



FIGURE 2.10 – Distance de Kullback-Leibler entre la densité réelle des classes sur la base de test, et la densité obtenue par application du modèle, en fonction du nombre d'itérations.

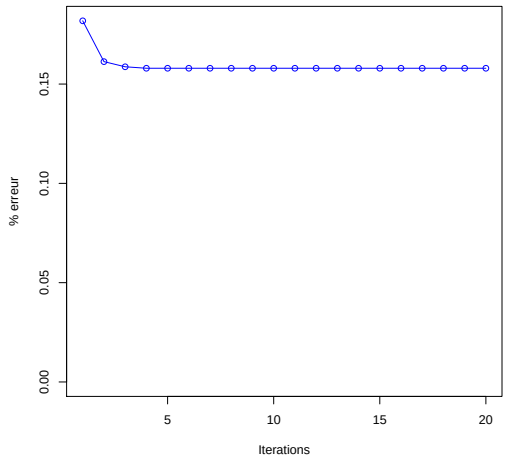



FIGURE 2.11 – Evolution de l'erreur du modèle en fonction du nombre d'itérations.

de l'approche proposée et les limites des modèles obtenus par application de la théorie du *covariate shift* et de l'adaptation de domaines dans le cadre de variables explicatives numériques. Nous avons en effet vu que les hypothèses en présence d'*explanatory shift* ne sont pas compatibles avec les hypothèses sous-jacentes aux autres modèles présentés dans la littérature, ce qui permet de conclure à la complémentarité de l'approche proposée par rapport aux modèles existants. Par ailleurs, les expériences sur données réelles dans le contexte d'un problème de classification montrent la pertinence de l'approche itérative présentée dans ce chapitre, et sa rapidité de convergence, lorsque les variables explicatives sont des données textuelles.

Dans le chapitre suivant, nous proposons d'approfondir le cas où les poids dépendent des données, comme le cadre du *covariate shift* ou de l'*explanatory shift* par exemple. Nous nous intéresserons plus précisément au cas où les poids peuvent diverger en fonction des propriétés des observations, si bien que l'existence de l'estimateur de minimisation du risque empirique n'est plus assurée. Nous verrons alors que l'erreur relative, qui peut être considérée comme une version pondérée de l'erreur absolue, en est un cas particulier. Nous développerons alors en détail le cas de l'erreur relative qui, en pratique, est une fonction de perte particulièrement utilisée pour évaluer la performance d'un modèle prédictif, en raison de sa facilité de compréhension et d'interprétation.

**Contribution scientifique** Ce chapitre a fait l'objet de la communication suivante :

— "Supervised Classification under explanatory shift", Arnaud De Myttenaere, Bénédicte Le Grand, Fabrice Rossi, Apr 2016, Vannes, France. Statlearn 2016

# Chapitre 3

# Minimisation de l'erreur relative moyenne

## 3.1 Introduction

### 3.1.1 Contexte

Afin d'estimer au mieux la qualité d'un modèle par l'évaluation hors-ligne, il est important que la fonction de perte choisie pour calibrer le modèle coïncide avec le critère choisi pour l'évaluation en ligne. En pratique, pour des raisons statistiques ou métier il arrive que le critère diffère selon l'évaluation réalisée. Par exemple, dans le cadre d'un modèle de régression, la fonction de perte utilisée est dans la majorité des cas le critère des moindres carrés (en anglais *Mean Square Error*), ou l'erreur absolue (en anglais *Mean Absolute Error*), respectivement définies pour tout modèle $m$, et pour tout couple $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, avec $d \in \mathbb{N}^\star$ par

$$l_{MSE}(y, m(x)) = (y - m(x))^2 \quad \text{et} \quad l_{MAE}(y, m(x)) = |y - m(x)|.$$

En revanche, deux critères souvent retenus pour l'évaluation en ligne sont l'erreur absolue ou l'erreur relative moyenne (en anglais *Mean Absolute Percentage Error*), notée $l_{MAPE}$, définie pour tout modèle $m$ et couple $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ par

$$l_{MAPE}(y, m(x)) = \frac{|y - m(x)|}{|y|}.$$

Par exemple, dans une compétition de data science organisée au premier trimestre 2014 sur le site Kaggle [1], plateforme internationale permettant aux entreprises d'organiser des compétitions de data science, le fabricant d'ordinateurs ASUS a choisi ce critère pour évaluer la pertinence d'un modèle visant à prédire le nombre de défaillances de certains

1. voir `www.kaggle.com/c/pakdd-cup-2014` pour plus de détails.

composants d'ordinateurs. Pour cela, l'entreprise a mis à disposition des participants un historique de défaillances enregistrées, ainsi que les caractéristiques techniques des composants concernés. Ce problème représente un enjeu stratégique pour le constructeur, afin de déterminer en amont les composants les plus susceptibles d'être défectueux, de les améliorer, d'anticiper d'éventuels retours de produits et d'améliorer la satisfaction de ses clients. Ce critère d'erreur a également été retenu par *GdF Ecometering*[2] afin de mesurer la qualité de modèles visant à prédire la consommation électrique de différents sites français, et par un organisme de cotation des prix de véhicules d'occasion pour mesurer la qualité des modèles dans le cadre d'une compétition sur datascience.net [3], afin de déterminer le prix optimal en fonction de nombreux paramètres tels que la marque de la voiture, le nombre de vitesses, les dimensions, la région de vente, etc.

**Evaluation en ligne ou hors-ligne : des besoins différents**

En pratique, il est cependant rare d'utiliser le même critère pour les évaluations en ligne et hors-ligne d'un modèle prédictif. Ce constat s'explique simplement par le fait que les besoins pour chaque évaluation sont très différents. D'une part, l'évaluation hors-ligne a pour but de calibrer au mieux un modèle à partir d'un échantillon de données constituant la base d'apprentissage, ce qui relève d'un problème d'optimisation. D'autre part, l'évaluation en ligne a pour but de vérifier la pertinence du modèle une fois mis en production, ce qui relève plus d'une évaluation métier. Ainsi, pour des raisons théoriques, le critère des moindres carrés est généralement préféré lors d'une évaluation hors-ligne, tandis que l'erreur relative est souvent choisie lors d'une évaluation en ligne.

L'utilisation de l'erreur quadratique moyenne comme fonction de perte possède de nombreux avantages en termes d'optimisation. La fonction $l_{MSE}$ est en effet définie et continue sur $\mathbb{R}^2$, mais elle est également dérivable en tout point de $\mathbb{R}^2$ et strictement convexe, ce qui permet d'assurer l'existence d'un modèle optimal unique vis-à-vis de cette fonction de perte. En revanche, ce critère s'avère moins pertinent dans le cadre d'une évaluation en ligne, pour des raisons d'interprétation. Par exemple, dans le cas où la valeur de variable cible $y$ associée à une observation $x$ vaut $y = 1000$ et que la prédiction obtenue par le modèle $m$ est $m(x) = 1010$, il est difficile d'interpréter rapidement la qualité de ce modèle en sachant que son erreur est de 100 vis-à-vis des moindres carrés. On préférera en effet dire que l'erreur (absolue) commise par le modèle est de 10, ou encore que le modèle commet une erreur (relative) de 1%. Par ailleurs l'erreur relative permet de comparer l'erreur commise par le modèle $m$ relativement à la valeur de la variable cible : si pour une autre observation $x'$ la valeur prédite par le modèle est de 510 alors que la variable cible est $y' = 500$, l'erreur absolue est toujours de 10 tandis que l'erreur relative est de 2%.

2. `https://www.datascience.net/fr/challenge/16/details`.
3. `https://www.datascience.net/fr/challenge/13/details`.

| Critère | Optimisation | Interprétation | Adapté aux petites valeurs | Adapté aux grandes valeurs |
|---|---|---|---|---|
| Moindres carrés, $l_{MSE}$ | $+$ | $-$ | $+$ | $-$ |
| Erreur absolue, $l_{MAE}$ | $+/-$ | $+/-$ | $+$ | $+$ |
| Erreur relative, $l_{MAPE}$ | $-$ | $+$ | $-$ | $+$ |

TABLE 3.1 – Comparaison de l'intérêt des différents critères selon le type d'évaluation.

Le critère de l'erreur absolue, c'est-à-dire la fonction de perte $l_{MAE}$ possède également de nombreux avantages dans le cadre d'un problème optimisation : continuité et stricte convexité sur $\mathbb{R}^2$ notamment. En revanche, cette fonction n'est pas dérivable en tout couple $(x, y)$ vérifiant $y = m(x)$, ce qui peut rendre son optimisation légèrement plus complexe que celle liée aux moindres carrés.

La table 3.1.1 donne une comparaison de l'intérêt des différentes fonctions de perte selon le type d'évaluation réalisée.

**L'erreur relative comme critère commun aux évaluations en ligne et hors-ligne**

Bien que l'erreur relative s'avère la plus interprétable en pratique, et donc souvent la plus utilisée en contexte industriel, cette dernière possède de nombreux désavantages en termes d'optimisation. D'une part, cette fonction n'est pas strictement convexe ni dérivable sur $\mathbb{R}^2$, ce qui, a priori, ne permet pas d'assurer l'existence d'un modèle optimal. D'autre part, cette fonction n'est pas définie en $y = 0$, ce qui rend délicate son optimisation. Il s'agit d'ailleurs d'une raison pour laquelle cette fonction de perte est couramment utilisée lorsque la quantité à prédire est connue pour être relativement grande.

En pratique, la détermination du modèle optimal vis-à-vis de l'erreur relative peut toutefois être acquise à partir de la détermination du meilleur modèle vis-à-vis de l'erreur absolue. En effet, nous avons rappelé dans les chapitres précédents que, étant donné une fonction de perte $l$ pour laquelle on sait trouver un modèle optimal, il est aisé d'obtenir un modèle optimal vis-à-vis d'une version pondérée de cette dernière, notée $l_\omega$. Or en considérant la pondération $\omega = 1/|y|$, on a pour tout modèle $m$ et pour tout couple $(x, y) \in \mathbb{R} \times \mathbb{R}^\star$ :

$$l_{\omega,MAE}(y, m(x)) = \frac{1}{|y|} \cdot |y - m(x)| = l_{MAPE}(y, m(x)).$$

L'erreur relative moyenne peut donc être considérée comme une version pondérée de l'erreur absolue, ce qui permet d'obtenir le modèle optimal (sous réserve de son existence) vis-à-vis de l'erreur relative en pratique, dès lors qu'on dispose d'un algorithme minimisant l'erreur absolue. De cette façon, il est possible d'utiliser le critère de l'erreur relative dans le cadre d'une évaluation hors-ligne et en ligne, et ainsi avoir une évaluation

en ligne qui est en adéquation avec l'évaluation de la qualité du modèle réalisée une fois le modèle en production.

Dans ce chapitre, nous nous intéressons au cas de l'erreur relative moyenne, comme une alternative aux moindres carrés et à l'erreur absolue afin d'assurer la cohérence entre la fonction de perte choisie pour les évaluations hors-ligne et en ligne.

La résolution pratique de la minimisation de l'erreur relative pouvant être traitée en considérant les fonctions de perte pondérées, nous nous intéresserons dans ce chapitre essentiellement aux considérations théoriques liées à la minimisation de l'erreur relative moyenne.

### 3.1.2 Notations et cadre général

Comme dans les chapitres précédents, nous utilisons les notations classiques du contexte de régression (Neter *et al.*, 1996), où les données sont entièrement décrites par une paire de variables aléatoires $Z = (X, Y)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$. Notre objectif est de trouver un modèle approchant les observations, qui soit une fonction mesurable $g$ de $\mathbb{R}^d$ dans $\mathbb{R}$ telle que $g(X)$ est "proche" de $Y$. Comme nous l'avons évoqué dans la section 3.1.1, dans le cadre classique des régressions la proximité entre $g(X)$ et $Y$ est mesurée par la fonction de perte $L_2$, aussi appelée erreur des moindres carrés, notée $L_{MSE}$ (*Mean Square Error*) et définie par

$$L_{MSE}(g) = L_2(g) = \mathbb{E}(g(X) - Y)^2. \tag{3.1}$$

Soit $m$ la fonction de régression, définie sur $\mathbb{R}^d$ et à valeurs dans $\mathbb{R}$ par

$$m(x) = \mathbb{E}(Y|X = x). \tag{3.2}$$

Il est alors connu (voir par exemple (Gyorfi *et al.*, 2002)) que la fonction de régression ainsi définie est le meilleur estimateur dans le cas des moindres carrés, dans le sens où il s'agit de la fonction minimisant $L_2(g)$ sur l'ensemble des fonctions mesurables $g$ de $\mathbb{R}^d$ dans $\mathbb{R}$.

Plus généralement, la qualité d'un modèle est mesurée via une fonction de perte $l$, de $\mathbb{R}^2$ dans $\mathbb{R}^+$. Alors l'erreur ponctuelle d'un modèle $g$ en $X$ est donnée par $l(g(X), Y)$ et le **risque** d'un modèle $g$ vis-à-vis d'une fonction de perte $l$ est défini par

$$L_l(g) = \mathbb{E}(l(g(X), Y)). \tag{3.3}$$

Par exemple, dans le cas des moindres carrés, la fonction de perte, $l_2 = l_{MSE}$, est définie par $l_2(p, y) = (p - y)^2$ pour tout $(p, y) \in \mathbb{R}^2$. Cela conduit au risque $L_{MSE}$ noté $L_{l_2}(g) = L_{MSE}(g)$. Dans le cas où l'on dispose de $n$ observations $(X_i, Y_i)_{i=1,\dots,n}$ indépendantes et identiquement distribuées selon la loi de $(X, Y)$, avec $n \in \mathbb{N}$, on appelle **risque empirique** la quantité définie par

$$\widehat{L}_l(g)_n = \frac{1}{n} \sum_{i=1}^{n} l(g(x_i), y_i).$$

Le **risque optimal** est l'infimum de $L_l$ sur l'ensemble des fonctions mesurables, c'est-à-dire

$$L_l^* = \inf_{g \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} L_l(g), \tag{3.4}$$

où $\mathcal{M}(\mathbb{R}^d, \mathbb{R})$ représente l'ensemble des fonctions mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$. Comme rappelé précédemment, nous avons

$$L_{MSE}^* = L_2^* = L_{l_2}^* = \mathbb{E}(m(X) - Y)^2 = \mathbb{E}\left[\left(\mathbb{E}\left(Y|X\right) - Y\right)^2\right].$$

Comme mentionné en introduction, nous nous intéresserons dans ce chapitre au cas de l'erreur relative moyenne, que nous noterons MAPE (*Mean Absolute Percentage Error*). Dans ce cas, la fonction de perte est définie pour tout $(p, y) \in \mathbb{R}^2$ par

$$l_{MAPE}(p, y) = \frac{|p - y|}{|y|}, \tag{3.5}$$

avec les conventions que pour tout $a \in \mathbb{R}^\star$, $\frac{a}{0} = \infty$ et $\frac{0}{0} = 1$. Alors le risque MAPE d'un modèle $g$ est donné par

$$L_{MAPE}(g) = L_{l_{MAPE}}(g) = \mathbb{E}\left(\frac{|g(X) - Y|}{|Y|}\right). \tag{3.6}$$

Si les études théoriques de l'existence et de la convergence de l'estimateur de minimisation du risque empirique s'inscrivent dans un contexte nouveau, nous verrons que l'utilisation pratique des régressions MAPE peut être vue comme un cas particulier des régressions médianes. Ainsi, nous comparerons régulièrement nos résultats avec ceux obtenus dans le cas des régressions médianes où la fonction de perte est donnée pour tout $(p, y) \in \mathbb{R}^2$ par $l_{MAE}(p, y) = |p - y|$ et le risque, noté $l_{MAE}$ (*Mean Absolute Error*) est défini par

$$L_{MAE}(g) = L_{l_{MAE}}(g) = \mathbb{E}(|g(X) - Y|). \tag{3.7}$$

Les conséquences pratiques et théoriques de l'utilisation de cette fonction de perte sur le modèle obtenu ont déjà été largement étudiées dans la littérature, puisque l'erreur absolue conduit à la régression médiane, qui est un cas particulier des régressions quantiles (Koenker and Bassett Jr, 1978).

### 3.1.3 Problèmes théoriques

Avant de traiter la minimisation de la fonction de perte $l_{MAPE}$ d'un point de vue pratique et de procéder à des expériences sur données simulées ou réelles (voir section 3.6), nous proposons d'étudier dans ce chapitre les propriétés théoriques de l'estimateur obtenu par minimisation du risque empirique MAPE. Les axes théoriques que nous étudierons s'articulent autour de deux problématiques.

Dans un premier temps nous étudierons l'existence du risque MAPE, c'est-à-dire sous quelles conditions sur le modèle $g$ et les variables aléatoires $X$ et $Y$ la quantité $\mathbb{E}\left(l(g(X), Y)\right)$ est bien définie. Cette étude est nécessaire afin de déterminer sous quelles hypothèses sur les observations l'utilisation du critère MAPE pour l'évaluation d'un modèle est possible.

Dans un second temps, nous nous intéresserons à la convergence, aussi appelée consistance, de l'estimateur obtenu par minimisation du risque MAPE. Cette étude théorique est nécessaire afin de déterminer sous quelles hypothèses sur les observations l'estimateur obtenu par minimisation du risque empirique MAPE est pertinent, dans le sens où il coïncide avec l'estimateur optimal dans le cas où le nombre d'observations à disposition tend vers l'infini.

**Existence** L'étude de l'existence du risque MAPE consiste à déterminer sous quelles conditions sur un modèle $g$ et les variables aléatoires $X$ et $Y$ la quantité $\mathbb{E}\left[\left|\frac{Y-g(X)}{Y}\right|\right]$ est bien définie.

En comparaison avec l'erreur absolue (*Mean Absolute Error*, MAE) et le critère des moindres carrés (*Mean Square Error*, MSE), la MAPE possède plusieurs particularités. On peut d'abord noter l'asymétrie entre $g(X)$ et $Y$, contrairement aux deux autres fonctions de perte. En particulier, la MAPE est très sensible à des erreurs sur les petites valeurs de $Y$, et peu sensible à des erreurs sur les grandes valeurs. Ce phénomène s'illustre très bien en se plaçant dans le cas univarié, où $Y$ suit une loi uniforme sur $[a; a+1]$, avec $a > 0$, avec l'objectif de trouver le modèle optimal vis-à-vis de la MAPE sans variable explicative. Dans ce cas, le problème consiste à trouver le réel $m$ minimisant le risque donné par

$$L_{MAPE}(m) = \int_a^{a+1} \left|\frac{y-m}{y}\right| dy$$

et on peut montrer que le modèle $m^\star_{MAPE}$ minimisant ce risque est

$$m^\star_{MAPE}(a) = \sqrt{a(a+1)}.$$

En revanche, dans le cas de l'erreur absolue (MAE) l'estimateur optimal correspond à la médiane de la variable aléatoire $Y$, et on a donc $m^\star_{MAE}(a) = a + \frac{1}{2}$. Par conséquent, dans le cas où $a = 1$ et $Y$ suit une loi uniforme sur $[1; 2]$, l'estimateur optimal dépend du critère utilisé et on a $m^\star_{MAPE}(a) = \sqrt{2}$, tandis que $m^\star_{MAE}(a) = 1.5$

Alors, les erreurs MAE et MAPE en fonction de $y$ sont données par

$$l_{MAE}(y, m^\star_{MAE}(a)) = \left|a + \frac{1}{2} - y\right| \text{ et } l_{MAPE}(y, m^\star_{MAPE}(a)) = \left|\frac{\sqrt{a(a+1)} - y}{y}\right|.$$

Ces deux fonctions de perte sont représentées à la figure 3.1, pour $a = 0.1$. Cette figure illustre bien l'asymétrie de la fonction de perte MAPE, ainsi que la divergence de

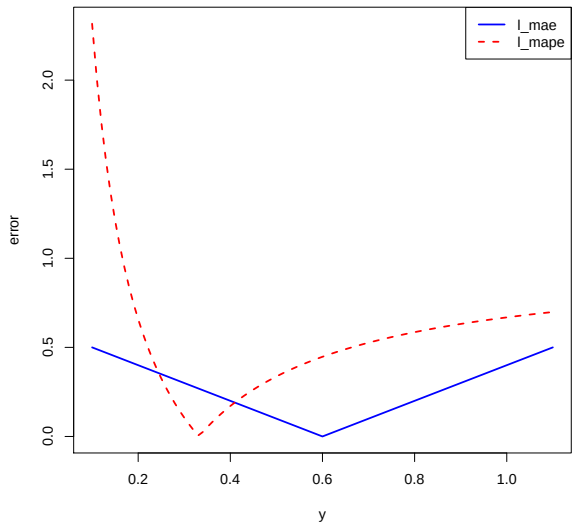



FIGURE 3.1 – Comparaison des comportements lorsque $y \to 0$ pour la MAE et la MAPE.

l'erreur pour des petites valeurs de $y$, contrairement au cas de la MAE. Dans la partie 3.2 nous verrons que, du fait de la divergence en 0, l'existence de $m^{\star}_{MAPE}$ ne peut être assurée que sous certaines hypothèses. En particulier, l'existence n'est pas assurée pour $a = 0$ (c'est pourquoi nous avons choisi une valeur de $a$ strictement positive, proche de zéro).

Les fonctions de perte MAE et MAPE ont par ailleurs des comportements différents lorsque $y$ tend vers 0 ou vers l'infini. En effet, pour $g(X)$ fixé, on a

$$\lim_{y \to 0} l_{MAPE}(g(X), y) = +\infty,$$

tandis que

$$\lim_{y \to 0} l_{MAE}(g(X), y) = |g(X)| < \infty \quad \text{et} \quad \lim_{y \to 0} l_{MSE}(g(X), y) = g(X)^2 < \infty.$$

En revanche,

$$\lim_{y \to \infty} l_{MAPE}(g(X), y) = 1,$$

tandis que

$$\lim_{y \to \infty} l_{MAE}(g(X), y) = +\infty \quad \text{et} \quad \lim_{y \to \infty} l_{MSE}(g(X), y) = +\infty.$$

Nous verrons dans ce chapitre que le rôle joué par $y$ en l'infini dans le cas de la MAE est très similaire à celui joué par $y$ en 0 dans le cas de la MAPE. Plus précisément, nous montrerons que, sous des hypothèses assez souples sur $Y$, il est possible d'assurer l'existence de la fonction de perte MAPE.

**Consistance** Une fois l'existence assurée, nous nous intéresserons à la propriété de consistance de la minimisation du risque empirique MAPE. Plus précisément, nous verrons sous quelles conditions l'estimation du risque optimal peut se faire en s'appuyant sur la minimisation du risque empirique

Pour rappel, étant donné $(X_i, Y_i)_{i=1,\dots,n}$ un échantillon de $n$ copies indépendantes et identiquement distribuées d'un couple de variables aléatoires $(X, Y)$ et $\mathcal{G}$ une famille de modèles, l'estimateur $\widehat{g}_{l,n}$ obtenu par minimisation du risque empirique est défini par

$$\widehat{g}_{l,n} = \arg\min_{g \in G} \widehat{L}_l(g)_n.$$

Pour obtenir la consistance de cet estimateur, il est nécessaire de contrôler les écarts entre le risque empirique et le risque sur une classe de fonctions, par exemple en obtenant que

$$\forall \epsilon > 0, \mathbb{P}\left\{ \lim_{n \to +\infty} \left( \sup_{g \in G} \left| \widehat{L}(\widehat{g}_{l,n})_n - L(g) \right| \right) > \epsilon \right\} = 0.$$

Ce résultat a déjà été établi pour quelques fonctions de perte (MSE et MAE notamment) sous certaines hypothèses (Gyorfi *et al.*, 2002), mais ne peut être généralisé au cas de la MAPE, qui possède des propriétés différentes de celles supposées (voir par exemple le lemme 17.6 dans Anthony and Bartlett (1999)). En particulier, nous verrons que la généralisation de ces résultats est rendue difficile d'une part par la non-existence de la MAPE lorsque $y$ prend des valeurs nulles, d'autre part par le fait qu'il s'agisse d'une fonction de perte non lipschitzienne en 0.

Dans ce chapitre, nous étudierons donc les conséquences des propriétés spécifiques à la MAPE sur la généralisation de la démonstration de la consistance de la méthode de minimisation du risque empirique.

### 3.1.4 Plan du chapitre

Dans la section 3.2, nous nous intéresserons à l'existence du risque MAPE. Nous évoquerons dans un premier temps les cas particuliers d'une variable aléatoire discrète, puis continue, avant de généraliser l'existence de la MAPE sous certaines hypothèses. Ensuite, nous verrons que l'unicité de la solution optimale n'est pas toujours acquise, et nous définirons donc une convention pour le choix de l'optimum.

La section 3.3 sera consacrée aux effets de l'utilisation de la MAPE sur le contrôle de la complexité. Nous verrons alors le rôle joué par les nombres de couverture sur le contrôle de l'estimateur obtenu par minimisation du risque empirique. Dans la section 3.4, nous détaillerons le cas de la dimension de Vapnik-Chervonenkis.

Ensuite, nous montrerons dans la section 3.5 la consistance de la méthode de minimisation du risque empirique vis-à-vis de la MAPE.

Enfin, la section 3.6 sera consacrée à des applications. Nous verrons tout d'abord une application sur un jeu de données public et déjà largement utilisé dans la littérature,

avant de développer une application dans un contexte industriel, sur des données réelles extraites du réseau social professionnel Viadeo.

## 3.2 Existence de la fonction de régression MAPE

Dans cette partie, nous étudierons l'existence d'un estimateur optimal vis-à-vis de la MAPE. Plus précisément, nous analyserons l'existence d'une fonction $m_{MAPE}$ minimisant le risque, c'est-à-dire telle que pour tout modèle $g$, $L_{MAPE}(g) \geq L_{MAPE}(m_{MAPE})$.

De façon évidente, pour tout couple de variables aléatoires $(X, Y)$ sur $\mathbb{R}^d \times \mathbb{R}$ et pour toute fonction $g$ définie sur $\mathbb{R}^d$ et à valeurs dans $\mathbb{R}$ nous avons

$$L_{MAPE}(g) = \mathbb{E}\left[\mathbb{E}\left(\frac{|g(X) - Y|}{|Y|}\middle| X\right)\right].$$

Par conséquent, une stratégie naturelle pour montrer l'existence d'une fonction $m_{MAPE}$ minimisant le risque MAPE consiste à considérer l'approximation ponctuelle. En d'autres termes, nous cherchons à résoudre, si possible, le problème d'optimisation

$$\forall x \in \mathbb{R}^d, \quad m_{MAPE}(x) = \arg\min_{m \in \mathbb{R}} \mathbb{E}\left(\frac{|m - Y|}{|Y|}\middle| X = x\right). \tag{3.8}$$

Afin de simplifier l'analyse, nous étudierons le problème d'optimisation suivant :

$$\min_{m \in \mathbb{R}} \mathbb{E}\left(\frac{|m - T|}{|T|}\right), \tag{3.9}$$

où $T$ est une variable aléatoire réelle. La quantité $\mathbb{E}\left(\frac{|m-T|}{|T|}\right)$ sera notée $J(m)$ dans la suite. En étudiant sous quelles conditions le problème (3.9) admet une solution pour toute variable aléatoire $T$, nous montrerons alors que pour tout $x \in \mathbb{R}^d$ le problème (3.8) admet une solution. Ainsi, nous assurerons l'existence d'une solution ponctuellement optimale, ce qui suffit à assurer l'existence de $m_{MAPE}$. La fonction $m_{MAPE}$ obtenue et définie sur $\mathbb{R}^d$ constituera alors le modèle minimisant l'erreur empirique MAPE, et nous l'appellerons régression MAPE.

En fonction de la distribution de la variable aléatoire $T$ et de la valeur de $m$, $J(m)$ n'est pas toujours une valeur finie, sauf dans le cas où $m = 0$. En effet, pour toute variable aléatoire $T$, $J(0) = 1$ d'après les conventions précédemment définies. Les deux exemples suivants illustrent, sur les cas particuliers des lois normales et uniformes, deux situations dans lesquelles l'existence de la MAPE n'est pas assurée.

**Cas gaussien** Considérons par exemple le cas où $T$ est une variable aléatoire gaussienne, de moyenne $\mu$ et de variance $\sigma^2$, et $m$ un réel strictement positif. Alors

$$J(m) = \int_{\mathbb{R}} \frac{|m-t|}{|t|} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt.$$

Or la fonction $g : t \mapsto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$ est continue et strictement positive sur $\mathbb{R}$. Donc il existe $\alpha > 0$ tel que pour tout $t \in [0; \frac{m}{2}]$, $g(t) > \alpha$.

Alors, pour tout $\epsilon > 0$, comme

$$J(m) > \int_{\epsilon}^{\frac{m}{2}} \frac{|m-t|}{|t|} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt,$$

On en déduit que

$$J(m) > \alpha \int_{\epsilon}^{\frac{m}{2}} \frac{|m-t|}{|t|} dt.$$

Ainsi,

$$J(m) > \underbrace{\alpha m \int_{\epsilon}^{\frac{m}{2}} \frac{1}{t} dt}_{\substack{\to \\ \epsilon \to 0}} \quad - \quad \underbrace{\alpha \cdot \frac{m}{2}}_{\text{valeur finie}} .$$

Cet exemple prouve que, sous l'hypothèse gaussienne, la fonction de régression MAPE n'existe pas puisque le risque n'est pas défini pour tout modèle $m$ non nul.

**Cas uniforme** Considérons ici le cas où $T$ est une variable aléatoire uniforme sur $[-1, 1]$. Alors :

$$J(m) = \frac{1}{2} \int_{-1}^{1} \frac{|m-t|}{|t|} dt.$$

De façon évidente, on a pour tout $\epsilon \in ]0; 1]$

$$\frac{|m-t|}{|t|} \geq \frac{|m-t|}{|t|} \mathbb{1}_{|t|>\epsilon},$$

et la fonction $g_\epsilon \mapsto \frac{|m-t|}{|t|} \mathbb{1}_{|t|>\epsilon}$ est intégrable sur $[-1; 1]$ en tant que fonction positive et majorée sur $[-1; 1]$.

Alors, pour tout $\epsilon \in ]0; 1]$ on a

$$J(m) \geq \frac{1}{2} \int_{-1}^{-\epsilon} \frac{|m-t|}{|t|} dt + \frac{1}{2} \int_{\epsilon}^{1} \frac{|m-t|}{|t|} dt.$$

Considérons le cas où $m \in ]0, 1]$ (le cas $m \in [-1; 0[$ se traitant de façon analogue). Alors, pour tout $\epsilon \in ]0; m]$

$$J(m) \geq \int_{\epsilon}^{1} \frac{|m-t|}{|t|} dt.$$

Or $m \in ]0; 1]$ donc

$$\int_{\epsilon}^{1} \frac{|m-t|}{|t|} dt \geq \int_{\epsilon}^{m} \frac{m-t}{t} dt$$

$$\geq \underbrace{m \log \frac{m}{\epsilon} - (m - \epsilon)}_{\substack{\to \\ \epsilon \to 0}} .$$

Et par conséquent $J(m) = +\infty$. De façon similaire, on montre que pour tout $m > 1$ ou $m < -1$ on a $J(m) = +\infty$.

D'après ces deux exemples, il semblerait donc que dès lors que $T$ peut prendre des valeurs aussi proches que possible de 0, alors $J(m) = \infty$ si $m \neq 0$. Dans ce cas, l'estimateur optimal vis-à-vis de la MAPE est $m_{MAPE} = 0$.

L'objectif des sections suivantes est d'étudier sous quelles conditions l'existence de la fonction de régression MAPE est assurée. D'abord, nous étudierons le cas particulier où $Y$ est une variable aléatoire discrète (section 3.2.1). Ensuite, nous traiterons le cas où $Y$ est une variable aléatoire continue (section 3.2.2), avant d'aborder le cas général (section 3.2.3). Enfin, nous aborderons dans la section 3.2.4 la question de l'unicité de l'optimum.

### 3.2.1 Cas discret

Dans le cas où $Y$ est une variable aléatoire discrète sur $\mathbb{R}$ prenant un nombre fini de valeurs distinctes, on peut montrer que la fonction définie par $J : m \mapsto \mathbb{E}\left[\left|\frac{Y-m}{Y}\right|\right]$ admet un minimum global sur $\mathbb{R}$. Ce résultat est décrit par la proposition 6.

**Proposition 6.** *Soit $n \in \mathbb{N}^*$ un entier naturel strictement positif et $Y$ une variable aléatoire réelle prenant ses valeurs dans l'ensemble $\{y_1, \ldots, y_n\}$, avec $y_i \in \mathbb{R}^\star$. On note $p_i = \mathbb{P}(Y = y_i)$ pour tout $i \in \{1, \ldots, n\}$.*

*Alors la fonction $J : m \mapsto \mathbb{E}\left[\left|\frac{Y-m}{Y}\right|\right]$ admet un minimum global sur $\mathbb{R}$, qui est atteint en un élément de $\{y_1, \ldots, y_n\}$.*

*Démonstration.* Par définition de $J$, on a pour tout $m \in \mathbb{R}$

$$J(m) = \sum_{i=1}^{n} P(Y = y_i) \left|\frac{y_i - m}{y_i}\right| .$$

Or pour tout $a \in \mathbb{R}^\star$, la fonction $m \mapsto \left|\frac{a-m}{a}\right|$ est convexe sur $\mathbb{R}$, et $P(Y = y_i) \geq 0$ pour tout $i \in \{1, \ldots, n\}$. En tant que combinaison linéaire à coefficients positifs de fonctions convexes définies sur $\mathbb{R}$, la fonction $J$ est donc convexe sur $\mathbb{R}$, et par conséquent continue.

Sans perte de généralité, on peut supposer que $y_i < y_{i+1}$ pour tout $i \in \{1, \ldots, n-1\}$.

Soit $m \in [y_1; y_n]$. Alors il existe $j_m \in \{1, \ldots, n-1\}$ tel que $y_{j_m} \leq m \leq y_{j_m+1}$, et

$$J(m) = \sum_{i=1}^{j_m} P(Y = y_i) \frac{m - y_i}{|y_i|} + \sum_{i=j_m+1}^{n} P(Y = y_i) \frac{y_i - m}{|y_i|}.$$

Ainsi, $J(m)$ est de la forme

$$J(m) = a_j \cdot m + b_j$$

avec

$$a_j = \sum_{i=1}^{j_m} \frac{P(Y = y_i)}{|y_i|} - \sum_{i=j_m+1}^{n} \frac{P(Y = y_i)}{|y_i|} \quad \text{et} \quad b_j \in \mathbb{R}$$

et donc, en tant que fonction affine sur $[y_{j_m}, y_{j_m+1}]$,

$$\forall m \in [y_{j_m}, y_{j_m+1}], J(m) \geq \min(J(y_{j_m}), J(y_{j_m+1})). \tag{3.10}$$

Par ailleurs, si $m < y_1$ on a $J(m) = \sum_{i=1}^{n} \frac{y_i - m}{|y_i|}$ et donc $\lim_{m \to -\infty} J(m) = +\infty$ et

$$\forall m \in ]-\infty, y_1], J(m) \geq J(y_1). \tag{3.11}$$

De même, si $m > y_n$ on a $J(m) = \sum_{i=1}^{n} \frac{m - y_i}{|y_i|}$ et donc $\lim_{m \to +\infty} J(m) = +\infty$ et

$$\forall m \in [y_n, +\infty[, J(m) \geq J(y_n). \tag{3.12}$$

La fonction $J$ est donc continue et affine par morceaux sur $\mathbb{R}$ (voir par exemple la figure 3.3). De plus, $\lim_{|m| \to +\infty} J(m) = +\infty$ donc la fonction $J$ est convexe et coercive et admet un minimum sur $\mathbb{R}$.

Enfin, d'après les équations 3.10, 3.11 et 3.12, on a

$$\forall m \in \mathbb{R}, J(m) \geq \min\left(J(y_1), \ldots, J(y_n)\right).$$

La fonction $J$ atteint donc son minimum en un de ses points de cassure $y_1, \ldots, y_n$. □

### 3.2.2 Cas continu

Dans le cas où $Y$ est une variable aléatoire réelle absolument continue (VARAC) de densité $\phi$, nous pouvons montrer que l'existence de la fonction de régression MAPE dépend de l'intégrabilité en 0 de la fonction $y \mapsto \frac{1}{|y|}\phi(y)$. Plus précisément, nous avons la proposition suivante :

**Proposition 7.** *Soit $Y$ une variable aléatoire réelle absolument continue (VARAC) de densité $\phi$.*

*Alors les trois assertions suivantes sont équivalentes :*

*i)* $\exists m \neq 0$ *tel que* $J(m) < \infty$,

*ii)* $\forall m \in \mathbb{R}, J(m) < \infty$

*iii)* $\int_{-1}^{1} \frac{1}{|y|}\phi(y)dy < \infty$

*Démonstration.* $\boxed{i \Rightarrow iii}$ Supposons qu'il existe un réel $m \neq 0$ tel que $J(m) < \infty$, et montrons que $\int_{-1}^{1} \frac{1}{|y|}\phi(y)dy < \infty$.

Notons $j_m(y) = \left|\frac{y-m}{y}\right| \phi(y)$ pour tout $y \in \mathbb{R}$.

Comme pour tout $\alpha > 0$ on a

$$\forall y \in \mathbb{R}, 0 \leq \mathbb{1}_{y\in[-\alpha;\alpha]} \cdot j_m(y) \leq j_m(y)$$

et que la fonction $j_m$ est intégrable sur $\mathbb{R}$ par hypothèse, on en déduit que

$$\int_{\mathbb{R}} \mathbb{1}_{y\in[-\alpha;\alpha]} j_m(y) < \infty. \tag{3.13}$$

De plus, comme $m \neq 0$, il existe un voisinage de 0 sur lequel la fonction $y \mapsto |y-m|$ est strictement positive. Ainsi, il existe $a \in ]0;1[$ et $c_a > 0$ tels que $|y-m| \geq c_a$ pour tout $y \in [-a, a]$.

En appliquant l'équation 3.13 pour $\alpha = a$ on obtient l'intégrabilité de la fonction $j_m$ sur $[-a;a]$ :

$$\int_{-a}^{a} j_m(y) < \infty. \tag{3.14}$$

Or par définition de $a$, pour tout $y \in [-a;a]$ on a $j_m(y) \geq \frac{c_a}{|y|}\phi(y) \geq 0$. Donc la fonction $y \mapsto \frac{c_a}{|y|}\phi(y)$ est intégrable sur $[-a;a]$.

Par ailleurs,

$$\forall y \in [-1,-a], \quad 0 \leq \frac{c_a}{|y|}\phi(y) \leq \underbrace{\frac{c_a}{|a|} \cdot \phi(y)}_{\text{intégrable sur } [-1;-a]} \quad ,$$

donc la fonction $y \mapsto \frac{c_a}{|y|}\phi(y)$ est intégrable sur $[-1;-a]$. De même, on montre que la fonction $y \mapsto \frac{c_a}{|y|}\phi(y)$ est intégrable sur [a ;1].

Par la relation de Chasles, on en déduit que $y \mapsto \frac{c_a}{|y|}\phi(y)$ est intégrable sur $[-1;1]$, avec $c_a > 0$. D'où

$$\int_{-1}^{1} \frac{1}{|y|}\phi(y)dy < \infty.$$

$\boxed{iii \Rightarrow ii}$ Soit $m \neq 0$. Supposons que $\int_{-1}^{1} \left|\frac{1}{y}\right| \phi(y)dy < \infty$ et montrons que $J(m) < \infty$.

Comme $\phi$ est une fonction positive sur $\mathbb{R}$, d'après l'inégalité triangulaire on a

$$\forall y \in [-1;1], \quad 0 \leq \left|\frac{y-m}{y}\right| \phi(y) \leq \phi(y) + \left|\frac{m}{y}\right| \phi(y).$$

Or $\phi$ est intégrable sur $[-1;1]$ (car il s'agit d'une fonction positive intégrable sur $\mathbb{R}$), et $\int_{-1}^{1} \left|\frac{m}{y}\right| \phi(y)dy \leq \infty$ par hypothèse. Donc la fonction $y \mapsto \phi(y) + \left|\frac{m}{y}\right| \phi(y)$ est intégrable sur $[-1;1]$, ce qui permet d'assurer que la fonction $y \mapsto \left|\frac{y-m}{y}\right| \phi(y)$ est intégrable sur $[-1;1]$ :

$$\int_{-1}^{1} \left|\frac{y-m}{y}\right| \phi(y)dy < \infty. \tag{3.15}$$

Par ailleurs, comme la fonction $y \mapsto \left|\frac{y-m}{y}\right|$ est continue sur $]-\infty;-1] \cap [1;+\infty[$ et telle que $\lim_{|y|\to\infty} \left|\frac{y-m}{y}\right| = 1$,

$$\exists M \in \mathbb{R}^+ \text{ tel que } \forall y \in ]-\infty;-1] \cap [1;+\infty[, \left|\frac{y-m}{y}\right| < M.$$

Et comme la fonction $\phi$ est positive, on en déduit que

$$\exists M \in \mathbb{R}^+ \text{ tel que } \forall y \in ]-\infty;-1] \cap [1;+\infty[, 0 \leq \left|\frac{y-m}{y}\right| \phi(y) < M\phi(y).$$

Or la fonction $y \mapsto M\phi(y)$ est intégrable sur $]-\infty;-1] \cap [1;+\infty[$, ce qui assure l'intégrabilité de la fonction $y \mapsto \left|\frac{y-m}{y}\right| \phi(y)$ sur $]-\infty;-1] \cap [1;+\infty[$.

Combiné avec l'équation 3.15, ce résultat permet de conclure que $J(m) < \infty$.

$\boxed{iii \Rightarrow ii}$ Immédiat.

□

Cette proposition assure l'existence du risque MAPE dans le cas où $Y$ est une variable aléatoire réelle absolument continue de densité $\phi$, ayant une décroissance en 0 *plus rapide* que la fonction linéaire $f : y \mapsto y$. Par ailleurs, dans le cas où $Y$ est une variable aléatoire réelle absolument continue, on peut montrer sous certaines hypothèses que le modèle $m_{MAPE}$ minimisant le risque MAPE vérifie $\int_{-\infty}^{m_{MAPE}} \frac{1}{y}\phi(y)dy = \frac{1}{2} \cdot \mathbb{E}\left[\frac{1}{Y}\right]$. Ce résultat fait l'objet de la proposition 8.

**Proposition 8.** *Soit $Y$ une variable aléatoire réelle absolument continue, de densité $\phi$ telle que $\int_{-1}^{1} \frac{1}{|y|}\phi(y)dy < \infty$, et $J(m) = \mathbb{E}\left|\frac{m-Y}{Y}\right|$.*

*Alors pour tout $m \in \mathbb{R}$ la fonction $y \mapsto \frac{1}{y}\phi(y)$ est intégrable sur $]-\infty;m]$. Par ailleurs la fonction $J$ est continue et deux fois dérivable sur $\mathbb{R}$, convexe, et atteint son minimum en un point $m^\star$ vérifiant*

$$\int_{-\infty}^{m^\star} \frac{1}{|y|}\phi(y)dy = \frac{1}{2} \cdot \mathbb{E}\left[\frac{1}{|Y|}\right].$$

*Démonstration.* Soit $Y$ une variable aléatoire réelle absolument continue, de densité $\phi$, et $s(y) = \frac{y}{|y|}$ pour tout $y \neq 0$.

Comme $\phi$ est une densité de probabilité, $\phi$ est intégrable sur tout sous-ensemble de $\mathbb{R}$ et donc en particulier sur tout intervalle $]-\infty;m]$ où $m \in \mathbb{R}$. Nous notons alors $\psi$ la primitive de $\phi$ telle que $\psi(m) = \int_{-\infty}^{m} \phi(y)dy$.

Par ailleurs, pour tout $m < 1$ on a $0 \leq \frac{1}{|y|}\phi(y) \leq \frac{1}{|m|}\phi(y)$ donc la fonction $y \mapsto \frac{1}{|y|}\phi(y)$ est intégrable sur $]-\infty;m]$. Et pour tout $m \geq 1$ on a $0 \leq \frac{1}{|y|}\phi(y) \leq \phi(y)$ pour tout $y \in [1;m]$ donc la fonction $y \mapsto \frac{1}{|y|}\phi(y)$ est intégrable sur $[1;m]$. Et comme par hypothèse on a $\int_{-1}^{1} \frac{1}{|y|}\phi(y)dy < \infty$, en appliquant la relation de Chasles on en déduit donc que

la fonction $y \mapsto \frac{1}{|y|}\phi(y)$ est intégrable sur $]-\infty; m]$ pour tout $m \in \mathbb{R}$. On note alors $F(m) = \int_{-\infty}^{m} \frac{1}{|y|}\phi(y)dy$.

Soit $m \in \mathbb{R}$ et calculons $J(m)$.

D'après la proposition 7 la fonction $y \mapsto \left|\frac{y-m}{y}\right| \phi(y)$ est positive et intégrable sur $\mathbb{R}$, et donc sur $]-\infty; m]$ et $[m; +\infty[$.

**Si** $m \geq 0$**.**

$$\begin{aligned}
J(m) &= \int_{-\infty}^{m} \frac{m-y}{|y|}\phi(y)dy + \int_{m}^{+\infty} \frac{y-m}{|y|}\phi(y)dy, \\
&= m\left(\int_{-\infty}^{m} \frac{1}{|y|}\phi(y)dy - \int_{m}^{+\infty} \frac{1}{|y|}\phi(y)dy\right) - \int_{-\infty}^{m} s(y)\phi(y)dy + \int_{m}^{+\infty} s(y)\phi(y)dy, \\
&= m\left(F(m) - (F(+\infty) - F(m))\right) + \int_{-\infty}^{0} \phi(y)dy - \int_{0}^{m} \phi(y)dy + \int_{m}^{+\infty} \phi(y)dy, \\
&= m\left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + \psi(0) - (\psi(m) - \psi(0)) + 1 - \psi(m), \\
&= m\left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + 1 - 2\psi(m) + 2\psi(0).
\end{aligned}$$

La fonction $J$ est donc dérivable sur $\mathbb{R}^+$ et pour tout $m \in \mathbb{R}^+$

$$\begin{aligned}
J'(m) &= \left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + m\left(\frac{2}{|m|}\phi(m)\right) - 2\phi(m), \\
&= 2F(m) - \mathbb{E}\frac{1}{|Y|}.
\end{aligned}$$

**Si** $m \leq 0$**.**

$$\begin{aligned}
J(m) &= \int_{-\infty}^{m} \frac{m-y}{|y|}\phi(y)dy + \int_{m}^{+\infty} \frac{y-m}{|y|}\phi(y)dy, \\
&= m\left(\int_{-\infty}^{m} \frac{1}{|y|}\phi(y)dy - \int_{m}^{+\infty} \frac{1}{|y|}\phi(y)dy\right) - \int_{-\infty}^{m} s(y)\phi(y)dy + \int_{m}^{+\infty} s(y)\phi(y)dy, \\
&= m\left(F(m) - (F(+\infty) - F(m))\right) + \int_{-\infty}^{m} \phi(y)dy - \int_{m}^{0} \phi(y)dy + \int_{0}^{+\infty} \phi(y)dy, \\
&= m\left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + \psi(m) - (\psi(0) - \psi(m)) + 1 - \psi(0), \\
&= m\left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + 1 + 2\psi(m) - 2\psi(0).
\end{aligned}$$

La fonction $J$ est donc dérivable sur $\mathbb{R}^-$ et pour tout $m \in \mathbb{R}^-$

$$J'(m) = \left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + m\left(\frac{2}{|m|}\phi(m)\right) + 2\phi(m),$$

$$= \quad 2F(m) - \mathbb{E}\frac{1}{|Y|}.$$

De façon générale, par croissance de la fonction $\psi$ sur $\mathbb{R}$ on a pour tout $m \in \mathbb{R}$

$$J(m) = m\left(2F(m) - \mathbb{E}\frac{1}{|Y|}\right) + 1 - 2\left|\psi(m) - \psi(0)\right|.$$

Par ailleurs, la fonction $J$ est dérivable sur $\mathbb{R}$ et pour tout $m \in \mathbb{R}$

$$J'(m) = 2F(m) - \mathbb{E}\frac{1}{|Y|}.$$

Ce résultat prouve d'ailleurs que la fonction $J$ est deux fois dérivable sur $\mathbb{R}$ et pour tout $m \in \mathbb{R}$ on a

$$J''(m) = \frac{2}{|m|}\phi(m).$$

Ainsi, pour tout $m \in \mathbb{R}$ on a $J''(m) \geq 0$ ce qui montre que la fonction $J$ est convexe sur $\mathbb{R}$.

Enfin, $F$ est une fonction continue sur $\mathbb{R}$ telle que $\lim_{m\to-\infty} F(m) = 0$, et $\lim_{m\to+\infty} F(m) = \mathbb{E}\frac{1}{|Y|}$. D'après le théorème des valeurs intermédiaires, il existe donc un réel $m^\star$ tel que $J(m^\star) = \frac{1}{2}\mathbb{E}\frac{1}{|Y|}$. Alors $J'(m^\star) = 0$ et donc par convexité de la fonction $J$ sur $\mathbb{R}$ on a $J(m) \geq J(m^\star)$ pour tout $m \in \mathbb{R}$. Ce résultat montre que la fonction $J$ admet un minimum en un point $m^\star$ tel que

$$F(m^\star) = \frac{1}{2} \cdot \mathbb{E}\frac{1}{|Y|}.$$

□

**Exemple : loi uniforme** Soit $Y$ une variable aléatoire uniforme sur $[a, b]$, avec $a$ et $b$ deux réels strictement positifs tels que $a < b$, et de densité $f$. Alors

$$\forall y \in \mathbb{R}, \quad f(y) = \frac{\mathbb{1}_{y\in[a;b]}}{b-a}.$$

Pour tout $m \in [a; b]$ l'espérance de la MAPE est donnée par

$$\begin{aligned}
J(m) &= \frac{1}{b-a}\int_a^b \left|\frac{y-m}{y}\right| dy, \\
J(m) &= \frac{1}{b-a}\left(\int_a^m \frac{m-y}{y}dy + \int_m^b \frac{y-m}{y}dy\right), \\
J(m) &= \frac{1}{b-a}\left(\int_a^m \left(\frac{m}{y} - 1\right)dy + \int_m^b \left(1 - \frac{m}{y}\right)dy\right),
\end{aligned}$$

$$\begin{aligned} J(m) &= \frac{1}{b-a}\left(m\log\frac{m}{a} - (m-a) + (b-m) - m\log\frac{b}{m}\right), \\ J(m) &= \frac{1}{b-a}\left(2m\log m - 2m + b + a - m\log(ab)\right). \end{aligned}$$

Alors

$$\frac{\partial J}{\partial m} = \frac{1}{b-a}\left(2\log m - \log(ab)\right).$$

Les conditions du premier ordre donnent

$$\frac{\partial J}{\partial m} = 0 \quad \Longleftrightarrow \; m = \sqrt{ab}.$$

Par conséquent, si $Y$ suit une loi uniforme sur $[a;b]$, alors le modèle $m_{MAPE}$ minimisant le risque MAPE est donné par $m_{MAPE} = \sqrt{ab}$.

### 3.2.3 Cas général

Dans cette section, nous étudions l'existence d'un minimum sur $\mathbb{R}$ pour la fonction $J : m \mapsto \mathbb{E}\left[\left|\frac{Y-m}{Y}\right|\right]$ dans le cas général d'une variable aléatoire $Y$ quelconque. La proposition suivante généralise l'existence de l'optimum sous certaines hypothèses :

**Proposition 9.** *Soit $T$ une variable aléatoire réelle absolument continue, et $J$ la fonction définie pour tout réel $m$ par $J(m) = \mathbb{E}\left|\frac{m-T}{T}\right|$. Alors $J(m) < \infty$ pour tout $m$ si, et seulement si*

1. $\mathbb{P}(T = 0) = 0$,
2. *et*

$$\mathbb{E}\left[-\frac{\mathbb{1}_{T\in[-1,0[}}{T}\right] < \infty, \qquad \mathbb{E}\left[\frac{\mathbb{1}_{T\in]0,1]}}{T}\right] < \infty, \tag{3.16}$$

*Si l'une de ces conditions n'est pas vérifiée, alors $J(m) = \infty$ pour tout $m \neq 0$.*

*Démonstration.* Soit $T$ une variable aléatoire réelle absolument continue, et $m \in \mathbb{R}$.

De façon immédiate on a

$$J(m) < \infty \iff \forall \mathcal{E} \subset \mathbb{R}, \quad \mathbb{E}\left[\left|\frac{\mathbb{1}_{T\in\mathcal{E}}}{T}\right|\right] < \infty,$$

et par définition la quantité $J(m)$ est finie si, et seulement si,

$$J(m) = \mathbb{E}\left(\mathbb{1}_{T=0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{1}_{T>0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{1}_{T<0}\frac{|m-T|}{|T|}\right).$$

Si $\mathbb{P}(T = 0) > 0$ alors pour tout $m \neq 0$, $J(m) = \infty$. Considérons le cas où $\mathbb{P}(T = 0) = 0$. Par symétrie, on peut supposer sans perte de généralité que $m > 0$. Alors

$$J(m) = \mathbb{E}\left(\mathbb{1}_{T>0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{1}_{T<0}\frac{|m-T|}{|T|}\right),$$

$$= \mathbb{E}\left(\mathbb{1}_{0<T\leq m}\frac{m-T}{|T|}\right) + \mathbb{E}\left(\mathbb{1}_{m<T}\frac{T-m}{|T|}\right) + \mathbb{E}\left(\mathbb{1}_{T<0}\frac{m-T}{|T|}\right).$$

Par linéarité de l'espérance, la quantité $J(m)$ est finie si, et seulement si,

$$J(m) = \mathbb{P}(T<0) + \mathbb{P}(T>m) - \mathbb{P}(T\in]0,m]) + m\mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,m]} + \mathbb{1}_{T<0} - \mathbb{1}_{T>m}}{|T|}\right).$$

Or $0 \leq m\mathbb{E}\left(\frac{\mathbb{1}_{T>m}}{T}\right) \leq \mathbb{P}(T>m)$ donc la quantité $m\mathbb{E}\left(\frac{\mathbb{1}_{T>m}}{T}\right)$ est finie.

Ceci montre que $J(m)$ est la somme de termes finis et de la quantité $m\mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,m]}+\mathbb{1}_{T<0}}{|T|}\right)$. En tant que somme de quantités positives et par linéarité de l'espérance, on a

$$\exists M\in\mathbb{R} \text{ tel que } \mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,m]} + \mathbb{1}_{T<0}}{|T|}\right) < M$$
$$\iff \exists M_1, M_2 \in \mathbb{R} \text{ tels que } \mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,m]}}{|T|}\right) < M_1 \text{ et } \mathbb{E}\left(\frac{\mathbb{1}_{T<0}}{|T|}\right) < M_2.$$

De même,

$$\exists M'\in\mathbb{R} \text{ tel que } \mathbb{E}\left(\frac{\mathbb{1}_{T<0}}{|T|}\right) < M'$$
$$\iff \exists M_1', M_2' \in \mathbb{R} \text{ tels que } \mathbb{E}\left(\frac{\mathbb{1}_{T\in[-m;0[}}{|T|}\right) < M_1' \text{ et } \mathbb{E}\left(\frac{\mathbb{1}_{T<-m}}{|T|}\right) < M_2'.$$

Or $0 \leq \mathbb{E}\left(\frac{\mathbb{1}_{T<-m}}{|T|}\right) \leq \frac{1}{|m|}P(T<-m)$ donc la quantité $\mathbb{E}\left(\frac{\mathbb{1}_{T<-m}}{|T|}\right)$ est finie.

Finalement,

$$\exists M\in\mathbb{R} \text{ tel que } J(m) < M$$
$$\iff \exists M_1, M_1' \in \mathbb{R} \text{ tels que } \mathbb{E}\left(\frac{\mathbb{1}_{T\in]0;m]}}{|T|}\right) < M_1 \text{ et } \mathbb{E}\left(\frac{\mathbb{1}_{T\in[-m;0[}}{|T|}\right) < M_1'.$$

Or $T$ est une variable aléatoire réelle absolument continue, donc la quantité $\mathbb{E}\left(\frac{\mathbb{1}_{T\in[0;m[}}{|T|}\right)$ (resp. $\mathbb{E}\left(\frac{\mathbb{1}_{T\in[-m;0[}}{|T|}\right)$) est finie si, et seulement si, la quantité $\mathbb{E}\left(\frac{\mathbb{1}_{T\in[0;1[}}{|T|}\right)$ (resp. $\mathbb{E}\left(\frac{\mathbb{1}_{T\in[-1;0[}}{|T|}\right)$) est finie, d'où le résultat. □

Si les conditions de la proposition 9 ne sont pas vérifiées, $J(m)$ est infinie sauf par convention dans le cas où $m=0$ et alors $\arg\min_{m\in\mathbb{R}} J(m) = 0$. En pratique, les conditions introduites à la proposition 9 peuvent être caractérisées de façon équivalente par des conditions sur les séries de terme général $\left(k\mathbb{P}\left(T\in\left]\frac{1}{k+1};\frac{1}{k}\right]\right)\right)_{k\geq 0}$ et $\left(k\mathbb{P}\left(T\in\left]-\frac{1}{k};-\frac{1}{k+1}\right]\right)\right)_{k\geq 0}$. Ce résultat fait l'objet de la proposition 10.

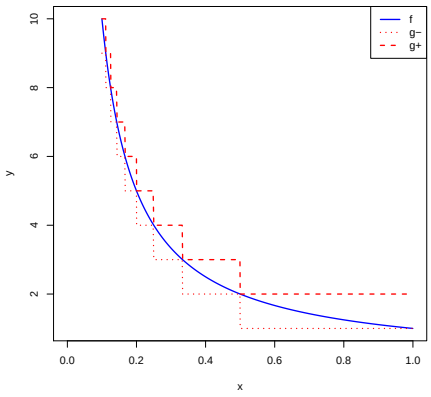



FIGURE 3.2 – Représentation graphique des fonctions $g^-$ et $g^+$.

**Proposition 10.** *Soit $T$ une variable aléatoire réelle absolument continue, et $J$ la fonction définie pour tout réel $m$ par $J(m) = \mathbb{E}\left|\frac{m-T}{T}\right|$. Alors $J(m) < \infty$ pour tout $m$ si, et seulement si*

1. $\mathbb{P}(T = 0) = 0$,
2. *et*

$$\sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left]\frac{1}{k+1}, \frac{1}{k}\right]\right) < \infty, \quad \sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left[-\frac{1}{k}, -\frac{1}{k+1}\right[\right) < \infty. \tag{3.17}$$

*Si l'une de ces conditions n'est pas vérifiée, alors $J(m) = \infty$ pour tout $m \neq 0$.*

*Démonstration.* Introduisons les fonctions suivantes, pour tout $x \in ]0;1]$ et $n \in \mathbb{N}$ :

$$f_k^-(x) = \begin{cases} 0 & \text{if } x \notin ]\frac{1}{k+1}, \frac{1}{k}], \\ k & \text{if } x \in ]\frac{1}{k+1}, \frac{1}{k}], \end{cases} \qquad f_k^+(x) = \begin{cases} 0 & \text{if } x \notin ]\frac{1}{k+1}, \frac{1}{k}], \\ k+1 & \text{if } x \in ]\frac{1}{k+1}, \frac{1}{k}], \end{cases}$$

$$g_n^- = \sum_{k=1}^{n} f_k^-(x), \qquad g_n^+ = \sum_{k=1}^{n} f_k^+(x),$$

$$g^- = \sum_{k=1}^{\infty} f_k^-(x), \qquad g^+ = \sum_{k=1}^{\infty} f_k^+(x).$$

Les fonctions $g^-$ et $g^+$ permettent d'encadrer la fonction $x \mapsto \frac{1}{x}$, comme représenté à la figure 3.2. Il est alors évident que pour tout $x \in ]0, 1]$, $g^-(x) \leq \frac{1}{x} \leq g^+(x)$. De plus,

$$\mathbb{E}(g_n^+(T)) = \sum_{k=1}^{n}(k+1)\mathbb{P}\left(T \in \left]\frac{1}{k+1}, \frac{1}{k}\right]\right),$$

$$\mathbb{E}(g_n^-(T)) = \sum_{k=1}^{n} k\mathbb{P}\left(T \in \left]\frac{1}{k+1}, \frac{1}{k}\right]\right) = \mathbb{E}(g_n^+(T)) - \mathbb{P}\left(T \in \left]\frac{1}{k+1}, 1\right]\right).$$

D'après le théorème de convergence monotone,

$$\mathbb{E}(g^+(T)) = \lim_{n\to\infty} \mathbb{E}(g_n^+(T)).$$

De plus, le lien entre $\mathbb{E}(g_n^-(T))$ et $\mathbb{E}(g_n^+(T))$ montre que, soit les deux quantités $\mathbb{E}(g^+(T))$ et $\mathbb{E}(g^-(T))$ sont finies, soit elles sont toutes les deux infinies. En outre, nous avons

$$\mathbb{E}(g^-(T)) \leq \mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,1]}}{T}\right) \leq \mathbb{E}(g^+(T)),$$

donc $\mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,1]}}{T}\right)$ est finie si, et seulement si, $\mathbb{E}(g^-(T))$ est finie. Par conséquent, une condition nécessaire et suffisante pour que $\mathbb{E}\left(\frac{\mathbb{1}_{T\in]0,1]}}{T}\right)$ soit finie est

$$\sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left]\frac{1}{k+1}, \frac{1}{k}\right]\right) < \infty.$$

Par symétrie, on montre de la même manière que $\mathbb{E}\left(-\frac{\mathbb{1}_{T\in]-1,0]}}{T}\right)$ est finie si, et seulement si,

$$\sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left[-\frac{1}{k}, -\frac{1}{k+1}\right[\right) < \infty.$$

La proposition 9 permet de conclure. □

Montrons maintenant la proposition suivante, qui assure que si les conditions de la proposition 9 sont vérifiées, $J(m)$ admet au moins un minimum global sur $\mathbb{R}$.

**Proposition 11.** *Soit $T$ une variable aléatoire réelle absolument continue et soit $J$ la fonction définie pour tout réel $m$ par $J(m) = \mathbb{E}\left|\frac{m-T}{T}\right|$. Si les hypothèses suivantes sont vérifiées :*

*i)* $\mathbb{P}(T = 0) = 0$,
*ii)* $\mathbb{E}\left[-\frac{\mathbb{1}_{T\in[-1,0[}}{T}\right] < \infty$
*iii)* $\mathbb{E}\left[\frac{\mathbb{1}_{T\in]0,1]}}{T}\right] < \infty$
*Alors $J$ est convexe et admet au moins un minimum global sur $\mathbb{R}$.*

*Démonstration. Convexité.* Remarquons tout d'abord que $J$ est convexe. En effet, pour tout $t \neq 0$, $m \mapsto \frac{|m-t|}{|t|}$ est clairement convexe, et donc par linéarité de l'espérance, $J$ est également convexe, sous réserve que la fonction $J$ soit finie en tout point, ce qui est assuré par la proposition 9.

*Coercivité.* Comme $\mathbb{P}(T = 0) = 0$, il existe deux nombres réels $a$ et $b$ vérifiant $a < b$ et tels que $\mathbb{P}(T \in [a, b]) > 0$, avec $a > 0$ ou $b < 0$. Supposons que $a > 0$ (le cas $b < 0$ se traite de façon symétrique). Alors pour tout $t \in [a, b]$, $\frac{1}{b} \leq \frac{1}{t} \leq \frac{1}{a}$. Par ailleurs, pour tout nombre réel $m > b$ et pour tout $t \in [a, b]$ on a

$$\frac{|m-t|}{|t|} = \frac{m}{t} - 1 \geq \frac{m}{b} - 1.$$

Donc pour tout nombre réel $m > b$ on a

$$\begin{aligned} J(m) &\geq \mathbb{E}\left(\frac{\mathbb{1}_{T\in[a,b]}|m-T|}{|T|}\right) \\ &\geq \left(\frac{m}{b}-1\right)\mathbb{P}(T\in[a,b]), \end{aligned}$$

et par conséquent $\lim_{m\to+\infty} J(m) = +\infty$.

De façon similaire, pour tout $m \in \mathbb{R}$ tel que $m < 0 < a$, et pour $t \in [a, b]$ on a

$$\frac{|m-t|}{|t|} = 1 - \frac{m}{t} \geq 1 - \frac{m}{b},$$

et donc

$$J(m) \geq \left(1 - \frac{m}{b}\right)\mathbb{P}(T \in [a, b]),$$

et par conséquent $\lim_{m\to-\infty} J(m) = +\infty$.

Ainsi, la fonction $J$ est coercive et admet un minimum local, qui est donc un minimum global par convexité. □

De façon générale, pour toute variable aléatoire $T$ nous avons montré qu'il existe un réel $m$ qui est un minimum global de la fonction $J(m) = \mathbb{E}\left(\left|\frac{m-T}{T}\right|\right)$. En revenant à notre problème, cela assure que la fonction de régression MAPE, $m_{MAPE}$, introduite à l'équation 3.8 est bien définie et prend des valeurs finies sur $\mathbb{R}^d$. Par ailleurs, comme $m_{MAPE}$ est ponctuellement optimale, il s'agit également d'un optimum global.

Cependant, comme $J$ n'est pas strictement convexe, le minimum n'est pas nécessairement unique. De façon générale, l'ensemble des minima globaux peut en effet être un intervalle de $\mathbb{R}$. Dans ce cas, on considérera par convention que la valeur moyenne de l'intervalle est la solution optimale.

### 3.2.4 Non-unicité de l'optimum

Afin d'illustrer la non-unicité de l'optimum global, considérons le cas où $T$ est une variable aléatoire discrète sur $\{1, 2, 3\}$, telle que $\mathbb{P}(T = 1) = 0.3$, $\mathbb{P}(T = 2) = 0.4$ et $\mathbb{P}(T = 3) = 0.3$. Alors le risque MAPE est donné pour tout $m \in \mathbb{R}$ par

$$J(m) = 0.3 \times |m-1| + 0.4 \times \left|\frac{m-2}{2}\right| + 0.3 \times \left|\frac{m-3}{3}\right|$$

et la figure 3.3 illustre bien qu'il existe une infinité de valeurs minimisant $J(m)$, à savoir tous les nombres réels compris dans l'intervalle $[1; 2]$. En effet, pour tout $m \in [1, 2]$, $J$ devient

$$J(m) = 0.3 \times (m-1) + 0.4 \times \frac{2-m}{2} + 0.3 \times \frac{3-m}{3},$$

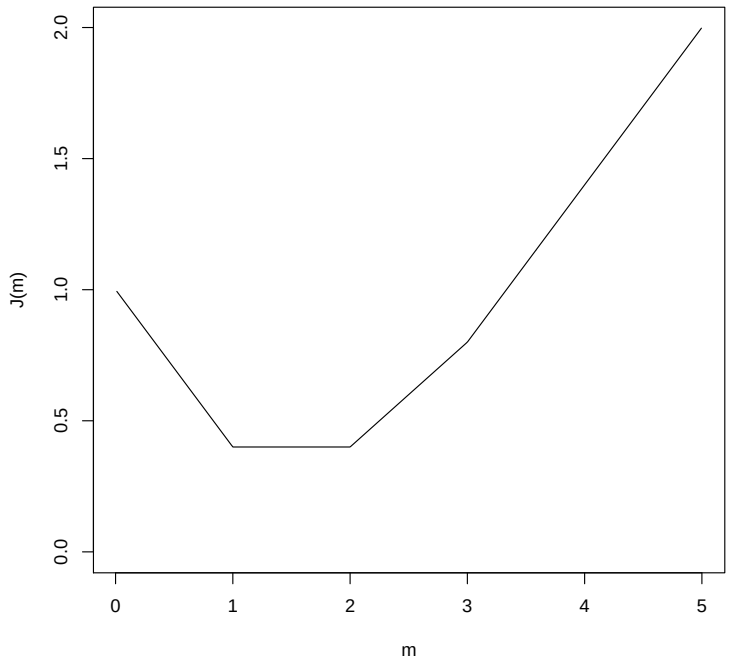



FIGURE 3.3 – Illustration d'une variable aléatoire avec une infinité de minima globaux pour le risque MAPE.

$$\begin{aligned} &= (0.3 - 0.2 - 0.1) \times m + (-0.3 + 0.4 + 0.3), \\ &= 0.4. \end{aligned}$$

qui est le minimum global sur $\mathbb{R}$.

Ici nous définissons alors par convention $\arg\min_m J(m) = \frac{3}{2}$, qui correspond à la valeur moyenne de l'intervalle $[1; 2]$ sur lequel la fonction $J$ atteint son minimum.

Dans cette section, nous avons vu qu'il est possible, sous certaines hypothèses, d'assurer l'existence de la fonction de régression MAPE. Dans la section suivante nous proposons de discuter de la valeur de l'estimateur obtenu par minimisation de la MAPE, par rapport à celui obtenu par minimisation de la MAE.

### 3.2.5 Comparaison des estimateurs MAE et MAPE

Soit (X,Y) un couple de variables aléatoires, $n$ un entier naturel strictement positif, $(X_i, Y_i)_{i=1,\dots,n}$ des copies indépendantes et identiquement distribuées du couple $(X, Y)$, et $\mathcal{G}$ une classe de modèles. Comme nous l'avons vu en introduction, l'estimateur MAE,

noté $\widehat{m}_{MAE}$ est celui minimisant l'erreur absolue moyenne, et est obtenu par

$$\widehat{m}_{MAE} = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} |y_i - g(x_i)| .$$

On peut montrer que, pour tout $x$, $\widehat{m}_{MAE}(x)$ correspond à la médiane de la variable aléatoire $Y$ conditionnellement à $X = x$. Plus formellement :

$$\mathbb{P}\left(Y > m_{MAE}(x)|X = x\right) = \mathbb{P}\left(Y \leq m_{MAE}(x)|X = x\right) = \frac{1}{2}.$$

La proposition 12 donne une relation entre les estimateurs MAE et MAPE. En particulier, cette proposition montre que, dans le cas où $Y$ est une variable aléatoire strictement positive, l'estimateur MAPE est toujours inférieur à l'estimateur MAE. Ce résultat est formalisé à la proposition 12.

**Proposition 12.** *Soit $T$ une variable aléatoire réelle positive, de médiane $\mu$, et $J$ la fonction définie pour tout réel $m$ par $J(m) = \mathbb{E}\left[\left|\frac{m-T}{T}\right|\right]$. Alors, en notant $m^\star = \arg\min_m J(m)$ on a*

$$m^\star \leq \mu.$$

*Démonstration.* Considérons une variable aléatoire $T$ réelle positive, de médiane $\mu$, $m \in [\mu, +\infty[$ et soit $J(m) = \mathbb{E}\left[\left|\frac{m-T}{T}\right|\right]$. Montrons que $J(m) \geq J(\mu)$.

Par définition,

$$J(m) = \mathbb{E}\left[\frac{m-T}{T}\mathbb{1}_{0\leq T\leq \mu}\right] + \mathbb{E}\left[\frac{m-T}{T}\mathbb{1}_{\mu\leq T\leq m}\right] + \mathbb{E}\left[\frac{T-m}{T}\mathbb{1}_{m\leq T}\right].$$

De même :

$$J(\mu) = \mathbb{E}\left[\frac{\mu-T}{T}\mathbb{1}_{0\leq T\leq \mu}\right] + \mathbb{E}\left[\frac{T-\mu}{T}\mathbb{1}_{\mu\leq T\leq m}\right] + \mathbb{E}\left[\frac{T-\mu}{T}\mathbb{1}_{m\leq T}\right].$$

Donc par linéarité de l'espérance, en notant $\delta(\mu, m) = J(m) - J(\mu)$ on a

$$\delta(\mu,m) = \underbrace{\mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{0\leq T\leq \mu}\right]}_{\delta_1} - \underbrace{\left(\mathbb{E}\left[\left(2-\frac{m+\mu}{T}\right)\mathbb{1}_{\mu\leq T\leq m}\right] + \mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{m\leq T}\right]\right)}_{\delta_2}.$$

Or

$$\begin{aligned}
\delta_2 &\leq \mathbb{E}\left[\frac{2m-(m+\mu)}{T}\mathbb{1}_{\mu\leq T\leq m}\right] + \mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{m\leq T}\right], \\
&= \mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{\mu\leq T\leq m}\right] + \mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{m\leq T}\right], \\
&= \mathbb{E}\left[\frac{m-\mu}{T}\mathbb{1}_{\mu\leq T}\right],
\end{aligned}$$

$$\leq \quad \mathbb{E}\left[\frac{m-\mu}{\mu}\mathbb{1}_{\mu\leq T}\right].$$

Par ailleurs,

$$\delta_1 \geq \mathbb{E}\left[\frac{m-\mu}{\mu}\mathbb{1}_{0\leq T\leq \mu}\right]$$

et par définition de $\mu$, on a

$$\mathbb{E}\left[\frac{m-\mu}{\mu}\mathbb{1}_{0\leq T\leq \mu}\right] = \mathbb{E}\left[\frac{m-\mu}{\mu}\mathbb{1}_{\mu\leq T}\right]$$

ce qui prouve que $\delta_1 \geq \delta_2$ et donc que

$$J(m) > J(\mu).$$

Ainsi, on a montré que la valeur de $m$ minimisant la fonction $J$ sur $[\mu, +\infty[$ est donnée par $m = \mu$. Comme

$$\arg\min_{m\in\mathbb{R}} J(m) \leq \arg\min_{m>\mu} J(m),$$

ce résultat suffit pour conclure que la valeur $m^\star$ minimisant la fonction $J$ sur $\mathbb{R}$ est inférieure à $\mu$. □

Dans le cas où $Y$ est une variable aléatoire strictement positive, l'estimateur MAPE est toujours inférieur à l'estimateur MAE. A l'inverse, on peut montrer de façon similaire que, pour une variable aléatoire $Y$ strictement négative, l'estimateur MAPE est toujours supérieur à l'estimateur MAE.

Plus généralement, on peut montrer que la valeur absolue de l'estimateur MAPE est toujours inférieure à la valeur absolue de l'estimateur MAE.

Après avoir établi l'existence du risque MAPE (section 3.2), nous avons vu qu'il existe une relation d'ordre entre les estimateurs MAE et MAPE. Dans la section suivante, nous proposons d'étudier l'impact de la fonction de perte MAE ou MAPE sur la consistance de l'estimateur de minimisation empirique.

## 3.3 Contrôle de complexité : nombres de couverture

### 3.3.1 Introduction générale

Dans cette section, nous proposons d'étudier l'influence du choix de la fonction de perte MAPE sur la consistance de l'estimateur obtenu par minimisation du risque empirique. Pour cela, nous nous placerons dans le cadre classique des régressions, en supposant des couples d'observations $Z = (X, Y)$, à valeurs dans $\mathcal{X} \times \mathbb{R}$ où $\mathcal{X}$ est un espace muni d'une métrique.

Comme précédemment, la qualité d'un modèle $g$ (fonction définie sur $\mathcal{X}$ et à valeurs dans $\mathbb{R}$) est mesurée à partir d'une fonction de perte $l$, qui est classiquement les moindres carrés (MSE : *Mean Square Error*), l'erreur absolue (MAE : *Mean Absolute Error*), ou l'erreur relative moyenne (MAPE : *Mean Absolute Percentage Error*) :

$$\forall p, y \in \mathbb{R}, \quad l_{MAPE}(p, y) = \left| \frac{p - y}{y} \right|,$$

avec les conventions que pour tout $a \neq 0$, $\frac{a}{0} = \infty$ et $\frac{0}{0} = 1$. Comme nous l'avons expliqué précédemment, le risque d'un modèle prédictif $g$ est défini comme l'espérance de la perte : $L_l(g) = \mathbb{E}(l(g(X), Y))$. Étant données $N$ copies indépendantes et identiquement distribuées du couple $(X, Y)$, notées $(X_i, Y_i)_{i=1,\dots,N}$, le risque empirique est alors la moyenne empirique de la fonction de perte calculée sur l'ensemble d'apprentissage :

$$\widehat{L}_l(g)_N = \frac{1}{N} \sum_{i=1}^{N} l(g(X_i), Y_i). \tag{3.18}$$

L'objectif de cette section est d'étudier la consistance des stratégies d'apprentissage lorsque la fonction de perte est la MAPE. Plus précisément, pour une fonction de perte $l$, nous définissons $L_l^* = \inf_g L_l(g)$, où le minimum est calculé sur l'ensemble des fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$, ainsi que $L_{l,G}^* = \inf_{g \in G} L_l(g)$, où le minimum est calculé sur une classe de modèles $G$. Nous nous intéressons à l'estimateur $\widehat{g}_{l,N}$ obtenu par la méthode de minimisation du risque empirique (en anglais *Empirical Risk Minimization*, notée ERM), soit $\widehat{g}_{l,N} = \arg\min_{g \in G_N} \widehat{L}_l(g)_N$, et nous montrons l'existance d'une loi des grands nombres uniforme, c'est-à-dire

$$\forall \epsilon > 0, \mathbb{P}\left\{ \lim_{N \to +\infty} \left( \sup_{g \in G_N} \left| \widehat{L}_{MAPE}(\widehat{g}_{l,N})_N - L_{MAPE}(g) \right| \right) > \epsilon \right\} = 0. \tag{3.19}$$

Ce résultat a déjà été établi pour certaines fonctions de perte ($l_{MSE}$ et $l_{MAE}$ par exemple), mais ne peut être généralisé au cas de la MAPE car les propriétés utilisées pour établir la consistance nécessitent l'hypothèse de continuité uniforme au sens de Lipschitz (voir par exemple le lemme 17.6 dans Anthony and Bartlett (1999)), qui n'est pas vérifiée dans le cas de la MAPE.

La preuve proposée de la consistance s'effectue en quatre étapes. D'abord nous montrerons dans la section 3.3.2 qu'il est possible de borner la probabilité à contrôler (équation 3.19) par une quantité dépendant du nombre de couverture $L_p$ de la classe de modèles considérée, que nous définirons dans cette même section. Puis, nous verrons dans la section 3.4.2 qu'il est possible de contrôler la borne obtenue par la dimension de Vapnik-Chervonenkis de la classe de fonctions considérée. Enfin, l'application de la loi uniforme des grands nombres, dans la section 3.5, nous fournira une nouvelle borne de la quantité à contrôler, à partir de laquelle nous pourrons assurer la consistance de l'estimateur obtenu par minimisation du risque empirique.

### 3.3.2 Lien avec les nombres de couverture

La consistance de la méthode de minimisation du risque empirique peut être obtenue en contrôlant la complexité de la famille de modèles $G$, où la notion de complexité peut être définie à l'aide des nombres de couverture (aussi appelés *covering number* en anglais), comme évoqué dans Gyorfi *et al.* (2002) par exemple. Le principe des couvertures est d'approcher un espace continu $\mathcal{E}$ par $n$ boules, de diamètre $\epsilon$, comme représenté à la figure 3.4. Le nombre de couverture de l'espace $\mathcal{E}$, pour un diamètre $\epsilon$ donné, noté $\mathcal{N}(\mathcal{E}, \epsilon)$, est alors le nombre minimal de boules nécessaires pour recouvrir $\mathcal{E}$. Ainsi, la complexité de l'espace $\mathcal{E}$ peut être mesurée par l'évolution de $\mathcal{N}(\mathcal{E}, \epsilon)$ en fonction de $\epsilon$.

Dans cette section, nous proposons donc d'étudier l'impact du choix de la MAPE sur les nombres de couverture d'une classe de fonction $G$.

#### Notations et définitions

Afin de comparer deux fonctions $g_1$ et $g_2$ de $G$, il convient d'introduire la notion de dissimilarité entre fonctions. Nous appellerons dissimilarité sur $G$, une fonction $\kappa$ positive et symétrique définie sur $G^2$ et à valeurs dans $\mathbb{R}^+$ qui mesure la dissimilarité entre deux fonctions appartenant à $G$ (en particulier, pour tout $g \in G$, on a $\kappa(g, g) = 0$). Les fonctions de dissimilarité peuvent être utilisées pour caractériser la complexité de $G$, en calculant le nombre de $\kappa$ $\epsilon$-couverture de $G$, noté $\kappa$ $\epsilon$-*covering number*.

Une dissimilarité classique est la norme infinie, définie de la façon suivante :

**Définition 3** (Norme infinie)**.** *Soit $\mathcal{Z}$ un sous-espace de $\mathbb{R}^d$, avec $d \in \mathbb{N}$, et $F$ une classe de fonctions positives, définies sur $\mathcal{Z}$ et à valeurs dans $\mathbb{R}$. La norme infinie sur $F$ est donnée par*

$$\|f\|_\infty = \sup_{z \in \mathcal{Z}} |f(z)|.$$

*et nous notons $\|F\|_\infty = \sup_{f \in F} \|f\|_\infty$. Il est immédiat que*

$$\forall f \in F, \forall z \in \mathcal{Z}, |f(z)| \leq \|F\|_\infty.$$

**Définition 4** ($\epsilon-$couverture)**.** *Soit $\mathcal{X}$ un sous-espace de $\mathbb{R}^d$, avec $d \in \mathbb{N}$, $F$ une classe de fonctions positives de $\mathcal{X}$ dans $\mathbb{R}^+$ et $\kappa$ une dissimilarité sur $F$. Pour tout $\epsilon > 0$ et $p$ un entier naturel, une $\epsilon$-couverture de $F$ de taille $p$ par rapport à $\kappa$ est une collection finie $f_1, \ldots, f_p$ d'éléments de $F$ telle que pour tout $f \in F$*

$$\min_{1 \leq i \leq p} \kappa(f, f_i) < \epsilon.$$

On définit alors le $\kappa$ $\epsilon$-covering number de $F$ de la façon suivante :

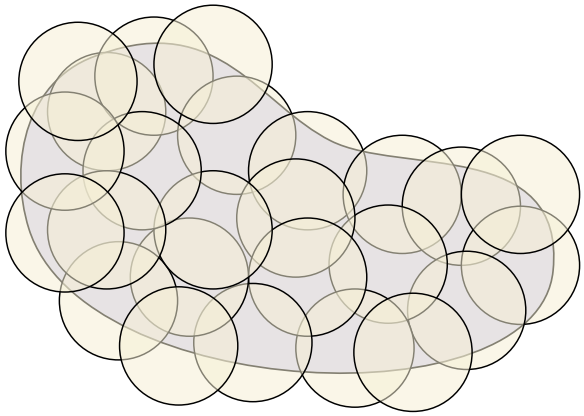


FIGURE 3.4 – Représentation graphique d'une $\epsilon-$couverture.

**Définition 5** (Nombre de couverture)**.** *Soit $\mathcal{X}$ un sous-espace de $\mathbb{R}^d$, avec $d \in \mathbb{N}$, $F$ une classe de fonctions positives de $\mathcal{X}$ dans $\mathbb{R}^+$, $\kappa$ une dissimilarité sur $F$ et $\epsilon > 0$. Alors le $\kappa$ $\epsilon$-covering number de $F$, noté $\mathcal{N}(\epsilon, F, \kappa)$, est la taille de la plus petite $\kappa$ $\epsilon$-couverture de $F$. Si une telle couverture n'existe pas, le $\kappa$ $\epsilon$-covering number de $F$ est infini.*

Le comportement de $\mathcal{N}(\epsilon, F, \kappa)$ par rapport à $\epsilon$ caractérise la complexité de $F$ à travers $\kappa$. Si la croissance de $\mathcal{N}(\epsilon, F, \kappa)$ lorsque $\epsilon$ tend vers 0 est suffisamment lente (pour un choix adapté de $\kappa$), alors la loi uniforme des grands nombres s'applique (voir le Lemme 1).

Afin d'alléger les notations, nous utiliserons les classes de fonctions dérivées, définies ci-dessous.

**Définition 6** (Classes de fonctions dérivées)**.** *Étant données une classe de modèle $G$ et une fonction de perte $l$, nous définissons les classes de fonctions dérivées, $H(G, l)$, par*

$$H(G, l) = \{h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^+,\ h(x, y) = l(g(x), y) \mid g \in G\}, \tag{3.20}$$

*et $H^+(G, l)$ par*

$$H^+(G, l) = \{h : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+,\ h(x, y, t) = \mathbb{1}_{t \leq l(g(x), y)} \mid g \in G\}. \tag{3.21}$$

En considérant les notations que nous avons introduites, nous pouvons réécrire le Lemme 9.1 de Gyorfi *et al.* (2002) de la façon suivante :

**Lemme 1** (Lemme 9.1 de Gyorfi *et al.* (2002))**.** *Soit $n, d \in \mathbb{N}$, $B \in \mathbb{R}^+$, $\mathcal{Z}$ un sous-ensemble de $\mathbb{R}^d$, $Z$ une variable aléatoire sur $\mathcal{Z}$ et $F_n$ une classe de fonctions de $\mathcal{Z}$ dans $[0, B]$. Notons $Z_1, \ldots, Z_n$ des copies indépendantes et identiquement distribuées de la variable aléatoire $Z$.*

*Alors pour tout* $\epsilon > 0$

$$\mathbb{P}\left\{\sup_{f\in F_n}\left|\frac{1}{n}\sum_{j=1}^{n} f(Z_j) - \mathbb{E}(f(Z))\right| > \epsilon\right\} \leq 2\mathcal{N}\left(\frac{\epsilon}{3}, F_n, \|.\|_\infty\right) e^{-\frac{2n\epsilon^2}{9B^2}}.$$

*Si en outre pour tout* $\epsilon > 0$ *on a*

$$\sum_{n=1}^{\infty} \mathcal{N}\left(\frac{\epsilon}{3}, F_n, \|.\|_\infty\right) < \infty,$$

*alors*

$$\lim_{n\to\infty} \sup_{f\in F_n}\left|\frac{1}{n}\sum_{j=1}^{n} f(Z_j) - \mathbb{E}(f(Z))\right| \overset{p.\ s.}{=} 0. \tag{3.22}$$

Un application du Lemme 1 à $H(G,l)$ donne directement

$$\mathbb{P}\left\{\sup_{g\in G}\left|\widehat{L}_l(g)_n - L_l(g)\right| > \epsilon\right\} \leq 2\mathcal{N}\left(\frac{\epsilon}{3}, H(G,l), \|.\|_\infty\right) e^{-\frac{2n\epsilon^2}{9B^2}},$$

sous réserve que le support de la norme infinie coïncide avec le support de $(X,Y)$ et que les fonctions appartenant à $H(G,l)$ soient bornées, ce qui nécessite que $\|h_1 - h_2\|_\infty < \infty$ pour tout $h_1, h_2 \in H(G,l)$. Dans le cas contraire, il est possible de se restreindre à un sous-espace de $\mathcal{Z}$ sur lequel la norme infinie est bien définie, comme nous le verrons dans les exemples qui suivent.

**Existence de $\|h_1 - h_2\|_\infty$ pour $h_1, h_2 \in H(G,l)$**

Pour des fonctions de perte classiques, la norme infinie est généralement mal définie sur $H(G,l)$. Par exemple, pour $h_1$ et $h_2$ deux fonctions sur $H(G, l_{MSE})$, générées par $g_1$ et $g_2$ (i.e. $h_i(x,y) = (g_i(x) - y)^2$ pour $i \in \{1,2\}$) :

$$\begin{aligned}|h_1(x,y) - h_2(x,y)| &= |(g_1(x) - y)^2 - (g_2(x) - y)^2|,\\ &= |g_1(x)^2 - g_2(x)^2 + 2y(g_2(x) - g_1(x))|.\end{aligned}$$

Si $G$ n'est pas réduit à une unique fonction, alors il existe deux fonctions $g_1$ et $g_2$ et une valeur de $x$ telle que $g_1(x) \neq g_2(x)$. Par conséquent $\sup_y |h_1(x,y) - h_2(x,y)| = \infty$.

Une situation similaire survient dans le cas de la MAPE. En effet, considérons deux fonctions $h_1$ et $h_2$ de $H(G, l_{MAPE})$, générées par $g_1$ et $g_2$ appartenant à $G$ (i.e., pour $i \in \{1,2\}$, $h_i(x,y) = \frac{|g_i(x)-y|}{|y|}$). Alors

$$\|h_1 - h_2\|_\infty = \sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \frac{||g_1(x) - y| - |g_2(x) - y||}{|y|}.$$

Remarquons que, si $G$ n'est pas un singleton, il existe toujours une valeur de $x$ et au moins deux fonctions $g_1$ et $g_2$ telles que $g_1(x) \neq 0$ et $|g_2(x)| \neq |g_1(x)|$.

Alors $\lim_{y\to 0} ||g_1(x)-y|-|g_2(x)-y|| = ||g_1(x)|-|g_2(x)|| > 0$ et par conséquent $\lim_{y\to 0+} \frac{||g_1(x)-y|-|g_2(x)-y||}{|y|} = +\infty$. Ainsi, $\|h_1 - h_2\|_\infty = +\infty$ pour $h_1$ et $h_2$ deux fonctions distinctes de $H(G, l_{MAPE})$.

Un moyen simple d'obtenir une valeur finie pour la norme infinie consiste à restreindre sa définition sur un sous-ensemble de $\mathcal{Z}$. Cela correspond en pratique à faire des hypothèses sur le support des données $(X, Y)$. Des hypothèses sur $G$ sont en général nécessaires, comme $\|G\|_\infty < \infty$.

Ces hypothèses dépendent généralement de la fonction de perte considérée. Par exemple, dans le cas des moindres carrés, il est naturel de supposer que $|Y|$ est **borné supérieurement** par $Y_U \in \mathbb{R}$ presque sûrement. Ainsi, si $(x, y) \in \mathbb{R}^d \times [-Y_U, Y_U]$ alors

$$|h_1(x,y) - h_2(x,y)| \leq 2\|G\|_\infty(\|G\|_\infty + Y_U),$$

et par conséquent la norme infinie est bien définie sur le sous-ensemble considéré.

Dans le cas de la MAPE, une hypothèse naturelle consiste à supposer que $|Y|$ est **borné inférieurement** par $Y_L$ presque sûrement.

Ainsi, pour $(x, y) \in \mathbb{R}^d \times (]-\infty, -Y_L] \cup [Y_L, \infty[)$ on a

$$|h_1(x,y) - h_2(x,y)| \leq 2 + 2\frac{\|G\|_\infty}{Y_L},$$

et par conséquent la norme infinie est bien définie sur le sous-ensemble considéré.

Le cas de l'erreur absolue, $l_{MAE}$ est légèrement différent. En effet, à $x$ fixé, on a

$$\forall y > \max(g_1(x), g_2(x)), \quad ||g_1(x)-y|-|g_2(x)-y|| = |g_1(x)-g_2(x)|.$$

De façon similaire, pour $y$ suffisamment grand négatif,

$$||g_1(x)-y|-|g_2(x)-y|| = |g_1(x)-g_2(x)|.$$

Ainsi, la norme infinie est bien définie sur $H(G, l_{MAE})$.

**Contrôle de la borne B** Si par exemple $\|G\|_\infty < \infty$, et si il existe $Y_U \in \mathbb{R}^{+\star}$ tel que $|Y| \leq Y_U$ presque sûrement, on a

$$\forall g \in G, (g(X)-Y)^2 \leq \|G\|_\infty^2 + Y_U^2 \quad (p.s.),$$

et

$$\forall g \in G, |g(X)-Y| \leq \|G\|_\infty + Y_U \quad (p.s.).$$

Alors le Lemme 1 peut être appliqué à $H(G, l_{MSE})$ (resp. à $H(G, l_{MAE})$) pour tout $B \geq \|G\|_\infty^2 + Y_U^2$ (resp. $B \geq \|G\|_\infty + Y_U$).

De même, dans le cas de la MAPE en supposant qu'il existe $Y_L \in \mathbb{R}^{+\star}$ tel que $|Y| \geq Y_L > 0$ presque sûrement, alors si $\|G\|_\infty$ est finie on a

$$\forall g \in G, \frac{|g(X)-Y|}{|Y|} \leq 1 + \frac{\|G\|_\infty}{Y_L} \quad (p.s.),$$

et par conséquent le Lemme 1 peut être appliqué à $H(G, l_{MAPE})$ pour tout $B \geq 1 + \frac{\|G\|_\infty}{Y_L}$.

Cette discussion montre que $Y_L$, la borne inférieure sur $|Y|$, joue un rôle très similaire pour la MAPE à celui joué par $Y_U$, la borne supérieure sur $|Y|$, pour la MAE et la MSE.

### 3.3.3 Lien entre les nombres de couverture MAE et MAPE

La relation entre les nombres de couverture associés à $H(G, l_{MAPE})$ et $H(G, l_{MAE})$ fait l'objet de la proposition suivante :

**Proposition 13.** *Soit $G$ une classe de modèles telle que $\|G\| < \infty$, et $Y_L > 0$. Soit $\|.\|_\infty^{Y_L}$ la norme infinie sur $H(G, l_{MAPE})$ définie par*

$$\|h\|_\infty^{Y_L} = \sup_{x \in \mathbb{R}^d, y \in ]-\infty, -Y_L] \cup [Y_L, \infty[} h(x, y).$$

*Soit $\epsilon > 0$, alors*

$$\mathcal{N}(\epsilon, H(G, l_{MAPE}), \|.\|_\infty^{Y_L}) \leq \mathcal{N}(\epsilon Y_L, H(G, l_{MAE}), \|.\|_\infty).$$

*Démonstration.* Soit $\epsilon > 0$ et $h'_1, \ldots, h'_k$ une $\epsilon Y_L$ couverture de $H(G, l_{MAE})$ (ainsi $k = \mathcal{N}(\epsilon Y_L, H(G, l_{MAE}), \|.\|_\infty)$). Soit $g_1, \ldots, g_k$ des fonctions appartenant à $G$ associées à $h'_1, \ldots, h'_k$ et soit $h_1, \ldots, h_k$ les fonctions correspondantes dans $H(G, l_{MAPE})$. Alors $h_1, \ldots, h_k$ est une $\epsilon-$couverture de $H(G, l_{MAPE})$.

En effet, soit $h$ un élément de $H(G, l_{MAPE})$ associé à $g$, et $h'$ la fonction correspondante dans $H(G, l_{MAE})$. Alors pour un entier $j$ fixé, $\|h' - h'_j\|_\infty \leq \epsilon Y_L$. Nous avons alors

$$\|h - h_j\|_\infty^{Y_L} = \sup_{x \in \mathbb{R}^d, y \in ]-\infty, -Y_L] \cup [Y_L, \infty[} \frac{||g(x) - y| - |g_j(x) - y||}{|y|}.$$

Or pour $y \in ]-\infty, -Y_L] \cup [Y_L, \infty[$, on a $\frac{1}{|y|} \leq \frac{1}{Y_L}$. Par conséquent,

$$\|h - h_j\|_\infty^{Y_L} \leq \sup_{x \in \mathbb{R}^d, y \in ]-\infty, -Y_L] \cup [Y_L, \infty[} \frac{||g(x) - y| - |g_j(x) - y||}{Y_L}.$$

Alors

$$\begin{aligned} \sup_{x \in \mathbb{R}^d, y \in ]-\infty, -Y_L] \cup [Y_L, \infty[} ||g(x) - y| - |g_j(x) - y|| &\leq \sup_{x \in \mathbb{R}^d, y \in \mathbb{R}} ||g(x) - y| - |g_j(x) - y||, \\ &\leq \|h' - h'_j\|_\infty, \\ &\leq \epsilon Y_L. \end{aligned}$$

On en déduit que

$$\|h - h_j\|_\infty^{Y_L} \leq \epsilon,$$

ce qui permet de conclure. □

Cette proposition montre que les nombres de couverture associés à une classe de fonctions $G$ dans le cas de la MAPE sont reliés aux nombres de couverture de cette même classe de fonctions dans le cas de la MAE, dès lors que $|Y|$ est inférieurement borné par un réel strictement positif.

### 3.3.4 Nombres de couverture $L_p$

Une analyse très similaire peut être faite en utilisant les nombres de couverture $L_p$ (en anglais $L_p$ *covering numbers*) qui, contrairement aux nombres de couverture infinis, ne dépendent pas de la norme infinie mais de la norme $L_p$, notée $\|.\|_p$ et qui dépend des copies observées de la variables aléatoire $Z$.

**Définition 7** (Norme $L_p$)**.** *Soit $\mathcal{Z}$ un sous-espace de $\mathbb{R}^d$, avec $d \in \mathbb{N}$, et $F$ une classe de fonctions positives, définies sur $\mathcal{Z}$ et à valeurs dans $\mathbb{R}$. Soit $Z_1, \ldots, Z_n$ des copies indépendantes et identiquement distribuées de la variable aléatoire $Z$, et $f_1, f_2$ deux éléments de $F$. Alors*

$$\|f_1 - f_2\|_p = \left( \frac{1}{n} \sum_{i=1}^{n} |f_1(Z_i) - f_2(Z_i)|^p \right)^{\frac{1}{p}}.$$

La proposition suivante donne une adaptation de la proposition 13 au cas des nombres de couverture $L_p$.

**Proposition 14.** *Soit $G$ une classe de fonctions arbitraire, et $(X_i, Y_i)_{i=1,\ldots,n}$ $n$ copies aléatoires et identiquement distribuées d'un couple de variable aléatoire $X, Y$, tels que $Y_i \neq 0$ pour tout $i \in \{1, \ldots, n\}$. Alors*

$$\mathcal{N}(\epsilon, H(G, l_{MAPE}), \|.\|_p) \leq \mathcal{N}(\epsilon \min_{1 \leq i \leq N} |Y_i|, H(G, l_{MAE}), \|.\|_p).$$

*Démonstration.* La preuve est similaire à celle proposée à la proposition 13. □

Alors, avec nos notations, le Théorème 9.1 de Gyorfi *et al.* (2002) se réécrit de la façon suivante :

**Théorème 1** (Théorème 9.1 de Gyorfi *et al.* (2002))**.** *Soit $n, d, p$ des entiers naturels strictement positifs, $B \in \mathbb{R}^+$, $\mathcal{Z}$ un espace arbitraire, $Z$ une variable aléatoire sur $\mathcal{Z}$ et $F$ une classe de fonctions de $\mathcal{Z}$ dans $[0, B]$.*

*Notons $Z_1, \ldots, Z_n$ des copies indépendantes et identiquement distribuées de la variable aléatoire $Z$.*

*Alors pour $\epsilon > 0$ :*

$$\mathbb{P}\left\{ \sup_{f \in F} \left| \frac{1}{n} \sum_{j=1}^{n} f(Z_j) - \mathbb{E}(f(Z)) \right| > \epsilon \right\} \leq 8\mathbb{E}\left\{ \mathcal{N}\left( \frac{\epsilon}{8}, F, \|.\|_{p, D_n} \right) \right\} e^{-\frac{n\epsilon^2}{128B^2}},$$

*où l'espérance des nombres de couverture est prise sur l'échantillon $D_n = (Z_i)_{1 \leq i \leq n}$.*

Comme dans le cas du Lemme 1, nous pouvons borner $\|H(G,l)\|_p$ via les hypothèses sur $G$ et $Y$. Par exemple, dans le cas de la MAE, on a

$$\mathbb{P}\left\{\sup_{g\in G}\left|\widehat{L}_{MAE}(g,D_n)-L_{MAE}(g)\right|>\epsilon\right\}\leq 8\mathbb{E}\left\{\mathcal{N}\left(\frac{\epsilon}{8},H(G,l_{MAE}),\|.\|_{p,D_n}\right)\right\}e^{-\frac{n\epsilon^2}{128(\|G\|_\infty+Y_U)^2}}, \quad (3.23)$$

et dans le cas de la MAPE

$$\mathbb{P}\left\{\sup_{g\in G}\left|\widehat{L}_{MAPE}(g,D_n)-L_{MAPE}(g)\right|>\epsilon\right\}\leq 8\mathbb{E}\left\{\mathcal{N}\left(\frac{\epsilon}{8},H(G,l_{MAPE}),\|.\|_{p,D_n}\right)\right\}e^{-\frac{n\epsilon^2 Y_L^2}{128(1+\|G\|_\infty)^2}}. \quad (3.24)$$

Ainsi, nous disposons d'une borne de la quantité à contrôler en fonction de l'espérance du nombre de couverture. Pour assurer la consistance, nous proposons maintenant d'exprimer cette borne en fonction de la dimension de Vapnik-Chervonenkis, notée *VC-dimension*, de la classe de fonctions considérée. Ce résultat nous permettra d'assurer la consistance de l'estimateur de minimisation du risque empirique en fonction de la complexité de la classe de fonctions choisie.

## 3.4 Contrôle de complexité : dimension de Vapnik-Chervonenkis

### 3.4.1 Introduction générale

Une façon de borner les nombres de couverture consiste à utiliser la dimension de Vapnik-Chervonenkis. Rappelons tout d'abord la définition du coefficient de pulvérisation d'une classe de fonctions.

**Définition 8** (Coefficient de pulvérisation)**.** *Soit $d\in\mathbb{N}$, $F$ une classe de fonctions de $\mathbb{R}^d$ dans $\{0,1\}$ et $n$ un entier positif. Soit $\{z_1,\dots,z_n\}$ un ensemble de $n$ points de $\mathbb{R}^d$. Soit*

$$s(F,\{z_1,\dots,z_n\})=|\{\theta\in\{0,1\}^n|\exists f\in F,\ \theta=(f(z_1),\dots,f(z_n))\}|,$$

*le nombre de vecteurs différents de taille $n$ engendrés par des fonctions de $F$ lorsqu'elles sont appliquées à $\{z_1,\dots,z_n\}$.*

*L'ensemble $\{z_1,\dots,z_n\}$ est dit **pulvérisé** par $F$ si $s(F,\{z_1,\dots,z_n\})=2^n$. Le $n$-ième coefficient de pulvérisation de $F$ est défini par*

$$\mathcal{S}(F,n)=\max_{\{z_1,\dots,z_n\}\subset\mathbb{R}^d} s(F,\{z_1,\dots,z_n\}).$$

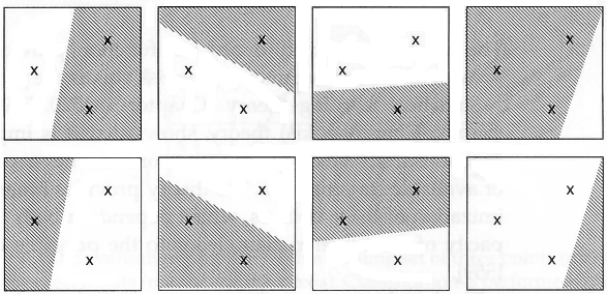


FIGURE 3.5 – Illustration de la dimension de Vapnik-Chervonenkis dans $\mathbb{R}^2$ dans le cas où le séparateur est une droite.

Alors la VC-dimension est définie de la façon suivante :

**Définition 9** (VC-dimension)**.** *Soit $F$ une classe de fonctions de $\mathbb{R}^d$ dans $\{0,1\}$. La VC-dimension de $F$ est définie comme*

$$VC_{dim}(F) = \sup\{n \in \mathbb{N}^+ \mid \mathcal{S}(F,n) = 2^n\}.$$

La figure 3.5 représente une illustration de la VC-dimension dans $\mathbb{R}^2$ dans le cas où le séparateur est une droite. Plus précisément, cette figure montre que pour tout $x_1, x_2, x_3 \in \mathbb{R}^2$, les points $x_1, x_2$ et $x_3$ peuvent être séparés de toutes les façons par une droite, le coefficient de pulvérisation dans ce cas est donc d'au moins 3. En revanche, tout ensemble de quatre points ne peut pas être séparé de toute les façons par une droite. Considérant par exemple les points $x_1 = (0,0), x_2 = (0,1), x_3 = (1,0)$ et $x_4 = (1,1)$, il est évident que les points $x_1$ et $x_4$ ne peuvent pas être isolés des points $x_2$ et $x_3$ à l'aide d'une droite. Le coefficient de pulvérisation est donc strictement inférieur à 4. Cette analyse permet de montrer que dans cet exemple la VC-dimension est vaut $VC_{dim} = 3$.

### 3.4.2 Contrôle des nombres de couverture $L_p$ par la VC-dimension

Comme présenté dans Gyorfi *et al.* (2002), il est possible de borner le $L^p$ covering number par une fonction dépendant de la VC-dimension :

**Théorème 2** (théorème 9.4 de Gyorfi *et al.* (2002))**.** *Soit $p \geq 1$, $G$ une classe de fonctions de $\mathbb{R}^d$ dans $[0;B]$, telle que $VC_{dim}(H^+(G,l)) \geq 2$, et $0 < \epsilon < \frac{\|H(G,l)\|_\infty}{4}$, alors :*

$$\mathcal{N}(\epsilon, H(G,l), \|.\|_{p,D_n}) \leq 3\left(\frac{2e\|H(G,l)\|_\infty^p}{\epsilon^p}\log\frac{3e\|H(G,l)\|_\infty^p}{\epsilon^p}\right)^{VC_{dim}(H^+(G,l))}. \quad (3.25)$$

En utilisant ces résultats, la partie droite de l'équation (3.23) est bornée supérieure-

ment par

$$24\left(\frac{2e(\|G\|_\infty + Y_U)^p}{\epsilon^p}\log\frac{3e(\|G\|_\infty + Y_U))^p}{\epsilon^p}\right)^{VC_{dim}(H^+(G,l_{MAE}))} e^{-\frac{n\epsilon^2}{128(\|G\|_\infty+Y_U)^2}}, \tag{3.26}$$

tandis que la partie droite de l'équation (3.24) est bornée supérieurement par

$$24\left(\frac{2e(1 + \|G\|_\infty)^p}{Y_L^p\epsilon^p}\log\frac{3e(1 + \|G\|_\infty))^p}{Y_L^p\epsilon^p}\right)^{VC_{dim}(H^+(G,l_{MAPE}))} e^{-\frac{n\epsilon^2 Y_L^2}{128(1+\|G\|_\infty)^2}}. \tag{3.27}$$

Comme nous le verrons dans la section 3.5, ce résultat nous permet d'assurer la consistance de l'estimateur obtenu par minimisation du risque empirique.

### 3.4.3 Lien entre les dimensions de Vapnik-Chervonenkis MAE et MAPE

On peut montrer que le fait de remplacer la MAE par la MAPE ne modifie pas la VC-dimension de la classe de fonctions considérée. Ce résultat fait l'objet de la proposition 15.

**Proposition 15.** *Soit $G$ une classe de modèles. Alors*

$$VC_{dim}(H^+(G, l_{MAPE})) \leq VC_{dim}(H^+(G, l_{MAE})).$$

*Démonstration.* Soit $(v_1, \ldots, v_k)$ un ensemble de $k$ points pulvérisés par $H^+(G, l_{MAPE})$, avec $v_j = (x_j, y_j, t_j)$ pour tout $j \in \{1, \ldots, k\}$. L'objectif est de construire un nouvel ensemble de points pulvérisé au sens de la MAE.

Par définition, pour chaque vecteur binaire $\theta \in \{0, 1\}^k$, il existe une fonction $h_\theta \in H(G, l_{MAPE})$ telle que $\forall j, \mathbb{1}_{t_j \leq h_\theta(x,y)}(x_j, y_j, t_j) = \theta_j$. Chaque $h_\theta$ correspond à une fonction $g_\theta \in G$, avec $h_\theta(x, y) = \frac{|g_\theta(x)-y|}{|y|}$.

Considérons $(w_1, \ldots, w_k)$ un nouvel ensemble de $k$ points tels que si $y_j \neq 0$, alors $w_j = (x_j, y_j, |y_j|t_j)$. Pour ces points et pour tout $g \in G$,

$$\mathbb{1}_{t_j \leq \frac{|g(x_j)-y_j|}{|y_j|}} = \mathbb{1}_{|y_j|t_j \leq |g(x_j)-y_j|},$$

et ainsi en notant $h'_\theta(x, y) = |g_\theta(x) - y|$ on a

$$\mathbb{1}_{t_j \leq h'_\theta(x,y)}(x_j, y_j, |y_j|t_j) = \mathbb{1}_{t_j \leq h_\theta(x,y)}(x_j, y_j, t_j) = \theta_j,$$

et pour tout $\omega_j$ tel que $y_j \neq 0$, il existe donc une fonction $h'_\theta$ telle que $\mathbb{1}_{t_j \leq h'_\theta(x,y)}(x_j, y_j, |y_j|t_j) = \theta_j$.

Considérons maintenant le cas $y_j = 0$.

Par définition,

$$g_\theta(x_j) = 0 \Rightarrow h_\theta(x_j, 0) = 1$$

et

$$g_\theta(x_j) \neq 0 \Rightarrow h_\theta(x_j, 0) = \infty.$$

Or, comme les points sont pulvérisés par $H^+(G, l_{MAPE})$, pour tout $\theta_j \in \{0, 1\}$ il existe une fonction $h_\theta \in H^+(G, l_{MAPE})$ telle que $\mathbb{1}_{t_j \leq h_\theta(x,y)}(x_j, y_j, t_j) = \theta_j$, donc nécessairement $t_j > 1$. Par ailleurs, si $\theta_j = 1$ alors $g_\theta(x_j) \neq 0$, et si $\theta_j = 0$ alors $g_\theta(x_j) = 0$.

Soit alors $w_j = (x_j, 0, \min_{\theta, \theta_j=1} |g_\theta(x_j)|)$. Remarquons que $\min_{\theta,\theta_j=1} |g_\theta(x_j)| > 0$ (puisqu'il y a un nombre fini de vecteurs binaires de dimension $k$).

Pour $\theta$ tel que $\theta_j = 0$, on a

$$h'_\theta(x_j, y_j) = |g_\theta(x_j) - y_j| = 0$$

et donc

$$h'_\theta(x_j, y_j) < \min_{\theta,\theta_j=1} |g_\theta(x_j)|, \ \textit{i.e.}\ \mathbb{1}_{t \leq h'_\theta(x,y)}(w_j) = 0 = \theta_j.$$

Pour $\theta$ tel que $\theta_j = 1$, on a :

$$h'_\theta(x_j, y_j) = |g_\theta(x_j)|$$

et donc

$$h'_\theta(x_j, y_j) \geq \min_{\theta,\theta_j=1} |g_\theta(x_j)|.$$

Alors

$$\mathbb{1}_{t \leq h'_\theta(x,y)}(w_j) = 1 = \theta_j.$$

Combiné au cas $y_j \neq 0$, ce résultat assure que pour tout vecteur binaire $\theta \in \{0, 1\}^k$, il existe une fonction $h'_\theta \in H(G, l_{MAE})$ telle que $\forall j,\ \mathbb{1}_{t \leq h'_\theta(x,y)}(w_j) = \theta_j$. Par conséquent les $w_j$ sont pulvérisés par $H^+(G, l_{MAE})$.

On en déduit que $VC_{dim}(H^+(G, l_{MAE})) \geq k$.

Si $VC_{dim}(H^+(G, l_{MAPE})) < \infty$, alors on peut prendre $k = VC_{dim}(H^+(G, l_{MAPE}))$ pour conclure la preuve.

Si $VC_{dim}(H^+(G, l_{MAPE})) = \infty$ alors $k$ peut être choisi arbitrairement grand et alors $VC_{dim}(H^+(G, l_{MAE})) = \infty$. □

## 3.5 Consistance

Afin d'obtenir la convergence uniforme presque sûre de $\widehat{L}_l(g, D_n)$ vers $L_l(g)$ sur $G$, les quantités majorantes définies aux équations (3.26) et (3.27) doivent être sommables (ce qui autorise l'application du Lemme de Borel-Cantelli). Pour des valeurs fixées de la VC-dimension, de $\|G\|_\infty$, $Y_L$ et $Y_U$, cela est toujours assuré.

Si ces quantités peuvent dépendre de $n$, alors il est évident que, comme c'était le cas pour les nombres de couvertures, $Y_U$ et $Y_L$ jouent des rôles symétriques pour la MAPE et la MAE. En effet, pour la MAE, une forte croissance de $Y_U$ avec $n$ peut empêcher

la borne d'être sommable. Par exemple, si $Y_U$ croît plus rapidement que $\sqrt{n}$, alors $\frac{n}{(\|G_n\|_\infty+Y_U)^2}$ ne converge pas vers 0 et la série n'est pas sommable. De façon similaire, si $Y_L$ converge trop rapidement vers zéro, par exemple en $\frac{1}{\sqrt{n}}$, alors $\frac{nY_L^2}{(1+\|G_n\|_\infty)^2}$ ne converge pas vers 0 et la série n'est pas sommable.

Dans cette section, nous développons avec plus de détails ces conditions dans le cas de la MAPE et nous montrons que, sous des hypothèses simples sur $(X, Y)$, par le principe de minimisation du risque empirique on peut construire un estimateur consistant du modèle optimal.

**Théorème 3.** *Soit $Z = (X, Y)$ une paire de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathbb{R}$ telle que $|Y| \geq Y_L > 0$ presque sûrement ($Y_L$ étant un nombre réel fixé). Soit $n \in \mathbb{N}$, et $(Z_n)_{n\geq 1} = (X_n, Y_n)_{n\geq 1}$ une série de copies indépendantes de $Z$.*

*Soit $(G_n)_{n\geq 1}$ une série de classes de fonctions mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$, telle que :*

1. *$G_n \subset G_{n+1}$ ;*
2. *$\bigcup_{n\geq 1} G_n$ est dense dans l'ensemble des fonctions $L^1(\mu)$ de $\mathbb{R}^d$ dans $\mathbb{R}$ pour toute mesure de probabilité $\mu$ ;*
3. *Pour tout $n$, $V_n = VC_{dim}(H^+(G_n, l_{MAPE})) < \infty$ ;*
4. *Pour tout $n$, $\|G_n\|_\infty < \infty$.*

*Si en outre*

$$\lim_{n\to\infty} \frac{V_n \|G_n\|_\infty^2 \log \|G_n\|_\infty}{n} = 0,$$

*et il existe $\delta > 0$ tel que*

$$\lim_{n\to\infty} \frac{n^{1-\delta}}{\|G_n\|_\infty^2} = \infty.$$

*Alors en notant $\widehat{g}_{l_{MAPE},G_n}$ l'estimateur de minimisation du risque empirique MAPE sur la classe de fonction $G_n$, $L_{MAPE}(\widehat{g}_{l_{MAPE},G_n})$ converge presque sûrement vers $L^*_{MAPE}$.*

*Démonstration.* Nous utilisons la décomposition classique entre erreur d'estimation et erreur d'approximation. Plus précisément, pour une classe de fonction $G$ et $g \in G$,

$$L_{MAPE}(g) - L^*_{MAPE} = \underbrace{L_{MAPE}(g) - L^*_{MAPE,G}}_{\text{erreur d'estimation}} + \underbrace{L^*_{MAPE,G} - L^*_{MAPE}}_{\text{erreur d'approximation}}.$$

La preuve de la consistance que nous proposons s'effectue alors en deux étapes. Dans un premier temps, nous montrerons que l'erreur d'approximation tend vers 0 lorsque $n$ tend vers l'infini. Puis, dans un second temps, nous montrerons que l'erreur d'estimation tend vers 0 lorsque $n$ tend vers l'infini.

**Contrôle de l'erreur d'approximation :** Montrons que

$$\lim_{n\to\infty} L^*_{MAPE,G_n} = L^*_{MAPE}.$$

Soit $n \in \mathbb{N}$, et $g_1, g_2 \in G_n$. Pour $x \in \mathbb{R}^d$ et $y \in ]-\infty, -Y_L] \cap [Y_L, +\infty[$, par convexité de la fonction $x \mapsto |x|$ l'application de l'inégalité de Jensen donne

$$\left|\mathbb{E}\left\{\frac{|g_1(X)-Y|}{|Y|} - \frac{|g_2(X)-Y|}{|Y|}\right\}\right| \leq \mathbb{E}\left\{\left|\frac{|g_1(X)-Y|}{|Y|} - \frac{|g_2(X)-Y|}{|Y|}\right|\right\}$$

Par ailleurs, d'après l'inégalité triangulaire on a pour tous réels $a, b$

$$|a| \leq |a-b| + |b| \text{ donc } |a| - |b| \leq |a-b|,$$

et de même

$$|b| \leq |b-a| + |a| \text{ donc } |a| - |b| \geq -|a-b|.$$

Donc pour tous $a, b \in \mathbb{R}$ on a $||a| - |b|| \leq |a-b|$. Par conséquent,

$$\left|\mathbb{E}\left\{\frac{|g_1(X)-Y|}{|Y|} - \frac{|g_2(X)-Y|}{|Y|}\right\}\right| \leq \mathbb{E}\left\{\left|\frac{|g_1(X)-g_2(X)|}{|Y|}\right|\right\}.$$

Et comme $|Y| \geq Y_L$ presque sûrement, par linéarité de l'espérance on a

$$\left|\mathbb{E}\left\{\frac{|g_1(X)-Y|}{|Y|}\right\} - \mathbb{E}\left\{\frac{|g_2(X)-Y|}{|Y|}\right\}\right| \leq \frac{1}{Y_L}\mathbb{E}\left\{|g_1(X)-g_2(X)|\right\},$$

donc

$$|L_{MAPE}(g_1) - L_{MAPE}(g_2)| \leq \frac{1}{Y_L}\mathbb{E}\left\{|g_1(X)-g_2(X)|\right\}. \tag{3.28}$$

Par ailleurs, pour toute fonction $g$ de MAPE finie, d'après le théorème de Fubini on a $x \mapsto \left|\frac{g(x)-y}{y}\right| \in L^1(\mathbb{P}_X)$ pour tout $y$ tel que $\mathbb{P}_Y(y) > 0$. Or, pour tout $y \neq 0$ on a $0 \leq \left|\frac{g(x)}{y}\right| \leq \left|\frac{g(x)-y}{y}\right| + 1$, ce qui permet d'assurer que $g \in L^1(\mathbb{P}_X)$. Pour tout $k \in \mathbb{N}$, il existe donc une fonction $g^*_k \in L^1(\mathbb{P}_X)$ telle que

$$L_{MAPE}(g^*_k) \leq L^*_{MAPE} + \frac{1}{k}$$

et comme $\bigcup_{n\geq 1} G_n$ est dense dans $L^1(\mathbb{P}_X)$, il existe une série de fonctions $(h^*_k)_{k\geq 1}$ de $\bigcup_{n\geq 1} G_n$ telle que $\mathbb{E}\left\{|h^*_k(X) - g^*_k(X)|\right\} \leq \frac{Y_L}{k}$.

Alors en appliquant l'équation 3.28 aux fonctions $h^*_k$ et $g^*_k$, on a pour tout $k \in \mathbb{N}$

$$|L_{MAPE}(h^*_k) - L_{MAPE}(g^*_k)| \leq \frac{1}{Y_L}\frac{Y_L}{k}$$

et donc

$$L_{MAPE}(h^*_k) \leq L_{MAPE}(g^*_k) + \frac{1}{k} \leq L^*_{MAPE} + \frac{2}{k}.$$

Soit $n_k = \min\{n \mid h_k^* \in G_n\}$. Par définition, $L^*_{MAPE,G_{n_k}} \leq L_{MAPE}(h_k^*)$.
En considérant alors $\epsilon > 0$ et $k \in \mathbb{N}$ tel que $\frac{2}{k} \leq \epsilon$, on a

$$L_{MAPE}(g_k^*) \leq L^*_{MAPE} + \epsilon$$

et

$$L^*_{MAPE,G_{n_k}} \leq L^*_{MAPE} + \epsilon.$$

Par conséquent, comme $G_n$ est une série d'ensembles croissants on a :

$$\forall n \geq n_k, \quad L^*_{MAPE,G_n} \leq L^*_{MAPE} + \epsilon.$$

Et donc

$$\lim_{n\to\infty} L^*_{MAPE,G_n} = L^*_{MAPE}.$$

**Contrôle de l'erreur d'estimation** Montrons que

$$\lim_{n\to\infty} |L_{MAPE}(\widehat{g}_{l_{MAPE},G_n}) - L^*_{MAPE,G_n}| = 0 \quad (p.s.).$$

Par application du Théorème 1 on a dans le cas particulier où la fonction de dissimilarité correspond à l'erreur absolue (*i.e.*, $p = 1$) :

$$\mathbb{P}\left\{\sup_{g\in G_n} \left|\widehat{L}_{MAPE}(g)_n - L_{MAPE}(g)\right| > \epsilon\right\} \leq D(n,\epsilon),$$

avec

$$D(n,\epsilon) = 8\mathbb{E}\left\{\mathcal{N}\left(\frac{\epsilon}{8}, H(G_n, l_{MAPE}), \|.\|_{1,D_n}\right)\right\} e^{-\frac{n\epsilon^2 Y_L^2}{128(1+\|G_n\|_\infty)^2}}.$$

Or d'après l'équation (3.27),

$$D(n,\epsilon) \leq 24\left(\frac{2e(1+\|G_n\|_\infty)}{\epsilon Y_L}\log\frac{3e(1+\|G_n\|_\infty))}{\epsilon Y_L}\right)^{V_n} e^{-\frac{n\epsilon^2 Y_L^2}{128(1+\|G_n\|_\infty)^2}}.$$

Et comme $\log(x) \leq x$ pour tout $x > 0$, on a

$$D(n,\epsilon) \leq 24\left(\frac{3e(1+\|G_n\|_\infty)}{\epsilon Y_L}\right)^{2V_n} e^{-\frac{n\epsilon^2 Y_L^2}{128(1+\|G_n\|_\infty)^2}},$$

c'est-à-dire

$$\begin{aligned} D(n,\epsilon) &\leq 24\exp\left(-\frac{n\epsilon^2 Y_L^2}{128(1+\|G_n\|_\infty)^2} + 2V_n\log\frac{3e(1+\|G_n\|_\infty)}{\epsilon Y_L}\right) \\ &\leq 24\exp\left(-\frac{n}{(1+\|G_n\|_\infty)^2}\left(\frac{\epsilon^2 Y_L^2}{128} - \frac{2V_n(1+\|G_n\|_\infty)^2\log\frac{3e(1+\|G_n\|_\infty)}{\epsilon Y_L}}{n}\right)\right). \end{aligned}$$

Comme $\lim_{n\to\infty} \frac{V_n \|G_n\|_\infty^2 \log \|G_n\|_\infty}{n} = 0$,

$$\lim_{n\to\infty} \frac{2V_n(1+\|G_n\|_\infty)^2 \log \frac{3e(1+\|G_n\|_\infty)}{\epsilon Y_L}}{n} = 0.$$

et comme $\lim_{n\to\infty} \frac{n^{1-\delta}}{\|G_n\|_\infty^2} = \infty$,

$$\lim_{n\to\infty} \frac{n^{1-\delta}}{(1+\|G_n\|_\infty)^2} = \infty.$$

En conséquence, pour $n$ assez grand, $D(n,\epsilon)$ est dominé par un terme de la forme

$$\alpha \exp(-\beta n^\delta),$$

avec $\alpha > 0$ et $\beta > 0$ (dépendant tous deux de $\epsilon$). Cela permet de conclure que $\sum_{n\geq 1} D(n,\epsilon) < \infty$. Alors le théorème de Borel-Cantelli implique que

$$\lim_{n\to\infty} \sup_{g\in G_n} \left| \widehat{L}_{MAPE}(g)_n - L_{MAPE}(g) \right| = 0 \quad (ps).$$

Par conséquent, pour tout $\epsilon > 0$ il existe un entier $N$ tel que pour tout entier $n \geq N$ et pour tout $g \in G_n$

$$\sup_{g\in G_n} \left| \widehat{L}_{MAPE}(g)_n - L_{MAPE}(g) \right| \leq \epsilon.$$

et donc pour tout $g \in G_n$

$$\widehat{L}_{MAPE}(g)_n \leq L_{MAPE}(g) + \epsilon.$$

Or par définition de $\widehat{g}_{l_{MAPE},G_n}$, pour tout $g \in G_n$ on a

$$\widehat{L}_{MAPE}(\widehat{g}_{l_{MAPE},G_n})_n \leq \widehat{L}_{MAPE}(g)_n,$$

et donc pour tout $g \in G_n$,

$$\widehat{L}_{MAPE}(\widehat{g}_{l_{MAPE},G_n})_n \leq L_{MAPE}(g) + \epsilon.$$

En prenant l'infimum sur $G_n$, on a par conséquent

$$\widehat{L}_{MAPE}(\widehat{g}_{l_{MAPE},G_n})_n \leq L^*_{MAPE,G_n} + \epsilon.$$

Ce résultat nous permet d'avoir un contrôle de la quantité $\delta_2$

Par ailleurs, pour $n$ assez grand on a

$$\widehat{L}_{MAPE}(\widehat{g}_{l_{MAPE},G_n})_n \geq L_{MAPE}(\widehat{g}_{l_{MAPE},G_n}) - \epsilon,$$

ce qui montre que

$$L_{MAPE}(\widehat{g}_{l_{MAPE},G_n}) \leq L^*_{MAPE,G_n} + 2\epsilon.$$

Or par définition de $L^*_{MAPE,G_n}$, on a $L^*_{MAPE,G_n} \leq L_{MAPE}(\widehat{g}_{l_{MAPE},G_n})$. D'où

$$\lim_{n\to\infty} |L_{MAPE}(\widehat{g}_{l_{MAPE},G_n}) - L^*_{MAPE,G_n}| = 0 \quad (p.s.).$$

**Conclusion de la preuve** La combinaison des résultats démontrés sur le contrôle des erreurs d'estimation et d'approximation permet de conclure que $L_{MAPE}(\widehat{g}_{l_{MAPE},G_n})$ converge presque sûrement vers $L^*_{MAPE}$. □

Dans ce chapitre, nous avons vu que l'existence du risque MAPE est bien définie sous certaines hypothèses sur $Y$. Puis, nous avons montré que la méthode de minimisation du risque empirique fournit un estimateur consistant vis-à-vis de la MAPE. Dans la section suivante, nous proposons deux applications des régressions MAPE.

## 3.6 Applications

Cette section a pour objectif d'illustrer les résultats précédemment obtenus, et de montrer l'intérêt des régressions MAPE.

Dans un premier temps, nous détaillerons dans la section 3.6.1 comment effectuer des régressions MAPE en pratique. Nous verrons que ces régressions s'inscrivent dans un cas particulier des régressions quantiles. Ensuite, nous illustrerons dans la section 3.6.2 l'intérêt de la régression MAPE sur une base de données publique, et comparerons les résultats avec ceux obtenus par les régressions médiane et moindres carrés. Enfin, dans la section 3.6.3, nous illustrerons un cas d'utilisation des régressions MAPE sur des données de Viadeo, afin de modéliser au mieux l'âge des utilisateurs n'ayant pas (ou mal) renseigné cette information sur leur profil personnel.

### 3.6.1 La MAPE en pratique

En pratique, construire un modèle de régression MAPE consiste à minimiser empiriquement la MAPE sur une classe de modèles. Dans le cas où l'on dispose de $n$ observations $(X_i, Y_i)_{i=1,\dots,n}$ indépendantes et identiquement distribuées selon la loi de $(X, Y)$, avec $n \in \mathbb{N}$, cela revient à résoudre

$$\widehat{g}_{l_{MAPE},G_n,D_n} = \arg\min_{g \in G_n} \frac{1}{n} \sum_{i=1}^{n} \frac{|g(X_i) - Y_i|}{|Y_i|}.$$

Du point de vue de l'optimisation, nous avons vu que ce problème peut être considéré comme un cas particulier des régressions médiane (qui sont également un cas particulier des régressions quantile). En effet, le quotient $\frac{1}{|Y_i|}$ peut être vu comme un poids fixé et, par conséquent, tout algorithme de régression quantile qui supporte les pondérations peut être utilisé pour trouver le modèle optimal. C'est par exemple le cas de la bibliothèque R `quantreg` (Koenker, 2013). Remarquons que, lorsque $G_n$ correspond à la classe des modèles linéaires, le problème d'optimisation est un problème de programmation linéaire qui peut être résolu par exemple par les méthodes de point intérieur (Boyd and Vandenberghe, 2004).

| speed | dist |
|---|---|
| 4.00 | 2.00 |
| 4.00 | 10.00 |
| 7.00 | 4.00 |
| 7.00 | 22.00 |
| 8.00 | 16.00 |
| 9.00 | 10.00 |

TABLE 3.2 – Extrait des premières lignes de la base *cars*, disponible dans la librairie *datasets* du logiciel *R*.

### 3.6.2 La base de données publique *cars*

Afin d'illustrer les résultats de la régression MAPE sur un jeu de données public, nous avons choisi de considérer la table *cars*, une base de données jouet disponible dans la bibliothèque *datasets* de R. Cette base, dont un extrait est représenté à la table 3.2, comporte deux variables renseignées sur un ensemble de 50 véhicules :

— *speed* : la vitesse des voitures avant freinage. Il s'agit d'une variable numérique, entière, comprise entre 4 et 25 miles par heure (*i.e.*, entre 6.4 km/h et 40.2 km/h).

— *dist* : la distance de freinage. Il s'agit d'une variable numérique, entière, comprise entre 2 et 120 pieds (*i.e.*, entre 0.6m et 36m).

Afin d'illustrer la qualité des différentes approches de régression, nous avons modélisé la distance de freinage des véhicules en fonction de leur vitesse par un modèle linéaire. Nous supposons donc que la relation entre la distance de freinage et la vitesse des véhicules s'écrit sous la forme

$$dist_i = a_\ell \cdot speed_i + b_\ell,$$

où $a_\ell$ et $b_\ell$ représentent les coefficients de la régression et dépendent de la fonction de perte utilisée : moindres carrés ($l_{MSE}$), erreur absolue ($l_{MAE}$) ou erreur relative ($l_{MAPE}$).

Les valeurs associées à la distance de freinage étant toutes strictement positives et finies, chacune des fonctions de perte, en particulier la MAPE, est bien définie sur l'ensemble des observations. Il est donc possible de procéder à l'estimation des coefficients par la méthode de minimisation du risque empirique. Nous estimons donc les coefficients $a_\ell$ et $b_\ell$ par la méthode de minimisation du risque empirique. Ainsi, les estimateurs des coefficients $a_\ell$ et $b_\ell$, notés $\hat{a}_\ell$ et $\hat{b}_\ell$ vérifient

$$(\hat{a}_\ell, \hat{b}_\ell) = \arg \min_{(a,b)\in\mathbb{R}^2} \ell(speed, a \cdot dist + b),$$

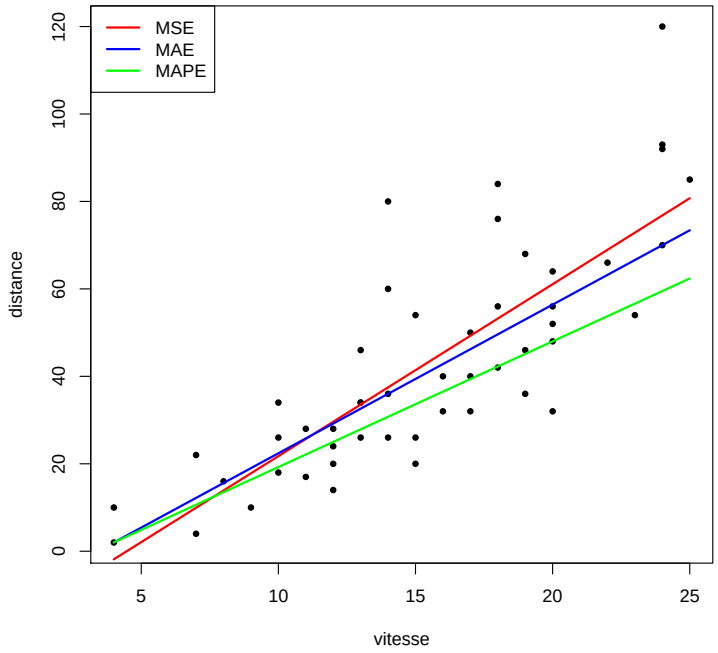



FIGURE 3.6 – Représentation graphique des différents modèles obtenus sur la base de données *cars* selon la fonction de perte utilisée.

où $\ell$ représente la fonction de perte utilisée pour calibrer le modèle, et le modèle associé à la fonction de perte est alors donné par

$$m_\ell(speed) = \hat{a}_\ell \cdot speed_i + \hat{b}_\ell.$$

Nous confronterons ici les résultats obtenus en utilisant trois fonctions de perte distinctes : MSE, MAE et MAPE.

**Comparaison des prédictions** La figure 3.6 propose une représentation graphique des données et des trois modèles distincts, dont les coefficients sont référencés dans la table 3.3.

On peut remarquer que les droites de régressions MAPE et MAE se coupent au point d'abscisse 4 et d'ordonnée 2. Ce point correspond au véhicule ayant la plus faible vitesse dans la base d'apprentissage. Par ailleurs, comme $b_{MAPE} < b_{MAE}$, les prévisions associées au modèle $m_{MAPE}$ sont toujours inférieures à celles associées au modèle $m_{MAE}$, ce qui est une conséquence du résultat démontré à la section 3.2.5.

| | $a_\ell$ | $b_\ell$ |
|---|---|---|
| $m_{MSE}$ | -17.579 | 3.932 |
| $m_{MAE}$ | -11.6 | 3.4 |
| $m_{MAPE}$ | -9.500 | 2.875 |

TABLE 3.3 – Valeurs numériques des coefficients $a_\ell, b_\ell$ associés à chaque modèle (MSE, MAE, MAPE).

| | | Modèle | | |
|---|---|---|---|---|
| | | $m_{MSE}$ | $m_{MAE}$ | $m_{MAPE}$ |
| Critère d'évaluation | $\ell_{MSE}$ | **227.07** | 239.77 | 325.03 |
| | $\ell_{MAE}$ | 11.58 | **11.28** | 12.37 |
| | $\ell_{MAPE}$ | 0.38 | 0.33 | **0.30** |

TABLE 3.4 – Résultats numériques associés aux différentes fonctions de perte sur la base de données *cars*.

**Comparaison des performances** Les résultats numériques sont présentés à la table 3.4. Cette table montre que la performance de chaque modèle dépend de la fonction de perte considérée. Plus précisément, si le critère d'évaluation est l'erreur relative (critère très souvent utilisé en pratique, pour sa facilité d'interprétation), le meilleur modèle est $m_{MAPE}$, et l'erreur relative moyenne est de 30%, alors qu'elle est de 33% pour $m_{MAE}$ et 38% pour $m_{MAPE}$. Ce résultat illustre bien l'intérêt de la MAPE en termes de qualité de prédiction lorsque le critère d'évaluation retenu est l'erreur relative.

En revanche, la table 3.4 montre également que l'utilisation des moindres carrés conduit au modèle le plus pertinent vis-à-vis des moindres carrés. De même la minimisation de l'erreur absolue conduit au modèle le plus pertinent vis-à-vis du critère $\ell_{MAE}$. Plus généralement, cette table illustre le fait que, pour un critère donné, l'utilisation d'une fonction de perte spécifique à ce critère conduit à un modèle plus performant.

Dans la section suivante, nous décrivons un cas d'utilisation des régressions MAPE chez Viadeo.

### 3.6.3 Cas pratique : modélisation de l'âge des membres du réseau social Viadeo

Afin d'illustrer la régression MAPE dans un contexte multivarié, nous proposons de modéliser l'année de naissance des membres de Viadeo en fonction de plusieurs variables explicatives : l'année d'inscription sur le site, l'âge moyen des contacts des membres, et

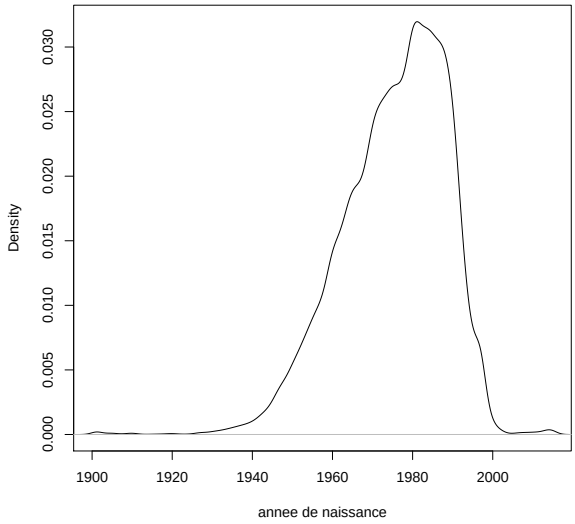



FIGURE 3.7 – Distribution de l'année de naissance des membres présents sur Viadeo (parmi ceux ayant renseigné l'information).

l'âge moyen des personnes portant le même prénom que les membres, d'après l'INSEE [4]. Un extrait de la base d'apprentissage est représenté à la table 3.5.

Ce sujet constitue une problématique importante chez Viadeo car l'année de naissance est renseignée pour 65% des membres seulement. Pour illustration, la figure 3.7 montre la distribution des années de naissance saisies par les membres. Cette information est néanmoins nécessaire pour améliorer la connaissance que Viadeo a de ses utilisateurs, et ainsi la performance de nombreux algorithmes (recommandation d'offres d'emploi, ciblage marketing, ...). Ceci permettra par exemple de prédire si un membre sera plus intéressé par une offre d'emploi destinée à un junior plutôt qu'une offre pour senior.

Pour améliorer la qualité des algorithmes utilisés chez Viadeo, il est donc pertinent d'inférer la date de naissance des membres n'ayant pas renseigné cette information.

Comme dans l'exemple précédent, afin de comparer les différents modèles nous supposons une relation linéaire entre l'âge des membres et les différentes variables explicatives. Les résultats numériques associés à chaque régression sont présentés dans la table 3.6.

On remarque à nouveau une cohérence entre la fonction de perte utilisée et le meilleur modèle associé à chaque critère d'évaluation. En particulier, vis-à-vis de l'erreur relative, qui est le critère retenu dans le cadre de ce modèle afin de faciliter la communication des résultats à l'ensemble des équipes, le modèle le plus performant est celui obtenu

4. voir `http://www.nosdonnees.fr/dataset?tags=insee`.

| Prénom | Année de naissance | Année d'inscription | Age moyen des contacts | Age moyen (selon INSEE) |
|---|---|---|---|---|
| dominique | 1976 | 2004 | 1972.69 | 1960.46 |
| david | 1977 | 2004 | 1971.51 | 1976.85 |
| yannick | NA | 2004 | 1978.78 | 1973.00 |
| diane | 1979 | 2004 | 1978.33 | 1981.88 |
| florence | NA | 2004 | 1977.11 | 1969.80 |
| olivier | 1976 | 2004 | 1949.00 | 1972.32 |

TABLE 3.5 – Extrait de la base de données utilisée pour modéliser l'âge des membres du réseau social Viadeo.

| | $m_{MSE}$ | $m_{MAE}$ | $m_{MAPE}$ |
|---|---|---|---|
| $\ell_{MSE}$ | **51.5671** | 52.1290 | 56.1162 |
| $\ell_{MAE}$ | 5.2900 | **5.2636** | 5.3796 |
| $\ell_{MAPE}$ | 0.1456 | 0.1434 | **0.1402** |

TABLE 3.6 – Comparaison des différents modèles obtenus pour l'inférence de l'âge des membres, selon la fonction de perte utilisée.

par la régression MAPE. Ceci illustre l'intérêt des régressions MAPE lorsque le critère utilisé pour mesurer la qualité d'un modèle est l'erreur relative moyenne.

## 3.7 Conclusion

Dans ce chapitre, nous avons vu que l'utilisation de la MAPE est possible pour la minimisation du risque empirique, et que l'estimateur obtenu est consistant sous certaines hypothèses. En particulier, nous avons vu que la consistance de l'estimateur est acquise si la densité de la variable à modéliser décroit suffisamment rapidement vers 0 en 0.

Par ailleurs, nous avons montré que la régression MAPE peut-être considérée comme un cas particulier des régressions médiane pondérées ; il est alors facile d'obtenir l'estimateur recherché à partir d'un algorithme de minimisation de l'erreur absolue. Les applications démontrent en effet la pertinence de l'approche proposée pour minimiser l'erreur relative.

Toutefois, pour certains modèles l'utilisation de pondération n'est pas simple. C'est par exemple le cas des modèles linéaires régularisés ou des modèles non paramétriques, qui permettent respectivement de limiter le sur-apprentissage et d'obtenir des modèles non linéaires. Nous aborderons ces modèles dans le chapitre suivant et détaillerons

en particulier l'impact de l'ajout de pondérations dans le problème d'optimisation consistant à minimiser l'erreur relative moyenne régularisée.

**Contributions scientifiques** Ce chapitre a fait l'objet des contributions et des communications scientifiques suivantes :

1. "Mean Absolute Percentage Error for Regression Models", Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi. Neurocomputing 2016.
2. "Using the Mean Absolute Percentage Error for Regression Models", Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi, Apr 2015, Bruges, Belgium. Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015).
3. "Consistance de la minimisation du risque empirique pour l'optimisation de l'erreur relative moyenne", 47èmes Journées de Statistique de la SFdS, Juin 2015,

# Chapitre 4

# Régression MAPE régularisée

## 4.1 Introduction

Dans le chapitre précédent, nous avons étudié le cas de la régression MAPE, à la fois d'un point de vue pratique et théorique. L'objectif de ce chapitre est d'étudier le cas de la régression MAPE régularisée. Dans le cadre de la régression régularisée, le problème de choix de modèle s'écrit de la façon suivante :

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, g(x_i)) + \lambda J(g)$$

où

— $J$ est le terme de régularisation. Il s'agit d'une fonction définie sur $\mathcal{G}$ et à valeurs dans $\mathbb{R}^+$, qui mesure la complexité du modèle $g$ : $J(g)$ est d'autant plus grand que le modèle $g$ est complexe.

— $\lambda$ règle le compromis entre l'attache aux données d'apprentissage et la régularisation. Il s'agit du poids accordé à la fonction de régularisation.

Les régressions régularisées permettent de pénaliser les modèles trop complexes au profit de modèles plus simples. L'intérêt majeur de ce type de modèles est d'augmenter la capacité de généralisation du modèle sur de nouvelles données, et ainsi limiter le sur-apprentissage.

Comme dans le chapitre précédent, nous comparerons les résultats obtenus dans le cadre de la MAPE à ceux obtenus avec la MAE, ou plus généralement avec la régression quantile. Dans le contexte de la régression quantile régularisée, la fonction de perte, aussi appelée *check-function* (Koenker and Bassett Jr, 1978), est définie, pour $\tau \in [0; 1]$, par

$$\rho_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau - 1)\xi & \text{sinon} \end{cases}$$

Cette fonction de perte est représentée à la figure 4.1.

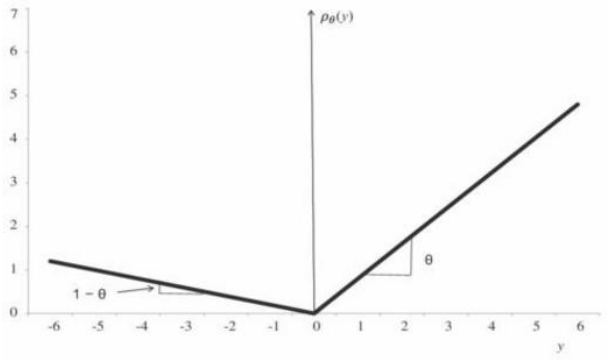



FIGURE 4.1 – Représentation de la *check-function*, aussi appelée *pinball loss*.

Étant donnée une famille de modèle $\mathcal{G}$, une fonction de régularisation $J$, un paramètre de régularisation $\lambda$, et $\tau$ le quantile pour lequel la fonction $g$ est optimisée, le problème de la régression quantile régularisée, traité par exemple dans Takeuchi *et al.* (2006); Li *et al.* (2007), est donc de trouver la fonction $g^\star \in \mathcal{G}$ qui vérifie

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda J(g), \tag{4.1}$$

Par exemple, le choix $\tau = 0.5$ conduit au cadre de la régression médiane. Dans le cadre de la régression MAPE régularisée, nous nous intéresserons au problème d'optimisation suivant :

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \left| \frac{y_i - g(x_i)}{y_i} \right| + \lambda J(g), \tag{4.2}$$

On note par ailleurs

$$\tilde{\ell}_\tau^{MAPE}(f, z_i) = \ell_\tau^{MAPE}(f, z_i) + \lambda J(f)$$

la fonction de perte régularisée.

La suite du chapitre est organisée en deux sections distinctes. D'abord, dans la section 4.2, nous détaillerons les questions pratiques liées à la régression MAPE régularisée. Nous verrons alors le problème d'optimisation associé à ce type de modèle, et comment le résoudre en pratique.

Puis, dans la section 4.3 nous verrons une application de la régression MAPE régularisée sur des données simulées.

## 4.2 Régression MAPE régularisée en pratique

### 4.2.1 Contexte

De façon similaire à Li *et al.* (2007), nous nous plaçons dans le cadre des régressions non paramétriques. Par rapport aux régressions classiques, les régressions non

paramétriques ont l'avantage d'explorer des relations non linéaires entre les variables explicatives $X$ et la variable cible $Y$.

Comme dans le chapitre précédent, l'objectif est de déterminer la meilleure fonction $g$ approximant la variable $Y$ conditionnellement à $X$, où $g$ appartient une famille $\mathcal{H}$ de fonctions prédéfinies. Dans le cadre des régressions non paramétriques, $\mathcal{H}$ est en fait un espace de Hilbert à noyau reproduisant (voir Rosipal and Trejo (2001) ou Berlinet and Thomas-Agnan (2011) par exemple), défini de la façon suivante :

**Définition 10** (Espace de Hilbert à noyau reproduisant)**.** *Soit $\mathcal{X}$ un espace arbitraire et $\mathcal{H}$ un espace de Hilbert de fonctions définies sur $\mathcal{X}$ et à valeurs dans $\mathbb{R}$.*

*On dit que $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant si, pour tout $x$ de $\mathcal{X}$, la forme linéaire $L_x$ définie sur $\mathcal{H}$ et à valeurs dans $\mathbb{R}$, telle que pour tout $f \in \mathcal{H}$ on a $L_x(f) = f(x)$, est continue sur $\mathcal{H}$.*

Dans la suite, nous noterons RKHS un espace de Hilbert à noyau reproduisant (en anglais *Reproducing Kernel Hilbert Space*). L'existence du noyau est alors assurée par le théorème de représentation de Riesz.

**Théorème 4** (Théorème de Riesz)**.** *Soit $\mathcal{H}$ est espace de Hilbert à noyau reproduisant, muni du produit scalaire $\langle ., . \rangle$.*

*Alors il existe une fonction $\phi \in \mathcal{H}$ telle que, pour tout $f \in \mathcal{H}$ il existe $\omega \in \mathcal{H}$ tel que $f(x) = \langle \phi(x), \omega \rangle$. Plus formellement :*

$$\exists \phi \in \mathcal{H} \text{ telle que } \forall f \in \mathcal{H}, \quad \exists \omega \in \mathcal{H} / \quad f(x) = \langle \phi(x), \omega \rangle.$$

La fonction noyau associée au RKHS $\mathcal{H}$ est alors donnée par

$$k(x, x') = \langle \phi(x), \phi(x') \rangle,$$

et pour un ensemble de données $x_1, \ldots, x_n$ on notera $K$ la matrice noyau dont le terme général est défini par $K_{i,j} = k(x_i, x_j)$.

Réciproquement, le théorème de Moore-Aronszajn (voir Aronszajn (1950)) permet d'assurer que tout noyau symétrique défini positif définit un unique espace de Hilbert à noyau reproduisant.

**Théorème 5** (Théorème de Moore-Aronszajn)**.** *Soit $K$ un noyau symétrique et défini positif sur un ensemble $\mathcal{E}$.*

*Alors il existe un unique espace de Hilbert $\mathcal{H}$ de fonctions sur $\mathcal{E}$ pour lequel $K$ est un noyau reproduisant.*

### 4.2.2 MAPE : problème primal

Dans le cadre des régressions régularisées, nous avons déjà vu que le problème d'optimisation s'écrit :

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \frac{\rho_\tau(y_i - f(x_i))}{y_i} + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2. \tag{4.3}$$

En utilisant les notations définies dans la section précédente et le théorème de représentation 4 on a que, pour tout $f \in \mathcal{H}$, il existe un $\omega \in \mathcal{H}$ et $b > 0$ tel que

$$f(x) = \langle \phi(x), w \rangle + b,$$

et en notant $\xi_i = y_i - f(x_i)$, $\xi_i^\star = 1 + f(x_i) - y_i$ et $C = \frac{1}{n\lambda}$, l'équation 4.3 peut se réécrire de la façon suivante :

$$\begin{array}{rl} \min\limits_{w,b,\xi_i,\xi_i^\star} & C\sum_{i=1}^m \frac{\tau\xi_i + (1-\tau)\xi_i^\star}{|y_i|} + \frac{1}{2}\|w\|^2, \\ \text{tels que} & y_i - \langle \phi(x_i), w \rangle - b \leq |y_i|\xi_i, \forall i, \\ & \langle \phi(x_i), w \rangle + b - y_i \leq |y_i|\xi_i^\star, \forall i, \\ & \xi_i \geq 0, \forall i, \\ & \xi_i^\star \geq 0, \forall i, \end{array} \tag{4.4}$$

Ce problème constitue le problème d'optimisation primal de la régression MAPE non paramétrique. La résolution de ce problème repose sur l'optimisation de nombreux paramètres : $\omega, b, \xi_i, \xi_i^\star$. Dans la section 4.2.3 nous proposons de calculer le problème d'optimisation dual, qui sera plus simple à résoudre.

### 4.2.3 MAPE : problème dual

Dans cette partie, nous montrons que le problème primal 4.4 est équivalent au problème dual suivant, où $K$ désigne la matrice noyau :

$$\begin{array}{rc} \max\limits_{\alpha} & \alpha^T y - \frac{1}{2}\alpha^T K \alpha \\ \text{tel que} & 1^T\alpha = 0 \\ & \forall i, \frac{C(\tau-1)}{|y_i|^2} \leq \alpha_i \leq \frac{C\tau}{|y_i|^2} \end{array} \tag{4.5}$$

La démonstration proposée s'effectue en plusieurs étapes. D'abord, nous calculerons le dual de Wolfe du problème d'optimisation 4.4. Ensuite, nous montrerons que, via plusieurs manipulations algébriques il est possible de passer du dual de Wolfe au problème 4.5, et réciproquement.

#### Dual de Wolfe

Soit $\theta = (w, b, \xi_i, \xi_i^\star)$ et $h$ et $g_{i,j}$ les fonctions définies par

$$\begin{array}{rcl} h(\theta) & = & C\sum\limits_{i=1}^m \frac{\tau\xi_i + (1-\tau)\xi_i^\star}{|y_i|} + \frac{1}{2}\|w\|^2, \\ \forall i, g_{i,1}(\theta) & = & y_i - \langle \phi(x_i), \omega \rangle - b - |y_i|\xi_i, \end{array}$$

$$\begin{aligned}
\forall i, g_{i,2}(\theta) &= \langle \phi(x_i), \omega \rangle + b - y_i - |y_i|\xi_i^\star, \\
\forall i, g_{i,3}(\theta) &= -\xi_i, \\
\forall i, g_{i,4}(\theta) &= -\xi_i.
\end{aligned}$$

Alors le problème dual de Wolfe au problème (4.4) est donné par

$$\begin{aligned}
\max_{u,\theta} \quad & h(\theta) + \textstyle\sum_{i=1}^m u_{i,1} g_{i,1}(\theta) + u_{i,2} g_{i,2}(\theta) + u_{i,3} g_{i,3}(\theta) + u_{i,4} g_{i,4}(\theta), \qquad (4.6) \\
\text{s. t.} \quad & \nabla f(\theta) + \textstyle\sum_{i=1}^m u_{i,1} \nabla g_{i,1}(\theta) + u_{i,2} \nabla g_{i,2}(\theta) + u_{i,3} \nabla g_{i,3}(\theta) + u_{i,4} \nabla g_{i,4}(\theta) = 0, \\
& u_{i,1}, u_{i,2}, u_{i,3}, u_{i,4} \geq 0, \forall i,
\end{aligned}$$

qui est, après quelques manipulations algébriques, équivalent au problème (4.7) :

$$\begin{aligned}
\max_{u,\theta} \quad & h(\theta) + \textstyle\sum_{i=1}^m u_{i,1} g_{i,1}(\theta) + u_{i,2} g_{i,2}(\theta) + u_{i,3} g_{i,3}(\theta) + u_{i,4} g_{i,4}(\theta), && (4.7) \\
\text{s. t.} \quad & w + \textstyle\sum_{i=1}^m (u_{i,1} - u_{i,1}) \phi(x_i) = 0, && (4.8) \\
& \textstyle\sum_{j=1}^m (u_{j,2} - u_{j,1}) = 0, && (4.9) \\
& \forall i, \tfrac{C\tau}{|y_i|} - |y_i| \cdot u_{i,1} - u_{i,3} = 0, && (4.10) \\
& \forall i, \tfrac{C(1-\tau)}{|y_i|} - |y_i| \cdot u_{i,2} - u_{i,4} = 0, && (4.11) \\
& \forall i, u_{i,1}, u_{i,2}, u_{i,3}, u_{i,4} \geq 0, \forall i. && (4.12)
\end{aligned}$$

Montrons maintenant que le problème 4.7 est équivalent au problème 4.5.

**Sens direct : de 4.7 à 4.5**

Ce problème peut être simplifié en introduisant la variable $\alpha_i = u_{i,1} - u_{i,2}$. Alors la valeur de $\omega$ est obtenue par la contrainte 4.8 et vaut $\omega = \sum_{i=1}^n \alpha_i \phi(x_i)$. Par ailleurs, la contrainte 4.9 s'écrit $1^T \alpha = 0$. En prenant en compte ces équations, la fonction à minimiser devient

$$\begin{aligned}
h(\theta) \quad + \quad & \sum_{i=1}^m u_{i,1} g_{i,1}(\theta) + u_{i,2} g_{i,2}(\theta) + u_{i,3} g_{i,3}(\theta) + u_{i,4} g_{i,4}(\theta) \\
= \quad & h(\theta) + \sum_{i=1}^n \alpha_i y_i - \|\omega\|^2 - \sum_{i=1}^n \xi_i (u_{i_1} |y_i| + u_{i,3}) - \sum_{i=1}^n \xi_i^\star (u_{i_2} |y_i| + u_{i,4}).
\end{aligned}$$

En utilisant les contraintes 4.10 et 4.11 les deux derniers termes se simplifient de la façon suivante :

$$\sum_{i=1}^n \xi_i (u_{i_1} |y_i| + u_{i,3}) + \sum_{i=1}^n \xi_i^\star (u_{i_2} |y_i| + u_{i,4}) = C \sum_{i=1}^n \frac{\tau \xi_i + (1-\tau) \xi_i^\star}{|y_i|}.$$

La fonction à minimiser s'écrit donc

$$\begin{aligned} h(\theta) &+ \sum_{i=1}^{m} u_{i,1} g_{i,1}(\theta) + u_{i,2} g_{i,2}(\theta) + u_{i,3} g_{i,3}(\theta) + u_{i,4} g_{i,4}(\theta), \\ &= \sum_{i=1}^{n} \alpha_i y_i - \frac{1}{2}\|\omega\|^2, \\ &= \sum_{i=1}^{n} \alpha_i y_i - \frac{1}{2}\alpha^T K \alpha. \end{aligned}$$

où $K_{ij} = k(x_i, x_j)$ est la matrice associée au noyau $k$. Ainsi, il est possible de réécrire la fonction à minimiser et les contraintes, de telle sorte qu'elles dépendent de $\alpha$ seulement.

**Réciproque : de 4.5 à 4.7**

Soit $\alpha_i$ un nombre réel. De façon évidente, il existe $u_{i,1} \geq 0$ et $u_{i,2} \geq 0$ tels que $\alpha_i = u_{i,1} - u_{i,2}$.

Alors, comme $\frac{C(\tau-1)}{|y_i|^2} \leq \alpha_i \leq \frac{C\tau}{|y_i|^2}$, *i.e.* $\frac{C(\tau-1)}{|y_i|} \leq |y_i|(u_{i,1} - u_{i,2}) \leq \frac{C\tau}{|y_i|}$, il existe $u_{i,3} \geq 0$ et $u_{i,4} \geq 0$ tels que

$$\forall i, \frac{C\tau}{|y_i|} - |y_i| \cdot u_{i,1} - u_{i,3} = 0 \quad \text{et} \quad \frac{C(1-\tau)}{|y_i|} - |y_i| \cdot u_{i,2} - u_{i,4} = 0.$$

De plus, en introduisant la variable $\omega = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, la réciproque est bien acquise. Ainsi, nous avons vu que le problème d'optimisation primal associé à la régression MAPE non paramétrique est équivalent au problème d'optimisation suivant :

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T y - \tfrac{1}{2}\alpha^T K \alpha \\ \text{tel que} \quad & 1^T \alpha = 0 \\ & \forall i, \tfrac{C(\tau-1)}{|y_i|^2} \leq \alpha_i \leq \tfrac{C\tau}{|y_i|^2} \end{aligned} \tag{4.13}$$

Dans Takeuchi *et al.* (2006), Takeuchi et al. montrent que, dans le cadre de la régression MAE non paramétrique, le problème d'optimisation est équivalent au suivant :

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T y - \tfrac{1}{2}\alpha^T K \alpha \\ \text{tel que} \quad & 1^T \alpha = 0 \\ & \forall i, C(\tau-1) \leq \alpha_i \leq C\tau \end{aligned} \tag{4.14}$$

Ces deux problèmes étant très proches, nous proposons de discuter de leurs spécificités dans la section suivante.

### 4.2.4 Comparaison des problèmes d'optimisation MAE et MAPE

En comparaison au problème (4.5), on peut remarquer que la modification de la fonction de perte (de l'erreur absolue à l'erreur relative) dans le problème primal revient à modifier l'espace d'optimisation dans le problème d'optimisation dual.

Plus précisément, ce changement de fonction de perte est équivalent à réduire (resp. augmenter) la taille de l'ensemble admissible de $\alpha_i$ si $y_i > 1$ (resp. $y_i < 1$).

De plus, en choisissant une très grande valeur pour $C$ (ou en faisant tendre $C$ vers l'infini), on peut s'assurer que les $\alpha_i$ optimaux, obtenus en résolvant le problème dual, convergent vers la même valeur pour la MAE et la MAPE. Ce constat, surprenant à première vue, est en fait équivalent à choisir une très petite valeur de $\lambda$ (ou de faire tendre $\lambda$ vers zéro), ce qui conduit à l'absence de terme de régularisation, et par conséquent à un fort sur-apprentissage des observations. Dans un tel cas de sur-apprentissage, $f(x_i) \simeq y_i$ et donc les valeurs des deux fonctions de perte sont équivalentes.

Remarquons que nous avons développé le problème d'optimisation dual dans le cas de la MAPE sans spécifier la valeur de $\tau$. Ainsi, le problème (4.5) s'applique de façon générale au cadre des *erreurs relatives quantiles*. L'efficacité de ces régressions ne sera cependant pas étudiée, et pour les simulations nous nous limiterons au cas où $\tau = 0.5$, ce qui correspond au cadre de la MAPE classique.

## 4.3 Applications

Dans cette section, nous illustrons la pertinence des régressions MAPE non paramétriques régularisées sur des données simulées, et nous comparons les résultats avec ceux obtenus par des régressions médianes à noyau.

### 4.3.1 Génération des données

Comme dans Takeuchi *et al.* (2006), nous avons simulé des données selon la fonction sinus cardinal, définie par

$$sinc(x) = \frac{sin(2\pi x)}{2\pi x}.$$

Cependant, pour illustrer la variation des prévisions en fonction de la proximité à zéro, nous avons ajouté un paramètre de translation, $a$, qui nous permet de définir la fonction sinus cardinal translatée, donnée par

$$sinc(x, a) = a + \frac{sin(2\pi x)}{2\pi x}.$$

Pour les expériences, nous avons engendré 1000 points qui représentent l'échantillon d'apprentissage, et 1000 autres points qui constituent un ensemble de test. Comme dans Takeuchi *et al.* (2006), le processus de génération des données est le suivant :

$$Y = sinc(X, a) + \epsilon(X),$$

avec $X \sim \mathcal{U}([-\infty;\infty])$ et $\epsilon(X) \sim \mathcal{N}\left(0, (0.1 \cdot \exp(1-X))^2\right)$.

Pour comparer les résultats entre l'estimation médiane et l'estimation MAPE, nous avons calculé $\widehat{f}_{MAPE,a}$ et $\widehat{f}_{MAE,a}$ pour plusieurs valeurs de $a$. Dans chaque cas, la valeur du paramètre de régularisation $C$ a été choisie par une validation croisée.

### 4.3.2 Résultats

| a | $MAPE(y, \widehat{f}_{MAE,a})$ (en %) | $MAPE(y, \widehat{f}_{MAPE,a})$ (en %) | $C_{MAE}$ | $C_{MAPE}$ |
|---|---|---|---|---|
| 0.00 | 128.62 | 94.09 | 0.01 | 0.10 |
| 0.10 | 187.78 | 100.10 | 0.05 | 0.01 |
| 0.50 | 72.27 | 57.47 | 5.00 | 10.00 |
| 1.00 | 51.39 | 39.53 | 10000.00 | 1.00 |
| 2.50 | 10.58 | 10.98 | 5.00 | 1.00 |
| 5.00 | 4.80 | 4.89 | 5.00 | 10.00 |
| 10.00 | 2.39 | 2.40 | 5.00 | 100.00 |
| 25.00 | 0.96 | 0.96 | 5.00 | 100000.00 |
| 50.00 | 0.48 | 0.48 | 5.00 | 1000.00 |
| 100.00 | 0.24 | 0.24 | 5.00 | 10000.00 |

Les résultats des expériences sont décrits à la table 4.3.2. Comme attendu, dans la plupart des cas l'erreur obtenue par $\widehat{f}_{MAPE,a}$ vis-à-vis de la MAPE est plus faible que celle obtenue par $\widehat{f}_{MAE,a}$. Ce constat est particulièrement vrai lorsque les valeurs de $y$ sont proches de zéro.

### 4.3.3 Illustration graphique

Quelques représentations graphiques de $\widehat{f}_{MAPE,a}$ et $\widehat{f}_{MAE,a}$ sont données à la figure 4.2. Cette figure permet d'illustrer plusieurs points intéressants :

— lorsque, pour une valeur de $x$, $y$ peut prendre des valeurs positives et négatives, la courbe rouge en trait plein (estimateur MAPE) est égale à 0, ce qui permet d'assurer une erreur relative de 100%, tandis que la courbe bleue en pointillés conduit à une erreur relative beaucoup plus importante ;

— la courbe associée à l'estimation médiane semble identique, à une translation près, quelle que soit la valeur de $a$, tandis que la courbe associée à l'estimation MAPE dépend très fortement de la proximité à 0. Cela s'explique par le fait que l'erreur absolue est inchangée par translation, ce qui n'est pas le cas de la MAPE ;

— la courbe associée à l'estimation MAPE est plus proche de 0 que la courbe bleue. Ceci est une conséquence du fait que la valeur absolue de l'estimateur MAPE

est toujours inférieure à la valeur absolue de la médiane, comme nous l'avons montré à la section 3.2.5.

## 4.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la régression MAPE non paramétrique. Après avoir exprimé le problème d'optimisation dans sa version primale puis duale, nous avons montré que le passage de la fonction de perte MAE à la fonction de perte MAPE se traduit par une modification des contraintes du problème dual.

Comme dans le chapitre précédent, nous avons donc vu que la résolution la régression MAPE non paramétrique possède de fortes similitudes avec le problème de la régression MAE non paramétrique.

Les résultats expérimentaux montrent la pertinence de l'approche présentée pour minimiser l'erreur relative moyenne régularisée. Par ailleurs, les résultats ont permis de mettre en évidence les différences entre les estimateurs obtenus par minimisation de l'erreur relative et de l'erreur absolue. En particulier, nous avons pu remarquer que les comportements divergent très fortement lorsque la variable cible peut prendre des valeurs proches de 0.

**Contributions scientifiques** Ce chapitre a fait l'objet des contributions et des communications scientifiques suivantes :

1. "Mean Absolute Percentage Error for Regression Models", Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi. Neurocomputing 2016.
2. "Mean Absolute Percentage Error : Minimization and Consistency of the ERM", présentation au séminaire du SAMM, université Paris 1 Panthéon-Sorbonne, décembre 2015.

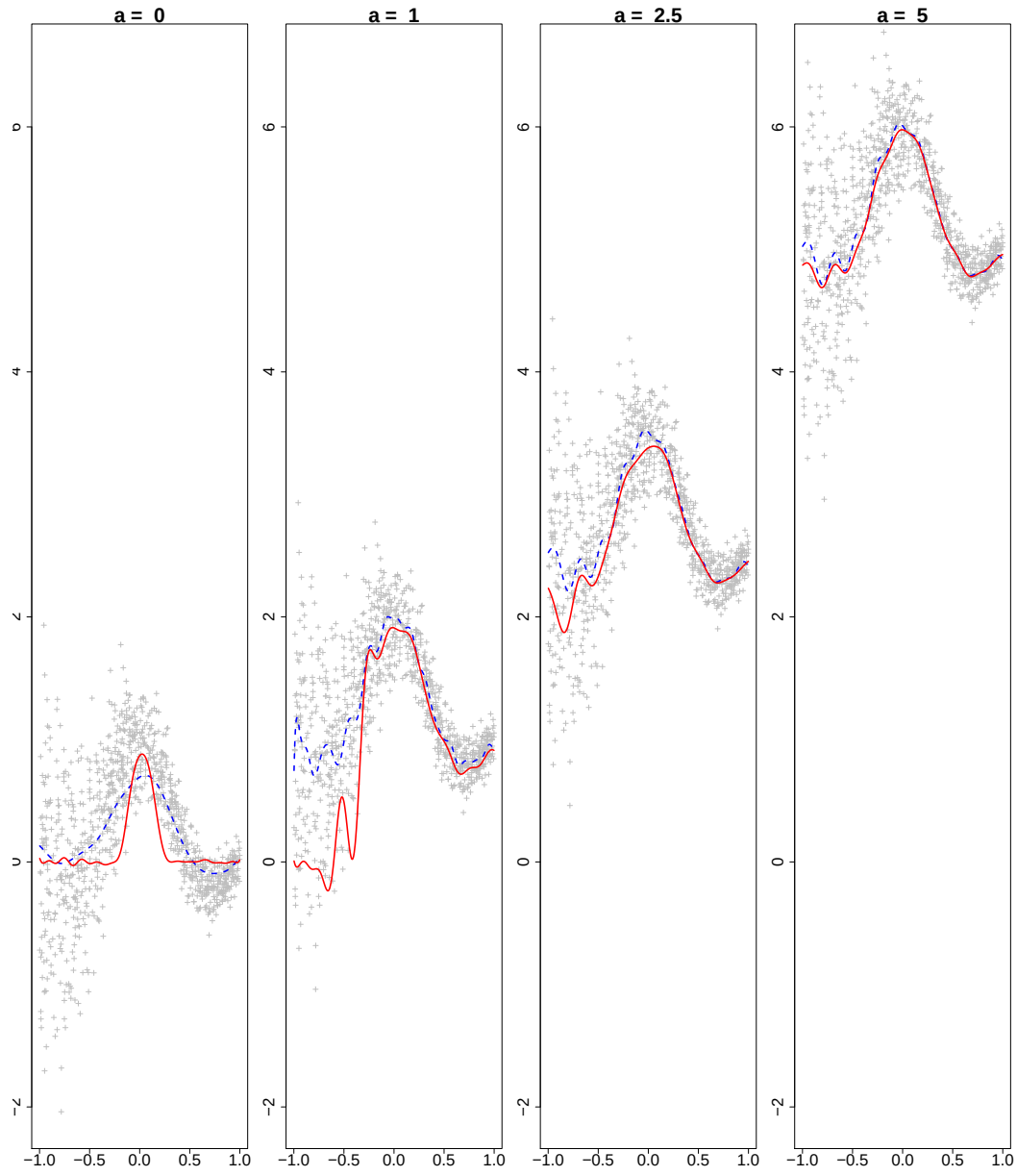



FIGURE 4.2 – Représentation graphique des modèles $\widehat{f}_{MAE,a}$ (en pointillés bleu) et $\widehat{f}_{MAPE,a}$ (en trait plein rouge).

# Conclusion et perspectives

Dans cette thèse, nous avons vu que l'évaluation d'un modèle joue un rôle très important dans l'élaboration de ce dernier, et que la phase d'évaluation peut être biaisée du fait de la qualité des données, notamment lorsque les bases d'apprentissage et de test possèdent des propriétés différentes. On parle alors de biais, ou *shift* en anglais. A l'aide de données réelles issues du réseau social professionnel Viadeo, nous avons mis en évidence le biais engendré par la méthodologie CRISP-DM (Wirth and Hipp, 2000) et nous avons proposé une nouvelle procédure d'évaluation hors-ligne d'un algorithme de recommandation permettant de réduire ce biais. Puis, nous avons vu de façon plus générale que plusieurs approches permettent de calibrer un modèle sur une base d'apprentissage biaisée tout en étant capable d'avoir une bonne capacité de prédiction sur une base de données test possédant des propriétés différentes de l'échantillon d'apprentissage. Deux solutions sont en effet souvent évoquées dans la littérature en fonction des hypothèses régissant le biais : d'une part la théorie du *covariate shift* (Shimodaira, 2000), et d'autre part les modèles d'adaptation de domaines (Villani, 2008).

Dans nos travaux, nous avons montré que ces approches dépendent d'hypothèses qui ne sont pas toujours vérifiées en pratique, et nous avons formalisé des hypothèses permettant de traiter un nouveau type de biais, appelé *explanatory shift*, portant sur la relation entre la variable cible et les variables explicatives, $P(Y|X)$. Nous avons alors proposé une nouvelle approche, qui permet de calibrer un modèle de classification de façon itérative dans le cas où la quantité $P(X|Y)$ est conservée entre les bases d'apprentissage et de test. Nous avons par ailleurs montré que, sous ces hypothèses, la fonction d'évaluation optimale fait intervenir des poids définis comme le ratio entre les densités de la variable cible sur les bases de test et d'apprentissage. Les expériences réalisées sur des données réelles issues du site de e-commerce *Cdiscount* et du jeu de données *Newsgroup* illustrent la pertinence de l'approche proposée dans le cas des modèles de classification. Une évolution naturelle de ces travaux à moyen terme pourrait être la généralisation de l'approche au cas des modèles de régression.

Ensuite, nous nous sommes intéressés au cas de l'erreur relative moyenne (MAPE, *Mean Absolute Percentage Error*) pour les modèles de régressions, qui peut être considérée

comme une version pondérée de l'erreur absolue. Dans ce cas, en notant $y$ la valeur de la variable cible $Y$, l'expression des poids est donnée par $1/|y|$, qui diverge lorsque $y$ tend vers 0. L'étude approfondie de cette fonction de perte n'avait pas été réalisée à notre connaissance dans la littérature. Nous nous sommes donc intéressés aux problèmes théoriques liés à l'utilisation de l'erreur relative moyenne comme fonction de perte pour les modèles de régression, comme cela avait déjà été réalisé pour des fonctions de perte plus classiques (moindres carrés et erreur absolue principalement, Devroye *et al.* (1996)).

Après avoir établi l'existence du risque empirique sous certaines hypothèses, nous avons démontré la consistance de l'estimateur obtenu par minimisation du risque empirique dans le cas où la valeur absolue de la variable cible est inférieurement bornée par un réel strictement positif. Ces résultats s'inscrivent dans la continuité de ceux présentés dans la littérature dans le cas des moindres carrés et de l'erreur absolue et permettent de faire une forte analogie entre les différentes fonctions de perte. A court terme, une évolution de nos travaux pourrait donc consister à assouplir l'hypothèse d'existence d'une borne inférieure sur la variable cible, afin de généraliser les résultats obtenus dans le cas où la variable cible peut s'approcher autant que possible de la valeur nulle. Une généralisation très similaire a déjà été obtenue dans la littérature (voir Devroye *et al.* (1996) par exemple) dans les cas des moindres carrés et de l'erreur absolue.

Enfin, nous avons détaillé les modèles de régression MAPE dans un cadre non paramétrique. Nous avons alors obtenu une expression explicite du problème d'optimisation associé à la régression MAPE non paramétrique ainsi que de sa version duale. Les résultats numériques nous ont permis de mettre en avant les spécificités de ces régressions, ainsi que les problèmes d'estimation liés à la proximité entre la variable cible et la valeur nulle. Une prolongation de ce chapitre à moyen terme pourrait être la preuve de la consistance de l'estimateur obtenu par résolution du problème dual dans le cas de la régression MAPE non paramétrique, déjà établie dans le cas de la régression non paramétrique minimisant l'erreur absolue (Li *et al.*, 2007).

# Publications et communications

1. “Reducing offline evaluation bias of collaborative filtering algorithms.”, A. De Myttenaere, Boris Golden, B. Le Grand, F. Rossi, Apr 2015, Bruges, Belgium. Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015).
2. ”Study of a bias in the offline evaluation of a recommendation algorithm.”, A. De Myttenaere, Boris Golden, B. Le Grand, F. Rossi. 11th Industrial Conference on Data Mining, ICDM 2015, Jul 2015, Hamburg, Germany. Ibai Publishing, pp.57-70, 2015, Advances in Data Mining.
3. ”Reducing offline evaluation bias in recommendation systems”, A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, 23rd annual Belgium-Dutch Conference on Machine Learning (Benelearn 2014), Bruxelles, juin 2014
4. ”Reducing offline evaluation bias in recommendation systems”, A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, séminaire du SAMM, université Paris 1 Panthéon Sorbonne, décembre 2014
5. “Supervised Classification under explanatory shift”, A. De Myttenaere, B. Le Grand, F. Rossi, Apr 2016, Vannes, France. Statlearn 2016.
6. ”Mean Absolute Percentage Error for Regression Models”, A. de Myttenaere, Boris Golden, B. Le Grand, F. Rossi. Neurocomputing 2016.
7. ”Using the Mean Absolute Percentage Error for Regression Models”, A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, Apr 2015, Bruges, Belgium. Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015).
8. ”Consistance de la minimisation du risque empirique pour l’optimisation de l’erreur relative moyenne”, 47èmes Journées de Statistique de la SFdS, Juin 2015.
9. "Mean Absolute Percentage Error : Minimization and Consistency of the ERM", présentation au séminaire du SAMM, université Paris 1 Panthéon-Sorbonne, décembre 2015.

# Références bibliographiques


G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6) :734–749, 2005.

M. Althoff, O. Stursberg, and M. Buss. Model-based probabilistic collision detection in autonomous driving. *Intelligent Transportation Systems, IEEE Transactions on*, 10(2) :299–310, 2009.

T. Anderson. *The theory and practice of online learning*. Athabasca University Press, 2008.

M. Anthony and P. L. Bartlett. *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, 1999.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3) :337–404, 1950.

J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 7–14. ACM, 2013.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

C. M. Bishop. Pattern recognition. *Machine Learning*, 2006.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Caphyon. Google organic ctr study. *Advanced Web Ranking*, 2014.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

M. Cuturi. Sinkhorn distances : Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Dataiku. Winning Kaggle, An introduction to Re-Ranking. `http://www.dataiku.com/blog/2014/01/14/winning-kaggle.html`, January 2014.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 21 of *Applications of Mathematics*. Springer, 1996.

E. F. Fama and M. H. Miller. *The theory of finance*, volume 3. Dryden Press Hinsdale, IL, 1972.

M. A. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3) :381–396, 2002.

H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of applied meteorology*, 11(8) :1203–1211, 1972.

R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition : An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006. IEEE, 2011.

L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

Hong. Unbiased Offline Evaluation of A/B Testing. `https://www.linkedin.com/today/post/article/20140423182505-20244634-unbiased-offline-evaluation-of-a-b-testing`, April 2014.

J. Rennie. 20 Newsgroups. `http://qwone.com/~jason/20Newsgroups/`, January 2008.

L. Kantorovitch. On the translocation of masses. *Management Science*, 5(1) :1–4, 1958.

P. A. Knight. The sinkhorn-knopp algorithm : Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1) :261–275, 2008.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica : journal of the Econometric Society*, pages 33–50, 1978.

R. Koenker. quantreg : Quantile regression. r package version 5.05, 2013.

L. Kohavi. *Online Controlled Experiments and A/B Tests*. Springer, 2015.

K. Lang. Newsweeder : Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

Langford. Vowpal Wabbit. `https://github.com/JohnLangford/vowpal_wabbit/wiki`, 2015.

Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, 102(477) :255–268, 2007.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

J. Mary, P. Preux, and O. Nicol. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 172–180, 2014.

G. J. McLachlan and K. E. Basford. Mixture models. inference and applications to clustering. *Statistics : Textbooks and Monographs, New York : Dekker, 1988*, 1, 1988.

G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough : how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101. ACM, 2006.

J. Michalakes, S. Chen, J. Dudhia, L. Hart, J. Klemp, J. Middlecoff, and W. Skamarock. Development of a next generation regional weather research and forecast model. In *Developments in Teracomputing : Proceedings of the Ninth ECMWF Workshop on the use of high performance computing in meteorology*, volume 1, pages 269–276. World Scientific, 2001.

J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10) :1345–1359, 2010.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2) :199–210, 2011.

D. H. Park, H. K. Kim, I. Y. Choi, and J. K. Kim. A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11) :10059–10072, 2012.

B. Pradel. *Evaluation des systèmes de recommandation à partir d'historiques de données*. PhD thesis, UPMC, 2013.

R. Rosipal and L. J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, 2(Dec) :97–123, 2001.

J.-B. Rudelle. *On m'avait dit que c'était impossible : Le manifeste du fondateur de Criteo*. Stock, 2015.

A. Said, B. Fields, B. J. Jain, and S. Albayrak. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1399–1408. ACM, 2013.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel. Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*, 1(1) :4–23, 2010.

G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

B. Shapira. *Recommender systems handbook*. Springer, 2011.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2) :227–244, 2000.

X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009 :4, 2009.

I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7 :1231–1264, 2006.

L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

S. J. Taylor. Modelling financial time series. 2007.

C. Villani. *Optimal transport : old and new*, volume 338. Springer Science & Business Media, 2008.

R. Wirth and J. Hipp. Crisp-dm : Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Citeseer, 2000.