UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR

THÈSE DE DOCTORAT UNIQUE

Année : 2018 - 2019 No d'ordre : 135

École Doctorale Mathématiques et Informatique

Mention

Mathématiques et Modélisation

Spécialité

Analyse, Statistiques et Applications

Présentée par

Badiassiatta Don Bosco DIATTA

SUJET :

Modélisation Bayésienne pour la Factorisation de Matrices Positives : Cas du Modèle Poisson-Gamma et Application en Reconstruction Automatique du Questionnaire Médical HSOPSC

Soutenue le 06 juillet 2019 devant le jury composé de :

Président :	M. Diaraf Seck	Professeur, UCAD
Examinateurs :	M. Gabriel Birame Ndiaye	Professeur, UCAD
	M. Abdoulaye Sène	Professeur, UCAD
Rapporteurs :	Mme. Sophie Dabo Niang	Professeur, Université de Lille
	M. Aliou Diop	Professeur, UGB
Directeurs de Thèse :	M. Papa Ngom	Maître de Conférences, UCAD
	M. Olivier François	Professeur, Ensimag Grenoble INP

Dédicaces

A toutes les personnes qui me sont chères je dédie cette Thèse,

A ma mère Jacqueline Diatta, à mon père Bernard,

A mes frères et sœurs, Romuald, Médard, Faguy, Béatrice et Clarisse,

A tous mes cousins et cousines, mention spéciale à toi

Prosper j'ai encore la nostalgie de nos pauses café à la BU,

A toi Élisabeth Basse j'exprime toute mon affection.

Hommage à toi Grand-mère Aourio, tu n'es certes plus
parmi nous, mais l'aboutissement de ce travail
est aussi le résultat de tes nombreuses
prières pour la réussite de tes
petits-enfants. Que ton
âme repose en paix.
Amen.

Tout obstacle renforce la détermination. Celui qui s'est fixé un but n'en change pas.

– Léonard De Vinci.

Remerciements

Je souhaite ici rendre hommage et exprimer ma profonde gratitude à tous ceux qui, de près ou de loin, ont contribué à la réalisation de cette thèse de Doctorat.

Mes remerciements s'adressent tout d'abord à mes deux directeurs de thèse que sont M. Papa Ngom, Maître de Conférences, Université Cheikh Anta Diop (UCAD, Dakar - Sénégal), et M. Olivier François, Professeur, Ensimag Grenoble INP (France) qui ont bien voulu accepter la direction scientifique de mes travaux, et dont les thèmes de recherches ont fortement inspiré cette thèse. Je leur suis reconnaissant pour leur compétence, rigueur intellectuelle, dynamisme et maitrise du sujet. Je vous remercie de vos encouragements tout au long du travail. Soyez assurés de mon attachement et de ma profonde gratitude.

A M. Papa Ngom, qui était également mon encadreur en Master 2 de Probabilités-Statistiques c'est le lieu ici de lui rendre un hommage appuyé pour la confiance qu'il a accordée à ma modeste personne. Alors jeune étudiant titulaire d'une Maîtrise en Mathématiques Appliquées Informatique et Finances (MAIF) à l'Université Gaston Berger de Saint-Louis (UGB), j'ai débarqué à l'Université Cheikh Anta Diop afin de poursuivre ma formation. L'option Probabilités-Statistiques a été sans doute le fruit de l'influence de ma trajectoire à l'UGB. Après le Master 2, M. Papa Ngom a bien voulu se lancer avec moi dans cette nouvelle aventure pour ne pas dire ce défit qu'est la Thèse.

A M. Olivier François je le remercie d'avoir permis la réalisation en 2016 d'un séjour de recherche au laboratoire TIMC-IMAG de l'Université Grenoble-Alpes, au sein de l'équipe BCM (Biologie Computationnelle et Mathématique). Je le remercie vivement de m'avoir obtenu un financement du LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) pour les frais de voyage aller-retour l'hébergement et la prise en charge. Par la même occasion je remercie toute l'équipe pour l'accueil chaleureux surtout Hélèna Martins avec qui je partageais le même bureau.

Je remercie également tous membres du laboratoire LMA (Laboratoire de Mathématiques Appliquées, UCAD) et ceux du projet NLAGA (Non Linear Analysis Geometry and Application) auquel j'appartiens, particulièrement son Responsable Professeur Diaraf Seck pour sa disponibilité, ses encouragements et ses précieux conseils de chercheur avisé qu'il ne cesse de nous prodiguer. Je remercie par la même occasion le projet NLAGA pour sa contribution financière à mon séjour de recherche à Grenoble.

Je suis très honoré de la présence à mon jury de thèse des membres suivants. Je tiens à leur assurer de ma profonde reconnaissance pour l'intérêt qu'ils portent à ce travail. Ainsi je tiens à remercier :

Mme Sophie Dabo NIANG, Professeur des Universités, au laboratoire LEM (Lille Économie et Management) de l'Université de Lille, pour l'honneur qu'elle m'a fait pour sa participation à mon jury de thèse en qualité de rapporteur de mon travail, pour le temps consacré à la lecture de cette thèse, et pour les suggestions et les remarques judicieuses qu'elle m'a indiquées.

M. Aliou DIOP, Professeur titulaire des Universités, au laboratoire LERSTAD (Laboratoire d'Etudes et de Recherches en Statistiques et Développement) de l'Université Gaston Berger, pour sa participation à mon jury de thèse en qualité de rapporteur de mon travail et pour toutes les remarques intéressantes qu'il m'a faites.

M. Diaraf SECK, Professeur titulaire des Universités, au laboratoire LMDAN (Laboratoire de Mathématiques de la Décision et d'Analyse Numérique) de l'Université Cheikh Anta Diop, pour l'honneur qu'il m'a fait en acceptant d'être membre de mon jury de thèse en sa qualité de président de jury.

M. Abdoulaye SENE, Professeur titulaire, au laboratoire LMA (Laboratoire de Mathématiques Appliquées) de l'Université Cheikh Anta Diop, pour sa participation à mon jury de thèse en qualité d'examinateur.

M. Gabriel Birame NDIAYE, Professeur titulaire et Directeur du laboratoire LMA de l'Université Cheikh Anta Diop, pour sa participation à mon jury de thèse en qualité d'examinateur.

Enfin je remercie mon père et ma mère qui m'ont toujours soutenu durant tout mon cursus universitaire, en particulier pendant les années de thèse qui correspondent aussi à beaucoup de sacrifices, merci pour votre soutien à tout point de vue et votre patience, sans lesquels je n'aurais pas pu travailler de façon sereine ou tout simplement je n'en serais pas là aujourd'hui.

Dans cette même vague de remerciements j'associe mes oncles Jonas Sambou et Madame ainsi que Laurent Diatta et Madame qui furent respectivement mes tuteurs à Saint-Louis pendant mes quatre années à l'Université Gaston Berger et à Dakar lorsque je venais poursuivre ma formation en Master 2 à l'Université Cheikh Anta Diop.

Résumé

Le phénomène des données manquantes est sujet actuel d'autant plus que sa récurrence est notée dans les questionnaires qui sont un outil essentiel de collecte de données. En santé publique les questionnaires sont des instruments d'auto-évaluation de plus en plus adoptés. En effet en 2004 l'agence américaine en santé publique AHRQ (Agency for Healthcare Research and Quality) a développé un célèbre questionnaire, nommé HSOPSC (Hospital Survey on Patient Safety Culture), destiné a améliorer la qualité de la prise en charge des patients dans les structures américaines publiques de santé et consistant à évaluer la culture de sécurité du patient du point de vue du personnel de ces établissements. Ce questionnaire a par la suite été adopté par plusieurs autres pays. Cependant de par sa longueur, le HSOPSC souffre d'un problème d'acceptabilité de la part de la cible et pourrait alors conduire à un fort taux données manquantes. Ce volume du questionnaire (42 items) allonge souvent la durée des enquêtes (19 mois pour le questionnaire HSOPSC réalisé en France entre 2013 et 2014), et dans certains cas des coûts pourraient être induits.

L'objectif général de notre travail est de reconstruire automatiquement un questionnaire HSOPSC dont nous avons artificiellement retiré certaines réponses suivant un mécanisme aléatoire. Ses motivations résident à deux niveaux, d'abord répondre efficacement à un problème de non-réponses qui se poserait pour d'autres questionnaires du même type, ensuite proposer une anticipation sur le problème d'acceptabilité du questionnaire mais aussi sur d'éventuels coûts ou durées des enquêtes. Pour ce second niveau, le questionnaire serait alors réduit au départ avant d'être soumis. Ce sous-échantillonnage du questionnaire vise à augmenter ses chances d'acceptabilité mais aussi à réduire les durées et éventuels coûts des enquêtes.

Nous proposons alors un modèle bayésien de factorisation de matrices positives. Le modèle en question est de type poisson-gamma, c'est-à-dire une vraisemblance de loi de poisson pour les données et des lois a priori gamma pour les paramètres. L'algorithme de reconstruction automatique que nous développons ici repose alors sur la procédure d'inférence sur les paramètres qui un échantillonneur de Gibbs. L'originalité de la contribution que nous proposons dans cette thèse réside dans le fait qu'un tel modèle n'a, à notre connaissance, jamais été utilisé auparavant pour une tache d'imputation sur des données de questionnaires, et en particulier ceux de type HSOPSC. Au delà de la particularité de notre approche, son intérêt est surtout de proposer un algorithme d'imputation avec des performances acceptables ou même supérieures à celles des principales méthodes récentes aussi basées sur le principe de factorisation matricielle.

Pour mener à bien ce travail, nous donnons d'abord quelques préliminaires résumés en deux chapitres. Le premier étudie quelques aspects essentiels de la modélisation bayésienne. Le second étudie la factorisation de matrices positives. Ensuite nous présentons la problématique des non-réponses et l'état de l'art au troisième chapitre, avant de donner la méthodologie que nous proposons au quatrième chapitre. Les performances de notre algorithme de reconstruction sont comparées à celles de plusieurs autres méthodes récemment développées : il s'agit de procédures d'analyses factorielles ou d'apprentissage automatique. Deux critères sont retenus pour évaluer la perte de l'information : la divergence Kullback-Leibler qui compare les histogrammes reconstruits et ceux originaux, et le RMSE (root mean square error) qui décrit la racine carrée de l'erreur quadratique moyenne. Si pour le premier critère notre algorithme donne des résultats acceptables jusqu'à 20% de réponses retirées, pour le second il présente les meilleures performances à partir

de 35% de taux de non-réponses.

Mots-clés : modélisation bayésienne, factorisation de matrices positives, modèle poisson-gamma, questionnaire, données manquantes, algorithme, échantillonnage de Gibbs, imputation, reconstruction automatique.

Abstract

The phenomenon of missing data is an important issue, especially in medical survey questionnaires, which are an essential tool for data collection or instruments of self-evaluation in public health. In 2004, the American agency for public health AHRQ (Agency for Healthcare Research and Quality) developed a survey questionnaire called HSOPSC (Hospital Survey on Patient Safety Culture), designed to improve the quality of patient care in US public health structures by assessing safety culture from the point of view of the personnel of these institutions. This survey was subsequently adopted by several others countries. Because of its length, the HSOPSC suffers from a problem of acceptability on the part of the medical workers and could lead to a high rate of missing data. In addition, the volume of the questionnaire lengthens the duration of the survey (19 months for the HSOPSC survey conducted in France between 2013 and 2014), and in some cases costs could be incurred.

The overall goal of our work is to automatically reconstruct an HSOPSC survey from which we have artificially removed some responses using a random mechanism. Its motivations reside on two levels. The first motivation is to answer to the problem of missing data which could arise for others surveys of the same type. The second motivation is to anticipate on the problem of acceptability of the survey and possible costs or durations of investigations, by reducing the length of the questionnaire. For this second level, the questionnaire would then be reduced initially before being submitted by subsampling specific sets of questions for each participant. The sub-sampling of the questionnaire aims to increase its chances of acceptability but also to reduce the durations and possible costs of investigations.

In this thesis, we propose a Bayesian model of factorization for nonnegative matrices. Our model is of Poisson-Gamma type, namely, assumes a Poisson distribution for the data and Gamma prior distributions for the parameters. The reconstruction algorithm developed here is based on the Gibbs sampler inference procedure for the parameters. The originality of our contribution is that such a model has, to our knowledge, never been used before for an imputation task on medical survey data, and in particular HSOPSC. Beyond the peculiarity of our approach, its main interest is to propose an imputation algorithm with acceptable or even higher performances than recent methods also based on the principle of matrix factorization.

To present this work, we first give some preliminaries summarized in two chapters. The first chapter examines some key aspects of Bayesian modeling. The second chapter provides details on nonnegative matrix factorization. In the third chapter, we introduce the problem of missing data and the state-of-the-art methods before presenting our new methodology in the fourth chapter. The performances of our reconstruction algorithm are compared with those of several others recently developed methods developed for factorial analysis or machine learning procedures. Two criteria are used to evaluate the loss of information : the Kullback-Leibler divergence comparing reconstructed and original histograms, and the root mean square error (RMSE). For the first criterion our algorithm gives acceptable results up to 20% of removed responses, for the second it presents the best performance from a rate of 35% of non-responses.

Keywords : Bayesian modeling, nonnegative matrix factorization, Poisson-Gamma model, survey questionnaire, missing data, algorithm, Gibbs sampling, imputation, automatic reconstruction.

Cette Thèse a été effectuée dans le cadre du projet NLAGA (Non Linear Analysis, Geometry and Applications) et réalisée aux laboratoires :

. Laboratoire LMA de (Laboratoire de Mathématiques Appliquées) de la Faculté des Sciences et Techniques de l'Université Cheikh Anta Diop.

. Laboratoire TIMC-IMAG (Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble).

Table des matières

Introduction générale

Lors de l'étude d'un phénomène, comme par exemple la cartographie d'une épidémie, une enquête sur le niveau de vie d'une population, ou un questionnaire auto-administré au sein d'une structure, la collecte de données est une phase cruciale qui nécessite une bonne méthodologie afin d'éviter une perte de données qui biaiserait les analyses statistiques. Toutefois, malgré ces précautions prises des cas de pertes de données sont notés dans la pratique. A côté de la perte de données, il y a d'autres situations causant l'absence de données. C'est le cas, par exemple, d'informations inexploitables considérées comme des données manquantes. Pour ce qui concerne les questionnaires d'enquêtes les données manquantes désignent les non-réponses. Alors plusieurs situations peuvent expliquer l'apparition des non-réponses dans des données d'enquêtes. En effet il peut s'agir pour celles-ci, dans certains cas, d'un refus de la part des répondants de renseigner sur certains points du questionnaire du fait de leur caractère sensible, ou, dans d'autres cas, d'une incompréhension de certaines questions posées. Par ailleurs d'autres contextes peuvent également expliquer les cas de non-réponses. Il s'agit des cas de non-réponses voulues ou contrôlées par les enquêteurs (décideurs, responsables de structures ...). Ceux-ci peuvent alors décider de réduire délibérément les questionnaires avant de les soumettre. Dans ce cas les non-réponses seraient donc majoritairement dues aux questions (retirées) non soumises. Deux motivations expliquent en général cette situation, d'une part le fait de vouloir augmenter les chances d'acceptabilité du questionnaire, d'autre part anticiper sur des contraintes de coûts et de durées allongées des enquêtes.

Le problème des données manquantes est donc récurrent et rencontré dans différents domaines des sciences fondamentales ou sociales, comme par exemple en épidémiologie, économie, biologie, génétique, climatologie, sociologie, ou encore rencontré dans le cadre des questionnaires d'enquêtes. Ces derniers font de plus en plus l'objet d'étude de la part des structures publiques de santé dans différents pays. C'est pourquoi la problématique des données manquantes est un sujet actuel et qui depuis longtemps a été un domaine de recherche sur lequel se sont intéressé plusieurs auteurs. En effet dès 1932 l'auteur Wilks a déjà proposé une modélisation de la présence des valeurs manquantes dans des données multivariées à l'aide d'une distribution bivariée normale (Wilks, 1932 [1]). Puis l'automatisation de ce type d'analyse a connu un tournant avec les études de Healy et Westmacott (1956) [2]. Les auteurs ont décrit une technique générale pour traiter les observations manquantes, applicable à toutes les analyses impliquant des estimations par la méthode des moindres carrées. Ensuite Anderson a proposé un modèle bivarié gaussien pour l'estimation de caractéristiques telles les moyennes, les variance ou encore les corrélations en présence de données manquantes, à l'aide de la méthode du maximum de vraisemblance (Anderson, 1957 [3]. Wilkinson en 1958 a déjà proposé une méthode d'estimation des valeurs manquantes par le biais d'équations où les inconnues sont celles-ci (Wilkinson, 1958 [4]). De leur côté Trawinski et Bargmann ont proposé une méthode d'estimation de paramètres, en présence de données manquantes par la méthode du maximum de vraisemblance (Trawinski et Bargmann, 1964 [5]). Les auteurs Afifi et Elashoff ont effectué une large revue de la littérature de données manquantes multivariées et ont montré comment certaines études déjà réalisées pourraient être simplifiées si la structure des données manquantes était bien identifiée et correspondait à certains modèles bien connus, par exemple la structure univariée où l'absence des données concerne une seule variable (Affi et Elashoff, 1966 [6]). Hartley et Hocking ont développé des méthodes plus générales donc adaptées à une structure quelconque et basée sur des modèles probabilistes avec utilisation d'une vraisemblance avec des techniques d'estimation s'appuyant sur le principe du

maximum de vraisemblance (Hartley et Hocking, 1971 [7]). Orchard et Woodbury ont également proposé des procédures d'imputation basées sur l'usage de la vraisemblance des données (Orchard et Woodbury, 1972 [8]).

Ainsi nous pouvons nous rendre compte que bien que qu'il y ait eu quelques modélisations déterministes pour le traitement des données manquantes, la majorité des études a été basée sur des modèles probabilistes, avec la considération d'une fonction de vraisemblance. Par la suite les modèles bayésiens sont venus enrichir les modèles probabilistes, avec la formalisation des connaissances antérieures sur les paramètres par des lois a priori. Les illustres précurseurs de ce type d'analyses sur les données manquantes sont Rubin et Little. D'abord Rubin a formalisé les mécanismes sous-jacents à l'absence de données, puis a donné quelques résultats importants sur l'inférence bayésienne en présence données manquantes (Rubin, 1976 [9]) et a développé la technique de l'imputation multiple qu'il a proposé comme alternative à la correction des désavantages de l'imputation simple. C'est une procédure de complétion de données, principalement des données de questionnaires, basée sur des méthodes d'inférence bayésiennes (Rubin, 1978, 1987 et 1988 [10, 11, 12]). Little a aussi modélisé les non-réponses dans les questionnaires (Little, 1982 [13]), avant de faire une étude commune avec Rubin pour un large éventail de méthodes d'analyses statistiques avec données manquantes (Little et Rubin, 1987 [14]. Leurs travaux ont inspiré beaucoup de recherches dans le domaine. En effet s'appuyant sur la théorie des mécanismes sous-jacents à l'absence de données, une étude sur des données incomplètes multivariées sous différents modèles (données continues gaussiennes, données catégorielles ...) a été faite par Schafer (1997) [15], une stratégie d'imputation multiple sous des modèles linéaires mixtes de données multivariées longitudinales incomplètes a été proposée par Schafer et Yucel (2002) [16]. van Buuren (1999) [17] a introduit une méthode d'imputation multiple, d'inférence bayésienne et basée sur des équations chaînées, qu'il a nommée MICE (Multivariate Imputation by Chained Equations) et qui a été utilisée dans des applications médicales (Resche-Rigon et White, 2018 [18]).

L'analyse statistique multivariée aura connu un tournant majeur avec l'avènement d'une nouvelle méthode de factorisation matricielle, comme alternative aux méthodes factorielles déjà existantes de décomposition en valeurs singulières, SVD (Sigular Values Decomposition), d'Analyse en Compsantes Principales (ACP), ou encore de quantification vectorielle, VQ (Vector Quantization). Cette nouvelle méthode, introduite par Lee et Seung et formalisée par $X \approx UV$, se démarque des autres par la contrainte de positivité sur les facteurs U et V, d'où son nom NMF (Nonnegative Matrix Factorization) désignant une factorisation de matrices positives (Lee et Seung, 1999 et 2001 [19, 20]). Au delà de la description des données, la contrainte de positivité positionne aussi la méthode NMF comme une procédure d'imputation de données. Les études faites par Lee et Seung ont alors boosté les approches déterministes de la complétion de données manquantes, qui jusqu'alors ne se limitaient qu'à des techniques de régressions linéaires multiples, où les variables comportant des données manquantes sont prédites de façon séquentielle (Little et Rubin, 1987 [14]). C'est ainsi qu'en 2004 Mao et Saul ont proposé un modèle de représentation et de prédiction de distances dans les réseaux internet à grande échelle par la factorisation NMF pondérée (Mao et Saul, 2004 [21]). Participant à la compétition *Netflix Prize* qui était un concours créé en 2006 et destiné à primer le meilleur algorithme de filtrage collaboratif permettant de prédire les notes des utilisateurs pour les films sur la base d'évaluations précédentes, les auteurs Kim et Choi ont repris le modèle de Mao et Saul, mais l'algorithme d'estimation a été substitué par la technique des moindres carrées alternées (Kim et Choi, 2009 [22]).

Bien que les modèles bayésiens et les factorisations NMF, sous-tendus par deux modélisations qui s'opposent (probabilistes et déterministes), semblent non conciliables il y a pourtant eu beaucoup de travaux qui ont associé les deux approches (Canny, 2004 [23], Zhang et al, 2006 [24], Virtanen, 2007 [25], Virtanen et al, 2008 [26], Salakhutdinov et Mnih, 2008 [27], Schmidt et al, 2009 [28], Cemgil, 2009 [29], Shan et Banerjee, 2010 [30], Porteous et al, 2010 [31], Gopalan et al, 2014 et 2015 [32, 33]...). Ce qui a permis d'avoir les modèles NMF bayésiens. Cependant malgré la multiplicité et la diversité de ces modèles, où nous avons remarqué une dualité des vraisemblances de lois de Poisson et de loi normales, peu d'entre eux ont eu à être appliqués au traitement

de données manquantes (Salakhutdinov et Mnih, 2008 [27] pour le modèle gaussien, Cemgil, 2009 [29] pour le modèle poissonien). Dans toutes ces modélisations NMF probabilistes citées la dualité des vraisemblances que nous avons évoquée se caractérise par le fait que si pour les unes une loi normale est affectée aux données X, pour les autres c'est la loi de Poisson qui est supposée. Ce choix des deux distributions n'est pas fortuit et explique ainsi les connexions très fortes entre les modèles NMF déterministes et ceux probabilistes. Pour le comprendre il faudrait remonter aux travaux des précurseurs de la méthodologie NMF, Lee et Seung, 1999 et 2001 [19, 20] (cf. Chapitre 2). En effet les auteurs ont articulé leur étude autour de deux fonctions-coûts que sont la divergence de Kullback-Leibler généralisée $F_1(U,V) = D_{KL}(X\|UV)$ et le carré de la distance euclidienne $F_2(U,V) = \|X-UV\|_F^2$ entre matrices, construite à partir de la norme de Frobenius. L'approximation factorielle $X \approx UV$ a donc consisté, suivant la fonction-coût choisie, à la résolution d'un problème de minimisation, c'est-à-dire trouver un couple de matrices positives (U^*, V^*) qui vérifient $(U^*, V^*) = \arg\max_{U,V} D_{KL}(X\|UV)$ ou $(U^*, V^*) = \arg\max_{U,V} \|X-UV\|_F^2$. Ainsi les modèles NMF probabilistes qui ont été développés à la suite des travaux de Lee et Seung sont dans leur quasi totalité et selon le type d'application soit d'inspiration la fonction-coût divergence de Kullback-Leibler généralisée (Canny, 2004 [23], Virtanen et al., 2008 [26], Cemgil, 2009 [29], Gopalan et al., 2014 et 2015 [32, 33] ...), soit d'inspiration la fonction-coût carré de la distance euclidienne (Zhang et al., 2006 [24], Salakhutdinov et Mnith, 2008 [27], Schmidt et al,2009 [28], Shan et Banerjee, 2010 [30], Portous et al, 2010 [31] ...). Il y a donc un lien intrinsèque fort entre, d'une part, la fonction-coût $D_{KL}(X\|UV)$ et la loi de Poisson et, d'autre part, la fonction-coût $\|X-UV\|_F^2$ et la loi normale. En effet il est apparu dans différentes études que résoudre les problèmes de minimisation $\arg\max_{U,V} D_{KL}(X\|UV)$ et $\arg\max_{U,V} \|X-UV\|_F^2$ revient respectivement et sous certaines conditions à maximiser une vraisemblance de loi Poisson $\mathcal{P}(X\,;\,UV)$ (Virtanen et al, 2008 [26]) et une vraisemblance de loi normale $\mathcal{N}(X\,;\,UV,\sigma)$ (Salakhutdinov et Mnih, 2008 [27])). S'agissant de la fonction objectif $D_{KL}(X\|UV)$ il a même été montré qu'on retrouve facilement les équations de mises à jour multiplicatives (cf. Chapitre 2 Eq. 2.5) si l'on utilise l'algorithme EM (Espérance-Maximisation) de maximisation d'une vraisemblance Poisson via un modèle de données augmentées (Cemgil, 2009 [29]). Les modèles NMF bayésiens peuvent ainsi être vus un prolongement des modèles NMF déterministes, avec des loi a priori sur les paramètres U et V comme contraintes de régularisation sur ces paramètres. Le modèle que nous étudions dans ce travail est de la classe des modèles NMF bayésiens poissoniens.

La motivation du choix de ce type de modèles réside d'abord sur le nombre très faible de modèles bayésiens NMF, en général, dédiés à l'imputation de données manquantes (Slakhutdinov et Mnih, 2008 [27], Cemgil, 2009 [29]), ensuite et à notre connaissance, sur l'inexistence de modèles bayésiens poissoniens NMF, en particulier, pour l'imputation de données incomplètes de questionnaires et particulièrement ceux d'enquêtes médicales, en l'occurrence le questionnaire *HSOPSC* (*Hospital Survey on Patient Safety Culture*). Le modèle bayésien poissonien NMF que nous développons dans cette thèse, et nommerons modèle NMF *Poisson-Gamma* pour la suite, est une approximation factorielle $X \approx UV$ spécifié par une vraisemblance de Poisson sur les données X et des lois a priori gamma sur les facteurs et paramètres du modèle que sont les matrices positives U et V. Notons cependant que malgré le nom générique *Poisson-Gamma* que nous lui donnons, les différents auteurs qui eu à le développer l'ont fait avec des spécificités aussi bien dans la conception que dans les applications. En effet il a d'abord été introduit par Virtanen et auteurs associés qui l'ont développé pour des applications de séparation et de détection de sources audio à un seul canal (Virtanen et al, 2008 [26]). Puis Cemgil, un des auteurs associés de Virtanen dans l'étude précédente a repris le modèle en l'adaptant au traitement de données incomplètes, notamment la reconstitution d'images de faces humaines dégradées issues de la base de données Olvetti (Cemgil, 2009 [29]). La méthode d'inférence utilisée alors par l'auteur pour l'estimation des paramètres U et V est l'algorithme variationnel d'estimation analytique de la loi a posteriori (Ghrahramani et Beal, 2000 [34]). Dans cette thèse la spécification de l'application du modèle concernera donc des données de questionnaires médicales.

En santé publique les questionnaires sont des instruments d'auto-évaluation de plus en plus

adoptés. En effet en 2004 l'agence américaine en santé publique AHRQ (Agency for Healthcare Research and Quality) a développé un célèbre questionnaire, nommé HSOPSC (Hospital Survey on Patient Safety Culture), destiné a améliorer la qualité de la prise en charge des patients dans les structures américaines publiques de santé et consistant à évaluer la culture de sécurité du patient du point de vue du personnel de ces établissements. Ce questionnaire a par la suite été adopté par plusieurs autres pays. Cependant bien que jouissant de bonnes propriétés psychométriques, le questionnaire HSOPSC souffre d'un problème d'acceptabilité de la part de la cible et pourrait alors conduire à un fort taux données manquantes. Par ailleurs le volume du questionnaire (42 items) pourrait d'une part, dans les cas où des coûts sont induits, alourdir les coûts inhérents, et d'autre part fortement allonger la durée de la collecte (19 mois pour le questionnaire HSOPSC réalisé en France).

L'objectif de notre travail est de reconstruire automatiquement un questionnaire HSOPSC dont nous retirerons délibérément et de manière artificielle et aléatoire certaines données de réponses. L'idéal est alors de disposer d'un questionnaire complet (ou presque). Les données du questionnaire HSOPSC sur lesquelles vont porter notre étude ont été recueillies au centre hospitalier universitaire (CHU) de Grenoble (France) entre 2013 et 2014. Celles-ci comportaient des données manquantes originelles d'un taux d'environ 1.8%. Ce faible pourcentage nous permet de qualifier le questionnaire de presque complet, et alors bien compatible avec l'étude que nous voulons mener dans cette thèse. Les motivations de notre travail de reconstruction automatique sur ce questionnaire HSOPSC trouvent leurs sources à deux niveaux. Il s'agit de proposer un algorithme d'imputation qui d'une part est capable de faire de la complétion de données manquantes sur d'autres questionnaires du même type ayant un taux élevé de non-réponses, et d'autre part capable de faire de la prédiction de réponses dans le cas où le questionnaire a été réduit en amont et que les non-réponses sont du fait des questions non soumises. Ce second cas est en général rencontré lorsque la structure de santé publique concernée décide d'augmenter les chances d'acceptabilité du questionnaire, ou aussi quand des préoccupations de coût ou de durée des enquêtes sont mises en avant. La collecte de données sur le HSOPSC ainsi sous-dimensionné, verrait ainsi sont coût et sa durée réduits à souhait puisqu'un taux de retrait d'items serait alors fixé et modifiable au besoin. Comme nous l'avons dit tantôt nos données d'application seront les données de réponses du HSOPSC version française (Occelli et al., 2013 [35]). Alors chaque point (ou item) du questionnaire comporte cinq réponses possibles ordonnées selon le niveau de l'accord du répondant. Les réponses possibles sont données par ordre croissant de l'accord comme suit : pas du tout d'accord (*strongly disagree*), pas d'accord (*disagree*), ni d'accord ni en désaccord (*neither agree nor disagree*), en accord (*agree*) tout à fait d'accord (*strongly agree*). Ces réponses sont alors respectivement associées à des valeurs entières allant de 1 à 5 selon le codage de Likert. Ce questionnaire a été soumis à l'hôpital universitaire de Grenoble, d'une capacité de 1836 lits et desservant une population de 675 000 habitants. L'enquête a été menée entre avril 2013 et septembre 2014. Les participants admissibles étaient des employés à temps plein ou à temps partiel ayant au moins six mois d'emploi dans les service cliniques, de laboratoire, de pathologie, de radiologies ou de pharmacie. Sur 5044 employés admissibles, 3888 ont finalement participé à l'étude.

La méthodologie que nous proposons alors, pour parvenir à ce travail de reconstruction automatiques, s'articule autour cinq points majeurs.
Le premier point consiste à la formalisation de notre modèle NMF bayésien Poisson-Gamma, c'est-à-dire la considération d'une vraisemblance de loi de Poisson sur les données X qui sont des données discrètes positives, d'où la compatibilité, et des a priori sur les paramètres résumés par des loi Gamma sur U et V. Nous noterons par X la matrice des données originales, l'introduction artificielle des données manquantes sera caractérisée par la considération d'un indicateur de non-réponses. Il s'agit d'une matrice $\delta = (\delta_{ij})$ définie par $\delta_{ij} = 0$ si la donnée x_{ij} est manquante et $\delta_{ij} = 1$ si la donnée x_{ij} est présente. L'approximation factorielle $X \approx UV$ consistera alors à une inférence bayésienne d'estimation des paramètres U et V tels que, avec l'hypothèse d'indépendance sur les données $x_{ij}\,, i = 1, \ldots, n\,; j = 1, \ldots, p$, chaque coefficient $\sum_\ell u_{i\ell} v_{\ell j}$ du produit UV est une valeur espérée de la donnée correspondante x_{ij}, c'est-à-dire $x_{ij} \sim \mathcal{P}(\sum_\ell u_{i\ell} v_{\ell j})$.

Par ailleurs, pour des commodités de calcul, des variables auxiliaires $S = (S^1, \ldots, S^\ell, \ldots, S^k)$, où chaque S^ℓ est une matrice de même taille que X, seront considérées et ajoutées au modèle. Celles-ci permettent en effet un calcul assez simple des lois conditionnelles a posteriori.
Le second point consistera au calcul des loi conditionnelles a posteriori et à l'implémentation de notre méthode d'inférence qui est un échantillonneur de Gibbs. Les lois conditionnelles a posteriori concerneront trois blocs de variables que nous appellerons *variables latentes* : il s'agit des paramètres U et V ainsi que des variables auxiliaires S qui seront estimées en même temps que les paramètres. L'hypothèse d'indépendance sera supposée sur ces variables latentes. Des loi conditionnelles a posteriori gamma seront trouvées pour paramètres U et V, alors que des lois multinomiales seront trouvées pour les variables S. L'identification des loi conditionnelles pour ces trois blocs de variables latentes permettra une implémentation simple de notre échantillonneur de Gibbs *PGNMF*. Cependant les lois a priori gamma des paramètres U et V comportent des paramètres $\theta = (a^u, b^u, a^v, b^v)$ qui constituent les hyperparamètres du modèle. Ceux-ci apparaissent comme des inconnues dans lois conditionnelles a posteriori doivent nécessairement être estimés.
Le troisième point sera donc celui où la méthodologie d'estimation des hyperparamètres sera développée. La fonction d'intérêt sera alors la vraisemblance marginale $p(X \mid \theta)$ et la procédure utilisée, la méthode variationnelle d'approximation.
Le quatrième et dernier point consistera, une fois les moyennes a posteriori $\bar{U}$ et $\bar{V}$ déterminées et le produit matriciel $\bar{U}\bar{V}$ calculé, à la procédure de correction des données imputées de $\bar{U}\bar{V}$. La procédure de correction sur cette matrice à coefficients non entiers, se ramènera dans un premier temps à un ré-échantillonnage des coefficients matriciels en leur attribuant l'entier le plus proche proportionnellement à la décimale de la valeur imputée ; dans un second temps, pour éviter la création de données non-présentes dans l'échantillon initial (données originales), les valeurs en dehors du rang sont ramenées aux valeurs minimales et maximales respectivement.
Pour finir nous comparerons notre algorithme *PGNMF* à des méthodes d'imputation récemment développées et qui sont en majorité des méthodes factorielles comme la notre. En effet Cemgil, 2009 [29] a dans son étude également montré comment une méthode EM (Espérance-Maximisation) sous un modèle Poisson-Gamma de données augmentées pouvait être adaptée pour une tâche de complétion. L'algorithme qui en découle pourrait être assimilé à celui des règles de mises à jour multiplicatives, MU (Multiplicative Updates) de Lee et Seung, 2001 [20] lorsqu'il est formaté pour pour un travail d'imputation. Une méthode de factorisation NMF déterministe et pondérée nommée WNMF (Weighted Nonnegative Matrix Factorization), dont la méthode d'inférence est basée sur les techniques des moindres carrées alternées ANLS (Alternating Nonnegative Matrix Factorization) de Zdunek et Cichocki, 2006 [36], Kim et al, 2007 [37], Lin, 2007 [38], Kim et Park, 2008[39], a été introduite par Kim et Choi, 2009 [22]. Leur modèle est en fait équivalent au modèle bayésien basic NMF de Salakhutdinov et Mnih, 2008 [27] ; et tous les deux dédiés au filtrage collaboratif proposé par Netflix. Une autre méthode d'imputation de données mixtes (continues et/ou catégorielles) nommée MissForest et basée sur la technique dite des forets aléatoires de Breiman, 2001 [40] a été développée par les auteurs Stekhoven et Bühlmann, 2011 [41]. Leur travail a été proposé comme une alternative à l'algorithme d'imputation multiple MICE de van Buuren et Oudshoorn, 1999 [17]. Enfin Josse et Husson, 2016 [42] puis Audigier et al, 2017 [43] ont présenté des méthodes d'imputation multiples par analyse de correspondance nommée MIMCA (Multiple Imputation by Multiple Correspondence Analysis), qui effectuent l'imputation de données catégorielles en utilisant l'analyse des correspondances multiples (ACM) (Josse et al., 2012 [44]).

Cette thèse se structure en quatre chapitres répartis en deux parties. La première constituant les préliminaires contient les deux premiers chapitres. La seconde partie donne la problématique, l'état de l'art et la méthodologie proposée qui sont articulés aussi en deux chapitres.

Le Chapitre 1 donne trois points majeurs quelques aspects essentiels de la modélisation bayésienne. Le premier point définit le paradigme bayésien qui s'appuie sur une modélisation

probabiliste dans un contexte d'analyse statistique, puis est caractérisé par l'inférence bayésienne qui consiste en une formalisation des connaissances antérieures sur le phénomène étudié par l'introduction des lois a priori avant de procéder à l'actualisation de ces connaissances grâce à l'information apportée par les données, c'est-à-dire la détermination des lois a posteriori. Le second point présente la théorie de la décision sous l'angle de l'inférence statistique classique en général et bayésienne en particulier. Une estimation sur un paramètre est alors considérée comme une décision dont l'utilité devra être optimale et inversement le coût minimisé. Enfin le troisième point présente les méthodes de calcul bayésiens, en particulier les méthodes Monte Carlo par Chaines de Markov, en l'occurrence l'échantillonneur de Gibbs.

Le Chapitre 2 étudie la factorisation matricielle NMF en deux axes majeurs. Le premier donne une présentation des algorithmes NMF classiques répartis en trois classes : les algorithmes de mise à jour multiplicative de Lee et Seung, les algorithmes du gradient de descente et les algorithmes de moindres carrés alternés. Le second donne quelques variantes qui constituent des aspects de factorisation pondérée. Les algorithmes qui en découlent sont inspirés des algorithmes de mise à jour multiplicative et de ceux des moindres carrés alternés.

Le Chapitre 3 traite, d'une part, de la problématique de la thèse qui est celle des données manquantes dans les questionnaires d'enquêtes, et d'autre part présente l'état de l'art. Ainsi le chapitre se structure en deux points. Le premier présente la typologie, à deux niveaux, des données manquantes. L'un catégorise les données manquantes selon la configuration des données non observées : on parle alors de structure univariée, monotone ou arbitraire. L'autre caractérise les données manquantes selon le mécanisme sous-jacent. Trois mécanismes ont été formalisés (MAR, MCAR et MNAR). Le second point donne l'état de l'art. En effet il y a eu dans la littérature différentes méthodes de traitement des données manquantes, que nous pouvons classer en deux catégories, d'un côté les analyses sans imputation et d'un autre les méthodes d'imputation. Pour ces dernières, un accent sera mis sur des méthodes récemment développées qui ont en commun avec la méthodologie que nous proposons, le principe de la factorisation matricielle. Une méthode d'apprentissage automatique basée sur des techniques de régression ainsi que quelques méthodes naïves d'imputation seront également étudiées.

Enfin le Chapitre 4 présente la méthodologie que nous proposons pour la reconstruction automatique d'un questionnaire HSOPSC, en l'occurrence celui du CHU de Grenoble. Il s'agit d'un questionnaire dont nous introduisons volontairement les non-réponses suivant un mécanisme MCAR, avant d'effectuer une tache d'imputation par le biais de l'algorithme *PGNMF* que nous développons. La perte d'information est mesurée à travers deux critères : le RMSE et la divergence de Kullback-Leibler. Le chapitre se structure en quatre axes majeurs. Le premier justifie le choix du modèle poissonien en donnant ses liens avec la fonction-coût divergence de Kullback-Leibler généralisée (Lee et Seung, 2001 [20]). Le second axe décrit d'abord les données du questionnaire HSOPSC, allant du cadre de leur recueil à leur codage suivant l'échelle de Likert ; puis il décrit le modèle bayésien Poisson-Gamma de factorisation NMF, la prise en charge des données manquantes étant caractérisée par la considération d'une matrice binaire où les valeurs 1 indique la présence de données alors que la valeur 0 indique leur absence. Le troisième axe présente notre méthodologie de reconstruction automatique dont l'algorithme résultant, *PGNMF*, s'appuie sur un échantillonneur de Gibbs. Enfin le quatrième axe donne les résultats décrivant les performances de notre algorithme. Celles-ci sont regardées sous l'angle de deux différents critères (le RMSE et la divergence de Kullback-Leibler), et seront comparées à celles des principales méthodes d'imputation étudiées dans l'état de l'art.

Première partie

Préliminaires

Chapitre 1

Modélisation bayésienne

Introduction

L'objet principal de la Statistique est de mener, grâce à l'observation d'un phénomène aléatoire, une inférence sur la distribution probabiliste à l'origine de ce phénomène, c'est-à-dire de fournir une analyse (ou une description) d'un phénomène passé, ou une prédiction d'un phénomène à venir de nature similaire. Ce phénomène pouvant être l'apparition ou l'expansion d'une épidémie, le métissage d'une population, la classification de document, le filtrage collaboratif, la reconstruction tomographique, l'apprentissage, l'imputation de données manquantes ... s'étudie par le biais d'une modélisation. Cette dernière désigne la représentation d'un système par un autre plus facile à appréhender. C'est donc un procédé qui grâce à ses caractéristiques et ses qualités peut servir de référence à l'imitation ou à la reproduction. Cependant, dans la plupart des cas, une caractéristique inhérente à la modélisation est sa « simplification » de la réalité. Son caractère « réducteur » est dû au fait qu'elle est une approximation de la réalité complexe. Cette propriété de la modélisation fait qu'elle perd une partie de la richesse du phénomène réel mais gagne, en contrepartie, en efficacité de par son aptitude à être facilement appréhendée par des outils mathématiques.

L'approche statistique est par essence formelle (ou mathématiquement structurée) parce qu'elle repose sur une formalisation poussée de la réalité objective. Notons par ailleurs que d'illustres auteurs considère la statistique comme l'interprétation du phénomène observé, plutôt que son explication (Robert, 2006 [45]). Si nous la regardons sous cet angle alors la critique de l'apposition d'un modèle probabiliste sur un phénomène inexpliqué, comme il est possible que le phénomène observé soit entièrement déterministe ou tout du moins sans rapport direct avec le modèle pré-supposé, n'aura guère de consistance. Il s'agit d'un point de vue fort qui s'illustre bien avec la modélisation de factorisation matricielle de Lee et Seung, 1999 et 2001 [19, 19], où approche déterministe et approche probabiliste conduisent aux mêmes équations d'estimations. Les modèles probabilistes formels permettent en effet d'incorporer simultanément les information disponibles sur le phénomène et les incertitudes inhérentes à ces information. Ils autorisent donc un discours qualitatif sur le problème en fournissant, à travers la théorie des probabilités, un véritable calcul de l'incertain qui permet de dépasser le stade descriptif des modèles déterministes. Notons que dans les modèles probabilistes d'analyses statistiques deux approches s'opposent : l'approche non paramétrique et celle paramétrique. La première suppose que l'inférence statistique doit prendre en compte la complexité du phénomène autant que possible et elle cherche donc à estimer la distribution sous-jacente du phénomène sous des hypothèses minimales, en ayant recours en général à l'estimation fonctionnelle(densité, fonction de régression ...). La seconde propose la représentation de la distribution des observations par une fonction de densité $f(x;\theta)$, où seul le paramètre θ (de dimension finie) est inconnu. Les deux approches ont leur avantages respectifs, cependant nous nous intéresserons dans ce chapitre aux modèles paramétriques.

La modélisation bayésienne, fait partie de la classe des approches probabilistes d'analyse statistique et doit sa formulation mathématique au révérend Thomas Bayes (1761) et à Pierre Simon Laplace (1773), même si l'histoire n'a retenu par la suite que le nom de Bayes, avec la

très connue *Formule de Bayes*. C'est une modélisation qui forme avec l'approche *fréquentiste* la principale dualité de l'analyse statistique. Bien qu'une vieille rivalité, née dans les années 1970, était entretenue par les « fréquentistes » et les « bayésiens » , les deux points de vue ne sont en réalité pas opposées puisque les probabilités fréquentistes et bayésiennes disent la même chose dès que l'on est sur des grands nombres (taille élevée des observations). Sur la finalité, leur différence apparaît dès que la taille des observations est faible, où l'inférence bayésienne se révèle plus utile. Dans la conception, la différence les deux écoles statistiques se manifeste par les fait que la statistique fréquentiste repose sur la loi des observations pour effectuer une inférence sur le paramètre, alors que la statistique bayésienne permet de combiner l'information apportée par les données avec les connaissances a priori, sur le paramètre, provenant soit d'études antérieures soit d'avis d'experts, dans le but d'obtenir une information a posteriori.

L'analyse Bayésienne est donc basée sur la formule de Bayes, selon laquelle, la distribution postérieure (loi a posteriori) d'un paramètre θ est proportionnelle à la distribution antérieure (loi a priori) paramètre θ par la vraisemblance de θ provenant des données collectées. Ainsi c'est ainsi un procédé d'affinement des croyances a priori sur un phénomène d'intérêt combinées aux informations tirées des données observées, pour arriver à des attentes postérieures mises à jour sur le phénomène. L'inférence bayésienne s'articule alors autour de trois points majeurs. D'abord la caractérisation des attentes a priori sur le phénomène étudié résumé par le paramètre, θ, du modèle. Cette caractérisation débouche sur la considération d'une loi a priori $\pi(\theta)$. Ensuite les informations sur les données observées sont modélisées par une distribution paramétrique $f(x;\theta)$, de paramètre θ et où $x = (x_1, \ldots, x_n)$ pouvant être un vecteur ou une matrice représente les données. Enfin la dernière étape consiste à la mise à jour des attentes a priori sachant les données observées. Cette mise à jour se caractérise par une distribution dite loi a posteriori du paramètre $\pi(\theta \mid x)$.

Dans ce chapitre un accent particulier sera mis sur les aspects décisionnels de l'inférence bayésienne parce que, d'une part, les analyses et/ou prédictions qu'elle effectue sont presque toujours motivées par un objectif (par exemple pour une entreprise, devrait-elle lancer un nouveau produit, pour un bateau marchand doit-on modifier sa trajectoire, pour un investisseur devrait-il vendre ses actions ...) ayant des conséquences mesurables (résultats financiers, durée du trajet, gains escomptés ...). D'autre part proposer des procédures inférentielles implique qu'on doit justifier le fait qu'elles soient préférables à d'autres. Il est donc nécessaire d'avoir un outil d'évaluation adapté à la comparaison de différentes procédures. Ces aspects décisionnels sont ainsi développés ici par la *théorie de la décision* dans un contexte d'analyse bayésienne.

Le dernier point qui sera abordé dans ce chapitre est l'ensemble des principaux outils d'inférence, désigné par la méthode du calcul bayésien. Celle-ci s'articule autour deux méthodologies que sont les méthodes d'intégration de Monte Carlo et les méthodes Monte Carlo par Chaînes de Markov. Les seconds cités engendrent des processus appelés chaînes de Markov par le biais principalement de deux algorithmes que sont l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs.

1.1 Le paradigme bayésien

1.1.1 Modélisation probabiliste en analyse statistique

Dans la plupart des domaines d'études et de recherches (mathématiques, biologie, cosmologie, médecine, épidémiologie économie ...), la modélisation mathématique est devenue presque incontournable. Principalement deux approches s'opposent, la modélisation déterministe et la modélisation probabiliste. Pour cette dernière qui nous intéresse ici nous donnons quelques exemples où elle s'applique à résoudre des problèmes de la vie réelle : la modélisation des filaments galactiques (van Lieshout et Stoica, 2003 [46]), la modélisation d'agrégats en épidémiologie animale (Erskine 2001 [47]), la modélisation du morcellement pour un environnement (Roques et Stoica, 2007 [48]), etc.

La modélisation probabiliste est ainsi un préalable à une bonne étude statistique en ce sens que l'objet de celle-ci est, grâce à l'observation d'un phénomène aléatoire, de faire une inférence

sur la distribution probabiliste à l'origine de ce phénomène. L'aléa dans le phénomène est porté par la part d'incertitude contenu dans les données. Cependant tout phénomène réel inexpliqué n'est pas toujours forcément sujet à une modélisation probabiliste car il peut arriver que le phénomène observé soit entièrement déterministe, sans que la fonction régulatrice du processus soit connue ni qu'il soit possible de la reconstruire à partir des observations. D'un autre côté si l'on regarde l'analyse statistique comme une interprétation du phénomène naturel observé et non explication, l'usage de la modélisation probabiliste se justifie d'avantage car les modèles qui en découlent permettent d'incorporer simultanément les informations disponibles sur le phénomène (facteurs déterminants, fréquence, amplitude, etc.) et les incertitudes inhérentes à ces informations. Ils autorisent donc un discours qualitatif sur le problème en fournissant, à travers la théorie des probabilités, un véritable calcul de l'incertain qui permet de dépasser le stade descriptif des modèles déterministes. C'est d'ailleurs la raison pour laquelle une interprétation probabiliste est nécessaire pour conduire une inférence statistique : elle donne un cadre qui permet de replacer le phénomène singulier observé dans la globalité d'un modèle et autorise ainsi les analyses et les généralisations. Il faut également noter qu'une modélisation probabiliste ne peut être défendue que si elle fournit une représentation suffisamment proche du phénomène observé. Toutefois un problème inhérent à la modélisation façon générale qu'est le caractère réducteur de la réalité complexe fait que les modèles probabilistes, en particuliers, ont une difficulté à connaître exactement la distribution probabiliste sous-jacente de la génération des observations, c'est-à-dire savoir s'il s'agit de la loi normale, exponentielle, binomiale, poisson ..., sauf dans des cas où la distribution des observations est parfaitement connue grâce à des considérations d'ordre physique, économique ou autres.

Ainsi le formalisme mathématique du modèle probabiliste est la donnée du triplet $(\mathcal{X}, \mathcal{A}, P)$ appelé espace probabilisé où $\mathcal{X}$ est l'espace des observations possibles, $\mathcal{A}$ est la tribu des évènements observables associée et P la mesure de probabilité définie sur $\mathcal{A}$. Comme nous l'avons dit, pour ce type de modèle les seules questions qui se posent sont de l'ordre du calcul de l'incertain. A partir du modèle probabiliste se construit donc le modèle statistique. La différence entre les deux modèles est que pour modèle statistique nous avons à la place de la probabilité P, une famille de probabilités. En effet le modèle statistique est considéré comme un outil mathématique associé à l'observation de données issues d'un phénomène aléatoire. Un travail fondamental préalable est l'expérience statistique qui consiste à recueillir une observation x d'un élément aléatoire X, à valeurs dans un espace $\mathcal{X}$ et dont on ne connait pas exactement la loi de probabilité P. Alors des considérations de la modélisation du phénomène observé amènent à admettre que P appartient à une famille $\mathcal{P}$ de lois de probabilité possibles. Le modèle statistique associé à cette expérience est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où $\mathcal{X}$ est l'espace des observations possibles, $\mathcal{A}$ est la tribu des évènements observables associée et $\mathcal{P}$ est une famille de lois de probabilités possibles définie sur $\mathcal{A}$. L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

Dans la modélisation statistique deux approches plus ou moins opposées sont adoptées : l'approche non-paramétrique et celle paramétrique.

Approche non-paramétrique. C'est une approche qui suppose que l'inférence statistique doit prendre en compte autant que possible la complexité du phénomène observé. Elle cherche donc à estimer la distribution sous-jacente du phénomène sous des hypothèses minimales. Notons également que pour ce procédé le modèle n'est pas décrit par un nombre fini de paramètres et que dans sa description du phénomène étudié plusieurs cas de figures peuvent se présenter. Nous pouvons avoir des cas où l'on s'autorise toutes les distributions possibles, c'est-à-dire ne faire aucune hypothèse sur la nature, la forme ou encore le type de la distribution des observations. On peut également travailler sur des espaces fonctionnels de dimension infinie (Exemple : l'espace des densités continues sur [0,1] ou l'espace des densités monotones sur $\mathbb{R}$), ou supposer non fixe le nombre de paramètres qui augmente avec le nombre d'observations, ou encore supposer le support de la distribution discret et qu'il augmente avec le nombre d'observations. L'approche non-paramétrique décrit donc le phénomène étudié par le biais d'estimation de la distribution de

probabilité des observations (estimation de fonction de répartition et de densité de probabilité), de fonctions de régression ou encore la réalisation de tests non-paramétriques (test d'adéquation à une loi, test de comparaison, test d'indépendance). La principale justification de l'analyse non-paramétrique est asymptotique et donc les méthodes sous-jacentes ne sont valables que lorsque la taille de l'échantillon devient infinie (ou en pratique fixée à une grande valeur). Notons cependant l'existence d'études sur des approches non-paramétriques comme les test de Hàjek et Sidák (1968) qui évacuent l'aspect d'estimation et les problèmes de tailles d'échantillons infinies par la construction de statistiques de test indépendantes des distributions bien que leurs applications restent limitées.

Approche paramétrique. Elle est basée sur le point de vue que la distribution des observations est conditionnée par un nombre fini de paramètres. C'est une procédé pragmatique dans la mesure où elle prend en compte la fait qu'un nombre fini d'observations (échantillon de taille fini) ne peut estimer qu'un nombre fini de paramètres. De plus la modélisation paramétrique permet une évaluation des outils inférentiels car elle travaille sur des échantillons de taille finie. Une définition formelle du modèle statistique paramétrique est la suivante.

Définition 1.1. *On dit qu'un modèle statistique est* ***paramétrique*** *s'il existe un entier p et un sous-ensemble Θ de $\mathbb{R}^p$ tels que la famille de probabilités $\mathcal{P}$ puisse être paramétrée par Θ, c'est-à-dire tels que l'application :*

$$\begin{array}{ccc} \Theta & \longrightarrow & \mathcal{P} \\ \theta & \longmapsto & P_\theta \end{array}$$

est surjective. On note $\mathcal{P} = \{P_\theta \ : \ \theta \in \Theta\}$.

Dans la plupart des cas le modèle est identifiable, quitte à prendre une autre paramétrisation. On supposera dans la suite que le modèle statistique est identifiable.

Définition 1.2. *Un modèle paramétrique $\mathcal{X}, \mathcal{A}, \mathcal{P}$ est dit* ***identifiable*** *si la fonction $\theta \mapsto P_\theta$ de la Définition 1.1 est de plus injective, c'est-à-dire $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$.*

La connaissance du support de la distribution paramétrique est aussi importante et sa définition est donnée par

Définition 1.3. *Soit P_θ une loi paramétrique de paramètre θ. On appelle* ***support*** *de la loi P_θ l'adhérence de l'ensemble des points en lesquels la densité de probabilité $f_\theta(x)$, dans le cas de v.a. continues, ou la probabilité $P_\theta(X = x)$, dans le cas de v.a. discrètes, ne s'annule pas. il est donné par :*

$$supp(P_\theta) = \overline{\{x \in \mathcal{X} : f_\theta(x) > 0\}} \quad ou \quad supp(P_\theta) = \overline{\{x \in \mathcal{X} : P_\theta(X = x) > 0\}}.$$

On constate qu'il est dénombrable dans le cas de v.a. discrètes et infini non dénombrable dans le cas de v.a. absolument continues. Ce support peut dépendre de θ. Il en est ainsi par exemple dans le cas du modèle uniforme $\left\{\mathcal{U}_{[0,\theta]} \ ; \ \theta > 0\right\}$.

Nous nous intéressons pour toute la suite à la modélisation statistique paramétrique dont nous donnons quelques exemples.

Exemple 1.1 (Problème de fiabilité et modèle de Bernoulli)**.** Considérons un problème de fiabilité où l'on étudie la durée de vie X d'un matériel. Il est raisonnable d'admettre que celle-ci est aléatoire et X est alors une variable aléatoire de fonction de répartition (f.d.r.) F. Supposons que l'on soit précisément intéressé par l'évaluation de la probabilité que le matériel soit en marche après un temps t_0 de fonctionnement, c'est à dire évaluer

$$\bar{F}(t_0) = P(X > t_0) = 1 - F(t_0).$$

Supposons que la v.a. X est à valeurs dans $\{0, 1\}$ pour modéliser l'état du matériel au temps t_0. On note $\{X = 1\}$ si le matériel est en marche et $\{X = 0\}$ s'il est en panne. On a $p_0 = P(X =$

$1) = \bar{F}(t_0)$ et $P(X = 0) = 1 - p_0$. La v.a. X est alors de loi de Bernoulli de paramètre p_0 inconnu dans $[0, 1]$. Le problème tel que modélisé donne à la v.a. X une infinité de loi de Bernoulli possibles $\mathcal{B}(1, p)$, avec p dans $[0, 1]$. L'inférence statistique consiste alors à trouver la vraie valeur p_0 à partir des observations $x_1, x_2, \ldots, x_n$ sur les n machines testées. Le modèle statistique est ici spécifié par $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où $\mathcal{X} = \{0, 1\}$, $\mathcal{A} = \mathcal{P}(\mathcal{X})$ et $\mathcal{P} = \{\mathcal{B}(1, p) \ : \ p \in [0, 1]\}$ et s'appelle modèle de Bernoulli. Il peut également servir de modélisation à l'expérience de lancer d'un pièce de monnaie ou encore au sondage d'intention de vote.
Le support de la loi de Bernoulli P_θ avec $P_\theta(X = x) = p^x(1 - p)^{1-x}$ est par, pour tout $x \in \text{supp}(P_\theta) = \{0, 1\}$.

Exemple 1.2 (Problème de contrôle de la qualité)**.** Considérons un entreprise de fabrique de vis. On constate que les mesures du diamètres X d'une vis varient d'une pièce à l'autre. Cet aléa peut être dû au procédé de fabrication et/ou aux éventuelles erreurs de mesure. Supposons que l'on ne connaisse pas la valeur moyenne du diamètre μ. Un modèle statistique adapté à une telle situation peut être la suivante. On suppose que l'aléa est symétrique et décroissant autour de la moyenne et que X admet une loi normale. Les données sont alors engendrées par le modèle suivant

$$X = \mu + \epsilon,$$

où ϵ est de loi normale $\mathcal{N}(0, \sigma^2)$. Autrement dit on a $X \sim \mathcal{N}(\mu, \sigma^2)$. Si l'on suppose dans un premier temps σ^2 connu alors le modèle statistique est alors donné par

$$\big(\mathbb{R},\ \mathcal{B}(\mathbb{R}),\ \big\{\mathcal{N}(\mu, \sigma^2) \ : \ \mu \in \mathbb{R}\big\}\big)$$

où $\mathcal{B}(\mathbb{R})$ est la tribu borélienne de $\mathbb{R}$. Remarquons que le seul paramètre du modèle est $\theta = \mu$ à valeurs dans $\Theta = \mathbb{R}$ puisque σ^2 est connu. Dans le cas contraire où σ^2 est lui aussi inconnu le modèle est alors donné par

$$\big(\mathbb{R},\ \mathcal{B}(\mathbb{R}),\ \big\{\mathcal{N}(\mu, \sigma^2) \ : \ \mu \in \mathbb{R},\ \sigma^2 > 0\big\}\big)$$

et l'on a $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ et $\theta = (\mu, \sigma^2)$, dans ce cas le paramètre est dit bi-dimensionnel. On peut aussi construire un modèle où l'espérance est connue et c'est la variance qui est inconnue.
Le support de la loi normale $P_\theta = \mathcal{N}(\mu, \sigma^2)$ est alors $\text{supp}(P_\theta) = \mathbb{R}$.

Modèle d'échantillonnage. Pour étudier un phénomène aléatoire, on a souvent intérêt à observer plusieurs réalisations indépendantes de celui-ci. On parle alors d'échantillon ou d'échantillonnage.

Définition 1.4. *On appelle n-**échantillon** de la loi P_θ, la donnée d'un vecteur $X^{(n)} = (X_1, \ldots, X_n)$ constitué de n v.a. indépendantes et identiquement distribuées (i.i.d.) de loi P_θ. On appelle **modèle d'échantillonnage**, le modèle*

$$(\mathcal{X}^n, \mathcal{A}^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n} \ : \ \theta \in \Theta\}),$$

où $\mathcal{X}^n$ est le produit cartésien de n espaces $\mathcal{X}_1, \ldots, \mathcal{X}_n$, $\mathcal{A}^{\otimes n}$ est la tribu produit (engendré par les pavés) sur $\mathcal{X}^n$ et $P_\theta^{\otimes n} = P_\theta \otimes \cdots \otimes P_\theta$ est la probabilité produit sur $(\mathcal{X}^n, \mathcal{A}^{\otimes n})$ qui est la loi du vecteur $X^{(n)} = (X_1, \ldots, X_n)$

Toutes les v.a. ont même loi, donc même valeur de θ pour P_θ. Un échantillon est un vecteur aléatoire dont les composantes son i.i.d. Sa réalisation est le résultat de n observations indépendantes du même phénomène et est notée $x^{(n)} = (x_1, \ldots, x_n)$. Un modèle d'échantillonnage est donc un modèle statistique particulier, où l'espace des observations $\mathcal{X}^n$ est le produit de n espaces, muni de sa tribu produit classique et de probabilités de la forme $P_\theta^{\otimes n}$. Grâce à l'indépendance et l'identique distribution, la probabilité jointe de l'échantillon, dans le cas discret, est alors :

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n \,;\, \theta) &= P_\theta^{\otimes n}(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P_\theta(X_i = x_i) \end{aligned}$$

Dans le cas continu, la densité de l'échantillon sous la loi P_θ est alors :

$$x^{(n)} = (x_1, \ldots, x_n) \mapsto \prod_{i=1}^{n} f_\theta(x_i)$$

Notation.

$$P_\theta(X = x) = P(X = x \,;\, \theta) \quad \text{et} \quad f_\theta(x) = f(x \,;\, \theta).$$

Dans les deux cas (discret et continu) si on considère le produit de droite non plus comme une fonction de x mais comme une fonction du paramètre θ, pour un $x^{(n)} = (x_1, \ldots, x_n)$ fixé, on parle de vraisemblance.

Vraisemblance

Définition 1.5. *Dans un modèle statistique paramétrique $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, on appelle* ***vraisemblance*** *de l'observation x la fonction*

dans le cas discret

$$\begin{aligned} L(.\,;\, x) \quad : \quad & \Theta \to \mathbb{R}^+ \\ & \theta \;\mapsto L(\theta; x) = P(X = x \,;\, \theta) \end{aligned}$$

dans le cas continu

$$\begin{aligned} L(\,.\,; x) \quad : \quad & \Theta \to \mathbb{R}^+ \\ & \theta \;\mapsto L(\theta; x) = f(x \,;\, \theta). \end{aligned}$$

Dans le cas d'un modèle d'échantillonnage, la vraisemblance de l'échantillon observé $x^{(n)} = (x_1, \ldots, x_n)$ s'écrit sous la forme,
dans le cas discret,

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} P(X_i = x_i \,;\, \theta)$$

dans le cas continu,

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \,;\, \theta)$$

C'est donc la loi conjointe du n-échantillon évaluée aux valeurs observées et considérée comme fonction du paramètre θ.

1.1.2 Inférence bayésienne

Comparée à la modélisation probabiliste, l'analyse statistique se ramène fondamentalement à une *inversion*[1], car elle doit déterminer les causes, réduites aux paramètres du mécanisme probabiliste générateur, à partir des effets résumés par les observations. En d'autres termes, quand nous observons un phénomène aléatoire contrôlé par le paramètre θ, une méthode statistique permet de déduire de ces observations une inférence (c'est-à-dire, en résumé, une caractérisation) sur θ, alors que la modélisation probabiliste caractérise le comportement des observations futures conditionnellement à θ. Ce caractère d'inversion propre à la Statistique apparaît de façon évidente

1. A l'époque de bayes et de Laplace, c'est-à-dire à la fin du XVIIIème siècle, la Statistique était appelée *Probabilités inverses*, à cause de cette perspective.

dans la notion de fonction de vraisemblance, car, d'un point de vue formel, il s'agit simplement d'une densité [2] réécrite dans le « bon ordre », $L(\theta \,|\, x) = f(x \,|\, \theta)$ soit donc comme fonction de θ, qui est inconnu, dépendant de la valeur observée x. Cette introduction du conditionnement dans la densité se justifie par le fait que dans le contexte bayésien la densité paramétrique est vue comme une loi de l'observation conditionnellement au paramètre θ, d'où $f(x\,;\,\theta) \equiv f(x \,|\, \theta)$.

Une description générale de l'inversion des probabilités est donnée par le *Théorème de Bayes* : Si A et E sont des événements tels que $P(E) \neq 0$, $P(A|E)$ et $P(E|A)$ sont reliés par

$$P(A|E) \quad = \quad \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}. \tag{1.1}$$

En particulier quand $P(B) = P(A)$ on a,

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)}, \tag{1.2}$$

Obtenir ces résultats à partir des axiomes de la Théorie des Probabilités est trivial. Il s'agit cependant de l'étape conceptuelle la plus importante dans l'histoire de la Statistique, constituant la première inversion des probabilités. L'équation (1.2) exprime le fait fondamental que, pour deux causes équiprobables, le rapport des probabilités pour un effet donné est égal au rapport des probabilités de ces deux causes. Ce théorème est aussi un principe d'actualisation, car il décrit la mise à jour de la vraisemblance de A de $P(A)$ vers $P(A|E)$, une fois que E a été observé. L'équation (1.1) pose ainsi les bases de l'inférence bayésienne.

1.1.2.1 Introduction de lois a priori

La philosophie de la méthode d'estimation bayésienne est très différente de celles des méthodes classiques (méthode du maximum de vraisemblance ou méthode des moments) où le paramètre θ est inconnu mais constant, déterministe. L'estimation est menée en considérant qu'on ignore tout de θ, mis à part son ensemble de définition.
Or parfois, on dispose d'une connaissance partielle sur θ. Cette information, dite a priori, peut provenir d'expériences similaires effectuées auparavant ou d'avis d'experts du phénomène étudié qui peuvent anticiper le résultat de l'expérience. Le principe de l'estimation bayésienne est de considérer que le paramètre θ est en fait la réalisation d'une variable aléatoire, et d'intégrer dans sa loi de probabilité toutes les informations a priori dont on dispose sur lui.

Des informations a priori aux lois a priori. Le point le plus critiquable et le plus critiqué de l'analyse bayésienne est le choix de la loi a priori. Car, une fois que cette loi a priori est connue, l'inférence peut être conduite d'une façon quasi mécanique en minimisant le coût a posteriori, en calculant les régions de plus forte densité a posteriori ou en intégrant les paramètres pour obtenir la distribution prédictive. La loi a priori est la clé de voute de l'inférence bayésienne et sa détermination est donc l'étape la plus importante dans la mise en œuvre de cette inférence. Dans une certaine mesure, c'est aussi la plus difficile. Évidemment, dans la ´ pratique, il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori, au sens où plusieurs lois de probabilité peuvent être compatibles avec cette information. Il y a plusieurs raisons pour cela : le décideur, le client ou le statisticien n'a pas forcément le temps ou les ressources (ni souvent la volonté) de chercher à construire un a priori exact (qui, de toute façon, peut tout simplement ne pas exister, au vu de l'information disponible) et doit compléter l'information partielle qu'il a rassemblée à l'aide de données subjectives afin d'obtenir une loi a priori.
Il est donc nécessaire le plus souvent de faire un choix (partiellement) arbitraire de loi a priori, ce qui peut avoir un impact considérable sur l'inférence qui en découle. En particulier, l'utilisation systématique de lois usuelles (normale, gamma, bêta, etc.) et la restriction plus forte encore aux lois conjuguées ne sont pas toujours justifiées, car la détermination subjective de la loi a priori qui

2. On parle de densité en supposant que la variable aléatoire est continue.

en résulte se fait au prix d'un traitement analytique plus fruste du problème, puisque ignorant une partie de l'information a priori. Certaines situations requièrent cependant une détermination partiellement automatisée de la loi a priori comme dans le cas extrême où l'information a priori est complètement absente. Nous considèrerons deux techniques usuelles : l'approche a priori conjuguée, qui nécessite une quantité limitée d'information, et l'approche non informative, qui est obtenue à partir de la distribution de l'échantillon.
Ces critiques contre l'approche bayésienne ont une certaine validité au sens où elles attirent l'attention sur le fait qu'il n'y a pas une façon unique de choisir une loi a priori, et que le choix de cette loi a un impact sur l'inférence résultante. Cet impact peut être négligeable, modéré ou énorme, puisqu'il est toujours possible de choisir une loi a priori qui donnera la réponse qu'on souhaite obtenir. Mais le point essentiel est ici que, premièrement, les lois a priori non fondées fournissent des inférences a posteriori non justifiées et, deuxièmement, le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers. Après des années de critiques, le travail de Jeffreys (1946) sur les a priori non informatifs apparut comme un don du ciel pour la communauté bayésienne, car il propose une méthode de construction de la loi a priori directement déduite de la distribution des observations. Certains bayésiens sont cependant en désaccord avec l'utilisation de méthodes automatisées (Lindley, 1971, 1990). Plus récemment, les avancées théoriques en robustesse et analyse de sensibilité ont aussi fourni une base solide à l'analyse bayésienne dans les cas d'information a priori incomplète, tandis que l'introduction de la modélisation hiérarchique permet de placer la sélection d'un a priori à un niveau plus éloigné, avec une diminution notable de l'impact sur l'inférence résultante.

Existence et Détermination de la loi a priori

Existence. Le mécanisme de génération du paramètre θ est généralement difficile à établir avec exactitude. Toutefois il peut arriver que le statisticien (ou le décideur) connaisse parfaitement le mécanisme sous-jacent tiré de considérations physiques, économiques, biologiques, etc. Dans ce cas la forme exacte ou même paramétrée pour la distribution a priori sur θ est connue. Notons que dans la plupart des cas, θ n'a pas de réalité propre (intrinsèque), mais correspond plutôt à une paramétrisation de la loi décrivant le phénomène aléatoire observé. La loi a priori π est alors un moyen de résumer l'information disponible sur ce phénomène, ainsi que l'incertitude liée à cette information. Ces situations impliquent évidemment des approximations de la vraie distribution a priori, si une vraie loi existe. Comme nous l'avons déjà dit les modèles statistiques sont le plus souvent des représentations simplifiées de ces phénomènes aléatoires et, puisqu'il n'existe pas de vrai modèle en général, mais seulement un modèle le plus proche du phénomène pour une distance appropriée, il est conceptuellement un peu difficile de parler de la vraie valeur de θ et, a fortiori, d'une vraie loi a priori.
D'un point de vue formel, il est possible de construire une distribution a priori en déterminant une échelle des vraisemblances respectives des valeurs du paramètre θ. Quand cette échelle est cohérente, c'est-à-dire respecte les axiomes dans Christian Robert, 2006 ([45], Chap. 3, Sect. 3.8.1) , l'existence d'une distribution a priori peut être déduite. L'existence d'une loi a priori subjective comme résultat d'un ordre des vraisemblances relatives est très important, car il nous permet d'échapper au cadre restrictif des justifications fréquentistes qui n'est pas toujours applicable à ce type de situations.

Détermination de la loi a priori. Plusieurs méthodes de détermination de loi a priori ont été développées. Il s'agit entre autres des méthodes de l'approximation directe, de l'entropie maximale, des loi a priori non informatives, des techniques de Bayes empiriques et hiérarchiques, mais aussi de la méthode des loi a priori conjuguées (Robert, 2006 [45]). C'est cette dernière citée qui portera ici notre intérêt.

Lois a priori conjuguées. Il s'agit d'une technique particulière de détermination de loi a priori paramétrée. En effet les cas où l'information est limitée nécessitent une telle approche. En

effet, quand l'information a priori sur le modèle est trop vague ou peu fiable, une construction subjective complète de la distribution a priori est évidemment impossible. D'autres raisons (retards, coûts à respecter, manque de communication entre statisticiens et décideurs, etc.) peuvent expliquer l'absence de distributions correctement définies. De plus, des exigences d'objectivité peuvent forcer le statisticien à fournir une réponse aussi neutre que possible, afin de fonder l'inférence sur le modèle d'échantillonnage uniquement. De tels cas semblent justifier le recours à des solutions non bayésiennes (estimateurs du maximum de vraisemblance, estimateurs sans biais optimaux, etc.).
D'abord, nous étudierons dans cette section une approche paramétrique classique qui implique un apport d'information subjective le plus limité possible et qui est à la base des deux techniques bayésiennes, hiérarchique et empirique. En dehors de l'exigence d'une contribution subjective minimale, les lois a priori conjuguées peuvent être considérées comme un point de départ pour l'élaboration de distributions a priori fondées sur une information a priori limitée, dont l'imprécision peut être déterminée grâce à des distributions a priori supplémentaires. Cependant, il faut garder à l'esprit le fait que l'impression commune que les lois conjuguées sont non informatives est fausse : le choix d'un a priori conjugué, bien qu'il soit défendable, est toujours un choix particulier et influence donc, dans une certaine mesure, l'inférence résultante. De plus, il peut obliger à ignorer une partie de l'information a priori si cette dernière n'est pas complètement compatible avec la structure de la loi a priori conjuguée. Enfin il existe d'autres lois a priori fondées sur la même information subjective limitée, mais avec une influence plus limitée sur l'inférence résultante. De plus, il peut obliger à ignorer une partie de l'information a priori si cette dernière n'est pas complètement compatible avec la structure de la loi a priori conjuguée.

Définition 1.6. *Une famille $\mathcal{F}$ de distributions de probabilité sur Θ est dite* ***conjuguée*** *(ou* ***fermé par échantillonnage****) par une fonction de vraisemblance $f(x\,|\,\theta)$ si, pour tout $\pi \in \mathcal{F}$, la distribution a posteriori $\pi(.\,|\,x)$ appartient également à $\mathcal{F}$.*

L'intérêt principal du caractère conjugué devient plus évident quand $\mathcal{F}$ est paramétrée. Effectivement, le passage de la distribution a priori à la distribution a posteriori se réduit dans ce cas à une mise à jour des paramètres correspondants. Cette seule propriété peut expliquer pourquoi les lois a priori conjuguées sont si populaires, car les distributions a posteriori sont toujours calculables (au moins jusqu'à un certain degré).

Justification des lois a priori conjuguées. L'approche a priori conjuguée, introduite par Raiffa et Schlaifer (1961), peut être justifiée partiellement par un raisonnement d'*invariance*. En fait, quand l'observation de $X \sim f(x\,|\,\theta)$ modifie $\pi(\theta)$ en $\pi(\theta\,|\,x)$, l'information transmise par x sur θ est évidemment limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $\pi(\theta)$, mais simplement de ses paramètres. Un changement plus radical de π est alors inacceptable et le choix des lois a priori devrait toujours être fait parmi les lois conjuguées, quelle que soit l'information a priori. D'une certaine façon, de Finetti (1974) avait un avis similaire parce qu'il considérait que l'information a priori pouvait être interprétée comme des observations passées virtuelles, ce qui mène forcément à des lois a priori conjuguées pour des familles exponentielles. Malheureusement, cette condition devient paradoxale dans les cas extrêmes où toute la distribution a priori est déjà disponible. Mais les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres. Une autre justification pour utiliser les lois a priori conjuguées est que certains estimateurs de Bayes sont alors linéaires, comme l'ont montré Diaconis et Ylvisaker (1979). Néanmoins, nous devons reconnaître que la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.
Ces lois a priori conjuguées sont parfois appelées *objectives* parce que le modèle d'échantillonnage, $f(x\,|\,\theta)$, détermine entièrement la classe des lois a priori, mais toute méthode qui produit de façon automatique des lois a priori à partir de la distribution d'échantillonnage serait tout aussi objective. A contrario, leur utilisation est fortement critiquée par certains bayésiens, car elle obéit à des contraintes techniques plutôt qu'à des impératifs d'adéquation à l'information a priori

disponible. Le rôle des lois a priori conjuguées est alors de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse.

Familles exponentielles. Les lois a priori conjuguées sont généralement associées à un type particulier de lois d'échantillonnage qui permet toujours leur obtention ; il est même caractéristique des lois a priori conjuguées comme nous le verrons ci-dessous. Ces lois constituent ce quoon appelle des *familles exponentielles* et sont étudiées en détail dans Brown (1986b).

Définition 1.7. *Soient μ une mesure σ-finie sur $\mathcal{X}$, Θ l'espace des paramètres, C et h des fonctions respectivement de $\mathcal{X}$ et Θ dans $\mathbb{R}_+$, et R et T des fonctions de Θ et $\mathcal{X}$ dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à μ)*

$$f(x\,|\,\theta) = C(\theta)h(x)\exp\left(R(\theta)T(x)\right) \tag{1.3}$$

est dite ***famille exponentiellement*** *de dimension k. Dans le cas particulier où $\Theta \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^k$ et*

$$f(x\,|\,\theta) = C(\theta)h(x)\exp(\theta x) \tag{1.4}$$

la famille est dite ***naturelle****.*

D'un point de vue analytique, les familles exponentielles ont certaines caractéristiques intéressantes (voir Brown, 1986b). En particulier, elles sont telles que, pour tout échantillon de (1.3), il existe une statistique exhaustive de dimension constante. En effet, si $X_1, \ldots, X_n \sim f(x\,|\,\theta)$, avec f satisfaisant (1.4), alors la statistique

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \quad \text{avec } \bar{X} \in \mathbb{R}^k$$

est exhaustive pour tout n. La réciproque de ce résultat a été aussi établie par Koopman (1936) et Pitman (1936).

De nombreuses distributions usuelles continues et discrètes appartiennent à des familles exponentielles.

Exemple 1.3. Si $\mathcal{S}_k$ est le simplexe de $\mathbb{R}^k$,

$$\mathcal{S}_k = \left\{\omega = (\omega_1, \ldots, \omega_k);\ \sum_{i=1}^{k}\omega_i = 1\,,\ \omega_i > 0\right\}\ ,$$

la loi de Dirichlet sur $\mathcal{S}_k$, $\mathcal{D}_k(\alpha_1, \ldots, \alpha_k)$, est une extension de la distribution bêta définie comme

$$f(p\,|\,\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\prod_{i=1}^{k} p_i^{\alpha_i - 1}\mathbb{1}_{\mathcal{S}_k}(p)$$

où $p(p_1, \ldots, p_k)$. Puisque

$$f(p\,|\,\alpha) = C(\alpha)h(p)\exp\left(\sum_{i=1}^{k}\alpha_i\log(p_i)\right)$$

la loi de Dirichlet constitue une famille naturelle exponentielle pour
$T(p) = (\log(p_1), \ldots, \log(p_k))$.

Exemple 1.4. Soit X un vecteur de loi normale multidimensionnelle $\mathcal{N}_p(\theta\,,\,\sigma^2 I_p)$, sa densité est donnée par

$$\begin{aligned} f(x|\theta) &= \frac{1}{\sigma^p}\frac{1}{(2\pi)^{p/2}}\exp\left(-\sum_{i=1}^{p}(x_i - \theta_i)^2/2\sigma^2\right) \\ &= C(\theta,\sigma)h(x)\exp(x(\theta/\sigma^2) + \|x\|^2(-1/2\sigma^2)). \end{aligned}$$

Ce qui montre que cette la loi normale appartient à une famille exponentielle de paramètres naturels θ/σ^2 et $-1/2\sigma^2$.

Dans ce dernier exemple, notons que l'espace des paramètres est de dimension $p+1$, tandis que la dimension des observables, X, est p. Bien que la dimension d'une famille exponentielle ne soit pas fixée, car il est toujours possible d'ajouter des combinaisons convexes des paramètres originaux comme des paramètres supplémentaires (et évidemment inutiles), une dimension minimale intrinsèque est associée à cette famille.

Définition 1.8. *Soit $f(x|\theta) = C(\theta)h(x)\exp(\theta x)$, une famille exponentielle naturelle. L'espace naturel des paramètres est*

$$E_N = \left\{\theta\,;\ \int_{\mathcal{X}} \exp(\theta x)h(x)\mathrm{d}\mu(x) < +\infty\right\}.$$

*La famille est dite **régulière** si E_N est un ensemble ouvert et **minimale** si $\dim(E_N) = \dim(K) = k$, où K est la clôture de l'enveloppe convexe du support de μ.*

Ainsi il est toujours possible de réduire une famille exponentielle à une forme standard et minimale de dimension m, et cette dimension m ne dépend aucunement de la paramétrisation choisie (Brown, 1986b).
Les familles exponentielles naturelles peuvent aussi être réécrites sous la forme

$$f(x|\theta) = h(x)\exp(\theta x - \psi(\theta)) \tag{1.5}$$

avec $\psi(\theta)$ qui est la fonction cumulante des moments. Une illustration de la forme (1.5) est donné dans l'exemple suivant avec la loi de Poisson.

Exemple 1.5. Si X est une v.a. de loi de Poisson $\mathcal{P}(\lambda)$, alors sa loi de probabilité

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{peut aussi s'écrire par} \quad f(x|\lambda) = \frac{1}{x!}e^{\theta x - e^\theta}$$

avec $h(x) = \frac{1}{x!}$, $\psi(\theta) = e^\theta$ et le paramètre naturel $\theta = \log\lambda$.

La structure régulière des familles exponentielles permet de nombreuses applications statistiques, comme en témoigne la vaste littérature sur ce sujet. (Voir Morris, 1982, Letac et Mora, 1990). Nous montrons dans la section suivant qu'elles autorisent également une construction simple de lois a priori conjuguées.

Lois conjuguées des familles exponentielles. Siot $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$, loi générique d'une famille exponentielle. Cette loi admet alors une famille conjuguée, comme le démontre le résultat suivant.

Théorème 1.1. *Une famille conjuguée pour $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$ est donnée par*

$$\pi(\theta\,|\,\mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda\psi(\theta))\,, \tag{1.6}$$

où $K(\mu,\lambda)$ est la constante de normalisation de la densité, $\lambda > 0$ et $\frac{\mu}{\lambda} \in \mathring{N}$.
La loi a posteriori correspondante est alors $\pi(\theta\,|\,\mu + x, \lambda + 1)$.

Nous donnons dans le tableau suivant les lois a priori conjuguées et les a posteriori correspondantes pour certaines lois usuelles des appartenant à des familles exponentielles.

1.1.2.2 Lois a posteriori

Le concept de l'inférence bayésienne est, comme nous l'avons dit plus haut dans la section 1.1, rendu possible grâce à l'inversion des probabilité conditionnelles permise par la théorie des probabilités. En réalité c'est la forme continue de (1.1) qui a été établie par Bayes et Laplace qui ont considéré que l'incertitude sur le paramètre θ d'un modèle peut être décrite par une loi de probabilité π sur Θ, appelée distribution a priori. Ce qui permet d'introduire la construction du modèle bayésien dont la contribution principale est de considérer en sus une distribution aléatoire pour les paramètres.

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathcal{N}(\theta,\sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	Normale $\mathcal{N}(\alpha(\sigma^2\mu+\tau^2x),\alpha\sigma^2\tau^2)$ $\alpha^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+\nu,\beta+x)$
Binomiale $\mathcal{B}(n,\theta)$	Bêta $\mathcal{B}e(\alpha,\beta)$	Bêta $\mathcal{B}e(\alpha+x,\beta+n-x)$
Binomiale négative $\mathcal{N}eg(m,\theta)$	Bêta $\mathcal{B}e(\alpha,\beta)$	Bêta $\mathcal{B}e(\alpha+m,\beta+x)$
Normale $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\alpha+1/2,\beta+(\mu-x)^2/2)$
Multinomiale $\mathcal{M}(n;\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$

TABLE 1.1: Lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles.

Définition 1.9. *Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique, $f(x|\theta)$, et d'une distribution a priori pour les paramètres, $\pi(\theta)$.*

L'inférence est alors fondée sur la distribution de θ conditionnelle à x, $\pi(\theta\,|\,x)$, appelée *loi a priori* et définie par

$$\pi(\theta|x)=\frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\mathrm{d}\theta} \tag{1.7}$$

En termes statistiques, le théorème de Bayes actualise donc l'information sur θ en extrayant l'information contenue dans l'observation x. Son impact provient de la décision audacieuse de mettre causes et effets sur le même niveau conceptuel, puisque les deux sont aléatoires. Du point de vue de la modélisation statistique, il y a donc peu de différences entre observations et paramètres, car les manipulations conditionnelles permettent l'échange de leurs rôles respectifs. Nous donnons dans les deux exemples (exemples historiques de Bayes et Laplace) qui suivent une illustration de l'usage de (1.7).

Exemple 1.6 (Bayes, 1763)**.** Une boule de billard W roule sur une ligne de longueur 1, avec une probabilité uniforme de s'arrêter n'importe où. Supposons qu'elle s'arrête en p. Une deuxième boule O roule alors n fois dans les mêmes conditions, et on note X le nombre de fois que la boule O s'arrête à gauche de W. Connaissant X, quelle inférence pouvons-nous mener sur p ?
Dans la terminologie moderne, le problème consiste à déterminer la distribution a posteriori de p conditionnellement à X. Par hypothèse, la loi a priori de p est la loi uniforme sur $[0,1]$ et X suit la loi binomiale $\mathcal{B}(n,p)$.
On a alors

$$\begin{aligned} P(X=x\,|\,p) &= C_n^x p^x(1-p)^{n-x}\,, \\ P(a<p<b\,,\,X=x) &= \int_a^b C_n^x p^x(1-p)^{n-x}\mathrm{d}p\,,\ \forall a<b\in[0,1] \\ P(X=x) &= \int_0^1 C_n^x p^x(1-p)^{n-x}\mathrm{d}p. \end{aligned}$$

Ce qui implique que

$$\begin{aligned} P(a < p < b \,|\, X = x) &= \frac{\int_a^b C_n^x p^x (1-p)^{n-x} \mathrm{d}p}{\int_0^1 C_n^x p^x (1-p)^{n-x} \mathrm{d}p} \\ &= \frac{\int_a^b p^x (1-p)^{n-x} \mathrm{d}p}{B(x+1, n-x+1)} \end{aligned}$$

donc la distribution de p conditionnellement à $X = x$ est la loi bêta, $\mathcal{B}e(x+1, n-x+1)$.

Dans le même esprit, Laplace introduit une modélisation probabiliste de l'espace des paramètres. Mais ses exemples sont plus avancés que ceux de Bayes au sens où les distributions a priori qu'il prend en compte sont fondées sur un raisonnement abstrait, plutôt que sur une justification physique.

Exemple 1.7 (Laplace, 1773)**.** Une urne contient un nombre n de cartes noires et blanches. Si la première carte sortie de l'urne est blanche, quelle est la probabilité que la proportion p de cartes blanches soit p_0 ?
Pour résoudre ce problème, Laplace suppose que tous les nombres de 2 à n?1 sont équiprobables comme valeurs de pn, donc que p soit uniformément distribué sur A_n, avec $A_n = \{2/n, \ldots, (n-1)/n\}$. La distribution a posteriori de p peut être alors calculée en utilisant le théorème de Bayes,

$$\begin{aligned} P(p = p_0 \,|\, \text{données}) &= \frac{p_0 \times 1/(n-2)}{\sum_{p=2/n}^{(n-1)/n} p \times 1/(n-2)} \\ &= \frac{np_0}{n(n-1)/2 - 1}\,. \end{aligned}$$

1.2 Approche bayésienne de la théorie de la décision

Si nous considérons que l'objectif général de la plupart des études inférentielles est de permettre au statisticien ou au client de prendre une *décision*, il peut être raisonnable d'établir un critère d'évaluation des procédures choisies. Ce critère prendra alors en compte les conséquences de chaque décision et dépendra des paramètres du modèle. Ces décisions peuvent être de différents types, par exemple acheter des capitaux selon leurs futurs rendements θ, interrompre une expérience agricole sur une nouvelle culture de productivité θ, déterminer si le nombre θ des sans domicile fixe a augmenté depuis le dernier recensement. Un autre type de décision est d'évaluer si une nouvelle théorie scientifique est compatible avec les données expérimentales disponibles.
Le critère d'évaluation est aussi habituellement appelé *coût* et est défini comme une fonction L de $\Theta \times \mathcal{D}$ dans $[0, +\infty[$, Θ l'espace des paramètres et $\mathcal{D}$ l'espace des décisions possibles. Cependant la plupart des exemples théoriques se concentrent sur le cas $\mathcal{D} = \Theta$, qui représente le cadre d'estimation standard. La fonction de coût est censée évaluer la pénalité (ou l'erreur) $L(\theta, d)$ associée à la décision d quand le paramètre prend la valeur θ. Ainsi dans un cadre traditionnel d'estimation du paramètre, lorsque $\mathcal{D} = \Theta$ ou $\mathcal{D} = h(\Theta)$, la fonction de coût $L(\theta, \delta)$ mesure l'erreur commise en évaluant $h(\theta)$ par δ. Il existe plusieurs axiomes de rationalité qui garantissent l'existence d'une telle fonction dans un cadre décisionnel. Dans la pratique, la détermination même de la fonction de coût est souvent difficile, en particulier parce que les conséquences de chaque action pour chaque valeur de θ sont souvent impossibles à déterminer quand $\mathcal{D}$ ou Θ sont de grands ensembles, par exemple quand ils contiennent un nombre infini d'éléments. De plus, dans les modèles qualitatifs, il peut être délicat de quantifier les conséquences de chaque décision. Nous verrons à travers des paradoxes comme le *paradoxe de Saint-Pétersbourg* que, même quand la fonction de coût semble évidente, par exemple lorsque des erreurs peuvent être exprimées comme pertes monétaires, la fonction de coût réelle peut être assez différente de son approximation linéaire et intuitive. La complexité de la détermination de la fonction de coût subjective du décideur incite souvent le statisticien à recourir aux fonctions de coût classiques [3],

3. Le terme classique est lié aux travaux de Laplace (1773) pour le coût absolu, de Gauss (1810) pour le coût quadratique.

choisies pour leur simplicité et leur souplesse mathématique. Ce type de fonction de coût est aussi nécessaire pour un traitement théorique de l'obtention des procédures optimales, quand il n'y a pas de motivation pratique pour le choix d'une fonction de coût en particulier.

Le modèle de la Théorie de la Décision bayésienne est est basée sur les trois facteurs suivants :
. la famille des distributions pour les observations, $f(x|\theta)$;
. la distribution a priori pour les paramètres, $\pi(\theta)$;
. le coût associé aux décision, $L(\theta, \delta)$.

Notons que pour le cas de la Théorie de la Décision classique (approche fréquentiste) la distribution a priori des paramètres n'est pas considérée. Cependant les partisans de l'approche bayésienne critiquent cette démarche fréquentiste de la théorie de la décision. En effet, ils considèrent l'existence d'une fonction de coût implique qu'une certaine information sur le problème considéré est disponible. Cette information peut donc être utilisée plus efficacement pour développer une distribution a priori. Ils estiment en réalité, que coût et a priori sont difficiles à dissocier et devraient être analysés simultanément (Lindley, 1985).

1.2.1 La fonction d'utilité

L'utilité est une notion utilisée en Statistique, en Économie mais aussi en Théorie des Jeux. Elle est définie comme l'opposée de la fonction de coût. Son appréhension nécessite d'ordonner les *conséquences* (ou *récompenses*) des décisions. Par *conséquences*, il faut comprendre que c'est l'ensemble des résultats émanant de l'action du décideur. Dans les cas les plus simples, il peut s'agir d'un gain ou d'un coût financier dû à cette décision. Dans le cas de l'estimation, l'utilité peut être une mesure de la distance entre l'estimation et la vraie valeur du paramètre. Les bases axiomatiques de l'utilité ont été attribuées à von Neumann et Morgenstern (1947) et ont mené à de nombreuses extensions, particulièrement en Théorie des Jeux. Dans un cadre statistique, cette approche a été considérée par Wald (1950) et Ferguson (1967).

Le cadre général sous-tendant la théorie de l'utilité considère que l'espace des récompenses $\mathcal{R}$, supposé complètement connu (par example, $\mathcal{R} = \mathbb{R}$). Nous supposons aussi qu'il est possible d'ordonner les récompenses, donc qu'il existe un ordre total, noté $\preceq$, sur $\mathcal{R}$ tel que si r_1, r_2 et r_3 sont dans $\mathcal{R}$,
(1) $r_1 \preceq r_2$ ou $r_2 \preceq r_1$;
(2) si $r_1 \preceq r_2$ et $r_2 \preceq r_3$, alors $r_1 \preceq r_3$.

Ces deux propriétés paraissent être des conditions minimales dans un cadre décisionnel. En particulier, la transitivité (2) est absolument nécessaire pour permettre une comparaison entre les procédures de décision. Sinon, nous pouvons nous retrouver avec des cycles tels que $r_1 \preceq r_2 \preceq r3 \preceq r_1$ et être dans l'incapacité de sélectionner la meilleure récompense parmi ces trois choix.

Pour avancer davantage dans la construction de la fonction d'utilité, il est nécessaire d'étendre l'espace des récompenses de $\mathcal{R}$ à $\mathcal{P}$, l'espace des distributions de probabilité dans $\mathcal{R}$. Ceci permet au décideur de prendre des décisions partiellement aléatoires ; de plus, l'espace des récompenses ainsi étendu est convexe. Ce genre d'espace de récompenses est tout à fait réaliste dans la mesure où les conséquences associées à une action ne sont pas connues au moment où la décision est prise ou, de façon équivalente, certaines décisions comportent une part de risque. Par exemple, en finance, le revenu financier d'actions cotées en Bourse n'est pas garanti au moment où les actionnaires doivent déterminer les entreprises dont ils devront acheter des actions. Dans ce cas, $\mathcal{D} = \{d_1, \ldots, d_n\}$, où d_k désigne l'action "*acheter des actions de la compagnie* k". Au moment de la décision, les gains associés aux différentes actions sont des dividendes aléatoires, connus seulement à la fin de l'année.

La relation d'ordre $\preceq$ est supposée disponible également dans $\mathcal{P}$. Par exemple, quand la récompense est monétaire, la relation d'ordre dans $\mathcal{P}$ peut être obtenue en comparant la moyenne des rendements associés à la distribution P. Il est donc possible de comparer des distributions de probabilité définies sur $\mathcal{R}$. Soient P_1, P_2 et P_3 des probabilités sur $\mathcal{R}$, les extensions des hypothèses (1) et (2) sur P sont données par :
(A_1) : $P_1 \preceq P_2$ ou $P_2 \preceq P_1$;
(A_2) : si $P_1 \preceq P_2$ et $P_2 \preceq P_3$, alors $P_1 \preceq P_3$.

L'existence de l'ordre $\preceq$ sur $\mathcal{P}$ est fondée sur l'hypothèse qu'il existe une récompense optimale, et donc qu'il existe au moins un ordre partiel sur les récompenses, même quand elles sont aléatoires. Ainsi s'il existe une fonction d'utilité U définie sur $\mathcal{R}$ associée à $\preceq$, alors la relation $P_1 \preceq P_2$ peut être caractérisée par l'inégalité des espérances correspondantes

$$\mathbb{E}_{P_1}[U(r)] \leq \mathbb{E}_{P_2}[U(r)].$$

Les conditions d'existence de la fonction d'utilité U sur $\mathcal{R}$ sont données ici. Pour cela caractérisons d'abord la notion de distribution de mélange. En effet, si nous considérons le groupe des distributions bornées $\mathcal{P}_b$, et P_1, P_2 deux distributions de $\mathcal{P}_b$, alors la *distribution de mélange* est définie par $P = \alpha P_1 + (1-\alpha)P_2$ et est vue comme la distribution qui génère une récompense de P_1 avec la probabilité α et une récompense de P_2 avec la probabilité $1-\alpha$. En plus de (A_1) et (A_2), des hypothèses (ou axiomes) supplémentaires nécessaires à l'existence d'une fonction d'utilité définie sur $\mathcal{R}$ sont données par :
(A_3) : si $P_1 \preceq P_2$, $\alpha P_1 + (1-\alpha)P \preceq \alpha P_2 + (1-\alpha)P$, pour tout $P \in \mathcal{P}$.
L'axiome (A_3) peut être illustré par le fait que si des actionnaires peuvent comparer deux compagnies avec des distributions des dividendes P_1 et P_2, ils doivent pouvoir garder le même classement s'il y a une probabilité $(1-\alpha)$ que les deux dividendes soient remplacées par des bons du Trésor avec une distribution de dividendes P. L'axiome qui suit établit que la relation d'ordre $\preceq$ doit être connexe (ou fermée) :
(A_4) : si $P_1 \preceq P_2 \preceq P_3$, il existe α et $\beta \in]0,1[$ tel que
$$\alpha P_1 + (1-\alpha)P_3 \preceq P_2 \preceq \beta P_1 + (1-\beta)P_3.$$
La condition (A_4) conduit alors au résultat suivant.

Lemme 1.1 (Ch. Robert, 2006)**.** *Si r_1, r_2 et r sont des récompenses dans $\mathcal{R}$, avec $r_1 \prec r_2$ et $r_1 \preceq r \preceq r_2$, il existe un seul v, $(0 \leq v \leq 1)$ tel que $r \sim vr_1 + (1-v)r_2$.*

Dans ce Lemme les relation $\prec$ et $\sim$ désignent respectivement l'ordre strict et la relation d'équivalence sur $\mathcal{R}$. Le Lemme est ainsi le point essentiel pour la construction de la fonction d'utilité, U, dans $\mathcal{R}$. En effet, pour r_1 et r_2 deux récompenses arbitraires telles que $r_2 \prec r_1$, nous pouvons définir la fonction U de la façon suivante.
Pour tout $r \in \mathcal{R}$ on a,
(i) $U(r) = v$ si $r_2 \preceq r \preceq r_1$ et $r \sim vr_1 + (1-v)r_2$;
(ii) $U(r) = \frac{-v}{1-v}$ si $r \preceq r_2$ et $r_2 \sim vr_1 + (1-v)r$;
(iii) $U(r) = \frac{1}{v}$ si $r_1 \preceq r$ et $r_1 \sim vr + (1-v)r_2$.
En particulier on a, $U(r_1) = 1$ et $U(r_2) = 0$. De plus, cette fonction U conserve la relation d'ordre sur $\mathcal{R}$ (DeGroot, 1970). L'extension de la définition de fonction d'utilité est faite non pas pour l'espace $\mathcal{P}$, mais pour l'espace $\mathcal{P}_b$) (Robert, 2006 [45]). L'espace $\mathcal{P}_b$ correspond aussi aux distribution (lois) à support borné, pour lesquelles il existe r_1 et r_2 de $\mathcal{R}$ tels que

$$[r_1, r_2] = \{r : r_1 \preceq r \preceq r_2\} \quad \text{et} \quad P([r_1, r_2]) = 1.$$

1.2.2 Fonction de coût en analyse statistique

La théorie de la décision est basée sur le modèle statistique vu sous l'angle décisionnel incluant ainsi trois espaces : l'espace des observations $\mathcal{X}$, l'espace des paramètres Θ et l'espace des décisions (ou espace d'action) $\mathcal{D}$. L'inférence statistique consiste alors à prendre une décision $d \in \mathcal{D}$ par rapport au paramètre $\theta \in \Theta$, fondée sur l'observation $x \in \mathcal{X}$. Dans la plupart des cas, la décision d consiste à évaluer (ou estimer) une fonction de θ, $h(\theta)$, le plus précisément possible. Une hypothèse forte de la Théorie de la Décision est que chaque action d peut être évaluée (ce qui signifie que la précision peut être quantifiée) et conduit à une récompense r, avec une utilité $U(r)$ (qui existe sous l'hypothèse de rationalité des décideurs). Pour la suite, cette utilité sera notée $U(\theta, d)$ pour insister sur le fait qu'elle dépend uniquement de ces deux facteurs. En supposant les récompenses r aléatoires et que celles-ci interviennent dans l'utilité U, nous écrirons cette dernière sous forme d'espérance : $U(\theta, d) = \mathbb{E}_{\theta,d}[U(r)]$. L'utilité peut alors être vue comme une mesure de proximité entre l'estimation proposée d et la vraie valeur $h(\theta)$.

Une fois la fonction d'utilité U établie, la fonction de coût L correspondante s'en déduit facilement puisque l'une est l'opposée de l'autre :

$$L(\theta, d) = -U(\theta, d)$$

En général, la fonction de coût est supposée positive, ce qui implique que la fonction d'utilité correspondante est négative : $U(\theta, d) \leq 0$. Ceci veut qu'il n'existe pas de décision ayant une utilité infinie. L'hypothèse de l'existence d'un minorant pour la fonction de coût L peut être critiquée comme trop stricte, mais elle évite des paradoxes. On peut aussi soutenir que, d'un point de vue statistique, la fonction de coût L représente bien le coût (ou l'erreur) dû à une mauvaise évaluation de la fonction d'intérêt $h(\theta)$. Par conséquent la meilleure évaluation possible de $h(\theta)$, lorsque θ est connu, peut entrainer au mieux un coût nul. Notons que sauf pour des cas triviaux, il est généralement impossible de minimiser uniformément, en d, la fonction de coût $L(\theta, d)$ quand θ est inconnu. Dans cette logique que l'approche fréquentiste propose de considérer plutôt le coût moyen, encore appelé *risque fréquentiste*, pour obtenir un critère de comparaison utilisable à partir d'une fonction de coût dans un contexte aléatoire,

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta(X))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x \end{aligned}$$

où $\delta(x)$ est la règle de décision, soit l'attribution d'une décision à chaque résultat $x \sim f(x|\theta)$. Alors $\delta(X)$ est appelée *estimateur*, tandis que la valeur $\delta(x)$ est appelée *estimation* de $h(\theta)$.

L'approche bayésienne de la Théorie de la Décision adopte une démarche plutôt différente, car considérant le paramètre ? comme inconnu, alors l'observation x tout à fait connu. Ce qui fait que pour le coût moyen l'intégrale est effectuée sur l'espace Θ, plutôt que sur l'espace $\mathcal{X}$,

$$\begin{aligned} \rho(\pi, \delta|X) &= \mathbb{E}^\pi[L(\theta, \delta(X))|X] \\ &= \int_\Theta L(\theta, \delta(x)) \pi(\theta|x) \mathrm{d}\theta \,, \end{aligned} \tag{1.8}$$

qui est appelé *coût moyen a posteriori*, car évaluant le coût moyen par rapport à la distribution a posteriori du paramètre θ conditionnellement à la valeur observée x. Ainsi pour un x donné, le coût moyen ou encore l'erreur moyenne résultant de la décision δ est donné par $\rho(\pi, \delta|x)$. Ce coût est ainsi une fonction de x, mais cette dépendance n'est pas gênante, contrairement à la dépendance fréquentiste du risque au paramètre puisque x, à la différence de θ, est connu.

En se donnant une distribution a priori π, il est aussi possible de définir le *risque intégré*, qui est le risque fréquentiste moyenné sur les valeurs de θ par rapport à leur distribution a priori. Ce risque est donné par,

$$\begin{aligned} r(\pi, \delta) &= \mathbb{E}^\pi[R(\theta, \delta)] \\ &= \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x \pi(\theta) \mathrm{d}\theta. \end{aligned}$$

Un intérêt particulier de ce deuxième concept est qu'il associe un nombre réel à chaque estimateur, et non une fonction de θ. Il induit donc un ordre total sur l'ensemble des estimateurs et permet une comparaison directe entre ces estimateurs. Cela implique que, quoique prenant en compte l'information a priori via la distribution a priori, l'approche bayésienne est suffisamment réductrice (dans un sens positif) pour atteindre une décision efficace.

Le risque intégré et le risque a posteriori sont en effet équivalents puisqu'ils conduisent à la même décision, comme le montre résultat suivant

Théorème 1.2 (Ch. Robert, 2006)**.** *Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in \mathcal{X}$, de la valeur $\delta(x)$ qui minimise le coût moyen a posteriori, $\rho(\pi, \delta|x)$, puisque*

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x) \mathrm{d}x, \quad \textit{où} \quad m(x) = \int_\Theta f(x|\theta) \pi(\theta) \mathrm{d}\theta.$$

Ce résultat mène à la définition suivante d'un estimateur de Bayes.

Définition 1.10. *Un estimateur de Bayes associé à une distribution a priori π et une fonction de coût L est un estimateur $\delta^\pi(X)$ minimisant $r(\pi, \delta)$. Pour chaque $X \in \mathcal{X}$, ce dernier est donné par*

$$\delta^\pi(X) = \arg\min_\delta \rho(\pi, \delta|X) \tag{1.9}$$

$r(\pi, \delta^\pi(X))$ est alors appelé ***risque de Bayes****.*

Le Théorème 1.2 fournit ainsi un outil constructif pour la détermination des estimateurs de Bayes. Notons que, d'un point de vue strictement bayésien, seul le coût moyen a posteriori $\rho(\pi, \delta \,|\, x))$ compte, puisque le paradigme bayésien est fondé sur une approche conditionnelle. Faire la moyenne sur toutes les valeurs possibles de x, alors que nous connaissons la valeur observée de x, semble être une perte d'information. Néanmoins, l'équivalence présentée par le Théorème 1.2 est important parce que, d'abord, elle montre que l'approche conditionnelle n'est pas nécessairement aussi dangereuse que les critiques fréquentistes peuvent l'indiquer. En effet, bien que l'approche bayésienne fonctionne de façon conditionnelle à l'observation x, elle inclut aussi les propriétés probabilistes de la distribution de l'observation $f(x|\theta)$. Ensuite, cette équivalence fournit une connexion entre les résultats classiques de la Théorie des Jeux et l'approche axiomatique bayésienne, fondée sur la distribution a posteriori.

Le calcul de tous ces risques nécessite la connaissance au préalable de la fonction de coût. Or en pratique, il est souvent difficile de construire la fonction de coût véritable associée au problème de décision. Dans ce cas des fonctions de coût classiques sont souvent utilisées.

1.2.3 Fonctions de coût usuelles

Quand le contexte d'une expérience ne permet pas une détermination de la fonction d'utilité (manque de temps, information, etc.), une alternative courante est de faire appel à des fonctions de coût classiques, qui sont mathématiquement simples et de propriétés connues. Bien entendu, cette approche est une approximation sous-jacente du modèle statistique et ne devrait être utilisée que quand la fonction d'utilité n'est pas disponible.

1.2.3.1 Le coût quadratique

Introduit par Legendre (1805) et Gauss (1810), ce coût est sans conteste le critère d'évaluation le plus commun. Fondant sa validité sur l'ambiguïté de la notion d'erreur dans un contexte statistique (soit erreur de mesure, soit variation aléatoire), il a aussi donné lieu à de nombreuses critiques, la plus fréquente étant sans doute le fait que le coût quadratique pénalise trop fortement les grandes erreurs. Dans son article de 1810, Gauss reconnaissait déjà l'arbitraire du coût quadratique mais le défendait au nom de la simplicité. Bien que les critiques concernant l'utilisation systématique de la fonction de coût quadratique soient fondées, son usage est néanmoins très répandu, car il donne en général des solutions bayésiennes qui sont celles naturellement fournies comme estimateurs pour une inférence non décisionnelle fondée sur une distribution a priori. En effet, les estimateurs de Bayes associés au coût quadratique sont les moyennes a posteriori. Cependant, notons que le coût quadratique n'est pas le seul coût à avoir cette caractéristique. Les fonctions de coût conduisant à la moyenne a posteriori comme estimateur de Bayes sont appelées fonctions de coût propres et ont été identifiées par Lindley (1985), Schervish (1989), der Meulen B. (1992), et Hwang et Pemantle (1994). Ce coût s'exprime par le carré de la différence entre la vraie valeur du paramètre θ (ou de $h(\theta)$) et de son estimation δ et est donnée par

$$L(\theta, \delta) = (\theta - \delta)^2 \tag{1.10}$$

Cependant, les fonctions de coût convexes comme (1.10) ont l'avantage incomparable d'éviter le paradoxe de *risk lovers*[4] et d'exclure les estimateurs randomisés.

4. Il s'agit d'un amoureux des risques est une personne qui est prête à prendre plus de risques tout en investissant afin de gagner des rendements plus élevés.

Des résultats relatifs à la détermination de l'estimateur de Bayes sous le coût quadratique sont ici donnés.

Proposition 1.1 (Ch. Robert, 2006). *L'estimateur de Bayes δ^π associé à la loi a priori $\pi(\theta)$ et au coût quadratique* (1.10) *est la moyenne a posteriori*

$$\delta^\pi(X) = \mathbb{E}^\pi[\theta|X] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}.$$

Les corollaires suivants se déduisent de manière immédiate.

Corollaire 1.1 (Ch. Robert, 2006). *L'estimateur de Bayes δ^π associé à $\pi(\theta)$ et au coût quadratique pondéré $L(\theta,\delta) = \omega(\theta)(\theta-\delta)^2$, où $\omega(\theta)$ est une fonction positive, est*

$$\delta^\pi(X) = \frac{\mathbb{E}^\pi[\omega(\theta)\theta|X]}{\mathbb{E}^\pi[\omega(\theta)|X]}.$$

Corollaire 1.2 (Ch. Robert, 2006). *Quand $\Theta \in \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à $\pi(\theta)$ et au coût quadratique $L(\theta,\delta) = (\theta-\delta)^t Q(\theta-\delta)$ est la moyenne a posteriori, $\delta^\pi = \mathbb{E}^\pi[\theta|X]$, pour toute matrice Q de taille $p \times p$, symétrique définie positive.*

Le coût quadratique est particulièrement intéressant lorsque l'espace des paramètres est borné et le choix d'un coût plus subjectif est impossible. En effet, ce coût est assez simple d'utilisation et l'erreur d'approximation est alors de faible importance. L'indétermination de la fonction de coût (et son remplacement par une approximation quadratique) est fréquente en évaluation de la précision, qui inclut par exemple l'estimation du coût (Rukhin, 1988a,b, Lu et Berger, 1989a,b, Hwang et al., 1992, Robert et Casella, 1993, 1994, et Fourdrinier et Wells, 1993).

1.2.3.2 Le coût absolu

C'est une mesure d'erreur proposée comme une alternative au coût quadratique en dimension 1. Il s'agit en effet d'un coût déjà évoqué par Laplace (1773) et qui s'exprime par

$$L(\theta,\delta) = |\theta-\delta| \tag{1.11}$$

Il désigne comme son nom l'indique la valeur absolue de la différence entre le paramètre θ et son estimation δ. Une généralisation de ce coût est donnée par une fonction linéaire par morceaux

$$L_{k_1,k_2} = \begin{cases} k_2(\theta-\delta) & \text{si } \theta > \delta, \\ k_1(\delta-\theta) & \text{sinon.} \end{cases} \tag{1.12}$$

Notons que si dans (1.12) si $k_1 = k_2$, nous retrouvons (1.11). De tels fonctions croissent plus lentement que le coût quadratique. Par, conséquent, tout en restant convexes, elles ne surpénalisent pas des erreurs grandes mais peu vraisemblables. Le résultat qui suit permet la détermination de l'estimateur de Bayes sous le coût (1.12).

Proposition 1.2 (Ch. Robert, 2006). *L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût linéaire par morceaux* (1.12) *est le fractile $k2/(k1+k2)$ de la loi a posteriori $\pi(\theta|x)$.*

Dans certains cas le besoin de mélanger coût quadratique et coût absolu est ressenti. C'est le cas avec Hubert (1964a) qui propose un mélange des deux coûts afin de maintenir une pénalisation quadratique aux alentours de 0. Ce coût est donné par

$$L(\theta,\delta) = \begin{cases} (\delta-\theta)^2 & \text{si } |\delta-\theta| < k, \\ 2\,k|\delta-\theta| - k^2 & \text{sinon.} \end{cases}$$

Ainsi bien que convexe, ce coût mixte ralentit la progression du coût quadratique pour des grandes erreurs et robustifie son effet. Malheureusement, il n'existe pas en général de formule explicite des estimateurs de Bayes sous cette fonction de coût.

1.2.3.3 Le coût 0 - 1

Ce coût est surtout utilisé dans l'approche classique des tests d'hypothèse, proposée par Neyman et Pearson. Plus généralement, c'est un exemple typique d'un coût non quantitatif. En effet, pour ce coût, la pénalité associée à un estimateur δ est 0 si la réponse est correcte et 1 sinon. Soit le test de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$, où H_0 est appelée l'hypothèse nulle. Alors $\mathcal{D} = \{0, 1\}$, où 1 représente l'acceptation de H_0 et 0 son rejet. En d'autres termes, la fonction de θ estimée est $\mathbb{1}_{\Theta_0}(\theta)$. Alors la fonction de coût 0 - 1 est définie par

$$L(\theta, \delta) = \begin{cases} 1 - \delta & \text{si } \theta \in \Theta_0 \\ \delta & \text{sinon.} \end{cases} \tag{1.13}$$

Le risque a posteriori correspondant est alors le suivant :

$$\rho(\pi, \delta|X) = \mathbb{1}_{\delta=0} P^\pi(\theta \in \Theta_0|X) + \mathbb{1}_{\delta=1} P^\pi(\theta \notin \Theta_0|X),$$

qui permet d'obtenir l'estimateur de Bayes associé

Proposition 1.3. *L'estimateur de Bayes associé à π et au coût* (1.13) *est*

$$\delta^\pi(X) = \begin{cases} 1 & si\ P^\pi(\theta \in \Theta_0|X) > P^\pi(\theta \notin \Theta_0|X), \\ 0 & sinon \end{cases}$$

Ainsi l'estimation permet d'accepter H_0 si c'est l'hypothèse la plus probable a posteriori, ce qui est une réponse naturelle.

1.3 Méthodes de calcul bayésien

Nous remarquons la place centrale qu'occupe la loi a posteriori dans la construction ou la recherche de l'estimateur de Bayes. D'ailleurs la version bayésienne du principe de vraisemblance implique que toute l'inférence sur θ repose entièrement sur la loi a posteriori $\pi(\theta|x)$. Cette dernière permet de décrire les propriétés du paramètre θ (qui est regardé comme une variable aléatoire). Alors les indicateurs résumant $\pi(\theta|x)$ tels que moyenne, mode, variance et médiane a posteriori sont par exemples des estimateurs potentiels de θ ou de $h(\theta)$. Notamment, lorsque la quantité d'intérêt est $h(\theta)$, un estimateur possible de $h(\theta)$ est la moyenne a posteriori $\mathbb{E}^\pi[h(\theta)|X]$. La diversité des estimateurs candidats à l'approximation de $h(\theta)$ pose la question du choix. Celui-ci est rendu possible par la considération d'une fonction de coût afin de comparer les différents estimateurs potentiels. Comme indiqué plus haut, sous le critère de cette fonction de coût, le meilleur choix, donné par (1.9), est appelé est estimateur de Bayes. Ce dernier dépend naturellement de la loi a priori mais aussi du critère coût choisi comme nous pouvons le voir avec les fonctions de coût classiques (coût quadratique, coût absolu, coût 0 - 1) abordées ci-dessus. Dans beaucoup de cas l'expression analytique de l'estimateur de Bayes est connue. C'est le cas notamment des lois a priori conjuguées des familles exponentielle de paramètres naturels. Sous le critère de coût quadratique les estimateurs de Bayes sont données pour certaines lois bien connues de familles exponentielles. Dans le tableau 1.2, nous remarquons que l'estimateur de Bayes correspond à la moyenne a posteriori du paramètre θ.

Cependant dans beaucoup d'autres cas la difficulté d'obtenir, pour un estimateur de Bayes, une expression analytique se pose. Cela peut être lié à plusieurs facteurs comme la distribution paramétrique des données $f(x|\theta)$, le choix de la loi a priori $\pi(\theta)$ ou aussi de la fonction de coût $L(\theta, \delta)$. La contribution de tous ces facteurs se résume dans le coût moyen a posteriori $\rho(\pi, \delta|X)$ exprimé dans (1.8). La forme intégrale de ce coût rend souvent difficile sa manipulation, notamment le calcul de l'intégrale par une procédure analytique. La première difficulté réside dans le choix de la fonction de coût. En effet lorsque le coût quadratique est choisi l'estimateur de Bayes est la moyenne a posteriori

$$\begin{aligned} \delta^\pi(X) &= \int_\Theta \theta \pi(\theta|x) \mathrm{d}\theta \\ &= \frac{\int_\Theta \theta \pi(\theta) f(x|\theta) \mathrm{d}\theta}{\int_\Theta \pi(\theta) f(x|\theta) \mathrm{d}\theta} \end{aligned}$$

Loi de x	Loi conjuguée	Moyenne a posteriori
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\frac{\mu\sigma^2+\tau^2 x}{\sigma^2+\tau^2}$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\frac{\alpha+x}{\beta+1}$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\frac{\alpha+\nu}{\beta+x}$
Binomiale $\mathcal{B}(n, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
Binomiale négative $\mathcal{N}eg(n, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\frac{\alpha+1}{\beta+(\mu-x)^2}$
Multinomiale $\mathcal{M}(n; \theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\frac{\alpha_i+x_i}{n+\sum_j \alpha_j}$

TABLE 1.2: Estimateurs de Bayes du paramètre θ sous coût quadratique pour les lois a priori conjuguées des familles exponentielles usuelles.

qui dans beaucoup de situations ne possède pas d'expression analytique du fait des intégrales. Aussi dans beaucoup de cas même si le numérateur de $\delta^\pi(X)$ est explicite le dénominateur $\int_\Theta \pi(\theta) f(x|\theta) d\theta$ peut être inconnu. C'est notamment le cas où la loi a posteriori $\pi(\theta|x)$ est seulement connue à une constante près, c'est-à-dire $\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$. La seconde difficulté, qui par ailleurs est déjà mise en évidence par le coût quadratique, est liée à l'expression de la loi a posteriori. Si l'expression analytique de $\pi(\theta|x)$ n'est pas connue il est souvent difficile d'expliciter $\rho(\pi, \delta|X)$ et de connaître ses propriétés afin de déterminer l'estimateur de Bayes $\delta^\pi(X)$ dans (1.9). Il arrive très souvent aussi que même si la loi a posteriori est entièrement connue, le coût moyen a posteriori $\rho(\pi, \delta|X)$ soit difficile à calculer analytiquement.
Alors l'alternative consistant à n'utiliser que des modèle d'échantillonnage, des lois a priori et des fonctions de coût qui mènent à des solutions explicites pour l'estimateur de Bayes est réductrice et n'est pas souvent adéquat au problème de décision posé. A ce titre plusieurs techniques, pour palier cette difficulté, ont été proposées. Il s'agit essentiellement de stratégies d'approximation de quantités a posteriori qui s'expriment sous forme d'intégrales. Ces méthodes sont principalement de deux ordres : les méthodes d'analyses numériques et les techniques statistiques. Ces dernières qui vont nous intéresser ici, peuvent être divisées en deux catégories : les méthodes d'*intégrations de Monte Carlo* et les méthodes de *Monte Carlo par Chaines de Markov (MCMC).*

1.3.1 Les méthodes Monte Carlo par Chaines de Markov

Les méthodes *Monte Carlo par Chaines de Markov* s'inscrivent dans une continuité des travaux de Metropolis dans le domaine de la physique de la matière condensée (Metropolis et al. (1953) [49]). Elles permettent l'estimation des moyennes de grandeurs physiques données par la formulation de Gibbs de la mécanique statistique sous la forme d'intégrales multidimensionnelles. Ce sont des méthodes de simulations qui reposent essentiellement sur la génération d'une chaîne de Markov et sont de deux types principalement : l'*algorithme de Metropolis-Hastings* et l'*échantillonneur de Gibbs*. Nous nous intéressons ici à ce dernier type d'échantillonnage.

1.3.1.1 Échantillonneur de Gibbs

Le nom *échantillonnage de Gibbs* est apparu pour la première fois dans l'article de Geman et Geman (1984) [50], qui sont les premiers à appliquer un échantillonneur de Gibbs sur un champ

aléatoire de Gibbs [5]. Il s'agit en fait d'un cas particulier de l'algorithme de Metropolis -Hastings. En effet les travaux de Geman et Geman (1984) s'appuyaient sur ceux de Metropolis et al. (1953), Hastings (1970) et Peskun (1973), et ont poussé Gelfand et Smith (1990) à écrire l'article qui a suscité le renouveau bayésien dans les années 90 grâces à ces algorithmes stochastiques.

Étant donné une distribution de probabilité π sur un espace d'états E qui est généralement un espace vectoriel (topologique) de dimension fini $n > 1$, cet algorithme génère une chaîne de Markov dont la distribution stationnaire (loi limite) est π. Il permet ainsi de tirer aléatoirement un élément de E selon π (on parle d'échantillonnage). On effectuer des tirages aléatoires suivant la loi π (loi stationnaire), ne pouvant pas le faire directement par des simulations i.i.d parce que soit π n'est pas une loi classique facile à simuler, soit π est seulement connue à une constante près, c'est-à-dire $\pi(x) \propto c\,\phi(x)$. Dans les deux cas il est nécessaire d'obtenir les lois conditionnelles de chacune des composantes du vecteur X. En effet si $E = \mathbb{R}^n$ alors $X = (X_1, X_2, \ldots, X_n)$, les lois conditionnelles sont toutes proportionnelles à la loi cible $\pi(x)$, dans le premier cas, c'est-à-dire $\pi_i(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \propto \pi(x)$, $i = 1, \ldots, n$ ou bien proportionnelles à $\phi(x)$, dans le second cas, c'est-à-dire, $\pi_i(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \propto \phi(x)$, $i = 1, \ldots, n$. Ainsi l'algorithme étant itératif, à chaque itération k, un vecteur $X^{(k)} = (X_1^{(k)}, X_2^{(k)}, \ldots, X_n(k))$ est généré composante par composante. Par exemple la composante $X_i^{(k)}$ est générée par la loi conditionnelle $\pi_i(x_i|x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)})$. Plus en détails nous avons l'algorithme suivant :

Données : Loi cible π ou bien loi proportionnelle à π, soit ϕ,
et les lois conditionnelles correspondantes
Résultat : Chaîne de Markov $(X_n,\ n \geq 0)$
Initialisation : choisir une valeur initiale pour le vecteur, soit $X^{(0)}$.;
pour k *allant de 1 à Kmax* **faire**
simuler la première composante $X_1^{(k)}$;
$X_1^{(k)} \sim \pi_1\big(x_1|x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}\big)$;
$\vdots$
simuler la $i$ème composante $X_i^{(k)}$;
$X_i^{(k)} \sim \pi_i\big(x_i|x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k-1)} \ldots, x_n^{(k-1)}\big)$;
$\vdots$
simuler la dernière composante $X_n^{(k)}$;
$X_n^{(k)} \sim \pi_n\big(x_n|x_1^{(k)}, x_2^{(k)}, \ldots, x_{n-1}^{(k)}\big)$;
fin

Algorithme 1 : Échantillonneur de Gibbs

On remplace donc le problème de génération de chaque vecteur $X^{(k)}$ par n sous-problèmes plus simples. La suite des vecteurs générés est une chaîne de Markov irréductible dont la distribution d'intérêt π est invariante. En effet, supposons que le vecteur $X^{(k)}$ est généré par la loi $\pi(x)$, soit $\big(X_1^{(k)}, X_2^{(k)}, \ldots, X_n^{(k)}\big) \sim \pi(x_1, x_2, \ldots, x_n)$. Alors

$$\begin{aligned}
\big(X_1^{(k+1)}, X_2^{(k)}, \ldots, X_n^{(k)}\big) &\sim \pi(x_1|x_2, \ldots, x_n)\pi(x_2, \ldots, x_n) \\
&\sim \pi(x_1, x_2, \ldots, x_n) \\
&\vdots \\
\big(X_1^{(k+1)}, X_2^{(k+1)}, \ldots, X_n^{(k+1)}\big) &\sim \pi(x_n|x_1, \ldots, x_{n-1})\pi(x_1, \ldots, x_{n-1}) \\
&\sim \pi(x_1, x_2, \ldots, x_n)\,,
\end{aligned}$$

5. Le champ aléatoire de Gibbs a ainsi donné le nom à cet algorithme

ce qui montre que la distribution d'intérêt $\pi(x)$ est bien invariante.

A la différence de l'algorithme de Metropolis-Hastings, l'échantillonneur de Gibbs possède n, $(n > 1)$ noyaux de transition π-invariants définis par

$$P_i(x, y) = \pi_i(y_i|x_{-i})\mathbb{1}_{\{x_{-i}=y_{-i}\}}\ ,\ \forall x, y \in \mathbb{R}^n$$

où $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ et $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, les $\pi_i(.|x_{-i})\ ,\ i = 1, \ldots, n$ sont les lois conditionnelles des composantes Y_i, $i = 1, \ldots, n$ d'un vecteur Y pour passer d'un état x à un état y. Nous distinguons principalement deux types d'échantillonneurs de Gibbs :

Échantillonneur par balayage systématique. Pour ce type d'algorithme, on visite séquentiellement l'ensemble des indices $N = \{1, 2, \ldots, n\}$, en relaxant à chaque pas i la valeur suivant la loi π_i conditionnelle à l'état courant. La transition de x à y s'écrit :

$$P(x, y) = \prod_{i=1}^{n} \pi_i(y_i|y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n).$$

C'est le type d'algorithme correspondant à l'Algorithme 1 et qui est le plus souvent utilisé pour décrire l'échantillonneur de Gibbs. Cependant à côté de celui-ci, il y a aussi autre algorithme.

Échantillonneur par balayage aléatoire. Soit ν une probabilité jamais nulle sur $N = \{1, 2, \ldots, n\}$. A chaque pas, un indice i est choisi avec la probabilité $\nu_i > 0$ et la valeur y est relaxée selon la loi conditionnelle π_i à l'état courant. La transition s'écrit :

$$P(x, y) = \sum_{i=1}^{n} \nu_i \pi_i(y_i|x_{-i})\mathbb{1}_{\{x_{-i}=y_{-i}\}}.$$

Si l'espace des états E est fini, la transition P est positive strictement.

D'après le théorème de Hammersley-Clifford la loi jointe $\pi(x)$ est caractérisée par ses lois conditionnelles.

Dimension 2. Si la loi jointe $\pi(x_1, x_2)$ (densité dans le cas continu) a des lois conditionnelles $\pi_i(x_1|x_2)$ et $\pi_(x_2|x_1)$ alors d'après Hammersley et Clifford (1970)

$$\pi_i(x_1, x_2) = \frac{\pi_1(x_1|x_2)}{\int \pi_1(y|x_2)/\pi_2(x_2|y)\mathrm{d}y}$$

Généralisation. Une généralisation à la dimension n est donnée. En effet sous l'hypothèse de positivité, une loi jointe $\pi(x_1, \ldots, x_n)$ peut s'écrire

$$\pi(x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \frac{\pi_{l_i}(x_{l_i}|x_{l_1}, \ldots, x_{l_{i-1}}, x'_{l_{i+1}}, \ldots, x'_{l_n})}{\pi_{l_i}(x'_{l_i}|x_{l_1}, \ldots, x_{l_{i-1}}, x'_{l_{i+1}}, \ldots, x'_{l_n})},$$

pour toute permutation l définie sur $\{1, \ldots, n\}$ et tout $x' \in \mathbb{R}^n$

L'échantillonneur de Gibbs est particulièrement bien adapté aux modèles hiérarchiques. Dans ce cas :

. les paramètres inconnus sont munis de lois a priori ainsi que les hyperparamètres associés ;

. en général on introduit des lois non informatives au dernier niveau de la hiérarchie ;

. dans certains cas les paramètres du dernier niveau de la hiérarchie sont estimés en maximisant la vraisemblance marginale (si possible), ou en les estimant en même temps que les autres paramètres par des algorithmes MCMC.

Un exemple illustratif de modèle pour lequel on peut obtenir les lois conditionnelles est le modèle paramétrique suivant :

Exemple 1.8.
. Données Poissoniennes

$$\begin{cases} X_i & \sim & \mathcal{P}(\lambda_1) & \text{pour} & i = 1, \ldots, m \\ X_i & \sim & \mathcal{P}(\lambda_2) & \text{pour} & i = m+1, \ldots, n, \end{cases}$$

avec m connu.
. Lois a priori sur les paramètres

$$\lambda_1 \sim \mathcal{G}(\alpha, \beta) \quad , \lambda_2 \sim \mathcal{G}(\alpha, \beta) \ , \quad \alpha = 2.$$

. Loi a priori sur les hyperparamètres

$$f(\beta) = \frac{1}{\beta} \mathbb{1}_{\mathbb{R}^+}(\beta)$$

. Loi jointe : c'est celle du triplet (X, λ, β) où chaque composante est un « bloc » avec $X = (X_1, \ldots, X_n)$, $\lambda = (\lambda_1, \lambda_2)$ et $\beta \in \mathbb{R}$. L'indépendance intra et inter blocs donne la loi jointe suivante :

$$f(x, \lambda, \beta) \propto \frac{1}{\beta} \prod_{i=1}^{m} \left(\frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1} \right) \prod_{i=m+1}^{n} \left(\frac{\lambda_2^{x_i}}{x_i} e^{-\lambda_2} \right) \prod_{i=1}^{2} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i}$$

. Lois conditionnelles : ces lois sont celles des paramètres du modèle :

. pour le paramètre λ,

$$\begin{aligned} \lambda_1 | \beta, X &\sim \mathcal{G}\left(\alpha + \sum_{i=1}^{m} x_i \ , \ \beta + m \right) \\ \lambda_2 | \beta, X &\sim \mathcal{G}\left(\alpha + \sum_{i=m+1}^{n} x_i \ , \ \beta + n - m \right) \end{aligned}$$

. pour le paramètre β,

$$\beta | \lambda, X \sim \mathcal{G}(2\alpha, \lambda_1 + \lambda_2)$$

Conclusion

Depuis sa formalisation par Bayes puis Laplace, la modélisation bayésienne s'est montrée comme une alternative à l'analyse statistique classique (approche fréquentiste). Dans la finalité si les deux méthodologies sont équivalentes lorsque la taille des observations est assez grande, la statistique bayésienne s'avère plus performante en petites dimensions. Rappelons que leur différence réside sur la manière dont le paramètre du modèle est regardé. Si l'approche classique considère le paramètre comme une simple valeur inconnue qui devra être estimée, l'approche bayésienne regarde par contre le paramètre comme une véritable variable aléatoire au même titre que les données d'observations. Cette dernière considération du paramètre offre beaucoup plus de souplesse au modèle en ce sens que diverses contraintes peuvent à la fois être imposées sur le paramètre, surtout avec les modèles bayésiens hiérarchiques.

Ainsi dans ce chapitre des aspects essentiels de la modélisation bayésienne ont été présentés. D'abord nous avons introduit le paradigme bayésien qui est sous-tendu par une modélisation probabiliste donc qui suppose l'existence de lois de probabilité sur les observations, et qui constituent la vraisemblance du modèle. La spécificité de ce paradigme est la considération du paramètre comme une variable aléatoire, avec l'attribution d'une loi a priori résultante des croyances et/ou connaissances antérieures sur le sujet. Nous avons alors montré comment passer des informations a priori aux lois a priori, puis avons donné les différentes techniques de déterminations des lois a priori, en mettant en relief la méthode des lois a priori conjuguées. La dernière étape clé du paradigme a été l'affinement des connaissances a priori sur le paramètre. Celle-ci a nécessité l'application de la formule de Bayes, et est caractérisée par l'établissement de la loi a posteriori

du paramètre. Nous avons illustré l'usage de la formule de Bayes par les exemples historiques de Bayes (1763) et de Laplace (1773).
Ensuite nous avons abordé un aspect important de l'analyse statistique, qu'est la théorie de décision, dans le contexte de l'approche bayésienne. Ainsi dans la modélisation la théorie de la décision propose, à côté de l'espace des observations $\mathcal{X}$ et de celui des paramètres Θ, un troisième espace qu'est celui des décisions $\mathcal{D}$. En effet elle considère l'inférence sur le paramètre θ comme une décision d ou une action sur celui-ci et ayant ses conséquences comme le coût $L(\theta, d)$ ou même le coût moyen aussi appelé risque. La meilleure inférence est alors la décision qui minimise ce risque.
Enfin le dernier point du chapitre a consisté à la présentation des méthodes de calcul bayésien qu'on peut aussi appeler les outils d'inférence. Il s'agit des méthodes d'intégration de Monte Carlo et des méthodes Monte Carlo par Chaînes de Markov.

Chapitre 2

Factorisation de matrices positives

Introduction

Les méthodes statistiques de description de données sont nombreuses et diverses. En fonction des besoins, différentes méthodologies ont été développées, parmi celles-ci les techniques de factorisations matricielles. Les méthodes d'analyses factorielles comme l'ACP (Analyse en Composantes Principales), la méthode de décomposition en valeurs singulières, SVD (Singular Value Decomposition) et la méthode de quantification vectorielle, VQ (Vector Quantization) en sont les plus populaires parmi les méthodes précurseures. Comme alternative à ces dernières une nouvelle méthode d'analyse factorielle a été introduite par Lee et Seung (1999) [19]. Il s'agit de la méthode de Factorisation de Matrices Positives, NMF (*Nonnegative Matrix Factorization*). En effet le problème posé par la méthode NMF est le suivant : $X \approx UV$, où il est question de trouver une approximation (satisfaisante) de la matrice de données positives X par le produit matriciel UV avec $U \geq 0$ et $V \geq 0$. La différence avec les méthodes antérieures citées réside dans la contrainte de positivité des matrices facteurs U et V. Elle ne s'applique donc qu'à des matrices des données positives. La contrainte de positivité introduite par Lee et Seung a offert une autre manière de voir par rapport à la description des données : il s'agit de la description d'un tout à partir de la perception que l'on a de ses parties. A ce titre, cette approche a ouvert la voie à de nombreuses applications importantes : l'apprentissage non supervisé, le traitement d'images (segmentation, reconstitution), le traitement de signal, la vision artificielle ...

Le problème de factorisation posé est en fait un problème de minimisation. Pour le résoudre les auteurs Lee et Seung (2001) [20] ont proposé deux fonctions-coûts $F_1(U,V) = D_{KL}(X\|UV)$ et $F_2(U,V) = \|X - UV\|_F^2$ qui désignent respectivement la divergence de Kullback généralisée (cas discret) et le carré de la distance euclidienne entre deux matrices construite à partir de la norme de Frobenius. On peut remarquer que ces deux fonctions-coûts sont convexes par rapport à chacune des variables U et V et non par rapport au couple (U,V), ce qui fait que les techniques d'optimisations numériques proposées ne sont adaptées qu'à la détermination de minima locaux. En effet, Lee et Seung (2001) ont développé des méthodes numériques de minimisations qu'on appelle algorithmes NMF qui reposent sur des problèmes de minimisation où il s'agit de trouver un couple de matrices (U^*, V^*) qui minimise la fonction objectif dans chacun des deux cas. Les algorithmes NMF ainsi construits sont des procédures itératives de mises à jour multiplicatives, MU (Multiplicative Updates).

D'autres algorithmes de détermination des matrices U et V ont été proposés. Il s'agit de la méthode des moindres carrées alternées et positives, ANLS(Alternating Nonnegative Least Squares) (Zdunek et Cichocki, 2006 [36], Kim et al, 2007 [37], Kim et Park (2008) [39]) et de la méthode du gradient de descente (GD). Ces deux types d'algorithmes s'appliquent sur la fonction-coût $F_2(U,V) = \|X - UV\|_F^2$. Les algorithmes de Lee et Seung ainsi que les méthodes ANLS et GD constituent la classe des algorithmes NMF classiques. Par ailleurs il existe beaucoup d'autres variantes NMF qui s'inspirent des méthodes classiques. Leur particularité réside dans le fait qu'elles imposent des conditions supplémentaires. Il s'agit entre autres de l'algorithme avec contrainte de *parcimonie*, de l'algorithme avec contrainte de *symétrie* et la méthode du *graphe*

régularisé.

Ce chapitre s'articulera autour de deux points majeurs. Dans un premier temps nous donnerons les algorithmes NMF classiques. Ceux sont constitués des algorithmes de Lee et Seung, des méthodes de factorisation des moindres carrés alternés ANLS et de la procédure du gradient de descente. Dans un second temps nous présenterons quelques variantes de la factorisation NMF. Ces adaptations répondent le plus souvent à des préoccupations pratiques d'applications.

2.1 Les algorithmes NMF classiques

Les méthodes NMF classiques que nous présentons ici sont les algorithmes de mises à jour multiplicatives de Lee et Seung (2001) [20], les méthodes ANLS de Kim et Park (2008) [39] avec les algorithmes ANLS et enfin la méthode GD (gradient de descente).

2.1.1 Méthodes de mise à jour multiplicative de Lee et Seung

2.1.1.1 Formalisation du problème NMF

Le problème de description des données occupe une place centrale dans presque tous les aspects de la vie quotidienne : santé, économie, science, sciences sociales (géographie, sociologie ...), etc. A ce titre le traitement des données requiert d'abord une phase de collecte, puis un procédé méthodique à savoir la disposition des données recueillies sur les individus (unités statistiques) sous forme d'une matrice en vue d'une exploitation judicieuse. Les méthodes d'analyses factorielles dont fait partie la classe des méthodes NMF sont incontournables à ce sujet. Ainsi les méthodes NMF font partie de la famille de la statistique multivariée, dont l'objectif est de décrire un ensemble de variables observées (les données), au moyen de variables latentes (non observées). Pour comprendre les motivations de la formalisation du problème NMF tel que décrit par les auteurs Lee et Seung il faut remonter à leur article de 1999 [19] où la méthode NMF était proposée comme une nouvelle approche d'analyse factorielle basée sur le principe que la description d'un *tout* (*entité entière*) peut reposer sur la description de l'ensemble de ses *parties*. A ce titre une double illustration en est faite avec un exemple sur les caractéristiques sémantiques à l'intérieur d'un texte et un autre sur le traitement d'images. C'est ce dernier exemple qui va nous intéresser pour expliquer les motivations de la méthode NMF surtout en ce qui concerne la contrainte de positivité des facteurs. Si des méthodes antérieures d'analyses factorielles comme les procédures VQ, ACP ou encore SVD décrivaient déjà les données au moyen de facteurs latents, la spécificité des méthodes NMF réside dans la contrainte de positivité imposée sur les matrices U et V. Cela implique que les données d'entrée, à savoir la matrice X, sont toutes les valeurs positives. En effet si X représente la matrice de p images où chacune est un $n \times 1$ vecteur colonne, alors X est une matrice de taille $n \times p$. Le but de la procédure NMF, consiste à trouver la structure cachée qui explique le mieux les données. Celle-ci est contenue dans les deux matrices $U \in \mathbb{R}^{n\times k}$ et $V \in \mathbb{R}^{k\times p}$ telles que leur produit vérifie,

$$X \approx UV\ , \quad U \geq 0,\ V \geq 0. \tag{2.1}$$

où k, le rang de la factorisation, est généralement choisi tel que $(n+p)k < np$. Ainsi la contrainte de positivité des facteurs latents U et V trouve tout son fondement sur le principe que la perception d'un tout est le résultat de la perception de toutes ses parties constituantes (Lee et Seung, 1999, [19]). Ce principe était déjà utilisé antérieurement dans les théories de reconnaissance des objets (Biederman (1987) [51] et de Ullman (1996) [52]), ainsi que les théories physiologiques de représentation des objets dans cerveau (Wachsmuth et al. (1994) [53] et Logothetis et al. (1996) [54]). Ainsi l'interprétation des facteurs latents dans la factorisation NMF est que d'une part la matrice U représente l'ensemble des images de base, où chacune est seulement composée d'un certain nombre de caractéristiques du visage humain (front, nez, bouche, yeux,...) ; chaque colonne de U décrit donc une image de base. D'autre part les colonnes de la matrice V sont appelés les encodeurs. Ceux-ci sont, en fait, des spécificateurs en ce sens que toute différence entre deux colonnes d'encodeurs se traduit une différence entre les images correspondantes dans la matrice

X. Notons que chaque colonne de V permet la génération de l'image correspondante de X, par le biais d'une combinaison linéaire des images de base. En effet si,

$$X = \left[X^1, X^2, \ldots, X^p\right], \quad U = \left[U^1, U^2, \ldots, U^k\right] \quad \text{et} \quad V = \left[V^1, V^2, \ldots, V^p\right],$$

tels que $\forall\, j = 1, \ldots, p,\ V^j = (v_{1j}, v_{2j}, \ldots, v_{kj})$ alors,

$$X^j \approx v_{1j}U^1 + v_{2j}U^2 + \cdots + v_{kj}U^k \tag{2.2}$$

où $X^j,\ V^j,\ j = 1, \ldots, p,\ U^\ell,\ \ell = 1, \ldots, k$ désignent les vecteurs colonnes respectivement des matrices X, V et U. L'équation (2.2) montre que chaque image est une combinaison linéaire non pas d'images entières de faces humaines (donc pas de superposition), mais plutôt des caractéristiques (éléments constitutifs de la face). La positivité montre bien comment une image peut être décrite à partir de ses constituants possibles (les vecteurs colonnes de U) ; d'où la procédure de combinaison additive de (2.2).

Bien que les motivations premières de la méthode NMF étaient de régler un problème spécifique qu'est l'apprentissage d'objets à partir de leurs parties constituantes, notamment avec des applications sur une étude de texte et sur des images des faces humaine, la procédure NMF est un type d'algorithme qui peut s'adapter à divers problèmes : la classification non supervisée, le traitement de signal, l'apprentissage automatique, la vision artificielle ... Résoudre le problème (2.1) revient à traiter un problème de minimisation, d'où la considération de fonctions-coûts.

2.1.1.2 Fonctions coûts

Les fonctions-coûts considérées par Lee et Seung (2001) sont de deux types comme nous l'avons déjà dit. Il s'agit de la divergence de Kullback généralisée (cas discret) et du carré de la distance euclidienne entre deux matrices.

Divergence de Kullback. Si nous considérons deux matrices $A = (a_{ij})$ et $B = (b_{ij})$ de même taille $n \times p$, la divergence de Kullback-Leibler généralisée, dans le cas discret, entre A et B est donnée par

$$D_{KL}(A\|B) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(a_{ij}\log\frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij}\right) \tag{2.3}$$

Même si elle est aussi appelée distance, il ne s'agit pas d'une distance à proprement parler puisqu'elle ne vérifie que les propriétés de positivité et de séparation de la distance mais pas la symétrie et l'inégalité triangulaire. Par ailleurs la mesure de Kullback-Leibler généralisée se ramène à la divergence classique bien connue de Kullback-Leibler lorsque $\sum_{ij} a_{ij} = \sum_{ij} b_{ij} = 1$, de telle sorte que les matrices A et B peuvent être considérées comme des probabilités. Comme toutes les mesures de divergence en général, celle-ci a aussi la propriété, d'une part de posséder une borne inférieure égale à 0 et d'autre part la propriété qui fait que lorsque B tend vers A alors elle tend vers 0. Ce qui en fait une fonction-coût appropriée et permet de voir le lien entre approximer une matrice X par un produit matriciel UV et minimiser la fonction objectif $F_1(U, V)$. Le problème de minimisation se pose alors comme suit :
Problème 1. Minimiser la fonction-coût $F_1(U, V) = D_{KL}(X\|UV)$ sous les contraintes $U \geq 0$ et $V \geq 0$.

Distance euclidienne. L'autre fonction-coût utilisée pour quantifier la qualité de l'approximation de la matrice X par la factorisation UV est le carré de la distance euclidienne entre deux matrices, soient A et B. Notons que c'est une fonction déjà utilisée par Paatero et Tapper (1997) [55] pour formuler un problème de moindres carrées. Elle se construit par le biais de la norme de Frobenius et se définit par

$$\|A - B\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(a_{ij} - b_{ij})^2\,. \tag{2.4}$$

Comme toute distance classique sa borne inférieure est 0 (propriété de positivité) et n'est égale à 0 que lorsque $A = B$. Ajouté à cela, la propriété de continuité de la distance permet à celle-ci aussi de quantifier la qualité d'approximation entre deux matrices ; ce qui en fait également une fonction objectif qui convient. Le second problème de minimisation est alors :
Problème 2. Minimiser la fonction-coût $F_2(U, V) = \|X - UV\|_F^2$ sous les contraintes $U \geq 0$ et $V \geq 0$.

2.1.1.3 Algorithmes et convergence

Pour la résolution des problèmes de minimisation 1 et 2 Lee et Seung (2001) ont proposé des techniques numériques d'optimisation conduisant à la recherche de minima locaux. Les algorithmes qui en ont découlé, appelés *Multiplicative Update rules* (MU) ou règles de mises à jour multiplicatives sont considérés par les auteurs comme un bon compromis entre la vitesse de convergence et la facilité d'implémentation.

Algorithmes. Il s'agit des principes itératifs de mises à jour des matrices U et V qui ont été donnés par Lee et Seung (2001) sous forme de théorèmes.

Théorème 2.1. *La divergence de Kullback-Leibler généralisée $F_1(U, V) = D_{KL}(X\|UV)$ est décroissante lorsque les matrices U et V sont mises à jour suivant les règles ainsi données :*

$$u_{i\ell} \leftarrow u_{i\ell} \frac{\sum_j \left(x_{ij} / \sum_{\ell'} u_{i\ell'} v_{\ell' j}\right) v_{\ell j}}{\sum_j v_{\ell j}} , \quad v_{\ell j} \leftarrow v_{\ell j} \frac{\sum_i u_{i\ell} \left(x_{ij} / \sum_{\ell'} u_{i\ell'} v_{\ell' j}\right)}{\sum_i u_{i\ell}}. \tag{2.5}$$

La divergence $D_{KL}(X\|UV)$ est invariante sous ces règles de mises à jour si et seulement si (U, V) a atteint un point stationnaire de la fonction.

Théorème 2.2. *La distance euclidienne $\|X - UV\|_F$ est décroissante lorsque U et V sont mises à jour par rapport aux règles suivantes :*

$$u_{i\ell} \leftarrow u_{i\ell} \frac{\sum_j x_{ij} v_{\ell j}}{\sum_j \sum_{\ell'} u_{i\ell'} v_{\ell' j} v_{\ell j}} , \quad v_{\ell j} \leftarrow v_{\ell j} \frac{\sum_i u_{i\ell} x_{ij}}{\sum_{\ell'} \sum_i u_{i\ell} u_{i\ell'} v_{\ell' j}}. \tag{2.6}$$

La distance euclidienne est invariante sous ces règles de mises à jour si et seulement si (U, V) a atteint un point stationnaire de la fonction.

Remarque : Dans le théorème 2.2, l'on peut remarquer que la distance n'a pas été élevée au carré. C'est parce que minimiser $\|X - UV\|_F$ revient à minimiser $\|X - UV\|_F^2$. Nous avons repris la fonction telle que écrite par Lee et Seung (2001) [20].

Convergence. Les équations (2.5) et (2.6) des théorèmes précédents génèrent respectivement pour chacune de deux fonctions-coûts deux suites de matrices $(U^{(n)})$ et $(V^{(n)})$ dont la convergence vers le couple optimum (U^*, V^*) a été démontrée. Pour y arriver les auteurs ont fait usage de fonctions auxiliaires comme dans le cas de l'algorithme Espérance-Maximisation (EM) de Dempster et al (1977) [56]. Si nous considérons une fonction $F(h)$, une fonction $G(h, h')$ est dite fonction auxiliaire pour $F(h)$ si elle vérifie les conditions suivantes : $G(h, h') \geq F(h)$ et $G(h, h) = F(h)$. L'importance de la fonction auxiliaire réside dans le fait qu'elle permet la génération d'une suite pour laquelle la fonction $F(h)$ est décroissante comme le montre le lemme suivant, et aussi illustrée dans graphique 2.1.

Lemme 2.1. *Si G est une fonction auxiliaire pour F, alors la fonction F est décroissante par rapport à la suite définie par*

$$h^{(t+1)} = \arg\min_h G(h, h^{(t)}). \tag{2.7}$$

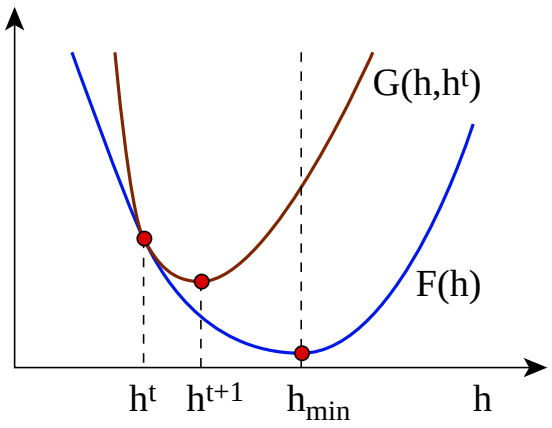

FIGURE 2.1: Illustration du rôle de la fonction auxiliaire dans la détermination du minimum de $F(h)$: minimiser la fonction auxiliaire $G(h, h') \geq F(h)$ garantit que $F(h^{(t+1)}) \leq F(h^{(t)})$ pour $h^{(t+1)} = \arg\min_h G(h, h^{(t)})$.

Les auteurs ont défini pour chacune des fonctions-coûts $F_1(U, V)$ et $F_2(U, V)$ les fonctions auxiliaires associées et on pourra remarquer que les règles de mises à jours multiplicatives (équations (2.5) et (2.6)) découlent de l'équation (2.7). Ainsi les deux lemmes qui vont suivre donnent respectivement la construction des fonctions auxiliaires pour les fonctions-coûts divergence $F_1(U, V)$ et carré de la distance euclidienne $F_2(U, V)$.

Lemme 2.2. *Si nous considérons les vecteurs colonnes des matrices X et V, soient $x = (x_1, \ldots, x_i, \ldots, x_n)^T$ une colonne quelconque de X et $v = (v_1, \ldots, v_\ell, \ldots, v_k)^T$ la colonne correspondante dans V, alors une fonction auxiliaire pour $v \mapsto F(v) = F_1(U, v)$ définie par*

$$F_1(U, v) = \sum_{i=1}^{n} x_i \log \left(\frac{x_i}{\sum_{\ell=1}^{k} u_{i\ell} v_\ell} \right) - x_i + \sum_{\ell=1}^{k} u_{i\ell} v_\ell \tag{2.8}$$

est la fonction

$$\begin{aligned} G(v, v^{(t)}) &= \sum_{i=1}^{n} (x_i \log x_i - x_i) + \sum_{i=1}^{n} \sum_{\ell=1}^{k} u_{i\ell} v_\ell \\ &\quad - \sum_{i=1}^{n} \sum_{\ell=1}^{k} x_i \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'=1}^{k} u_{i\ell'} v_{\ell'}^{(t)}} \left(\log u_{i\ell} v_\ell - \log \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'=1}^{k} u_{i\ell'} v_{\ell'}^{(t)}} \right) \end{aligned} \tag{2.9}$$

Lemme 2.3. *Sous les hypothèses du lemme précédent, si $K(v^{(t)})$ est la matrice diagonale de taille $k \times k$ et définie par*

$$K_{\ell\ell'}(v^{(t)}) = \delta_{\ell\ell'} (U^T U v^{(t)})_\ell / v_\ell^{(t)} \tag{2.10}$$

où δ est la matrice unité, alors

$$G(v, v^{(t)}) = F(v^{(t)}) + (v - v^{(t)})^T \nabla F(v^{(t)}) + \frac{1}{2} (v - v^{(t)})^T K(v^{(t)}) (v - v^{(t)}) \tag{2.11}$$

est une fonction auxiliaire pour $v \mapsto F(v) = F_2(U, v)$ définie par

$$F_2(U, v) = \frac{1}{2} \sum_{i=1}^{n} (x_i - \sum_{\ell=1}^{k} u_{i\ell} v_\ell)^2. \tag{2.12}$$

Les preuves de ces lemmes sont données dans Lee et Seung (2001) [20]. Nous remarquons dans ces lemmes que les fonctions objectifs considérées découlent des fonctions $F_1(U,V)$ et $F_2(U,V)$ appliquées au couple (U,v) et de variable le vecteur colonne v, la matrice U étant fixée. Cependant notons que $F_2(U,v)$ découle rigoureusement de $\frac{1}{2}F_2(U,V)$, c'est parce que minimiser $F_2(U,V)$ équivaut à minimiser $\frac{1}{2}F_2(U,V)$. Ce principe adopté par les auteurs consiste à diviser le problème d'approximation (2.1) en p (nombre de colonnes de X) sous-problèmes où il s'agit d'approcher chaque colonne x de la matrice X par le produit Uv, c'est-à-dire $x \approx Uv$, avec v la colonne correspondante de V. Cela se revient diviser chacun des problèmes 1 et 2 en p sous-problèmes par la considération des fonctions-coûts $F_1(U,v)$ et $F_2(U,v)$. Un travail analogue qui consiste à approximer la matrice X ligne par ligne le produit uV, où u est la ligne associée de la matrice U, permettra alors de considérer les fonctions-coûts $F_1(u,V)$ et $F_2(u,V)$. Il s'agit en fait de permuter les rôles des matrices U et V.

La convergence des règles MU dans (2.5) et (2.6) a été prouvé par les auteurs. Il s'agit pour chacune des fonctions auxiliaires données par (2.9) et (2.11) de résoudre le problème posé dans (2.7).

Preuve du Théorème 2.1. Le minimum de la fonction auxiliaire $v \mapsto G(v, v^{(t)})$ donnée par (2.9) est déterminé en résolvant les équations du gradient nul suivantes

$$\frac{\partial G(v, v^{(t)})}{\partial v_\ell} = 0\ ,\ \ell = 1, \ldots, k.$$

on obtient alors,

$$-\frac{1}{v_\ell}\sum_i x_i \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} + \sum_i u_{i\ell} = 0\,,\ \ell = 1, \ldots, k.$$

d'où on a,

$$v_\ell^{(t+1)} = \frac{v_\ell^{(t)}}{\sum_i u_{i\ell}} \sum_i \frac{u_{i\ell} x_i}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} \tag{2.13}$$

D'après le lemme 2.1 la suite de vecteurs $(v^{(t)})$ ainsi générée par (2.13) rend décroissante la fonction $F_1(U,v)$ et lui fait tendre vers son minimum. Si on pose $v = v_{.j}$ et $x = x_{.j}$, $j = 1, \ldots, p$, on obtient alors la mise à jour multiplicative de (2.5) des coefficients de la matrice V

$$v_{\ell j}^{(t+1)} \leftarrow v_{\ell j}^{(t)} \frac{\sum_i u_{i\ell} x_{ij} / \sum_{\ell'} u_{i\ell'} v_{\ell' j}^{(t)}}{\sum_i u_{i\ell}}.$$

En permutant les rôles des matrices U et V dans les lemmes 2.1 et 2.2 on obtient aussi de manière facile la mise à jour multiplicative des coefficients de U

$$u_{i\ell}^{(t+1)} \leftarrow u_{i\ell}^{(t)} \frac{\sum_j v_{\ell j} x_{ij} / \sum_{\ell'} u_{i\ell'}^{(t)} v_{\ell' j}}{\sum_j v_{\ell j}}\ ,\ \ell = 1, \ldots, k.$$

□

Si on pose $u = u_{i.}$, $i = 1, \ldots, n$, alors la suite de vecteurs lignes $(u^{(t)})$ ainsi construite fait décroître et converger la fonction $F(u) = F_1(u,V)$ convexe vers son minimum.

Preuve du Théorème 2.2. Considérons la fonction auxiliaire $G(v, v^{(t)})$ donnée par (2.11) la résolution du problème (2.7) donne

$$v^{(t+1)} = v^{(t)} - K(v^{(t)})^{-1} \nabla F(v^{(t)}) = v_\ell^{(t)} \frac{\sum_i u_{i\ell} x_i}{\sum_i \sum_{\ell'} u_{i\ell} u_{i\ell'} v_{\ell'}^{(t)}} \tag{2.14}$$

D'après le lemme 2.1, la fonction $F_2(U,v)$ décroît par rapport à la suite de vecteurs $(v^{(t)})$ générée par (2.14) et tend vers son minimum, puisqu'elle est convexe en v. Si on considère les vecteurs v

et x comme la j ième colonne respectivement des matrices V et X, c'est-à-dire $v = v_{\cdot j}$ et $x = x_{\cdot j}$, $j = 1, \ldots, p$, alors on retrouve les règles multiplicatives de mise à jour de V pour le carré de la distance euclidienne

$$v_{\ell j}^{(t+1)} \leftarrow v_{\ell j}^{(t)} \frac{\sum_i u_{i\ell} x_{ij}}{\sum_i \sum_{\ell'} u_{i\ell} u_{i\ell'} v_{\ell' j}^{(t)}}.$$

De même en permutant les rôles des matrices U et V dans les Lemmes 2.1 et 2.3 nous pouvons facilement obtenir la mise à jour de la matrice U

$$u_{i\ell}^{(t+1)} \leftarrow u_{i\ell}^{(t)} \frac{\sum_j x_{ij} v_{\ell j}}{\sum_{\ell'} \sum_j u_{i\ell'}^{(t)} v_{\ell' j} v_{\ell j}} \ , \ \ell = 1, \ldots, k.$$

Pour chaque ligne $u = u_{i\cdot}$, $i = 1, \ldots, n$, la suite $(u^{(t)})$ ainsi générée fait décroître et converger la fonction $F(u) = F_2(u, V)$ ver son minimum. □

2.1.2 Méthodes du gradient de descente

La méthodologie du *gradient de descente* est une technique d'optimisation basée sur le *principe de Cauchy* présenté sa *Méthode générale pour la résolution des systèmes d'équations simultanées* (Cauchy, 1847 [57]). Les méthodes du *gradient de descente* (*GD*) constituent la seconde classe des algorithmes NMF classiques après celle des algorithmes MU de Lee et Seung (2001) [20]. Ces derniers ont même fait usage de cette technique simple même si les algorithmes MU sont principalement basés sur la considération de fonctions auxiliaires.

La méthodologie du gradient de descente fait elle même partie de la famille des *algorithmes à directions de descente*. Une telle famille regroupe des algorithmes d'optimisation différentiable destinés à minimiser une fonction réelle différentiable définie sur un espace euclidien (le plus souvent l'espace $\mathbb{R}^n$, l'espace des n-uplets de nombres réels, muni d'un produit scalaire), ou, plus généralement, sur un espace hilbertien. L'algorithme est itératif et procède donc par améliorations successives. Au point courant, un déplacement est effectué le long d'une direction de descente, de manière à faire décroître la fonction. Le déplacement le long de cette direction est déterminé par la technique numérique connue sous le nom de recherche linéaire.

2.1.2.1 Algorithmes à directions de descente.

Le cadre de la méthode est décrit comme suit. On cherche un point x^* qui minimise une fonction différentiable f définie sur un espace de hilbertien E et à valeurs réelles. On note par $\langle ., . \rangle$ le produit scalaire défini sur E, $\|.\|$ la norme associée, $f'(x)$ la dérivée et $\nabla f(x)$ le gradient de f en x si bien que pour tout $d \in E$, on a $f'(x) \, . \, d = \langle \nabla f(x), d \rangle$. Alors l'algorithme cherche un minimum de f en générant une suite de points $(x_k)_{k \geq 1}$, appelés itérés, qui approchent de mieux en mieux un minimum x^* du critère f. Cette suite est construite en se fondant sur deux constructions :
. calcul d'une direction de descente $d_k \in E$,
. détermination d'un pas α_k (recherche linéaire), qui est un nombre réel strictement positif, le long de la direction de descente de telle sorte que le nouvel itéré donne au critère une valeur inférieure à celle qu'il a en l'itéré courant ; le nouvel itéré est de la forme suivante $x_{k+1} := x_k + \alpha_k d_k$.
Les règles de recherche linéaire sont diverses. Elles permettent à chaque itération de déterminer la valeur du paramètre α_k. Celles-ci consistent, pour la plupart, à trouver la valeur qui minimise la fonction-coût $q(\alpha) = f(x_k + \alpha d_k)$. Considérant que d_k est une direction de descente, on obtient $q'(0) = \nabla f(x_k).d_k < 0$, ce qui permet de déterminer le comportement de q en fonction des valeurs de α. Il convient toutefois d'être prudent :
. en choisissant α trop grand, on ne parviendra pas à faire décroître les valeurs de q ou au pire on obtient un algorithme oscillant ;
. en choisissant α trop petit, l'algorithme aura une convergence lente.
Les objectifs majeurs visés par les règles de recherches linéaires sont au nombre de deux :

Le premier objectif est de faire décroître suffisamment la fonction f. Cela se traduit le plus souvent par la réalisation d'une inégalité de la forme $f(x_k + \alpha_k d_k) \leq f(x_k) + \eta_k$, où η_k est un terme négatif qui joue un rôle-clé dans la convergence de l'algorithme utilisant la recherche linéaire. L'argument est le suivant. Si $f(x_k)$ est minorée (c'est-à-dire il existe une constante C telle que $f(x_k) \geq C\,, \forall\, k$), alors le terme négatif η_k tend nécessairement vers 0. C'est souvent à partir de la convergence vers 0 de cette suite que l'on parvient à montrer que le gradient lui-même doit tendre vers 0. Le terme négatif devra prendre une forme bien particulière si l'on veut pouvoir en tirer de l'information.
Le second objectif de la recherche linéaire est d'empêcher le pas $\alpha_k > 0$ d'être trop petit, voire trop proche de 0, car ceci peut entraîner une « fausse convergence », c'est-à-dire la convergence des itérés vers un point non stationnaire.

Les règles de recherches linéaires du pas peuvent être classées en deux catégories : les règles dites *exactes* et les règles dites *inexactes*.

Recherches linéaires exactes. Les règles de recherches linéaires exactes développées, et que nous donnons ici sont : la *règle de Cauchy* et la *règle de Curry*.
Ainsi comme l'on cherche à minimiser la fonction f, il semble naturel de chercher à minimiser le critère le long de la direction d_k et donc de déterminer le pas α_k comme solution du problème de minimisation :

$$\alpha^* = \arg\min_{\alpha \geq 0} q(\alpha)$$

La détermination de α_k par ce problème, est appelée *règle de Cauchy* et le pas déterminé par cette règle est appelé pas de Cauchy ou pas optimal.
Dans certains cas, on préfèrera le plus petit point stationnaire de $q(\alpha)$ qui fait décroître cette fonction, c'est-à-dire $\alpha_k = \inf\{\alpha \geq 0 \,:\, q'(\alpha) = 0,\ q(\alpha) < q(0)\}$. On parle alors de *règle de Curry* et le pas déterminé par cette règle est appelé pas de Curry.

A côté de ces règles de recherches linéaires exactes qui s'efforcent d'identifier un minimum local de $q(\alpha)$ pour $\alpha > 0$, et qui conduisent souvent à des algorithmes longs, il a été proposé une autre classe de méthodes de recherches linéaires qui sont les règles de recherches linéaires inexactes qualifiées de plus économiques que les précédentes.

Recherches linéaires inexactes. Parmi les règles de recherches linéaires inexactes nous décrivons brièvement les plus connues.
Règle d'Armijo :
La règle d'Armijo se base sur le choix d'un paramètre $0 < m < 1$ et détermine une valeur approchée de α_k par la condition : $q(\alpha) \leq q(0) + m\alpha q'(0)$.
Règle de Goldstein :
Golstein a proposé en 1967 une méthode basée sur le choix cette fois-ci de deux paramètres $0 < m_1 < m_2 < 1$ et détermine les valeurs approchées de α_k par deux conditions : $q(\alpha) \leq q(0) + m_1\alpha q'(0)$ et $q(\alpha) \geq q(0) + m_2\alpha q'(0)$.
Règle de Wolfe :
Wolfe a proposé en 1969 une méthode sur le choix de deux paramètres $0 < m_1 < m_2 < 1$ et détermine les valeurs approchées de α_k par deux conditions : $q(\alpha) \leq q(0) + m_1\alpha q'(0)$ et $q(\alpha) \geq m_2 q'(0)$.

Les différents algorithmes à directions de descente portent en général le nom de leur direction de descente. A ce titre, l'algorithme de gradient est celui qui utilise la direction du gradient car le vecteur d_k est colinéaire au gradient, c'est-à-dire $d_k = -\nabla f(x_k)$.

2.1.2.2 Algorithmes du gradient de descente

Le schéma des algorithmes du gradient de descente s'inspire de celui des algorithmes à directions de descente. Alors, un itéré initial $x_0 \in E$ est d'abord choisi, puis un seuil de tolérance $\epsilon \geq 0$ est défini. Ensuite à chaque itération, k, la direction de descente $d_k = -\nabla f(x_k)$ est calculée ainsi

que le pas $\alpha_k > 0$ qui est déterminé par une règle de recherche linéaire sur f en x_k le long de la direction $-\nabla f(x_k)$, avant de calculer un nouvel itéré $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$. Plus en détails nous avons l'algorithme suivant :

Données : Fournir la fonction $f(x)$.
Résultat : Obtenir l'estimation d'un minimum local x^*.
. Initialisation : choisir le point initial x_0 avec l'itération initiale $k = 1$;
tant que ($\|\nabla f(x_{k-1})\| \geq \epsilon$) **faire**
  . Calcul du pas : α_{k-1};
  . Calcul du nouvel itéré : $x_k = x_{k-1} - \alpha_{k-1}\nabla f(x_{k-1})$;
  . Calcul du gradient : $\nabla f(x_k)$;
  . $k = k + 1$;
fin

Algorithme 2 : Algorithme standard du gradient de descente

Cet algorithme structurellement très simple repose sur le fait que, dans le voisinage d'un point x, la fonction f décroît le plus fortement dans la même direction mais en sens opposé à celui du gradient de f en x, à savoir $-\nabla f(x)$. En effet si $f'(x) \neq 0$, la direction $d = -\nabla f(x)$ est une direction de descente de f en x, puisque $f'(x) \,.\, d = -\|\nabla f(x)\|^2 < 0$, si bien que pour tout $\alpha > 0$ assez petit, on a $f(x - \alpha \nabla f(x)) < f(x)$. A chaque itération, puisque $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, on a bien $f(x_{k+1}) < f(x_k)$.

Cependant il convient de noter que la direction de descente dépend fortement du produit scalaire que l'on se donne sur l'espace hilbertien E. En effet, si $\sigma : (u, v) \mapsto \sigma(u, v)$ est un autre produit scalaire sur E, il existe un opérateur linéaire continu $S : E \to E$ auto-adjoint et défini positif tel que $\sigma(u, v) = \langle Su, v \rangle$, si bien que le gradient de f en x pour ce dernier produit scalaire s'écrit $S^{-1}\nabla f(x)$, ce qui montre explicitement la dépendance du gradient au produit scalaire. Il n'y a donc pas une unique direction de descente relative au gradient. On peut même voir que toute direction de descente de f en x, c'est-à-dire toute direction d telle que $f'(x) \,.\, d < 0$, est l'opposé du gradient de f en x pour un certain produit scalaire. L'efficacité de l'algorithme du gradient dépendra donc du choix de ce produit scalaire.

L'algorithme 2 peut être considéré comme le prototype ou la forme standard des algorithmes du gradient de descente, et pouvant être spécifié selon la règle de recherche linéaire de pas choisie. Comme exemples nous donnons les algorithmes du gradient de descente avec la règle d'Armijo et du gradient de descente avec approximation du premier ordre (Ho, 2008 [58]). L'algorithme avec recherche linéaire de règle d'Armijo a besoin de deux paramètres σ et β qui peuvent affecter sa convergence. Le critère d'Armijo dans les algorithmes du gradient est également abordé dans Bertsekas (1999) [59] et Lin (2007) [38].

Données : Fournir la fonction $f(x)$.
Résultat : Obtenir l'estimation d'un minimum local x^*.
. Initialisation : x_0 , σ , β , $\alpha_0 = 1$ et $k = 1$;
répéter
  . $\alpha_k = \alpha_{k-1}$;
  . $y = x_k - \alpha_k \nabla f(x_k)$;
  si $(f(y) - f(x_k) > \sigma\langle \nabla f(x_k), y - x_k \rangle)$ **alors**
    répéter
      . $\alpha_k = \alpha_k \,.\, \beta$;
      . $y = x_k - \alpha_k \nabla f(x_k)$;
    jusqu'à $(f(y) - f(x_k) \leq \sigma\langle \nabla f(x_k), y - x_k \rangle)$;
  fin
  sinon
    répéter
      . $lasty = y$;
      . $\alpha_k = \alpha_k / \beta$;
      . $y = x_k - \alpha_k \nabla f(x_k)$;
    jusqu'à $(f(y) - f(x_k) > \sigma\langle \nabla f(x_k), y - x_k \rangle)$;
    $y = lasty$
  fin
  . $x_{k+1} = y$;
  . $k = k + 1$;
jusqu'à *Critère d'arrêt atteint*;

Algorithme 3 : Gradient de descente avec règle d'Armijo

Le second exemple est celui de la méthode du gradient avec approximation du premier ordre. Celle-ci est basée sur le développement limité du premier ordre de la fonction-coût f. En effet si nous considérons le problème de minimisation suivant :

$$x^* = \arg\min_x f(x),$$

à chaque itération, k et pour chaque itéré x_k une approximation de $f(x)$ est donnée par $\bar{f}(x)$, avec

$$\bar{f}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x_k - x\|_2^2 \,,$$

où L est une constante de Lipschitz, telle que sa plus petite valeur satisfait $f(x) \leq \bar{f}(x)$, $\forall x$. Par le fait de cette inégalité, la solution x_{k+1} du problème suivant

$$x^* = \arg\min_x \bar{f}(x),$$

est aussi un point de décroissance de la fonction $f(x)$ puisque

$$f(x_{k+1}) \leq \bar{f}(x_{k+1}) \leq \bar{f}(x_k) = f(x_k).$$

Ce qui établit la descente de la fonction f à chaque itération avec $f(x_{k+1}) \leq f(x_k)$. S'agissant de la constante L puisqu'elle n'est pas connue a priori, elle est estimée à chaque itération en même temps que les itérés x_k. La méthode est décrite plus en détails dans l'algorithme suivant :

Données : Fournir la fonction $f(x)$.
Résultat : Obtenir une estimation d'un minimum local x^*.
. Initialisation : x_0 , L_0 et $k = 0$;
répéter
. $y = x_k - \frac{1}{L_k}\nabla f(x_k)$;
tant que $(f(y) - f(x_k) > \langle \nabla f(x_k), y - x_k \rangle + \frac{L_k}{2}\|y - x_k\|_2^2)$ **faire**
. $L_k = L_k/\beta$;
. $y = x_k - \frac{1}{L_k}\nabla f(x_k)$;
fin
. $x_{k+1} = y$;
. $L_{k+1} = L_k \,.\, \beta$;
. $k = k + 1$;
jusqu'à *Critère d'arrêt atteint*;

Algorithme 4 : Gradient de descente avec approximation du premier ordre.

2.1.2.3 Méthodes du gradient de descente pour la factorisation NMF

Tous les algorithmes du gradient évoqués ou décrits dans la section précédente sont adaptables à la fonction-coût carré de la distance Euclidienne :

$$F_2(U, V) = \|X - UV\|_F^2.$$

Ce qui permet d'obtenir la seconde classe des algorithmes NMF classiques que sont les algorithmes NMF du gradient de descente. Ainsi si on considère la fonction $F_2(U, V))$, le principe de ces algorithmes consiste d'abord à regarder celle-ci par rapport à chacune de ses variables U et V. Les fonctions résultantes sont alors $U \mapsto F_2(U, V)$, où la variable V est fixée et considérée comme une constante et $V \mapsto F_2(U, V)$, où la variable U est fixée et considérée comme une constante. Alors à chaque itération, k, l'itéré $(U^{(k)}, V^{(k)})$ est donné dans un premier temps par les équations de mise à jour suivantes

$$U^{(k)} = U^{(k-1)} - \alpha_k \nabla_U F_2(U^{(k-1)}, V^{(k-1)}) \,, \quad V^{(k)} = U^{(k-1)} - \alpha_k \nabla_V F_2(U^{(k-1)}, V^{(k-1)}) \quad (2.15)$$

avec $\nabla_x f(x, y)$ est le gradient de $f(x, y)$ par rapport à la variable x. Puis dans un second temps la contrainte de positivité de U et V est prise en charge par une opération de projection respectivement dans les orthants positifs $\mathbb{R}_+^{n\times k}$ et $\mathbb{R}_+^{k\times p}$. La projection consiste à ramener à 0 chacun des coefficients négatifs éventuels des matrices $U^{(k)}$ et $V^{(k)}$ (Chu et al, 2005 [60] ; Berry et al, 2007 [61] ; Lin, 2007 [38]). Ainsi si nous notons par $[\,.\,]_+$ l'opérateur de projection d'une matrice A, l'algorithme basic du gradient de descente est décrit comme suit :

Données : Fournir la matrice des données positives X et la tolérance $\epsilon > 0$.
Résultat : Obtenir les estimations de U et V.
. Initialisation : $k = 1$, $U^{(0)} \geq 0$, $V^{(0)} \geq 0$;
tant que $(\|\nabla_U F_2(U^{(k-1)}, V^{k-1})\| > \epsilon)$ *ou* $(\|\nabla_V F_2(U^{(k-1)}, V^{(k-1)})\| > \epsilon)$ **faire**
. Calcul des pas : α_{k-1}^U et α_{k-1}^V;
. Calcul de l'itéré : $(U^{(k)}, V^{(k)})$
. $U^k = [U^{(k-1)} - \alpha_{k-1}\nabla_U F_2(U^{(k-1)}, V^{k-1})]_+$;
. $V^{(k)} = [U^{(k-1)} - \alpha_{k-1}\nabla_V F_2(U^{(k-1)}, V^{(k-1)})]_+$;
. Calcul des gradients : $\nabla_U F_2(U^{(k)}, V^{(k)})$ et $\nabla_V F_2(U^{(k)}, V^{(k)})$;
. Incrémentation :
. $k = k + 1$;
fin

Algorithme 5 : Algorithme NMF du gradient de descente basique

Dans certains cas, à la place de la boucle **tant que ... faire** dans l'algorithme 5 peut être remplacée par une simple boucle **pour ... faire** (Berry et al, 2007 [61]). Certains algorithmes

initialisent les pas à la valeur 1, puis les multiplient par 1/2 à chaque itération (Hoyer, 2004 [62]). Il s'agit la d'une procédure simple mais pas idéale car elle impose des restrictions aux itérés $U^{(k)}$ et $V^{(k)}$. Par ailleurs, sans un choix judicieux pour les pas α_k^U et α_k^V, on peut difficilement parler de la convergence des méthodes du gradient de descente. De plus, la procédure de la projection dans les orthants positifs pour les matrices U et V rend la convergence encore plus difficile. Les algorithmes du gradient de descente qui utilise une simple règle géométrique pour la mise à jour du pas, comme le fait de le multiplier par une fraction, à chaque itération, produit souvent des factorisations qui ne sont pas efficientes. Dans ce cas les algorithmes sont très sensibles à l'initialisation des matrices U et V. Cependant Chu et al (2005) [60] ont proposé repris la méthode de Shepherd comme technique du gradient de descente qui peut accélérer sa convergence par les choix du pas, même si le support théorique sur la convergence n'a pas été bien fourni.

2.1.3 Algorithmes de moindres carrés alternés

Après la classe des algorithmes de mise à jour multiplicative MU de Lee et Seung, et les méthodes du gradient de descente GD, la dernière classe des algorithmes NMF classiques est celle des algorithmes de moindres carrés alternés. Cette dernière peut être divisée en deux sous-classes : les algorithmes ALS d'une part et les algorithme ANLS d'autre part. La particularité de ces deux sous-classes d'algorithmes est qu'elles utilisent toutes la technique du *processus alterné* qui est une procédure ancienne d'optimisation utilisée depuis des décennies et connue sous différents noms comme *variables alternées* (alternating variables), *recherche de coordonnées* (coordinate search) ou encore *méthode de variation locale* (method of local variation) (Nocedal et Wright, 1999 [63]). L'approche des moindres carrés alternés est aussi communément appelée méthode de descente par block de coordonnées (*block coordinate descent*) dans un problème d'optimisation avec contrainte de borne (*bound-constrained optimization*) comme on peut le voir dans Bertsekas (1999) [59] et Lin (2007) [38]. L'autre spécificité des algorithmes des moindres carrés alternés réside dans la considération du Problème 2 posé par Lee et Seung et qui consiste à la minimisation de la fonction-coût $F_2(U,V) = \|X - UV\|_F^2$ sous les contraintes $U \geq 0$ et $V \geq 0$, ou de manière équivalente à la considération de la fonction-coût $\frac{1}{2}\|X - UV\|_F^2$.

2.1.3.1 Les algorithmes ALS

La méthodologie ALS (*Alternating Least Squares*) désignant une procédure de moindres carrés alternés a été pour la première fois employée en factorisation matricielle par Paatero et Tapper (1994) [64]. Notons que même si la dénomination de la méthode ne fait pas référence à la positivité des matrices facteurs U et V, cette contrainte est bien présente dans le problème de minimisation que tente de résoudre la procédure ALS. Le problème dont se proposent de résoudre les algorithmes ALS est donc le suivant :

$$(U^*, V^*) = \arg \min_{U \geq 0, V \geq 0} \|X - UV\|_F^2 \tag{2.16}$$

La non-convexité du problème dans (2.16) par rapport au couple de matrices (U, V) et la difficulté posée par les méthodes analytiques ont encouragé les méthodes numériques qui sont des procédures itératives comme le cas des algorithmes ALS. Cependant la stratégie ALS exploite le fait que ce problème d'optimisation dans (2.16) ne soit convexe que par rapport à la variable U uniquement ou la variable V seule. Alors à chaque itération les matrices U et V sont mises à jour d'abord par des solutions analytiques de sous-problèmes de minimisation sans contraintes, puis la contrainte de positivité est prise en charge par une procédure qui met à 0 les éventuels coefficients négatifs de chacune de matrices (Berry et al, 2007 [61]). La méthode ALS subdivise donc la version sans contrainte du problème (2.16) en deux sous-problèmes :

$$U^* = \arg \min_U \|X - UV\|_F^2 \quad \text{et} \quad V^* = \arg \min_V \|X - UV\|_F^2. \tag{2.17}$$

Le processus alterné de la méthode se caractérise par le fait que les deux sous-problèmes dans (2.17) sont résolus en alternance par l'approche des moindres carrés où la variable endogène n'est

plus un vecteur colonne mais une matrice (la matrice X), et cela dans une procédure itérative où l'une des matrices U et V est fixée par rapport à l'autre. L'algorithme type élémentaire ALS est donc de la forme suivante :

Données : Fournir la matrices des données X.
Résultat : Obtenir les estimations des matrices U et V.
Initialisation de la matrice U;
pour *i allant de* 1 *à Maxiter* **faire**
  Trouver V par la méthode des moindres carrés via l'équation matricielle $U^TUV = U^TX \quad (1)$;
  Mettre tous les coefficients négatifs de V à 0;
  Trouver U par la méthode des moindres carrés via l'équation matricielle $VV^TU^T = VX^T \quad (2)$;
  Mettre tous les coefficients négatifs de U à 0;
fin

Algorithme 6 : ALS classique basic

Les équations (1) et (2) de l'algorithme ALS basic dérivent respectivement des *équation normales* de la méthode des moindres carrés ordinaires appliquée aux deux problèmes de minimisations dans (2.17). Alors les estimations, à chaque itération, de U et V sont respectivement données par $\hat{U}^T = (VV^T)^{-1}VX^T$ et $\hat{V} = (U^TU)^{-1}U^TX$, puis les éventuels coefficients négatifs sont ramenés à la valeur 0. Cette technique simple présente également quelques avantages supplémentaires : elle favorise la parcimonie ; de plus, elle permet aux itérées d'apporter une flexibilité supplémentaire non disponible dans d'autres algorithmes, notamment la classe des algorithmes MU. Un inconvénient de ces derniers est qu'une fois qu'un coefficient dans U ou V devient égal à 0, il y reste coincé. Ce verrouillage à 0 des coefficients est restrictif, ce qui signifie qu'une fois que l'algorithme commence à se diriger vers un point fixe, même s'il est un point fixe médiocre, il doit continuer dans cette veine. Les algorithmes ALS eux sont plus flexibles, permettant au processus itératif d'échapper à ce verrouillage.

Convergence La convergence des algorithmes ALS est discutée dans ce paragraphe. Comme pour les méthodes de variables alternées en général où la convergence n'est pas prouvée, toutes les études sur la méthodologie ALS n'ont pas aussi démontré sa convergence (Paatero, 1994 et 1999 [65, 65] ; Langville et al, 2014 [66]). Il a même été démontré que les algorithme ALS converge plutôt vers un point fixe, mais que ce point fixe est soit un extremum local ou un point de selle (Finesso et Spreij, 2004 [67] ; Gonzalez et Zhang, 2005 [68] ; Lin, 2007 [38]). L'une des explications de cette non convergence vers un minimum local (forcément) est la procédure non analytique de prise en charge des contraintes de positivité des matrice U et V. Cette procédure ad-hoc de prise en compte de la contrainte de positivité, bien qu'elle accélère considérablement la vitesse l'algorithme (et améliore la parcimonie), ne favorise cependant pas la convergence vers un minimum local.

2.1.3.2 Les algorithmes ANLS

La seconde sous-classe des algorithmes de moindres carrés alternés est celle des algorithmes ANLS (*Alternating Nonnegative Least Squares*) qui désignent les algorithmes de moindres carrés positifs et alternés. A l'image de la précédente sous-classe, elle utilise le procédé itératif d'alternance de la résolution de deux problèmes de minimisation. Le problème de minimisation d'origine est

$$(U^*, V^*) = \arg \min_{U \geq 0, V \geq 0} \frac{1}{2}\|X - UV\|_F^2. \tag{2.18}$$

Les deux problèmes de minimisation que résolvent en alternance les algorithmes ANLS sont en effet des sous-problèmes de (2.18) qui sont donnés par :

$$U^* = \arg \min_{U \geq 0} \frac{1}{2}\|X - UV\|_F^2 \quad \text{et} \quad V^* = \arg \min_{V \geq 0} \frac{1}{2}\|X - UV\|_F^2 \tag{2.19}$$

La méthodologie ANLS décompose alors chacun des deux sous-problèmes de minimisation dans (2.19) en plusieurs sous-problèmes également. La formulation de ces dernier est alors de la forme

$$x^* = \arg\min_{x \geq 0} \frac{1}{2}\|y - Ax\|_2^2. \tag{2.20}$$

où y et x sont des vecteurs colonnes de tailles respectives $n \times 1$ et $p \times 1$, et A une matrice de dimension $n \times p$. Le problème (2.20) est nommé NNLS (*Non-Negative Least Squares*), désignant un problème de minimisation par des moindres carrés positives. Le problème NNLS est un problème de moindres carrés classique avec contraintes de positivité qui a été posé et résolu pour la première fois dans l'ouvrage de Lawson et Hanson (1974) [69]. Plusieurs méthodes de résolution de (2.20), et qui prennent correctement en charge la contrainte de positivité sur la base de considérations théoriques, ont été proposées par divers auteurs. Ces fondements théories des méthodes de résolution du problème NNLS en font la différence majeure avec la classe des algorithmes ALS qui sont des procédures ad-hoc qui très souvent gagnent plus en vitesse mais moins en précision. L'ensemble des méthodes résolvant le problème (2.20) peut être appelé classe des algorithmes NNLS (Langville et al, 2014 [66]). Bien qu'étant des procédures à vitesse très lente, plusieurs études ont montré que les algorithmes NNLS convergent vers un minimum local (Dhillon et Sra, 2005 [70] ; Lin, 2007 [38]). Plusieurs algorithmes NNLS ont alors été développés. Les précurseurs Lawson et Hanson ont proposé l'algorithme classique de résolution du problème NNLS, qu'ils ont nommé *Active Set Method* (Algorithme 7). Il s'agit d'une procédure d'optimisation qui recherche dans l'ensemble des points admissibles, des variables actives et passives de façon optimal en échangeant une variable entre deux sous-ensembles de travail. Notons que si des variables passives (c'est-à-dire strictement positives) de la solution sont connues à l'avance, alors le problème NNLS peut simplement être résolu par une procédure de moindres carrés sans contraintes sur ces variables passives. Cependant cet algorithme est très lent en pratique à cause du calcul de l'inverse $((A^P)^T A^P)^{-1}$. C'est pourquoi beaucoup d'autres méthodes ont été proposées afin d'améliorer le temps de convergence. Bierlaire et al, 1991 [71] ont adopté la méthode du *Gradient Projeté*. Bro et de Jong (1997) [72] puis van Benthem et Keenan (2004) [73] ont proposé des versions rapides de l'algorithme standard *Active Set*, avec les algorithmes FNNLS (*Fast Non-Negativity constrained linear Least Squares*). Chu et al (2005) [60] ont suggéré la méthode de *Newton Projeté*. Merrit et Zhang (2005) [74] ont présenté la méthode du *Gradient à Point-Intérieur*. Kim et al (2006) [75] ont développé une nouvelle méthode de *Quasi-Newton Projeté*.

Données : Matrice de valeurs réelles A de taille $n \times p$,
Vecteur colonne de valeurs réelles y de taille $n \times 1$,
Valeur réelle ϵ, désignant la tolérance pour le critère d'arrêt.
Résultat : Estimation du vecteur x.
Initialisation :
. Considérer l'ensemble $P = \emptyset$;
. Considérer l'ensemble $R = \{1, \ldots, p\}$;
. Considérer un vecteur colonne x à coefficients tous nuls de taille $p \times 1$;
. Calculer le vecteur $w = A^T(y - Ax)$;
tant que $((R \neq \emptyset)$ *et* $(\max(w) > \epsilon))$ **faire**
1. Considérer j dans R comme l'index du $\max(w)$ dans w ;
2. Ajouter j à P;
3. Retirer j de R;
4. Considérer A^P une sous-matrice de A et dont les colonnes sont indexées par l'ensemble P;
5. Soit s un vecteur de même taille que x, dont s^P un sous-vecteur indexé par P et s^R un sous-vecteur indexé par R;
6. Calculer $s^P = ((A^P)^T A^P)^{-1}(A^P)^T y$;
7. Mettre tous les coefficients de s^R à 0;
8. **tant que** $(\min(s^P) \leq 0)$ **faire**
. Calculer $\alpha = \min(\frac{x_i}{x_i - s_i})$ pour i dans P avec $s_i \leq 0$;
. Affecter à x le vecteur $x + \alpha(s - x)$;
. Retirer de R tous les indices j de P tels que $x_j = 0$;
. Calculer $s^P = ((A^P)^T A^P)^{-1}(A^P)^T y$;
. Mettre tous les coefficients de s^R à 0;
fin
9. Affecter à x le vecteur s;
10. Affecter à w le vecteur $A^T(y - Ax)$;
fin

Algorithme 7 : Algorithme Active set

Revenons à l'équation (2.19) pour considérer le second sous-problème de minimisation :

$$V^* = \arg\min_{V \geq 0} \frac{1}{2}\|X - UV\|_F^2 ,$$

celui-ci peut être divisé, si l'on fixe la matrice U et considère les colonnes de V, en p problèmes NNLS de la forme (2.20) :

$$v_*^1 = \arg\min_{v^1 \geq 0} \frac{1}{2}\|x^1 - Uv^1\|_2^2 , \quad \ldots , \quad v_*^p = \arg\min_{v^p \geq 0} \frac{1}{2}\|x^p - Uv^p\|_2^2 \tag{2.21}$$

où v_*^j, et v^j $j = 1, \ldots, p$, sont respectivement le j ième vecteur colonne des matrices V^* et V. S'agissant du premier sous-problème de minimisation dans (2.19)

$$U^* = \arg\min_{U \geq 0} \frac{1}{2}\|X - UV\|_F^2$$

un opérateur de transposition de matrice, $T[.]$ où $T[A] = A^T$ avec A une matrice, est nécessaire pour retrouver la forme du problème (2.20). Ce sous-problème sera alors équivalent à ce problème qui suit

$$U_*^T = \arg\min_{U^T \geq 0} \frac{1}{2}\|X^T - V^T U^T\|_F^2 \tag{2.22}$$

où la transposition fait basculer l'inconnue U ou encore U^T du côté droit du produit matriciel. Ce qui donne de la manière que dans (2.21) les n sous-problèmes NNLS de (2.22) :

$$u_*^{T\,(1)} = \arg\min_{u^{T\,(1)} \geq 0} \frac{1}{2}\|x^{T\,(1)} - V^T u^{T\,(1)}\|_2^2 , \quad \ldots , \quad u_*^{T\,(n)} = \arg\min_{u^{T\,(n)} \geq 0} \frac{1}{2}\|x^{T\,(n)} - V^T u^{T\,(n)}\|_2^2 \tag{2.23}$$

où $u_*^{T\,(i)}$, $u^{T\,(i)}$ et $x^{T\,(i)}$, $i = 1, \ldots, n$, sont respectivement le i ième vecteur colonne des matrices U_*^T, U^T et X^T.

Données : Fournir la matrices des données X.
Résultat : Obtenir les estimations des matrices U et V.
Initialisation de la matrice U avec des coefficients positifs;
tant que *le critère d'arrêt n'est atteint* **faire**
  1. Fixer la matrice U courante et trouver V;
  pour j *allant de* 1 *à* p **faire**
    Résoudre le j ième problème NNLS $v_*^j = \arg\min_{v^j \geq 0} \frac{1}{2}\|x^j - Uv^j\|_2^2$;
  fin
  2. Fixer la matrice V courante et trouver U;
  pour i *allant de* 1 *à* n **faire**
    Résoudre le i ième problème NNLS $u_*^{T\,(i)} = \arg\min_{u^{T\,(i)} \geq 0} \frac{1}{2}\|x^{T\,(i)} - V^T u^{T\,(i)}\|_2^2$;
  fin
fin

Algorithme 8 : ANLS classique

2.2 Quelques variantes : Algorithmes de factorisation pondérée

A la suite des algorithmes NMF classiques, d'autres types d'algorithmes ont été développés. Il s'agit de variantes des algorithmes classiques, et dont la différence avec celles-ci réside sur la modification des fonctions-coûts classiques que sont la divergence de Kullback-Leibler généralisée $F_1(U,V)$ et le carré de la distance euclidienne $F_2(U,V)$ (Lee et Seung, 2001 [20]) ou encore $\frac{1}{2}F_2(U,V)$ (dans beaucoup d'autres études NMF). Notons que les deux dernières fonctions-coûts $F_2(U,V)$ et $\frac{1}{2}F_2(U,V)$ sont équivalentes pour les problèmes de minimisation. Les motivations ayant conduit au développement de ces variantes résident dans les préoccupations additionnelles que cherche à prendre en charge la factorisation NMF. Cela se traduit par la considération d'une nouvelle fonction-coût à partir d'une classique : $F_1(U,V)$ ou $F_2(U,V)$ ou encore $\frac{1}{2}F_2(U,V)$. Cette nouvelle fonction-coût porte alors en elle les contraintes additionnelles, en plus de la positivité, que l'on cherche à imposer sur les facteurs U ou/et V de l'approximation $X \approx UV$. C'est ainsi que nous avons les contraintes de symétrie, de parcimonie, de lissage ... A côté de ces types de variantes une autre classe de factorisation NMF a été développée : il s'agit de la factorisation NMF pondérée.

La factorisation NMF pondérée, encore connue sous l'acronyme WNMF (*Weighted Nonnegative Matrix Factorization*) est une méthodologie dont les motivations ou encore les préoccupations qu'elle cherche à résoudre sont diverses même si la problématique des données manquantes en occupe la plus grande partie. La formalisation du problème WNMF consiste alors à incorporer une matrice de *poids* W, à coefficients positifs, de même taille que la matrice des données X, dans l'une des fonctions-coûts classiques $F_1(U,V) = D_{KL}(X\|UV)$ (Guillamet et al, 2001 [76] ; Blondel et al, 2008 [77]) ou $F_2(U,V) = \|X - UV\|_F^2$ (Paatero, 1997 [65] ; Mao et Saul, 2004 [21] ; Zhang et al, 2006 [24]) ou $\frac{1}{2}F_2(U,V) = \frac{1}{2}\|X - UV\|_F^2$ (Blondel et al, 2008 [77] ; Kim et Choi, 2009 [22]). Les fonctions-coûts qui en résultent sont alors la *divergence de Kullback-Leibler généralisée pondérée* $D_{KL,W}(X\|UV)$, définie par

$$D_{KL,W}(X\|UV) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(w_{ij}\left(x_{ij}\log\frac{x_{ij}}{\sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}} - x_{ij} + \sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right)\right) \tag{2.24}$$

et le *carré de la distance euclidienne pondérée* $\|X - UV\|_{F,W}^2$.

$$\|X - UV\|_{F,W}^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(w_{ij}\left(x_{ij} - \sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right)^2\right), \tag{2.25}$$

ou encore sa fonction-coût équivalente

$$\frac{1}{2}\|X-UV\|^2_{F,W}=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(w_{ij}\left(x_{ij}-\sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right)^2\right), \tag{2.26}$$

où $W=(w_{ij})$ est la matrice de poids (Blondel et al, 2008 [77]). Notons que les fonctions-coûts classiques peuvent être regardées comme des cas particuliers de ces fonctions fonctions-coûts pondérées avec tous les coefficients w_{ij} de la matrice W étant égaux à 1, c'est-à-dire $w_{ij}=1\,,\forall\ i,\ j$. Dans ce cas la factorisation NMF classique est un cas particulier de la factorisation WNMF, où $W=\mathbf{1}_{n\times p}$ est une matrice où tous les coefficients sont égaux à 1.

2.2.1 Les algorithmes MU avec la factorisation WNMF

Dans cette section nous présentons les extensions des algorithmes MU de Lee et Seung (2001) [20], faites par Blondel et al (2008) [77] dans le cas de la factorisation WNMF. De tous les auteurs qui ont étudié la factorisation NMF pondérée ces auteurs sont ceux qui donné une étude exhaustive des extensions des règles MU de Lee et Seung. Les extensions sont faites pour les deux fonctions-coûts classiques de Lee et Seung que sont la divergence de Kullback-Leibler généralisée et le carré de la distance euclidienne pour donner des fonctions-coûts pondérées sur lesquels Blondel et al (2008) ont établi de nouvelles règles de mises à jour multiplicatives. Notons cependant qu'ils ne sont pas les premiers à donner ou encore à établir ces règles multiplicatives. En effet, avant eux Mao et Saul (2004) [21], puis Zhang et al (2006) [24] ont donné les règles MU dans le cas pondéré pour ce qui concerne la fonction-coût carré de la distance euclidienne.

2.2.1.1 Cas de la divergence de Kullback-Leibler généralisée pondérée

Les auteurs ont commencé par un rappel des idées basiques de la méthodologie de Lee et Seung, 2001 [20] qui, notons-le ont utilisé l'approche du gradient de descente. En effet la minimisation de la fonction-coût $F_1(U,V)$ sous les contraintes $U\geq 0\,,\ V\geq 0$ requiert la construction de gradients

$$\nabla_U F_1(U,V)=-\left(\frac{X}{UV}-\mathbf{1}_{n\times p}\right)V^T\,,\quad \nabla_V F_1(U,V)=-U^T\left(\frac{X}{UV}-\mathbf{1}_{n\times p}\right) \tag{2.27}$$

où $\mathbf{1}_{n\times p}$ est une matrice de taille $n\times p$ et dont tous les coefficients sont égaux à 1. Les conditions de Kuhn-Tucker sont données par

$$U\geq 0\,,\qquad V\geq 0\,, \tag{2.28}$$

$$\nabla_U F_1(U,V)\geq 0\,,\qquad \nabla_V F_1(U,V)\geq 0\,, \tag{2.29}$$

$$U\circ\nabla_U F_1(U,V)=0\,,\quad V\circ\nabla_V F_1(U,V)=0. \tag{2.30}$$

où $A\circ B$ est le produit matriciel de Hadamard entre A et B. Le théorème qui suit généralise alors le théorème 2.1 dans le cas pondéré. Notons que Blondel et al (2008) ont écrit les mises à jour pondérées MU en écriture matricielle, cependant pour nous nous conformons à la notation des règles MU dans les théorèmes 2.1 et 2.2, pour faciliter les comparaisons.

Théorème 2.3. *La divergence de Kullback-Leibler généralisée pondérée $D_{KL,W}(X\|UV)$ est décroissante sous les règles de mise à jour suivantes :*

$$u_{i\ell}\leftarrow u_{i\ell}\frac{\sum_j\left(w_{ij}x_{ij}/\sum_{\ell'}u_{i\ell'}v_{\ell' j}\right)v_{\ell j}}{\sum_j w_{ij}v_{\ell j}}\,,\quad v_{\ell j}\leftarrow v_{\ell j}\frac{\sum_i u_{i\ell}\left(w_{ij}x_{ij}/\sum_{\ell'}u_{i\ell'}v_{\ell' j}\right)}{\sum_i u_{i\ell}w_{ij}} \tag{2.31}$$

La divergence pondérée $D_{KL,W}(X\|UV)$ est invariante sous ces règles de mise à jour si et seulement si les conditions (2.28) *et* (2.30) *sont vérifiées.*

Preuve. Comme dans Lee et Seung, 2001 [20], le théorème est prouvé seulement pour V, le cas de la matrice U pouvant s'en déduire facilement par analogie. Les auteurs ont alors considéré les divergences partielles par rapport aux colonnes correspondantes des matrices V, W et X notées par v, w et x :

$$\begin{aligned} F(v) &= D_{KL,w}(x\|Uv) && (2.32)\\ &= \sum_i w_i \left(x_i \log x_i - x_i + \sum_\ell u_{i\ell} v_\ell - x_i \log \sum_\ell u_{i\ell} v_\ell \right) && (2.33) \end{aligned}$$

Cette divergence partielle a pour fonction auxiliaire :

$$\begin{aligned} G(v, v^{(t)}) &= \sum_i \left(w_i \left(x_i \log x_i - x_i + \sum_\ell u_{i\ell} v_\ell \right.\right.\\ &- \left.\left. x_i \sum_\ell \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} \left(\log u_{i\ell} v_\ell - \log \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} \right) \right) \right) && (2.34) \end{aligned}$$

A cause de la convexité de la fonction $-\log(x)$ et puisque $\sum_\ell \frac{u_{i\ell} v_\ell^{(t)}}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} = 1$, l'inégalité $G(v, v^{(t)}) \geq F(v)\,,\ \forall\, v$ est obtenue. En plus $G(v^{(t)}, v^{(t)}) = F(v^{(t)})$, alors les relations suivantes entre la fonction-coût $F(v)$ sa fonction auxiliaire associée sont obtenues

$$F(v^{(t)}) = G(v^{(t)}, v^{(t)}) \geq \min_v G(v, v^{(t)}) = G(v^{(t+1)}, v^{(t)}) \geq F(v^{(t+1)}). \quad (2.35)$$

Ainsi pour établir les règles pondérées MU, les auteurs ont montré qu'il suffit, dans un premier temps, d'établir les équations qui annulent le gradient en v de $G(v, v^{(t)})$:

$$\frac{\partial G(v, v^{(t)})}{\partial v_\ell} = \sum_i w_i u_{i\ell} - \frac{v_\ell^{(t)}}{v_\ell} \sum_i w_i x_i \frac{u_{i\ell}}{\sum_{\ell'} u_{i\ell'} v_{\ell'}^{(t)}} = 0\,,\ \ell = 1, \ldots, k. \quad (2.36)$$

Puis dans un second temps le minimum de $G(v, v^{(t)})$ est obtenu à partir de (2.36) et donné par :

$$v^{(t+1)} = \frac{v^{(t)}}{U^T w} \circ \left(U^T \frac{[x \circ w]}{U v^{(t)}} \right). \quad (2.37)$$

Si on spécifie les vecteurs colonnes quelconques v, x et w, c'est-à-dire on pose $v = v_{.j}$, $x = x_{.j}$ et $w = w_{.j}$, $j = 1, \ldots, p$ on obtient l'équation de mise à jour de la matrice V dans (2.31)

$$v_{\ell j}^{(t+1)} \leftarrow v_{\ell j}^{(t)} \frac{\sum_i u_{i\ell} \left(w_{ij} x_{ij} / \sum_{\ell'} u_{i\ell'} v_{\ell' j}^{(t)} \right)}{\sum_i u_{i\ell} w_{ij}}.$$

La relation (2.35) montre que la divergence pondérée est décroissante sous les règles de mise à jour de V. En utilisant (2.37) et en considérant la relation

$$\nabla F(v^{(t)}) = U^T w - U^T \frac{[x \circ w]}{U v^{(t)}} \quad (2.38)$$

on obtient simplement l'égalité $v^{(t+1)} = v^{(t)}$ si et seulement si $v^{(t)} \circ \nabla F(v^{(t)}) = 0$; et finalement la contrainte de positivité du vecteur colonne $v^{(t)}$ est automatiquement satisfaite. □

2.2.1.2 Cas du carré de la distance euclidienne pondérée

Comme dans le cas précédent, les gradients associés à la fonction-coût classique $\frac{1}{2} F_2(U, V)$ sont d'abord donnés :

$$\nabla_U \frac{1}{2} F_2(U, V) = -(X - UV) V^T\,, \quad \nabla_V \frac{1}{2} F_2(U, V) = -U^T (X - UV). \quad (2.39)$$

Puis en remplaçant $F_1(U,V)$, par $\frac{1}{2}F_2(U,V)$ dans les équations (2.29) et (2.30) on obtient les conditions de Kuhn-Tucker pour la fonction-coût $\frac{1}{2}F_2(U,V)$:

$$U \geq 0\,, \qquad V \geq 0\,, \tag{2.40}$$

$$\nabla_U \frac{1}{2}F_2(U,V) \geq 0\,, \quad \nabla_V \tfrac{1}{2}F_2(U,V) \geq 0\,, \tag{2.41}$$

$$U \circ \nabla_U \frac{1}{2}F_2(U,V) = 0\,, \quad V \circ \nabla_V \tfrac{1}{2}F_2(U,V) = 0. \tag{2.42}$$

Comme préalable à la généralisation du théorème 2.2 (Théorème 1 dans Lee et Seung, 2001), les auteurs Blondel et al (2008) ont d'abord établit un résultat sous forme de lemme.

Lemme 2.4. *Soit A une matrice positive symétrique et v un vecteur colonne positif, alors la matrice $\hat{A} = diag\left(\frac{Av}{v}\right) - A$ est semi-définie positive.*

La preuve du lemme est donnée dans Blondel et al (2008) [77]. L'extension du théorème 2.2 est alors donnée par le théorème suivant :

Théorème 2.4. *La fonction-coût carré de la distance euclidienne pondérée $\frac{1}{2}\|X - UV\|^2_{F,W}$ est décroissante sous les règles de mise à jour suivantes :*

$$u_{i\ell} \leftarrow u_{i\ell}\frac{\sum_j x_{ij}w_{ij}v_{\ell j}}{\sum_j w_{ij}\sum_{\ell'} u_{i\ell'}v_{\ell' j}v_{\ell j}}\,, \quad v_{\ell j} \leftarrow v_{\ell j}\frac{\sum_i u_{i\ell}x_{ij}w_{ij}}{\sum_i u_{i\ell}w_{ij}\sum_{\ell'} u_{i\ell'}v_{\ell' j}} \tag{2.43}$$

La fonction-coût pondérée $\frac{1}{2}\|X - UV\|^2_{F,W}$ est invariante si et seulement si les conditions (2.40) *et* (2.42) *sont vérifiées.*

Preuve. De même que dans le cas précédent Blondel et al (2008) ont seulement montré comment établir les règles pondérées MU pour la matrice V, ceux de la matrice U pouvant après s'en déduire simplement. Alors la fonction-coût pondérée $\frac{1}{2}\|X-UV\|^2_{F,W}$ est subdivisée en p fonctions-coûts conduisant ainsi à p sous-problèmes minimisation indépendants où les variables sont les p vecteurs colonnes, v, de V. Les autres vecteurs colonnes associés sont x et w, respectivement des matrices X et W. La fonction-coût partielle pour une colonne v est :

$$\begin{aligned} F(v) &= \frac{1}{2}\|x - Uv\|^2_{F,W} && (2.44)\\ &= \frac{1}{2}\sum_i \left(w_i\left(x_i - [Uv]_i\right)^2\right) && (2.45)\\ &= \frac{1}{2}\left(x - Uv\right)^T D_w (x - Uv) && (2.46) \end{aligned}$$

où $D_w = diag(w)$. Soit $v^{(t)}$ est l'approximation courante du minimiseur de $F(v)$, alors l'on peut réécrire $F(v)$ sous la forme quadratique suivante :

$$F(v) = F(v^{(t)}) + (v - v^{(t)})^T \nabla_v F(v^{(t)}) + \frac{1}{2}(v - v^{(t)})^T U^T D_w U (v - v^{(t)}) \tag{2.47}$$

où $\nabla_v F(v^{(t)})$ est explicitement donné par

$$\nabla_v F(v^{(t)}) = -U^T D_w (x - Uv^{(t)}). \tag{2.48}$$

La fonction-coût partielle $F(v)$ a alors pour fonction auxiliaire

$$G(v, v^{(t)}) = F(v^{(t)}) + (v - v^{(t)})^T \nabla_v F(v^{(t)}) + \frac{1}{2}(v - v^{(t)})^T D(v^{(t)})(v - v^{(t)}) \tag{2.49}$$

où $G(v^{(t)}, v^{(t)}) = F(v^{(t)})$ et $D(v^{(t)})$ est une matrice diagonale choisie telle que $D(v^{(t)}) - U^T D_w U$ est semi-définie positive impliquant ainsi que $G(v, v^{(t)}) - F(v) \geq 0\,,\ \forall\, v$. Le choix de $D(v^{(t)})$ est similaire à celui proposé par Lee et Seung :

$$D(v^{(t)}) = diag\left(\frac{U^T D_w U v^{(t)}}{v^{(t)}}\right) \tag{2.50}$$

Le lemme 2.4 assure le fait que la matrice $D(v^{(t)}) - U^T D_w U$ est semi-définie positive. Il en résulte que :

$$F(v^{(t)}) = G(v^{(t)}, v^{(t)}) \geq \min_v G(v, v^{(t)}) = G(v^{(t+1)}, v^{(t)}) \geq F(v^{(t+1)}) \tag{2.51}$$

où $v^{(t+1)}$ est trouvé en résolvant l'équation $\frac{\partial G(v, v^{(t)})}{\partial v} = 0$, donnant ainsi :

$$\begin{aligned}
v^{(t+1)} &= v^{(t)} - D(v^{(t)})^{-1} \nabla F(v^{(t)}) && (2.52)\\
&= v^{(t)} + diag\left(\frac{v^{(t)}}{U^T D_w U v^{(t)}}\right) U^T D_w (x - U v^{(t)}) && (2.53)\\
&= v^{(t)} + v^{(t)} \circ \frac{U^T D_w (x - U v^{(t)})}{U^T D_w U v^{(t)}} && (2.54)\\
&= v^{(t)} \circ \frac{U^T D_w x}{U^T D_w U v^{(t)}} && (2.55)\\
&= v^{(t)} \circ \frac{U^T (w \circ x)}{U^T (w \circ (U v^{(t)}))} && (2.56)
\end{aligned}$$

L'équation (2.56) donne à l'échelle des coefficients de chaque vecteur $v = v_{.j}, j = 1, \ldots, p$, les règles MU pour V dans (2.43). La relation (2.51) montre que la fonction-coût pondérée $\frac{1}{2} F_{2,W}(U, V) = \frac{1}{2} \|X - UV\|^2_{F,W}$ est décroissante par rapport aux règles pondérées MU pour V, et l'équation (2.52) montre que $v^{(t+1)} = v^{(t)}$ si et seulement si $v^{(t)} \circ \nabla F(v^{(t)}) = 0$. Finalement, la positivité de $v^{(t)}$ est automatiquement satisfaite. □

2.2.1.3 Lien entre les deux fonctions-coûts

Les auteurs Blondel et al (2008) ont souligné une relation existante entre les deux fonctions-coûts pondérées. Alors pour le cas de la fonction-coût $D_{KL,W}(X \| UV)$, considérant les règles MU pour V, en écriture matricielle, qu'ils ont réécrit comme suit :

$$\begin{aligned}
V \leftarrow \frac{V}{U^T W} \circ \left(U^T \frac{[W \circ X]}{UV}\right) &= V \circ \left(\frac{U^T \frac{[W \circ X]}{UV}}{U^T \frac{[W \circ (UV)]}{UV}}\right)\\
&= V \circ \left(\frac{U^T \left[\frac{W}{UV} \circ X\right]}{U^T \left[\frac{W}{UV} \circ (UV)\right]}\right), && (2.57)
\end{aligned}$$

l'on peut remarquer, comme le disent les auteurs, que l'équation (2.57) montre que les règles MU, qui actualisent à chaque itération la matrice V, pour le cas de la divergence de Kullback-Leibler généralisée pondérée sont équivalentes à celles qui mettent à jour V dans le cas d'une fonction-coût carré de la distance euclidienne pondérée, où sa matrice de poids est $W_{UV} = \frac{W}{UV}$. Toutefois pour cette nouvelle fonction-coût pondérée, la matrice poids dépendant des matrices U et V, est actualisée à chaque itération du fait de la mise à jour de celles-ci. Le même principe permet d'établir les règles MU, pour la matrice U, qui mettent en relation la divergence de Kullback-Leibler généralisée pondérée de matrice de poids W et la fonction-coût carré de la distance euclidienne pondérée de matrice de poids W_{UV}.

Inversément, l'on peut voir que les règles pondérées MU pour les matrices U et V, dans le cas de la fonction-coût carré de la distance euclidienne pondérée de matrice de poids W sont équivalentes aux règles pondérées MU dans le cas d'une divergence de Kullback-Leibler généralisée pondérée de matrice de poids $W_{UV} = W \circ (UV)$.

Le tableau 2.1 résume les règles MU pour la factorisation classique NMF et la factorisation pondérée WNMF concernant les deux types de fonctions-coûts. L'écriture matricielle des règles MU permet aisément de comparer les deux types de factorisations matricielles pour chacune des fonctions-coûts correspondantes. On peut donc voir que la seule différence réside dans la spécification de la matrice de poids qui est $\mathbf{1}_{n \times p}$ pour la factorisation NMF et W_1 ou W_2 quelconque pour la factorisation WNMF.

	Divergence KL généralisée (DKLG)	Distance Euclidienne (DE)
NMF	$V \leftarrow \frac{V}{U^T \mathbf{1}_{n\times p}} \circ \left(U^T \frac{[\mathbf{1}_{n\times p} \circ X]}{UV} \right)$	$V \leftarrow V \circ \frac{U^T(\mathbf{1}_{n\times p} \circ X)}{U^T(\mathbf{1}_{n\times p} \circ (UV))}$
WNMF	$V \leftarrow \frac{V}{U^T W_1} \circ \left(U^T \frac{[W_1 \circ X]}{UV} \right)$	$V \leftarrow V \circ \frac{U^T(W_2 \circ X)}{U^T(W_2 \circ (UV))}$
DKLG⇔ED	$W_2 = \frac{W_1}{UV}$	

TABLE 2.1: Tableau récapitulatif des règles MU pour NMF et WNMF.

2.2.2 Les algorithmes ANLS sous la factorisation WNMF

La classe des algorithmes des moindres carrés alternés est aussi adaptable à la factorisation pondérée WNMF. Nous nous intéressons ici à la sous-classe des algorithmes ANLS développés par rapport à la factorisation WNMF, particulièrement l'algorithme ANLS-WNMF proposé par les auteurs Kim et Choi (2008) [22]. En effet les motivations avancées par les auteurs étaient d'apporter une alternative à des algorithmes WNMF existants (Mao et Saul, 2004 [21], Zhang et al, 2006 [24]) jugés lents en temps de convergence mais de précision pouvant être améliorée, malgré leur facilité d'implémentation.

Formalisation. Dans Kim et Choi (2009) [22], le problème de factorisation est formulé par $X \approx UV^T$, c'est-à-dire en considérant la transposée de la matrice V en lieu et place de V comme nous le faisons dans cette thèse. C'est pourquoi pour nous conformer à nos notations nous adaptons ici leurs formulations.

La fonction-coût considérée par les auteurs est $\frac{1}{2}\|X - UV\|^2_{F,W}$, comme définie dans (2.26), et pour laquelle la matrice de poids W est une matrice binaire définie par

$$w_{ij} = \begin{cases} 1 & \text{si } x_{ij} \text{ est observée} \\ 0 & \text{si } x_{ij} \text{ est manquante.} \end{cases}$$

Cette spécification de la matrice de poids W s'explique par le fait que les auteurs ont proposé dans leur étude de résoudre un problème de *données manquantes*, par l'imputation, dans une tâche de prédiction pour le système de recommandation collaboratif proposé par la firme cinématographique Netflix. Cependant nous pouvons considérer ici la matrice de poids, à coefficients positifs, quelconque pour faire une étude plus générale concernant la factorisation pondérée.

Alors par analogie au problème NMF classique de minimisation, comme dans (2.18), le problème NMF pondéré correspondant est donné par

$$(U^*, V^*) = \arg \min_{U\geq 0, V\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}. \tag{2.58}$$

Ensuite puisque la méthode ANLS est une procédure de minimisation alternée par bloc de coordonnées (*block coordinate descent method*), où chacune des deux matrices U et V représente un bloc de coordonnées, alors le problème (2.58) peut être subdivisé en deux sous-problèmes comme dans (2.19) :

$$U^* = \arg\min_{U\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}, \quad \text{et} \quad V^* = \arg\min_{V\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}. \tag{2.59}$$

Puis chacun des sous-problèmes dans (2.59) est subdivisé, à son tour, en problèmes NNLS. Alors on a :

$$V^* = \arg\min_{V\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W} \tag{2.60}$$

est partagé en p problèmes NNLS pondérés donnés par :

$$v^1_* = \arg\min_{v^1\geq 0} \frac{1}{2}\|x^1 - Uv^1\|^2_{2,W}\,,\ \ldots,\quad v^p_* = \arg\min_{v^p\geq 0} \frac{1}{2}\|x^p - Uv^p\|^2_{2,W} \tag{2.61}$$

ou de manière équivalente par les p problèmes NNLS suivants :

$$v^1_* = \arg\min_{v^1\geq 0} \frac{1}{2}\|(D^1x^1) - (D^1U)v^1\|^2_2\,,\ \ldots,\quad v^p_* = \arg\min_{v^p\geq 0} \frac{1}{2}\|(D^px^p) - (D^pU)v^p\|^2_2 \tag{2.62}$$

où $D^j = diag[(w^j)^{1/2}]$ est une matrice diagonale de taille $n \times n$ construite à partir du j ième vecteur colonne, w^j, de la matrice W, et pour lequel la racine carrée de chacune de ses composantes est considérée. Cette réécriture, proposée par les auteurs, des problèmes de (2.61) en ceux de (2.62) ont permis de passer de la norme euclidienne pondérée $\|\,.\,\|_{2,W}$ à la norme euclidienne classique $\|\,.\,\|_2$. Ce qui permet ainsi de retrouver, pour chacun des p problèmes de (2.62), la forme NNLS classique comme dans (2.20). Alors résoudre le problème (2.60) revient à résoudre les p problèmes NNLS de (2.62).

L'autre bloc de coordonnées est la matrice U qui permet aussi la minimisation de la fonction-coût dans (2.58), lorsque V est fixé. Il s'agit alors de considérer le problème suivant :

$$U^* = \arg\min_{U\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}. \tag{2.63}$$

Alors, là aussi, par analogie à la méthode ANLS de la factorisation NMF classique, on fait appel à l'opérateur de transposition matricielle pour obtenir, comme dans (2.22), le problème suivant :

$$U^T_* = \arg\min_{U^T\geq 0} \frac{1}{2}\|X^T - V^TU^T\|^2_{F,W} \tag{2.64}$$

Ainsi, de la même manière que pour V, les problèmes NNLS dérivant de la décomposition de (2.64) sont donnés par :

$$\begin{aligned} u^{T\,(1)}_* &= \arg\min_{u^{T\,(1)}\geq 0} \frac{1}{2}\|(D_1x^{T\,(1)}) - (D_1V^T)u^{T\,(1)}\|^2_2\,, \\ &\ldots \\ u^{T\,(n)}_* &= \arg\min_{u^{T\,(n)}\geq 0} \frac{1}{2}\|(D_nx^{T\,(n)}) - (D_nV^T)u^{T\,(n)}\|^2_2 \end{aligned} \tag{2.65}$$

où $D_i = diag[(w_i)^{1/2}]$ est une matrice diagonale de taille $p \times p$ construite à partir du i ième vecteur ligne, w_i, de la matrice W, et pour lequel la racine carrée de chacune de ses composantes est considérée. Alors résoudre résoudre le problème de minimisation (2.63) équivaut à résoudre le problème (2.64) ou de manière équivalente les n problèmes NNLS de (2.65).

Algorithme. Pour la résolution des problèmes NNLS, Kim et Choi (2009) [22] ont proposé la méthode de Newton projeté. Notons cependant que d'autres méthodes, comme la procédure Active Set de Lawson et Hanson (1974) [69], la méthode du Gradient projeté de Bro et de Jong (1997) [72], la méthode de Quasi-Newton projeté de Kim et al (2006) [75], ou toute autre méthode de minimisation adaptée, peuvent bien être utilisées.
Alors les p problèmes NNLS de (2.62) sont d'abord résolus pour trouver le premier bloc de coordonnées, V, qui rend décroissante la fonction-coût dans le problème (2.58), la matrice courante U étant fixée. Puis les n problèmes NNLS de (2.65) sont résolus afin de trouver le second bloc de coordonnées, U, qui fait décroître à son tour, la fonction-coût de (2.58), la matrice V nouvelle étant fixée. Ces deux mécanismes sont alors réalisés de façon alternée et itérative jusqu'à la convergence vers un minimum local (U^*, V^*) du problème (2.58) (Algorithme 9).

Données : Fournir la matrice des données positives X.
Résultat : Obtenir les estimations de U et V.
. Initialisation : $U^{(0)}$, $V^{(0)}$: initialiser les matrices par des coefficients positifs;
pour *t allant de* 1 *à Maxiter* **faire**
  . Mettre à jour la matrice V colonne par colonne
  pour *j allant de* 1 *à p* **faire**
    . $v^j_{(t+1)} \leftarrow \arg\min_{v^j \geq 0} \frac{1}{2}\|(D^j x^j) - (D^j U_{(t+1)})v^j\|_2^2$;
  fin
  . Mettre à jour la matrice U ligne par ligne
  pour *i allant de* 1 *à n* **faire**
    . $u^{T\,(i)}_{(t+1)} \leftarrow \arg\min_{u^{T\,(i)} \geq 0} \frac{1}{2}\|(D_i x^{T\,(i)}) - (D_i V^T_{(t)})u^{T\,(i)}\|_2^2$;
  fin
fin

Algorithme 9 : Algorithme pondéré ANLS-WNMF.

Conclusion

L'introduction de la méthode de factorisation de matrices positives NMF a été un tournant majeur pour les analyses statistiques multivariées. Introduisant alors un nouveau point de vue par rapport à des analyses factorielles déjà existantes et bien connues que sont l'ACP (Analyse en Composantes Principales), la méthode de décomposition en valeurs singulières, SVD (Singular Value Decomposition) et la méthode de quantification vectorielle, VQ (Vector Quantization), les auteurs Lee et Seung (1999 et 2001) ont permis aux analyses de descriptions des données multivariées de prendre une dimension supérieure. La modélisation de divers phénomènes aussi importants les uns que les autres dans divers domaines, comme dans le domaine médical, de l'industrie cinématographique avec le filtrage collaboratif, du traitement de signal, des questionnaires d'enquêtes ..., a alors été possible. Ainsi à la suite des algorithmes originels de Lee et Seung (2001), plusieurs autres algorithmes NMF ont été développées par divers auteurs. Alors ces différentes méthodes de factorisation NMF peuvent être classées en deux catégories que sont les algorithmes NMF classiques d'une part et les variantes de celles-ci d'autre part.

Dans ce chapitre nous avons présenté les principales méthodes de factorisation NMF classiques et quelques unes de leurs variantes, en l'occurrence les factorisations NMF pondérées. A ce titre, le chapitre a été articulé principalement autour de deux points.

D'abord les algorithmes NMF classiques ont été présentés. Il s'agit des trois principalement connus : les algorithmes de mise à jour multiplicative , les algorithmes du gradient de descente, et les algorithmes des moindres carrés alternés. S'agissant des procédures de mise à jour multiplicative MU, proposées par les auteurs Lee et Seung (2001) [20], pour l'estimation des matrices positives U et V de l'approximation factorielle $X \approx UV$, la formalisation du problème d'abord été faite. Puis la définition des fonctions-coûts telles que présentées par les auteurs a été donnée : il s'agit de la divergence de Kullback-Leibler généralisée et du carré de la distance euclidienne matricielle. Enfin les règles MU ont été données sous forme de résultats de théorèmes des auteurs et la convergence des algorithmes a aussi été discuté. Pour ce qui concerne les méthodes du gradient de descente pour la factorisation NMF, elles sont inspirées des méthodes numériques de recherche de minimum que sont les algorithmes à directions de descente sur des fonctions réelles, $f(x)$, différentiables et définies sur un espace euclidien ou plus généralement un espace hilbertien. C'est pourquoi la présentation des méthodes à directions de descente, de manière générale, et celle des méthodes du gradient de descente, en particulier, ont été faites. Ces dernières ont alors été adaptées à la factorisation NMF, où la fonction-coût, ici $f(U, V)$, est égale à la fonction carré de la distance euclidienne $\|X - UV\|_F^2$. La dernière classe des algorithmes NMF classiques est celle des algorithmes des moindres carrés alternés. Celle-ci a deux sous-classes essentiellement que sont les algorithmes ALS et ANLS. Ces derniers, rappelons-le, ont pour point commun la procédure d'alternance dans la mise à jours des matrices U et V. Toutefois les algorithmes ALS

sont plus simples à implémenter que ceux ANLS, car pour les premiers la mise à jour de chacune des deux matrices se fait de façon globale alors que pour les secondes les matrices sont mises à jour colonne par colonne ou ligne par ligne.

Le second principal point qui a été abordé dans ce chapitre a concerné quelques variantes des algorithmes classiques. Nous nous sommes cependant intéressés à la classe des algorithmes pondérés WNMF, même si plusieurs autres variantes existent encore. L'extension qui a été opérée sur les algorithmes classiques, a consisté à l'introduction d'une matrice de poids positive dans chacune des fonctions-coûts classiques, conduisant ainsi à l'obtention de fonctions-coûts pondérées. Alors les algorithmes WNMF qui ont découlés de la minimisation de ces fonctions-coûts pondérées, ont été déterminés de manière analogue à ceux de la factorisation NMF classique. C'est ainsi que Mao et Saul (2004) [21], Zhang et al (2006)[24], ou encore Blondel et al (2008) [77] ont proposé une extension des règles classiques MU de Lee et Seung (2001) [20], dans un contexte de factorisation pondérée WNMF. Par ailleurs, les auteurs Kim et Choi (2009) [22] ont développé l'extension des algorithmes de moindres carrés alternés ANLS, dans le cadre d'une factorisation pondérée, pour obtenir les algorithmes ANLS-WNMF. La détermination de ces derniers s'est aussi inspirée de celle des algorithmes ANLS classiques.

L'intérêt porté sur la classe des algorithmes pondérés WNMF, réside sur le fait qu'ils sont adaptables au traitement de données manquantes, en l'occurrence l'imputation. Notons que la problématique des données manquantes est le thème général abordé au chapitre suivant, et dans lequel s'inscrit le sujet particulier de notre thèse.

Deuxième partie

Problématique, État de l'art et Méthodologie proposée

Chapitre 3

Problématique des données manquantes dans les questionnaires d'enquêtes et État de l'art

Introduction

Lors de l'étude d'un phénomène, comme par exemple la cartographie d'une épidémie, une enquête sur le niveau de vie d'une population, ou un questionnaire administré à une structure, la collecte de données est une phase cruciale qui nécessite une bonne méthodologie afin d'éviter une absence de données qui biaiserait l'analyse statistique. Ainsi dans les questionnaires d'enquêtes plusieurs causes de données manquantes sont notées. Il s'agit d'abord des cas de pertes de données qui arrivent très souvent dans la pratique. D'autres causes expliquant l'absence de données sont aussi très souvent remarquées. Il s'agit des cas de refus de la part des répondants de renseigner sur certains points du questionnaire, soit du fait de leur caractère sensible, soit d'une incompréhension de certaines questions posées. Il s'y ajoutent aussi les cas de réponses inexploitables. Notons aussi qu'il peut très souvent arriver des situations où les enquêteurs (décideurs, instituts d'enquêtes ...), vu les coûts élevés et/ou les longues durées très souvent inhérents aux enquêtes peuvent choisir, suivant l'objectif recherché, d'ajuster les questionnaires en les réduisant en amont par le retrait aléatoire de certains points afin, par la suite, de procéder à une reconstruction automatique sans perte majeure d'information.

Ainsi l'on peut aisément constater la diversité des causes d'absence de données. Notons alors pour la suite, dans le cas spécifique des questionnaires, une donnée manquantes est aussi appelée *non-réponse*. La présence des non-réponses dans les données d'enquêtes constitue, comme nous l'avons dit, un réel problème dans les analyses. Alors la question naturelle serait de savoir comment utiliser les méthodes statistiques classiques (cas où les données sont complètes) face à la contrainte des non-réponses, ou aussi quelle(s) stratégie(s) pour résoudre ce problème. A ce titre plusieurs études se sont intéressées à la problématique des non-réponses. Différentes procédures d'approches ont alors été développées. Celles-ci peuvent être classées en deux catégories : les *méthodes sans complétion* et les *méthodes de complétion* encore appelées *méthodes d'imputation*. S'agissant des premières elles sont, pour l'essentiel, composées de deux procédures très connues que sont les *études de cas complets* et les *études de cas disponibles* (Little et Rubin, 1987 [14], Demissie et al, 2003 [78]). Quant aux secondes elles consistent à remplacer les données manquantes par des estimations que sont des valeurs les plus vraisemblables possibles afin de tenter de recouvrer l'information perdue du fait de l'absence de données. Elles ont été développées par divers auteurs, proposées comme alternatives aux méthodes sans complétion et sont incontournables lorsque le taux de données manquantes est élevé car permettant à défaut d'éliminer, de réduire considérablement le biais qui subsisterait dans les résultats des analyses sans complétion.

Les aspects théoriques des différentes études effectuées considèrent, en général, l'ensemble des données recueillies (observées ou non) disposées sous forme d'une matrice que nous notons ici par X et dont nous supposons de taille $n \times p$, où n désigne le nombre d'individus et p

le nombre de variables d'enquêtes. Il existe alors différentes matrices de données manquantes, suivant que l'on les regarde du point de vue de la configuration de l'absence des données ou des causes à l'origine des données manquantes. On parle alors de typologie des non-réponses. Celle-ci se regarde à deux niveaux. Le premier niveau concerne la répartition des non-réponses au sein de la matrice de données : on parle alors de structures de non-réponses. Trois principales structures ont été notées, il s'agit des structures univariées, monotones et arbitraires (Little et Rubin, 1987 [14]). Le second regroupe tous les mécanismes sous-jacents à l'absence de données. Essentiellement trois mécanismes (MAR, MCAR et MNAR) sont connus et ont été formalisés par Rubin (1976) [9]. . En effet, la dépendance à la structure de données manquantes a déjà été évoquée dans l'étude faite par les auteurs Afifi et Elashoff (1966) [6], qui ont effectué une large revue de la littérature de données manquantes multivariées et ont montré comment certaines études déjà réalisées pourraient être simplifiées si la structure des données manquantes était bien identifiée et correspondrait à des modèles bien connus, comme par exemple la structure univariée. Par ailleurs, Hartley et Hocking (1971) [7] ont développé des méthodes plus générales donc adaptées à une structure quelconque et basée sur des modèles probabilistes avec utilisation d'une vraisemblance avec des techniques d'estimation s'appuyant sur le principe du maximum de vraisemblance.

Parmi les nombreux domaines de la problématique des données manquantes largement étudiés dans la littérature, les cas des données manquantes liées à des questionnaires sont récurrents. En effet Little (1982) [13] a revisité plusieurs méthodes d'imputation de données manquantes en les spécifiant dans le cadre de données issues de questionnaires. En outre la méthodologie d'imputation multiple de non-réponses dans les données de questionnaires de Rubin (1987) [11], a été intensivement appliquée dans le domaine médical (van Buuren et Oudshoorn, 1999 [17], Taylor et al, 2002 [79], Joseph et al, 2004 [80], Shrive et al, 2006 [81], van Buuren, 2007 [82], van Ginkel et al, 2007 [83], White et al, 2011 [84], Resseguier et al, 2013 [85] ...). D'autres méthodes d'imputation compatibles avec des données de questionnaires ont été récemment développées. En effet l'algorithme EM (Espérance-Maximisation) de Dempster et al (1977) [56] a été développé sous un modèle poissonien de données augmentées par Cemgil (2009) [29] et adapté à l'imputation de données. Les auteurs Kim et Choi (2009) [22] ont revisité les méthodes de factorisation pondérées WNMF précédemment étudiées, d'abord par Mao et Saul (2004) [21], puis par Zhang et Wang (2006) [24]. Cependant, contrairement à Mao et Saul qui ont utilisé les règles MU pondérées (cf. Chap. 2, sect. 2.2.1), ou à Zhang et Wang qui ont développé une procédure EM, la méthode d'inférence proposée par Kim et Choi (2009) a été l'algorithme *ANLS-WNMF* (cf Chap. 2, sect. 2.2.2). Par ailleurs, sur la base de la méthode dites des *forêts aléatoires* de Breiman (2001) [40], les auteurs Stekhoven et Bühlmann (2011) [41] ont développé la méthode *MissForest* qui est une stratégie d'imputation de données mixtes (continues et/ou catégorielles) proposée comme une alternative à l'algorithme MICE (Multiple Imputation by Chained Equations) de van Buuren et Oudshoorn (1999) [17]. Plus récemment Josse et Husson (2016) [42] puis Audigier et al, 2017 [43] ont développé des méthodes d'imputations multiples par analyse des correspondances, *MIMCA* (*Multiple Imputation by Multiple Correspondence Analysis*), qui effectuent l'imputation de données catégorielles en utilisant l'analyse des correspondances multiples (ACM) (Josse et al, 2012 [44]).

Le chapitre est principalement articulé autour de trois points. D'abord le premier point analyse les causes pratiques de l'absence des données. Ces causes sont étudiées dans le contexte de questionnaires d'enquêtes où l'absence de données est caractérisée pour l'essentiel par la perte ou les non-réponses. Ces dernières sont de deux ordres : les non-réponses partielles et les non-réponses totales.
Ensuite le second point présente la typologie des données manquantes qui se manifeste à deux niveaux. D'un côté nous avons la structure des données manquantes et d'un autre nous avons les différents types de mécanismes qui causent l'absence de données.
Enfin l'état de l'art est discuté dans le dernier point. Il s'agit de donner une revue des principales méthodes de traitement des données manquantes, en insistant sur les méthodes d'imputation, avec essentiellement les différentes procédures récemment développées.

3.1 Typologie des données manquantes

Les solutions proposées aux différents problèmes de données manquantes, notamment dans les questionnaires, sont diverses et multiples. Comme nous l'avons évoqué dans l'introduction, des méthodes sans imputation et des procédures d'imputation ont été développées. Ces différentes stratégies de résolution tiennent compte de la typologie des données manquantes, c'est-à-dire, de la structure et/ou du mécanisme à l'origine de l'absence de données. Nous présentons alors les deux aspects de la typologie des données manquantes.

Formalisation et Notations. La problématique des données manquantes, dans ce chapitre, est étudiée dans le contexte de questionnaires d'enquêtes. Alors la formalisation du problème est telle que le données recueillies font intervenir le couple individu/variable. En effet l'enquête est adressée à des individus d'une population, et est structurée en plusieurs points qui sont les variables d'enquêtes (aussi appelées questions ou items). Les données sont alors disposées sous de matrice X de taille $n \times p$, où n est le nombre d'individus et p le nombre de variables. Nous notons alors les individus, qui représentent les lignes, par $X_1, X_2, \ldots, X_i, \ldots, X_n$, et les variables, qui désignent les colonnes, par $X^1, X^2, \ldots, X^j, \ldots, X^p$. Nous avons alors ces trois représentations possibles de la matrice X :

$$X = \left[X^1, X^2, \ldots, X^p\right], \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \qquad X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}$$

où $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, \ldots, n$ et $X^j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, $i = 1, \ldots, n$. Avec la considération des données manquantes, la matrice X sera aussi désignée par l'ensemble des données observées, noté par X^{obs}, et l'ensemble des données manquantes, noté par X^{mis}, c'est-à-dire $X = \{X^{obs}, X^{mis}\}$.

3.1.1 Structures des données manquantes

Comme nous l'avons déjà évoqué dans l'introduction, il existe trois sortes de structures des données manquantes : il s'agit des structures univariées, monotones et arbitraires.

Structure univariée. La structure est dite *univariée* lorsque les données manquantes apparaissent une seule variable. Considérons. Beaucoup des situations conduisent à ce cas de figure. Considérons le cas le plus simple de deux variables X^1 et X^2 où X^1 est entièrement observée alors X^2 sujette à des absences de valeurs. Un exemple illustratif simple est celui d'une enquête à deux variables X^1 et X^2, où la première variable désigne l'âge et la seconde représentant le revenu. Dans une situation expérimentale la variable X^1 peut être une variable entièrement observée ou une variable fixe contrôlée par l'expérimentateur, comme l'indicateur de traitement dans un modèle aléatoire. Les données sur X^2 peuvent être incomplètes en raison d'évènements incontrôlés au cours de la collecte de données, tels que la non-réponse, les valeurs contradictoires ou erreurs dans l'enregistrement des données. Les méthodes de traitements appropriées dans le cas d'une structure univariée peuvent être les *études de cas complets* ou les *études de cas disponibles.* S'agissant des premières, si nous considérons la matrice de données $X = \left[X^1, X^2, \ldots, X^p\right]$, où seule la variable X^p est incomplète , avec ses m premières composantes $x_{1p}, x_{2p}, \ldots, x_{mp}$, seules observées, tandis que ses $n - m$ valeurs restantes sont manquantes (illustration tableau 3.1), et que l'analyse statistique consiste à calculer les moyennes et les variances empiriques de chaque variable alors la variable incomplète X^p est écartée pour ne travailler qu'avec les variables complètes. Quant aux secondes méthodes, alors pour la variable X^p la moyenne et la variance sont évaluées sur ses m valeurs observées. Ces deux procédures sont appropriées, sous l'hypothèse que le mécanisme causant l'absence de données est ignorable et est du type MCAR. A l'opposé, si le mécanisme était non-ignorable, c'est-à-dire MNAR, l'étude de cas disponible induirait un biais

dans l'analyse, et dans ce cas d'autres méthodes de traitement comme l'imputation pourraient être envisagées.

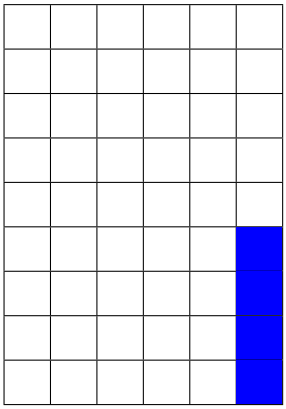

TABLE 3.1: Structure univariée où la couleur bleue représente les données manquantes.

Structure monotone. La structure est dite *monotone* lorsque les variables incomplètes sont regroupées et ordonnées, de façon croissante, en fonction de la proportion de données manquantes que chacune d'elles contient. Les cas de structures monotones fréquemment rencontrés sont les observations longitudinales incomplètes. Les données manquantes résultent alors d'une interruption de suivi, c'est-à-dire lorsqu'un évènement cause la sortie d'étude d'un sujet : on parle de phénomène d'attrition. Les patients qui quittent alors l'étude sont listés dans une partie à part du questionnaire CRF (*Case Report Form*) qui est une enquête en version papier ou électronique spécialement utilisée dans les domaines de recherches comme les essais cliniques. Les motifs spécifiés peuvent alors être un effet néfaste, une non efficacité du produit d'investigation, une maladie non liée à l'étude, un patient non coopérant, une violation du protocole, etc.
Lorsque le nombre de variables incomplètes est très petit par rapport au nombre total de variables, les méthodes de traitement comme les *études de cas complets* ou les *études de cas disponibles*, sous l'hypothèse du mécanisme MCAR sont également appropriées ici. Dans ce cas, supposons qu'il y a k variables entièrement observées et donc $p-k$ variables incomplètes comme dans le tableau 3.2. Alors pour les études de cas complets ces $p-k$ variables sont tout simplement retirées de l'étude, tandis que pour les études de cas disponibles les moyennes et variances sont calculées pour ces variables incomplètes par rapport à leurs valeurs observées respectives. Par

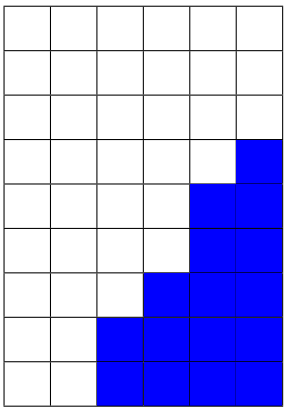

TABLE 3.2: Structure monotone où la couleur bleue représente les données manquantes.

contre lorsque leur nombre est grand, la perte massive de données inhérentes à ces méthodes pourrait conduire à la recherche d'autres alternatives comme l'imputation de données. Une première méthode serait alors de combler les cases vides de la matrice X, pour chacune des $p-k$ variables incomplètes, par les moyennes respectives de ces variables. On parle alors de *complétion*

par la moyenne, et l'on peut remarquer que les moyennes résultantes pour chaque variable sont les mêmes que celles des études de cas disponibles. Cependant la matrice de covariances résultante, est biaisée car les moyennes imputées sous-estiment les variances et covariances. Notons que la plupart des stratégies de résolution plus élaborées supposent les variables de loi normale multivariée et où les moyennes et les matrices de covariances par la méthode du maximum de vraisemblance.

Structure arbitraire. La structure *arbitraire*, aussi appelée structure *non-monotone* ou aléatoire est celle où les données manquantes apparaissent de façon aléatoire. La disposition des données manquantes est dans ce cas irrégulière comme nous pouvons le voir le tableau 3.3. Si nous regardons toujours le cas des exemples d'enquête sur les essais cliniques, la structure arbitraire se manifeste lorsqu'un patient ne se présente pas à une visite pour une raison quelconque, mais reste dans l'étude. Il se présentera donc aux visites suivantes. D'autres cas pratiques où la configuration de structures arbitraires apparaît sont les phénomènes de pertes de données ou encore les non-réponses aléatoires. Pour ce qui concerne les pertes, elles peuvent se manifester pendant ou après les enquêtes. En effet pendant la collecte, faute de procédures judicieuses, des fiches d'enquêtes où sont déjà relevées des informations recueillies peuvent être perdues ou partielles abimées par les enquêteurs. A la fin des enquêtes, lors du dépouillement et de l'enregistrement des données peuvent aussi être perdues. S'agissant des non-réponses aléatoires, elles peuvent être le fait d'une incompréhension de certains points du questionnaire par certains individus de l'échantillon soumis à l'enquête, ou d'un refus de répondre à des questions jugées sensibles ou indiscrètes par un certain nombre d'individus.
Comme la structure arbitraire a de commun avec celle monotone, l'apparition de l'absence de valeurs sur plusieurs variables, l'on peut tenter de retrouver la structure monotone en permutant les individus dans la matrice de données X. Les méthodes de traitement appropriées aux structures monotones peuvent alors être appliquées.

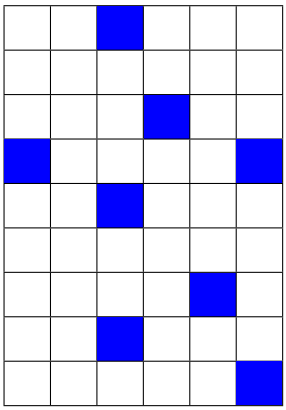

TABLE 3.3: Structure arbitraire où la couleur bleue représente les données manquantes.

Dans le cas échéant, lorsqu'il n'y a aucun réarrangement possible pour retrouver la structure monotone, alors l'on a une structure strictement arbitraire. Beaucoup de méthodes de résolution de données manquantes pour de telles structures ont été proposées. Alors comme dans le cas des structures monotones, les procédures d'inférences, comme par exemple l'estimation de la moyenne du vecteur des variables et de la matrice de covariances, reposent le plus souvent sur la méthode du maximum de vraisemblance. Les variables sont telles que le vecteur $(X^1, \ldots, X^j, \ldots, X^p)$ est supposée de loi normale multivariée, alors la méthode du maximum de vraisemblance est basée sur l'algorithme EM (Espérance-Maximisation) qui est une technique générale pour trouver l'estimateur du maximum de vraisemblance dans le contexte de données manquantes.

3.1.2 Mécanisme de génération des données manquantes

3.1.2.1 Quelques causes pratiques à l'origine des non-réponses

La collecte de données en vue de l'étude d'un phénomène est une phase très importante. En statistique la donnée constitue la matière essentielle pour appréhender un phénomène étudié. A ce titre, son recueil dans les meilleures conditions possibles est nécessaire pour une disposition en quantités suffisantes afin d'effectuer de bonnes inférences. Cependant dans la pratique des cas de données manquantes sont très souvent notés et les causes, comme nous l'avons dit en Introduction, sont d'origines diverses. Nous notons la présence du phénomène dans plusieurs champs disciplinaires ou secteurs d'activité. En effet dans le domaine médical et en particulier dans le cadre des essais cliniques, les données manquantes sont du fait des problèmes de non-compliance au traitement. On observe le même phénomène, dans des systèmes de surveillance, lorsqu'il y a des défauts de déclaration ou encore des déclarations incomplètes. L'apparition de données manquantes est également notée dans beaucoup d'autres domaines.
Toutefois nous mettons ici en relief le phénomène de non-réponse dans les questionnaires d'enquêtes. Pour celles-ci les causes pratiques à l'origine des non-réponses sont multiples et diverses. Ainsi les pertes de données pendant les enquêtes ou a posteriori lors des processus de conservation est un des facteurs récurrents. D'autres types de non-réponses sont fréquents et divers, dépendant ainsi des mécanismes qui les engendrent. En effet le refus de répondre, de la part de certains individus de la population enquêtée, sur certains points d'un questionnaire à cause leur incompréhension, est souvent noté. Un autre cas récurrent où des non-réponses sont enregistrées est celui lié au caractère sensible (sexualité, consommation de drogue, etc) de certains points du questionnaire. Dans certaines situations aussi il arrive qu'il y ait des réponses inexploitables, celles-ci sont alors considérées comme des non-réponses. Nous notons deux sortes de non-réponses : les non-réponses totales et les non-réponses partielles.

Une *non-réponse totale* est celle pour laquelle aucune information n'est recueillie sur une unité échantillonnée, c'est-à-dire lorsqu'il arrive qu'un individu interrogé n'a répondu à aucun point du questionnaire, soit lorsque les variables d'intérêt n'ont toutes enregistrées aucune valeur alors que les informations sur les variables auxiliaires ont été recueillies. Les variables auxiliaires constituent l'ensemble des informations marginales qui sont celles qui permettent de voir ce qui différencie les non-répondants des individus répondants. Alors il peut s'agir d'informations collectées de visu sur le terrain comme, par exemple, le type de logement (individuel ou collectif). Il peut également s'agir de données sociodémographiques ou médico-administratives récoltées au préalable. Notons cependant que les cas de non-réponses totales sont marginaux donc peu rencontrés dans la pratique. C'est pourquoi notre intérêt est porté sur les cas des non-réponses partielles qui sont les plus fréquemment rencontrés.

Une non-réponse est dite *partielle* lorsque, pour un individu enquêté, une partie seulement de l'information est relevée. C'est le cas où l'individu renseigne sur certains points du questionnaire et ne répond pas sur d'autres. Les causes de non-réponses partielles sont variées et il est important de pouvoir appréhender les mécanismes sous-jacents afin de permettre l'élaboration d'analyses adéquates. Ainsi suivant que l'on se place du côté de la population cible ou des structures chargées des enquêtes, on distingue des mécanismes involontaires et volontaires qui conduisent à des non-réponses respectivement de mêmes natures.

Les mécanismes sont qualifiés d'*involontaires* s'ils ne dépendent pas du fait que l'individu ait choisi de ne pas répondre à certains items pour une raison ou pour une autre, mais plutôt de facteurs extérieurs comme, par exemple, l'oubli, la barrière linguistique ou encore les conditions d'entretien. D'autres mécanismes involontaires sont notés dans la pratique. En effet une cause involontaire peut être la complexité de certaines questions qui conduisent naturellement à une méconnaissance de la réponse à donner. Cet état de fait est le plus souvent dû à une non adaptation du questionnaire par rapport à la cible. Une autre source de non-réponse découle, dans certains cas, des problèmes de mémorisation du fait de l'ancienneté des informations à fournir. Il peut s'agir par exemple de questions sur des consommations alimentaires plus ou moins

anciennes. Par ailleurs, pour un bon déroulement des enquêtes, la stabilité des condition d'entretien est aussi requise. En effet lorsque les conditions de recueil d'informations sont instables la collecte de données peut être interrompue entrainant du coup l'obtention d'une information partielle. C'est par exemple le cas lors d'enquêtes téléphoniques avec une défaillance du réseau ou d'enquêtes réalisées auprès de populations marginalisées telles que les sans domicile fixe, les usagers de drogues ou les populations carcérales avec lesquelles des difficultés très souvent rencontrées sur le terrain.

A l'opposé, les mécanismes *volontaires* eux sont caractérisés par le fait que la non-réponse est une conséquente directe de la rétention de l'information de la part du répondant à un questionnaire. Ce sont des mécanismes qui dépendent des variables d'enquêtes. A ce titre, le refus de donner l'information est directement lié à la nature même des questions posées. Deux cas se posent principalement. Le premier est celui où le refus de donner l'information concernant une variable, de la part du répondant, ne dépend que des informations qu'il fournit sur une ou plusieurs autres variables, et non sur la variable en question. Le second est celui où le refus de renseigner par rapport à une variable dépend de la variable elle-même, et éventuellement d'autres variables. Ce dernier cas correspond à celui des questions à caractère sensible (exemple, sexualité, consommation d'alcool ou de tabac, déclaration au fisc, etc). Toutefois l'appréciation du degré de sensibilité de certains points d'un questionnaire peut varier d'un individu à un autre ou encore d'une catégorie socioculturelle à une autre.
Si l'on se situe du côté des enquêteurs par contre les mécanismes volontaires sont ceux pour lesquels non-réponses sont dues à un sous-échantillonnage du questionnaire en amont. Les non-réponses sont donc du fait des questions non soumises. Il s'agit d'un cas intéressant que nous examinerons au dernier chapitre.

Nous voyons ainsi que les facteurs à l'origine de l'absence des données sont multiples et diverses. Alors pour éviter au maximum le risque de non-réponses lors d'une enquête, plusieurs précautions devraient être prises. Il s'agit entre autres d'une bonne formation d'abord des enquêteurs, ensuite d'une bonne qualité du questionnaire en évitant les questions mal conçues, trop longues et inciter les enquêtés à répondre par le biais de questions courtes concises impliquant ainsi des réponses courtes et claires. Aussi veiller à bien placer les questions délicates en évitant des les mettre en début de questionnaire, sinon cela pourrait conduire le répondant à abréger son contact avec l'enquêteur, ou dans le cas particulier d'une enquête par téléphone, à couper l'entretien téléphonique. Enfin il faudrait bien choisir la cible afin de pouvoir assurer un suivi des non-répondants jusqu'à ce qu'ils puissent apporter les réponses nécessaires.

3.1.2.2 Formalisation des mécanismes conduisant à l'absence de données

Modélisation du mécanisme de non-réponse. La connaissance du mécanisme à l'origine des valeurs manquantes est une considération très importante pour l'application d'une méthode de résolution appropriée. Alors pour des études théoriques basées sur des fondements mathématiques il a été trouvé nécessaire de formaliser les différentes causes pratiques d'absence de données. Cette formalisation introduite par Rubin (1976) [9] puis encore plus expliquée par Little et Rubin (1987) [14], a été développée autour de trois types de mécanismes (MAR, MCAR et MNAR). Rappelons que la matrice X représente l'ensemble de données recueillies et comportant des données manquantes. Alors si nous reprenons l'écriture de la matrice de données $X = \{X^{obs}, X^{mis}\}$, pour chaque individu i , $i = 1, \ldots, n$ de l'échantillon, on note X_i^{obs} l'ensemble des valeurs renseignées et X_i^{mis} l'ensemble des valeurs manquantes le concernant. Donc les données sur l'individu, i peuvent être désignées par $X_i = \{X_i^{obs}, X_i^{mis}\}$. A chaque vecteur X_i on associe un autre vecteur $R_i = (r_{i1}, r_{i2}, \ldots, r_{ip})$ de même taille $1 \times p$, et qui renseigne sur la nature des informations, suivant qu'elles soient des réponses ou des non-réponses de l'individu i. Chaque composante r_{ij}, $j = 1, \ldots, p$ de R_i vaut 1 si la valeur x_{ij} correspondante est observée, et vaut 0 si x_{ij} est manquante. Dans certains cas la matrice X peut contenir un ensemble de variables complètes. Nous distinguerons alors parmi les variables $X^1, \ldots, X^l, X^{l+1}, \ldots, X^p$, les covariables entièrement observées $X^1, \ldots, X^l$ d'un côté, et les covariables incomplètes $X^{l+1}, \ldots, X^p$, d'un autre, que nous notons par $Y^1, \ldots, Y^k$, avec respectivement $X^{l+1} = Y^1, \ldots, X^p = Y^k$. et donc

on a $p = l + k$. Dans ce cas la matrice X est donnée par :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1l} & y_{11} & y_{12} & \cdots & y_{1k} \\ x_{21} & x_{22} & \cdots & x_{2l} & y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nl} & y_{n1} & y_{n2} & \cdots & y_{nk} \end{pmatrix}$$

La sous-matrice Y désigne les k covariables incomplètes pour les n individus. L'ensemble des vecteurs lignes R_i , $i = 1, \ldots, n$ modélisent en fait les non-réponses des individus (0 à chaque composante en cas de non-réponse) et forment la matrice des indicateurs de non-réponses R donnée par :

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{bmatrix}$$

L'ensemble des données observées X^{obs} et des données manquantes X^{mis} peuvent alors s'exprimer en fonction des indicateurs de réponses R :

$$X^{obs} = X\mathbb{1}_{\{R=1\}} \quad \text{et} \quad X^{mis} = X\mathbb{1}_{\{R=0\}} \quad \text{avec} \quad \begin{cases} \mathbb{1}_{\{R=1\}} = 1 \\ \mathbb{1}_{\{R=0\}} = 0 \end{cases}$$

Ainsi, l'usage des méthodes de résolution efficaces, nécessite l'appréhension de la typologie des données manquantes dans sa globalité, c'est-à-dire, en plus de la structure (univarié, monotone ou arbitraire), connaître le mécanisme sous-jacent qui explique l'absence de certaines valeurs. Toutefois, le plus souvent dans la pratique, les données en elles mêmes ne fournissent pas d'information quant au mécanisme (Rubin, 1978 [10]). D'où l'importance de disposer des informations a priori sur le mécanisme. Deux principales modélisations du mécanisme de données manquantes sont retenus dans les analyses statistiques. Il s'agit des *mécanismes ignorables*, qui renferment les deux types de mécanismes *MAR* et *MCAR*, et des *mécanismes non-ignorables*, constitués du type de mécanisme *MNAR*.

3.1.2.3 Les mécanismes ignorables : MAR et MCAR

La notion de *mécanisme ignorable* dans l'étude des données manquantes a été introduite par Rubin (1976) [9]. La formalisation du mécanisme s'est alors faite par la construction d'une distribution de probabilité, $g(R \,|\, X, \phi)$, sur les indicateurs non-réponses, R, conditionnellement aux données X et à un paramètre ϕ. Alors le concept d'*ignorabilité* de mécanisme sous-jacent à l'absence de données a été étudié par Rubin dans un contexte d'inférence sur un paramètre. En effet, si les données X sont supposées distribuées suivant une loi de densité $f(X \,|\, \theta)$, où θ est le paramètre qui conditionne les données, le travail de l'auteur a consisté a montré comment le fait d'*ignorer* le mécanisme à l'origine des données manquantes, lors de l'inférence sur le paramètre θ peut être considérée comme une procédure correcte, c'est-à-dire une méthode *appropriée*. Pour cela des conditions ont été nécessaires sur le mécanisme, ou plus précisément la distribution $g(R \,|\, X, \phi)$ à des valeurs fixées $\tilde{R}$ et $\tilde{X}^{obs}$ pour respectivement les indicateurs de non-réponses R et les données observées X^{obs}, c'est-à-dire $g(\tilde{R} \mid \tilde{X}^{obs}, \phi)$.

Le constat fait par Rubin était alors que la plupart des analyses antérieures avec données manquantes dont le mécanisme était non contrôlé par le statisticien (Afifi et Elashoff, 1966 [6] ; Anderson, 1957 [3] ; Hartley et Hocking, 1971 [7] ; Healy et Westmacott, 1956 [2] ; Wilkinson, 1958 [4]), ignoraient ce mécanisme sans pour autant se poser la question qui consiste à s'assurer si l'inférence y afférent était appropriée. C'est la question à laquelle a tenté de répondre Rubin (1976) [9] en formalisant d'abord l'ignorabilité du mécanisme dans le contexte d'inférence d'un paramètre d'une distribution de probabilité. Rappelons que le paramètre, θ en question est celui d'une distribution, de densité $f(X \mid \theta)$. L'inférence sur θ est effectuée à partir de valeurs fixées

des données observées $\tilde{X}^{obs}$ et des indicateurs, correspondants, de non-réponses $\tilde{R}$. Notons que ces derniers caractérisent aussi la structure observée des données manquantes. Ainsi *ignorer le mécanisme* revient d'abord (1) à fixer la variable aléatoire R à sa valeur observée $\tilde{R}$, puis (2) à supposer que les données observées $\tilde{X}^{obs}$ sont distribuées suivant la loi marginale des données X^{obs} :

$$f(X^{obs} \,|\, \theta) = \int f(X^{obs}, X^{mis} \,|\, \theta) \mathrm{d}X^{mis} \tag{3.1}$$

L'équation (3.1) montre bien l'omission de la distribution du mécanisme de non-réponse, $g(R \,|\, X, \phi)$, dans l'inférence du paramètre θ basée sur la distribution marginale des données observées, $f(X^{obs} \,|\, \theta)$, conditionnée uniquement par le paramètre θ.

L'inférence *appropriée* (ou *correcte*), est celle qui prend en compte le mécanisme de données manquante, c'est-à-dire celle qui considère la distribution $g(R \,|\, X, \phi)$, dans le processus d'inférence sur θ, puisqu'il y a effectivement des données manquantes. Alors la question centrale est de savoir comment l'inférence avec *ignorabilité* du mécanisme peut être considérée comme une inférence *appropriée*. Autrement dit, pour une configuration fixée, $\tilde{R}$, de données manquantes et pour des données X, quelles conditions devrait-on imposer sur la probabilité $g(\tilde{R} \mid X, \phi)$ pour que le fait d'*ignorer* le mécanisme, lors de l'inférence sur θ, soit considéré comme une procédure *correcte*. Pour y répondre trois conditions sur la probabilité $g(\tilde{R} \mid X, \phi)$ ont été établies par Rubin (1976) :

- condition 1 : les données manquantes le sont aléatoirement ;
- condition 2 : les données observées le sont aléatoirement ;
- condition 3 : les paramètres ϕ et θ sont distincts ;

qui sont respectivement ainsi définies,

Définition 3.1. *Les données manquantes le sont aléatoirement (missing at random) si pour chaque valeur de ϕ, la probabilité $g(\tilde{R} \mid X, \phi)$ est constante, quelles que soient les données manquantes de X^{mis}.*

Définition 3.2. *Les données observées le sont aléatoirement (observed at random) si pour chaque valeur de ϕ et de X^{mis}, la probabilité $g(\tilde{R} \mid X, \phi)$ est constante, quelles que soient les données observées X^{obs}.*

Définition 3.3. *Le paramètre ϕ est distinct du paramètre θ si leur espace joint Ω se factorise en Ω_ϕ et Ω_θ , ($\Omega = \Omega_\theta \times \Omega_\phi$) et si ϕ et θ sont indépendants lorsque les distributions a priori respectives sont spécifiées.*

Deux types d'inférences sur le paramètre θ ont alors servi à l'étude. Il s'agit de l'inférence à partir de la distribution d'échantillonnage et de l'inférence basée sur la vraisemblance du paramètre conditionnellement aux observations.

Inférence basée sur la distribution d'échantillonnage. C'est une méthode d'inférence qui permet de comparer la valeur observée d'une statistique (estimateur, critère de test, ou intervalle de confiance) par rapport à sa distribution d'échantillonnage, sous diverses hypothèses sous-jacentes. Il s'agit donc d'appréhender la distribution d'une statistique par rapport à sa valeur observée $S(\tilde{R}, \tilde{X}^{obs})$. L'inférence *correcte*, pour ce type d'analyse, est alors celle qui prend en compte la distribution qui conditionne l'absence de données $g(R \mid X, \phi)$. Dans ce cas l'inférence sur θ via la distribution d'échantillonnage de la statistique observée dépend de la loi marginale des données observées X^{obs} conditionnellement aux indicateurs de non-réponses $\tilde{R}$ et aux paramètres θ et ϕ. Celle-ci est obtenue en intégrant, par rapport aux données manquantes X^{mis}, la loi conditionnelle des données X sachant les indicateurs $\tilde{R}$.

$$f(X^{obs} \mid \tilde{R}, \theta, \phi) = \int f(X \mid \theta) g(\tilde{R} \mid X, \phi) / K_{\theta,\phi}(\tilde{R}) \, \mathrm{d}X^{mis} \tag{3.2}$$

où $K_{\theta,\phi}(\tilde{R}) = \int f(X \,|\, \theta) g(\tilde{R} \,|\, X, \phi) \mathrm{d}X$ est la probabilité marginale que les indicateurs R prennent la valeur $\tilde{R}$. La distribution d'échantillonnage de la valeur observée de la statistique $S(\tilde{R}, \tilde{X}^{obs})$

est alors déterminée à partir de la distribution $f(X^{obs} \mid \tilde{R}, \theta, \phi)$, et l'équation (3.2) montre bien la considération du mécanisme sous-jacent. Par ailleurs, l'inférence sur θ en *ignorant* le mécanisme consiste à déterminer la distribution d'échantillonnage de la statistique observée $S(\tilde{R}, \tilde{X}^{obs})$ à partir de la loi de densité $f(X^{obs} \mid \theta)$ dans (3.1). Alors la question est de savoir comment les deux approches peuvent être équivalentes. Pour y répondre des conditions parmi celles définies ci-dessus ont été considérées par Rubin et le résultat donné sous forme de théorème :

Théorème 3.1. *Supposons que (a) les données manquantes sont manquantes aléatoirement et (b) les données observées sont observées aléatoirement. Alors la distribution d'échantillonnage de la statistique observée $S(\tilde{R}, \tilde{X}^{obs})$ liée à $f(X \mid \theta)$ en ignorant le mécanisme qui cause les données manquantes, c'est-à-dire calculée à partir de la densité* (3.1)*, est égale à la distribution d'échantillonnage de $S(\tilde{R}, \tilde{X}^{obs})$ liée à $f(X \mid \theta)g(R \mid X, \phi)$, c'est-à-dire qui est calculée à partir de la densité* (3.2) *en supposant que $K_{\theta,\phi}(\tilde{R}) > 0$.*

Preuve. Sous les hypothèses (a) et (b), pour chaque valeur de ϕ la probabilité $g(\tilde{R} \mid X)$ ne varie pas quelles que soient les données X. Par conséquent $K_{\theta,\phi}(\tilde{R}) = g(\tilde{R} \mid X)$, et donc les deux densités dans (3.1) et (3.2) sont égales. Ce qui fait que la distribution d'échantillonnage de toute statistique $S(\tilde{R}, \tilde{X}^{obs})$ et dépendant de la densité (3.1) est égale à celle dépendant de la densité (3.2). □

Inférence basée sur la vraisemblance. Le second type d'inférence sur θ est celui faite à partir de la vraisemblance du paramètre conditionnellement aux données. Alors deux sortes d'analyses ont été faites : l'inférence directe et l'inférence bayésienne.

Inférence directe : L'approche d'inférence directe proposée par Rubin a consisté au calcul du ratio de vraisemblance pour différentes valeurs du paramètre. Il s'agit de comparer l'adéquation de l'ajustement de deux modèles afin de déterminer celui qui offre le meilleur ajustement pour les données. Dans le contexte de l'étude faite par Rubin, les paramètres θ et ϕ prennent leurs valeurs dans l'espace joint des paramètres $\Omega_{\theta,\phi}$

Ainsi lorsque l'inférence sur le paramètre θ est effectuée, *ignorer le mécanisme* des données manquantes revient, comme nous l'avons déjà dit, à considérer la densité de (3.1). Alors la vraisemblance qui en découle est celle du paramètre θ par rapport aux données observées $\tilde{X}^{obs}$, et qui est donnée par :

$$L(\theta \mid \tilde{X}^{obs}) = \delta(\theta, \Omega_\theta) f(\tilde{X}^{obs} \mid \theta) \tag{3.3}$$

où $\delta(a, \Omega)$ est la fonction indicatrice de Ω.

A l'opposé, avec l'inférence *appropriée* le mécanisme est pris en compte, donc la distribution $g(R \mid X, \phi)$ indicateurs de non-réponses R. Dans ce cas, distribution du modèle est la loi jointe des indicateurs R et des données X conditionnellement aux paramètres θ et ϕ,

$$f(X, R \mid \theta, \phi) = f(X \mid \theta) g(R \mid X, \phi) \tag{3.4}$$

La loi jointe d'intérêt est celle marginale du couple de variables (X^{obs}, R) qui est donnée par

$$f(X^{obs}, R \mid \theta, \phi) = \int f(X^{obs}, X^{mis} \mid \theta) g(R \mid X^{obs}, X^{mis}, \phi) \mathrm{d}X^{mis} \tag{3.5}$$

Alors par rapport aux valeurs fixées $(\tilde{X}^{obs}, \tilde{R})$ l'inférence directe, via le rapport de vraisemblances, sur le couple (θ, ϕ) nécessite la considération de la fonction de vraisemblance définie par :

$$L(\theta, \phi \mid \tilde{X}^{obs}, \tilde{R}) = \delta\left((\theta, \phi), \Omega_{\theta,\phi}\right) f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi) \tag{3.6}$$

où $\delta\left((\theta, \phi), \Omega_{\theta,\phi}\right)$ est la fonction indicatrice de l'espace des paramètres $\Omega_{\theta,\phi}$.

Là également des conditions parmi celles définies plus haut on été supposées, par Rubin, sur le mécanisme sous-jacent pour obtenir une équivalence des deux approches ; autrement dit pour que l'inférence avec *ignorabilité* du mécanisme soit une inférence *appropriée*. Le résultat a été donné sous forme de théorème :

Théorème 3.2. *Supposons (a) que les données manquantes sont manquantes aléatoirement et (b) que le paramètre ϕ est distinct du paramètre θ. Alors l'inférence par le rapport de vraisemblance $\frac{L(\theta_1 \mid \tilde{X}^{obs})}{L(\theta_2 \mid \tilde{X}^{obs})}$ en* ignorant *le mécanisme qui cause les données manquantes, est équivalent à l'inférence* correcte *via le rapport de vraisemblances $\frac{L(\theta_1,\phi \mid \tilde{X}^{obs},\tilde{R})}{L(\theta_2,\phi \mid \tilde{X}^{obs},\tilde{R})}$, $\forall \phi \in \Omega_\phi$ tel que $g_\phi(\tilde{R} \mid \tilde{X}) > 0$.*

Preuve. Pour $\tilde{X}^{obs}$ et $\tilde{R}$ et fixées dans (3.3), la condition (a) permet d'obtenir la relation suivante

$$f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi) = g(\tilde{R} \mid \tilde{X}^{obs}, \phi) \int f(\tilde{X}^{obs}, X^{mis} \mid \theta) \mathrm{d}X^{mis}$$

qui implique que

$$f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi) = g(\tilde{R} \mid \tilde{X}^{obs}, \phi) f(\tilde{X}^{obs} \mid \theta)$$

La condition (b) permet de factoriser la fonction indicatrice de l'espace joint $\Omega_{\theta,\phi}$, c'est-à-dire

$$\delta((\theta, \phi), \Omega_{\theta,\phi}) = \delta(\theta, \Omega_\theta)\, \delta(\phi, \Omega_\phi)$$

De ces deux dernières équations permettent bien d'obtenir la relation qui existe entre la vraisemblance lorsque le mécanisme est ignoré et la vraisemblance approprié, c'est-à-dire si le mécanisme est pris en compte, comme le montre la relation suivante :

$$L(\theta, \phi \mid \tilde{X}^{obs}, \tilde{R}) = \delta(\phi, \Omega_\phi)\, g(\tilde{R} \mid \tilde{X}^{obs}, \phi)\, L(\theta \mid \tilde{X}^{obs})$$

Cette relation montre que les deux rapports de vraisemblances, pour θ_1 et θ_2 fixés, sont égaux. Ce qui permet d'avoir l'équivalence des deux approches. □

Inférence bayésienne : La seconde méthode d'inférence basée sur la vraisemblance est l'inférence bayésienne. Les paramètres θ et ϕ sont dans ce cas regardés comme des variables aléatoires auxquelles sont affectées des lois a priori. La loi du couple (θ, ϕ) a été spécifiée par le produit $p(\theta)p(\phi \mid \theta)$, et l'inférence a consisté à la détermination des lois a posteriori des paramètres. Pour ce dernier type d'analyse également, il sera étudié les conditions pour lesquelles les deux sortes de procédures que sont l'inférence avec mécanisme ignoré et l'inférence avec prise en compte du mécanisme (approche appropriée)

L'inférence bayésienne qui *ignore* le mécanisme à l'origine des données manquantes est celle effectuée uniquement sur le paramètre θ et qui considère donc sa loi a priori $p(\theta)$, puis que les données observées sont distribuées suivant la loi de densité donnée dans (3.1). Par conséquent la loi a posteriori de θ est proportionnelle au produit de ces deux distributions :

$$\pi(\theta \mid \tilde{X}^{obs}) \propto p(\theta) f(\tilde{X}^{obs} \mid \theta) \tag{3.7}$$

L'inférence bayésienne *appropriée* est celle qui cherche à estimer le couple de paramètres (θ, ϕ), puisque le processus qui cause les données manquantes est bien considéré. Dans ce cas le modèle est tel que la distribution jointe des données observées $\tilde{X}^{obs}$ et de la structure observée de données manquantes $\tilde{R}$ est donnée par (3.5). La loi a posteriori du couple (θ, ϕ) est alors donnée par :

$$\pi(\theta, \phi \mid \tilde{X}^{obs}, \tilde{R}) \propto p(\theta)p(\phi \mid \theta) f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi) \tag{3.8}$$

Pour que les deux approches d'inférence soient équivalentes, des conditions ont également été supposées sur le mécanisme à l'origine de données manquantes $g(R \mid X, \phi)$: (a) les données manquantes sont manquantes aléatoirement et (b) le paramètre ϕ est distinct du paramètre θ. Cela permettra alors considérer indifféremment les inférences basées respectivement sur (3.7) et (3.8) pour l'estimation du paramètre θ, comme le montre le théorème suivant.

Théorème 3.3. *Supposons (a) que les données manquantes le sont aléatoirement et (b) que le paramètre ϕ est distinct θ. Alors l'inférence effectuée avec la distribution a posteriori de θ en* ignorant *le mécanisme qui cause les données manquantes, c'est-à-dire celle qui est calculée à partir de* (3.7)*, est équivalente à l'inférence faite avec la distribution a posteriori* correcte *de θ, c'est-à-dire celle qui est calculée à partir de* (3.8)*.*

Preuve. Considérons l'expression de $f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi)$ dans (3.5), c'est-à-dire,

$$\int f(\tilde{X}^{obs}, X^{mis} \mid \theta) g(\tilde{R} \mid \tilde{X}^{obs}, X^{mis}, \phi) \mathrm{d}X^{mis},$$

alors l'hypothèse (a) implique que $g(\tilde{R} \mid \tilde{X}^{obs}, X^{mis}, \phi) = g(\tilde{R} \mid \tilde{X}^{obs}, \phi)$, d'où

$$f(\tilde{X}^{obs}, \tilde{R} \mid \theta, \phi) = g(\tilde{R} \mid \tilde{X}^{obs}, \phi) \int f(\tilde{X}^{obs}, X^{mis} \mid \theta) \mathrm{d}X^{mis} \quad (*)$$

et l'hypothèse (b) faite sur les paramètres ϕ et θ et qui stipule qu'ils soient distinctes, implique qu'elles sont indépendantes étant donné que leurs densités a priori respectives $p(\phi)$ et $p(\theta)$ sont spécifiées ; ce qui entraîne alors que

$$p(\theta)p(\phi \mid \theta) = p(\theta)p(\phi) \quad (**)$$

Des équations $(*)$ et $(**)$, on obtient que le membre de droite de l'équation (3.8) est égale à

$$(p(\phi) g(\tilde{R} \mid \tilde{X}^{obs}, \phi)) (p(\theta) \int f(\tilde{X}^{obs}, X^{mis} \mid \theta) \mathrm{d}X^{mis})$$

cette expression étant proportionnelle à $\pi(\theta \mid \tilde{X}^{obs})$ on obtient donc, par rapport à θ, une relation de proportionnalité entre les lois a posteriori (3.7) et (3.8). Ce qui veut dire que l'inférence sur θ peut se faire indifféremment avec l'une ou l'autre des lois a posteriori, d'où l'équivalence des deux approches d'inférences. □

L'équivalence entre les deux types d'inférences cités dans ce théorème se traduit par la proportionnalité des lois a posteriori correspondantes, comme c'était le cas avec les vraisemblances. D'ailleurs les mêmes hypothèses sont conservées aussi bien pour l'inférence directe que pour l'inférence bayésienne.

L'étude faite par Rubin (1976) [9] a été un tournant majeur dans les analyses statistiques multivariées avec données manquantes. Elle a permis de « corriger une erreur » commise par plusieurs études antérieures, qui consistait à ignorer, d'une manière ou d'une autre, le mécanisme sous-jacent à l'absence de données, sans pour autant s'assurer que la procédure était appropriée. Notons qu'elle a été faite dans des analyses d'inférences sur des paramètres de modèles statistiques en présence de données manquantes. Aussi elle a marqué le début de la catégorisation des processus causant les données manquantes en mécanismes *ignorables* et *non-ignorables*. Notons cependant que ni la dénomination de *mécanismes ignorables* ni la spécification des ceux-ci n'a été faite par Rubin (1976) [9]. L'intérêt porte surtout sur les conditions 1 et 2 qui ont permis de spécifier les mécanismes ignorables dans une étude postérieure. En effet Little et Rubin (1987) [14] ont définit, d'une part, les mécanismes ignorables (*MAR* et *MCAR*) en les identifiant aux conditions pour lesquels les inférences associées sont qualifiées d'appropriées (cf. théorèmes 3.1, 3.2 et 3.3), et d'autre part les mécanismes non-ignorables (*MNAR*).

Mécanisme MAR. Le mécanisme *MAR* (*Missing At Random*) désigne que les données manquantes le sont au hasard. Il correspond donc naturellement à la condition 1 (hypothèse (a) dans les théorèmes). Alors comme on peut le voir dans la définition 3.1, la formalisation mathématique du mécanisme s'est faite par le biais d'une loi de probabilité. Il s'agit en fait d'une loi discrète puisque les valeurs possibles des indicateurs de non-réponses R sont 0 et 1. Ainsi dans un jeu de données, les valeur manquantes suivent le mécanisme *MAR*, si la loi de probabilité des indicateurs de non-réponses R conditionnellement aux données X, ne dépend que des valeurs observées X^{obs} :

$$P(R \mid X^{obs}, X^{mis}, \phi) = P(R \mid X^{obs}, \phi), \ \forall \ \phi, \ X^{obs} \text{ et } X^{mis}. \quad (3.9)$$

Nous pouvons donc identifier la probabilité $P(R \mid X, \phi)$ à la probabilité $g(R \mid X, \phi)$ précédemment définie comme loi du mécanisme, c'est-à-dire que $P(R \mid X, \phi) = g(R \mid X, \phi)$. L'hypothèse

MAR sur les données est très souvent rencontrées dans les études de traitement des données manquantes, notamment les inférences à partir des fonctions de vraisemblances. C'est aussi une typologie compatible avec les inférences basées sur la vraisemblance(inférences directes ou bayésiennes).
Un cas illustratif simple pour comprendre ce mécanisme est de l'exemple d'un jeu de données recueillies d'une enquête à deux variable, c'est-à-dire $X = [X^1, X^2]$ où X^1 désigne l'âge qui est entièrement observée et X^2, le revenu qui contient des valeurs manquantes. Ainsi si la probabilité de réponse sur le revenu pour un individu i ne dépend pas du revenu, mais dépend de son âge et éventuellement de l'âge des autres individus, alors les données sont dites *MAR*. Dans ce cas les valeurs observées de X^2 (le revenu) ne sont pas nécessairement un sous-échantillon aléatoire de l'échantillon global des données de X^2, mais sont un sous-échantillon tiré conditionnellement à des sous-classes définies par les valeurs de X^1 (l'âge).

Mécanisme MCAR Le mécanisme *MCAR* (*Missing Completely At Random*) désigne que les données manquantes le sont complètement au hasard. Il est la résultante de l'association des conditions 1 et 2 (hypothèses (a) et (b) du théorème 3.1), donc mécanisme selon lequel les données manquantes le sont au hasard et les données observées le sont aussi au hasard. Le mécanisme peu donc être vu comme un cas particulier du mécanisme *MAR*, puisque la condition 1 leur est commune. Ainsi dans un jeu de données, les valeurs manquantes suivent le mécanisme *MCAR*, si la loi de probabilité des indicateurs de non-réponses R conditionnellement aux données X, ne dépend pas des données et ne varie pas pour tout R :

$$P(R \mid X, \phi) = P(R \mid \phi) = p, \ \forall\, X \text{ et } \phi. \tag{3.10}$$

où $0 < p < 1$ est une constante. Dans ce cas les valeurs observées de chaque variable incomplète forment un sous-échantillon aléatoire de l'ensemble des valeurs échantillonnées de la variable. Le mécanisme *MCAR* est compatible aussi bien avec les inférences faites avec distribution d'échantillonnage que celles basées sur la fonction de vraisemblance.

3.1.2.4 Les mécanismes non-ignorables : MNAR

Les mécanismes non-ignorables sont ceux dont le processus d'apparition des données manquantes est complètement appréhendé. Autrement dit l'on connaît comment les données manquantes sont survenues. Il s'agit de mécanismes plus connus sous l'appelation *MNAR* (*Missing Not At Random*) qui désigne que les données manquantes ne le sont pas au hasard. Ainsi un jeu de données, les valeurs manquantes suivent le mécanisme *MNAR*, si la loi de probabilité des indicateurs de non-réponses R conditionnellement aux données X, dépend des données manquantes X^{mis} :

$$P(R \mid X, \phi) = P(R \mid X^{mis}, \phi), \ \forall\, X^{mis} \text{ et } \phi, \tag{3.11}$$

dans certains cas la probabilité peut aussi dépendre, en plus des données manquantes, de certaines données observées. Pour les méthodes de traitement, dans un contexte *MNAR*, ne pas tenir compte des non-réponses peut engendrer un biais important dans l'analyse (par exemple, estimation d'un paramètre). Pour pallier ce problème, on peut modéliser conjointement les indicateurs de non-réponses et la variable d'intérêt, à travers une distribution conjointe paramétrique, et effectuer l'estimation du paramètre par la méthode du maximum de vraisemblance. Dans certains cas cette estimation se fait de manière directe par une approche analytique, dans plusieurs autres cas l'estimation s'effectue de façon itérative par des techniques d'analyse statistiques comme l'algorithme d'Espérance-Maximisation (EM) de Dempster et al (1977) [56].

Les mécanismes de type *MNAR* sont constitués de deux catégories de modèles : les modèles de mécanismes *non-ignorables connus* et les modèles de mécanismes *non-ignorables inconnus* (Little et Rubin, 1987 [14]). La différence entre ces deux types de modèles réside sur la distribution du mécanisme de données manquantes. Dans le premier cas cette distribution ne dépend d'aucun paramètre inconnu ϕ et donc est entièrement connu ; alors que dans le second cas la distribution

dépend d'un paramètre inconnu ϕ. Ce qui fait que les modèles non-ignorables connus sont plus simples à modéliser que ceux non-ignorables inconnus.

Supposons que nous avons la structure des données où la matrice des données X est de la forme

$$X = \begin{array}{c} \begin{matrix} X^1 & \dots & X^\ell & Y^1 & \dots & Y^k \end{matrix} \\ \begin{pmatrix} x_{11} & \dots & x_{1\ell} & y_{11} & \dots & y_{1k} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{n\ell} & y_{n1} & \dots & y_{nk} \end{pmatrix} \end{array} \tag{3.12}$$

où $X^1, X^2, \dots, X^\ell$ sont les variables entièrement observées et $Y = [Y^1, Y^2, \dots, Y^k]$ la sous-matrice des variables incomplètes qui sont les variables d'intérêt. Deux approches de formulation des modèles *non-ignorables* sont données par Little and Rubin (1987) [14].
Une première approche consiste à modéliser la distribution jointe des indicateurs de non-réponses R et des variables incomplètes Y, $f(Y, R \,|\, X^1, \dots, X^\ell, \theta, \phi)$ qui est factorisée sous la forme suivante

$$f(Y, R \,|\, X^1, \dots, X^\ell, \theta, \phi) = f(Y \,|\, X^1, \dots, X^\ell, \theta) f(R \,|\, Y, X^1, \dots, X^\ell, \phi) \tag{3.13}$$

où le premier facteur $f(Y \,|\, X^1, \dots, X^\ell, \theta)$ est la distribution conditionnelle des variables incomplètes Y par rapport aux covariables $X^1, \dots, X^\ell$ et au paramètre θ. Le second facteur $f(R \,|\, Y, X^1, \dots, X^\ell, \phi)$ est la distribution du mécanisme de données manquantes avec son paramètre ϕ.
Une seconde approche est de factoriser $f(Y, R \,|\, X^1, \dots, X^\ell, \theta, \phi)$ comme suit

$$f(Y, R \,|\, X^1, \dots, X^\ell, \theta, \phi) = f(Y \,|\, X^1, \dots, X^\ell, R, \theta) f(R \,|\, X^1, \dots, X^\ell, \phi) \tag{3.14}$$

La factorisation (3.13) est une façon de modélisation de la loi jointe de R et Y où cette fois-ci la distribution conditionnelle des variables incomplètes Y dépend, en plus des covariables X^j , $j = 1, \dots, \ell$ et du paramètre θ, des indicateurs R. L'introduction de ces indicateurs de non-réponse permet de pouvoir séparer le groupe des répondants à celui des non-répondants, avec une distribution pour chaque groupe. C'est notamment le cas où l'approche bayésienne prédictive est utilisée pour des estimations des paramètres et des données manquantes. Le second facteur $f(R \,|\, X^1, \dots, X^\ell,)$ n'est pas la distribution du mécanisme de non-réponse, puisqu'au conditionnement il y a les variables incomplètes Y qui manquent. Si on considère $f(R \,|\, X^1, \dots, X^\ell,)$ comme la distribution du mécanisme de non-réponse on serait dans le cas du mécanisme MAR, alors tel n'est pas le cas ici.

Mécanismes non-ignorables connus Les modèles de mécanismes *non-ignorables connus* désignent les mécanismes pour lesquels la distribution de probabilité dépend des données manquantes, éventuellement des données observées mais ne contient pas de paramètre inconnu ϕ.

$$f(R \,|\, X^{mis}, X^{obs}, \phi) = f(R \,|\, X^{mis}, X^{obs}) \tag{3.15}$$

Ici la distribution $f(R \,|\, X^{mis}, X^{obs})$ est entièrement déterminée. L'absence ici du paramètre ϕ qui conditionne généralement cette distribution du mécanisme des données manquantes marque la simplicité de ces modèles, avec seul le paramètre θ de la distribution des données $f(X \,|\, \theta)$ à estimer. Nous donnons quelques exemples, relativement simples, qui permettent d'illustrer ces modèles (Little et Rubin, 1987 [14]). La méthode d'estimation du paramètre θ utilisée est, comme nous l'avons dit, le maximum de vraisemblance. Cependant la critique principale, dans certains cas (exemple, modèles de régression linéaires multiples), de cette méthode est que l'estimation par la méthode du maximum de vraisemblance est basée sur l'hypothèse difficilement vérifiable de la normalité de erreurs pour le modèle impliquant la variable d'intérêt. Le premier exemple de modèle de mécanisme *non-ignorable connu* que nous donnons ici concerne des données univariées censurées de loi exponentielle de paramètre θ.

Exemple 3.1. *Les données* $X = (x_1, x_2, \ldots, x_n)^T$ *sont supposées indépendantes et identiquement distribuées suivant la loi exponentielle de paramètre* θ *:*

$$f(X \,|\, \theta) = \theta^{-n} \exp\left(- \sum_{i=1}^{n} \frac{x_i}{\theta} \right).$$

On suppose que l'enquête est réalisée sur une population de n individus et sur une seule variable d'enquête. Les données comportent des valeurs manquantes avec la disposition suivante : les m premières valeurs sont observées, $X^{obs} = (x_1, \ldots, x_m)^T$ *et les valeurs restantes sont manquantes* $X^{mis} = (x_{m+1}, \ldots, x_n)^T$*. Comme il s'agit d'un* modèle non-ignorable connu *et que les données sont censurées, le point de censure c est connu de tel sorte que toutes les données supérieures ou égales à cette valeur sont manquantes. La distribution du mécanisme de données manquantes est entièrement connue est donnée par*

$$f(R \,|\, X, \phi) = \prod_{i=1}^{n} f(R_i \,|\, x_i, \phi) = \prod_{i=1}^{n} f(R_i \,|\, x_i)$$

où

$$f(R_i \,|\, x_i) = \begin{cases} 1 & \textit{si } R_i = 1 \textit{ et } x_i < c, \textit{ ou bien } R_i = 0 \textit{ et } x_i > c \\ 0 & \textit{sinon.} \end{cases}$$

Dans cet exemple l'inférence sur θ, consiste d'abord à considérer la distribution jointe des données observées et des indicateurs de non-réponses. La technique, elle, consiste à factoriser cette distribution jointe peut alors se décomposer en deux blocs de facteurs

$$\begin{aligned} f(X^{obs}, R \,|\, \theta) &= \prod_{i=1}^{m} f(x_i, R_i \,|\, \theta) \prod_{i=m+1}^{n} f(R_i \,|\, \theta) \\ &= \prod_{i=1}^{m} f(x_i \,|\, \theta) f(R_i \,|\, x_i) \prod_{i=m+1}^{n} \mathbb{P}(x_i > c \,|\, \theta) \\ &= \theta^{-m} \exp\left(- \sum_{i=1}^{m} \frac{x_i}{\theta} \right) \exp\left(- \frac{(n-m)c}{\theta} \right) \end{aligned} \tag{3.16}$$

où $P(x_i > c \,|\, \theta) = \exp(-\frac{c}{\theta})$. Étant donné que nous sommes dans le cas d'un mécanisme de données manquantes non-ignorable nous utilisons la vraisemblance *correcte* du paramètre θ conditionnellement aux données observées X^{obs}

$$L(\theta \,|\, X^{obs}, R) \propto f(X^{obs}, R \,|\, \theta)$$

La maximisation de cette vraisemblance par le biais de l'équation (3.16) permet d'obtenir un estimateur de θ,

$$\hat{\theta} = \frac{1}{m} \left(\sum_{i=1}^{m} x_i \, + \, (n-m)c \right).$$

Le second exemple illustre de la modélisation du mécanisme *non-ignorable connu* dans un travail d'inférence sur un paramètre.

Exemple 3.2. *Le modèle consiste à un échantillonnage de n données univariées suivant la loi exponentielle de paramètre* θ*, soit* $X = (x_1, x_2, \ldots, x_n)^T$*. L'échantillon comporte également des données manquantes (censurées) ; mais à la différence du précédent exemple les données censurées sont regroupées en catégories. Ainsi si les données observées* $X^{obs} = (x_1, \ldots, x_m)^T$ *sont constituées de m valeurs, les* $n - m$ *données restantes* $X^{mis} = (x_{m+1}, \ldots, x_n)^T$ *sont manquantes et sont alors regroupées en J catégories de sorte que pour chaque entier* $j = 1, \ldots, J$*, la j ième catégorie contient* r_j *valeurs de* X^{mis} *comprises entre deux valeur fixées* a_j *et* b_j*, avec* $a_j > 0$ *et*

$b_j > 0$.
Le vecteur des indicateurs de non-réponses $R = (R_1, R_2, \ldots, R_n)^T$ *est alors ainsi redéfini*

$$R_i = \begin{cases} 1 & , \textit{ si } 1 \leq i \leq m \\ j+1 & , \textit{ si } m < i \leq n \textit{ et } x_i \textit{ tombe dans la } j \textit{ ième catégorie.} \end{cases}$$

Dans cet exemple, l'estimation du paramètre θ se fait par la méthode du maximum de vraisemblance par le biais de l'algorithme Espérance-Maximisation (EM). Rappelons que celui est une procédure souvent utilisée lorsque l'on effectue une inférence sur le(s) paramètre(s) du modèle et que les données comportent des valeurs qui manquent. C'est une méthode qui consiste de manière itérative à maximiser vraisemblance (des données observées) $L(\theta \,|\, X^{obs})$ ou plus précisément la log-vraisemblance des données $\log(L(\theta \,|\, X^{obs}))$ qui se décompose en deux quantités si son espérance est calculée par rapport à la loi des données manquantes X^{mis} conditionnellement aux données observées X^{obs}, aux indicateurs de non-réponses R et au paramètre θ, $f(X^{mis} \,|\, X^{obs}, R, \theta)$

$$\log(L(\theta \,|\, X^{obs})) = Q(\theta; \theta^{(t)}) - H(\theta; \theta^{(t)})$$

Le premier terme $Q(\theta\,;\,\theta^{(t)})$ est celui qui requiert notre intérêt pour l'étape E (calcul de l'Espérance) car la suite $(\theta^{(t)})$ telle que $\theta^{(t+1)} = \arg\max Q(\theta\,;\,\theta^{(t)})$, vérifie aussi $Q(\theta\,;\,\theta^{(t+1)}) \geq Q(\theta\,;\,\theta^{(t)})$ et converge vers l'estimateur $\hat{\theta}$ de θ, qui maximise la log-vraisemblance $\log(L(\theta \,|\, X^{obs}))$ ($\hat{\theta} = \arg\max \log(L(\theta \,|\, X^{obs}))$). La quantité $Q(\theta\,;\,\theta^{(t)})$ désigne, en fait, l'espérance de la log-vraisemblance complétée $\log(L(\theta \,|\, X^{obs}, X^{mis}, R))$ par rapport à $f(X^{mis} \,|\, X^{obs}, R, \theta)$,

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \int \log L(\theta \,|\, X^{obs}, X^{mis}, R) f(X^{mis} \,|\, X^{obs}, R, \theta^{(t)}) \mathrm{d}X^{mis} \\ &= \mathbb{E}\left(\log L(\theta \,|\, X^{obs}, X^{mis}, R) \,|\, X^{obs}, R, \theta^{(t)}\right) \end{aligned}$$

L'expression de $\log L(\theta \,|\, X^{obs}, X^{mis}, R)$ est donnée par

$$\log L(\theta \,|\, X^{obs}, X^{mis}, R) = \log f(X \,|\, \theta) + \log f(R \,|\, X) \tag{3.17}$$

La loi du mécanisme de données manquantes $f(R \,|\, X)$ est ici aussi entièrement connue. Par rapport à la vraisemblance complétée qui est une fonction du paramètre θ, $f(R \,|\, X)$ est peut être regardée comme une constante. Ce qui fait que dans (3.17) seul le premier terme nous intéressera pour le calcul de l'espérance (étape E de la méthode EM). Son expression est donné par

$$\begin{aligned} \log f(X \,|\, \theta) &= \log\left(\theta^{-n} \exp\Big(-\sum_{i=1}^{n} \frac{x_i}{\theta}\Big)\right) \\ &= -n \log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i \end{aligned}$$

d'où son espérance est donnée par

$$\mathbb{E}\left(\log f(X \,|\, \theta) \,|\, X^{obs}, R, \theta^{(t)}\right) = -n\log\theta - \frac{1}{\theta}\mathbb{E}\left(\sum_{i=1}^{n} x_i \,|\, X^{obs}, R, \theta^{(t)}\right),$$

où

$$\mathbb{E}\left(\sum_{i=1}^{n} x_i \,|\, X^{obs}, R, \theta^{(t)}\right) = \sum_{i=1}^{m} x_i + \sum_{j=1}^{J} r_j \hat{x}_j^{(t)}$$

avec les valeurs prédites $\hat{x_j}^{(t)}$, à l'itération t, données par

$$\begin{aligned} \hat{x}_j^{(t)} &= \mathbb{E}\left(x \,|\, a_j \leq x < b_j\,;\,\theta^{(t)}\right) \\ &= \int_{a_j}^{b_j} x \frac{1}{\theta^{(t)}} \exp\left(-\frac{x}{\theta^{(t)}}\right) \mathrm{d}x \end{aligned}$$

Par la technique de l'intégration par parties on a

$$\hat{x}_j^{(t)} = \theta^{(t)} + \frac{b_j e^{-b_j/\theta^{(t)}} - a_j e^{-a_j/\theta^{(t)}}}{e^{-b_j/\theta^{(t)}} - e^{-a_j/\theta^{(t)}}}$$

L'étape maximisation (M) permet d'obtenir la mise à jour du paramètre :

$$\theta^{(t+1)} = \frac{1}{n}\left(\sum_{i=1}^{m} x_i + \sum_{j=1}^{J} r_j \hat{x}_j^{(t)}\right)$$

Si $b_j = \infty$, la valeur prédictive des données censurées est donnée par $\hat{x}_j^{(t)} = \theta^{(t)} + a_j$. La valeur de $\theta^{(t+1)}$ se réécrit alors comme suit

$$\theta^{(t+1)} = \frac{1}{n}\left(\sum_{i=1}^{m} x_i + \sum_{j=1}^{J} r_j(\theta^{(t)} + a_j)\right)$$

La convergence de la suite $(\theta^{(t)})_t$ vers sa limite, l'estimateur $\hat{\theta}$, donne à partir d'un rang t_0, $\theta^{(t)} \approx \theta^{(t+1)} \approx \hat{\theta}\,, \forall\, t \geq t_0$, et donc on a

$$\hat{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m} x_i + \sum_{j=1}^{J} r_j a_j\right)$$

En particulier, si $a_j = c\,, \forall\, j$, toutes les observations ont le même point de censure, et alors

$$\hat{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m} x_i + (n-m)c\right)$$

qui est le même estimateur du paramètre θ que dans l'exemple précédant.

Mécanismes non-ignorables inconnus. Les modèles de mécanismes *non-ignorables inconnus* sont ceux pour lesquels la distribution de probabilité des indicateurs de non-réponses dépend des données manquantes X^{mis}, d'un paramètre inconnu ϕ et éventuellement des données observées X^{obs} :

$$f(R \,|\, X, \phi) = f(R \,|\, X^{mis}, X^{obs}, \phi) \tag{3.18}$$

Les modèles de censures stochastiques (extensions du modèle Tobit) peuvent être considérés comme des modèles à mécanismes non-ignorables inconnus. La modélisation consiste alors à une régression linéaire de deux variables d'intérêt Y^1 et Y^2, de taille $n \times 1$, par rapport à des covariables complètes $X^1, \ldots, X^\ell$. La variable Y^1 est incomplètement observée alors que la variable Y^2 est complètement inobservée. Les valeurs observées de Y^1 vérifient la propriété suivante : pour un individu i, $i = 1, \ldots, n$, la valeur y_{i1} est observée si et seulement si la valeur y_{i2} de Y^2 dépasse un seuil donné (par exemple 0). Un exemple illustratif est le modèle bivarié normal de censure stochastique.

Exemple 3.3. *La description du modèle est la suivante. Soient Y^1 et Y^2 deux variables respectivement incomplètement observée et complètement inobservée, soient $X^1, \ldots, X^\ell$ les ℓ covariables complètes du modèle. La distribution normale est supposée sur les variables à données manquantes $Y = [Y^1, Y^2]$. Alors la distribution paramétrique de Y conditionnellement aux covariables, $f(Y \,|\, X^1, \ldots, X^\ell, \theta)$ est donnée par*

$$f(Y \,|\, X^1, \ldots, X^\ell, \theta) = \prod_{i=1}^{n} f(Y_i \,|\, X^1, \ldots, X^\ell, \theta) \tag{3.19}$$

avec

$$f(Y_i \,|\, X^1, \ldots, X^\ell, \theta) = \mathcal{N}\,(y_i\,;\, \mu_i\,,\, \Sigma) \tag{3.20}$$

où Y_i est la ligne i de la matrice Y et $\mathcal{N}(y_i\,;\,\mu_i\,,\,\Sigma)$ est une loi normale bivariée, avec

$$y_i = (y_{i1}, y_{i2})\,,\ \mu_i = (x_i\beta_1, x_i\beta_1) \text{ et } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}$$

où $x_i = (x_{i0}, x_{i1}, \ldots, x_{i\ell})$ est constitué du terme constant $x_{i0} = 1$ et des prédicteurs $x_{i1}, \ldots, x_{i\ell}$ de l'individu i, $\beta_1 = (\beta_{01}, \ldots, \beta_{\ell 1})^T$ et $\beta_2 = (\beta_{02}, \beta_{12}, \ldots, \beta_{\ell 2})^T$ sont les vecteurs des coefficients de régression. La matrice de covariances est Σ où paramètres réels σ_1 et ρ sont des paramètres. La loi conditionnelle des indicateurs de non-réponses est $f(R\,|\,Y, X^1, \ldots, X^\ell, \phi)$ qui est spécifiée par une distribution dégénérée avec donc les coefficients de R qui sont donnés par :

$$R_{i1} = \begin{cases} 1\,, & si\ y_{i2} > 0 \\ 0\,, & si\ y_{i2} \leq 0 \end{cases} \qquad et \quad R_{i2} \equiv 0. \tag{3.21}$$

où $R_{ij} = 1$ si y_{ij} est observée et $R_{ij} = 0$ lorsque y_{ij} est manquante.

Dans cet exemple l'estimation des paramètres a nécessité une autre version de la factorisation (3.13), qui est intéressante pour ce modèle. Alors la distribution jointe des indicateurs de non-réponses et variables incomplètes $f(R, Y^1, Y^2\,|\,X^1, \ldots, X^p, \theta, \phi)$ par rapport à la variable totalement manquante Y^2, pour obtenir

$$f(R, Y^1\,|\,X^1, \ldots, X^p, \theta, \phi) = f(Y^1\,|\,X^1, \ldots, X^p, \theta) f(R\,|\,Y^1, X^1, \ldots, X^p, \phi)$$

Le modèle est alors simplifié avec R qui désigne uniquement maintenant le vecteur R^1. Les paramètres du modèle sont alors $\theta = (\beta_1, \sigma_1^2)$ et $\phi = (\beta_1, \beta_2, \rho)$. Des équations (3.20) et (3.21) on obtient la distribution $f(R\,|\,Y^1, X^1, \ldots, X^p, \phi)$ qui est spécifiée, pour chaque individu i, par une une loi de Bernoulli $\mathcal{B}(p_i)$ de paramètre p_i qui est la probabilité conditionnelle de réponse pour i et qui est donnée par

$$\begin{aligned} p_i &= P(R_i = 1\,|\,y_{i1}, \tilde{x}_i) \\ &= P(y_{i2} > 0\,|\,y_{i1}, \tilde{x}_i) \\ &= \Phi\left(\frac{\tilde{x}_i\beta_2 + \rho\sigma_1^{-1}(y_{i1} - \tilde{x}_i\beta_1)}{\sqrt{1-\rho^2}}\right) \end{aligned}$$

L'algorithme EM est égale utilisé pour l'estimation des paramètres θ et ϕ. Dans le cas où il n'y pas de contraintes placées sur les coefficients β_1, l'étape (E) consiste au calcul des espérances suivantes, déterminées par le biais des propriétés de la loi normale bivariée,

$$\begin{aligned} \mathbb{E}(y_{i2}\,|\,y_{i2} \leq 0) &= \mu_{i2} - \lambda(-\mu_{i2}) \\ \mathbb{E}(y_{i2}\,|\,y_{i2} > 0) &= \mu_{i2} + \lambda(\mu_{i2}) \\ \mathbb{E}(y_{i1}\,|\,y_{i2} \leq 0) &= \mu_{i1} - \rho\sigma_1\lambda(-\mu_{i2}) \\ \mathbb{E}(y_{i2}^2\,|\,y_{i2} \leq 0) &= 1 + \mu_{i2}^2 - \mu_{i2}\lambda(-\mu_{i2}) \\ \mathbb{E}(y_{i2}^2\,|\,y_{i2} > 0) &= 1 + \mu_{i2}^2 + \mu_{i2}\lambda(\mu_{i2}) \\ \mathbb{E}(y_{i1}^2\,|\,y_{i2} \leq 0) &= \mu_{i1}^2 + \sigma_1^2 - \rho\sigma_1\lambda(-\mu_{i2})(2\mu_{i1} - \rho\sigma_1\mu_{i2}) \\ \mathbb{E}(y_{i1}y_{i2}\,|\,y_{i2} \leq 0) &= \mu_{i1}(\mu_{i2} - \lambda(-\mu_{i2})) + \rho\sigma_1. \end{aligned}$$

Dans ces expressions, la fonction $\lambda(.)$ est définie par le rapport $\lambda(x) = \varphi(x)/\Phi(x)$ qui est l'inverse du ratio de Mills, le conditionnement par rapport aux données $\tilde{x}_i$ et aux paramètres est implicites, les expressions de μ_{i1} et μ_{i2} sont respectivement $\tilde{x}_i\beta_1$ et $tildex_i\beta_2$, et $y_{i2} \leq 0$ s'applique aux cas où Y^1 et Y^2 sont manquantes, $y_{i2} > 0$ aux cas où Y^1 est observé et Y^2 manquante. Les valeurs manquantes pour chaque individu i, y_{i1}, y_{i2}, $y_{i1}y_{i2}$ et y_{i1}^2 sont remplacées par leurs espérances respectivement conditionnellement aux paramètres et aux données observées.
L'étape (M) consiste à la maximisation de l'espérance $Q(\theta, \phi\,;\,\theta^{(t)}, \phi^{(t)})$ et se résume en ces trois sous-étapes suivantes :

(1) Régression de Y^2 sur $\tilde{X}$, avec estimation des coefficients $\hat{\beta}_2$.
(2) Régression de Y^1 sur Y^2 et $\tilde{X}$, avec les coefficients estimés $\hat{\delta}$ pour Y^2 et $\hat{\beta}_1^*$ pour $\tilde{X}$, et avec la variance résiduelle $\hat{\sigma}^2_{1\cdot 2}$.
(3) Mettre à jour les paramètres restant : $\hat{\beta}_1 = \hat{\beta}_1^* + \hat{\delta}\hat{\beta}_2,\ \hat{\sigma}_1^2 = \hat{\sigma}^2_{1\cdot 2} + \hat{\delta}^2$ et $\hat{\rho} = \hat{\delta}/\hat{\sigma}_1$.

Cette première section a permis de comprendre la problématique des données manquantes, particulièrement dans les questionnaires d'enquêtes. Ainsi la connaissance au mieux de ce type de données, en ce qui concerne leur typologie (structures des données manquantes et mécanismes sous-jacents), est très importante pour le choix ou la construction de méthodes d'analyses appropriées.

3.2 État de l'art sur la problématique.

La problématique des données manquantes a intéressé beaucoup de travaux de recherche en analyses statistiques multivariées. A ce titre plusieurs méthodes de traitement des données manquantes ont été proposées dans la littérature. Deux catégories de méthodes ont été ainsi proposées : les *analyses sans imputation* et les *méthodes d'imputation* (encore appelées *méthodes de complétion*). Les analyses sans imputation sont l'ensemble procédures d'analyses effectuées en présence de données manquantes. Nous pouvons diviser celles-ci en sous-classes. La première étant constituée des deux méthodes avec suppression que sont les *études de cas complets* et les *études de cas disponibles* qui sont considérées comme des procédures ad-hoc et rapides (Little et Rubin, 1987 [14]). Leur point commun est le retrait éventuel de certaines données observées du processus d'analyse statistique (exemple calcul de moyennes et de variances), d'où leur appellation de méthodes avec suppression. La seconde, quant à elle, est constituée de l'ensemble des stratégies d'inférence en présence de données manquantes, plus structurées, basées sur des modèles statistiques paramétriques et dont l'outil essentiel est la fonction de vraisemblance. En effet, diverses études ont été menées par divers auteurs. Ainsi Anderson a proposé un modèle bivarié gaussien pour l'estimation de caractéristique telles les moyennes, les variance ou encore les corrélations en présence de données manquantes, à l'aide de la méthode du maximum de vraisemblance (Anderson, 1957 [3]). Les auteurs Tranwinski et Bargmann ont proposé une méthode d'estimation de paramètres, en présence de données manquantes par la méthode du maximum de vraisemblance (Trawinski et Bargmann, 1964 [5]). Hartley et Hocking ont développé des méthodes plus générales donc adaptées à une structure quelconque et basée sur des modèles probabilistes avec utilisation d'une vraisemblance avec des techniques d'estimation s'appuyant sur le principe du maximum de vraisemblance (Hartley et Hocking, 1971 [7]). Orchard et Woodbury ont également proposé des procédures d'imputation basées sur l'usage de la vraisemblance des données (Orchard et Woodbury, 1972 [8]). Cependant pour ces méthodes qui ne complètent pas les données manquantes, nous présenterons seulement, ici, la première sous-classe qui est celle des analyses avec suppression de données, puisque l'accent sera principalement mis dans cette section, sur les procédures d'imputation.

Ces méthodes de complétion sont constituées de l'ensemble des procédés qui consistent à remplacer les données manquantes par des valeurs d'estimation, à partir d'une modélisation bien structurée (déterministe ou probabiliste). Ainsi plusieurs auteurs se sont intéressés à la résolution du problème de données manquante par le biais de l'imputation. En effet, Rubin (1987) [11] a introduit la méthode de l'imputation multiple, proposée comme alternative à l'imputation simple (exemple l'imputation par la moyenne), afin de capter les différentes variabilités possibles pour l'estimation des valeurs manquantes. Puis la méthode a été intensivement revisité dans la littérature notamment dans le domaine médical comme dans Van Buurren (2007) [82], Van Ginkel et al (2007) [83], Shrive et al (2006) [81], Taylor et al [79], White et al (2011) [84], etc. L'imputation multiple est adaptée de manière générale à tous les types de problèmes de données manquantes (structure univariée, monotone et arbitraire), et s'appuie sur une vraisemblance (inférence direct ou bayésienne).

Toutefois nous nous intéressons ici à quelques unes des principales méthodes d'imputation

récemment développées. L'intérêt porté sur ces méthodes réside d'une part sur leur diversité de part les différents modèles sous-jacents, d'autre part sur des points de similitudes d'approches qu'elles ont avec notre algorithme d'imputation *PGNMF*. En effet, d'abord le point commun que notre algorithme partage avec toutes ces méthodes est l'approche factorielle, ensuite le second point commun est l'approche d'analyse statistique qu'il partage avec certaines, enfin la troisième caractéristique commune est la factorisation NMF qu'il partage avec d'autres. Ainsi les procédures d'approches statistiques sont la méthode *PEM* (*Poisson Espérance-Maximisation*), ainsi nommé parce qu'elle s'appuie sur notre modèle poissonien et est donc une spécification de l'algorithme EM de Dempster et al (1977) [56], la méthode *MISSF* (*MissForest*) de Stekhoven and Bühlmann (2011) [41] basée sur la technique des *forêts aléatoires* de Breiman (2001) [40]) qui consiste en une régression linéaire, et la méthode *MIMCA* (*Multiple Imputation with Multiple Correspondence Analysis*) de Audigier et al (2017) [43] qui est une procédure d'imputation de données catégorielles. Quant aux méthodes de factorisation NMF, on l'algorithme *PEM* précédemment cité et l'algorithme *WNMF* (*Weighted Nonnegative Matrix Factorization*) Kim and Choi (2009) [22] qui est une méthode de factorisation pondérée déjà étudiée au chapitre précédent (Chap. 2, Sect. 2.2.2).
Nous achèverons cette section de l'état de l'art, par la présentation des principales méthodes dites naïves (ou rudimentaires). Il s'agit des stratégies de complétion par la moyenne, la médiane, le mode et aussi la méthode *LOCF* (*Last Observation Carry Forward*). Cette dernière consiste à compléter chaque valeur manquante par la dernière valeur observée.

3.2.1 Les analyses sans imputation

Comme nous l'avons déjà dit, les analyses sans imputation que nous présentons ici sont celles dites avec suppression de données : les *études de cas complets* et les *études de cas disponibles*. Rappelons que la qualification de méthodes de suppression réside dans le fait qu'elles consistent à retirer de l'analyse statistique des données observées d'individus pour lesquels des données données manquantes sont enregistrées. Notons qu'en général, l'utilisation de l'une ou l'autre de ces techniques repose sur l'hypothèse que les données manquantes sont du mécanisme MCAR. En d'autres termes, faire appel à l'étude de cas complets ou l'étude de cas disponibles suppose que la probabilité de manquer de données sur sa variable dépendante n'est pas liée à d'autres variables indépendantes ni à la variable dépendante elle-même. Ainsi ce sont des méthodes considérées comme des procédures rapides de traitement de données manquantes, présentées dans Little et Rubin (1987) [14] et que nous donnons plus en détails comme suit :

Étude de cas complets (listwise deletion). C'est une méthode de traitement de données manquantes qui supprime un enregistrement entier dès qu'il y a une valeur qui manque. Un enregistrement est ici l'ensemble des données recueillies sur un individus. En effet si nous considérons la matrice de données $X = \left[X^1, X^2, \ldots, X^p\right] = [X_1, X_2, \ldots, X_n]^T$ de taille $n \times p$, où n est le nombre d'individus avec $X_i\,,\ i = 1, \ldots, n$ le i ième individu et p le nombre de variables avec $X^j\,,\ j = 1, \ldots, p$ la j ième variable, alors la procédure consiste à retirer tout individu $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ possédant au moins une valeur qui manque. C'est le cas par exemple, lors d'un questionnaire et que l'individu X_i n'a pas répondu à un ou plusieurs questions. Une conséquence immédiate de cette méthode est la perte de beaucoup de données car pour chaque individu présentant des valeurs qui manquent toutes les autres données observées le concernant sont également perdues par le fait de la suppression. Cette perte massive de données affecte la puissance statistique des analyses effectués. En effet le pouvoir statistique dépend en partie de la grande taille de l'échantillon. Étant donné que la méthode exclut toutes les données observées, pour un individu i dont certaines valeurs sont manquantes, elle réduit l'échantillon analysé statistiquement. C'est une méthode qui impose ainsi de ne travailler qu'avec une matrice où toutes les données sont entièrement observées. Une illustration est l'exemple où on veut modéliser, pour un échantillon de 10 individus, le revenu en fonction de l'âge et du genre. L'étude de cas complet consiste d'abord à retirer de l'analyse les sujets 3, 4 et 8 qui ont enregistré des données manquantes (NA), puis à effectuer l'analyse statistique souhaitée sur un tableau complet de données.

Sujet	Age	Genre	Revenu
1	29	M	\$40 000
2	45	M	\$36 000
3	81	M	NA
4	22	NA	\$16 000
5	41	M	\$98 000
6	33	F	\$60 000
7	22	F	\$24 000
8	NA	F	\$81 000
9	33	F	\$55 000
10	45	F	\$80 000

TABLE 3.4: Revenu en fonction de l'âge et du genre.

Outre la faiblesse statistique de l'analyse posée par la suppression massive de données, un autre problème lié à l'usage de l'*étude de cas complets* est la restriction de celle-ci au seul mécanisme MCAR, qui pourtant dans beaucoup de cas pratiques n'est pas réaliste. En effet la supposition des mécanismes MAR ou MNAR favorise l'introduction de biais dans l'analyse. Pour le cas du mécanisme MNAR qui peut s'observer dans les questionnaires lorsque des non-réponses sur certains points sont liées à la nature même de ceux-ci (questions à caractère sensible), le retrait des individus non-répondants à ces points a de fortes chances d'introduire un biais dans les analyses statistiques qui s'ensuivent puisque ces individus peuvent être un groupe qui présente des caractéristiques différentes du reste de l'échantillon enquêté. Une illustration de ce cas de figure est l'exemple d'un questionnaire qui inclus des questions sensibles sur l'usage de drogue, le revenu ou la sexualité des individus enquêtés. Alors beaucoup de sujets dans l'échantillon peuvent ne pas répondre en raison de la nature intrusive des questions, mais peuvent répondre à tous les autres points du questionnaire. Cette méthode de traitement exclura ces répondants de l'analyse. Cela peut créer un biais car les participants qui divulguent ces informations peuvent présenter des caractéristiques différentes des participants qui n'en refusent de répondre. Ainsi cette mise à l'écart de groupes d'individus ayant des caractéristiques différentes d'un groupe à l'autre réduit fortement la variabilité des informations recueillies, ce qui introduit le biais dans les résultats.

Études de cas disponibles (pairewise deletion). Le second type de méthodes d'études sans complétion est constitué des *études de cas disponibles*. Il s'agit de procédures qui limitent la suppression massive de données provoquée par le type d'analyse précédent, car elles prennent en comptent toutes données observées. En effet, pour ce type d'études, chaque variable X^j est considérée par rapport à toutes ses valeurs observées. Alors le calcul de caractéristiques comme la moyenne ou la variance se fait, pour chaque variable X^j , $j = 1, \ldots, p$ en considérant toutes ses valeurs observées tout en ignorant les valeurs qui manquent. Les études de cas disponibles présentent cependant un inconvénient qui est la variabilité les échantillons impliqués dans le calcul des caractéristiques. Cette variabilité dépend de la structure des données manquantes (univariée, monotone ou arbitraire). En effet si nous considérons la matrice de données $X = \left[X^1, X^2, \ldots, X^p\right]$ où l'on s'intéresse au calcul de la moyenne pour les différentes variables X^j , $j = 1, \ldots, p$, un individu peut être impliqué dans l'échantillon servant au calcul de la moyenne pour une variable X^j sans pour autant l'être pour une autre variable X^k. Ce qui entraine la non stabilité ou de manière équivalent la variabilité des échantillons de calculs. En d'autres termes les moyennes ne sont pas calculées sur la base d'un échantillon fixe (nombre, n, d'individus). Cela peut alors poser un problème de compatibilité des analyses univariées, surtout lorsque différents groupes existent

à l'intérieur de l'échantillon d'origine de taille n. Les échantillons de données observées pour chaque variable sont, en fait, des sous-échantillons du n-échantillon sur lequel l'étude (enquête, sondage) est réalisée au départ.

Si pour les analyses univariées précédentes le problème de variabilité des échantillons s'est posé, les analyses multivariées elles en plus du même problème se trouvent confrontées à une autre difficulté qui, d'ailleurs est une conséquence du premier problème. Il s'agit de la différence de taille des sous-échantillons de valeurs observées dans les différentes variables. Par exemple, pour des analyses bivariées où les caractéristiques calculées sont les covariances et les coefficients de corrélations, l'on a besoin de fixer une même taille d'échantillon pour la paire de variables considérées. Alors les études de cas disponibles s'adaptent pour trouver un échantillon commun à chacune des paires de variables concernées : on parle alors d'*études de cas disponibles par paires* (*pairwise available-case methods* ou *suppression par paire de variables* (*pairewise deletion*). Dans ce cas il y a bien suppression de données, mais dans une proportion moindre que celle des études de cas complets. En effet les covariations (covariances et corrélations) entre deux variables X^j et X^k sont alors calculées en incluant dans l'échantillon commun que les individus i, $i = 1, \ldots, n$, pour lesquels les valeurs x_{ij} et x_{ik} sont à la fois observées. Les valeurs manquantes, pour la paire de variables, sont ignorées ainsi que les valeurs observées seules. Ces dernières correspondent au cas où pour les deux pour variables X^j et X^k et pour un individu donné, i la valeur x_{ij} est observée alors que x_{ik} est manquante, et vice versa. L'échantillon commun à la paire est donc l'ensemble des individus i, pour lesquels les valeurs x_{ij} et x_{ik} sont à la fois observées. La taille de cet échantillon est alors $n^{(jk)}$. En particulier les covariances sont données par (en gardant le notations de Little and Rubin (1987) [14])

$$S_{jk}^{(jk)} = \frac{1}{n^{(jk)} - 1} \sum_{(jk)} \left(x_{ij} - \bar{x}_j^{(jk)} \right) \left(x_{ik} - \bar{x}_k^{(jk)} \right) \tag{3.22}$$

avec $n^{(jk)}$ est le nombre des cas où les variables $X^{(j)}$ et $X^{(k)}$ sont à la fois observées et les moyennes $\bar{x}_j^{(jk)}$ et $\bar{x}_k^{(jk)}$ calculées sur l'effectif $n^{(jk)}$ de l'échantillon commun à la paire de variables $X^{(j)}$ et $X^{(k)}$. Une première idée pourrait être de calculer les variances sur $S_{jj}^{(j)}$ et $S_{kk}^{(k)}$ sur les effectifs respectifs $n^{(j)}$ et $n^{(k)}$ où les toutes les valeurs sont disponibles pour chacune des deux variables pour obtenir les coefficients de corrélation suivants :

$$r_{jk}^{*} = \frac{S_{jk}^{(jk)}}{\sqrt{S_{jj}^{(j)} S_{kk}^{(k)}}} \tag{3.23}$$

Un inconvénient de ce procédé est que les valeurs r_{jk}^{*} ne sont plus comprises entre -1 et 1. Pour corriger cela les variances $S_{jj}^{(j)}$ et $S_{kk}^{(k)}$ dans l'équation (3.23) peuvent alors être remplacées par les variances $S_{jj}^{(j)}$ et $S_{kk}^{(k)}$ estimées à partir de la taille $n^{(jk)}$ de l'échantillon commun aux deux variables. Ce qui donne les nouveaux coefficients de corrélation suivants

$$r_{jk}^{(jk)} = \frac{S_{jk}^{(jk)}}{\sqrt{S_{jj}^{(jk)} S_{kk}^{(jk)}}} \tag{3.24}$$

Les nouvelles estimations des covariances sont alors données par

$$S_{jk}^{*} = r_{jk}^{(jk)} \sqrt{S_{jj}^{(j)} S_{kk}^{(k)}} \tag{3.25}$$

Les *études de cas disponibles* qui analysent les données par paires de variables (*pairwise available-case*) comme les estimations des matrices covariances et coefficients de corrélations (équations (3.22) à (3.25)) tentent ainsi de récupérer au mieux que possible les informations perdues avec les *études de cas complets*.

Cependant elles possèdent aussi des limites dans beaucoup de cas pratiques, notamment avec les estimations r_{jk}^{*} dans (3.23) qui ne sont pas toujours bien définies avec des valeurs pouvant sortir de l'intervalle $[-1\,;\,1]$. Par ailleurs bien que les estimations corrigées des corrélations $r_{jk}^{(jk)}$ vérifient $-1 \leq r_{jk}^{(jk)} \leq 1$ il peut y avoir des situations où la matrice des corrélations n'est pas définie positive. En effet considérons l'exemple artificiel suivant (Little et Rubin, 1987 [14]) où les données contiennent des valeurs manquantes et sont réparties en trois variables $X = \left[X^{(1)}, X^{(2)}, X^{(3)}\right]$. On a alors la matrice suivante où les tirets (-) désignent les valeurs qui manquent.

$$X^T = \begin{pmatrix} 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & - & - & - & - \\ 1 & 2 & 3 & 4 & - & - & - & - & 1 & 2 & 3 & 4 \\ - & - & - & - & 1 & 2 & 3 & 4 & 4 & 3 & 2 & 1 \end{pmatrix}$$

A partir de l'équation (3.24) nous obtenu les coefficients suivants $r_{12}^{(12)} = 1$, $r_{13}^{(13)} = 1$ et $r_{23}^{(23)} = -1$. Ces estimations sont clairement insatisfaisantes, puisque $Corr(X^{(1)}, X^{(2)}) = Corr(X^{(1)}, X^{(3)}) = 1$ implique que $Corr(X^{(2)}, X^{(3)}) = 1$ et non -1. De même les matrices de covariances calculées à partir des équations (3.22) ou (3.25) ne sont sont pas toujours définies positives. Étant donné que plusieurs analyses basées sur la matrice de covariance comme la régression multiple, nécessitent une matrice définie positive, des modifications ad-hoc s'imposent lorsque cette condition n'est pas satisfaite.

En définitive, puisque les *études de cas disponibles* impliquent beaucoup plus de données, l'on pourrait s'attendre à ce qu'elles soient toujours plus efficientes que les *études de cas complets*. En réalité tel n'est pas toujours le cas. En effet lorsque l'hypothèse MCAR est supposée sur les données et que les corrélations sont faibles, Kim et Curry (1977) [86] ont montré par le biais de simulation qu'ils ont effectuées que les études de cas disponibles peuvent donner des résultats plus satisfaisant que les études de cas complets. D'autres simulations par contre (Haitovsky, 1968 [87] ; Azen et Van Guilder, 1981 [88]) ont montré dans le cas où les corrélations sont fortes, la supériorité des études de cas complets. Ainsi pour conclure, aucune des deux procédures ne donne en général une satisfaction entière, d'où le besoin de les combiner parfois.

3.2.2 Les méthodes d'imputation

3.2.2.1 La méthode Poisson Espérance-Maximisation (PEM)

La méthode *PEM* a été développée par Cemgil (2009) [29]. Elle est une spécification de l'algorithme EM de Dempster et al (1977) [56], sous un modèle poissonien, comme celui par rapport auquel nous développons notre méthodologie. Rappelons la méthode EM de Dempster est un algorithmes qui permet l'estimation des paramètres d'un modèle statistique par le maximum de vraisemblance lorsque dépend de variables latentes non observables ou possède des données manquantes. En général, l'algorithme EM est méthode sans imputation et qui consiste à faire de l'inférence sur un paramètre d'un modèle comportant des données manquantes. Cependant dans le cas précis du modèle poissonien de factorisation NMF, sa spécification qui est l'algorithme *PEM* permet de faire de l'imputation de données par le biais du principe de la factorisation $X \approx UV$, où X désigne la matrice des données et U, V les paramètres à estimer.

Nous rappelons à présent le principe de fonctionnement de l'algorithme EM, avant de le spécifier pour la procédure PEM. Si l'on note par $p(X \,|\, \theta)$, avec $X = (X^{obs}, X^{mis})$, la loi jointe de l'ensemble des données X conditionnellement au paramètre θ, alors la log-vraisemblance complétée est donnée par $L(\theta \mid X) = \log p(X^{obs}, X^{mis} \,|\, \theta)$. L'idée avec la méthode EM est de factoriser la loi jointe données observées et données manquantes est donnée par $p(X \,|\, \theta) = p(X^{obs}, X^{mis} \,|\, \theta) = p(X^{mis} \,|\, X^{obs}, \theta) p(X^{obs} \,|\, \theta)$. Remarquons dans cette factorisation la loi prédictive des données manquantes conditionnellement aux données observées, $p(X^{mis} \,|\, X^{obs}, \theta)$. La log-vraisemblance complétée peut alors se réécrire par $L(\theta \,|\, X) = \log p(X^{mis} \,|\, X^{obs}, \theta) + \log p(X^{obs} \,|\, \theta)$, ce qui donne la décomposition cherchée de la log-vraisemblance des données observées $L(\theta \,|\, X^{obs}) =$

$\log p(X^{obs} \,|\, \theta)$,

$$L(\theta \,|\, X^{obs}) = \log p(X^{obs}, X^{mis} \,|\, \theta) - \log p(X^{mis} \,|\, X^{obs}, \theta) \tag{3.26}$$

L'algorithme *EM* étant une procédure itérative basée sur l'espérance de la log-vraisemblance complétée, par rapport à la loi des données manquantes X^{mis} conditionnellement au données observées et au paramètre courant $\theta^{(t)}$, alors (3.26) peut se réécrire par

$$L(\theta \,|\, X^{obs}) = \mathbb{E}\left(\log p(X^{obs}, X^{mis} \,|\, \theta) \,|\, X^{obs}, \theta^{(t)}\right) - \mathbb{E}\left(\log p(X^{mis} \,|\, X^{obs}, \theta) \,|\, X^{obs}, \theta^{(t)}\right) \tag{3.27}$$

Le premier terme du membre de droite dans (3.27) est l'espérance en question que nous noter par $Q(\theta\,;\,\theta^{(t)})$. Alors la suite suite construite par $\theta^{(t+1)} = \arg\max_\theta Q(\theta\,;\,\theta^{(t)})$ tend vers un maximum de la log-vraisemblance de $L(\theta \,|\, X^{obs})$ qui est une estimation $\hat{\theta}$ du paramètre θ.
Ainsi l'algorithme, comme son nom l'indique, possède deux étapes principales que sont l'évaluation de l'espérance (E) et la maximisation de la quantité $Q(\theta\,;\,\theta^{(t)})$. qui sont effectuée de façon itérative.
La description de l'algorithme *PEM*, nécessite d'abord une brève présentation du modèle poissonien. Alors les données X sont supposées indépendantes et distribuées suivant une loi de Poisson :

$$p(X \mid U, V) = \prod_{i=1}^{n}\prod_{j=1}^{p} \mathcal{P}\left(x_{ij}\,;\,\sum_{\ell=1}^{k} u_{i\ell}v_{\ell j}\right) \tag{3.28}$$

où $\mathcal{P}(y\,;\,\lambda)$ est la loi de Poisson d'une variable Y de paramètre λ. Une première spécification peut alors être faite, c'est-à-dire $\theta = (U, V)$. Des variables auxiliaires $S = (S^1, \ldots, S^\ell, \ldots, S^k)$ considérées comme des variables latentes ont été introduites dans le modèle, telles que pour chaque $n \times p$ matrice S^ℓ les coefficients sont indépendants et de loi de Poisson. On a donc

$$p(S \mid U, V) = \prod_{\ell=1}^{k} p(S^\ell \mid U, V) = \prod_{\ell=1}^{k}\prod_{i=1}^{n}\prod_{j=1}^{p} \mathcal{P}\left(s_{ij}^{\ell}\,;\,u_{i\ell}v_{\ell j}\right) \tag{3.29}$$

Alors la loi jointe des données observées et variables latentes est donnée par :

$$p(X^{obs}, S \mid U, V) = \prod_{x_{ij}\text{ observée}} p(s_{ij}^{1:k} \mid u_{i,1:k}, v_{1:k,j}) p(x_{ij} \mid s_{ij}^{1:k}) \tag{3.30}$$

où $s_{ij}^{1:k} = (s_{ij}^1, \ldots, s_{ij}^k)$, $u_{i,1:k} = (u_{i1}, \ldots, u_{ik})$ et $v_{1:k,j} = (v_{1j}, \ldots, _{kj})$.
Par analogie à l'équation (3.26), on a :

$$L(U, V \mid X^{obs}) = \log p(X^{obs}, S \mid U, V) - \log p(S \mid X^{obs}, U, V) \tag{3.31}$$

avec les variables latentes S qui remplacent les variables manquantes X^{mis}. Alors la quantité d'intérêt $Q(U, V\,;\,U^{(t)}, V^{(t)})$ est donnée par :

$$Q\left(U, V\,;\,U^{(t)}, V^{(t)}\right) = \mathbb{E}\left(\log p(X^{obs}, S \mid U, V) \mid X^{obs}, U^{(t)}, V^{(t)})\right) \tag{3.32}$$

qui l'espérance de la log-vraisemblance complétée des données observées et variables latentes calculée à partir de la distribution des variables latentes S conditionnellement aux données observées X^{obs} et paramètres courants $U^{(n)}$, $V^{(n)}$, $p(S \mid X^{obs}, U^{(t)}, V^{(t)})$. Cette distribution est obtenue à partir de la relation

$$p(S \mid X^{obs}, U, V) = \frac{p(X^{obs}, S \mid U, V)}{p(X^{obs} \mid U, V)}$$

et est décrite par une loi multinomiale :

$$p(S \mid X^{obs}, U, V) = \prod_{x_{ij}\text{ observée}} \mathcal{M}(s_{ij}^{1:k}\,;\,x_{ij}\,,\,p_{ij}^1, \ldots, p_{ij}^k)$$

où x_{ij} et p_{ij}^ℓ , $\ell = 1, \ldots, k$ sont les paramètres tels que $\sum_\ell s_{ij}^\ell = x_{ij}$ et $p_{ij}^\ell = u_{i\ell}v_{\ell j} / \sum_{\ell'} u_{i\ell'}v_{\ell' j}$ sont les probabilités de somme égale à 1. Ce qui permet d'obtenir l'algorithme suivant :

Données : Matrice des données X, comportant des données manquantes.
Résultat : Matrice imputée $\tilde{X} \approx \tilde{U}\tilde{V}$, où $\tilde{U}$ et $\tilde{V}$ sont les estimations de U et V.
Initialization : $t = 0$, $U^{(0)}$ et $V^{(0)}$;
tant que *l'algorithme n'a pas convergé* **faire**

. **(E) : Espérance**, $Q\left(U, V\,;\, U^{(t)}, V^{(t)}\right)$, à évaluer ;

. **(M) : Maximisation** de l'espérance et mise à jour du couple (U, V) par $\left(U^{(t+1)}, V^{(t+1)}\right) = \arg\max_{(U,V)} Q\left(U, V; U^{(t)}, V^{(t)}\right)$;

. $u_{i\ell}^{(t+1)} \leftarrow u_{i\ell}^{(t)} \dfrac{\sum_{j=1}^{n} \delta_{ij} x_{ij} v_{\ell j}^{(t)} \,/\, \sum_{\ell'=1}^{k} u_{i\ell'}^{(t)} v_{\ell' j}^{(t)}}{\sum_{j=1}^{n} \delta_{ij} v_{\ell j}^{(t)}}, \quad i = 1, \ldots, n\,,\ \ell = 1, \ldots, k,$

. $v_{\ell j}^{(t+1)} \leftarrow v_{\ell j}^{(t)} \dfrac{\sum_{i=1}^{p} \delta_{ij} x_{ij} u_{i\ell}^{(t)} \,/\, \sum_{\ell'=1}^{k} u_{i\ell'}^{(t)} v_{\ell' j}^{(t)}}{\sum_{i=1}^{p} \delta_{ij} u_{i\ell}^{(t)}}, \quad \ell = 1, \ldots, k\,,\ j = 1, \ldots, p.$

. $t \leftarrow t + 1$;

fin

Algorithme 10 : Poisson Espérance-Maximisation (PEM).

Dans l'algorithme 10 les valeurs $\delta_{ij}\,,\ i = 1, \ldots, n,\ j = 1, \ldots, p$, sont les coefficients d'une matrice binaire δ, avec $\delta_{ij} = 1$ si la donnée x_{ij} est observée et $\delta_{ij} = 0$ si la donnée x_{ij} est manquante.

3.2.2.2 La méthode des forêts aléatoires : MissForest

La méthode *MissForest* est une technique d'imputation de données mixtes (mélange de données continues, discrètes et qualitatives). Récemment introduite par [41], elle est proposée comme une alternative à des méthodes du même type (traitant des données mixtes) comme la très connue méthode *MICE (Multivariate Imputation by Chained Equations)* de [17].
C'est une méthode basée sur la technique de régression de [40] dite *Forêt Aléatoire*[1] *(FA)*, en vue faire des prédictions sur des données manquantes. Cependant si la méthode *FA* nécessite la présence au moins d'une variable complètement observée, *MissForest* se suffit des valeurs observées des variables.
Supposons que les données sont disposées sous forme de matrice $X = [X^1, X^2, \ldots, X^p]$ de dimension $n \times p$, c'est-à-dire des données recueillies sur n individus et concernant p variables. La méthode est alors basée sur la méthode dit des *forêts aléatoires* de [40] qui est une méthode de régression en présence de données manquantes et qui utilise pour chaque variable la fréquence des valeurs observées. Cependant si elle nécessite la présence au moins d'une variable complètement observée pour l'apprentissage de la forêt, la procédure *MissForest*,elle se suffit des valeurs observées des variables. Supposons que la matrice de données est $X = [X^1, X^2, \ldots, X^p]$ de dimension $n \times p$. Si l'on considère une variable arbitraire X^s qui contient des valeurs manquantes indexées par $\mathbf{i}_{mis}^{(s)} \subseteq \{1, \ldots, n\}$, alors les données de la matrice X peuvent être séparées en quatre parties : (1) les valeurs observées de la variable X^s, notées par $\mathbf{y}_{obs}^{(s)}$; (2) les valeurs manquantes de la variable X^s, notées par $\mathbf{y}_{mis}^{(s)}$; (3) les variables autres que X^s avec des observations indexées par $\mathbf{i}_{obs}^{(s)} = \{1, \ldots, n\} \setminus \mathbf{i}_{mis}^{(s)}$ qu'on note par $\mathbf{x}_{obs}^{(s)}$; (4) les variables autres que X^s avec des observations $\mathbf{i}_{mis}^{(s)}$ qu'on note par $\mathbf{x}_{mis}^{(s)}$.
L'algorithme consiste dans un premier temps à compléter les données manquantes par une méthode naïve, comme par exemple l'imputation par la moyenne. Ensuite choisir les variables $X^s\,,\ s = 1, \ldots, p$ par ordre croissant du nombre de valeurs manquantes. Pour chaque variable X^s, les valeurs manquantes sont imputées par la régression des *forêts aléatoires* avec la réponse $\mathbf{y}_{obs}^{(s)}$ et les prédicteurs $\mathbf{x}_{obs}^{(s)}$, ensuite données manquantes $\mathbf{y}_{mis}^{(s)}$ sont prédites par *forêts aléatoires*

1. Méthode plus connue sous sa version anglaise,*Random Forest*

d'apprentissage appliquées à $\mathbf{x}_{mis}^{(s)}$. La procédure d'imputation est répétée jusqu'à atteindre le critère d'arrêt.
Les détails de l'algorithme sont donnés dans l'Algorithme 11.

Données : La matrice X et le critère d'arrêt κ
Résultat : La matrice imputée X^{imp}
Initialisation des données manquantes (exemple par la moyenne);
$\mathbf{k} \leftarrow$ vecteur des indices des colonnes de X tirées par ordre croissant des données manquantes;
répéter
  $\mathbf{X}_{old}^{imp} \leftarrow$ matrice précédemment imputée;
  pour *s parcourant le vecteur* **k faire**
    Ajuster par **forêt aléatoire** : $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$;
    Prédire $\mathbf{y}_{mis}^{(s)}$ en utilisant $\mathbf{x}_{mis}^{(s)}$;
    $\mathbf{X}_{new}^{imp} \leftarrow$ mise à jour de la matrice imputée en utilisant les valeurs prédites $\mathbf{y}_{mis}^{(s)}$;
  fin
  mettre à jour le critère η.
jusqu'à *le critère η atteint*;

Algorithme 11 : Imputation par MissForest

3.2.2.3 La méthodologie de factorisation NMF pondérée : WNMF

Il s'agit d'une classe d'algorithmes de factorisation NMF. Ses constituants sont des variantes des algorithmes NMF classiques, appelés algorithmes de factorisation NMF pondérés et nommés *WNMF* (*Weighted Nonnegative Matrix Factorization*). Plusieurs auteurs se sont intéressés à la factorisation pondérée WNMF. En effet, ses illustres précurseurs, Mao et Saul (2004) [21] ont juste donné et utilisé l'algorithme d'estimation associé, dans un modèle de représentation et de prévision des distances dans les réseaux internet à grande échelle, où des données manquantes sont notées. L'algorithme d'estimation qu'ils ont proposé est dérivé des règles de mise à jour multiplicative, *MU* (*Multiplicative Update*) de Lee et Seung (2001) [20], associées à la fonction coût carré de la distance euclidienne. Les auteurs Zhang et al (2006) [24] ont utilisé le même algorithme *MU* pondéré, dans un modèle de filtrage collaboratif associé à un système de recommandation de produit ou service, et dont les évaluations des utilisateurs comportent des données manquantes. L'algorithme d'estimation en question a été étudié plus en détails, par Blondel et al (2008) [77], avec un formalisme analogique à celui de Lee et Seung (2001). La présentation de cette étude, dans cette thèse, est faite au Chapitre 2 (Sect. 2.2.1.2).

Cependant l'algorithme *WNMF* auquel on s'intéresse, ici, est celui développé par Kim et Choi (2009) [22], dans une modélisation également de système de recommandation où la matrice des données d'évaluation de produit(s) renseignées par des utilisateurs, comporte des données manquantes. L'application associée a été la prédiction collaborative et la tâche, l'estimation des valeurs manquantes dans les données d'évaluation afin de prédire les préférences d'un utilisateur sur un produit, notamment les films de la base Netflix dans ce cas précis. La procédure d'estimation qu'ils ont proposée s'appuie sur la classe des algorithmes des moindres carrés alternées positives *ANLS* (*Alternating Nonnegative Least Squares*) qui est développée par divers auteurs et étudiée au Chapitre 2 (Sect. 2.1.3.2). Le choix de cette méthode s'explique par le fait qu'elle est plus récente et s'est posée comme alternative aux études antérieures que celles de Mao et Saul (2004) et de Zhang et al (2006), car jugée améliorant le temps de convergence mais aussi la précision des estimations.
La méthode de factorisation pondérée *ANLS-WNMF* de Kim et Choi a déjà été étudiée au Chapitre 2 (Sect. 2.2.2) où il s'est agit d'étudier la factorisation matricielle NMF. Nous rappellerons seulement quelques aspects de l'étude faite.

Pour décrire l'algorithme donnons d'abord le problème de factorisation matricielle qui consiste, si l'on dispose d'une d'une matrice de données X de dimension $n \times p$, à trouver deux matrices

$U \in \mathbb{R}_+^{n\times k}$ et $V \in \mathbb{R}_+^{k\times p}$ telles que $X \approx UV$. Alors le problème *ANLS-WNMF* à résoudre la problème de minimisation suivant :

$$(U^*, V^*) = \arg\min_{U\geq 0, V\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W} \tag{3.33}$$

qui est équivalent, de façon plus explicite, au problème suivant :

$$(U^*, V^*) = \arg\min_{U\geq 0, V\geq 0} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}\left(x_{ij} - \sum_{\ell=1}^{k} u_{il}v_{\ell j}\right)^2 \tag{3.34}$$

où $W = (w_{ij})$ est la matrice binaire des poids qui formalise la présence des données manquantes et est ainsi définie : $w_{ij} = \begin{cases} 1 & \text{, si } x_{ij} \text{ est observée} \\ 0 & \text{sinon.} \end{cases}$ Puisque la méthode *ANLS* est une procédure de minimisation alternée par bloc de coordonnées (*block coordinate descent method*), où chacune des deux matrices U et V représente un bloc de coordonnées, alors le problème (3.33) peut être subdivisé en deux sous-problèmes :

$$U^* = \arg\min_{U\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}\,, \quad \text{et} \quad V^* = \arg\min_{V\geq 0} \frac{1}{2}\|X - UV\|^2_{F,W}. \tag{3.35}$$

Puis chacun des sous-problèmes est subdivisé en problèmes NNLS. Alors le second sous-problème est partagé en p problèmes NNLS pondérés donnés par :

$$v^1_* = \arg\min_{v^1\geq 0} \frac{1}{2}\|x^1 - Uv^1\|^2_{2,W}\,, \ \ldots, \quad v^p_* = \arg\min_{v^p\geq 0} \frac{1}{2}\|x^p - Uv^p\|^2_{2,W} \tag{3.36}$$

ou de manière équivalente par les p problèmes NNLS suivants :

$$v^1_* = \arg\min_{v^1\geq 0} \frac{1}{2}\|(D^1x^1) - (D^1U)v^1\|^2_2\,, \ \ldots, \quad v^p_* = \arg\min_{v^p\geq 0} \frac{1}{2}\|(D^px^p) - (D^pU)v^p\|^2_2 \tag{3.37}$$

où $D^j = diag[(w^j)^{1/2}]$ est une matrice diagonale de taille $n \times n$ construite à partir du j ième vecteur colonne, w^j, de la matrice W, et pour lequel la racine carrée de chacune de ses composantes est considérée.

L'autre bloc de coordonnées est la matrice U qui permet aussi la minimisation de la fonction-coût dans (3.33), lorsque V est fixé. Ce qui correspond au premier sous problème de (3.35). De la même manière que pour le second, les problèmes NNLS dérivant de la décomposition du premier sous-problème sont donnés par :

$$\begin{aligned} u^{T(1)}_* &= \arg\min_{u^{T(1)}\geq 0} \frac{1}{2}\|(D_1x^{T(1)}) - (D_1V^T)u^{T(1)}\|^2_2\,, \\ &\ldots \\ u^{T(n)}_* &= \arg\min_{u^{T(n)}\geq 0} \frac{1}{2}\|(D_nx^{T(n)}) - (D_nV^T)u^{T(n)}\|^2_2 \end{aligned} \tag{3.38}$$

où $D_i = diag[(w_i)^{1/2}]$ est une matrice diagonale de taille $p \times p$ construite à partir du i ième vecteur ligne, w_i, de la matrice W, et pour lequel la racine carrée de chacune de ses composantes est considérée. Alors résoudre résoudre le problème de minimisation (3.36) équivaut à résoudre le problème (3.37) ou de manière équivalente les n problèmes NNLS de (3.38). Ce qui donne l'algorithme *ANLS-WNMF* suivant :

Données : Fournir la matrice des données positives X, avec données manquantes.
Résultat : Matrice imputée $\tilde{X} \approx \tilde{U}\tilde{V}$, où $\tilde{U}$ et $\tilde{V}$ sont les estimations de U et V.
. Initialisation : $U^{(0)}$, $V^{(0)}$: initialiser les matrices par des coefficients positifs;
pour t *allant de* 1 *à Maxiter* **faire**
  . Mettre à jour la matrice V colonne par colonne
  pour j *allant de* 1 *à* p **faire**
    . $v^j_{(t+1)} \leftarrow \arg\min_{v^j \geq 0} \frac{1}{2}\|(D^j x^j) - (D^j U_{(t+1)})v^j\|_2^2$;
  fin
  . Mettre à jour la matrice U ligne par ligne
  pour i *allant de* 1 *à* n **faire**
    . $u^{T\,(i)}_{(t+1)} \leftarrow \arg\min_{u^{T\,(i)} \geq 0} \frac{1}{2}\|(D_i x^{T\,(i)}) - (D_i V^T_{(t)})u^{T\,(i)}\|_2^2$;
  fin
fin

Algorithme 12 : Algorithme pondéré ANLS-WNMF.

3.2.2.4 Imputation par analyse des correspondances multiples : MIMCA

L'analyse factorielle *MCA* (*Multiple Correspondence Analysis*) est une méthode d'analyse de données multivariées qui se spécifie par la recherche de composantes principales dans le but de décrire, résumer et visualiser des données catégorielles disposées sous forme de matrices. Ainsi, si l'on considère une matrice $X = [X^1, \ldots, X^j, \ldots, X^p]$ dont les donnés des p variables sont relevées sur n individus, l'*ACM* peut être vue comme une extension de l'*ACP* (*Analyse en Composantes Principales* utilisant des métriques autres que $\mathbf{D} = \frac{1}{n}\mathbb{1}_n$ définie sur l'espace des variables $\mathbb{R}^n$ et $\mathbf{M} = diag(S^2_{X^1}, \ldots, S^2_{X^p})$, où $S^2_{X^j}$ désigne la variance empirique de X^j, définie sur l'espace des individus $\mathbb{R}^p$. La présence de variables qualitatives dans ce cas impose un recodage des variables car la décomposition en valeurs singulières ne peut être appliquée que sur une matrice constituée de données quantitatives. Ainsi, dans le cadre de l'ACM, qui est la méthode adaptée à des variables qualitatives, l'ensemble des variables est recodé sous la forme d'un tableau disjonctif complet. Chaque variable qualitative est ainsi remplacée par autant d'indicatrices que le nombre de modalités de réponses qu'elle possède. L'illustration est faite dans les deux suivants tableaux suivants (Table 3.5), où celui de gauche contient les modalités de p variables qualitatives, alors que celui de droite désigne le tableau disjonctif complet associé. Ce dernier tableau à valeurs numériques est celui auquel l'on s'intéresse dans le cadre de l'*ACM* et doit ensuite être centré. Il possède n lignes et J colonnes où J est le nombre total de modalités de réponses.

X^1	...	X^p
a	...	b
a	...	b
b	...	a
a	...	c
b	...	b
b	...	c

X^1_a	X^1_b	...	X^p_a	X^p_b	X^p_c
1	0	...	0	1	0
1	0	...	0	1	0
0	1	...	1	0	0
1	0	...	0	0	1
0	1	...	0	1	0
0	1	...	0	0	1

TABLE 3.5: Recodage d'un jeu de données qualitatives sous la forme d'un tableau disjonctif complet. A gauche le jeu qualitatif, à droite le tableau disjonctif complet correspondant.

Ainsi, on s'intéresse ici à l'espace des indicatrices et à celui des individus. La métrique adoptée sur l'espace des individus est $\mathbf{M} = \frac{1}{J}{\mathbf{D}_\Sigma}^{-1}$, avec ${\mathbf{D}_\Sigma}^{-1} = diag(n_1/n, \ldots, n_j/n, \ldots, n_J/n)$, la matrice diagonale avec les proportions des individus par modalité pour éléments diagonaux.

Ainsi la distance entre deux individus est donnée par :

$$d_{i,i'} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{x_{ij} - x_{i'j}}{\sqrt{n_J}}\right)^2. \tag{3.39}$$

Cette métrique implique notamment que deux individus ne prenant pas la même modalité de réponse sont davantage éloignés si l'une des modalités est rare. Ceci permet d'identifier facilement les individus atypiques car ceux-ci se retrouvent éloignés des autres. La métrique sur l'espace des indicatrices est inchangée ($\mathbf{D} = \frac{1}{n}\mathbb{1}_n$). Alors comme pour l'ACP, si on note par $\mathbf{X}$ la matrice recodé centrée, les composantes principales sont obtenues en effectuant la décomposition en valeurs singulières, SVD (singular-value decomposition) du triplet $(\mathbf{X}, \mathbf{M}, \mathbf{D})$.

Cette spécificité de l'*ACM*, a permis de regarder la méthode non pas uniquement sous l'angle de la représentation graphique de la matrice de reconstitution par la formule de factorisation de la SVD, afin de comprendre l'information portée dans le jeu de données, mais d'utiliser cette information pour prédire des valeurs manquantes de la matrice de données (Audigier et al., 2017 [43]. En effet des travaux précédents, permettant d'estimer les paramètres de l'ACM (Josse et al., 2012 [44]) en présence de données manquantes, laissaient déjà entrevoir un moyen de faire de l'imputation pour des données qualitatives. Néanmoins, les algorithmes itératifs proposés n'ont jamais été étudiés, jusqu'alors, en termes de qualité de prédiction des données manquantes. C'est pourquoi Audigier et al. (2017) [43] ont proposé une classe de méthodes ACM d'imputation de données catégorielles (c'est-à-dire issues de variables qualitatives) que sont l'imputation simple et l'imputation multiple toutes deux implémentées dans le package **missMDA** du logiciel `R`, par le biais respectivement des fonctions *imputeMCA* et *MIMCA*. Nous nous intéresserons, ici cependant, à la description de la second fonction. Pour cela commençons d'abord par le rappel du cadre théorique du problème. La matrice $\mathbf{X}$ est celle des données recueillies sur p variables qualitatives, pour n individus. Le tableau disjonctif correspondant est $\mathbf{Z}$ est de taille $n \times r$, où r est le nombre total de modalités de réponses. La métrique sur l'espace des individus est définie par le biais de la matrice diagonale

$$\frac{1}{p}\mathbf{D}_\Sigma^{-1}, \quad \text{où} \quad \mathbf{D}_\Sigma = diag\left(p_1^{X^1}, \ldots, p_{q_1}^{X^1}, \ldots, p_1^{X^p}, \ldots, p_{q_p}^{X^p}\right)$$

avec $\mathbf{D}_\Sigma$ une matrice diagonale de dimension $r \times r$ et $p_\ell^{X^j}$ est la proportion des observations donnant la modalité ℓ sur la variable X^j. Une autre métrique est $\frac{1}{n}\mathbb{1}_n$ construite à partir de la matrice identité $\mathbb{1}_n$ de taille $n \times n$ est définie sur l'espace des variables, ou de manière équivalente à celui des indicatrices, c'est-à-dire des colonnes de $\mathbf{Z}$. Soit $\mathbf{M}$ une matrice de taille $n \times r$ où chaque ligne est égale au vecteur des moyennes de chaque colonne de $\mathbf{Z}$. L'analyse *MCA* consiste alors à chercher une matrice $\hat{\mathbf{Z}}$ de rang inférieure s aussi proche que possible du tableau disjonctif $\mathbf{Z}$ dans le sens de ces métriques déjà définies. Il s'agit donc d'effectuer la décomposition SVD mettant en œuvre le triplet de matrice $(\mathbf{Z} - \mathbf{M}, \frac{1}{p}\mathbf{D}_\Sigma^{-1}, \frac{1}{n}\mathbb{1}_n)$, ce qui est équivalent à la relation suivante :

$$\mathbf{Z} - \mathbf{M} = U\Lambda^{1/2}V^T \tag{3.40}$$

où U est une matrice de taille $n \times r$ dont les colonnes sont des vecteurs singuliers satisfaisant la relation $U^T \text{diag}(1/n, \ldots, 1/n)U = \mathbb{1}_r$, V une matrice de taille $r \times r$ dont les colonnes sont des vecteurs singuliers satisfaisant la relation $V^T\frac{1}{p}\mathbf{D}_\Sigma^{-1}V = \mathbb{1}_r$, et $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})$ la matrice diagonale de taille $r \times r$ des valeurs singulières.

Alors effectuer l'imputation multiple par le procédé MCA , il faut tenir compte de l'incertitude concernant les composantes principales. Pour cela une approche bootstrap non paramétrique basée sur les spécificités de la procédure MCA. En effet celle-ci permet d'attribuer un poids à chaque individu. Cette possibilité d'inclure une pondération pour un individu est très utile lorsque les mêmes lignes de l'ensemble de données apparaissent plusieurs fois. Au lieu de stocker chaque réplique, l'on peut utiliser un poids proportionnel au nombre d'occurrences de chaque ligne, ce qui permet de stocker uniquement les lignes différentes. Ainsi, un bootstrap non paramétrique peut facilement être effectué simplement en modifiant le poids des individus : si un individu

n'appartient pas à la réplique bootstrap, son poids est nul sinon, son poids est proportionnel au nombre de fois où l'observation a lieu dans la réplique. Notez que les individus avec un poids égal à zéro sont classiquement appelés *individus supplémentaires* dans le cadre l'analyse *MCA* (Greenacre, 1984 [89]).

Données : Matrice incomplète de données catégorielles.
Résultat : Matrice imputée.
1. Refléter la variabilité sur l'ensemble des paramètres du modèle d'imputation : tirer n valeurs, avec le principe de remise, dans $\{1, \ldots, n\}$ et définir un poids r_i, pour chaque individu, qui est proportionnel au nombre de fois où l'individu i est tiré ;

2. Imputer le tableau disjonctif à partir des poids précédemment tirés :

. Initialisation : $t = 0$, recoder la matrice de variables qualitatives $\mathbf{X}$ en tableau disjonctif $\mathbf{Z}$, substituant les données manquantes dans $\mathbf{Z}$ par des valeurs initiales (les proportions) et calculer les métriques $\mathbf{M}^{(0)}$ et $\mathbf{D}_\Sigma^{(0)}$ définies à partir de ce tableau complété;

. tant que *le critère de convergence n'est pas atteint* **faire**
- $t \leftarrow t + 1$;
- Effectuez la factorisation SVD sur le triplet
  $\left(\mathbf{Z}^{(t-1)} - \mathbf{M}^{(t-1)}, \frac{1}{p}\left(\mathbf{D}_\Sigma^{(t-1)}\right)^{-1}, \mathrm{diag}(r_1, \ldots, r_n)\right)$;
  pour obtenir $\hat{U}^{(t)}$, $\hat{V}^{(t)}$ et $(\hat{\Lambda}^{(t)})^{1/2}$;
- Identifier le rang s, les vecteurs associés calculer la matrice ajustée;
  $\hat{\mathbf{Z}}^{(t)} = \left(\hat{U}^{(t)}\left(\hat{\Lambda}_s^{(t)}\right)^{1/2}\left(\hat{V}^{(t)}\right)^T\right) + \mathbf{M}^{(t-1)}$;
  où $\left(\hat{\Lambda}_s^{(t)}\right)^{1/2}$ est la matrice diagonal des s premières valeurs singulières
- Mettre à jour la matrice $\mathbf{Z}^{(t)}$ par
  $\mathbf{Z}^{(t)} \leftarrow W * \mathbf{Z} + (1 - W) * \hat{Z}^{(t)}$;
  où * est le produit matricielle de Hadamard, W une matrice de poids,
  avec $w_{ij} = 0$ si z_{ij} est manquante et $z_{ij} = 1$ sinon.
- Mettre à jour les matrices $\mathbf{D}_\Sigma^{(t)}$ et $\mathbf{M}^{(t)}$ à partir de $\mathbf{Z}^{(t)}$

fin
3. Imiter la distribution du jeu de données catégoriques en utilisant le lancer de pièce sur $\mathbf{Z}^{(t)}$:

. Modifier si nécessaire les valeurs de $\mathbf{Z}^{(t)}$: les valeurs négatives sont remplacées par 0 et les valeurs supérieures à 1 sont remplacées par 1.

. Pour les cellules imputées codant une valeur manquante, tirer une modalité selon la distribution multinomiale.

4. Créer M ensembles de données imputées :
pour *m allant de 1 à M* **faire**
| Alterner les étapes **1**, **2** et **3.**
fin

Algorithme 13 : Algorithme MIMCA.

Alors l'algorithme *MIMCA* est décrit comme suit. D'abord un nombre M d'ensembles de poids pour les individus sont tirés. Alors M imputations simples sont effectuées : : dans un premier temps, l'algorithme *MCA itératif régularisé* est utilisé pour imputer le tableau disjonctif incomplet en utilisant les poids précédents pour les individus ; ensuite, la pièce de monnaie est utilisée pour obtenir des données catégoriques et imiter la distribution des données catégorielles. À la fin, M ensembles de données imputées sont obtenus et n'importe quelle méthode statistique peut être appliquée à chacun d'eux. La procédure plus détaillée est ici donnée par l'algorithme

13.

Nous terminons cette section de l'état de l'art par la présentation quelques méthodes dites *naïves*, ainsi nommées parce que, intrinsèquement, elles n'effectuent aucune étude, sur les données, en amont ou pendant le processus d'analyse. Il s'agit des plus illustres souvent utilisées dans la pratique.

3.2.2.5 Les méthodes naïves

Il s'agit essentiellement des méthodes d'imputation par la *moyenne*, la *médiane*, le *mode* ou encore la méthode *LOCF*. Ce sont des méthodes d'imputation intuitives qui induisent très souvent des biais dans par rapport aux estimations en raison de leur caractères rudimentaires (absence d'analyse).

La méthode d'imputation par la *moyenne* est une procédure qui consiste à compléter les valeurs manquantes par la moyenne calculée sur l'ensemble des valeurs présentes. C'est un procédé d'imputation pouvant se faire avec la fonction *impute* implémentée dans le package **Hmisc** du logiciel `R`. Il s'agit d'une fonction dont l'une des entrées, à savoir l'argument *fun* sert à spécifier la méthode (sous forme de fonction) par devrait être effectuée, en l'occurrence ici la fonction *mean*.

La méthode d'imputation par la *médiane* est une procédure qui consiste à compléter les valeurs manquantes par la médiane calculée sur l'ensemble des données présentes. C'est une stratégie d'imputation pouvant aussi s'effectuer par la fonction *impute*.

La méthode d'imputation par le *mode* est un procédé qui consiste à compléter les données manquantes par le mode calculé sur l'ensemble des valeurs observées. C'est une procédure d'imputation qui peut aussi s'effectuer par la fonction *impute*.

La méthode *LOCF* (*Last Observation Carried Forward*) elle complète chaque valeur manquante par la dernière valeur observée en parcourant la matrice des données. Il s'agit d'une procédure d'imputation implémentée, sous forme d'une fonction *na.locf*, dans le package **zoo** du logiciel `R`.

Cependant pour les trois premières procédures d'imputation, afin de tenter d'augmenter leurs performances respectives en comparaison avec l'algorithme d'imputation *PGNMF* que nous développerons au Chapitre 4, nous effectuerons l'imputation pour chaque variable à partir de la caractéristique (moyenne, médiane ou mode) calculée sur les valeurs observées de celle-ci.

Conclusion

La problématique des données manquantes a longtemps été un thème qui a intéressé de nombreuses études sur des phénomènes aussi divers que variés, et concernant plusieurs domaines comme l'économie, la biologie, la génétique, les questionnaires d'enquête, ... Les analyses statistiques multivariées ont alors développées diverses méthodologies pour s'adapter à l'absence de certaines valeurs dans les données recueillies. Ainsi plusieurs modélisations statistiques, d'approche paramétrique ont permis l'inférence sur les paramètres et en présence de données manquantes, comme par exemple la très connue méthode EM (Espérance-Maximisation) et plusieurs autres stratégies basées sur la vraisemblance des paramètres sur les données observées. Cependant lorsque le taux de données manquantes est assez élevé, les analyses statistiques risquent fort d'être biaisées ; c'est pourquoi de nombreuses études ont proposé comme solution l'imputation des données manquantes.
Plusieurs auteurs se sont très tôt intéressés aux inférences statistiques en présence de données manquantes, par exemple Wilks, 1932 [1], Healy et Westmacott, 1956 [2], Anderson, 1957 [3]. Toutefois pour voir l'engouement et la convergence massive autour de la problématique des non-réponses, il faut attendre la formalisation mathématique des causes d'absence des données par Rubin, 1976 [9]. En effet les trois mécanismes sous-jacent à l'absence de données qui ont été formalisés ont été classés en deux catégories : il s'agit des causes MAR et MCAR pour les mécanismes dits ignorables et des causes de type MNAR

qui constituent les mécanismes non-ignorables. A côté de cette catégorisation, une autre non moins importante a été la celle sur la structure des données manquantes. Ainsi trois types de structures ont été donnés (Rubin, 1987 [11], Little et Rubin, 1987 [14]) : les structures univariées, monotones et arbitraires. La prise en compte de cette typologie à deux niveaux (mécanismes et structures) des données manquantes est un préalable à une imputation efficiente. Avec cette formalisation, une véritable étude théorique de la statistique des données manquantes a été développée (Rubin, 1978 [11], Little, 1982 [13], Rubin, 1987 [11], Little et Rubin, 1987 [14], Schafer, 1999, [15], Schafer et Yucel, 2002 [16] ...)

Dans ce chapitre nous avons ainsi présenté, d'une part, la problématique des données manquantes, dans le contexte particulier mais aussi récurrent des questionnaires d'enquêtes, et d'autre part l'état de l'art, en mettant en relief les méthodes d'imputation d'analyses factorielles. A ce titre le chapitre est articulé autour de deux points majeurs.

D'abord nous présenté la double typologie des données manquantes. La première a concerné la répartition de ce type de données à l'intérieur des variables disposées sous forme de matrice (ou tableau). Cette répartition peut concerner une seule variable, on parlera alors de structure univariée. La disposition des données manquantes peut aussi intéresser plusieurs variables, et se faire sous forme étagée suivant leur nombre par variable, on dira alors que la structure est monotone. La disposition des données manquantes peut être de telle sorte ces données apparaissent de façon hasardeuse à l'intérieur des variables, la structure est alors dite aléatoire. La seconde typologie a concerné les catégories de données manquantes selon la cause sous-jacente. Nous avons, d'abord donné un grand aperçu des causes pratiques d'absence des données, particulièrement pour ce qui concerne les données de questionnaires, allant de la perte de certaines réponses à l'incompréhension de certains points, en passant par la sensibilité de certaines questions et beaucoup d'autres causes encore. Nous avons ensuite donné la formalisation des ces différents mécanismes qui sont de trois types selon la classification de Rubin (1976) [9] (*MAR*, *MCAR* et *MNAR*). Nous avons ainsi .

Ensuite nous avons donné l'état de l'art par la présentation d'une large revue des méthodes de résolution développées, en insistant sur les méthodes d'imputation qui seront comparées à l'algorithme résultant de la méthodologie que nous développerons au chapitre suivant. Les analyses statistiques, faites sur données comportant des valeurs manquantes, que nous avons présenté sont deux catégories : les analyses sans imputation et les procédures d'imputation. La première catégorie est constituée des études de cas complets et des études de cas disponibles. La seconde catégorie a comporté diverses méthodes d'imputation basées sur le principe de l'analyse factorielle, ainsi que quelques procédures de complétion naïves.

Si les données manquantes apparaissent pour différentes catégories de variables multivariées, comme des variables biologiques, génétiques, économiques ou encore des variables environnementales, leur récurrence est surtout notée sur les données de questionnaires qui ont constitué essentiellement l'objet de ces études ci-dessus citées. Ainsi parmi la diversité des causes des non-réponses pour il y en a qui sont spécifiques et liées à la nature même du questionnaire. En effet dans différents pays les structures de santé publique sont évaluées sur la qualité de services dispensés aux patients. Si certaines ont développé des *systèmes d'informations cliniques* (*SIC*), comme c'est le cas de l'Hôpital européen Georges Pompidou de Paris ou encore le Centre Hospitalier Universitaire de Sherbrooke au Québec (Palm, 2010 [90]), pour une prise en charge efficiente, d'autres ont mis en place un instrument d'évaluation de la sécurité du patient par le personnel de structures. L'instrument consiste en un questionnaire auto-administré, c'est-à-dire dont la structure de santé adresse à son personnel. L'enquête-type nommé HSOPSC (Hospital Survey on Patient Safety Culture) a été réalisé par l'agence américaine AHRQ (Agency for Healthcare Research and Quality) qui est une structure fédérale chargée de conduire et de supporter la recherche sur la sécurité des patients et la qualité des soins médicaux pour tous les citoyens américains (Sorra et Nieva, 2004 [91]). Cependant c'est un questionnaire qui souffre d'un problème « d'acceptabilité » de par la nature des questions, et pourrait conduire à un taux élevé de non-réponses. La version française du HSOPSC (Occelli, 2013 [35]) soumis à l'hôpital universitaire de Grenoble, entre avril 2013 et septembre 2014 constituera l'ensemble de nos données réelles d'applications. Au regard de la contrainte liée à la durée de l'enquête et du problème d'acceptabilité du questionnaire, une alternative serait le sous-échantillonnage du HSOPSC, en vue d'une reconstruction automatique, sujet du dernier chapitre de cette thèse.

Chapitre 4

Factorisation NMF Poisson-Gamma et reconstruction automatique du questionnaire médical HSOPSC

Introduction

Dans ce dernier chapitre il s'agit de l'articulation d'un modèle bayésien et d'une modélisation de factorisation matricielle NMF, pour ainsi dire que ces deux modèles d'analyses statistiques multivariées peuvent être associés afin de répondre à une problématique posée : l'imputation de données manquantes. En effet c'est le chapitre qui constitue la jonction des chapitres 1 et 2 pour proposer une solution à la problématique des données manquantes posée au chapitre 3. Rappelons que les chapitres 1 et 2 ont respectivement abordé quelques aspects fondamentaux de la modélisation bayésienne et la méthodologie déterministe de la factorisation matricielle NMF. Ces deux approches ont séparément été utilisées pour répondre à la problématique des données manquantes dans des domaines d'applications spécifiques : nous avons les études de Rubin, 1978 [10], Rubin, 1988 [12], Schafer, 1997 [15], van Buuren, 2007 [82], Resche-Rigon et White, 2016 [18] ... pour ce concerne la modélisation bayésienne et ceux de Mao et Saul, 2004 [21], Kim et Choi, 2009 [22] ... pour la factorisation déterministe NMF. Les approches combinées, de résultantes les modèles bayésiens NMF n'ont jusqu'ici pas beaucoup été utilisées pour l'imputation de données manquantes (Salakhutdinov et Mnih, 2008 [27]), et en particulier les modèles Poisson (Cemgil, 2009 [29]). Notre modèle de vraisemblance de loi Poisson avec des lois a priori Gamma sur les paramètres se présente alors comme une contribution à ce type de modèle mixte. Toutefois, avec l'algorithme d'inférence qui en découle caractérisé par un échantillonneur de Gibbs et que nous nommons PGNMF (Poisson-Gamma Nonnegative Matrix Factorization), il se veut novateur dans l'application à l'imputation de données de questionnaire et en l'occurrence le questionnaire type HSOPSC (Sorra et Nieva, 2004 [91], Occelli et al, 2013 [35]).

Le modèle bayésien NMF de type Poisson-Gamma a déjà été développé par divers auteurs. D'abord Canny, 2004 [23] a été l'un des précurseurs de la factorisation NMF probabiliste ; cependant son modèle n'a pa comporté pas d'aspects bayésiens, mais une vraisemblance comme produit de lois Poisson et Gamma et la méthode d'inférence l'algorithme EM (Espérance-Maximisation). Ensuite les auteurs Virtanen et associés ont établi un modèle bayésien NMF pour la factorisation $X \approx UV$, où les données X sont de loi Poisson et les paramètres U et V de loi a priori Gamma. Les auteurs ont montré l'équivalence de ce modèle basique à celui de Lee et Seung, 2001 [20] ; cependant dans le soucis d'obtenir des résultats consistants dans la modélisation du traitement de signal, ils ont introduit une chaine de Makov Gamma par le biais de variables auxiliaires Z qui sont estimées en même temps que les paramètres U et V avec l'algorithme MAP (Maximum a Posteriori) (Virtanen et al, 2008 [26]). Le même modèle basique a été repris par Cemgil, qui était d'ailleurs un des auteurs associés de Virtanen. Cependant Cemgil a adapté le modèle à complétion de données manquantes, et appliqué à la reconstitution d'images dégradées via l'algorithme variationnel d'estimation de la loi a posteriori (Cemgil, 2009 [29]). Notre modèle s'inspire des études de Cemgil (2009), et est aussi dédié, dans cette étude, à la complétion de données manquantes. Toutefois la différence réside dans la méthode d'inférence des paramètres U et V en ce sens nous avons utilisé un échantillonneur de Gibbs (algorithme PGNMF) à la place. Comme nous l'avons dit l'autre particularité de notre modèle est qu'il n'as pas encore fait l'objet d'une application sur des données de questionnaires et particulièrement le questionnaire HSOPSC.

La problématique des données manquantes, comme nous l'avons vu au Chapitre 3, a fait l'objet de

beaucoup d'études (Afifi et Elashoff, 1966 [6], Hartley et Hocking, 1971 [7], Rubin, 1976 [9], Little et Rubin, 1987 [14], Schafer, 1997 [15], Schafer et Yucel, 2002 [16] ...), en particulier sur les données de questionnaires (Rubin, 1978 [10], Little, 1982 [13], Rubin, 1987 et 1988 [11, 12], Schafer et Graham, 2002 [92] ...) et principalement des questionnaires du domaine médical (Cohen, 1982 [93], Taylor et al, 2002 [79], Shrive et al, 2006 [81], van Ginkel et al, 2007 [83], Resseguier et al, 2013 [85] ...). Des méthodes récentes de complétion et comptatibles avec des données de questionnaire ont aussi été développées par différents auteurs. En effet l'algorithme EM de Dempster et al, 1977 [56] a été implémenté sous un modèle Poisson de données augmentées (Cemgil, 2009 [29]. Une méthode de factorisation NMF pondérée, nommée WNMF (Weighted Nonnegative Matrix Factorization), basée sur les techniques des moindres carrées alternées ANLS (Alternating Nonnegative Matrix Factorization) (Zdunek et Cichocki, 2006 [36], Kim et al, 2007 [37], Lin, 2007 [38], Kim et Park, 2008 [39]) a été introduite par Kim et Choi, 2009 [22]. Sur la base de la méthode dites des forêts aléatoires de Breiman, 2001 [40], les auteurs Stekhoven et Bühlmann ont proposé une stratégie d'imputation de données mixtes (continues et/ou catégorielles) qu'ils ont nommée MissForest. Par ailleurs Josse et Husson, 2016 [42] puis Audigier et al, 2017 [43] ont présenté des méthodes d'imputations multiples par analyse des correspondance, MIMCA (Multiple Imputation by Multiple Correspondence Analysis), qui effectuent l'imputation de données catégorielles en utilisant l'analyse des correspondances multiples (ACM) (Josse et al, 2012 [44]).

La méthode d'imputation PGNMF que nous proposons sera comparée en termes de performance à ces méthodes récentes citées. En effet l'algorithme présente de meilleurs résultats suivant le critère RMSE (Root Mean Square Error) lorsque le taux de données manquantes dépasse le seuil de 40% mais donne aussi des performances acceptables par rapport au critère marginal divergence de Kullback-Leibler (KL) lorsque le taux de données manquantes est faible avec un taux de moins de 50%. En réalité le problème que nous cherchons à résoudre est celui d'une reconstruction automatique de questionnaire. Ainsi nous supposons que les enquêteurs ont délibérément réduit le questionnaire en retirant ces certains points et en le spécifiant aléatoirement pour chaque individu enquêté. Ce qui nous amène à considérer le mécanisme MCAR (Missing Completely at Random) (Rubin, 1976 [9]) comme sous-jacent à l'absence de données. Ainsi les données manquantes ne sont pas le fait d'une perte ou d'un refus de réponse mais plutôt le fait de points non soumis à l'individu.

L'enquête type repose sur un instrument appelé HSOPSC (Hospital Survey on Patient Safety Culture) développé en 2004 par l'agence américaine AHRQ (Agency for Healthcare Research and Quality) (Sorra et Nieva, 2004 [91]). Cette enquête a été élaborée sur la base d'une revue de la littérature, raffinée selon la théorie psychométrique et soutenue par des analyses psychométriques effectuées en 2004 sur un personnel de 1437 agents travaillant dans 21 hôpitaux américains. Certaines méthodes d'analyse du questionnaire ne tiennent pas compte des données manquantes, selon la méthode de calcul des scores proposée par l'agence AHRQ. De telles approches tendent à réduire le nombre de données et la puissance des enquêtes, tout en favorisant l'introduction de biais (Rotnitzky et Wypij, 1994 [94], Demissie et al, 2003 [78], Joseph et al, 2004 [80]). Le questionnaire HSOPSC est conçu pour évaluer la culture de la sécurité des patients du point de vue du personnel hospitalier. Il est auto-administré et comporte quarante-deux points (ou items). Bien que les propriétés psychométriques du questionnaire soient bien caractérisées, l'enquête souffre d'un problème d'acceptabilité et peut posséder un taux élevé de non-réponses. Ce qui motive ici la considération d'un questionnaire réduit au départ, pour procéder ensuite à la reconstruction automatique de l'intégralité des réponses. Le questionnaire considéré est la version française du HSOPSC (Occelli et al, 2013 [35]) effectué à l'hôpital universitaire de Grenoble entre 2013 et 2014.

Le chapitre est structuré autour quatre points majeurs. D'abord la section 4.1 reprend l'étude faite par Virtanen et al, 2008 [26] et montre la connexion très étroite qui existe entre le modèle Poisson NMF que nous considérons ici et le modèle NMF classique de Lee et Seung, 2001 [20]. Ensuite la section 4.2 présente le questionnaire HSOPSC et notre modèle bayésien NMF de reconstruction matricielle. Puis dans la section 4.3 nous donnons une brève présentation méthodes alternatives déjà décrites an chapitre 3, ainsi que quelques méthodes naïves. Enfin dans la section 4.4 nous comparons notre algorithme PGNMF à plusieurs méthodes récemment développées, par le biais de deux critères d'évaluation RMSE (Root Mean Square Error) et KL (divergence Kullback-Leibler).

4.1 Lien avec le modèle NMF de fonction-coût la divergence de Kullback

La vraisemblance de loi Poisson considérée montre le lien très étroit qui existe entre certaines méthodes NMF déterministes d'une part et probabilistes d'autre part. Pour le comprendre il faut remonter au premier article traitant de la factorisation NMF de Lee et Seung, dans lequel les auteurs ont évoqué la

relation qui existe entre la fonction-coût

$$F(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}\log\sum_{\ell=1}^{k}u_{i\ell}v_{\ell j} - \sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right) \tag{4.1}$$

qu'ils avaient considérée et une vraisemblance de loi Poisson (Lee et Seung, 1999 [19]). Ils ont ainsi interprété le problème de maximisation de la fonction F, via les équations mises à jour multiplicatives (Ref. Chap. 2, Eq. (2.5)) comme la maximisation d'une vraisemblance de loi Poisson. Cette relation a été clairement expliquée par Virtanen et auteurs associés (Virtanen et al, 2008 [26]) qui ont établi la relation existante la fonction donnée par (4.1), la vraisemblance Poisson et la divergence de Kullback généralisée (Ref. Chap. 2, Eq. (2.3)). En effet les auteurs ont montré que si nous considérons la fonction-coût (2.3) appliquée au couple (X, UV),

$$D_{KL}(X\|UV) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}\log\frac{x_{ij}}{\sum_{\ell}u_{i\ell}v_{\ell j}} - x_{ij} + \sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right), \tag{4.2}$$

alors le premier problème de minimisation (Problème 1, Chap. 2, Sect. 2.1.1.2) que nous reprenons ici

$$(U^*,V^*) = \arg\min_{U\geq 0, V\geq 0} D_{KL}(X\|UV) \tag{4.3}$$

est équivalent au problème suivant

$$(U^*,V^*) = \arg\min_{U\geq 0, V\geq 0} -F(U,V). \tag{4.4}$$

Ce qui revient à écrire

$$\arg\min_{U\geq 0, V\geq 0} D_{KL}(X\|UV) = \arg\min_{U\geq 0, V\geq 0} -F(U,V) \tag{4.5}$$

Nous observons bien ici la cohérence avec la maximisation de $F(U,V)$ évoquée par Lee et Seung (1999). La relation entre la divergence de Kullback-Leibler $D_{KL}(X\|UV)$ et la fonction-coût $F(U,V)$ est donnée par les équations suivantes

$$\begin{aligned} D_{KL}(X\|UV) &= \sum_{ij}\left(x_{ij}\log x_{ij} - x_{ij}\log\sum_{\ell=1}^{k}u_{i\ell}v_{\ell j} - x_{ij} + \sum_{\ell=1}^{k}u_{i\ell}v_{\ell j}\right) \\ &= \sum_{ij}\left(-x_{ij}\log\sum_{\ell}u_{i\ell}v_{\ell j} + \sum_{\ell}u_{i\ell}v_{\ell j}\right) + \sum_{ij}\left(x_{ij}\log x_{ij} - x_{ij}\right) \\ &= -F(U,V) + \sum_{ij}\left(x_{ij}\log x_{ij} - x_{ij}\right). \end{aligned} \tag{4.6}$$

L'équation (4.6) peut se réécrire comme suit :

$$D_{KL}(X\|UV) = -F(U,V) + c_1 \tag{4.7}$$

où $c_1 = \sum_{ij}(x_{ij}\log x_{ij} - x_{ij})$ est une constante en ce sens que la fonction-coût divergence de Kullback-Leibler est regardée comme une fonction du couple (U,V). Ce qui montre bien que minimiser $D_{KL}(X\|UV)$ revient à minimiser $-F(U,V)$ ou encore de manière équivalente à maximiser $F(U,V)$). Virtanen et al. ont montré que cette fonction-coût $F(U,V)$ de Lee et Seung (1999) constitue la jonction entre la divergence $D_{KL}(X\|UV)$ et la vraisemblance de loi Poisson. En effet supposons que les données de la matrice X sont toutes indépendantes et de loi Poisson, c'est-à-dire $x_{ij} \sim \mathcal{P}(x_{ij}\,;\,\sum_{\ell}u_{i\ell}v_{\ell j})$. Posons

$$p(x_{ij}\mid U,V) = \mathcal{P}(x_{ij}\,;\,\sum_{\ell}u_{i\ell}v_{\ell j}) \;\text{ et donc }\; p(X\mid UV) = \mathcal{P}(X\,;\,UV)$$

on a alors,

$$\begin{aligned} \log p(X\mid UV) &= \sum_{ij}\left(-\sum_{\ell}u_{i\ell}v_{\ell j} + x_{ij}\log\sum_{\ell}u_{i\ell}v_{\ell j} - \log(\Gamma(x_{ij}+1))\right) \\ &= \sum_{ij}\left(x_{ij}\log\sum_{\ell}u_{i\ell}v_{\ell j} - \sum_{\ell}u_{i\ell}v_{\ell j}\right) + \sum_{ij}\Big(\log\big(\Gamma(x_{ij}+1)\big)\Big) \\ &= F(U,V) + c_2 \end{aligned} \tag{4.8}$$

où $c_2 = \sum_{ij} \log\big(\Gamma(x_{ij}+1)\big)$ est regardé comme une constante puisque ne dépendant pas de U et V. L'équation (4.8) peut se réécrire comme suit,

$$\log p(X \mid UV) = \log \exp\left(F(U,V)\right) + c_2 \tag{4.9}$$

L'équation (4.9) est équivalente à

$$p(X \mid UV) \propto \exp\left(F(U,V)\right) \tag{4.10}$$

Dans (4.10) les fonctions $p(X \mid UV)$ et $\exp(F(U,V))$ sont proportionnelles et positives donc elles ont la même variation. De même $\exp(F(U,V))$ et $F(U,V)$ ont même variation, d'où $p(X \mid UV)$ et $F(U,V)$ varient dans les mêmes sens. Cela revient à dire que maximiser la vraisemblance de loi Poisson revient à maximiser la fonction-coût $F(U,V)$ ou de manière équivalente à minimiser $-F(U,V)$, d'où

$$\arg \max_{U\geq 0, V\geq 0} p(X \mid UV) = \arg \min_{U\geq 0, V\geq 0} -F(U,V). \tag{4.11}$$

Les équations (4.5) et (4.11) permettent alors d'établir que,

$$\arg \min_{U\geq 0, V\geq 0} D_{KL}(X\|UV) = \arg \max_{U\geq 0, V\geq 0} p(X \mid UV) \tag{4.12}$$

qui montre bien que minimiser la fonction-coût divergence de Kullback $D_{KL}(X\|UV)$ (Lee et Seung, 2001 [20]) revient à interpréter le problème avec un modèle probabiliste de loi Poisson $p(X \mid UV)$ de paramètre le produit UV, avec l'hypothèse d'indépendance sur les coefficients. Une présentation détaillée du modèle est faite dans la section 4.2.2.

4.2 Description des données et du modèle Poisson-Gamma

4.2.1 Description des données

Nos données d'application sont celles du questionnaire-type HSOPSC (Hospital Survey on Patient Safety Culture). Comme nous l'avons dit en Introduction le HSOPSC est un instrument développé en 2004 par l'agence américain AHRQ (Agency for Healthcare Research and Quality). C'est une agence pour la recherche et la qualité des services en santé publique. Elle est la principale agence fédérale chargée de mener et de soutenir des recherches visant à améliorer la sécurité des patients et la qualité des soins de santé pour tous les Américains (Sorra et Nieva, 2004 [91]). L'objectif de l'AHRQ est de promouvoir une culture de la sécurité et de l'amélioration de la qualité dans le système de santé des États-Unis, ce qui contribuera à accélérer l'adoption des résultats de la recherche dans la pratique et les politiques. Ainsi l'agence a permis le développement de cette enquête sur la culture de la sécurité des patients. C'est un outil utile pour évaluer la culture de sécurité d'un hôpital dans son ensemble ou pour des unités spécifiques au sein d'un hôpital. En outre, l'enquête peut être utilisée pour suivre l'évolution de la sécurité des patients au fil du temps et pour évaluer l'impact des interventions en matière de sécurité des patients.

En outre, depuis 2001, l'AHRQ a soutenu un large éventail d'autres recherches sur la sécurité des patients afin de mettre au point des méthodes novatrices de collecte, d'analyse et de communication des données sur la sécurité des patients ; comprendre l'impact des conditions de travail sur la sécurité des patients, y compris les sciences de l'ergonomie et les facteurs humains ; et encourager l'utilisation des technologies de l'information pour réduire les erreurs médicales.

Pour développer le questionnaire, les chercheurs ont effectué une revue de la littérature sur la sécurité, les accidents, les erreurs médicales, le signalement des erreurs, la culture de sécurité. En outre, les chercheurs ont passé en revue les enquêtes existantes publiées et non publiées sur la culture de sécurité. Ils ont également mené des entretiens en personne (face à face) et aussi par le biais de communication téléphoniques avec le personnel d'hôpitaux. L'enquête a été soumise à un test préalable auprès du personnel d'hôpitaux afin de s'assurer que les éléments étaient facilement compris et pertinents par rapport à la sécurité des patients en milieu hospitalier. Enfin, l'enquête a été testée auprès de plus de 1 400 employés de 21 hôpitaux à travers les États-Unis.

Des données pilotes ont d'abord été collectées puis analysées, en examinant les statistiques des items ainsi que la fiabilité et la validité des échelles de la culture de la sécurité. Aussi la structure factorielle de l'enquête a été examinée par des analyses factorielles exploratoires et confirmatoires. Sur la base de l'analyse des données pilotes, le questionnaire a été révisée en ne retenant que les meilleurs éléments et échelles. Ainsi le HSOPSC résultant est jugé avoir de bonnes propriétés psychométriques pour les éléments et les échelles inclus.

L'enquête et le matériel de la boîte à outils qui l'accompagne sont conçus pour fournir aux responsables hospitaliers les connaissances de base et les outils nécessaires pour mener une évaluation de la

culture de la sécurité, ainsi que des idées pour l'utilisation des données. Ainsi le HSOPSC est divisé en deux parties. La première partie présente des problèmes inhérents au processus de collecte de données et à l'organisation globale du projet. La deuxième partie comprend le formulaire d'enquête, suivi d'un aperçu séparé des éléments inclus, regroupés en fonction des dimensions de la culture de sécurité qu'ils sont censés mesurer et des résultats de fiabilité dérivés des données pilotes.

Après les États-Unis qui l'on conçu, le HSOPSC a par la suite été adopté par plusieurs autres pays qui trouvent en lui un moyen efficace d'évaluation de leurs différents systèmes de santé publique. Ainsi dans cette thèse, les données de notre étude sont celles provenant de la version française du HSOPSC (Occelli et al, 2013 [35]) soumis à l'hôpital universitaire de Grenoble, d'une capacité de 1836 lits et desservant une population de 675 000 habitants. L'enquête a été menée de façon anonyme et bénévole entre avril 2013 et septembre 2014. Les participants admissibles étaient des employés à temps plein ou à temps partiel ayant au moins 6 mois d'emploi dans les services cliniques, de laboratoire, de pathologie, de radiologie ou de pharmacie. Sur 5044 employés admissibles, 3888 (77.08%) ont participé à l'étude. Les données issues de l'enquête présentent environ 1.8% de non-réponses.

Pour mieux comprendre le questionnaire-type HSOPSC et ses principaux aspects , nous donnons ici un aperçu de la notion de culture de sécurité, les douze (12) dimensions de la culture de la sécurité que l'enquête cherche à mesurer et la procédure de conduite de l'enquête (Sorra et Nieva, 2004 [91] ; Jones et al, 2008 [95]).

4.2.1.1 Notion de la culture de sécurité

Les différentes définitions de la culture de sécurité contiennent plusieurs éléments communs (Health and Safety Commission, 1993 [96] ; Wiegmann et al, 2002 [97] ; Jones et al, 2008 [95]). La culture de la sécurité fait référence aux croyances et pratiques persistantes et partagées des membres de l'organisation concernant sa volonté de détecter les erreurs et d'en tirer des leçons. L'institut IOM (Institute of Medicine, 2004 [98]) déclare qu'une culture de la sécurité dans les soins de santé nécessite trois éléments :

✓ Une conviction que, bien que les processus de soins de santé présentent un risque élevé, ils peuvent être conçus pour éviter les échecs.

✓ Un engagement au niveau organisationnel pour détecter les erreurs et en tirer les leçons.

✓ Un environnement réglementaire (donc perçu comme juste) parce que la discipline s'applique lorsqu'un employé augmente sciemment le risque pour les patients et les autres agents.

Une culture de la sécurité est présente dans les organisations à haute fiabilité, caractérisées par des processus complexes et risqués, mais des taux d'erreur très faibles. De telles organisations atteignent une fiabilité élevée, car elles sont préoccupées par les échecs, sensibles à la manière dont chaque membre de l'équipe affecte un processus, permettent à ceux qui connaissent le mieux un processus de prendre des décisions et résistent à la tentation de blâmer les individus pour des erreurs dans des processus complexes.

La sécurité des patients est une composante essentielle de la qualité des soins de santé. Alors que les organisations de soins de santé cherchent continuellement à s'améliorer, on reconnaît de plus en plus l'importance d'instaurer une culture de la sécurité. Pour instaurer une culture de la sécurité, il faut comprendre les valeurs, les convictions et les normes relatives à ce qui est important pour une organisation et connaître les attitudes et les comportements liés à la sécurité des patients qui sont attendus et appropriés (Sorra et Nieva, 2004 [91]). La sécurité du patient est définie comme l'évitement et la prévention des blessures du patient ou des événements indésirables résultant des processus de prestation des soins de santé.

Ainsi une définition de la *culture de la sécurité* est donnée par l'organisation britannique HSC (Health and Safety Commission) dans son troisième rapport sur l'Organisation de la Sécurité (Health and Safety Commission, 1993 [96]).

La culture de sécurité d'une organisation est le produit de valeurs, d'attitudes, de perceptions, de compétences et de modèles de comportement individuels et collectifs qui déterminent l'engagement, le style et les compétences de gestion de la santé et de la sécurité d'une organisation. Les organisations ayant une culture de sécurité positive se caractérisent par des communications fondées sur la confiance

mutuelle, par des perceptions communes de l'importance de la sécurité et par la confiance dans l'efficacité des mesures de prévention.

4.2.1.2 Les dimensions de la culture de sécurité mesurées dans le questionnaire

L'enquête met l'accent sur les problèmes de sécurité des patients et sur le signalement des erreurs et des événements. Un *événement* est défini comme tout type d'erreur, d'incident, d'accident ou de déviation, qu'il cause ou non un préjudice à un patient. Ainsi le HSOPSC est composé de quarante deux (42) items (questions) classés en douze (12) dimensions, lesquelles sont regroupées en trois (03) catégories. La première comporte deux (02) dimensions qui sont des mesures de résultats : les perceptions générales de la sécurité et la fréquence des événements rapportés. La seconde catégorie regroupe sept (07) dimensions qui mesurent la culture de la sécurité au niveau des unités de travail (les services). Il s'agit des attentes du superviseur/responsable et actions favorisant la sécurité des patients, de l'apprentissage organisationnel, du travail d'équipe au sein des services, de l'ouverture à la communication, du retour d'information et la communication sur erreur, de la réponse non punitive à une erreur et de la dotation en personnel. La dernière catégorie renferme trois (03) dimensions qui mesurent la culture de la sécurité au niveau de l'entité toute entière (la structure hospitalière). Il s'agit du soutien de la direction de l'hôpital pour la sécurité des patients, du travail d'équipe dans les départements de l'hôpital, des transferts et transitions entre hôpitaux. Par ailleurs des essais pilotes ont permis de s'assurer que l'enquête avait de bonnes propriétés psychométriques (Sorra et Nieva, 2004 [91]).

Dimensions	**Description**
Dimension 1	**Perceptions globales de la sécurité**
A10	C'est par hasard que des erreurs plus graves ne se produisent pas ici.
A15	La sécurité des patients n'est jamais sacrifiée pour faire plus de travail.
A17	Nous avons des problèmes de sécurité des patients dans cette unité.
A18	Nos procédures et systèmes sont efficaces pour éviter les erreurs de se produire.
Dimension 2	**Fréquence des comptes rendus d'événements**
D1	Lorsqu'une erreur est commise mais qu'elle est interceptée et corrigée avant d'affecter le patient, à quelle fréquence est-elle rapportée ?
D2	Lorsqu'une erreur est commise, mais qu'elle n'a pas le potentiel de nuire au patient, à quelle fréquence est-elle rapportée ?
D3	Lorsqu'une erreur est commise qui pourrait nuire au patient mais ne le fait pas, à quelle fréquence est-elle rapportée ?

Dimension 3	**Attentes des superviseurs / gestionnaires et actions en faveur de la sécurité des patients**
B1	Mon superviseur / responsable dit un bon mot quand il / elle voit un travail effectué conformément aux procédures établies pour la sécurité des patients.
B2	Mon superviseur/responsable prend sérieusement en compte les suggestions du personnel pour améliorer la sécurité des patients.
B3	Chaque fois que la pression augmente, mon supérieur hiérarchique ou directeur souhaite que nous travaillions plus rapidement, même si cela implique de prendre des raccourcis.
B4	Mon superviseur/responsable oublie les problèmes de sécurité des patients qui se répètent
Dimension 4	**Apprentissage organisationnel - amélioration continue**
A6	Nous agissons activement pour améliorer la sécurité des patients
A9	Les erreurs ont conduit à des changements positifs ici
A13	Après avoir apporté des modifications pour améliorer la sécurité des patients, nous évaluons leur efficacité.
Dimension 5	**Travail d'équipe au sein des unités**
A1	Les gens se soutiennent dans cette unité.
A3	Lorsque beaucoup de travail doit être fait rapidement, nous travaillons en équipe pour faire le travail.
A4	Dans cette unité, les gens se traitent avec respect.
A11	Lorsqu'un secteur de cette unité devient vraiment occupé, d'autres aident.
Dimension 6	**Ouverture de la communication**
C2	Le personnel s'exprimera librement s'il voit quelque chose qui peut avoir un effet négatif sur les soins des patients.
C4	Le personnel est libre de remettre en question les décisions ou les actions de ceux qui ont plus d'autorité.
C6	Le personnel a peur de poser des questions lorsque quelque chose ne semble pas bien.
Dimension 7	**Retour d'information et communication sur les erreurs**
C1	Nous recevons des commentaires sur les changements mis en place basés sur les rapports d'événements.
C3	Nous sommes informés des erreurs qui se produisent dans cette unité.
C5	Dans cette unité, nous discutons des moyens d'éviter que des erreurs ne se reproduisent.

Dimension 8	**Réponse non punitive à une erreur**
A8	Le personnel a le sentiment que ses erreurs lui sont reprochées.
A12	Lorsqu'un événement est rapporté, on a l'impression que le compte rendu est fait sur la personne et non sur le problème.
A16	Le personnel craint que ses erreurs ne soient conservées dans son dossier personnel.
Dimension 9	**Recrutement**
A2	Nous avons suffisamment de personnel pour gérer la charge de travail.
A5	Le personnel de cette unité travaille plus longtemps que le meilleur temps assigné aux soins des patients.
A7	Nous utilisons plus d'agence ou de personnel temporaire que ce qui est préférable pour les soins aux patients.
A14	Nous travaillons en " mode crise" en essayant de faire trop, trop vite.
Dimension 10	**Soutien de la direction de l'hôpital pour la sécurité des patients**
F1	La direction de l'hôpital crée un climat de travail propice à la sécurité des patients.
F8	Les actions de la direction de l'hôpital montrent que la sécurité des patients est une priorité absolue.
F9	La direction de l'hôpital ne semble s'intéresser à la sécurité des patients qu'après qu'un événement indésirable s'est produit.
Dimension 11	**Travail d'équipe dans les unités de l'hôpital.**
F2	Les unités de l'hôpital ne se coordonnent pas bien.
F4	Il existe une bonne coopération entre les unités de l'hôpital qui ont besoin de travailler ensemble.
F6	Travailler avec le personnel d'autres unités de l'hôpital est souvent désagréable.
F10	Les unités de l'hôpital travaillent bien ensemble pour fournir les meilleurs soins aux patients.

Dimension 12	**Transferts et transitions d'hôpitaux**
F3	Les choses " tombent entre les mailles du filet" lors du transfert de patients d'une unité à une autre.
F5	Des informations importantes sur les soins aux patients sont souvent perdues lors des changements de quart.
F7	Des problèmes se posent souvent lors de l'échange d'informations entre les unités de l'hôpital.
F11	Les changements de quart sont problématiques pour les patients de cet hôpital.

TABLE 4.1: Les quarante deux (42) questions (items) réparties en douze (12) dimensions de la culture de la sécurité mesurées dans le HSOPSC.

Dans le tableau 4.1 sont décrites les douze (12) dimensions de la culture de la sécurité évaluées dans le questionnaire-type HSOPSC. Chaque dimension comporte un certain nombre de questions (items) posées. Ce qui fait un total de quarante deux (42) questions soumises à l'enquête : par exemple pour la dimension 1 les items constituants sont A10, A15, A17 et A18 (Sorra et Nieva, 2004 [91] ; Jones et al, 2008 [95]). Les réponses possibles aux questions sont indiquées dans le questionnaire. Ainsi deux catégories de réponses sont fournies. L'une indique le niveau de l'accord de l'agent par rapport à un item (par exemple A15 : " *La sécurité des patients n'est jamais sacrifiée pour faire plus de travail.* ") décrivant un problème de sécurité des patients. Alors les cinq (05) réponses possibles pour cette catégorie sont données par ordre croissant de l'accord : *pas du tout d'accord (strongly disagree)*, *pas d'accord (disagree)*, *ni d'accord ni en désaccord (neither agree nor disagree)*, *en accord (agree) tout à fait d'accord (strongly agree)*. Ce qui renseignera sur le niveau de l'évaluation de l'item en question et par équivalence de la préoccupation de sécurité pointée. L'autre catégorie indique la fréquence à la quelle les comptes rendus d'évènements (erreurs) sont rapportés. Alors là aussi les cinq (05) réponses possibles associées à cette catégorie sont rangées par ordre d'importance par rapport au niveau de la fréquence de l'évènement. Ainsi nous avons les réponses possibles suivantes : *jamais (never)*, *rarement (rarely)*, *parfois (sometimes)*, *la plupart du temps (most of the time)* et *toujours (always)*. Un exemple d'item de cette catégorie est C3 : " *Nous sommes informés des erreurs qui se produisent dans cette unité.* ".

Parallèlement à cet étiquetage (par rapport aux réponses apportées) des items, le codage de Likert sur une échelle de 1 à 5, a été faite dans le questionnaire en ce qui concerne le niveau de réponse. Ce qui rend dès lors possible le traitement quantitatif des données (réponses). Ce qui justifie l'application de notre approche à données quantitatives discrètes et positives : notre algorithme de reconstruction automatique PGNMF. Pour revenir au codage Likert, à chaque réponse de l'une ou de l'autre catégorie, est associée un certain nombre points désignant le niveau d'évaluation de l'item du point de vue de l'agent répondant. A ce titre les points sont distribués par ordre " d'importance" de l'accord ou de la fréquence. C'est ainsi qu'aux réponses *pas du tout d'accord* et *jamais* sont attribuées respectivement un (01) point ; aux réponses *pas d'accord* et *rarement* deux (02) points respectivement, ainsi de suite jusqu'aux réponses *en accord* et *la plupart du temps* qui ont respectivement cinq (05) points (Sorra et Nieva, 2004 [91]).

4.2.1.3 Procédure de conduite de l'enquête

La procédure de conduite de l'enquête comporte plusieurs étapes à suivre allant de l'identification de la cible aux méthodes de collectes des données en passant par la sélection de l'échantillon.

La cible. Le HSOPSC examine la culture de la sécurité des patients du point de vue du personnel hospitalier. L'enquête peut être complétée par tous les types de personnel hospitalier, des agents de l'entretien ménager aux infirmières et médecins, en passant par la sécurité. L'enquête convient toutefois

mieux aux éléments suivants :

✓ Le personnel hospitalier en contact direct ou en interaction directe avec les patients (personnel clinique, tel que les infirmières, ou le personnel non clinique, tel que les commis d'unité).

✓ Le personnel hospitalier qui peut ne pas avoir de contact direct ni d'interaction avec les patients mais dont le travail affecte directement les soins aux patients (personnel dans des unités telles que pharmacie, laboratoire / pathologie).

✓ Les médecins employés dans les hôpitaux qui passent la plupart de leurs heures de travail à l'hôpital (urgentologues, hospitalistes, pathologistes).

✓ Les superviseurs, gestionnaires et administrateurs d'hôpitaux.

Notez que certains médecins ont des privilèges dans les hôpitaux mais ne sont pas des employés d'hôpitaux et peuvent passer la majorité de leur temps de travail dans des milieux ambulatoires non hospitaliers. Par conséquent, ces types de médecins peuvent ne pas être parfaitement au courant de la culture de sécurité de l'hôpital et ne devraient généralement pas être invités à répondre au sondage. Il convient d'accorder une attention particulière au choix des médecins à inclure ou à exclure du sondage

La sélection de l'échantillon. La population dans laquelle est sélectionnée l'échantillon est le personnel de l'hôpital ou du système hospitalier cible. Les sondages peuvent être administrés à toute la population du personnel hospitalier ou à un sous-ensemble de celle-ci. Bien qu'interroger tout le personnel peut sembler souhaitable, le temps et les ressources supplémentaires nécessaires peuvent éliminer cette option. C'est pourquoi si la direction de l'hôpital prend le choix de n'interroger qu'une partie de son personnel, un certain nombre de règles sont à suivre.

Déterminer les individus à questionner. Tous les membres du personnel de l'hôpital ou du système hospitalier représentent la population. Dans cette population, l'administration peut interroger le personnel de tous les secteurs de l'hôpital ou se concentrer sur des unités spécifiques, des catégories de personnel ou des niveaux de personnel. Il y a diverses façons de sélectionner un échantillon dans une population. Plusieurs types d'échantillons sont décrits ci-dessous. Il s'agit alors de sélectionnez le type qui correspond le mieux à aux besoins de la structure, en tenant compte de ce qui est pratique et aussi des ressources disponibles.

✓ **Personnel dans certaines catégories de personnel.** L'administration peut être intéressé uniquement par le sondage auprès du personnel dans des catégories spécifiques, telles que les soins infirmiers. Avec cette approche, elle peut sélectionner tout le personnel dans une catégorie ou choisir un sous-ensemble du personnel. Cependant, cette approche à elle seule peut ne pas être suffisante pour représenter les points de vue de tous les membres du personnel de l'hôpital.

✓ **Personnel dans des unités particulières.** L'administration peut interroger le personnel de certaines unités, en particulier, comme les urgences, la pharmacie etc.

✓ **Une approche combinée.** Si possible il est recommandé aux enquêteurs d'utiliser une combinaison des deux précédentes approches d'échantillonnage. Par exemple, ils peuvent être intéressé par l'enquête auprès de toutes les infirmières (une catégorie d'effectifs), mais seulement d'un sous-ensemble d'effectifs de chaque unité de l'hôpital (à l'exception des soins infirmiers). L'utilisation d'une combinaison des types d'échantillons permet d'échantillonner de manière sélective certains types de personnel afin de représenter de manière exhaustive la diversité du personnel de l'hôpital.

Déterminer la taille de l'échantillon. La taille de l'échantillon dépendra des personnes que l'hôpital souhaite interroger et des ressources disponibles. Bien que les ressources puissent limiter le nombre de membres du personnel à interroger, plus l'on interroge de personnel, plus l'on a de chances de représenter correctement la population du personnel.
Pour déterminer la taille de l'échantillon, la direction de l'hôpital réfléchira à son budget et au nombre de réponses qu'elle souhaite recevoir (c'est-à-dire son objectif de réponse), parce que tout le monde ne répond pas à l'ensemble des questions posées. Alors elle peut s'attendre à recevoir des sondages complétés d'environ 30 à 50% de son échantillon. Par conséquent, pour atteindre son objectif de réponse, la taille de son échantillon doit être au moins le double du nombre de réponses qu'elle souhaite recevoir.

Par exemple si le nombre de réponses qu'elle souhaite obtenir est 200 réponses complètes, elle devrait administrer des sondages à au moins 400 membres du personnel. Cependant une alternative proposée à cet échantillonnage " surdimensionné ", avec toutes ses implications, est un travail de complétion a posteriori des réponses, tâche proposée dans cette thèse avec la reconstruction automatique du questionnaire. Il s'agit alors d'échantillonner à hauteur du nombre voulu de réponses (dimension normale) et les non-réponses enregistrées seront alors automatiquement complétées par l'algorithme PGNMF (Diatta et al, 2018 [99]).

Compiler la liste complète du personnel. Après avoir déterminé qui doit être interrogé et avoir arrêté la taille de l'échantillon, l'équipe de l'enquête doit compiler une liste du personnel à partir duquel l'échantillon devrait être tiré. Lors de la compilation plusieurs éléments d'information sont à inclure pour chaque membre du personnel, comme le nom et prénom, l'adresse mail interne de l'hôpital ou ceux du bureau ou de la maison pour les questionnaires soumis par mail, le département de compétence, la catégorie de personnel, le titre du poste.
Si l'on doit sélectionnez tous les membres du personnel d'une catégorie d'effectifs, d'un département de l'hôpital en particulier, aucun type d'échantillonnage n'est nécessaire, il faut simplement compiler une liste de tout ce personnel. Par contre l'on sélectionne un échantillon appartenant à une catégorie de personnel ou à une unité particulière de l'hôpital, l'on devrait utiliser une méthode telle qu'un *échantillonnage aléatoire simple* ou un *échantillonnage systématique*.

Déterminer les méthodes de collectes des données. Une fois que les ressources disponibles ont été déterminées, ainsi que la portée du projet de même que la mise en place d'une équipe de projet, et une fois l'échantillon choisi, il reste à présent de savoir comment collecter les données. Les auteurs Sorra et Nieva, 2004 [91] ont préconisé un guide dans les décisions relatives aux méthodes de collecte de données. Les méthodes choisies pour l'envoi et le retour des enquêtes ont une incidence sur la façon dont le personnel perçoit la confidentialité de leurs réponses, ce qui affectera le taux de réponses à l'enquête. Ainsi pour obtenir des taux de réponses maximum parmi tout le personnel de l'hôpital, il est recommandé d'utiliser une méthode de collecte de données sur papier.

Décider comment les questionnaires seront distribués et retournés. Lorsque la direction de l'hôpital décide de la manière dont les enquêtes seront distribuées et renvoyées, elle doit prendre en compte toute expérience antérieure de l'hôpital avec les sondages, en tentant de répondre aux questions suivantes. Des enquêtes antérieures auprès d'hôpitaux ont-elles été postées aux adresses personnelles des employés ou administrées par le système de courrier interne au travail ? Les questionnaires ont-ils été retournés par le biais de personnes de contact, du système de courrier interne, dans des "boîtes de dépôt" de l'hôpital ou par courrier en utilisant des enveloppes de retour pré-affranchies ? Les enquêtes ont-elles été renvoyées à un endroit à l'hôpital ou à un fournisseur extérieur ? Quels étaient les taux de réponse au sondage auprès des employés ? Si possible, il est préférable d'utiliser des méthodes qui ont déjà été utilisées avec succès dans l'hôpital.

✓ **Distribution des questionnaires.** Les questionnaires peuvent être envoyés directement aux adresses personnelles du personnel ou gérés via un système de messagerie interne au travail. Si des enquêtes sont envoyées aux domiciles, l'équipe d'enquête doit vérifier qu'elle dispose des adresses correctes et mises à jour des membres du personnel et qu'elle a pris en compte les frais d'envoi et de retour dans son budget. Si des enquêtes sont administrées auprès du personnel au travail, il est recommandé de donner des instructions explicites et de permettre au personnel de répondre à l'enquête pendant les heures de travail pour souligner le soutien de l'administration de l'hôpital à la collecte de données.

✓ **Retour des questionnaires.** Si le budget de l'hôpital est limité, les sondages remplis peuvent être renvoyés à une personne de contact de l'hôpital désignée par le système de courrier interne ou aux lieux de collecte des sondages au sein de l'hôpital. Cette méthode de retour des sondages peut toutefois susciter des inquiétudes chez les employés quant à la confidentialité de leurs réponses.
Si l'hôpital a peu d'expérience dans l'administration de sondages auprès des employés ou si la direction pense qu'il y a des problèmes de confidentialité, il est préférable de demander aux membres du personnel de poster leurs questionnaires complétés directement à un fournisseur externe ou à une adresse située à l'extérieur de l'hôpital dans des enveloppes de retour pré-affranchies. Si un fournisseur n'est pas utilisé, la direction devrait envisager de renvoyer les questionnaires à l'adresse du siège social afin que le personnel soit assuré que personne dans son hôpital ne verra les réponses remplies.

Établir les points de contact au sein de l'hôpital. Il est recommandé à la direction d'établir des personnes à l'hôpital qui serviront de points de contact pour l'enquête. Celles-ci augmentent la visibilité de l'enquête en montrant leur soutien à l'effort et en aidant à répondre aux questions sur l'enquête. Il faut également décidez combien de points de contact sont nécessaires en tenant compte du nombre d'employés ou d'unités participant à l'enquête. Il est aussi recommandé d'utiliser au moins deux types de points de contact.

✓ **Un point de contact principal de l'hôpital.** Au moins un point de contact principal de l'hôpital doit être désigné dans l'équipe de projet afin que le personnel dispose d'une source centrale pour ses questions ou préoccupations concernant l'enquête. l'équipe de projet est chargée d'inclure les coordonnées du point de contact principal de l'hôpital dans la lettre de notification préalable ou la lettre d'accompagnement de l'enquête envoyée au personnel (c'est-à-dire numéro de téléphone, adresse électronique, numéro de bureau).

✓ **Désigner des points de contact secondaires.** L'équipe de projet peut décider de recruter des points de contact pour chaque unité de l'hôpital, ou catégorie de personnel incluse dans l'échantillon. Un point de contact au niveau d'une unité est chargé de promouvoir et d'administrer l'enquête dans son unité et de rappeler au personnel de l'unité de remplir l'enquête, sans les contraindre en aucune façon. Une lettre d'information décrivant ces tâches et l'ensemble du processus d'enquête doit être envoyée aux contacts potentiels avant de commencer l'administration de l'enquête. Les contacts au niveau de chaque unité se situent généralement au niveau de la direction ou de la supervision, tels que les infirmières gestionnaires, chefs de service ou chefs de quart.

4.2.2 Description du modèle

Dans cette section, nous décrivons notre modèle. Il s'applique à des données discrètes positives donc compatible aux données du codage Likert (valeurs entières de 1 à 5) des réponses au questionnaire de type HSOPSC. Le modèle que nous nommons *Poisson-Gamma* est ici d'usage tout à fait justifié. En effet comme nous l'avons évoqué dans la section 4.1, la vraisemblance de loi de Poisson sur les données X est intrinsèquement liée à la fonction-coût divergence de Kullback-Leibler généralisée de Lee et Seung (2001) [20]. Nous y avons montré que modéliser une approximation matricielle, $X \approx UV$, où $X \geq 0$, $U \geq 0$ et $V \geq 0$, par le biais de cette fonction-coût de Lee et Seung (c'est-à-dire minimiser $F(U,V) = D_{KL}(X\|UV)$), revient à maximiser la vraisemblance associée de loi de Poisson $p(X \mid UV)$. Les lois de probabilités supposées sur les paramètres U et V du modèle montre l'envie d'aller plus loin qu'une simple estimation par maximum de vraisemblance, et de profiter des avantages de la modélisation bayésienne. L'approche des lois a priori conjuguées a permis d'obtenir les lois Gamma comme lois a priori pour U et V.

Supposons que l'on dispose d'un échantillon de n individus interrogés sur p items, dont les réponses sont codées par des entiers positifs (en l'occurrence ici les données du HSOPSC du CHU de Grenoble (France, 2014)). Les réponses au questionnaire sont disposées sous forme d'une matrice, X, de taille $n \times p$. Soit k un entier supérieur à 1 (et inférieur au minimum de n et p). Le modèle probabiliste que nous considérons fait l'hypothèse que les coefficients de la matrice X sont des réalisations de la loi de Poisson, dont la moyenne matricielle est décrite par le produit matriciel UV, où U est une matrice de dimension $n \times k$ et V est une matrice de dimension $k \times p$. Pour l'observation x_{ij}, le modèle génératif est décrit par :

$$x_{ij} \mid U, V \sim p(x_{ij} \mid U, V) = \mathcal{P}\left(x_{ij}\,;\, \sum_{\ell=1}^{k} u_{i\ell} v_{\ell j}\right) \tag{4.13}$$

où $\mathcal{P}(x\,;\,\lambda)$ désigne la loi de Poisson de paramètre λ. Conditionnellement aux matrices U et V, les variables x_{ij} sont indépendantes. La vraisemblance du modèle $L(U,V \mid X)$ s'exprime alors par le biais de la distribution jointe des observations X conditionnellement aux paramètres, $p(X \mid U,V)$.

$$\begin{aligned} L(U,V \mid X) &= p(X \mid U,V) && (4.14)\\ &= \prod_{i=1}^{n}\prod_{j=1}^{p} p(x_{ij} \mid U,V) && (4.15)\\ &= \prod_{i=1}^{n}\prod_{j=1}^{p} \exp\Big(x_{ij}\log[UV]_{ij} - [UV]_{ij} - \log\Gamma(x_{ij}+1)\Big). && (4.16) \end{aligned}$$

Les paramètres du modèle, U, V, sont des variables latentes dont les coefficients sont intra et inter

indépendants et de lois a priori Gamma

$$u_{i\ell} \sim \mathcal{G}(u_{i\ell}\,;\, a^u,\, b^u/a^u)\,, \quad i = 1, \ldots, n\,, \quad \ell = 1, \ldots, k, \tag{4.17}$$

et

$$v_{\ell j} \sim \mathcal{G}(v_{\ell j}\,;\, a^v,\, b^v/a^v)\,, \quad j = 1, \ldots, p\,, \quad \ell = 1, \ldots, k. \tag{4.18}$$

La forme choisie pour la densité des lois Gamma est :

$$\mathcal{G}(x\,;\, a, b) = \exp\left((a-1)\log x - \frac{x}{b} - \log\Gamma(a) - a\log b\right) \tag{4.19}$$

où $a \in \mathbb{R}_+^*$ est le paramètre de forme et $b \in \mathbb{R}_+^*$ est le paramètre d'échelle.

Nous noterons par $\theta = (\theta^u, \theta^v)$ l'ensemble des hyperparamètres a^u, b^u, a^v et b^v, soit $\theta^u = (a^u, b^u)$ et $\theta^v = (a^v, b^v)$, à partir desquels s'expriment les paramètres de formes et d'échelles loi Gamma dans (4.17) et (4.18).

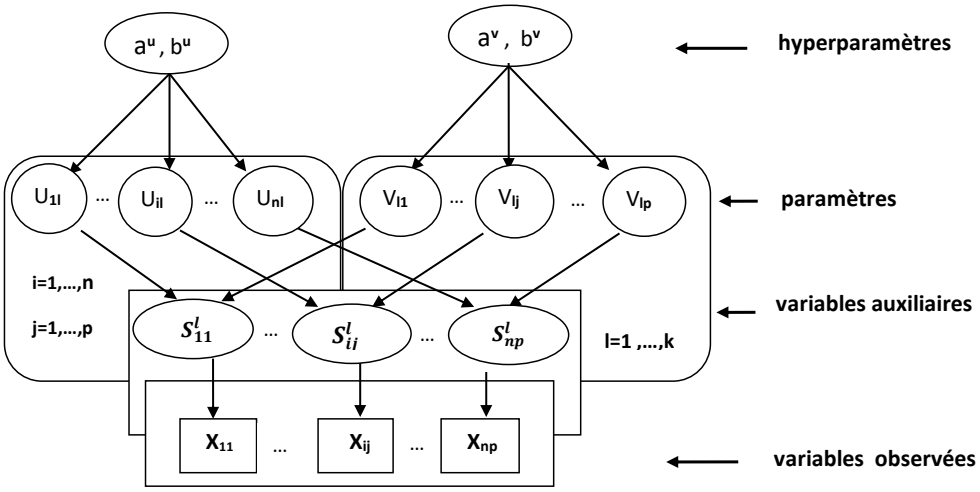

FIGURE 4.1: Diagramme DAG hiérarchique du modèle Poisson-Gamma.

Pour terminer le modèle, nous formalisons ici la prise en compte des données manquantes. Pour cela nous considérons une matrice binaire δ de coefficients δ_{ij} tels que $\delta_{ij} = 0$ si la donnée x_{ij} est manquante et $\delta_{ij} = 1$ si la donnée x_{ij} est observée. Alors la distribution jointe des données observées conditionnellement aux paramètres est donnée par :

$$p(X^{obs} \mid U, V) = \prod_{i=1}^{n}\prod_{j=1}^{p} p(x_{ij} \mid U, V)^{\delta_{ij}}\,. \tag{4.20}$$

Une illustration du modèle bayésien NMF Poisson-Gamma est faite dans la figure 4.1. Il s'agit du diagramme DAG du modèle montrant les liaisons hiérarchiques entre variables. Les variables auxiliaires $S = (S^1, \ldots, S^\ell, \ldots, S^k)$ sont introduites pour nous faciliter le calcul des lois a posteriori conditionnelles (les détails sont donnés dans la section 4.3.1).

4.3 Méthodologie d'imputation sous le modèle NMF Poisson-Gamma

La méthodologie d'imputation que nous proposons repose essentiellement sur l'échantillonneur de Gibbs. Il s'agira alors de compléter des données manquantes artificiellement introduites, dans un jeu de données complet, par un mécanisme *MCAR* que nous décrirons plus loin.

4.3.1 Échantillonneur de Gibbs

L'échantillonneur de Gibbs nécessite la détermination des lois a posteriori conditionnelles des paramètres. Toutefois pour des commodités de calculs nous introduisons dans le modèle des variables auxiliaires $S = (S^1, \ldots, S^\ell, \ldots, S^k)$, où $S^\ell\,, \ell = 1, \ldots, k$ est une matrice de même taille que la matrice des

données X, vérifiant la relation $X = \sum_{\ell=1}^{k} S^{\ell}$ et telles que pour tout coefficient de S^{ℓ}

$$s_{ij}^{\ell} \mid U, V \sim p(s_{ij}^{\ell} \mid U, V) = \mathcal{P}\left(s_{ij}^{\ell} \,;\, u_{i\ell} v_{\ell j}\right) \tag{4.21}$$

Ainsi chaque donnée x_{ij} peut être considérée comme la somme de k réalisations de variables indépendantes de loi de Poisson, c'est-à-dire $x_{ij} = \sum_{\ell=1}^{k} s_{ij}^{\ell}$. L'introduction des variables auxiliaires ne change pas le modèle de factorisation matricielle proposé précédemment mais facilite plutôt le calcul des lois a posteriori conditionnelles. L'inférence s'effectuera maintenant sur le triplet de bloc de variables latentes $(S^{X^{obs}}, U, V)$. Le modèle augmenté s'applique alors au couple $(X^{obs}, S^{X^{obs}})$ et sa loi jointe est donnée par :

$$p(X^{obs}, S^{X^{obs}} \mid U, V) = p(X^{obs} \mid S^{X^{obs}}) p(S^{X^{obs}} \mid U, V) = \prod_{i=1}^{n} \prod_{j=1}^{p} \left(p(x_{ij} \mid s_{ij}^{1:k}) p(s_{ij}^{1:k} \mid u_{i,1:k}, v_{1:k,j})\right)^{\delta_{ij}} \tag{4.22}$$

où $s_{ij}^{1:k} = (s_{ij}^{1}, \ldots, s_{ij}^{\ell}, \ldots, s_{ij}^{k})$, $u_{i,1:k} = (u_{i1}, \ldots, u_{i\ell}, \ldots, u_{ik})$ et $v_{1:k,j} = (v_{1j}, \ldots, v_{\ell j}, \ldots, v_{kj})$.

La loi a posteriori des variables latentes est alors proportionnelle à la loi jointe des données observées et des variables latentes ,

$$\pi(S^{X^{obs}}, U, V \mid X^{obs}, \theta) \propto p(X^{obs}, S^{X^{obs}}, U, V \mid \theta) \tag{4.23}$$

avec cette dernière qui se factorise comme suit :

$$p(X^{obs}, S^{X^{obs}}, U, V \mid \theta) = p(X^{obs} \mid S^{X^{obs}}) p(S^{X^{obs}} \mid U, V) p(U \mid \theta^{u}) p(V \mid \theta^{v}). \tag{4.24}$$

où $p(X^{obs} \mid S^{X^{obs}}) = \prod_{ij} \left[\delta\left(x_{ij} - \sum_{\ell} s_{ij}^{\ell}\right)\right]^{\delta_{ij}}$, avec $\delta(x)$ étant la fonction delta de Kronecker définie par $\delta(x) = 1$ si $x = 0$ et $\delta(x) = 0$ sinon.

Alors la loi conditionnelle a posteriori des variables latentes $S^{X^{obs}}$, correspondant aux données observées, est obtenue à partir des relations de proportionnalités suivantes :

$$p(S^{X^{obs}} \mid X^{obs}, U, V, \theta) \propto p(X^{obs}, S^{X^{obs}}, U, V \mid \theta) \propto p(X^{obs} \mid S^{X^{obs}}) p(S^{X^{obs}} \mid U, V) \tag{4.25}$$

ce qui permet d'obtenir

$$p(S^{X^{obs}} \mid X^{obs}, U, V, \theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{p} \delta\left(x_{ij} - \sum_{\ell=1}^{k} s_{ij}^{\ell}\right)^{\delta_{ij}} \exp\left(\sum_{\ell=1}^{k} \left(\delta_{ij} s_{ij}^{\ell} \log(u_{i\ell} v_{\ell j}) - \delta_{ij} \log \Gamma\left(s_{ij}^{\ell} + 1\right)\right)\right) \tag{4.26}$$

et d'identifier ainsi un produit de lois multinomiales :

$$p(S^{X^{obs}} \mid X^{obs}, U, V, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{p} \mathcal{M}\Big((s_{ij}^{1}, \ldots, s_{ij}^{k}) \,;\, \delta_{ij} x_{ij} \,,\, (p_{ij}^{1}, \ldots, p_{ij}^{k})\Big) \tag{4.27}$$

où les probabilités p_{ij}^{ℓ} sont données par

$$p_{ij}^{\ell} = \frac{u_{i\ell} v_{\ell j}}{\sum_{\ell'=1}^{k} u_{i\ell'} v_{\ell' j}}\,, \quad \ell = 1, \ldots, k.$$

Les lois conditionnelles a posteriori des coefficients du paramètre U sont déterminées à partir des relations de proportionnalité suivantes :

$$p(U \mid X^{obs}, S^{X^{obs}}, V, \theta) \propto p(X^{obs}, S^{X^{obs}}, U, V \mid \theta) \propto p(U \mid \theta^{u}) p(S^{X^{obs}} \mid U, V) \tag{4.28}$$

ce qui donne

$$p(U \mid X^{obs}, S^{X^{obs}}, V, \theta) \propto \prod_{i=1}^{n} \prod_{\ell=1}^{k} \exp\left(-u_{i\ell}\Big(\frac{a^{u}}{b^{u}} + \sum_{j=1}^{p} \delta_{ij} v_{\ell j}\Big) + \Big(a^{u} - 1 + \sum_{j=1}^{p} \delta_{ij} s_{ij}^{\ell}\Big) \log u_{i\ell}\right) \tag{4.29}$$

et permet d'identifier un produit lois gamma :

$$p(U \mid X^{obs}, S^{X^{obs}}, V, \theta) = \prod_{i=1}^{n} \prod_{\ell=1}^{k} \mathcal{G}\Big(u_{i\ell} \,;\, \alpha_{i\ell}^{u} \,,\, \beta_{i\ell}^{u}\Big) \tag{4.30}$$

où $\alpha_{i\ell}^u = a^u + \sum_j \delta_{ij} s_{ij}^\ell \,, \quad \beta_{i\ell}^u = (a^u/b^u + \sum_j \delta_{ij} v_{\ell j})^{-1}$.

De même les lois conditionnelles a posteriori des coefficients du paramètre V sont à partir des relations de proportionnalité suivantes :

$$p(V \mid X^{obs}, S^{X^{obs}}, U, \theta) \propto p(X^{obs}, S^{X^{obs}}, U, V \mid \theta) \propto p(V \mid \theta^v) p(S^{X^{obs}} \mid U, V), \tag{4.31}$$

ce qui donne de façon plus explicite

$$p(V \mid X^{obs}, S^{X^{obs}}, U, \theta) \propto \prod_{\ell=1}^{k} \prod_{j=1}^{p} \exp\left(-v_{\ell j} \Big(\frac{a^v}{b^v} + \sum_{i=1}^{n} \delta_{ij} u_{i\ell} \Big) + \Big(a^v - 1 + \sum_{i=1}^{n} \delta_{ij} s_{ij}^\ell \Big) \log v_{\ell j} \right) \tag{4.32}$$

et aussi d'identifier un produit de lois gamma :

$$p(V \mid X^{obs}, S^{X^{obs}}, U, \theta) = \prod_{\ell=1}^{k} \prod_{j=1}^{p} \mathcal{G}\Big(v_{\ell j} \,;\, \alpha_{\ell j}^v \,,\, \beta_{\ell j}^v \Big) \tag{4.33}$$

où $\alpha_{\ell j}^v = a^v + \sum_i \delta_{ij} s_{ij}^\ell \,, \quad \beta_{\ell j}^v = (a^v/b^v + \sum_i \delta_{ij} u_{i\ell})^{-1}$.

Une fois les lois a posteriori conditionnelles déterminées nous pouvons établir l'échantillonneur de Gibbs qui en découle (algorithme 14).

Données : La matrice de données X et les valeurs des hyper-paramètres a^u, b^u, a^v et b^v.
Résultat : Les estimations $\hat{U}$ et $\hat{V}$ des matrices paramètres U et V.
Initialisation : $u_{i\ell}^{(0)} \sim \mathcal{G}(u_{i\ell} \,;\, a^u, b^u/a^u)$ et $v_{\ell j}^{(0)} \sim \mathcal{G}(v_{\ell j} \,;\, a^v, b^v/a^v)$;
pour *t de* 1 *à* $t_{\max}$ **faire**
  Simulation des variables auxiliaires S ;
  pour *i de 1 à n et j de 1 à p* **faire**
    pour *ℓ de 1 à k* **faire**
      $p_{ij}^{\ell\,(t)} = \frac{u_{i\ell}^{(t-1)} v_{\ell j}^{(t-1)}}{\sum_{\ell'=1}^{k} u_{i\ell'}^{(t-1)} v_{\ell' j}^{(t-1)}}$;
    fin
    $p_{ij}^{1:k\,(t)} = \Big(p_{ij}^{1,(t)}, \ldots, p_{ij}^{k,(t)} \Big)$;
    $s_{ij}^{1:k\,(t)} \sim \mathcal{M}\left(s_{ij}^{1:k} \,;\, \delta_{ij} x_{ij} \,,\, p_{ij}^{1:k\,(t)} \right)$;
  fin
  pour *i de 1 à n et j de 1 à p* **faire**
    pour *ℓ de 1 à k* **faire**
      Simulation des coefficients de U ;
      $\alpha_{i\ell}^{u\,(t)} = a^u + \sum_j \delta_{ij} s_{ij}^{\ell\,(t)}$;
      $\beta_{i\ell}^{u\,(t)} = \left(a^u/b^u + \sum_j \delta_{ij} v_{\ell j}^{(t-1)} \right)^{-1}$;
      $u_{i\ell}^{(t)} \sim \mathcal{G}\left(u_{i\ell} \,;\, \alpha_{i\ell}^{u\,(t)}, \beta_{i\ell}^{u\,(t)} \right)$;
      Simulation des coefficients de V ;
      $\alpha_{\ell j}^{v\,(t)} = a^v + \sum_i \delta_{ij} s_{ij}^{\ell,(t)}$;
      $\beta_{\ell j}^{v\,(t)} = \left(a^v/b^v + \sum_i \delta_{ij} u_{i\ell}^{(t)} \right)^{-1}$;
      $v_{\ell j}^{(t)} \sim \mathcal{G}\left(v_{\ell j} \,;\, \alpha_{\ell j}^{v,(t)}, \beta_{\ell j}^{v,(t)} \right)$;
    fin
  fin
fin

Algorithme 14 : Échantillonneur de Gibbs associé au modèle Poison-Gamma.

Les paramètres inconnus dans cet algorithme sont les hyperparamètres du modèle, a^u, b^u, a^v, b^v. Il convient donc de les estimer avant de l'exécuter.

4.3.2 Estimation des hyperparamètres du modèle

Les hyperparamètres du modèle Poisson-Gamma sont décrits par $\theta = (a^u, b^u, a^v, b^v)$. Dans une optique bayésienne empirique, ceux-ci sont estimés en effectuant une approximation variationnelle de la log-vraisemblance marginale des données observées $\log p(X^{obs} \mid \theta)$, puis en calculant le maximum de la fonction approchée [34, 100, 29, 32, 33]. La méthode que nous utilisons est inspirée d'algorithmes développés dans plusieurs travaux antérieurs, notamment par Cemgil (2009) [29]. Les détails des calculs peuvent être trouvés dans cette référence bibliographique. Nous résumons les étapes principales de la méthode d'estimation des hyperparamètres ci-dessous. Tout d'abord, la log-vraisemblance marginale peut être minorée de la manière suivante :

$$\log p(X^{obs} \,|\, \theta) \geq \mathbb{E}_q \left[\log \frac{p(X^{obs}, S^{X^{obs}}, U, V \,|\theta)}{q(S^{X^{obs}}, U, V \mid X^{obs}, \theta)} \right] \tag{4.34}$$

où l'espérance $\mathbb{E}_q[.]$ est calculée par rapport à une loi instrumentale, $q(S^{X^{obs}}, U, V \,|X^{obs}, \theta)$. Cette dernière est déterminée de sorte à approcher la loi a posteriori au sens de la divergence de Kullback-Leibler. Le membre de droite dans (4.34) est appelée borne variationnelle. L'idée développée dans la méthode variationnelle empirique (Bihop, 2006 [100] ; Cemgil, 2009 [29]) est d'approcher, par le biais de cette borne que nous notons $B_{\mathrm{v}}[q(S^{X^{obs}}, U, V \mid X^{obs}, \theta)]$, la log-vraisemblance marginale. Notons que nous avons l'égalité dans (4.34) lorsque la loi instrumentale $q(S^{X^{obs}}, U, V \mid X^{obs}, \theta)$ est égale à la loi a posteriori $p(S^{X^{obs}}, U, V \mid X^{obs}, \theta)$. L'approximation (par une expression analytique) de cette dernière permet ainsi d'approcher $\log p(X^{obs} \mid \theta)$ par la fonctionnelle $B_{\mathrm{v}}[q(S^{X^{obs}}, U, V \mid X^{obs}, \theta)]$ qui est plus facile à manipuler et sur laquelle des méthodes d'analyse comme la maximisation sont possibles. Maximiser cette quantité comme fonction de θ reviendra aussi à maximiser la log-vraisemblance marginale. L'approximation de la loi a posteriori repose sur un produit de loi marginales

$$q(S^{X^{obs}}, U, V \,|\, X^{obs}, \theta) = q(S^{X^{obs}} \,|\, X^{obs}, \theta) q(U \,|\, X^{obs}, \theta) q(V \,|\, X^{obs}, \theta)\,, \tag{4.35}$$

où les facteurs sont déterminés à partir des équations de mise à jour suivantes :

$$q(S^{X^{obs}}|X^{obs}, \theta)^{(t+1)} \quad \propto \quad \exp\Big(\Big\langle \log P(X^{obs}, S^{X^{obs}}, U, V|\theta) \Big\rangle_{q(U|X^{obs},\theta)^{(t)} q(V|X^{obs},\theta)^{(t)}} \Big) \tag{4.36}$$

$$q(U|X^{obs}, \theta)^{(t+1)} \quad \propto \quad \exp\Big(\Big\langle \log P(X^{obs}, S^{X^{obs}}, U, V|\theta) \Big\rangle_{q(S^{X^{obs}}|X^{obs},\theta)^{(t+1)} q(V|X^{obs},\theta)^{(t)}} \Big) \tag{4.37}$$

$$q(V|X^{obs}, \theta)^{(t+1)} \quad \propto \quad \exp\Big(\Big\langle \log P(X^{obs}, S^{X^{obs}}, U, V|\theta) \Big\rangle_{q(S^{X^{obs}}|X^{obs},\theta)^{(t+1)} q(U|X^{obs},\theta)^{(t+1)}} \Big) \tag{4.38}$$

avec $\langle . \rangle_q$ [1], désignant l'espérance suivant la loi q. Les lois marginales sont déterminées par une procédure itérative aboutissant à une approximation de la loi a posteriori $p(S^{X^{obs}}, U, V \mid X^{obs}, \theta)$. Puis par analogie à l'algorithme EM (Espérance-Maximisation) la fonction $B_{\mathrm{v}}[q(S^{X^{obs}}, U, V \mid X^{obs}, \theta)]$, approchant la log-vraisemblance $\log p(X^{obs} \mid \theta)$ est décomposée en une somme de deux quantités dont celle d'intérêt est l'espérance

$$Q(\theta \,;\, \theta^{(t)}) = \mathbb{E}_{q^{(t)}} \left[\log p(X^{obs}, S^{X^{obs}}, U, V \,|\theta) \right]. \tag{4.39}$$

où $q^{(t)} = q(S^{X^{obs}}, U, V \mid X^{obs}, \theta^{(t)})$. La maximisation de cette espérance en la variable θ et la réévaluation de l'espérance avec le maximum trouvé, de façon itérative engendre une suite $(\theta^{(t)})_t$ qui converge vers un maximum local de la log-vraisemblance marginale $\log p(X^{obs} \mid \theta)$. L'espérance dans (4.39) est alors

1. Parfois pour des commodités d'écriture, nous utiliserons la notation $\langle X \rangle_q$ à la place de celle traditionnelle $\mathbb{E}_q(X)$.

exprimée comme suit :

$$\begin{aligned} Q(\theta;\theta^{(t)}) &= \sum_i \sum_j \delta_{ij} \sum_\ell \Big(- \mathbb{E}_{q^{(t)}}[u_{i\ell}]\mathbb{E}_{q^{(t)}}[v_{\ell j}] + \mathbb{E}_{q^{(t)}}[s_{ij}^\ell]\mathbb{E}_{q^{(t)}}[\log u_{i\ell}] \\ &\quad + \mathbb{E}_{q^{(t)}}[s_{ij}^\ell]\mathbb{E}_{q^{(t)}}[\log v_{\ell j}] - \mathbb{E}_{q^{(t)}}[\log \Gamma(s_{ij}^\ell + 1)]\Big) \\ &\quad + \sum_i \sum_j \delta_{ij} \mathbb{E}_{q^{(t)}}[\log \delta(x_{ij} - \sum_\ell s_{ij}^\ell)] \\ &\quad + \sum_i \sum_\ell \Big((a^u - 1)\mathbb{E}_{q^{(t)}}[\log u_{i\ell}] - \frac{a^u}{b^u}\mathbb{E}_{q^{(t)}}[u_{i\ell}] - \log\Gamma(a^u) - a^u \log \frac{b^u}{a^u} \Big) \\ &\quad + \sum_\ell \sum_j \Big((a^v - 1)\mathbb{E}_{q^{(t)}}[\log v_{\ell j}] - \frac{a^v}{b^v}\mathbb{E}_{q^{(t)}}[v_{\ell j}] - \log\Gamma(a^v) - a^v \log \frac{b^v}{a^v} \Big). \end{aligned}$$

où $\delta(x)$ est la fonction de Kronecker. L'estimation des hyperparamètres consiste alors à résoudre itérativement le problème de maximisation suivant :

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta\,;\,\theta^{(t)}). \tag{4.40}$$

Lors d'expériences préliminaires, nous avons obtenu de meilleurs résultats lorsque les paramètres de forme étaient fixés à des valeurs garantissant la parcimonie (valeurs inférieures à 1, fixées à $a^v = 0.8$ et $a^u = 0.5$ dans nos calculs). Dans ce cas, le travail d'estimation des hyperparamètres se simplifie et se ramène à l'estimation des moyennes, b^u et b^v, des lois a priori. Le problème (4.40) revient alors à mettre à jour b^u et b^v de la manière suivante :

$$b^{u,(t+1)} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{\ell=1}^{k} \mathbb{E}_{q^{(t)}}[u_{i\ell}]\,,$$

et

$$b^{v,(t+1)} = \frac{1}{kp} \sum_{\ell=1}^{k} \sum_{j=1}^{p} \mathbb{E}_{q^{(t)}}[v_{\ell j}]\,.$$

L'algorithme est exécuté avec plusieurs valeurs initiales $b^{u,(0)}$ et $b^{v,(0)}$. Les estimations donnant la valeur la plus élevée de la borne variationnelle sont retenues pour lancer ensuite l'algorithme de Gibbs. Notons que l'approximation variationnelle fournit aussi des estimations des matrices latentes U et V, mais nous avons observé que ces estimations étaient moins bonnes que celles données par l'algorithme de Gibbs.

4.3.3 Algorithme de reconstruction automatique

L'algorithme que nous développons ici, et nommons *PGNMF* (*Poisson-Gamma Nonnegative Matrix Factorization*, est celui de reconstruction automatique de données. Son principe repose, d'abord, sur le retrait de façon délibérée d'une certaine proportion de données d'un jeu complet de nature discrète, suivant un mécanisme ignorable, en l'occurrence ici le mécanisme *MCAR*. Ensuite procéder à sa reconstruction par le biais de l'algorithme qui repose essentiellement sur l'échantillonneur de Gibbs (Algorithme 14). Le retrait de données se fait de façon artificielle, par la considération de la matrice binaire δ évoquée dans la description du modèle. Rappelons que pour celle-ci un coefficient δ_{ij} qui vaut 0 correspond à une donnée x_{ij} qui est manquante, alors que quand il vaut 1 la donnée associée est observée. Alors le mécanisme *MCAR* que nous proposons repose sur une loi de Bernoulli $\mathcal{B}(p)$, avec $p = 1 - \pi$ et $0 < \pi < 1$ qui désignera le pourcentage de données manquantes (c'est-à-dire les non-réponses dans le questionnaire). La matrice indicatrice de réponses/non-réponses δ est alors générée suivant la loi cette loi de Bernoulli. La description de l'algorithme PGNMF est faite ci-dessous (Algorithme 15).

4.4 Résultats

La mise en œuvre pratique de notre algorithme *PGNMF* demande d'observer, pour l'échantillonneur de Gibbs associé, une période de chauffe (*burn-in*) comprenant entre 100 et 300 cycles de mise à jour des paramètres avant d'entrer dans une phase stationnaire. La période de chauffe est suivie de 1500 à 2000 cycles de mise à jour pour calculer les estimateurs (moyennes a posteriori empiriques). La performance

Données : Matrice, X, de données discrètes et positives.

Résultat : Matrice imputée $Impute(\hat{U}\hat{V})$.

1. Lancer l'échantillonneur de Gibbs (Algorithme 14);

2. Observer une période de chauffe (*burn-in*) pour permettre à l'algorithme d'entrer dans sa phase stationnaire ;

3. Puis opérer N cycles de mise à jour en stockant pour chacun des paramètres U et V, les multiples matrices générées, $U^{(1)}, \ldots, U^{(N)}$ et $V^{(1)}, \ldots, V^{(N)}$;

4. Calculer les valeurs estimées et reconstruire la matrice originale X.

- Calculer les estimations des moyennes a posteriori $\hat{U}$ et $\hat{V}$;
- Reconstruction de la matrice X par le produit matriciel, plus précisément $X \approx Impute(\hat{U}\hat{V})$, où $Impute(.)$ est une fonction qui rend une copie de X et où les données manquantes sont complétées par les coefficients correspondant dans la matrice $\hat{U}\hat{V}$;

5. Procédure de « conversion en entiers » des données imputées qui sont très souvent des décimaux :

- Les coefficients décimaux de $Impute(\hat{U}\hat{V})$ sont ré-échantillonnés en leur attribuant l'entier le plus proche proportionnellement à la décimale de la valeur imputée;
- Pour éviter la création de données non-présentes (qui sont du reste en pourcentage très faible) dans le jeu initial, les valeurs en dehors du rang (valeurs entre 1 et 5) sont ramenées aux valeurs minimales et maximales respectivement;

Algorithme 15 : Imputation par PGNMF.

de notre algorithme de reconstruction automatique *PGNMF* est évaluée à travers deux critères, et est comparée à celle plusieurs autres méthodes. Il s'agit des procédures d'imputation précédemment étudiées au chapitre 3 (sect. Etat de l'art) : la méthode espérance-maximisation sous le modèle poissonien de données augmentées *PEM*, la méthode de factorisation matricielle pondérée *ANLS-WNMF*, mais que nous nommerons ici simplement *WNMF*, la procédure *missForest* que nous nommerons *MissF*, l'analyse factorielle *MIMCA* et les méthodes naïves d'imputation. Ces dernières citées sont les méthodes d'imputation par la moyenne (*Moy*), par la médiane (*Med*), par le mode (*Mod*) et la procédure *LOCF*.

4.4.1 Critères d'évaluation

Les deux critères qui permettent d'évaluer les performances des différentes méthodes sont la racine de l'erreur quadratique moyenne, RMSE (root mean square error) et la divergence de Kullback-Leibler, D_{KL}, moyenne.

Le critère RSME mesure la différence entre les valeurs reconstruites et celles observées dans la matrice originale (complète).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Vec}(X)_i - \mathrm{Vec}(\hat{X})_i\right)^2}$$

où $\hat{X} = Impute(\hat{U}\hat{V})$, et $\mathrm{Vec}(X)_i$ et $\mathrm{Vec}(\hat{X})_i$ représentent respectivement les N valeurs retirées de la matrice originale et les N valeurs correspondantes dans la matrice reconstruite. Le critère RMSE est très souvent utilisé pour évaluer les méthodes de reconstruction matricielle en apprentissage automatique lorsqu'il s'agit, par exemple, d'évaluer des systèmes de recommandation. C'est un critère classique utilisé en imputation ou en prédiction de données. Il mesure ici la qualité globale de la reconstruction automatique sur l'ensemble des 42 items.

Le critère Kullback-Leibler D_{KL}, lui, évalue la perte d'information sur chaque variable reconstituée. En effet outre le calcul des scores effectué pour chacune des douze (12) dimensions du HSOPSC (Sorra et Nieva, 2004 [91]) l'analyse des données de l'enquête peut aussi se présenter sous la forme d'un compte-rendu dans lequel les histogrammes des réponses aux questions sont discutées séquentiellement item par item. Ce qui justifie notre intérêt pour cette second critère qui est introduit dans le but de refléter une des pratiques mises en œuvre lors de l'analyse des résultats d'un tel questionnaire en santé publique. Nous utilisons la divergence de Kullback-Leibler pour estimer la perte d'information induite par la reconstruction des histogrammes de réponse à chaque item du questionnaire. Pour un item particulier, la divergence de Kullback-Leibler est définie de la manière suivante

$$D_{KL}(P\|Q) = \sum_{j=1}^{J} p_j \log(p_j/q_j)$$

où P représente la loi empirique des réponses codées dans le questionnaire complet et Q représente la loi empirique des réponses codées dans le questionnaire reconstruit, J étant le nombre de choix possibles. La mesure finale, que nous considérons ici, est la moyenne des valeurs de divergence calculées pour chacun des p items du questionnaire.

La divergence D_{KL} mesure la qualité de reconstruction des lois marginales de la matrice originale, tandis que l'erreur RMSE peut tenir compte des structures de corrélation existant dans la matrice originale.

4.4.2 Résultats sur données simulées

Dans un premier temps, les différentes méthodes sont évaluées sur des données synthétiques, simulées à partir le modèle génératif Poisson-Gamma présenté dans la section 4.2.2 (Eq. (4.13)). L'idée ici est de simuler des données très proches de celles du questionnaire HSOPSC à partir de notre modèle génératif. Cependant vouloir obtenir des entiers exclusivement entre 1 et 5 a été difficile pour nous. C'est pourquoi nous avons trouvé un compromis jusqu'au rang 7, c'est-à-dire des entiers entre 1 et 7, avec un rejet des 0. Le rejet des 0 introduit donc un léger biais qui se apparaitra dans les résultats. Ainsi les coefficients des matrices U et V ont été tirés aléatoirement et indépendamment suivant les lois Gamma $\mathcal{G}(u\,;\,65, 0.8/65)$ et $\mathcal{G}(v\,;\,50, 0.7/50)$ respectivement, avec un nombre de facteurs fixé à $k = 2$. Finalement les données synthétiques sont supposées tirées sur un échantillon de $n = 1500$ individus sur lesquels l'intérêt a porté sur $p = 30$ variables. Le travail de reconstruction est effectué sur une série de huit matrices dont les données sont artificiellement retirées suivant un pourcentage π de 10% à 80%, sous le mécanisme *MCAR* déjà défini. Toutefois une légère contrainte veillant à garder, pour chaque variable, au moins une valeur non nulle a été introduite. Donc l'hypothèse qu'un même item n'a été répondu par tous les individus est ici à écarter.

	10%	20%	30%	40%	50%	60%	70%	80%
locf	1.2130	1.2185	1.2186	1.2037	1.2139	1.2080	1.2016	1.2082
med	1.0559	1.0515	1.0526	1.0576	1.0499	1.0636	1.0553	1.0548
missf	1.0478	1.0483	1.0513	1.0834	1.1086	1.1306	1.1644	1.2101
pem	0.9761	0.9805	0.9781	0.9837	0.9843	0.9945	1.0023	1.0241
pgmnf	0.9780	0.9791	0.9778	0.9821	**0.9778**	**0.9838**	**0.9818**	**0.9889**
wnmf	0.9969	0.9964	0.9983	1.0043	1.0135	1.0247	1.0424	1.0592
mimca*	**0.9726**	**0.9733**	**0.9753**	**0.9819**	0.9835	0.9987	-	-

TABLE 4.2: Données simulées à partir du modèle Poisson-Gamma. Evaluation de l'erreur de reconstruction (RMSE) en fonction du taux de données supprimées. *L'algorithme n'a pas convergé pour les taux 70-80%. Les valeurs pour les taux 10-60% ont été obtenues en augmentant la valeur du paramètre de régularisation à 100.

Pour les matrices simulées selon le modèle Poisson-Gamma, les valeurs de RMSE varient entre 0.973 (*MIMCA*) pour un taux de suppression de 10% et 1.210 (*MissF*) pour un taux de suppression de 80% (Table 4.2). Comme attendu, les méthodes naïves présentent les performances les plus faibles. Les algorithmes reposant sur la factorisation matricielle (*PEM*, *PGNMF*, *WNMF* et *MIMCA*) obtiennent les

valeurs de RMSE les plus basses. Une faible variabilité des erreurs de reconstruction est observée pour l'ensemble de ces méthodes. La méthode *MIMCA* obtient les meilleures performances pour des taux de suppression d'items situés entre 10% et 40% avec des valeurs de RMSE situées entre 0.9726 et 0.9819. L'algorithme *PGNMF* obtient les meilleurs résultats lorsque le taux de suppression est plus important. Pour des taux compris entre 50% et 80%, les valeurs de RMSE pour *PGNMF* se situent entre 0.9778 et 0.9889 (Table 4.2).

Concernant la divergence de Kullback-Leibler, les valeurs moyennées sur les 30 items varient entre 0.002 (*PGNMF*) pour un taux de suppression de 10% et 0.489 (*Med*) pour un taux de suppression de 80% (Table 4.3). À nouveau, les méthodes naïves présentent des performances plus faibles que les méthodes matricielles. La méthode *PGNMF* obtient les meilleures performances pour des taux de suppression situés entre 10% et 30% avec des valeurs situées entre 0.002 et 0.012. L'algorithme *MissF* obtient les meilleurs résultats lorsque le taux de suppression est plus important. Pour des taux compris entre 40% et 80%, les valeurs de divergence pour *MissF* se situent entre 0.018 et 0.057 (Table 4.3). Relativement aux méthodes de reconstruction précédentes, les résultats des simulations mettent en évidence les bonnes propriétés de reconstruction matricielle de l'algorithme *PGNMF* et valident son implémentation numérique.

	10%	20%	30%	40%	50%	60%	70%	80%
moy	0.0039	0.0136	0.0273	0.0525	0.0790	0.1181	0.1707	0.2450
med	0.0051	0.0189	0.0411	0.0789	0.1249	0.1912	0.3038	0.4891
missf	0.0028	0.0091	0.0167	**0.0183**	**0.0179**	**0.0203**	**0.0339**	**0.0570**
pem	0.0038	0.0128	0.0257	0.0484	0.0675	0.0934	0.1186	0.1313
pgmnf	**0.0018**	**0.0060**	**0.0117**	0.0244	0.0348	0.0581	0.0942	0.1589
wnmf	0.0032	0.0106	0.0198	0.0373	0.0471	0.0640	0.0698	0.0702
mimca*	0.0040	0.0137	0.0274	0.0539	0.0866	0.1227	-	-

TABLE 4.3: Données simulées à partir du modèle Poisson-Gamma. Evaluation de la divergence de Kullback-Leibler en fonction du taux de données supprimées. *L'algorithme n'a pas convergé pour les taux 70-80%. Les valeurs pour les taux 10-60% ont été obtenues en augmentant la valeur du paramètre de régularisation à 100.

4.4.3 Résultats sur données réelles : questionnaire HSOPSC

Dans un second temps, nous avons évalué la possibilité de reconstruire automatiquement des questionnaires réduits obtenus à partir de l'enquête HSOPSC effectuée entre avril 2013 et septembre 2014 à l'hôpital universitaire de Grenoble. Une analyse par ACM (Analyse des Correspondances Multiples) des données du questionnaire a tout d'abord été effectuée après une imputation des données manquantes utilisant la méthode de Josse et Husson (2016) [42] (Figure 4.2). Le diagramme des variances a permis de mettre en évidence trois axes principaux dans les données. Dans la suite de l'étude de simulation, nous avons utilisé $k = 3$ facteurs pour l'ensemble des méthodes reposant sur la factorisation matricielle (méthodes *PEM, PGNMF, WNMF, MIMCA*). Afin de quantifier la perte d'information liée à la reconstruction automatique du questionnaire, nous avons calculé la corrélation entre les premiers axes principaux de la matrice disjonctive originale et ceux des tableaux de probabilités des matrices reconstruites à partir de données masquées (Josse et Husson, 2016 [42]). Les valeurs obtenues pour des répliques par bootstrap du jeu de données montrent que la variabilité est faible. Pour un questionnaire de type HSOPSC, les résultats indiquent qu'il est possible de restituer une part importante de l'information manquante lorsque de grands nombres d'items sont aléatoirement supprimés. Ce résultat justifie les valeurs élevées des taux de suppression considérés dans le reste de notre étude.

Selon le critère RMSE, les valeurs de performance varient entre 0.950 (*MissF*) pour un taux de suppression de 10% et 1.3354 (*LOCF*) pour un taux de suppression de 80% (Table 4.4). Les méthodes naïves présentent des performances plus faibles que les méthodes factorielles. La méthode *WNMF* obtient les meilleures performances pour des taux de suppression situés entre 20% et 30% (valeurs de RMSE situées entre 0.9737 et 0.9829). L'algorithme *PGNMF* obtient les meilleurs résultats lorsque le taux de

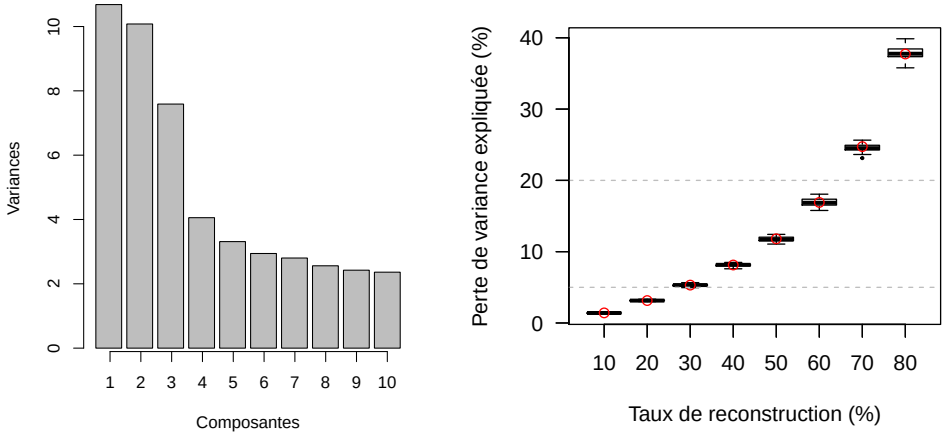

FIGURE 4.2: Gauche : Variances correspondant aux axes principaux de l'analyse des correspondances multiples des données du questionnaire HSOPSC. Droite : Perte de variance expliquée par les 3 axes principaux lorsque des données sont masquées. Les lignes pointillées correspondent aux pourcentages 5% et 20%. Les boîtes correspondent aux valeurs obtenues par bootstrap à partir du jeu de données original.

suppression est supérieur ou égal à 40% (valeur de RMSE situées entre 0.9866 et 1.0220, Table 4.4). Notons la proximité des résultats des algorithmes *PGNMF* et *PEM* (Figure 4.3) qui traduit la " similarité " des deux modélisations sous-jacentes. Rappelons que ces dernières s'appliquent sur la même vraisemblance de loi de Poisson.

	10%	20%	30%	40%	50%	60%	70%	80%
locf	1.3117	1.3066	1.3083	1.3057	1.3076	1.3144	1.3268	1.3354
moy	1.0522	1.0443	1.0494	1.0475	1.0508	1.0477	1.0483	1.0482
med	1.0073	1.0136	1.0180	1.0208	1.0239	1.0185	1.0171	1.0220
mod	1.0902	1.0993	1.1063	1.0984	1.1098	1.1037	1.0982	1.0856
missf	**0.9500**	0.9819	1.0090	1.0223	1.0551	1.0971	1.1374	1.2034
pem	0.9750	0.9819	0.9881	0.9874	0.9918	0.9994	1.0094	1.0343
pgmnf	0.9746	0.9815	0.9874	**0.9866**	**0.9908**	**0.9973**	**1.0052**	**1.0220**
wnmf	0.9607	**0.9737**	**0.9829**	0.9891	1.0009	1.0272	1.0586	1.1234
mimca	0.9807	0.9880	0.9956	0.9974	1.0001	1.0031	1.0093	1.0255

TABLE 4.4: Questionnaire HSOPSC. Erreur de reconstruction RMSE en fonction du taux de données supprimées.

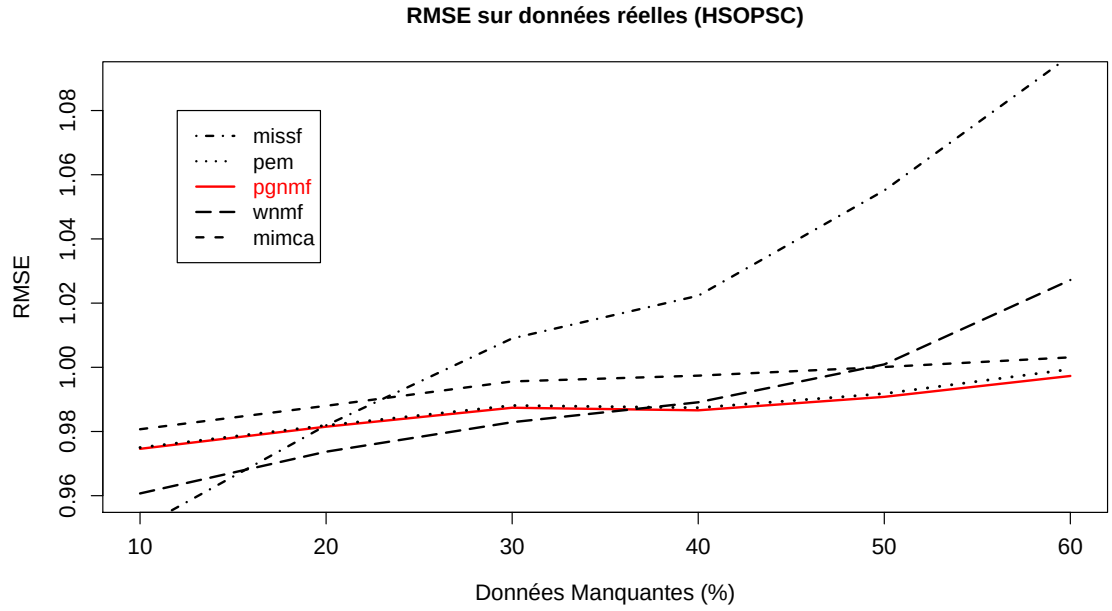

FIGURE 4.3: Questionnaire HSOPSC. Représentation graphique de la performance des méthodes de reconstruction automatique (RMSE) en fonction du taux de données supprimées.

Selon le critère de divergence de Kullback-Leibler, les valeurs de performance varient entre 0.0003 (*MissF*) pour un taux de suppression de 10% et 0.6964 (*Med*) pour un taux de suppression de 80% (Table 4.5). La méthode *MissF* et *WNMF* obtient les meilleures performances pour l'ensemble des taux de suppression (valeurs de divergence situées entre 0.0003 et 0.0214). L'algorithme *PGNMF* obtient des résultats presque similaires à ceux de PEM mais moins bonnes que ceux des autres méthodes factorielles *WNMF* et *MIMCA*. Cependant lorsque le taux de suppression est faible (entre 10% et 30%) les écart entre les méthodes factorielles sont minimes (Figure 4.4).

	10%	20%	30%	40%	50%	60%	70%	80%
locf	0.0024	0.0075	0.0143	0.0228	0.0321	0.0432	0.0550	0.0693
moy	0.0046	0.0174	0.0404	0.0722	0.1163	0.1776	0.2656	0.4107
med	0.0081	0.0313	0.0714	0.1262	0.2022	0.3086	0.4641	0.6964
mod	0.0075	0.0292	0.0660	0.1169	0.1930	0.2951	0.4293	0.6659
missf	**0.0003**	**0.0012**	**0.0026**	**0.0044**	**0.0061**	**0.0089**	**0.0131**	**0.0214**
pem	0.0021	0.0085	0.0197	0.0340	0.0549	0.0764	0.1014	0.1228
pgmnf	0.0021	0.0085	0.0196	0.0337	0.0548	0.0763	0.1015	0.1266
wnmf	0.0015	0.0057	0.0126	0.0217	0.0331	0.0405	0.0492	0.0473
mimca	0.0017	0.0068	0.0155	0.0269	0.0441	0.0635	0.0880	0.1143

TABLE 4.5: Questionnaire HSOPSC. Divergence de Kullback-Leibler en fonction du taux de données supprimées.

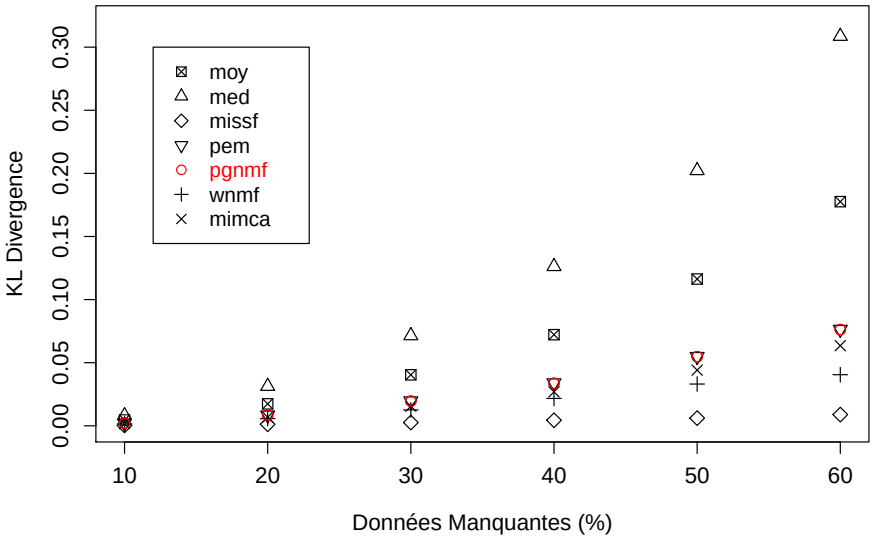

FIGURE 4.4: Questionnaire HSOPSC. Représentation graphique de la performance des méthodes de reconstruction automatique (Divergence de Kullback-Leibler) en fonction du taux de données supprimées.

4.4.4 Résultats sur répliques boostrap

Nous avons effectué cinq (05) répliques boostrap à partir des réponses du questionnaire HSOPSC original pour compléter notre étude. L'idée ici a été de créer cinq autres questionnaires virtuels HSOPSC plus ou moins similaires à l'original afin de voir le comportement notre algorithme *PGNMF* par rapport aux autres. Il s'agissait alors de voir si les tendances dans le HSOPSC original se confirmeraient ou non. Ainsi par le biais du critère RMSE nous avons comparé sur chacune des répliques les erreurs de reconstruction des méthodes les plus performantes. Alors notre algorithme *PGNMF* est comparé aux autres méthodes factorielles, *PEM*, *WNMF* et *MIMCA*, et à la méthode d'apprentissage automatique *MissF*.

Les résultats obtenus ont montré que l'algorithme *PGNMF* a donné les meilleures performances RMSE pour l'ensemble des huit (08) matrices de données manquantes construites à partir de chacune des répliques boostrap. Remarquons également comme l'on pouvait s'y attendre la proximité de ses résultats avec la procédure *PEM*. Nous donnons ici les résultats pour deux (02) répliques boostrap seulement, les résultats complets étant donnés en annexe.

Boostrap 1

	10%	20%	30%	40%	50%	60%	70%	80%
missf	1.2336	1.2316	1.2332	1.2424	1.2733	1.3101	1.3474	1.3865
pem	1.1208	1.1195	1.1265	1.1339	1.1373	1.1460	1.1582	1.1869
pgmnf	**1.1211**	**1.1196**	**1.1259**	**1.1329**	**1.1354**	**1.1431**	**1.1524**	**1.1726**
wnmf	1.1563	1.1610	1.1712	1.1853	1.2002	1.2281	1.2652	1.3395
mimca	1.4672	1.4658	1.4689	1.4688	1.4699	1.4718	1.4784	1.4888

TABLE 4.6: Performances de différentes méthodes d'imputation par le biais du RMSE en fonction du taux de non-réponses sur la réplique boostrap 1.

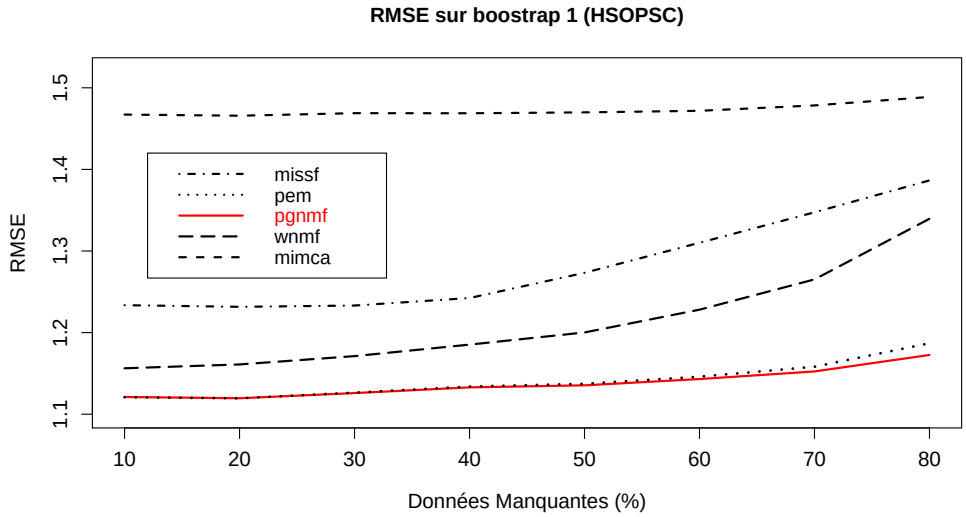

FIGURE 4.5: Performances suivant le critère RMSE sur la réplique boostrap 1.

Boostrap 2

	10%	20%	30%	40%	50%	60%	70%	80%
missf	1.2361	1.2395	1.2411	1.2568	1.2758	1.2930	1.3370	1.3887
pem	1.1232	1.1264	1.1278	1.1316	1.1369	1.1482	1.1615	1.1894
pgmnf	**1.1228**	**1.1257**	**1.1270**	**1.1307**	**1.1353**	**1.1451**	**1.1556**	**1.1748**
wnmf	1.1594	1.1655	1.1761	1.1881	1.2028	1.2310	1.2718	1.3434
mimca	1.4720	1.4750	1.4711	1.4736	1.4740	1.4751	1.4822	1.4899

TABLE 4.7: Performances de différentes méthodes d'imputation par le biais du RMSE en fonction du taux de non-réponses sur la réplique boostrap 2.

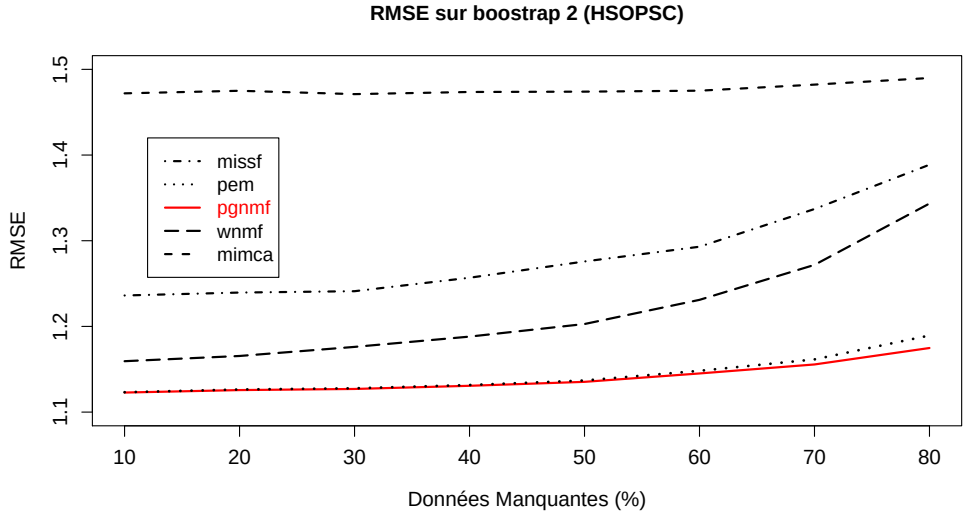

FIGURE 4.6: Performances suivant le critère RMSE sur la réplique boostrap 2.

Conclusion

En utilisant les résultats d'une enquête HSOPSC réalisée au centre hospitalier universitaire de Grenoble auprès de travailleurs médicaux, nous avons envisagé de réduire aléatoirement le questionnaire destiné à évaluer la culture de la sécurité des patients afin d'augmenter son acceptabilité auprès du personnel hospitalier. Nous avons estimé qu'une part importante de l'information contenue dans le questionnaire était redondante et pouvait être restituée automatiquement. Nous avons délibérément masqué une partie des réponses des participants, et comparé notre algorithme à plusieurs autres méthodes d'imputation de données manquantes pour reconstituer l'information masquée. Pour y parvenir, nous avons décrit un modèle bayésien de factorisation matricielle pour l'imputation des données discrètes. Concernant l'estimation des paramètres de ce modèle, nous avons proposé un nouvel algorithme d'échantillonnage de Gibbs,précédé d'une implémentation d'une approche variationnelle pour les hyperparamètres.

Le chapitre a ainsi été articulé autour de quatre points majeurs.
D'abord nous avons montré comment le modèle proposé était de type NMF en donnant sont lien intrinsèque avec la fonction-coût divergence de Kullback-Leibler de la factorisation classique de Lee et Seung (2001).
Ensuite nous avons décrit nos données d'application qui sont issues d'un questionnaire HSOPSC (CHU de Grenoble, France). Nous avons également donné la description de notre modèle bayésien NMF Poisson-Gamma, sur lequel s'appuie l'approche que nous avons proposée.
Puis nous avons donné la méthodologie d'imputation sous ce modèle, en trois points. Dans un premier temps nous avons spécifié l'échantillonneur de Gibbs associé au modèle, après après avoir bien identifié les lois conditionnelles a posteriori. Dans un second temps, nous avons présenté une méthodologie d'estimation des hyperparamètres basées sur une approche variationnelle. Dans troisième temps, nous avons décrit notre algorithme de reconstruction automatique.
Enfin les performances de notre algorithme *PGNMF* ont été données en comparaison à d'autres méthodes de factorisation matricielle, d'apprentissage automatique ou encore des méthodes naïves. Les résultats de ces évaluations ont été faites par le biais deux critères : le RMSE et la Divergence KL.

Concernant l'enquête HSOPSC de Grenoble, nous avons obtenu les résultats suivants. Les méthodes naïves de reconstruction automatique (imputation par la moyenne (*moy*), la médiane (*med*), le mode (*mod*), la méthode *locf*) ont globalement donné les performances les plus basses par rapport aux deux mesures étudiées (RMSE et KL). Nous proposons donc leur non utilisation pour les tâches de complétion de données dans les questionnaires HSOPSC. A l'opposé les méthodes factorielles (*PEM*, *PGNMF*, *WNMF*, *MIMCA*) et d'apprentissage automatique (*MissF*) ont présenté dans l'ensemble les performances les plus élevées. Pour la reconstruction des histogrammes de réponses aux items, *MissF* a donné les meilleurs résultats. Les méthodes factorielles l'ont cependant remporté sur la méthode d'apprentissage automatique en ce qui concerne le critère RMSE. Notre algorithme *PGNMF* a alors donné les meilleurs résultats surtout lorsque le taux de données masquées est relativement élevé. Mieux qu'une tendance confirmée les résultats RMSE sur cinq (05) répliques boostrap ont placé *PGNMF* comme meilleure procédure de reconstruction automatique non seulement sur le questionnaire HSOPSC de Grenoble mais aussi des enquêtes HSOPSC virtuelles similaires. Le premier fait marquant est que les méthodes factorielles et celle d'apprentissage automatique produisent des résultats satisfaisants, supérieurs aux méthodes naïves traditionnellement utilisées dans ce domaine. Le second fait marquant est distinction des algorithmes *MissF* et *PGNMF*. Alors que le premier est plus performant en reconstruction des histogrammes marginales (critère KL), le second lui l'emporte sur la mesure globale de l'erreur de reconstruction (critère RMSE). Pour une tâche de reconstruction automatique ou même d'imputation de données d'un questionnaire HSOPSC il parait donc souhaitable de les combiner.

En définitive, les méthodes de factorisation matricielles et la méthode d'apprentissage automatique *MissF* permettent de reconstruire de grandes quantités de données supprimées de manière efficace. Leur utilisation lors d'enquêtes médicales portant sur la culture de sécurité des patients permettrait d'envisager une gestion optimisée des questionnaires hospitaliers. En effet, nos résultats suggèrent que des enquêtes médicales similaires à celle effectuée à Grenoble pourraient être réalisées en réduisant substantiellement le nombre de questions posées à chaque travailleur médical avec une perte limitée des interprétations de l'enquête.

Conclusion générale

Cette thèse a traité d'un aspect très actuel de la large problématique des données manquantes. Il s'agit des non-réponses dans les questionnaires d'enquêtes, en l'occurrence le questionnaire médical HSOPSC. L'objectif général de notre travail a été de voir comment le questionnaire HSOPSC du centre hospitalier universitaire de Grenoble presque complet (environ 1.8% de données manquantes originelles) pourrait être automatiquement reconstruit sans perte majeure de l'information, après avoir été délibérément sous-échantillonné de manière aléatoire. Ses motivations ont résidé à deux niveaux : (1) d'abord répondre efficacement à un problème de non-réponses qui se poserait pour d'autres questionnaires du même type, vu la nature et la grande taille du questionnaire, (2) ensuite, d'un côté, anticiper sur un problème de non-réponses sur un questionnaire HSOPSC donné, d'un autre, répondre à d'éventuelles contraintes de coûts et/ou de durées des enquêtes. En effet la première motivation a consisté à pouvoir traiter en aval et de manière assez efficiente des non-réponses qui pourraient être dues à un problème d'acceptabilité, de la part de la cible, qui est surtout lié au volume du questionnaire (42 items). Ce problème pourrait alors conduire à un fort taux de données manquantes non voulues par la structure de santé publique ayant soumis l'enquête à son personnel. La seconde motivation a été d'une part de proposer une anticipation de résolution d'un très probable problème de non-réponses. Il s'est agit alors d'une proposition d'augmentation de l'acceptabilité du questionnaire par sa réduction et sa spécification individuelle. D'autre part elle a été de proposer une alternative à d'éventuels problèmes de coûts liés à la collecte de données et/ou à des durées d'enquêtes relativement longues (19 mois pour le questionnaire HSOPSC de France, Grenoble). Ainsi les préoccupations qui sous-tendent cette seconde motivation de notre travail ont pour point commun le sous-échantillonnage en amont du questionnaire. En effet, la structure en santé publique qui s'auto-administrerait un questionnaire type HSOPSC, pourrait le réduire tout en l'individualisant aléatoirement avant de le soumettre à son personnel. Alors la collecte de données sur le HSOPSC ainsi sous-dimensionné, verrait son coût et sa durée réduits proportionnellement au taux de retrait des items. Les non-réponses seront alors en majorité dues aux questions non soumises.
L'algorithme *PGNMF* que nous avons développé pourrait alors être utilisé, soit pour effectuer une tâche d'*imputation* dans le cas général, soit un travail de *prédiction*, dans le cas particulier du questionnaire sous-échantillonné, des réponses qui auraient été données si le questionnaire avait été entièrement soumis. Dans cette tâche de reconstruction automatique de l'algorithme *PGNMF* l'aspect auquel nous nous sommes intéressé a été la mesure de la perte d'information, en ce sens que la valeur imputée n'est pas toujours égale à la vraie valeur manquante ou retirée. Deux critères ont alors été retenus : la mesure d'erreur RMSE et la divergence de Kullback-Leibler.

Pour mener à bien ce travail de reconstruction automatique, nous avons choisi une approche de factorisation matricielle NMF dans un cadre de modélisation bayésienne. Le modèle bayésien NMF Poisson-Gamma que nous avons proposé, à notre connaissance, n'avait été appliqué que pour la reconstruction d'images de faces humaines dégradées suivant un axe vertical, (Cemgil, 2009 [29]). Son usage pour une tâche d'imputation de questionnaire, spécifiquement de reconstruction automatique de questionnaire médical HSOPSC, n'a jamais été constaté auparavant. Ce qui, avec l'algorithme de reconstruction proposé, a constitué la particularité de notre travail, parmi la multiplicité et la diversité des modèles NMF et de leurs applications.

Cette thèse a été articulée autour de quatre chapitres répartis en deux parties. La première a constitué les *préliminaires* et contenu les deux premiers chapitres visant familiariser dans un premier temps le lecteur avec la modélisation bayésienne d'un côté et la théorie sur la factorisation de matrices positives NMF. La seconde a constitué les deux derniers chapitres, dont le troisième a étudié la *problématique* des données manquantes issue de questionnaires ainsi que l'*état de l'art* sur la question. Enfin le quatrième chapitre considéré comme la *méthodologie* de résolution du problème posé, a décrit notre modèle NMF Poisson-Gamma, puis a donné les différentes étapes de la conception, la mise en œuvre et les résultats de performance de notre algorithme de reconstruction automatique *PGNMF*. Plus en détails le déroulement des principales étapes de notre travail s'est effectué comme suit :

Le chapitre 1 a étudié quelques aspects principaux théoriques de la modélisation bayésienne. Il a été

structuré en trois points majeurs. D'abord le paradigme bayésien s'appuyant sur une modélisation probabiliste dans un cadre d'analyse statistique, décrit par une inférence bayésienne dont les spécificités sont (1) la considération d'informations antérieurement acquises sous une modélisation par des lois a priori, (2) l'actualisation de ces informations conditionnellement aux observations : calcul des loi a posteriori. Ensuite la théorie de la décision dans le contexte de l'approche bayésienne a été donné ; la décision ici consiste en une action sur le(s) paramètre(s) du modèle. Cette action désigne, en fait, à l'estimation des paramètres et dont la procédure optimale est déterminée à l'aide d'une fonction de coût. Enfin les méthodes du calcul bayésien ont été données, notamment les méthodes Monte Carlo par Chaînes de Markov (MCMC) en l'occurrence la méthodologie de l'échantillonneur de Gibbs.

Le chapitre 2 a étudié la théorie quelques aspects pratiques de la factorisation de matrices positives. Il a été articulé principalement en points. Dans un premier temps les algorithmes NMF classiques ont été donnés, puis quelques variantes ont été étudié dans un second temps. En effet nous avons donné les trois principales classes d'algorithmes classiques. La première désigne les algorithmes à règles de mises à jour multiplicatives, *MU* (*Multiplicative Update*). Alors la formalisation du problème NMF a évoqué des aspects historiques donnés par les précurseurs Lee et Seung (1999 et 2001) [19, 20], comme par exemple la justification de la contrainte de positivité, par le fait que l'appréhension que l'on a d'une entité découle de l'assemblage (ou sommation) de la conception que l'on a de ses différentes parties. La détermination du couple de matrices positives dont le produit approxime une matrice de données positives a reposé sur la considération de deux fonction-coûts classiques : la divergence de Kullback-Leibler généralisée et le carré de la distance euclidienne. Pour finir les algorithmes et résultats de convergence ont été donnés. La seconde classe est constituée des méthodes du gradient descente dont les algorithmes NMF découlant se sont inspirés des procédures du gradient de descente appliquées sur une fonction différentiable définie sur un espace hilbertien et à valeurs réelles. La dernière classe a été celle des algorithmes des moindres carrés alternés, décrites par deux sous-classes : les algorithmes *ALS* et *ANLS*. Dans un second temps nous avons étudié quelques variantes des algorithmes, en l'occurrence les algorithmes de factorisations pondérées. Nous nous sommes intéressés à celles découlant des mises à jour *MU* et des algorithmes *ANLS*. L'intérêt porté sur ce type de variantes, est que la factorisation pondérée permet le traitement de données manquantes qui a été abordé au chapitre suivant.

Le chapitre 3 a été consacré à l'étude de la problématique du sujet traité dans cette thèse, qui est celle des données manquantes ou plus précisément celle des non-réponses dans les questionnaires, et à l'état de l'art sur la question. Il a été agencé en deux axes majeurs. Le premier a consisté à l'étude de la typologie des données manquantes. Celle-ci est caractérisée à deux niveaux : d'abord par la configuration des données manquantes donnant ainsi les trois structures étudiées (univariées, monotones et arbitraires), puis par la nature du mécanisme de génération des données manquantes. Ce dernier type de caractérisation de la typologie a été la formalisation des processus à l'origine de l'absence de données. Ainsi deux classes de mécanismes sous-jacents se sont distinguées : les mécanismes *ignorables* et ceux *non-ignorables*. Les premiers nommés sont constitués des mécanismes *MAR* selon lesquels les données manquantes le sont aléatoirement alors que les données observées ne le sont pas aléatoirement, et *MCAR* selon lesquels les données manquantes le sont aléatoirement et les données observées le sont aléatoirement. Les seconds nommés sont spécifiés par les mécanismes de type *MNAR*, selon lesquels les données manquantes ne le sont pas aléatoirement et les données observées également ne le sont pas aléatoirement. Le second axe a consisté à l'étude de l'état de l'art. Deux types d'études ont été notés dans la littérature. Il s'agit des analyses sans imputation d'une part et des méthodes d'imputation d'autre part. S'agissant des premières elles sont décrites par les inférences, en présence de données manquantes, sur des paramètres de modèles statistiques, et par les analyses avec suppression de données. Les secondes méthodes sont celles qui consistent à remplacer les données manquantes par des valeur d'estimation, suivant un principe ou un modèle donné. L'accent a été mis sur celles-ci, et particulièrement sur des méthodes a majorité basées sur l'analyse factorielle, en l'occurrence la factorisation de matricielle, mais aussi sur une méthode d'apprentissage automatique s'appuyant sur la technique de régression, et de quelques méthodes simples dites naïves. Les différents algorithmes résultant ont été comparé à la procédure *PGNMF* que nous avons développée au chapitre suivant.

Enfin le chapitre 4 a développé notre méthodologie dédiée à la résolution d'un aspect de la large problématique des données manquantes. Il s'est agit de la reconstruction automatique de non-réponses dans un questionnaire de type HSOPSC. Ce chapitre a été structuré en quatre axes majeurs. D'abord la justification du modèle de loi de Poisson a été faite en montrant, avec des aspects historiques, son lien intrinsèque, avec la fonction-coût divergence de Kullback-Leibler généralisée de Lee et Seung (1999 et 2001) [19, 20]. Ensuite les données d'application ont été présentées. Il s'agit de données issues d'un questionnaire de type HSOPSC dont la description a été faite. En effet les données du questionnaire, ont été recueillies en France au centre hospitalier universitaire de Grenoble en avril 2013 et septembre 2014, soit une durée de 19 mois environ. Le questionnaire est constitué de 42 items soumis à un per-

sonnel de 3888 travailleurs. La technique de codage utilisée à été celle de Likert, avec une échelle à cinq niveaux donnant ainsi cinq types de réponses possibles codées en valeurs entières de 1 à 5. Le modèle Poisson-Gamma de factorisation NMF $X \approx UV$ a également été décrit où la loi de Poisson a été supposée sur les données X alors que des distributions Gamma ont été imposées aux facteurs U et V, la stratégie des lois a priori conjuguées (cf Chap. 1 sect 1.1.2.1). Pour compléter le modèle la présence et la prise en charge des données manquantes ont été caractérisées par l'introduction d'une matrice binaire qui caractérise les données en réponses ou non-réponses. Puis notre méthodologie d'imputation a été développé en s'appuyant sur un échantillonneur de Gibbs spécifié à partir de lois conditionnelles a posteriori dérivées du modèle. Les hyperparamètres ont été estimés par une approximation variationnelle de la log-vraisemblance marginale puis une recherche du maximum de la fonction résultante. Alors l'algorithme de reconstruction automatique *PGNMF* été décrit, complétant ainsi des données artificiellement retirées selon un mécanisme *MCAR*. Enfin les résultats de la reconstruction automatique ont été donnés. Deux critères ont servi à l'évaluation des performances de notre algorithme : le RMSE et la divergence de Kullback-Leibler. Le premier nommé est une mesure globale de la qualité de la reconstruction. Le second cité compare deux à deux les items correspondants entre ceux reconstitués et les originaux. Les performances du *PGNMF* ont été comparées à celles des méthodes d'imputation étudiées au chapitre précédent.

Nous avons donc proposé dans cette thèse un algorithme de reconstruction automatique qui s'applique des valeurs entières strictement positives, en l'occurrence les données du questionnaire HSOPSC du CHU de Grenoble, sous un modèle bayésien NMF Poisson-Gamma. Ainsi la particularité de l'algorithme *PGNMF* résultant est qu'il n'a, à notre connaissance, jamais été utilisé dans des travaux antérieurs pour une tâche d'imputation. Alors la question centrale a été d'évaluer l'efficacité de l'algorithme à travers la mesure de la perte d'information par le biais de deux critères : la mesure RMSE et la divergence de Kullback-Leibler (KL). Pour ceux-ci, en l'absence de valeurs de références absolues, l'efficacité du *PGNMF* a donc été évaluée en comparaison d'autres méthodes. Il s'agit quelques structurées d'imputation récemment développées et partageant en commun avec notre algorithme le principe de la factorisation matricielle (*PEM*, *WNMF* et *MIMCA*), mais aussi d'un algorithme d'apprentissage automatique couramment utilisé basé sur les techniques de régression des forêts aléatoires (*MissF*), et de quelques méthodes naïves (imputation par la moyenne, (*moy*), la médiane (*med*), le mode (*mod*) et la méthode *locf*) souvent utilisées dans les analyses de questionnaires médicaux incomplets.

Une première évaluation a été faite sur données synthétiques simulées à partir de notre modèle génératif *Poisson-Gamma*. Les méthode naïves ont donné les valeur RMSE et KL les plus élevées, donc inversément les performances les plus faibles. Les méthodes factorielles et d'apprentissage ont fourni les meilleurs résultats. Comme on s'y attendait notre algorithme *PGNMF* a globalement donné les recouvrements d'information les plus satisfaisants. Les quelques faiblesses notées sont du fait du biais introduit lors de la génération même des données.

Concernant l'enquête HSOPSC de Grenoble, les méthodes naïves de reconstruction automatique ont globalement donné les performances les plus basses par rapport aux deux mesures d'erreurs étudiées (RMSE et KL). A l'opposé les méthodes factorielles (*PEM*, *PGNMF*, *WNMF*, *MIMCA*) et d'apprentissage automatique (*MissF*) ont présenté dans l'ensemble les performances les plus élevées. Pour la reconstruction des histogrammes de réponses aux items, *MissF* a donné les meilleurs résultats. Les méthodes factorielles l'ont cependant remporté sur la méthode d'apprentissage automatique en ce qui concerne le critère RMSE. Notre algorithme *PGNMF* a alors donné les meilleurs résultats surtout lorsque le taux de données masquées est relativement élevé. Mieux qu'une tendance confirmée les résultats RMSE sur cinq (05) répliques boostrap ont placé *PGNMF* comme meilleure procédure de reconstruction automatique non seulement sur le questionnaire HSOPSC de Grenoble mais aussi des enquêtes HSOPSC virtuelles similaires. Le premier fait marquant est que les méthodes factorielles et celle d'apprentissage automatique produisent des résultats satisfaisants, supérieurs aux méthodes naïves traditionnellement utilisées dans ce domaine. Le second fait marquant est distinction des algorithmes *MissF* et *PGNMF*. Alors que le premier est plus performant en reconstruction des histogrammes marginales (critère KL), le second lui l'emporte sur la mesure globale de l'erreur de reconstruction (critère RMSE).

En définitive, pour des travaux ultérieurs sur un questionnaire-type HSOPSC, comme celui du CHU de Grenoble, nous proposons que les méthodes naïves ne soient pas utilisées pour la reconstruction automatique, ou l'imputation, ou la prédiction (c'est selon le contexte) des réponses manquantes aux items. En outre il parait souhaitable de combiner les algorithmes *MissF* et *PGNMF* suivant le critère de performance que l'on met en avant. Le point faible de notre algorithme *PGNMF* demeure sur sa capacité à reconstituer de manière efficiente les histogrammes de réponses aux items du HSOPSC. Notons qu'à la différence du RMSE, le critère divergence de Kullback-Leibler évalue non pas forcément et directement les écarts de reconstruction mais plutôt compare les lois marginales des items originaux à ceux reconstruits. Ce type d'analyse de la perte de l'information, par le critère KL, est souvent utilisé pour ce genre de don-

nées médicales d'enquêtes, d'où la perspective d'amélioration de notre algorithme dans des travaux futurs.

Si nous considérons la mesure de la perte de l'information, calculée à travers la divergence de Kullback-Leibler, sur les données simulées (Table 4.3), où notre algorithme a présenté les meilleurs résultats jusqu'à 30% de données retirées, l'on peut entrevoir la possibilité d'une amélioration des performances de notre algorithme sur les données du type HSOPSC et éventuellement d'autres questionnaires. En effet lors de la construction de notre modèle nous avions opté pour une simplification en liant l'ensemble des coefficients de chaque paramètre, c'est-à-dire que respectivement les coefficients des matrices U et V, ont la même loi gamma (cf. Chap.4, Eq. (4.17) et (4.18)). Cette construction des lois a priori, de la sorte, n'aidait pas à bien prendre les relations entre variables (items) d'une part, et les similarités entre individus selon les réponses données d'autre part. Dans ce cas une perspective d'étude serait alors de ne pas lier les coefficients des paramètres du modèle, mais de chercher les éventuels groupes entre individus par un travail de classification non supervisée.

Annexe 1

Hospital Survey on Patient Safety

Instructions

This survey asks for your opinions about patient safety issues, medical error, and event reporting in your hospital and will take about 10 to 15 minutes to complete.

If you do not wish to answer a question, or if a question does not apply to you, you may leave your answer blank.

- *An **"event"** is defined as any type of error, mistake, incident, accident, or deviation, regardless of whether or not it results in patient harm.*
- ***"Patient safety"** is defined as the avoidance and prevention of patient injuries or adverse events resulting from the processes of health care delivery.*

SECTION A: Your Work Area/Unit

In this survey, think of your "unit" as the work area, department, or clinical area of the hospital where you spend *most* of your work time or provide *most* of your clinical services.

What is your primary work area or unit in this hospital? Select ONE answer.

- ☐ a. Many different hospital units/No specific unit
- ☐ b. Medicine (non-surgical)
- ☐ c. Surgery
- ☐ d. Obstetrics
- ☐ e. Pediatrics
- ☐ f. Emergency department
- ☐ g. Intensive care unit (any type)
- ☐ h. Psychiatry/mental health
- ☐ i. Rehabilitation
- ☐ j. Pharmacy
- ☐ k. Laboratory
- ☐ l. Radiology
- ☐ m. Anesthesiology
- ☐ n. Other, please specify: _____

Please indicate your agreement or disagreement with the following statements about your work area/unit.

Think about your hospital work area/unit…	**Strongly Disagree** ▼	**Disagree** ▼	**Neither** ▼	**Agree** ▼	**Strongly Agree** ▼
1. People support one another in this unit	☐1	☐2	☐3	☐4	☐5
2. We have enough staff to handle the workload	☐1	☐2	☐3	☐4	☐5
3. When a lot of work needs to be done quickly, we work together as a team to get the work done	☐1	☐2	☐3	☐4	☐5
4. In this unit, people treat each other with respect	☐1	☐2	☐3	☐4	☐5
5. Staff in this unit work longer hours than is best for patient care	☐1	☐2	☐3	☐4	☐5

SECTION A: Your Work Area/Unit (continued)

Think about your hospital work area/unit…	Strongly Disagree ▼	Disagree ▼	Neither ▼	Agree ▼	Strongly Agree ▼
6. We are actively doing things to improve patient safety	☐1	☐2	☐3	☐4	☐5
7. We use more agency/temporary staff than is best for patient care	☐1	☐2	☐3	☐4	☐5
8. Staff feel like their mistakes are held against them	☐1	☐2	☐3	☐4	☐5
9. Mistakes have led to positive changes here	☐1	☐2	☐3	☐4	☐5
10. It is just by chance that more serious mistakes don't happen around here	☐1	☐2	☐3	☐4	☐5
11. When one area in this unit gets really busy, others help out	☐1	☐2	☐3	☐4	☐5
12. When an event is reported, it feels like the person is being written up, not the problem	☐1	☐2	☐3	☐4	☐5
13. After we make changes to improve patient safety, we evaluate their effectiveness	☐1	☐2	☐3	☐4	☐5
14. We work in "crisis mode" trying to do too much, too quickly	☐1	☐2	☐3	☐4	☐5
15. Patient safety is never sacrificed to get more work done	☐1	☐2	☐3	☐4	☐5
16. Staff worry that mistakes they make are kept in their personnel file	☐1	☐2	☐3	☐4	☐5
17. We have patient safety problems in this unit	☐1	☐2	☐3	☐4	☐5
18. Our procedures and systems are good at preventing errors from happening	☐1	☐2	☐3	☐4	☐5

SECTION B: Your Supervisor/Manager

Please indicate your agreement or disagreement with the following statements about your immediate supervisor/manager or person to whom you directly report.

	Strongly Disagree ▼	Disagree ▼	Neither ▼	Agree ▼	Strongly Agree ▼
1. My supervisor/manager says a good word when he/she sees a job done according to established patient safety procedures	☐1	☐2	☐3	☐4	☐5
2. My supervisor/manager seriously considers staff suggestions for improving patient safety	☐1	☐2	☐3	☐4	☐5
3. Whenever pressure builds up, my supervisor/manager wants us to work faster, even if it means taking shortcuts	☐1	☐2	☐3	☐4	☐5
4. My supervisor/manager overlooks patient safety problems that happen over and over	☐1	☐2	☐3	☐4	☐5

SECTION C: Communications

How often do the following things happen in your work area/unit?

Think about your hospital work area/unit…	**Never** ▼	**Rarely** ▼	**Some-times** ▼	**Most of the time** ▼	**Always** ▼
1. We are given feedback about changes put into place based on event reports	☐1	☐2	☐3	☐4	☐5
2. Staff will freely speak up if they see something that may negatively affect patient care	☐1	☐2	☐3	☐4	☐5
3. We are informed about errors that happen in this unit	☐1	☐2	☐3	☐4	☐5
4. Staff feel free to question the decisions or actions of those with more authority	☐1	☐2	☐3	☐4	☐5
5. In this unit, we discuss ways to prevent errors from happening again	☐1	☐2	☐3	☐4	☐5
6. Staff are afraid to ask questions when something does not seem right	☐1	☐2	☐3	☐4	☐5

SECTION D: Frequency of Events Reported

In your hospital work area/unit, when the following mistakes happen, *how often are they reported?*

	Never ▼	**Rarely** ▼	**Some-times** ▼	**Most of the time** ▼	**Always** ▼
1. When a mistake is made, but is *caught and corrected before affecting the patient*, how often is this reported?	☐1	☐2	☐3	☐4	☐5
2. When a mistake is made, but has *no potential to harm the patient*, how often is this reported?	☐1	☐2	☐3	☐4	☐5
3. When a mistake is made that *could harm the patient*, but does not, how often is this reported?	☐1	☐2	☐3	☐4	☐5

SECTION E: Patient Safety Grade

Please give your work area/unit in this hospital an overall grade on patient safety.

☐	☐	☐	☐	☐
A Excellent	**B** Very Good	**C** Acceptable	**D** Poor	**E** Failing

SECTION F: Your Hospital

Please indicate your agreement or disagreement with the following statements about your hospital.

Think about your hospital…	**Strongly Disagree** ▼	**Disagree** ▼	**Neither** ▼	**Agree** ▼	**Strongly Agree** ▼
1. Hospital management provides a work climate that promotes patient safety	☐1	☐2	☐3	☐4	☐5
2. Hospital units do not coordinate well with each other	☐1	☐2	☐3	☐4	☐5
3. Things “fall between the cracks” when transferring patients from one unit to another	☐1	☐2	☐3	☐4	☐5
4. There is good cooperation among hospital units that need to work together	☐1	☐2	☐3	☐4	☐5

SECTION F: Your Hospital (continued)

Think about your hospital…	Strongly Disagree ▼	Disagree ▼	Neither ▼	Agree ▼	Strongly Agree ▼
5. Important patient care information is often lost during shift changes	☐1	☐2	☐3	☐4	☐5
6. It is often unpleasant to work with staff from other hospital units	☐1	☐2	☐3	☐4	☐5
7. Problems often occur in the exchange of information across hospital units.....	☐1	☐2	☐3	☐4	☐5
8. The actions of hospital management show that patient safety is a top priority	☐1	☐2	☐3	☐4	☐5
9. Hospital management seems interested in patient safety only after an adverse event happens.....	☐1	☐2	☐3	☐4	☐5
10. Hospital units work well together to provide the best care for patients	☐1	☐2	☐3	☐4	☐5
11. Shift changes are problematic for patients in this hospital.....	☐1	☐2	☐3	☐4	☐5

SECTION G: Number of Events Reported

In the past 12 months, how many event reports have you filled out and submitted?

- ☐ a. No event reports
- ☐ b. 1 to 2 event reports
- ☐ c. 3 to 5 event reports
- ☐ d. 6 to 10 event reports
- ☐ e. 11 to 20 event reports
- ☐ f. 21 event reports or more

SECTION H: Background Information

This information will help in the analysis of the survey results.

1. How long have you worked in this hospital?

- ☐ a. Less than 1 year
- ☐ b. 1 to 5 years
- ☐ c. 6 to 10 years
- ☐ d. 11 to 15 years
- ☐ e. 16 to 20 years
- ☐ f. 21 years or more

2. How long have you worked in your current hospital work area/unit?

- ☐ a. Less than 1 year
- ☐ b. 1 to 5 years
- ☐ c. 6 to 10 years
- ☐ d. 11 to 15 years
- ☐ e. 16 to 20 years
- ☐ f. 21 years or more

3. Typically, how many hours per week do you work in this hospital?

- ☐ a. Less than 20 hours per week
- ☐ b. 20 to 39 hours per week
- ☐ c. 40 to 59 hours per week
- ☐ d. 60 to 79 hours per week
- ☐ e. 80 to 99 hours per week
- ☐ f. 100 hours per week or more

SECTION H: Background Information (continued)

4. What is your staff position in this hospital? Select ONE answer that best describes your staff position.

- ☐ a. Registered Nurse
- ☐ b. Physician Assistant/Nurse Practitioner
- ☐ c. LVN/LPN
- ☐ d. Patient Care Asst/Hospital Aide/Care Partner
- ☐ e. Attending/Staff Physician
- ☐ f. Resident Physician/Physician in Training
- ☐ g. Pharmacist
- ☐ h. Dietician
- ☐ i. Unit Assistant/Clerk/Secretary
- ☐ j. Respiratory Therapist
- ☐ k. Physical, Occupational, or Speech Therapist
- ☐ l. Technician (e.g., EKG, Lab, Radiology)
- ☐ m. Administration/Management
- ☐ n. Other, please specify:

5. In your staff position, do you typically have direct interaction or contact with patients?

- ☐ a. YES, I typically have direct interaction or contact with patients.
- ☐ b. NO, I typically do NOT have direct interaction or contact with patients.

6. How long have you worked in your current specialty or profession?

- ☐ a. Less than 1 year
- ☐ b. 1 to 5 years
- ☐ c. 6 to 10 years
- ☐ d. 11 to 15 years
- ☐ e. 16 to 20 years
- ☐ f. 21 years or more

SECTION I: Your Comments

Please feel free to write any comments about patient safety, error, or event reporting in your hospital.

THANK YOU FOR COMPLETING THIS SURVEY.

Annexe 2

RÉSULTATS SIMULATIONS SUR REPLIQUES BOOSTRAP

Boostrap 3

	10%	20%	30%	40%	50%	60%	70%	80%
missf	1.2366	1.2390	1.2498	1.2621	1.2689	1.3094	1.3395	1.3919
pem	1.1248	1.1263	1.1317	1.1369	1.1376	1.1475	1.1633	1.1904
pgmnf	**1.1237**	**1.1259**	**1.1309**	**1.1357**	**1.1355**	**1.1447**	**1.1571**	**1.1799**
wnmf	1.1622	1.1676	1.1759	1.1910	1.2036	1.2266	1.2712	1.3496
mimca	1.4725	1.4734	1.4757	1.4742	1.4741	1.4786	1.4850	1.4982

Table 1: Performances de différentes méthodes d'imputation par le biais du RMSE en fonction du taux de non-réponses sur le boostrap 3.

Boostrap 4

	10%	20%	30%	40%	50%	60%	70%	80%
missf	1.2480	1.2396	1.2480	1.2406	1.2723	1.2967	1.3319	1.3817
pem	1.1288	1.1206	1.1291	1.1313	1.1384	1.1444	1.1571	1.1912
pgmnf	**1.1283**	**1.1196**	**1.1284**	**1.1304**	**1.1359**	**1.1413**	**1.1513**	**1.1755**
wnmf	1.1622	1.1577	1.1737	1.1863	1.2015	1.2271	1.2629	1.3380
mimca	1.4717	1.4658	1.4692	1.4690	1.4705	1.4734	1.4786	1.4890

Table 2: Performances de différentes méthodes d'imputation par le biais du RMSE en fonction du taux de non-réponses sur le boostrap 4.

Boostrap 5

	10%	20%	30%	40%	50%	60%	70%	80%
missf	1.2446	1.2424	1.2348	1.2498	1.2873	1.2937	1.3358	1.3842
pem	1.1319	1.1279	1.1256	1.1327	1.1407	1.1475	1.1610	1.1887
pgmnf	**1.1312**	**1.1277**	**1.1252**	**1.1313**	**1.1386**	**1.1441**	**1.1551**	**1.1733**
wnmf	1.1679	1.1680	1.1711	1.1888	1.2041	1.2277	1.2708	1.3426
mimca	1.4773	1.4734	1.4697	1.4692	1.4708	1.4736	1.4794	1.4916

Table 3: Performances de différentes méthodes d'imputation par le biais du RMSE en fonction du taux de non-réponses sur le boostrap 5.

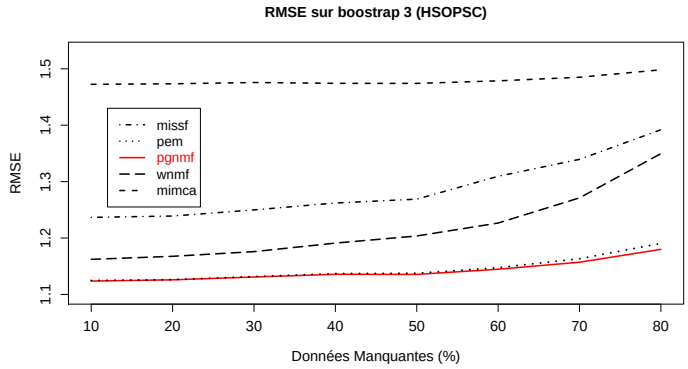

Figure 1: Performances suivant le critère RMSE sur la réplique boostrap 3.

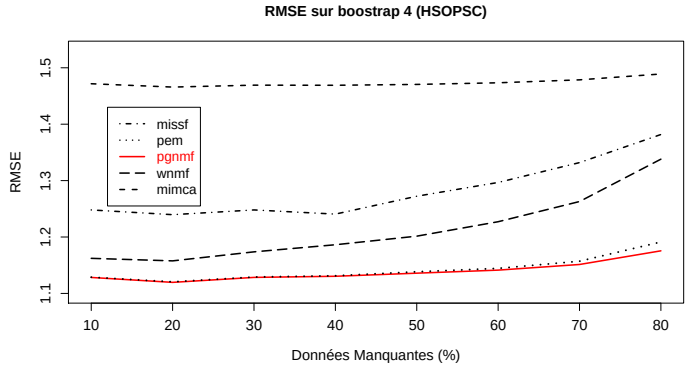

Figure 2: Performances suivant le critère RMSE sur la réplique boostrap 4.

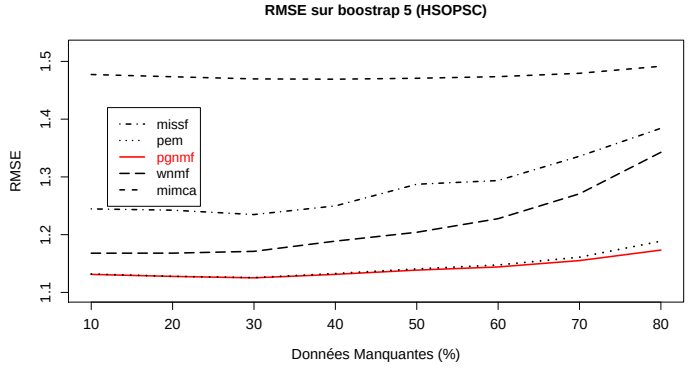

Figure 3: Performances suivant le critère RMSE sur la réplique boostrap 5.

Bibliographie

[1] S. S. Wilks. Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3) :163–195, 1932.

[2] M. J. R. Healy and M. Westmacott. Missing values in experiments analyzed on automatic computers. *Journal of the Royal Statistical Society*, 5(3) :203–206, 1956.

[3] T. W. Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observation are missing. *Journal of the American Statistical Association*, 52(278) :200–203, 1957.

[4] G. N. Wilkinson. Estimation of missing values for the analysis of incomplete data. *Biometrics*, 14(2) :257–286, 1958.

[5] I. M. Trawinski and R. E. Bargmann. Maximum likelihood estimation with incomplete multivariate data. *The Annals of Mathematical Statistics*, 35(2) :647–657, 1964.

[6] A. A. Afifi and R. M. Elashoff. Missing Observation in Multivariate Statistics : I. Review of the Literature. *Journal of the American Statistical Association*, 61(315) :595–604, 1966.

[7] H. O. Hartley and R. R. Hocking. The analysis of incomplete data. *Biometrics*, 27(4) :783–823, 1971.

[8] T. Orchard and M. A. Woodbury. A missing information principle : theory and applications. In *Proceedings Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 697–715, Berkeley, California, 1972. University of California Press.

[9] D. B. Rubin. Inference in Missing Data. *Biometrika*, 63 :581–592, 1976.

[10] D. B. Rubin. Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 20–34, 1978.

[11] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics, New York, 1987.

[12] D. B. Rubin. An Overview of Multiple Imputation. In *Proceedings of the Survey Research Section, American Statistical Association*, pages 79–84, 1988.

[13] R. J. A. Little. Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, 77(378) :237–250, 1982.

[14] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New York, 1987.

[15] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, New York, 1997.

[16] J. L. Schafer and R. Yucel. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11 :437–457, 2002.

[17] S. Van Buuren and K. Oudshoorn. Flexible Multivariate Imputation by mice. *TNO Prevention Center, Leiden, The Netherlands*, pages 1–20, 1999.

[18] M. Resche-Rigon and I. R. White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6) :1634–1649, 2018.

[19] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Letters to Nature*, 401 :788–791, 1999.

[20] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. *Adv. Neural Inform. Process. Systems*, 13 :556–562, 2001.

[21] Y. Mao and L. K. Saul. Modeling distances in large-scale networks by matrix factorization. In *Proceedings of the AC Internet Measurement Conference*, Sicily, Italy, 2004.

[22] Y. D. Kim and S. Choi. Weighted nonnegative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1541–1544, Taipei, 2009. IEEE.

[23] J. Canny. A Factor Model for Discrete Data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129, 2004.

[24] S. Zhang, W. Wang, J. Ford, and F. Makedon. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 549–553, 2006.

[25] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio Speech and Language Processing*, 15(3) :1066–1074, 2007.

[26] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorization for audio signal modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1825–1828, Las Vegas, 2008. IEEE.

[27] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887, Helsinki, Finland, 2008.

[28] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, Richard Petersens Plads, Building 321, 2009. Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

[29] A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience*, 2009 :1–17, 2009.

[30] H. Shan and A. Banerjee. Generalized probabilistic matrix factorizations for collaborative filtering. In *2010 IEEE 10th Data Mining International Conference*, pages 1025–1030. IEEE, 2010.

[31] I. Porteous, A. U. Asuncion, and M. Welling. Bayesian matrix factorization with side information and dirichlet process mixtures. In *Conference of the Association for the Advancement of Artificial Intelligence*. AAAI Press, 2010.

[32] P. Gopalan, L. Charlin, and D. Blei. Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems 27*, pages 3176–3184, 2014.

[33] P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *UAI'15 Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 326–335, Amsterdam, Netherlands, 2015.

[34] Z. Ghahramani and M. Beal. Propagation Algorithms for Variational Bayesian Learning. In *Advances in Neural Information Processing Systems*, pages 507–513, Cambridge, Massachusetts, USA, 2001. MIT Press.

[35] P. Occelli, J.-L. Quenon, M. Kret, and al. Validation of the French version of the Hospital Survey on Patient Safety Culture questionnaire. *International Journal for Quality in Health Care*, 25(4) :459–468, 2013.

[36] R. Zdunek and A. Cichocki. Non-negative matrix factorization with quasi-Newton optimization. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Soft Computing*, Zakopane, Poland, 2006.

[37] D. Kim, S. Sra, and I. S. Dhillon. Fast Newton-type methods for the least squares nonnegative matrix approximation problem. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 343–354, 2007.

[38] C. J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19 :2756–2779, 2007.

[39] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30 :713–730, 2008.

[40] L. Breiman. Random Forests. *Machine Learning*, 45 :5–32, 2001.

[41] D. J. Stekhoven and P. Bühlmann. Missforest-nonparametric missing value imputation for mixed-type data. *Bioinformatics Advance Access*, 28 :112–118, 2011.

[42] J. Josse and F. Husson. missMDA : a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1) :1–31, 2016.

[43] V. Audigier, F. Husson, and J. Josse. MIMCA : multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2) :501–518, 2017.

[44] J. Josse, M. Chavent, B. Liquet, and F. Husson. Handling missing values with Regularized Iterative Multiple Correspondence Analysis. *Journal of Classification*, 29(1) :91–116, 2012.

[45] C. Robert. *Le choix bayésien, Principes et pratique.* Springer-Verlag, Paris, 2006.

[46] M. N. M. van Lieshout and R. S. Stoica. The candy model revisited : properties and inference. *Statistica Neerlandica*, 57 :1–30, 2003.

[47] R. J. Erskine. Mastitis control in dairy herds. In In O.M. Radostis editor, editor, *Herd health : food animal production medecine*, pages 397–433, Philadelphia, 2001. 3rd ed. W.B. Saunders Company.

[48] L. Roques and R. S. Stoica. Species persistence decreases with habitat fragmentation : an analysis in periodic stochastic environments. *Journal of Mathematical Biology*, 55(2) :189–205, 2007.

[49] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21 :1087, 1953.

[50] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In *IEEE Transactions on Pattern Analysis and Mathematical Intelligence*, volume 6, 1984.

[51] I. Biederman. Recognition-by-components : a theory of human image understanding. *Psychol. Rev.*, 94 :115–147, 1987.

[52] S. Ullman. *High-Level Vision : Object Recognition and Visual Cognition.* MIT Press, Cambridge, Massachusetts, 1996.

[53] E. Wachsmuth, M. W. Oram, and D. I. Perrett. Recognition of objects and their component parts : responses of single units in the temporal cortex of the macaque. *Cereb. Cortex*, 4 :509–522, 1994.

[54] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annu. Rev. Neurosci*, 19 :577–621, 1996.

[55] P. Paatero and U. Tapper. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37 :23–35, 1997.

[56] 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*, 39 :1–38, 1977.

[57] A. L. Cauchy. Methode générale pour la résolution des systèmes d'équations simultanées. *Oeuvres Complètes*, 10(383) :399–402, 1847.

[58] N.-D. Ho. *Nonnegative Matrix Factorization Algorithms and Applications.* PhD thesis, Université Catholique de Louvain, 2008.

[59] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, 1999.

[60] M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation and interpretation of nonnegative matrix factorizations. http ://www.wfu.edu/plemmons/papers/chu-ple.pdf, 2005.

[61] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52 :155–173, 2007.

[62] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5 :1457–1469, 2004.

[63] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, Berlin, 1999.

[64] P. Paatero and U. Tapper. Positive matrix factorization : a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 :111–126, 1994.

[65] P. Paatero. A weighted non-negative least squares algorithm for three-way parafac factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 38(2) :223–242, 1997.

[66] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling. Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization. *arXiv :1407.7299 Computer Science, Numerical Analysis*, 2014.

[67] L. Finesso and P. Spreij. Approximate nonnegative matrix factorization via alternating minimization. In *Sixteenth International Symposium on Mathematical Theory of Networks and Systems*, Leuven, 2004.

[68] E. F. Gonzalez and Y. Zhang. Accelerating the Lee-Seung algorithm for nonnegative matrix factorization. Technical report, Rice University, March 2005. Technical Report TR-05-02.

[69] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems.* Prentice-Hall, 1974.

[70] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Proceeding of the Neural Information Processing Systems (NIPS) Conference*, 2005.

[71] M. Bierlaire, P. L. Toint, and D. Tuyttens. On Iterative Algorithms for Linear Least Squares Problems with Bound constraints. *Linear Algebra and its Applications*, 143 :111–143, 1991.

[72] R. Bro and de Jong S. A fast non-negativity constrained linear least squares algorithm. *Journal of Chemometrics*, 11(5) :393–401, 1997.

[73] M. H. van Benthem and M. R. Keenan. Fast algorithm for the solution of large-scale nonnegativity-constrained least squares problems. *Journal of Chemometrics*, 18 :441–450, 2004.

[74] M. Merritt and Y. Zhang. Interior-Point Gradient Method for Large-Scale Totally Nonnegative Least Squares Problems. *Journal of Optimization Theory and Applications*, 126(1) :191–202, 2005.

[75] D. Kim, S. Sra, and I. S. Dhillon. A New Projected Quasi-Newton Approach for the Non-negative Least Squares Problem. Technical report, The Univ. of Texas at Austin, 2006. Technical Report TR-06-54, Computer Sciences.

[76] D. Guillamet, M. Bressan, and J. Vitrià. A weighted nonnegative matrix factorization for local representations. In *IEEE Computer Society Conference on Compute Vision and Pattern Recognition*, Kauai, HI, USA.

[77] V. D. Blondel, N.-D. Ho, and P. van Dooren. Weighted Nonnegative Matrix Factorization and Face Feature Extraction. In *Image and Vision Computing*, pages 1–17, 2008.

[78] S. Demissie, M. P. LaValley, N. J. Horton, R. J. Glynn, and L. A. Cupples. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistic in Medicine*, 22(4) :545–557, 2003.

[79] J. M. G. Taylor, K. L. Cooper, J. T. Wei, A. V. Sarma, T. E. Raghunathan, and S. G. Heeringa. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American. *American Journal of Epidemiology*, 156(8) :774–782, 2002.

[80] L. Joseph, P. Belisle, H. Tamim, and J. S. Sampalis. Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. *Journal of Clinical Epidemiology*, 57(2) :147–153, 2004.

[81] F. M. Shrive, H. Stuart, H. Quan, and W. A. Ghali. Dealing with missing data in a multi-question depression scale : a comparison of imputation methods. *BMC Medical Research Methodology*, 6 :57, 2006.

[82] S. van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16 :219–242, 2007.

[83] J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma. Multiple imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric results. *Multivariate behavioral research*, 42(2) :387–414, 2007.

[84] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations : issues and guidance for practice. *Statistics in Medicine*, 30(4) :377–399, 2011.

[85] N. Resseguier, H. Verdoux, R. Giorgi, F. Clavel-Chapelon, and X. Paoletti. Dealing with missing data in the Center for Epidemiologic Studies Depression self-report scale : a study based on the French E3N cohort. *BMC Medical Research Methodology*, 13 :28, 2013.

[86] J. O. Kim and J. Curry. The treatement of missing data in multivariate analysis. *Social Meth. Res.*, 6 :215–240, 1977.

[87] Y. Haitovsky. Missing data in regression analysis. *Journal of Royal Statistique Society*, 30 :67–81, 1968.

[88] S. Azen and M. Van Guilder. Conclusions regarding algorithms for handling incomplete data. In *Proceedings of the Statistical Computing Section,*, pages 53–56. American Statistical Association, 1981.

[89] M. J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, London, 1984.

[90] S. J.-M. Palm. *Les Facteurs d'Acceptabilité d'un Système d'Information Clinique (SIC) : Evaluation Comparative France (HEGP) - Québec (CHUS)*. PhD thesis, Université Pierre et Marie Curie - Université de Sherbrooke, 2010.

[91] J. S. Sorra and V. F. Nieva. *Hospital Survey on Patient Safety Culture. (Prepared by Westat, under Contract No. 290-96-0004)*. AHRQ Publication No. 04-0041, Rockville, MD : Agency for Healthcare Research and Quality, 2004.

[92] J. L. Schafer and J. W. Graham. Missing Data : Our View of the State of the Art. *Psychological Methods*, 7(2) :174–177, 2002.

[93] S. B. Cohen. An analysis of alternative imputation strategies for individuals with partial data in the National Medical Care Expenditure Survey. *Review of public data use*, 10(3) :153–165, 1982.

[94] A. Rotnitzky and D. Wypij. A note on the bias of estimators with missing data. *Biometrics*, 50(4) :1163–1170, 1994.

[95] K. J. Jones, A. Skinner, L. Xu, J. Sun, and K. Mueller. The AHRQ Hospital Survey on Patient Safety Culture : A tool to plan and evaluate patient safety programs. *Advances in Patient Safety : New Directions and Alternative Approaches*, 2, 2008.

[96] Health and Safety Commission. *Organising for safety : Third report of the human factors study group of ACSNI*. HSE Books, Sudbury, UK.

[97] D. A. Wiegmann, H. Zhang, and T. et al. von Thaden. A synthesis of safety culture and safety climate research. Technical report, Federal Aviation Administration, 2002. . Technical Report No. : ARL-02-3/FAA-02-2. Contract No. DTFA 01-G-015.

[98] Institute of Medicine. *Patient safety : Achieving a new standard of care*. The National Academies Press, Washington DC.

[99] B. D. B. Diatta, P. Ngom, B. Boussat, and O. François. Reconstruction automatique de formulaires d'enquête médicale sur la culture de sécurité des patients par une méthode de factorisation matricielle bayésienne. *Journal de la Société Française de Statistique*, 159(2) :111–127, 2018.

[100] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, USA, 2006.

[101] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64 :583–616, 2002.

[102] P. Paatero. The multilinear engine-a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *J. Comput.Graphical Statis*, 8(4) :1–35, 1999.

[103] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis. *Bioinformatics*, 23 :1495–1502, 2007.

[104] M. H. Van Benthem and M. R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of Chemometrics*, 18 :441–450, 2004.

[105] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14 :1303–1347, 2013.

[106] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37 :183–233, 1999.

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

[108] M. Reilly. Data Analysis Using Hot Deck Multiple Imputation. *Journal of the Royal Statistical Society*, 42(3) :307–313, 1993.

[109] R. M. Yucel. Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, 11(4) :351–370, 2011.

[110] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1) :41–55, 1983.

[111] J. M. Etchegaray and E. J. Thomas. Comparing two safety culture surveys : safety attitudes questionnaire and hospital survey on patient safety. *BMJ Quality and Safety*, 21(6) :490–498, 2012.

[112] K. Kemp, S. Warren, N. Chan, B. McCormack, M. Santana, and H. Quan. Qualitative complaints and their relation to overall hospital rating using an H-CAHPS-derived instrument. *BMJ Quality and Safety*, 25(10) :770–777, 2016.

[113] R. S. Stoica, P. Gregori, and J. Mateu. Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115 :1860–1882, 2005.

[114] R. S. Stoica, V. J. Martinez, and E. Saar. A three dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 55 :189–205, 2007.

[115] R. S. Stoica, V. J Martinez, and E. Saar. Filaments in observed and mock galaxy catalogues. *Astronomy and Astrophysics,*, 510 :1–12, 2010.

[116] E. Tempel, R. S. Stoica, and E. Saar. Evidence for spin alignment of spiral and elliptical galaxies in filaments. *Monthly Notices of the Royal Astronomical Society*, 428 :1827–1836, 2013.

[117] W. K. Hastings. Monte Carlo Sampling Methods using Markov Chains and their applications. *Biometrika*, 57(1) :97–109, 1970.

[118] D. L. Streiner. Missing data and the trouble with LOCF. *Evid Based Ment Health*, 11(1) :3–5, 2008.

[119] A. T. Cemgil and O. Dikmen. Conjugate Gamma Markov random fields for modelling nonstationary sources. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, pages 697–705, London, UK, 2007.

[120] Y. Chen, J. Mairal, and Z. Harchaoui. Fast and Robust Archetypal Analysis for Representation Learning. In *CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1478–1485, Washington, DC, USA, 2014.

[121] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, USA, 2003. IEEE.

[122] S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of 5th International Conference on Music Information Retrieva*, pages 318–325, Barcelona, Spain, 2004.

[123] A. J. Feelders. Handling Missing Data in Trees : Surrogate Splits or Statistical Imputation. In *PKDD '99 Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pages 329–334, London, UK, 1999. Springer-Verlag.

[124] X. Descombes, R. S. Stoica, and J. Zerubia. Two markov point processes for simulating line networks. In *International Conference on Image Processing Proceedings*, Kobe, Japan, 1999. IEEE.

[125] R. S. Stoica, X. Descombe, and J. Zerubia. Road extraction in remote sensed images using a stochastic geometry framework. In *Bayesian inference and maximum entropy methods in science and engineering : 20th International Workshop*, volume 568, pages 531–542. AIP Conference Proceeding, 2001.

[126] P. Gregori, J. Mateu, and R. S. Stoica. A marked point process for modelling three dimensional patterns. In *Spatial Point Process Modelling and its Applications*, page 91, Castellon, Spain, 2004. University Jaume I.

[127] J. Carpenter and M. Kenward. *Multiple imputation and its application.* John Wiley and Sons, New York, 2012.

[128] N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. *Statistical Package for the Social Science.* McGraw-Hill Inc, USA, 2nd edition edition, 1975.

[129] A. W. F. Edwards. *Likelihood.* Cambridge University Press, 1972.

[130] V. J. Martinez and E. Saar. *Statistics of the galaxy distribution.* Chapman and Hall, 2002.

[131] J. Moller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes.* Chapman and Hall/CRC, Boca Raton, 2004.

[132] R. S. Stoica. *Processus Ponctuels pour l'Extraction de Réseaux Linéïques.* PhD thesis, Université de Nice Sophia-Antipolis, 2001.

[133] R. S. Stoica. *Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées.* PhD thesis, Université Lille 1, 2014.

[134] L. Ben Othman Amroussi. *Conception et validation d'une méthode de complétion des valeurs manquantes fondée sur leurs modèles d'apparition.* PhD thesis, Université de Caen, Basse-Normandie, Ecole doctorale SIMEM, 2011.