# Imputation multiple par analyse factorielle : Une nouvelle méthodologie pour traiter les données manquantes

Vincent Audigier

N° ordre : 2015-25
N° série : G-15

# THÈSE / AGROCAMPUS OUEST

Sous le sceau de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

**DOCTEUR DE L'INSTITUT SUPÉRIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE**

Spécialité : Mathématiques Appliquées

**Ecole doctorale : MATISSE**

présentée par

**Audigier Vincent**

# Imputation multiple par analyse factorielle

## Une nouvelle méthodologie pour traiter les données manquantes

Soutenue le **25 novembre 2015** devant la commission d'Examen :

| | | |
|---|---|---|
| Bouveyron Charles | Université de Paris Descartes (France) | Rapporteur |
| Van Mechelen Iven | Université de Louvain (Belgique) | Rapporteur |
| Biernacki Christophe | Université de Lille 1 (France) | Président |
| Resche-Rigon Matthieu | Université Paris Diderot (France) | Examinateur |
| François Husson | Agrocampus Ouest, Rennes (France) | Directeur de thèse |
| Julie Josse | Agrocampus Ouest, Rennes (France) | Directeur de thèse |

# Imputation multiple par analyse factorielle
## Une nouvelle méthodologie pour traiter les données manquantes

Les données manquantes sont fréquentes dans la pratique statistique. Elles sont problématiques car la plupart des méthodes ne peuvent pas être appliquées sur un tableau de données incomplet. Une solution classique pour gérer les données manquantes consiste à recourir à l'imputation multiple.

Bien que de nombreuses techniques soient disponibles dans la littérature, celles-ci présentent encore de nombreux manques. En particulier, il existe très peu de solutions satisfaisantes pour compléter des jeux de données avec des variables qualitatives ou mixtes, ou des jeux de données avec beaucoup de variables, ou peu d'individus par rapport au nombre de variables. Cela s'explique notamment par les problèmes de surajustement engendrés par un trop grand nombre de paramètres à estimer.

Cette thèse est centrée sur le développement de nouvelles méthodes d'imputation multiples, basées sur des techniques d'analyse factorielle. Ces dernières sont classiquement utilisées pour résumer et visualiser des données multidimensionnelles quantitatives, qualitatives ou mixtes. L'étude des méthodes factorielles, ici en tant que méthodes d'imputation, offre de grandes perspectives en termes de diversité du type de données imputées d'une part, et en termes de dimensions de jeux de données imputés d'autre part. Leur propriété de réduction de la dimension limite en effet le nombre de paramètres estimés.

Dans un premier temps, nous détaillons une méthode d'imputation simple par analyse factorielle de données mixtes. Les propriétés de cette méthode sont étudiées, en particulier sa capacité à gérer la diversité des liaisons mises en jeu et à prendre en compte les modalités rares. Sa qualité de prédiction est illustrée en la comparant à la méthode de référence pour prédire des données mixtes : l'imputation par forêts aléatoires.

Ensuite, une méthode d'imputation multiple pour des données quantitatives par analyse en composantes principales (ACP) est présentée. Celle-ci repose sur une approche Bayésienne du modèle d'ACP. Contrairement aux méthodes classiques basées sur modèle Gaussien, elle permet d'estimer des paramètres en présence de données manquantes y compris quand le nombre d'individus est petit devant le nombre de variables, ou quand les corrélations entre variables sont fortes.

Enfin, une méthode d'imputation multiple pour des données qualitatives par analyse des correspondances multiples (ACM) est proposée. La variabilité de prédiction des données manquantes est reflétée par une procédure de bootstrap non-paramétrique. L'imputation multiple par ACM offre une réponse au problème majeur de l'explosion combinatoire qui met en défaut la majorité des méthodes concurrentes dès lors que le nombre de variables ou de modalités est élevé.

Le recours aux méthodes d'analyse factorielle ouvre ainsi de nouvelles perspectives pour l'inférence en présence de données manquantes, via des méthodes d'imputation multiple innovantes.

Mots clés : Données manquantes, Données mixtes, Données qualitatives, Imputation multiple, Imputation simple, Analyse factorielle des données mixtes, Analyse en composantes principales, Analyse des correspondances multiples, Bayésien, Bootstrap

# Multiple imputation using principal component methods
## A new methodology to deal with missing values

Missing data are common in the domain of statistics. They are a key problem because most statistical methods cannot be applied to incomplete data sets. Multiple imputation is a classical strategy for dealing with this issue.

Although many multiple imputation methods have been suggested in the literature, they still have some weaknesses. In particular, it is difficult to impute categorical data, mixed data and data sets with many variables, or a small number of individuals with respect to the number of variables. This is due to overfitting caused by the large number of parameters to be estimated.

This thesis proposes new multiple imputation methods that are based on principal component methods, which were initially used for exploratory analysis and visualisation of continuous, categorical and mixed multidimensional data. The study of principal component methods for imputation, never previously attempted, offers the possibility to deal with many types and sizes of data. This is because the number of estimated parameters is limited due to dimensionality reduction.

First, we describe a single imputation method based on factor analysis of mixed data. We study its properties and focus on its ability to handle complex relationships between variables, as well as infrequent categories. Its high prediction quality is highlighted with respect to the state-of-the-art single imputation method based on random forests.

Next, a multiple imputation method for continuous data using principal component analysis (PCA) is presented. This is based on a Bayesian treatment of the PCA model. Unlike standard methods based on Gaussian models, it can still be used when the number of variables is larger than the number of individuals and when correlations between variables are strong.

Finally, a multiple imputation method for categorical data using multiple correspondence analysis (MCA) is proposed. The variability of prediction of missing values is introduced via a non-parametric bootstrap approach. This helps to tackle the combinatorial issues which arise from the large number of categories and variables. We show that multiple imputation using MCA outperforms the best current methods.

Using principal component methods therefore opens up new interesting perspectives for multiple imputation.

Keywords : Missing data, Mixed data, Categorical data, Multiple Imputation, Single Imputation, Factorial analysis of mixed data, Principal component analysis, Multiple correspondence analysis, Bayesian, Bootstrap

# REMERCIEMENTS

Je tiens tout d'abord à remercier Iven van Michelen et Charles Bouveyron pour avoir accepté d'être les rapporteurs de cette thèse. Je remercie également Christophe Biernacki et Matthieu Resche-Rigon pour leur participation à ce jury en qualité d'examinateurs.

J'adresse par ailleurs de grands remerciements à mes deux directeurs de thèse François Husson et Julie Josse pour m'avoir tout d'abord initié à la recherche, puis pour m'avoir fait confiance durant ces trois années. Merci pour votre bienveillance, votre encadrement, votre disponibilité, vos qualités humaines, vos conseils tant dans les aspects scientifiques, que de communication ou de pédagogie. J'espère (également) que nous travaillerons encore ensemble à l'avenir.

Merci également à tous les membres de l'équipe du LMA2, permanents ou temporaires, d'aujourd'hui ou d'hier. Par ordre alphabétique, merci tout d'abord à David. Ton objectivité mêlée à ton humour ravageur constitue un ensemble que j'ai beaucoup apprécié. Merci aussi à Ekeina. Tu as su me redonner goût à la bettrave, et pour ça je ne te remercierai jamais assez... Merci à Elisabeth, pour sa sympathie, mais surtout pour son foie gras hors pair que mes papilles ne sont pas prêtes d'oublier (une piqûre de rappel ne faisant toutefois jamais de mal). Bien évidemment, je remercie également Emeline, Geoffroy, pour avoir accompagné cette fin de thèse et rendu cette période de rédaction plus distrayante. Merci aussi à Emeline, Perthame, fille de Benoît Perthame. Merci beaucoup, merci pour tout et merci encore. Mes remerciements vont également à Guillaume. Ce fut un réel plaisir de rencontrer quelqu'un avec des valeurs telles que les tiennes. Merci à Jérôme pour m'avoir fait partager son expertise vis-à-vis de l'AFM. Merci aussi pour vos nombreuses anecdotes, toujours captivantes, qui ont animé beaucoup de pauses café. Je remercie également Karine qui a toujours assuré quand j'ai eu besoin de ses services au cours de ces trois années. Merci aussi à Leslie. Si la salle D0 était notre résidence principale, je suis ravi d'avoir été ton colocataire pendant un an. Mes remerciements vont également à Linda, une de mes lectrices les plus fidèles. Thank you very much for your help ! Je tiens aussi à remercier Magalie qui a guidé mes premiers pas dans l'enseignement. Merci aussi pour ta constante sympathie. Margot, tu sais tout le mal que je pense de toi, je pense t'en faire part assez souvent pour qu'il ne soit pas nécessaire de le mentionner ici. Disons simplement que je te remercie de me donner envie de partir de Rennes. Marie,

merci pour ce magnifique gabarit ! Tu es en somme la *Mère* de ce manuscrit. Et puis merci aussi pour ces bons moments passés en ta compagnie. Merci aussi à Mathieu. J'espère que les problèmes de données manquantes en génétique trouverons un jour une solution multiple. Je remercie aussi Pavlo ! Merci pour toute l'aide que tu as pu m'apporter ces derniers mois. Merci également de m'avoir fait partager l'étendue de tes connaissances. Je pense, avec une probabilité non-négative, que nous échangerons encore à l'avenir. Je tiens aussi à remercier Sébastien. En accueillant sous ton propre toit les miséreux doctorants égarés, tu as remis une génération de loosers sur le chemin de la gagne. Merci beaucoup pour ton hospitalité. Enfin, je remercie quelqu'un qui avait une vie de thésard, une vie de père, de mari, une vie de guitariste, la liste est longue. On arrive facilement à sept vies, quoi de plus normal pour un chat ? Merci Tam.

Par ailleurs je remercie mes amis Rennais qui ont égayé mes soirées et week-ends. Merci à Cyril, Gaspar, Marie-line. Je remercie aussi mes amis Picto-Charentais pour avoir su accepter mon éloignement durant ces quelques années et, pour certains d'entre eux, pour m'avoir rendu visite dans mon vaste appartement. Merci à Cousin, la Mexicaine, Juliette, Anastasia, Arnaud, JB, el Portouguèche, Pouzette, Moussa, Armelle, Byran, Jéjé, mon boudin et tous ceux que j'oublie honteusement.

Enfin, je remercie ma mère, mon père, et ma sœur pour être toujours là si le besoin s'en fait sentir. Je profite de ces remerciements familiaux pour adresser une pensée à Albert qui est en somme un deuxième père pour moi.

Si je suis le *Créateur* des chapitres qui suivent, ce tout n'aurait pas pu voir le jour sans l'ensemble des personnes précédemment citées. Merci encore.

# TABLE DES MATIÈRES

# CHAPITRE 1

# INTRODUCTION

Sciences sociales, sciences médicales, sciences du comportement, domaine bancaire, chimiométrie,... bien rares sont les domaines où les jeux de données sont exempts de données manquantes. Face à des données incomplètes, l'attitude généralement adoptée par les utilisateurs est de limiter leur analyse statistique aux individus complets, méthode appelée *suppression par liste* ou *étude du cas complet* (Little et Rubin, 2002). Cette méthode est également utilisée par défaut dans la plupart des logiciels. Or, cette façon de procéder n'est pas satisfaisante pour au moins deux raisons. La première est que les individus complets ne constituent pas nécessairement un échantillon représentatif du jeu de données. Ceci implique que l'inférence menée via la méthode du cas complet est généralement biaisée. La seconde raison est que se limiter à un sous-échantillon amène à augmenter la variabilité des estimateurs utilisés. Ceci n'est pas un réel problème sur un jeu de données où le nombre d'individus incomplets ne représente qu'une minorité de l'ensemble des individus, mais le nombre d'individus incomplets tend à être rapidement élevé dès lors que le nombre de variables est grand. En effet, dans la mesure où chaque variable est sujet à être incomplète, la probabilité qu'un individu soit complet est faible. Par exemple, supposons que les données manquantes soient disposées complètement au hasard et que 5% des valeurs soient manquantes pour chaque variable, alors un jeu composé de 50 variables contient en moyenne 8% d'individus complets. Ainsi, la méthode du cas complet n'est généralement pas raisonnable et les données manquantes doivent faire l'objet d'une attention particulière.

Les premières solutions apportées par les chercheurs pour gérer simplement le problème des données manquantes ont été d'utiliser des méthodes d'*imputation simple*, c'est-à-dire de remplacer chaque donnée manquante par une valeur plausible. L'imputation est une façon commode de se ramener au cadre classique de données complètes sans pour autant supprimer des observations. Toutefois, l'objectif reste d'appliquer une méthode statistique sur un tableau incomplet et l'imputation effectuée n'est pas sans lien avec l'analyse statistique appliquée ensuite. La Figure 1 illustre différentes méthodes d'imputation pour imputer un jeu de données constitué de variables quantitatives $(X_1, X_2)$. Les données ont été générées selon une loi normale bivariée. Seule la variable $X_2$ du jeu contient des don-

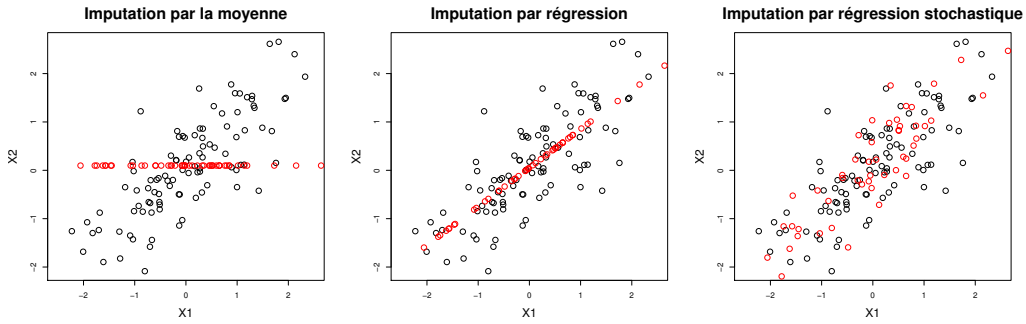


FIGURE 1 – Représentation d'un jeu de données bivarié imputé selon trois méthodes différentes : imputation par la moyenne, imputation par régression et imputation par régression stochastique. Les valeurs imputées sont représentées en rouge et les données complètement observées en noir.

nées manquantes, disposées de façon complètement aléatoire. Supposons que l'objectif ici soit d'inférer sur les paramètres de la loi normale. La figure de gauche représente le résultat d'une imputation par la moyenne. Il apparaît clairement qu'une telle méthode d'imputation n'est pas adaptée pour estimer la variance de $X_2$ d'une part, ou la covariance entre $X_1$ et $X_2$ d'autre part, car la liaison entre les deux variables n'a pas été utilisée pour l'imputation. En revanche, pour estimer les moyennes des variables, cette méthode peut convenir. La seconde méthode d'imputation est une imputation par régression. On a estimé les paramètres du modèle à partir des individus complets, puis on a imputé les données manquantes selon les valeurs prédites par le modèle de régression. Cette méthode tient compte des relations entre les variables, et permet d'estimer correctement la covariance entre les deux variables. En revanche, la variance de $X_2$ n'est pas assez reflétée ce qui amène à sous-estimer celle-ci et à surestimer le coefficient de corrélation entre les deux variables. L'inférence peut être largement améliorée en ajoutant un résidu aléatoire sur la prédiction par régression, ceci permet d'estimer correctement chacun des paramètres de la loi normale. Cette imputation par régression stochastique correspond à la figure de droite.

Le problème des méthodes d'imputation simple, même si elles sont adaptées à la méthode statistique employée par la suite, est qu'aucune distinction n'est faite entre les données observées et les données imputées. Or, les données imputées sont incertaines et cette incertitude n'est pas reflétée au travers des données imputées. Ceci a pour conséquence de sous-estimer la variabilité des estimateurs mis en œuvre. Pour corriger cela, la solution est d'imputer plusieurs fois le tableau de données de façons différentes, on parle alors d'*imputation multiple*. Pour autant, l'imputation multiple ne se résume pas à une succession d'imputations simples. En effet, les paramètres du modèle d'imputation sont estimés à partir des données et sont donc incertains. Cette incertitude sur les paramètres du modèle d'imputation doit être reflétée au travers des différents tableaux imputés. Dans ce but, chaque tableau est imputé selon des paramètres différents.

L'idée de l'imputation multiple remonte à 1977 où Donald Rubin a proposé cette tech-

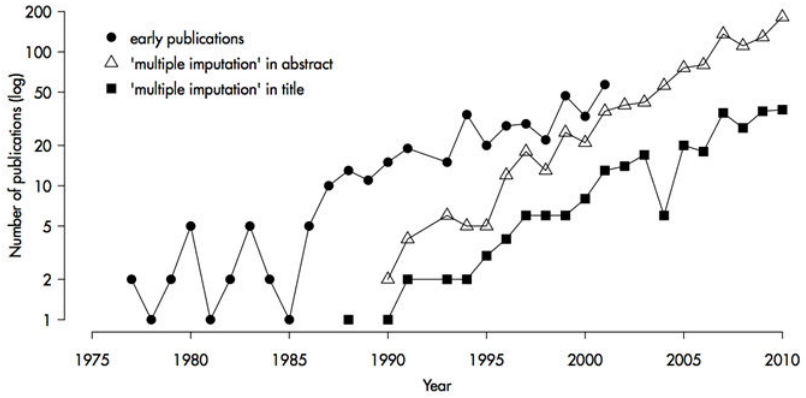


FIGURE 2 – Nombre de publications en échelle logarithmique à propos de l'imputation multiple entre 1977 et 2010 selon trois méthodes de comptage : nombre de publications dans la collection disponible à l'adresse `www.multiple-imputation.com` ; nombre de publications avec le terme "imputation multiple" dans le titre, le résumé ou les mots-clés ; nombre de publications avec le terme "imputation multiple" dans le titre uniquement. Source : `www.stefvanbuuren.nl`

nique dans le cadre d'une étude sur les ménages américains pour laquelle les données sur les salaires étaient souvent manquantes. Les premiers travaux théoriques sur l'imputation multiple ont ensuite vu le jour en 1987 (Rubin, 1987). Cette idée extrêmement novatrice a alors fait l'objet de critiques dans les années 90 (Fay, 1991, 1992, 1993, 1994, 1996) ce qui a permis d'améliorer encore la théorie sur la méthode. Depuis 2005, l'imputation multiple est acceptée par la communauté scientifique (Van Buuren, 2012), et le nombre de publications à ce sujet croît de façon exponentielle (*cf.* Figure 2).

Aujourd'hui, les méthodes d'imputation multiple sont nombreuses. Elle se différencient notamment par les modèles d'imputation qu'elles utilisent : modèles joints ou modèles conditionnels. L'imputation par un modèle joint consiste à faire l'hypothèse d'une distribution multivariée à l'ensemble des données. Les données manquantes sont alors imputées simultanément conditionnellement aux données observées. Le modèle Gaussien est par exemple le modèle joint le plus communément utilisé dans le cadre de données quantitatives.

L'imputation selon des modèles conditionnels, appelée *imputation par équations enchaînées* (Van Buuren, 2012; Van Buuren *et al.*, 2006; Van Buuren, 2014), quant à elle, impute les variables les unes après les autres. Par exemple, pour des variables quantitatives, chacun des modèles conditionnels peut être un modèle de régression multivarié. Les variables sont alors imputées successivement selon le modèle de régression qui leur est attribué. La procédure est répétée plusieurs fois de façon à obtenir une convergence. Plus précisément, l'imputation par équations enchaînées repose sur un algorithme de Gibbs (Geman et Geman, 1984; Gelfand et Smith, 1990). Celui-ci permet de simuler une distribution jointe à partir de la simulation successive des lois conditionnelles. La convergence évoquée est donc une convergence en distribution.

L'imputation par modèles joints et par équations enchaînées sont parfois équivalentes comme dans le cas où le modèle joint est Gaussien et les modèles conditionnels sont des modèles de régression incluant l'ensemble des variables (Hughes *et al.*, 2014). Toutefois, elles présentent des différences dans le cas général (Kropko *et al.*, 2014; Lee et Carlin, 2010).

Les modèles joints ont l'avantage de présenter des propriétés théoriques plus solides que l'imputation par équations enchaînées. En effet, l'imputation par équations enchaînées ne mène à une convergence vers une distribution jointe que si les modèles conditionnels sont *compatibles* (Besag, 1974), c'est-à-dire s'il existe une distribution jointe qui possède les distributions conditionnelles spécifiées, ce qui n'est pas toujours le cas. Aussi, même si les modèles conditionnels sont compatibles, la convergence n'est pas pour autant assurée (Roberts, 1996). L'approche par modèle joint est également plus rapide que les équations enchaînées qui nécessitent d'estimer les paramètres d'un nombre de modèles potentiellement grand, un par variable incomplète. De plus, cette estimation doit être répétée plusieurs fois pour atteindre la convergence.

Pour autant, il n'est pas forcément évident de proposer un modèle joint bien adapté à la structure des données. Par exemple une des variables peut avoir pour distribution une loi de Student (Van Buuren, 2012, p.66), tandis que les autres sont distribuées normalement. Le recours à des modèles conditionnels permet alors de proposer un modèle plus proche de la nature des données.

Un des problèmes actuels en imputation multiple est la construction de modèles joints pour des données non distribuées normalement, c'est par exemple le cas des données qualitatives ou mixtes. Le modèle loglinéaire a été proposé pour les variables qualitatives, le "general location model" pour les données mixtes (Schafer, 1997; Olkin et Tate, 1961), mais ces modèles d'imputation sont rapidement surparamétrés dès lors que le nombre de variables qualitatives excède la dizaine.

L'adaptation de l'imputation par le modèle Gaussien pour imputer des variables qualitatives a fait l'objet d'une littérature abondante (Allison, 2005; Bernaards *et al.*, 2007; Yucel *et al.*, 2008; Demirtas, 2009, 2010; Schafer, 1997; Horton *et al.*, 2003; King *et al.*, 2001). Les données qualitatives sont imputées selon la loi normale puis, soit laissées telles quelles, soit affectées à une modalité en arrondissant à la modalité la plus proche ou en effectuant un tirage. Ces méthodes se distinguent principalement par la façon d'affecter les modalités. Ces adaptations du modèle Gaussien permettent l'imputation de données qualitatives ou mixtes mais restent limitées à des variables qualitatives binaires ou ordonnées.

Une autre gamme de modèles joints repose sur l'hypothèse d'une structure latente Gaussienne (Boscardin *et al.*, 2008; He, 2012). Ces modèles combinent un modèle normal multivarié pour les variables quantitatives et un modèle probit multivarié pour les variables binaires et ordinales. Plus précisément, le modèle probit fait l'hypothèse de variables latentes Gaussiennes dont les variables binaires et ordinales constituent les variables observées. Ces modèles peuvent être vus comme des cas particuliers des copules Gaussiennes (Nelsen, 2006; Joe, 1997; Sklar, 1959). Les copules sont des modèles joints définis à partir des fonctions de répartition des lois marginales des

variables et d'une structure de dépendance liant ces variables entre elles. Par exemple, la dépendance pour une copule Gaussienne est définie par une matrice de covariance, mais d'autre forme de dépendance peuvent être utilisées via les copules de Gumbel, Krank, Clayton, etc. Les copules fournissent des modèles joints très souples, car les marginales peuvent être adaptées à la nature des données (loi Beta pour des proportions, loi de Bernoulli pour des variables binaires, etc). Mais en dehors des travaux de Boscardin *et al.* (2008) et He (2012) l'imputation par les copules n'a été étudiée qu'en tant que méthode d'imputation simple (Di Lascio *et al.*, 2015; Käärik et Käärik, 2009). Ces modèles restent également limités à des variables quantitatives, binaires, ou ordinales.

Une autre difficulté en imputation multiple est l'imputation de jeux de données dont le nombre de variables est important, ou au moins important relativement au nombre d'individus. En effet, la prise en compte des relations entre toutes les variables amène à utiliser des modèles complexes qui deviennent rapidement surparamétrés, amenant à des problèmes d'inférence ou à des problèmes d'ordre informatique. Les modèles de référence sont notamment des modèles rapidement surparamétrés quand le nombre de variables est grand : modèle normal pour les données quantitatives, modèle log-linéaire pour les données qualitatives, general location model pour les données mixtes (Schafer, 1997).

Pour faire face à ce problème, deux solutions sont classiquement envisagées. La première est de découper le problème en imputant les variables les unes après les autres par équations enchaînées. D'une part, ceci permet de limiter le nombre de paramètres estimés à chaque étape, ce qui peut résoudre des problèmes informatiques. Par exemple cela évite de stocker l'intégralité des paramètres en mémoire. D'autre part, il est possible de sélectionner les variables retenues pour chaque modèle conditionnel ce qui permet de réduire globalement le nombre de paramètres estimés. Toutefois, les équations enchaînées nécessitent un temps de calcul important quand le nombre de variables est élevé.

L'autre solution est d'utiliser un modèle joint pour l'ensemble des variables mais en imposant des contraintes sur les paramètres du modèle. Différents travaux ont été effectués dans ce sens. Song et Belin (2004) ont proposé d'utiliser un modèle d'analyse en facteurs pour des variables quantitatives, le nombre de paramètres du modèle est ainsi contrôlé par le nombre de facteurs retenus. He (2012) a étendu ce modèle à des données à la fois quantitatives, binaires, et ordinales. Inspiré des travaux de Boscardin et Zhang (2004) et de ceux de (Zhang *et al.*, 2008), He (2012) a également proposé de réduire la dimension du modèle à structure latente Gaussienne en utilisant une modélisation Bayésienne contraignant la distribution de la matrice de variance covariance. Liu (2010) a repris le modèle d'analyse en facteurs pour l'appliquer à des données quantitatives, binaires ou ordonnées, mais cette fois-ci structurées par thèmes. Les variables de chaque thème y sont supposées indépendantes conditionnellement aux facteurs latents continus, propres à chaque thème. Cependant, chacun de ces modèles stipule une distribution normale pour les variables. Ainsi, les extensions proposées restent limitées à des variables binaires ou qualitatives ordonnées et la présence de variables qualitatives nominales demeure problématique sur des jeux de données où le nombre de variables est important.

Enfin, un des enjeux important en imputation multiple est la prise en compte de liaisons complexes dans l'imputation, typiquement des effets d'interaction. A titre d'exemple, une imputation par la loi normale ne permet pas de préserver des interactions. Une solution pour les préserver est de découper le jeu de données et d'imputer indépendamment chaque partie (Carpenter et Kenward, 2013, p.147-151). Par exemple s'il existe une interaction des variables $X_1$ et $X_2$ sur $X_3$, et que $X_1$ est qualitative, alors on peut découper le jeu de données selon $X_1$ et imputer indépendamment chaque jeu. Cette solution s'adapte à des données qualitatives mais suppose aussi que les paramètres du modèle puissent être estimés pour un faible nombre d'individus. Une autre possibilité est d'utiliser des équations enchaînées et de spécifier des modèles d'imputation avec interactions, mais ceci peut être fastidieux si le nombre de modèles est important. Le problème de la prise en compte des effets d'interaction suscite un intérêt particulier de la communauté scientifique ces derniers temps (Vermunt *et al.*, 2008; Vidotto *et al.*, 2014; Si et Reiter, 2013; Murray et Reiter, 2014; Doove *et al.*, 2014; Shah *et al.*, 2014). Si et Reiter (2013) ont notamment proposé une méthode d'imputation à partir d'un modèle de mélange pour des variables qualitatives. Le modèle utilisé est un modèle à classes latentes. Celui-ci repose sur l'hypothèse d'une indépendance des variables conditionnellement à la classe à laquelle appartiennent les individus. La modélisation Bayésienne adoptée pour ce modèle a l'avantage de choisir de façon automatique le nombre de classes. Ce modèle permet de prendre en compte les effets d'interaction, mais l'hypothèse d'indépendance conditionnelle rend le modèle inadapté à des jeux de données où les liaisons entre variables sont fortes (Marbac-Lourdelle, 2014). Dans le cadre des équations enchaînées, des modèles conditionnels utilisant les forêts aléatoires (Breiman, 2001) ont récemment été étudiés (Doove *et al.*, 2014; Shah *et al.*, 2014). Les forêts aléatoires permettent de prendre en compte des liaisons complexes entre variables mais leur coût calculatoire est prohibitif.

L'objet de cette thèse est de proposer de nouvelles méthodes d'imputation multiples qui sont basées sur des modèles qui réduisent la dimension. Il s'agit des méthodes d'analyse factorielle, utilisées initialement pour l'analyse exploratoire de données multidimensionnelles. Ces méthodes sont à l'origine purement géométriques, elles consistent à rechercher un sous-espace sur lequel l'inertie projetée du nuage de points est maximale. Chacune d'entre elles est adaptée à un type de données particulier : l'analyse en composante principales (ACP) permet d'analyser un jeu de données constitué de variables quantitatives (Jolliffe, 2002) ; l'analyse des correspondances multiples (ACM) permet elle d'analyser un jeu de données qualitatives (Greenacre et Blasius, 2006; Lebart *et al.*, 1984) ; l'analyse factorielle des données mixtes (AFDM) permet de traiter des variables à la fois quantitatives et qualitatives (Escofier, 1979; Kiers, 1991; Pagès, 2015) ; l'analyse factorielle multiple (AFM) analyse des variables de n'importe quelle nature, structurées par thèmes (Pagès, 2015). Bien que ces méthodes s'appliquent sur des variables de natures différentes, elles peuvent toutes être vues comme des méthodes de projection pour des métriques particulières. Ainsi, l'étude de ces méthodes en tant que méthodes d'imputation offre de grandes perspectives en termes de diversité du type de données imputées d'une part, et en termes de dimensions de jeux de données imputés d'autre part, car, via

la réduction de la dimension, le nombre de paramètres estimés est faible. Le travail effectué dans cette thèse offre des réponses à différents problèmes actuels rencontrés en imputation multiple : proposition de modèles joints pour des données qui s'éloignent de la normalité, gestion d'un nombre de variables important, prise en compte des interactions.

Tout d'abord, trois modèles d'imputation joints sont proposés : imputation par ACP pour des données quantitatives, ACM pour des variables qualitatives, AFDM pour des données mixtes. Ces modèles d'imputation sont étudiés simultanément au travers de l'imputation simple par AFDM. La qualité de prédiction de ces modèles est étudiée par simulation et comparée à l'imputation simple par forêts aléatoires (Stekhoven et Bühlmann, 2012) considérée comme la méthode de référence pour l'imputation simple de données mixtes. Les propriétés de l'imputation par analyse factorielle sont ainsi mises en avant. En particulier, la qualité de prédiction sur des jeux réels est étudiée, permettant d'illustrer les performances de ces modèles quand les liaisons entre variables sont potentiellement complexes et où le nombre d'individus est parfois inférieur au nombre de variables.

Ensuite, une méthode d'imputation multiple pour des données quantitatives est présentée. Il s'agit d'une méthode d'imputation reposant sur la modélisation Bayésienne du modèle d'ACP proposée par Verbanck *et al.* (2013). La méthode est comparée aux méthodes les plus performantes pour imputer ce type de données : imputation par équations enchaînées utilisant des régressions (Van Buuren, 2012) et imputation par modèle joint Gaussien (King *et al.*, 2001). La comparaison s'appuie sur une étude très approfondie à la fois sur des jeux simulés et sur des jeux réels. Cette méthode d'imputation offre une solution au problème des jeux de données où le nombre de variables est grand et où le nombre d'individus ne l'est pas nécessairement.

Enfin, une méthode d'imputation multiple pour des données qualitatives binaires, ordinales ou nominales est proposée. La méthode d'imputation est basée sur l'ACM. Cette méthode est comparée à 5 méthodes différentes sur la base de jeux réels : imputation par modèle joint selon la loi normale (King *et al.*, 2001), imputation par le modèle joint log-linéaire (Schafer, 1997), imputation par le modèle joint à classes latentes (Si et Reiter, 2013), imputation par équations enchaînées utilisant des modèles de régression logistiques (Van Buuren, 2012), ou les forêts aléatoires (Shah *et al.*, 2014). Ces comparaisons mettent en évidence que l'imputation multiple par ACM est applicable sur des jeux où les variables comportent de nombreuses modalités où que le nombre d'individus est petit devant le nombre de variables. De plus, la méthode a le gros avantage d'être applicable en un temps très raisonnable.

Nous commencerons dans le Chapitre 2 par préciser le type de données manquantes auxquelles nous nous intéresserons et repositionnerons l'imputation multiple par rapport aux autres méthodes reconnues pour effectuer de l'inférence dans ce cadre. Nous insisterons notamment sur l'importance d'un point de vue théorique à savoir prendre en compte un grand nombre de variables pour inférer dans ce contexte et renforcerons ainsi le poids du problème du nombre de variables en imputation multiple. A partir des chapitres suivants, nous nous intéresserons aux méthodes d'imputation par les méthodes d'analyse factorielle. Le Chapitre 3 présentera les méthodes d'imputation simple et s'appuiera sur l'article Audigier *et al.* (2014) (à paraître dans *Advances in Data Analysis and Classifica-*

*tion*). Cela permettra d'illustrer les propriétés des modèles d'imputation, indispensables en vue de mener une inférence. Une fois ces propriétés établies, nous nous intéresserons aux méthodes d'imputation multiple. La difficulté pour passer de l'imputation simple à l'imputation multiple repose sur la façon de refléter l'incertitude sur les paramètres du modèle d'imputation. Cette difficulté nous amènera à présenter dans un premier temps, en Chapitre 3, l'imputation multiple de variables quantitatives par ACP. L'incertitude sur les paramètres sera reflétée par une approche Bayésienne. Ce chapitre intégrera l'article Audigier *et al.* (2015b) (à paraître dans *Journal of Statistical Computation and Simulation*). Puis, le Chapitre 4 portera sur l'imputation multiple de variables qualitatives par ACM. L'incertitude sur les paramètres sera prise en compte via une approche Bootstrap. Ce chapitre intègrera l'article (Audigier *et al.*, 2015a) (en révision mineure pour la revue *Statistics and Computing*).

# CHAPITRE 2

# LES DONNÉES MANQUANTES ET LEUR GESTION

Dans ce chapitre, l'imputation multiple est présentée en détail et positionnée par rapport aux autres méthodes permettant d'effectuer une inférence en présence de données manquantes : approches par pondération et approches par vraisemblance. Les caractéristiques des données manquantes auxquelles ces méthodes font appel sont d'abord rappelées et illustrées par des exemples. Les méthodes sont ensuite présentées dans le cadre de données manquantes au hasard. En particulier, l'imputation multiple y est détaillée en s'appuyant sur les outils développés dans le cadre des approches par vraisemblance. Elle apparaît comme la méthode d'inférence la plus en phase avec l'hypothèse de données manquantes au hasard dans la mesure où les modèles d'imputation peuvent considérer un grand nombre de variables.

**Contents**



# 1 Classification des données manquantes

Les données manquantes peuvent être classifiées selon leur répartition, appelée *dispositif*, et leur origine, appelée *mécanisme*. Afin de préciser ces termes, il est nécessaire d'introduire dans un premier temps le vocabulaire propre au domaine des données manquantes.

## 1.1 Terminologie

On considère un tableau $\mathbf{X}_{n\times p}$ à $n$ lignes et $p$ colonnes, appelé matrice des *données complètes*. On considère également $\mathbf{R} = (r_{ik})_{\substack{1\leq i\leq n\\ 1\leq k\leq p}}$ la matrice indiquant la présence ou l'absence de données manquantes telle que $r_{ik} = 1$ si l'on observe la valeur pour l'individu $i$ sur la variable $k$ et $r_{ik} = 0$ sinon. $\mathbf{R}$ est appelée le *dispositif* des données manquantes. $\mathbf{R}$ est parfaitement connue et permet de partitionner la matrice $\mathbf{X}$ en sa partie observée, notée $\mathbf{X}^{obs}$, et sa partie non observée notée $\mathbf{X}^{miss}$. On note $\mathbf{X} = \left(\mathbf{X}^{obs}, \mathbf{X}^{miss}\right)$. Les *données observées*, dont dispose le chercheur, correspondent alors au couple ($\mathbf{X}^{obs}$,$\mathbf{R}$). Les différentes formes que peut prendre $\mathbf{R}$, indépendamment de $\mathbf{X}$ joue un rôle important dans la gestion des données manquantes.

Par ailleurs, les matrices **X** et **R** peuvent être présentées d'un point de vue probabiliste comme les réalisations de variables aléatoires $X$ et $R$ respectivement. La variable $R$ est appelée *mécanisme générant les données manquantes*. On note alors $f(X, R; \gamma, \psi)$ la densité du couple de variables aléatoires $(X, R)$ où $\gamma$ désigne l'ensemble des paramètres qui régit la distribution de $X$ et $\psi$ celui qui régit la distribution de $R$. La distribution des données observées est obtenue en intégrant la densité sur les données manquantes :

$$f\left(X^{obs}, R; \gamma, \psi\right) = \int f\left(X, R; \gamma, \psi\right) dX^{miss}. \tag{1}$$

Inférer sur les paramètres de cette distribution est complexe dans la mesure où il est nécessaire de modéliser la loi jointe des données et du mécanisme. Selon le mécanisme considéré, des simplifications importantes peuvent être opérées.

## 1.2 Dispositifs de données manquantes

La Figure 1 illustre différents dispositifs de données manquantes. Ainsi on distingue d'une

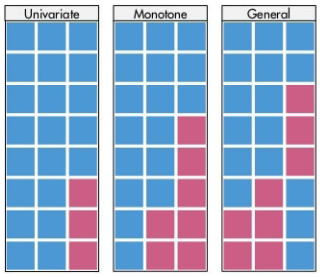


FIGURE 1 – Illustration de différents dispositifs de données manquantes : univarié, monotone, quelconque. Les données observées sont en bleu, les données manquantes en rouge. Source `www.stefvanbuuren.nl`.

part les configurations *univariées*, où une seule variable possède des données manquantes, des configurations *multivariées*, où plusieurs variables sont incomplètes. Le premier dispositif en Figure 1 est univarié alors que les autres sont multivariés. Cette distinction est importante. Par exemple si on souhaite estimer les paramètres d'un modèle de régression, où seule la variable réponse est manquante, alors il est envisageable d'estimer ces paramètres à partir des individus complètement observés. En effet, eux seuls nous renseignent sur le lien entre les covariables et la réponse. En revanche, si le dispositif est multivarié, alors des individus peuvent posséder des données manquantes sur une des covariables. Ces individus incomplets apportent une information sur le lien entre la réponse et les covariables observées, mais il paraît plus difficile d'estimer les paramètres du modèle en considérant ces individus.

D'autre part, on distingue les configurations *monotones* des configurations *non-monotones*. Une configuration est dite monotone s'il existe une permutation des colonnes de **R** telle que si $\mathbf{X}_k$ est manquant, alors les autres variables $\mathbf{X}_{k'}$ telles que $k < k'$ sont également manquantes. Ce type de configuration est fréquent quand les variables

correspondent à une mesure à différents temps d'observation. En effet, un individu qui sort de l'étude au temps $k$ est manquant aux temps suivants. Les deux premières configurations de la Figure 1 sont monotones alors que la dernière ne l'est pas. Les dispositifs monotones offrent généralement des simplifications théoriques importantes. En particulier, les méthodes de pondérations abordées dans la Section 2.1 sont adaptées à ce type de dispositif.

## 1.3 Mécanismes à l'origine des données manquantes

Les mécanismes à l'origine des données manquantes peuvent être classés en trois groupes : les données générées complètement au hasard dites *MCAR* pour *missing completely at random*, les données générées au hasard dites *MAR* pour *missing at random*, et les données non générées au hasard, dites *NMAR* pour *non missing at random* (Rubin, 1976; Little, 1995). Les méthodes employées pour gérer les données manquantes sont conditionnées par le type de mécanisme qui affecte le jeu de données.

### 1.3.1 Mécanisme MCAR

On appelle *données manquantes générées complètement au hasard* des données manquantes dont la probabilité d'occurrence est sans lien avec les données complètes. Ce type de mécanisme est typiquement rencontré dans les enquêtes où des individus sont amenés à remplir un questionnaire. En effet, il se peut que certains individus aient oublié de répondre à des questions ou que certaines réponses n'aient pas été saisies manuellement au moment de la numérisation des questionnaires papier. Sous l'hypothèse MCAR, le mécanisme $R$ est donc indépendant de $X^{obs}$ et $X^{miss}$, on a alors

$$f\left(R|X^{obs}, X^{miss}; \gamma\right) = f\left(R; \gamma\right). \tag{2}$$

Cette hypothèse sur le mécanisme, même si elle est légitime dans certains cas, est assez forte.

Un mécanisme MCAR permet de considérer les individus complets comme un sous-échantillon représentatif des individus du jeu de données. Ainsi, les données manquantes complètement au hasard permettent d'appliquer les méthodes statistiques usuelles sur le jeu de données restreint à ses individus complètement observés sans engendrer de biais. Toutefois, cela conduit à réduire la taille de l'échantillon considéré et donc à construire des estimateurs avec des variances plus grandes.

### 1.3.2 Mécanisme MAR

De façon intuitive, les données générées au hasard correspondent à des données manquantes générées indépendamment des données elles-mêmes ($\mathbf{X}^{miss}$) mais pouvant dépendre de la partie observée ($\mathbf{X}^{obs}$). Un exemple de données MAR est le cas d'une enquête de satisfaction auprès d'actifs sans emploi vis-à-vis du service offert par l'Agence pour l'emploi. Une première série de questions est posée en Janvier aux bénéficiaires. On suppose que toutes les personnes ont répondu à ces questions. Les personnes ayant répondu "non" à la question "êtes vous globalement satisfait du service proposé ?", sont alors

soumises une nouvelle fois à cette série de questions en Février, les autres personnes ne sont pas réinterrogées. Le jeu de données est constitué des réponses aux questions posées en Janvier et Février. La présence de données manquantes sur les données de Février dépend de la réponse en Janvier à la question “êtes vous globalement satisfait du service proposé”. Le mécanisme est donc MAR. De façon formelle, l'hypothèse MAR est définie par :

$$f\left(R|X^{obs}, X^{miss}; \gamma\right) = f\left(R|X^{obs}; \gamma\right) \tag{3}$$

Sous ce mécanisme, les individus complets ne sont plus un sous-échantillon représentatif de la population et l'inférence sur ces individus peut être biaisée.

L'hypothèse MAR généralise l'hypothèse MCAR, mais est moins restrictive. Elle permet de factoriser la densité des données observées (Equation 1) de la façon suivante :

$$\begin{aligned} f\left(X^{obs}, R; \gamma, \psi\right) &= f\left(R|X^{obs}; \psi\right) \times \int f\left(X; \gamma\right) dX^{miss} \\ &= f\left(R|X^{obs}; \psi\right) f\left(X^{obs}; \gamma\right) \end{aligned} \tag{4}$$

Cette factorisation simplifie donc l'expression de la distribution des données observées en l'exprimant comme un facteur dépendant des paramètres d'intérêt $\gamma$ et un autre dépendant des paramètres de nuisance $\psi$ lié aux données manquantes.

L'hypothèse MAR peut parfois être évidente, comme dans l'exemple précédent, quand on connaît le mécanisme générant les données manquantes, mais en général les données manquantes ne dépendent pas de l'expérimentateur et on ne sait pas si cette hypothèse est vérifiée. De plus, il est impossible de vérifier cette hypothèse (Fitzmaurice *et al.*, 2014, p.9). Ainsi, à défaut de pouvoir la vérifier, il est recommandé d'inclure des *variables auxiliaires* dans le jeu de données (Schafer, 1997; van der Palm *et al.*, 2014). On entend par là des variables (avec peu ou pas de données manquantes) qui ne sont pas d'un intérêt scientifique, mais qui permettent d'expliquer la présence de données manquantes et rendre ainsi l'hypothèse MAR valide.

### 1.3.3 Mécanisme NMAR

Par opposition au mécanisme MAR, un mécanisme est dit NMAR si la probabilité d'apparition de données manquantes est liée à la partie non observée des données $X^{miss}$. Ces mécanismes sont fréquents dans le cadre d'enquêtes sur des sujets sensibles comme les revenus, la consommation d'alcool, l'usage de drogues, etc. Par exemple, un individu fortuné aura plutôt tendance à ne pas répondre à une question portant sur ses revenus. La probabilité d'apparition d'une donnée manquante est ici liée à la partie non observée des données. Plus formellement, un mécanisme est dit NMAR si $f\left(R|X^{obs}, X^{miss}, \gamma\right) \neq f\left(R|X^{obs}, \gamma\right)$. Dans ce cas, l'expression de la distribution des données observées (Equation 1) ne se simplifie pas et l'inférence nécessite généralement de modéliser le couple $(X, R)$. Dans ce cas la modélisation est d'une part plus complexe, et d'autre part, l'inférence sur les paramètres du modèle est plus difficile car le nombre de paramètres à estimer est alors accru, ce qui peut poser des problèmes d'identifiabilité.

Bien que ce mécanisme soit le plus général, les travaux pour gérer les données manquantes dans ce cadre sont assez spécifiques à leur application (Allison, 2002, p.5) et la généralisation reste un sujet de recherche assez ouvert. Par la suite, nous décrivons uniquement les méthodes pour gérer les données manquantes pour des mécanismes MAR (et a fortiori MCAR).

# 2 Méthodes pour gérer les données manquantes

Trois grandes familles de méthodes sont communément utilisées pour inférer à partir de jeux incomplets : les approches par pondération, les approches basées sur la vraisemblance et les méthodes d'imputation multiple. La philosophie de ces méthodes est assez différente. Les méthodes par pondération, parfois qualifiées de semi-paramétriques, portent sur une modélisation du mécanisme à l'origine des données manquantes. Elles consistent à repondérer les individus complètement observés par rapport à leur probabilité de présenter des données manquantes. Au contraire les deux autres familles de méthodes modélisent la loi jointe des données observées. Les approches par vraisemblance visent à inférer directement sur les paramètres de cette loi tandis que les approches par imputation passent par une étape de simulation pour remplacer les données manquantes et inférer ensuite sur le jeu rendu complet.

## 2.1 Approches par pondération

Le principe des approches par pondération pour gérer les données manquantes consiste à attribuer un poids aux individus sans données manquantes de façon à corriger le biais observé dans la mise en œuvre de la méthode du cas complet. Cette idée est issue de la théorie des sondages (Horvitz et Thompson, 1952) où, pour limiter le nombre de personnes interrogées, la probabilité qu'un individu fasse partie de l'étude n'est pas uniforme mais définie selon un plan, dit *plan de sondage*. Par la suite, pour inférer sur la population, les individus sont repondérés de façon inversement proportionnelle à la probabilité qu'ils avaient de faire partie de l'échantillon. Typiquement, si une population est divisée en $J$ strates homogènes de taille $N_j$, et que $n_j$ individus sont sélectionnés dans cette strate, alors un individu $i$ issu de la strate $j$ représente $\pi_i = \frac{N_j}{n_j}$ unités de la population. Ainsi, si on cherche à estimer la valeur moyenne d'une variable $Y$ de la population, un tel individu sera pondéré par un poids $w_i$ inversement proportionnel à $\pi_i$ :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i \tag{5}$$

avec

$$w_i = \frac{n\pi_i^{-1}}{\sum_{k=1}^{n} \pi_k^{-1}} \tag{6}$$

Notons que l'expression de $w_i$ correspond à une normalisation de $\pi_i^{-1}$ de façon à ce que les poids somment au nombre d'unités tirées $n$.

Cette approche s'étend naturellement au cas où certaines unités ont des données manquantes sur $Y$. Notant $\phi$ la probabilité de réponse d'une unité sélectionnée, la probabilité de tirer une unité $i$ et que celle-ci réponde est donnée, selon la règle de Bayes, par le produit $\pi_i\phi_i$. On a alors

$$\bar{y} = \frac{1}{r}\sum_{i=1}^{r} w_i y_i \tag{7}$$

avec $r$ le nombre de répondants et

$$w_i = \frac{r\left(\pi_i\phi_i\right)^{-1}}{\sum_{k=1}^{r}\left(\pi_k\phi_k\right)^{-1}} \tag{8}$$

Dans le cas classique où l'échantillon dont on dispose est représentatif de la population, et non issu d'un plan de sondage, on pose $\pi_i = 1$ pour tout $i$. En pratique, la probabilité de réponse $\phi$ n'est pas connue et doit donc être estimée. Pour ce faire, la façon la plus simple consiste à classifier les unités sur la base de variables complètes, éventuellement en utilisant des variables auxiliaires, et estimer $\phi$ par la proportion de données manquantes au sein de la classe.

Ainsi, les méthodes de pondération nécessitent de modéliser le mécanisme à l'origine des données manquantes car c'est lui qui permet de définir les poids utilisés. En revanche, il n'est pas nécessaire de spécifier la distribution des données. Ceci est une spécificité importante des méthodes par pondération. Dans l'exemple précédent aucune hypothèse sur la distribution des données n'a été faite, on a simplement estimer l'espérance de la variable $Y$. Ces méthodes de pondération des observations sont attrayantes par leur aspect intuitif car elles consistent à reprendre les estimateurs classiques en affectant un poids aux individus.

Toutefois, les raisons pour lesquelles les données sont manquantes sont souvent mal connues (Brick, 2013; Rotnitzky, 2009) ce qui rend l'estimation des poids très difficile en pratique. L'amélioration des méthodes de pondération par le biais des méthodes dites "double robust" (Scharfstein *et al.*, 1999) motive de nombreuses recherches récentes. Ces méthodes reposent d'une part sur la modélisation du mécanisme générant les données manquantes et d'autre part sur la modélisation des données complètes. Elles proposent un moyen de construire un estimateur consistant pour une quantité d'intérêt (par exemple une espérance) si l'une ou l'autre de ces modélisations est correcte (Tsiatis, 2006, p.147).

Si les méthodes "double robust" ont des propriétés théoriques très intéressantes, leurs performances en pratique ont fait l'objet de critiques dès lors que les modèles choisis pour les données et le mécanisme ne sont plus parfaitement spécifiés (Kang et Schafer, 2007). L'amélioration de ces méthodes fait l'objet de recherches actuelles (voir Rotnitzky et Vansteelandt (2014) pour une revue détaillée).

Plus généralement, les méthodes de pondération sont beaucoup moins attrayantes en présence d'un dispositif multivarié et en particulier quand celui-ci est non-monotone. Il est alors nécessaire de définir une pondération des individus différente pour chaque variable

incomplète (Schafer et Graham, 2002) et la mise en œuvre de ces méthodes devient plus compliquée (Tsiatis et Davidian, 2014, p.179). Aussi, les poids des individus sont estimés à partir des données, mais généralement, l'incertitude portant sur ces estimations n'est pas prise en compte dans le calcul de la variabilité des estimateurs. Les conséquences d'un point de vue pratique sont obscures (Little et Rubin, 2002, p.53).

Ces méthodes suscitent l'intérêt de différentes communautés comme en biostatistique où les dispositifs monotones sont fréquents (Prague *et al.*, 2015), ou dans la communauté des sondages (Haziza *et al.*, 2012) où ces méthodes sont utilisées depuis longtemps. Toutefois, nous ne nous y intéresserons pas davantage, notamment parce qu'elles sont construites sur l'hypothèse que l'on dispose de cas complets.

## 2.2 Approches basées sur la vraisemblance

Contrairement aux approches par pondération, les approches basées sur la vraisemblance visent à inférer directement sur le paramètre $\gamma$ de la distribution jointe des données (ou sur une fonction de ce paramètre) en présence de données manquantes. Notons que cet objectif est commun à celui des méthodes d'imputation simple qui nécessitent de définir les paramètres du modèle qui sert à imputer en présence de données manquantes (*cf.* Figure 1 Chapitre 1). D'un point de vue théorique, ces approches sont les mêmes que celles utilisées dans le cadre standard sans données manquantes, à savoir l'estimation par maximum de vraisemblance et l'inférence Bayésienne. La différence entre ces deux approches réside dans la définition même d'un paramètre. L'inférence par maximum de vraisemblance se place sous le paradigme fréquentiste où le paramètre est vu comme une caractéristique d'une population. Le but est d'estimer la vraie valeur du paramètre et d'y associer un intervalle de confiance.

Au contraire, le paradigme Bayésien voit le paramètre $\gamma$ régissant la distribution des données comme une variable aléatoire possédant une distribution et non comme une valeur fixe. L'objectif ici est de déterminer la distribution du paramètre conditionnellement aux données observées en se basant sur la distribution marginale de celui-ci dite *loi a priori*. Cette distribution conditionnelle aux données, la loi *a posteriori*, nous renseigne alors sur la variabilité du paramètre par rapport à sa valeur moyenne. Ainsi, l'espérance de cette loi a posteriori fournit une estimation ponctuelle de ce paramètre et permet de construire un intervalle de crédibilité associé.

En présence de données manquantes, la distribution des données observées $\left(X^{obs}, R\right)$ ne dépend pas uniquement de $\gamma$ mais également de $\psi$ ce qui complexifie le problème d'inférence. Plus précisément, l'approche par maximum de vraisemblance consiste à maximiser la *vraisemblance des données observées*. Sous l'hypothèse de réalisations indépendantes, la vraisemblance des données observées est donnée par :

$$L\left(\gamma, \psi; X^{obs}, R\right) = \prod_{i=1}^{n} f\left(x_i^{obs}, r_i; \gamma, \psi\right). \tag{9}$$

Dans le cadre Bayésien, il s'agit de déterminer la loi a posteriori des paramètres

$$p\left(\gamma, \psi | X^{obs}, R\right) = \frac{f\left(X^{obs}, R | \gamma, \psi\right) p\left(\gamma, \psi\right)}{\int\int f\left(X^{obs}, R | \gamma, \psi\right) p\left(\gamma, \psi\right) d\gamma d\psi} \tag{10}$$

où $p(\gamma, \psi)$ est la loi a priori sur les paramètres. Cette loi a posteriori est alors proportionnelle au produit de la loi conditionnelle du couple $\left(X^{obs}, R\right)$ sachant $\gamma, \psi$, en d'autres termes la vraisemblance observée, multipliée par la loi a priori sur les paramètres.

L'inférence sur $\gamma$, basée sur la vraisemblance observée ne semble pas pouvoir être effectuée indépendamment du paramètre de nuisance $\psi$ alors que celui-ci n'est pas d'un intérêt scientifique ici. Ceci est possible sous l'hypothèse que le mécanisme est *ignorable*.

### 2.2.1 Ignorabilité

Un mécanisme est dit ignorable s'il vérifie deux propriétés : celle de générer des données manquantes *au hasard* et celle de *distinction* des paramètres (Little et Rubin, 2002; Rubin, 1987).

L'hypothèse MAR a déjà été présentée en Section 1.3.2. Sous cette hypothèse on a

$$f\left(X^{obs}, R; \gamma, \psi\right) = f\left(R|X^{obs}; \psi\right) f\left(X^{obs}; \gamma\right). \tag{11}$$

Ainsi, si l'espace des paramètres joint $(\psi, \gamma)$ est le produit cartésien des espaces marginaux, c'est-à-dire que la connaissance de $\psi$ ne donne aucune information sur $\gamma$ et réciproquement, alors la vraisemblance des données observées est proportionnelle à la vraisemblance ignorant le mécanisme à l'origine des données manquantes $L\left(\gamma|X^{obs}\right)$

$$L\left(\gamma, \psi|X^{obs}, R\right) \propto L\left(\gamma|X^{obs}\right). \tag{12}$$

Sous cette deuxième condition de *distinction* des paramètres du mécanisme et des paramètres régissant les données, un mécanisme MAR est alors dit *ignorable*.

Dans un cadre Bayésien, les paramètres $\gamma$ et $\psi$ ne sont plus fixes mais vus comme des variables aléatoires. Ainsi, cette deuxième condition s'exprime comme une indépendance entre les loi a priori de $\gamma$ et $\psi$. Si le mécanisme est ignorable, on a alors

$$p\left(\gamma, \psi|X^{obs}, R\right) \propto f\left(X^{obs}|\gamma\right) p\left(\gamma\right) \tag{13}$$

En présence d'un mécanisme ignorable, il n'est plus nécessaire de modéliser la distribution du mécanisme à l'origine des données manquantes, ou plus exactement, il n'est plus nécessaire de modéliser la loi jointe de $(X, R)$, mais seulement une des distributions marginales.

Dans la plupart des situations, l'hypothèse que les paramètres du mécanisme n'apportent pas d'information sur les paramètres régissant les données complètes est raisonnable, la réciproque l'est aussi (Schafer, 1997; Little et Rubin, 2002). Ainsi, l'hypothèse de paramètres distincts admise. En pratique, l'hypothèse d'un mécanisme ignorable et de données MAR sont donc utilisées de façon interchangeable, raison pour laquelle les mécanismes MAR et MCAR sont qualifiés d'ignorables. Cette hypothèse d'ignorabilité sera également faite dans le cadre de l'imputation multiple.

Toutefois, l'ignorabilité n'est pas indispensable pour pouvoir inférer sans modéliser le mécanisme. Par exemple considérons un jeu de données bivarié, où seule la seconde variable possède des données manquantes. Supposons que le mécanisme à l'origine des données manquantes est lié uniquement à la partie non-observée des données. L'hypothèse MAR n'est donc pas vérifiée, le mécanisme n'est pas ignorable. Pour autant, il apparaît clairement qu'il n'est pas nécessaire de modéliser le mécanisme à l'origine des données manquantes pour inférer sur la loi marginale de la première variable étant donné que celle-ci est complète.

L'hypothèse MAR telle qu'elle a été définie fait référence à l'ensemble des paramètres $\gamma$ qui régissent les données, mais on voit ici que si l'inférence porte sur un sous-ensemble de ces paramètres, alors celle-ci reste appropriée sans avoir besoin modéliser le mécanisme. Pour cette raison, Little et Zanganeh (2013) ont récemment introduit la définition de mécanisme *MAR pour un paramètre*. Si on écrit $\gamma = (\gamma_1, \gamma_2)$ où $\gamma_1$ et $\gamma_2$ sont des sous ensembles de paramètres, le mécanisme est dit MAR pour $\gamma_1$ si la vraisemblance observée peut être décomposée sous la forme :

$$L\left(\gamma_1, \gamma_2, \psi | X^{obs}, R\right) = L\left(\gamma_1 | X^{obs}, R\right) \times L\left(\gamma_2, \psi | X^{obs}, R\right). \tag{14}$$

Si de plus $\gamma_1$ et $(\gamma_2, \psi)$ sont distincts, alors celui-ci est dit *ignorable pour* $\gamma_1$. Dans le cadre Bayésien, si le mécanisme est MAR pour $\gamma_1$ et que les lois a priori de $\gamma_1$ et $(\gamma_2, \psi)$ sont indépendantes, alors le mécanisme est ignorable pour $\gamma_1$ et la loi a posteriori pour $\gamma_1$ est donnée par :

$$p\left(\gamma_1 | X^{obs}, R\right) \propto f\left(X^{obs}, R | \gamma_1\right) p\left(\gamma_1\right). \tag{15}$$

Cette définition permet d'élargir les cas où on peut toujours mener une inférence sans modéliser le mécanisme à l'origine des données manquantes. Par exemple, dans un cas moins trivial que l'exemple précédent, si $X_1$ et $X_2$ sont deux groupes de variables, éventuellement incomplets, tels que la probabilité d'apparition de données manquantes sur le premier groupe dépend des données observées sur le premier groupe uniquement, mais que la probabilité d'apparition de données manquantes sur le second groupe dépend de l'ensemble des variables des deux groupes, observées ou non. Le mécanisme est clairement NMAR, mais on montre facilement qu'il est MAR pour $\gamma_1$ (Little et Zanganeh , 2013) et on peut donc inférer sur ce paramètre sans modéliser le mécanisme.

Par mesure de simplification, nous nous placerons dans le cadre de données manquantes générées au hasard uniquement, mais il faut garder à l'esprit que ces méthodes peuvent s'étendre à certains cas non-ignorables.

Sous l'hypothèse d'ignorabilité, l'estimation par maximum de vraisemblance consiste à maximiser $L\left(\gamma | X^{obs}\right)$ par rapport à $\gamma$, tandis que l'inférence Bayésienne porte sur la loi $f\left(X^{obs} | \gamma\right) p\left(\gamma\right)$.

### 2.2.2 Maximum de vraisemblance

L'estimation par maximum de vraisemblance peut être difficile, même sans données manquantes, car les équations de vraisemblance n'ont pas toujours de solutions explicites. En

présence de données manquantes cette difficulté est accrue. Dans les cas des dispositifs monotones il est parfois possible d'obtenir des expressions analytiques pour les estimateurs du maximum de vraisemblance, mais en règle générale, l'utilisation d'algorithmes itératifs est nécessaire.

Par exemple, reprenons le cas d'une distribution normale bivariée présenté en introduction. On se donne un jeu $X = (X_1, X_2)$ tel que

$$X \sim \mathcal{N}\left(\mu = (\mu_{X_1}, \mu_{X_2}), \Sigma = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} \end{pmatrix}\right). \tag{16}$$

Dans le cas d'un dispositif monotone (univarié) où les individus 1 à $r$ sont complets et les autres sont incomplets sur $X_2$, l'expression de la vraisemblance se décompose sous la forme

$$L(\gamma|X^{obs}) = \prod_{i=1}^{r} f(x_{i1}, x_{i2}|\gamma) \prod_{i=r+1}^{n} f(x_{i1}|\gamma) \tag{17}$$

En l'état, cette expression ne permet pas de déterminer les estimateurs du maximum de vraisemblance car les paramètres qui régissent la distribution de $(X_1, X_2)$ ne sont pas distincts de ceux qui régissent la distribution de $X_1$. Dans le cas monotone, une reparamétrisation permet de factoriser la vraisemblance en un produit de vraisemblances avec des paramètres distincts (Anderson, 1957). Ainsi, on peut se ramener à un cadre standard d'estimation par maximum de vraisemblance comme dans un cas complet, l'inférence en présence de données manquantes ne présente pas de difficulté.

Faisons à présent l'hypothèse de données manquantes au hasard sur $X_1$ et $X_2$ tel que les $r$ premiers individus sont complets, les $t$ suivants incomplets sur $X_2$ seulement, les derniers incomplets sur $X_1$ seulement. La vraisemblance observée a alors pour expression

$$L(\gamma|X^{obs}) = \prod_{i=1}^{r} f(x_{i1}, x_{i2}|\gamma) \prod_{i=r+1}^{r+t+1} f(x_{i1}|\gamma) \prod_{i=r+t+2}^{n} f(x_{i2}|\gamma). \tag{18}$$

Aucune paramétrisation menant à une factorisation en produit de vraisemblances avec des paramètres distincts n'est possible.

Dans ce cas d'un dispositif quelconque, le recours à des algorithmes itératifs devient indispensable. L'algorithme EM (Dempster *et al.*, 1977) est l'un d'entre eux. Celui-ci alterne deux étapes appelées étape E, pour "Expectation" et étape M pour "Maximisation". Après une initialisation des paramètres, les statistiques suffisantes intervenant dans l'expression de la vraisemblance des données complètes son remplacées par leur espérance, c'est l'étape E. L'étape M consiste alors à calculer les estimateurs du maximum de vraisemblance en remplaçant les termes non observés par l'espérance définie à l'étape E. Ces étapes sont alors répétées jusqu'à ce que les estimations des paramètres ne varient plus d'une itération à l'autre. En procédant ainsi, la vraisemblance des paramètres courants croît à chaque étape et assure donc une convergence vers un maximum, au moins local.

Plus formellement, l'algorithme repose sur une réécriture de la log-vraisemblance $l\left(\gamma|X^{obs}\right)$. En effet, on peut écrire

$$l(\gamma|X) = l\left(\gamma|X^{miss}, X^{obs}\right) = l\left(\gamma|X^{obs}\right) + ln(f(X^{miss}|X^{obs}; \gamma)) \tag{19}$$

d'où

$$l\left(\gamma|X^{obs}\right) = l\left(\gamma|X\right) - ln(f(X^{miss}|X^{obs};\gamma)). \tag{20}$$

En intégrant chaque membre par rapport à la distribution des données manquantes conditionnellement aux données observées $f(X^{miss}|X^{obs},\gamma^{(t)})$ pour une valeur donnée de $\gamma$ notée $\gamma^{(t)}$, on a

$$l\left(\gamma|X^{obs}\right) = Q(\gamma|\gamma^{(t)}) - H(\gamma|\gamma^{(t)}) \tag{21}$$

avec

$$Q(\gamma|\gamma^{(t)}) = \int l\left(\gamma|X\right) f(X^{miss}|X^{obs},\gamma^{(t)})dX^{miss} \tag{22}$$

et

$$H(\gamma|\gamma^{(t)}) = \int ln(f(X^{miss}|X^{obs},\gamma)f(X^{miss}|X^{obs},\gamma^{(t)})dX^{miss}. \tag{23}$$

L'initialisation de l'algorithme consiste à fixer $\gamma^{(t)}$ à l'itération $t = 0$. L'étape E consiste à calculer $Q$. L'étape M consiste ensuite à maximiser cette vraisemblance par rapport à $\gamma$. A l'issue de cette étape, on dispose d'une nouvelle valeur pour $\gamma$, notée $\gamma^{(t+1)}$ qui vérifie $l\left(\gamma^{(t+1)}|X^{obs}\right) \geq l\left(\gamma^{(t)}|X^{obs}\right)$.

La vraisemblance croît donc à chaque itération, mais seule une convergence vers un maximum local est assurée. En changeant d'initialisation des paramètres, on peut obtenir une estimation différente. Une façon de limiter cette sensibilité à l'initialisation est d'utiliser une stratégie SMALL-EM (Biernacki *et al.*, 2003) qui consiste à initialiser l'algorithme aléatoirement plusieurs fois, de lancer l'algorithme sur quelques itérations (10 par exemple), puis de retenir l'initialisation qui conduit à la vraisemblance la plus élevée pour lancer à nouveau l'algorithme sur un nombre d'itérations plus grand.

Cet algorithme peut être appliqué dans l'exemple précédent. Les statistiques suffisantes pour estimer $\gamma$, les paramètres de la loi normale, sont les quantités $\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \sum_{i=1}^{n} x_{i1}x_{i2}, \sum_{i=1}^{n} x_{i1}^2, \sum_{i=1}^{n} x_{i2}^2$. L'étape E consiste à calculer $\mathbb{E}\left[x_{i2}|x_{i1};\gamma\right]$, $\mathbb{E}\left[x_{i1}x_{i2}|x_{i1};\gamma\right]$,$\mathbb{E}\left[x_{i2}^2|x_{i1};\gamma\right]$ pour $x_{i2}$ manquant et $\mathbb{E}\left[x_{i1}|x_{i2};\gamma\right]$, $\mathbb{E}\left[x_{i1}x_{i2}|x_{i2};\gamma\right]$,$\mathbb{E}\left[x_{i1}^2|x_{i2};\gamma\right]$ pour $x_{i1}$ manquant. Disposant d'une valeur pour $\gamma$, ces calculs sont classiques dans le cadre de la loi normale. L'étape M consiste alors à calculer les estimateurs du maximum de vraisemblance à partir des statistiques suffisantes précédemment calculées, ce qui ne présente plus de différence par rapport au cadre où les données sont complètes.

L'algorithme EM est particulièrement important pour la gestion des données manquantes en général. Sa mise en œuvre n'est pas toujours aussi simple que dans le cas d'une distribution normale. Par exemple, le paramètre $\gamma$ peut être soumis à des contraintes. L'étape $M$ de maximisation devient alors plus complexe (Meng et Rubin, 1993).

L'inférence par maximum de vraisemblance est une approche qui s'adapte facilement à tout type de dispositifs de données manquantes, l'algorithme EM permettant de gérer les dispositifs non-monotones. Toutefois, cette inférence repose sur les propriétés asymptotiques de l'estimateur du maximum de vraisemblance et n'est donc pas adaptée à de petits échantillons. Au contraire, l'inférence Bayésienne ne nécessite pas de grands échantillons. De plus, elle fournit directement une estimation de la variance associée à l'estimation ponctuelle via la loi a posteriori dont on dispose sur les paramètres.

### 2.2.3 Estimation Bayésienne

Contrairement à l'approche par maximum de vraisemblance, l'estimation Bayésienne permet d'introduire de l'information supplémentaire sur les paramètres. Celle-ci peut par exemple résulter d'une connaissance acquise sur d'autres données. D'autre part, pour des lois a priori conjuguées avec la vraisemblance, *i.e.* appartenant à la même famille de distribution, la loi a posteriori est facile à calculer. L'approche Bayésienne fournit ainsi une solution pour mener à bien les calculs.

La mise en œuvre de cette estimation en présence de données manquantes se heurte aux mêmes difficultés que l'approche fréquentiste. L'estimation Bayésienne consiste à déterminer la loi a posteriori des paramètres $p(\gamma|X)$ qui sous l'hypothèse d'ignorabilité est proportionnelle à $p(\gamma)L(\gamma|X^{obs})$. Quand la vraisemblance est factorisable, on peut obtenir une expression explicite pour la loi a posteriori, mais en règle générale, cette loi n'est pas accessible. Un algorithme itératif tel que l'algorithme de Data Augmentation (Tanner et Wong, 1987) peut alors être employé. Celui-ci permet d'effectuer des tirages indépendants des paramètres dans leur loi a posteriori. En répétant les tirages, on peut ainsi obtenir une estimation ponctuelle des paramètres et y associer un intervalle de crédibilité en se basant sur les percentiles de l'échantillon.

Le principe de l'algorithme de Data-Augmentation (Tanner et Wong, 1987) est de se ramener à un jeu de données complet pour lequel la loi a posteriori peut être plus facilement simulée. Les données sont ainsi "augmentées" en attribuant des valeurs aux données manquantes. Plus précisément, l'algorithme est basé sur l'alternance de deux étapes :

$$\text{Etape I : Tirer } X^{miss}_{(t+1)} \text{ dans } f\left(X^{miss}|X^{obs},\gamma^{(t)}\right)$$
$$\text{Etape P : Tirer } \gamma^{(t+1)} \text{ dans } p\left(\gamma^{(t)}|X^{miss}_{(t+1)},X^{obs}\right)$$

L'étape I est l'étape d'imputation où les données manquantes sont tirées dans leur distribution prédictive pour le paramètre courant $\gamma^{(t)}$. L'étape P est l'étape de tirage dans la loi a posteriori des paramètres. En alternant ces deux étapes, Tanner et Wong (1987) ont montré que la suite $\left(\gamma^{(t)}\right)_{1\leq t\leq T}$ converge vers la loi a posteriori $p\left(\gamma|X^{obs}\right)$. Ainsi, cet algorithme permet de simuler la distribution a posteriori des paramètres. On peut noter que l'algorithme de Data-Augmentation est un échantillonneur de Gibbs particulier (Geman et Geman, 1984; Gelfand et Smith, 1990). L'échantillonneur de Gibbs permet de simuler une loi jointe à partir des lois conditionnelles. Ici, on simule la loi jointe de $\left(\gamma, X^{miss}|X^{obs}\right)$ à partir de la loi de $\left(X^{miss}|X^{obs},\gamma\right)$ et celle $\left(\gamma|X^{miss},X^{obs}\right)$. La particularité de l'algorithme de Data-Augmentation étant que les données manquantes

sont vues comme des paramètres.

On note que les étapes I et P de l'algorithme de Data-Augmentation sont très proches des étapes E et M de l'algorithme EM. En effet, l'étape E de l'algorithme EM consiste à calculer l'espérance des statistiques suffisantes intervenant dans l'écriture de la vraisemblance des données complètes, l'étape I de l'algorithme de Data-Augmentation simule les données ce qui permet d'obtenir ces statistiques suffisantes. Aussi l'étape M de l'algorithme EM maximise la vraisemblance des données complètes, l'étape P de l'algorithme de Data-Augmentation effectue un tirage à partir de cette vraisemblance.

L'initialisation de l'algorithme de Data-Augmentation est généralement effectuée en fixant $\gamma^{(0)}$ à son estimation par maximum de vraisemblance, à l'aide d'un algorithme EM. En effet l'estimateur de maximum de vraisemblance tend à être proche du centre de la loi a posteriori, en particulier quand l'a priori choisi est non-informatif. Ainsi, ce choix évite d'initialiser la suite par une valeur qui se trouverait en queue de distribution, ce qui accélère généralement la convergence de l'algorithme (Enders, 2010; Schafer, 1997).

Le critère d'arrêt de l'algorithme est toutefois moins évident à définir. En effet, il s'agit ici d'une convergence en loi, c'est-à-dire que la distribution du paramètre est stable au bout d'un nombre d'itérations assez grand, mais les valeurs successives de $\gamma^{(t)}$ ne tendent pas vers une valeur fixe. Bien que des méthodes sophistiquées pour vérifier la convergence existent, elles restent complexes à mettre en œuvre et la vérification de la convergence est généralement effectuée empiriquement. Pour cela on représente graphiquement la suite $\left(\gamma^{(t)}\right)_{1\leq t\leq T}$, pour chaque composante de $\gamma$, et on choisit un nombre d'itérations assez grand de façon à ce qu'elle atteigne son régime stationnaire.

La construction d'estimations ponctuelles et d'intervalles de crédibilité nécessite également que les tirages soient effectués de façon indépendantes. Or, les valeurs successives des éléments de la suite sont corrélées. Le nombre d'itérations nécessaires pour considérer que les éléments sont indépendants est également déterminé graphiquement, selon les autocorrélogrammes. A nouveau, cette approche est empirique.

Enfin, une question importante dans ce contexte est le nombre de réalisations indépendantes nécessaires pour limiter l'erreur de simulation et ainsi obtenir une approximation suffisamment précise de l'espérance d'une composante du paramètre (ou d'une fonction de ce paramètre), ainsi que des bornes de l'intervalle de crédibilité (Schafer, 1997, p.132). Ce nombre est fréquemment de l'ordre de plusieurs milliers.

Dans le cas de la loi normale, et sans connaissance sur les données, un choix classique pour la loi a priori sur les paramètres de la loi normale $(\mu, \Sigma)$ est l'a priori de Jeffrey (Jeffrey, 1946) qui est un a priori non informatif. Avec ce choix d'a priori, la loi a posteriori pour les paramètres est une loi de Wishart inverse pour $\Sigma$ et une loi normale pour $\mu|\Sigma$ dont les paramètres dépendent des données (Gelman *et al.*, 2003). Pour une telle modélisation, l'algorithme de Data-Augmentation s'applique de la façon suivante : pour une valeur courante de $(\mu, \Sigma)$ , d'abord obtenue par maximum de vraisemblance, l'étape I consiste à imputer les données manquantes par régression stochastique. L'étape P consiste ensuite à calculer les paramètres de la loi a posteriori selon les nouvelles valeurs imputées, puis

à effectuer un tirage de $\Sigma$ dans la loi de Wishart inverse correspondante, puis à tirer $\mu$ conditionnellement au tirage $\Sigma$ précédemment effectué. Ces étapes sont alors répétées jusqu'à convergence.

En règle générale, la loi a posteriori n'a pas toujours une forme simple, et il est souvent nécessaire d'utiliser des algorithmes itératifs, tel que l'algorithme de Gibbs, pour pouvoir effectuer un tirage dans cette loi. Le choix de la loi a priori est également délicat en l'absence d'une expertise réelle sur les données. Dans pareil cas, une solution classique est de choisir les paramètres de la loi a priori à partir des données, en les estimant par maximum de vraisemblance par exemple. Cette approche est appelée *Bayésien empirique*. Une autre solution est de faire appel à des lois non-informatives comme l'a priori de Jeffrey.

Comme l'estimation par maximum de vraisemblance, l'estimation Bayésienne s'adapte assez bien à tout type de dispositifs de données manquantes, l'algorithme de Data-Augmentation permettant de gérer les dispositifs non-monotones. Aussi, elle est particulièrement intéressante sur des échantillons de petites tailles, car elle ne repose pas sur des hypothèses asymptotiques, au contraire de l'approche par maximum de vraisemblance. On pourrait objecter que si l'échantillon est petit, alors l'influence de la loi a priori sur la loi a posteriori est importante, ce qui n'est pas souhaitable en l'absence d'une connaissance réelle sur les données. Toutefois, un a priori peu informatif conduit généralement à une inférence plus précise qu'une inférence par maximum de vraisemblance sur un petit échantillon (Fitzmaurice *et al.*, 2014, p.15).

Les approches par vraisemblance présentées ici reposent sur l'hypothèse d'ignorabilité, et en particulier sur l'hypothèse d'un mécanisme MAR. Or, il n'est pas possible de tester cette hypothèse contre l'hypothèse NMAR. Le seul moyen de rendre cette hypothèse crédible est d'introduire des variables auxiliaires susceptibles d'expliquer la présence de données manquantes. Une possibilité pour inclure des variables auxiliaires dans ces approches serait tout simplement de modifier la modélisation des données de façon à intégrer ces variables, mais ceci n'est généralement pas souhaitable. Prenons l'exemple où l'on estime des paramètres de régression par maximum de vraisemblance. Si on considère les variables auxiliaires comme des variables explicatives, alors on modifie complètement l'interprétation que l'on fait des paramètres de la régression. Or, on ne souhaite pas modifier l'interprétation du modèle, seulement intégrer ces variables auxiliaires pour améliorer l'inférence. Des approches ont été proposées pour "transmettre" l'information des variables auxiliaires sans modifier l'interprétation des paramètres (Graham, 2003; Savalei et Bentler, 2009; Yuan et Bentler, 2000). Toutes ces méthodes restent encore limitées. L'approche "saturated correlates model" (Graham, 2003) qui repose sur les modèles d'équations structurelles est la plus simple à utiliser. Dans l'exemple du modèle de régression précédent, cette méthode gère les variables auxiliaires en construisant un modèle d'équations structurelles saturé où les variables auxiliaires sont corrélées à toutes les variables, exceptée la réponse du modèle. De cette façon, la variable réponse n'est expliquée que par les covariables, mais l'information portée par les variables auxiliaires est bien prise en compte. Cette méthode est toutefois mise en défaut quand le nombre de variables auxiliaires devient important ou que celles-ci

comportent des données manquantes (Enders, 2010).

La dernière approche abordée pour gérer le problème des données manquantes est l'imputation multiple. Elle possède en particulier l'avantage de pouvoir intégrer facilement des variables auxiliaires.

## 2.3 L'imputation multiple

### 2.3.1 Fondements théoriques

L'imputation multiple repose sur une approche Bayésienne du modèle d'inférence : on cherche ici à estimer l'espérance de la loi a posteriori d'une quantité d'intérêt $Q$, ainsi que sa variance de façon à construire un intervalle de crédibilité pour $Q$. La différence avec l'approche Bayésienne présentée précédemment est que $Q$ n'est pas nécessairement le paramètre qui régit la distribution des données $\gamma$. $Q$ est par exemple une moyenne, une proportion, un coefficient de corrélation. Par mesure de simplification on supposera temporairement que $Q$ est scalaire, mais dans le cas général $Q$ peut être multidimensionnel.

On peut écrire la loi a posteriori de $Q$ sous la forme

$$p\left(Q|X^{obs},R\right)=\int p\left(Q|X^{obs},X^{miss}\right)f\left(X^{miss}|X^{obs},R\right)dX^{miss} \tag{24}$$

et sous l'hypothèse d'ignorabilité

$$p\left(Q|X^{obs},R\right)=\int p\left(Q|X^{obs},X^{miss}\right)f\left(X^{miss}|X^{obs}\right)dX^{miss}. \tag{25}$$

Cette décomposition est importante car elle exprime $p(Q|X^{obs},R)$, qui n'est pas simple à calculer, comme la combinaison de deux lois a posteriori dans lesquelles il est plus facile d'effectuer des tirages. Le premier terme de cette intégrale est la loi a posteriori de $Q$ conditionnellement au jeu de données complet, ce calcul est généralement simple à effectuer ; le second est la distribution prédictive des données manquantes.

Si on sait générer les données manquantes dans leur distribution prédictive, autrement dit imputer les données manquantes selon $f\left(X^{miss}|X^{obs}\right)$, alors cette intégrale peut être approchée par la moyenne des lois a posteriori évaluées en les données générées $\left(X_m^{miss}\right)_{1\leq m\leq M}$ :

$$p(Q|X^{obs},R)\approx\frac{1}{M}\sum_{m=1}^{M}p\left(Q|X_m^{miss},X^{obs}\right). \tag{26}$$

L'espérance de la loi a posteriori est alors approchée par

$$\begin{aligned}\mathbb{E}\left[Q|X^{obs}\right]&\approx\int Q\frac{1}{M}\sum_{m=1}^{M}p\left(Q|X_m^{miss},X^{obs}\right)dQ\\&=\bar{Q}\end{aligned} \tag{27}$$

avec $\bar{Q}=\frac{1}{M}\sum_{m=1}^{M}\widehat{Q}_m$ et $\widehat{Q}_m\approx\mathbb{E}\left[Q|X_m^{miss},X^{obs}\right]$, l'estimation de $Q$ pour le jeu imputé $m$. $\bar{Q}$ est l'estimateur agrégé à partir de $\left(\widehat{Q}_m\right)_{1\leq m\leq M}$.

La variance de la loi a posteriori est quant à elle estimée par

$$\begin{aligned}\mathbb{V}\left[Q|X^{obs}\right] &\approx \left(\frac{1}{M}\sum_{m=1}^{M}\bar{U}_m\right) + \left(\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{Q}_m - \bar{Q}\right)^2\right) \\ &= \bar{U} + B \end{aligned} \tag{28}$$

où $\bar{U}_m \approx \mathbb{V}\left[Q|X_m^{miss}, X^{obs}\right]$ est l'estimation de la variance de $Q$ pour le jeu imputé $m$. $\bar{U}$ est l'estimation de la variance intra-imputation, qui correspond à l'estimation de la variabilité due à l'échantillonnage, et $B$ est l'estimation de la variabilité inter-imputation qui correspond à la variabilité attribuable à la présence de données manquantes.

L'approximation de la variance peut être raffinée en multipliant $B$ par $\left(1+\frac{1}{M}\right)$ afin de prendre en compte que pour un nombre fini de tableaux $M$, $\bar{Q}$ n'est qu'une approximation de $\mathbb{E}\left[\widehat{Q}|X\right]$ et qu'en conséquence une variabilité supplémentaire de $B/M$ doit être ajoutée (Rubin, 1987). Ce terme correspond donc à l'erreur de simulation qui est de cette façon intégrée dans le calcul de la variabilité totale. Cette estimation est notée $T = \bar{U} + \left(1+\frac{1}{M}\right)B$.

Ainsi, l'imputation multiple se résume à trois étapes :

1. l'imputation des $M$ tableaux,
2. l'analyse de chaque tableau imputés, correspondant au calcul de $\widehat{Q}_m$ et $\bar{U}_m$,
3. et l'étape d'agrégation correspondant au calcul de $\bar{Q}$ et de la variabilité totale $T$.

Sous l'hypothèse que $Q$ est distribué normalement pour $X$ fixé, on peut en déduire un intervalle de crédibilité $\bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}$ avec $t_{\nu,1-\alpha/2}$ le quantile de niveau $1-\alpha/2$ pour la loi de Student à $\nu$ degrés de liberté. $\nu$ est donné par (Rubin, 1987; Rubin et Schenker, 1986)

$$\nu = (M-1)\left(1/\left(\frac{\left(1+\frac{1}{M}\right)B}{T}\right)^2\right). \tag{29}$$

Dans le cas habituel où les données sont complètes, le nombre de degrés de liberté croît avec la taille de l'échantillon et la loi de Student tend vers une loi normale. Ici, on remarque que le nombre de degrés de liberté croît avec le nombre de tableaux et décroît avec le rapport

$$\lambda = \frac{\left(1+\frac{1}{M}\right)B}{T}. \tag{30}$$

Ce rapport s'interprète comme la part de variabilité attribuable à la présence de données manquantes. Pour une taille d'échantillon faible et peu de données manquantes, le nombre de degrés de liberté donné par (29) peut donc être très au-delà de celui du cas complet, situation jugée problématique (Barnard et Rubin, 1999). Ainsi, une correction a été proposée par Barnard et Rubin (1999) de façon à ce que le nombre de degrés de liberté n'excède pas celui obtenu sur un jeu complet :

$$\nu^{Bar} = \left(\frac{1}{\nu} + \frac{1}{\tilde{\nu}}\right)^{-1} \tag{31}$$

avec

$$\tilde{\nu} = (1 - \lambda) \frac{\nu_{com} + 1}{\nu_{com} + 3} \nu_{com} \tag{32}$$

et $\nu_{com}$ le nombre de degré de liberté pour un jeu complet.

Une propriété remarquable de l'imputation multiple par rapport à l'approche Bayésienne est que le nombre de tableaux imputés $M$ peut être relativement faible devant les milliers de tirages que peut nécessiter l'approche Bayésienne. En effet, l'intervalle de crédibilité tient ici compte de l'erreur de simulation et est donc valide dès $M \geq 2$. Au contraire, l'algorithme de Data-Augmentation produit suffisamment de simulation pour pouvoir négliger cette erreur dans la détermination des bornes de l'intervalle de crédibilité. Remarquons que l'imputation multiple suggère ainsi une solution pour limiter le nombre de simulations en Data-Augmentation dans le cas où la quantité d'intérêt $Q$ correspond aux paramètres du modèle $\gamma$.

Toutefois, plus le nombre de tableaux est grand, moins l'erreur de simulation est importante, et donc plus l'intervalle de crédibilité est petit. Rubin (1987) a montré que la variance à $M$ fixé est reliée à celle obtenue pour un nombre infini de tableaux par

$$T_M = \left(1 + \frac{FMI}{M}\right) T_\infty \tag{33}$$

où $FMI$ désigne la fraction d'information manquante

$$FMI = \frac{\nu + 1}{\nu + 3} \lambda + \frac{2}{\nu + 3} \tag{34}$$

et s'est interprète comme $\lambda$. Ainsi, supposons que la fraction d'information manquante vaille 0.3, ce qui pour fixer les idées correspond au cas de l'estimation d'une moyenne pour une variable indépendante avec 30% de données manquante. Alors pour $M = 5$, on obtient $T_M = 1.06 \times T_\infty$ ce qui signifie que l'intervalle de confiance est de 6% plus grand que celui obtenu pour un nombre infini de tableaux, correspondant à la variabilité minimale qu'on puisse espérer atteindre.

Le nombre de tableaux à imputer reste fonction de la fraction d'information manquante, est liée notamment au pourcentage de données manquantes. Ainsi, il est recommandé de choisir un nombre de tableaux plus élevé si le jeu de données comporte de nombreuses données manquantes. Avec les moyens calculatoires actuels, on recommande généralement de prendre une vingtaine de tableaux (Van Buuren, 2012).

### 2.3.2 Lien entre modèle d'imputation et modèle d'analyse

La caractéristique majeure de l'imputation multiple est qu'elle sépare le problème d'inférence en deux : celui de l'imputation des données et celui de l'analyse des données imputées. Ces étapes peuvent être effectuées par des personnes différentes. Par exemple un statisticien peut compléter le jeu de données de $M$ façons, et différents experts peuvent ensuite analyser les données imputées indépendamment. Par conséquent le modèle d'imputation peut être différent du modèle d'analyse. Ainsi, le modèle qui sert à imputer peut

comprendre des variables qui n'appartiennent pas au modèle d'analyse. Se pose alors la question importante de l'adéquation entre l'analyse et l'imputation effectuée, c'est ce qu'on appelle la *congénialité* (Meng, 1994; Schafer, 2003).

Différentes situations peuvent être distinguées. Premièrement le cas où le modèle d'imputation et le modèle d'analyse reposent sur les même distributions et font appel aux mêmes variables. Dans ce cas, le modèle d'imputation apporte la même information que celle utilisée pour le modèle d'analyse. L'inférence obtenue est asymptotiquement la même que celle obtenue en procédant directement par maximum de vraisemblance pour le modèle d'analyse. Par exemple, on peut imputer un jeu quantitatif selon la loi normale et estimer par la suite des paramètres d'une régression multiple expliquant une variable en fonction de toutes les autres. Les deux modèles font alors appel à une même distribution pour la loi conditionnelle de la réponse sachant les covariables.

Le deuxième cas est celui où le modèle d'imputation utilise les mêmes variables que le modèle d'analyse, mais que le modèle d'imputation suppose une distribution plus générale que le modèle d'analyse. Par exemple, on peut imputer par la loi normale, mais estimer les paramètres d'un modèle d'analyse en facteurs communs et spécifiques (Bartholomew *et al.*, 2011) qui est un modèle à effet aléatoires où la loi jointe des données est une loi normale multivariée avec une matrice de variance-covariance structurée. Dans ce cas, le modèle d'analyse repose sur les mêmes variables que le modèle d'imputation, mais le modèle d'analyse est moins général. Dans ce cas, l'imputation multiple conduit à une inférence non biaisée, mais la variabilité des estimateurs sera plus large que celle qu'on aurait obtenue en procédant directement par maximum de vraisemblance.

Le troisième cas est celui où le modèle d'imputation est plus restrictif que le modèle d'analyse. Par exemple, supposons un modèle de régression avec interaction comme modèle d'analyse et un modèle Gaussien, donc sans interaction, pour le modèle d'imputation. Le modèle d'imputation fait alors l'hypothèse plus restrictive de l'absence d'interaction. Si des relations d'interaction sont bien présentes entre les variables du jeu de données, alors l'inférence menée à la suite de l'imputation sera biaisée. Dans le cas contraire, l'inférence sera non biaisée et la variabilité des estimateurs sera plus faible que celle qu'on aurait obtenue en adoptant une approche par maximum de vraisemblance sur le modèle d'analyse. Cette propriété est appelée *superefficiency* (Rubin, 1996).

Le dernier cas est celui où les variables utilisées diffèrent entre les deux modèles. L'imputation multiple permet en effet d'intégrer davantage de variables que celles du modèle d'analyse. Cette situation peut par exemple arriver si le statisticien dispose d'une base de données alors que l'expert qui va analyser les données n'est intéressé qu'à quelques variables sur une thématique particulière. L'ajout d'information supplémentaire peut grandement améliorer l'inférence si ces variables sont liées aux variables du modèle d'analyse ou si elles sont liées au mécanisme générant les données manquantes.

L'exemple choisi par (Enders, 2010, p.337-338) illustre très bien la première situation. Cet exemple porte sur un score de dépression calculé à partir des réponses que des individus ont fournies dans le cadre d'un questionnaire à choix multiples. Ce score est ensuite utilisé en tant que variable réponse d'un modèle de régression multiple pour d'autres variables. Les données manquantes n'apparaissent que dans les réponses aux questionnaires. Les individus incomplets, même s'il ne leur manque qu'une seule donnée,

n'ont pas de score, ce qui oblige à devoir gérer les données manquantes dans l'inférence des paramètres du modèle de régression. L'imputation multiple offre une réponse simple à ce problème en imputant d'abord les variables du questionnaire, puis en calculant le score correspondant pour chaque tableau, et enfin en procédant à l'analyse, c'est-à-dire en calculant les coefficients de la régression et leur variabilité et en agrégeant les estimations. Les approches par vraisemblance rendent beaucoup plus complexe l'utilisation des données du questionnaire pour inférer sur les paramètres du modèle de régression.

Aussi, dans la situation ou des variables liées au mécanisme générant les données manquantes sont disponibles, l'imputation multiple offre un moyen simple d'intégrer ces variables auxiliaires pour rendre l'hypothèse MAR plus crédible.

Le modèle d'imputation doit donc être choisi en fonction de l'analyse qui sera fait par la suite. La connaissance des propriétés du modèle d'imputation utilisé est ainsi essentielle. Pour que le modèle d'imputation soit au moins aussi riche que le modèle d'analyse, et donc que l'inférence soit la moins biaisée possible et la moins variable, il est bon d'inclure des variables auxiliaires. Il est donc important qu'une méthode d'imputation puisse être mise en œuvre en présence d'un nombre élevé de variables.

### 2.3.3 Imputation proper

L'étape d'analyse ne présente pas de difficulté particulière par rapport au cadre standard sans données manquantes. La difficulté principale en imputation multiple est d'imputer les données. Pour cela il faut imputer selon la distribution prédictive $p\left(X^{miss}|X^{obs}\right)$ des données manquantes. Une méthode d'imputation qui génère $M$ tableaux indépendants issus de cette distribution est alors dite *Bayesian proper* (Schafer, 1997, p.105). Or, la distribution prédictive $p\left(X^{miss}|X^{obs}\right)$ dépend de paramètres inconnus. En réécrivant cette distribution sous la forme

$$p\left(X^{miss}|X^{obs}\right) = \int p\left(X^{miss}|X^{obs},\gamma\right) p\left(\gamma|X^{obs}\right) d\gamma, \tag{35}$$

il apparaît que celle-ci s'écrit comme la distribution prédictive à $\gamma$ fixé, moyennée par rapport à la loi a posteriori de $\gamma$. Une méthode d'imputation multiple nécessite donc de refléter l'incertitude sur $X^{miss}$ pour un paramètre $\gamma$ donné, ainsi que de refléter l'incertitude sur les paramètres de ce modèle.

Reprenons l'exemple du modèle Gaussien. Les données sont distribuées selon une loi normale multivariée $X \sim \mathcal{N}(\mu,\Sigma)$ régie par $\gamma = (\mu,\Sigma)$. Une façon classique de tirer $\gamma$ dans sa distribution a posteriori, et sûrement la plus naturelle ici, est d'utiliser un algorithme de Data-Augmentation. En choisissant un a priori de Jeffrey pour le paramètre $\gamma$, l'algorithme de Data-Augmentation permet d'effectuer $M$ tirages indépendants de $\gamma$ dans $p\left(\gamma|X^{obs}\right)$. On impute ensuite par régression l'ensemble des données manquantes selon $p\left(X^{miss}|X^{obs}\right)$. On peut noter que les étapes P succédant aux étapes I pour lesquelles les paramètres ont été retenus fournissent directement les $M$ imputations dans la distribution prédictive à $\gamma$ fixé. On peut ainsi retenir ces tableaux plutôt que de réimputer les données après avoir généré $M$ tirages indépendants de $\gamma$. Cette méthode est Bayesian proper. Au

contraire, si on avait fixé le paramètre $\gamma$ à son estimation par maximum de vraisemblance, alors la procédure d'imputation multiple n'aurait pas vérifié cette propriété. Par exemple, l'imputation par régression aléatoire, illustrée en Figure 1 dans le Chapitre 1, n'est pas Bayesian proper.

En pratique, le modèle pour les données n'est pas connu, on n'a donc pas de certitude sur le caractère Bayesian proper d'une méthode d'imputation. D'autre part, le modèle d'imputation n'est pas toujours régit par un paramètre $\gamma$, comme par exemple l'imputation par forêts aléatoires (Breiman, 2001; Stekhoven et Bühlmann, 2012; Shah *et al.*, 2014; Doove *et al.*, 2014) qui utilise des prédictions par arbres pour imputer les données. Les méthodes d'imputation multiple paramétriques et non paramétriques dites *proper* (Rubin, 1987) pour $\left(\widehat{Q}, U\right)$ mènent à une inférence *valide* pour la quantité d'intérêt $Q$, *i.e.* non biaisée et ne sous-estimant pas la variabilité (Van Buuren, 2012, p.36). Celles-ci satisfont à trois propriétés qui sont fonction du paramètre considéré. La première est que l'espérance de l'estimateur agrégé $\bar{Q}$, sur tous les dispositifs de données manquantes possibles, est égale à l'estimation obtenue si le jeu avait été complet. La seconde est que, l'espérance de la variance d'échantillonnage sur l'ensemble des dispositifs est égale à celle qu'on aurait observée si le jeu avait été complet. Enfin, la dernière condition assure que la variance due à la présence de donnée manquante $B$ est plus faible que la variance d'échantillonnage. Ainsi, si la méthode d'imputation est *proper* et que la méthode d'analyse est *valide*, alors la procédure d'inférence est *valide* (Rubin, 1987).

Pour approcher la distribution prédictive des données manquantes, quand le modèle dont sont issues les données n'est pas connu, ou que le modèle d'imputation utilisé n'est pas paramétré, on peut avoir recours à une approche bootstrap plutôt qu'à une approche Bayésienne. Le principe consiste à rééchantillonner le jeu de données, utiliser cet échantillon comme échantillon d'apprentissage, puis prédire les données manquantes sur le jeu incomplet d'origine. Cette procédure peut être utilisée sur l'exemple Gaussien : on peut par exemple appliquer un bootstrap non-paramétrique en tirant avec remise $n$ individus du jeu de données, puis, l'échantillon étant incomplet, utiliser l'algorithme EM présenté en Section 2.2.2 pour obtenir une estimation de $\gamma$. On répète alors l'opération $M$ fois pour obtenir un jeu de $M$ paramètres. On utilise ensuite chacun de ces paramètres pour produire $M$ imputations du jeu de données.

Hormis dans des cas très simples, les conditions qui font qu'une méthode d'imputation est proper sont impossibles à vérifier (Van Buuren, 2012; Schafer, 1997). Le caractère proper d'une méthode d'imputation multiple est donc évalué par des simulations : pour différentes quantités d'intérêt, on vérifie que l'inférence est sans biais et que la variabilité des estimateurs n'est pas sous-estimée.

# 3 Discussion

L'inférence en présence de données manquantes peut être abordée selon les trois approches présentée. L'approche par pondération, très intuitive, a la particularité de n'être vraiment pertinente que pour des dispositifs monotones. Au contraire, les approches par vraisemblance, via des algorithmes itératifs sont applicables sur des dispositifs quelconques. Ces algorithmes ne sont pas pour autant toujours simples à mettre en

œuvre, ce qui rend leur utilisation un peu lourde car, à chaque nouvelle inférence, un nouvel algorithme doit être mis en place. Au contraire, l'imputation multiple est plus commode car elle sépare le problème de gestion des données manquantes de celui de l'analyse sur le jeu complet. Ceci permet d'effectuer des analyses différentes pour une même imputation, mais nécessite aussi de veiller au problème de congénialité.

Parce qu'elle utilise d'une part un modèle d'imputation et d'autre part un modèle d'analyse, l'imputation multiple offre aussi un moyen simple d'inclure davantage de variables que n'en contient le modèle d'analyse. Cette capacité à utiliser des variables supplémentaires permet d'intégrer facilement des variables expliquant le mécanisme à l'origine des données manquantes, rendant l'hypothèse MAR plus crédible. Il permet aussi d'intégrer des variables liées aux données observées.

En plus d'intégrer beaucoup de variables, le modèle d'imputation doit potentiellement prendre en compte un grand nombre de relations entre les variables de façon à être au moins aussi large que le modèle d'analyse. Or, un modèle large avec beaucoup de variables est un modèle généralement complexe, dont les paramètres ne sont pas toujours estimables, du fait du nombre éventuellement limité d'observations, amplifié par le nombre éventuellement grand de données manquantes. Beaucoup de modèles d'imputation actuels sont ainsi mis en défaut à cause d'un nombre d'individus trop faible par rapport au nombre de variables ou d'un nombre de données manquantes important.

Dans les chapitres suivants nous verrons notamment en quoi les méthodes d'imputation multiple basées sur les méthodes d'analyse factorielles proposent une solution à ce problème. Pour cela, nous commencerons par identifier les propriétés du modèle d'imputation sur lesquelles repose ces méthodes. Il ne s'agira pas ici de mener une inférence, mais d'étudier la qualité de prédiction des données manquantes, autrement dit, l'imputation des données manquantes la plus plausible selon ce modèle. Ceci permettra d'identifier le type d'analyse qui peut être mise en œuvre par la suite. Parce qu'on ne s'intéressera qu'à une seule imputation du tableau, cette méthode correspond à une méthode d'imputation simple. Elle sera l'objet du Chapitre 3. Ce n'est qu'une fois ces propriétés étudiées que nous reviendrons au problème de l'inférence dans les Chapitres 4 et 5. Le Chapitre 4 présentera l'imputation multiple par ACP permettant d'effectuer des inférences sur un jeu de données dont les variables sont quantitatives. Nous vérifierons par simulation le caractère proper de cette procédure. Le Chapitre 5 quant à lui portera sur une procédure d'imputation multiple par ACM afin d'inférer sur des données qualitatives. Celle-ci sera également évaluée par simulation.

# CHAPITRE 3

# IMPUTATION SIMPLE PAR LES MÉTHODES D'ANALYSE FACTORIELLE

DANS CE CHAPITRE, nous proposons trois modèles d'imputation joints selon la nature des variables du jeu de données : ACP pour des données quantitatives, ACM pour des variables qualitatives, AFDM pour des données mixtes. Ces modèles sont construits à partir des liaisons entre variables et des ressemblances entre individus et ne nécessite qu'un nombre réduit de paramètres. L'imputation sous ces modèles est étudiée au travers de l'imputation par AFDM. Cette étude illustre les propriétés des méthodes d'imputation par les méthodes d'analyse factorielles : ces méthodes sont particulièrement adaptées à la présence de relations linéaires entre variables quantitatives, elles peuvent être mise en œuvre même si le nombre d'individus est inférieur au nombre de variables, leurs performances sont très intéressantes pour imputer des modalités rares, elle sont adaptées à des mécanismes MCAR et MAR.

**Contents**



La construction d'une méthode d'imputation multiple commence par le choix d'un modèle d'imputation. La compréhension des propriétés de ce modèle permet de cibler le type d'inférence qui pourra être mené par la suite sur un jeu imputé à partir de ce modèle. Ainsi, l'objet de ce chapitre est de présenter les méthodes d'imputation simple, pour des données quantitatives, qualitatives ou mixtes, utilisant les méthodes d'analyse factorielle. Nous commencerons par présenter les méthodes d'analyse factorielle, puis expliquerons comment en estimer les paramètres en présence de données manquantes. Enfin, nous présenterons les méthodes d'imputation par composantes principales et illustrerons leurs propriétés.

# 1 Méthodes d'analyse factorielle

Les méthodes d'analyse factorielle sont des méthodes d'analyse exploratoire multidimensionnelles utilisées pour identifier les relations entre variables ainsi que les ressemblances entre individus sur des jeux de données où le nombre de variables, et éventuellement le nombre d'individus, sont trop élevés pour envisager d'effectuer cette tâche via une succession d'analyses univariées ou bivariées. Ces méthodes reposent sur une réduction de la dimension. Le principe est de rechercher un sous-espace qui maximise l'inertie de projection du nuage des individus ou de façon équivalente, du nuage des variables. La visualisation de la projection des points sur ce sous-espace permet ensuite d'identifier les relations entre variables et les ressemblances entre individus. Nous nous intéressons

ici à trois méthodes d'analyse factorielle : l'analyse des composantes principales (ACP), l'analyse des correspondances multiples (ACM) et l'analyse factorielle des données mixtes (AFDM). Chacune de ces méthodes est adaptée à un type de données particulier (quantitatives, qualitatives, mixtes, respectivement), mais leur principe général est très proche. Nous présentons ces méthodes par le biais de l'une d'entre elle : l'ACP.

Les $n$ individus et les $p$ variables quantitatives sont décrits par une matrice $\mathbf{X}_{n\times p}$ que l'on suppose centrée. Cette matrice peut être représentée comme un nuage de $n$ points dans $\mathbb{R}^p$ appelé *espace des variables*, chacun des points de cet espace étant affecté d'un poids $d_i\,(1 \leq i \leq n)$. De façon équivalente on peut représenter cette matrice par un nuage de $p$ points dans l'*espace des individus* $\mathbb{R}^n$, chacun des points étant muni d'un poids $m_k\,(1 \leq k \leq p)$. Afin de déterminer le sous-espace qui maximise l'inertie de projection de ces points, il convient de munir chacun de ces espaces d'une métrique : **D** pour l'espace des variables, qui définit également le poids des individus, et **M** pour l'espace des individus, qui définit aussi le poids des variables.

Dans le cadre de l'ACP ces métriques sont généralement définies par $\mathbf{D} = \frac{1}{n}\mathbb{I}_n$ et $\mathbf{M} = diag\left(s^2_{X_1}, \ldots, s^2_{X_p}\right)$ où $s^2_{X_k}$ désigne la variance empirique de la variable $X_k$. Ce choix de **M** correspond à une ACP *normée*. Il donne à chaque variable la même importance dans la définition de la distance entre deux individus quelque soit sa variance :

$$d^2_{i,i'} = \sum_{k=1}^{p}\left(\frac{x_{ik} - x_{i'k}}{s_{X_k}}\right)^2. \tag{1}$$

Ce choix de métrique se justifie en particulier quand les variables n'ont pas les mêmes unités. En revanche, quand les variables ont les mêmes unités on peut adopter la métrique $\mathbf{M} = \mathbb{1}_p$ qui correspond à une ACP dite *non normée*. Le choix de $\mathbf{D} = \frac{1}{n}\mathbb{I}_n$ quant à lui assure un poids uniforme sur l'ensemble des individus.

Le sous-espace qui maximise l'inertie de projection des points, est aussi celui qui minimise la distance entre les points et leur projections. Ainsi, l'ACP consiste à rechercher une matrice $\widehat{\mathbf{X}}$ de rang inférieur $S$ qui approche au mieux la matrice **X** au sens de la norme $\|\cdot\|_{\mathbf{M}\otimes\mathbf{D}}$

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}^S\in\mathcal{M}_{n,p}(\mathbb{R})}{\arg\min} \;\| \mathbf{X}^S - \mathbf{X} \|^2_{\mathbf{M}\otimes\mathbf{D}} = trace\left(\left(\mathbf{X}^S - \mathbf{X}\right)^\top \mathbf{D}\left(\mathbf{X}^S - \mathbf{X}\right)\mathbf{M}\right) \tag{2}$$

où $\mathbf{X}^S$ désigne une matrice réelle à $n$ lignes et $p$ colonnes de rang $S$. Notons que cette matrice $\widehat{\mathbf{X}}$ peut être vue comme une estimation d'une matrice $\tilde{\mathbf{X}}$ de rang $S$, portant l'information que l'on cherche à analyser, mais dont on n'observe qu'une version bruitée **X** de plein rang. Ainsi, $\widehat{\mathbf{X}}$ est une approximation en un sens géométrique de **X** mais aussi une estimation de $\tilde{\mathbf{X}}$. Le calcul de $\widehat{\mathbf{X}}$ s'effectue via une décomposition en valeurs singulière (SVD) du triplet $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ (Eckart et Young, 1936). En décomposant la matrice **X** sous la forme $\mathbf{X} = \mathbf{U}_{n\times p}\mathbf{\Lambda}^{1/2}_{p\times p}\mathbf{V}^\top_{p\times p}$ tel que $\mathbf{U}_{n\times p}$ soit **D**-orthonormée $\left(\mathbf{U}^\top\mathbf{D}\mathbf{U} = \mathbb{I}_p\right)$, $\mathbf{V}_{p\times p}$ soit **M**-orthonormée $\left(\mathbf{V}^\top\mathbf{M}\mathbf{V} = \mathbb{I}_p\right)$, $\mathbf{\Lambda}^{1/2} = \mathbf{diag}\left(\lambda_\mathbf{1}^{\mathbf{1/2}}, \ldots, \lambda_\mathbf{p}^{\mathbf{1/2}}\right)$ soit une matrice diagonale, $\widehat{\mathbf{X}}$ est alors donnée par la formule dite de *reconstitution* :

$$\widehat{\mathbf{X}} = \widehat{\mathbf{U}}_{n\times S}\widehat{\mathbf{\Lambda}}^{1/2}_{S\times S}\widehat{\mathbf{V}}^\top_{S\times S} \tag{3}$$

où $\widehat{\mathbf{U}}$, $\widehat{\mathbf{\Lambda}}^{1/2}$ et $\widehat{\mathbf{V}}$ correspondent aux matrices $\mathbf{U}$, $\mathbf{\Lambda^{1/2}}$ et $\mathbf{V}$ restreintes à leurs $S$ premiers éléments. La matrice $\widehat{\mathbf{X}}$ est alors de rang $S$. Les colonnes de $\widehat{\mathbf{V}}$, les vecteurs propres, définissent une base du sous-espace de $\mathbb{R}^p$ auquel les individus de $\widehat{\mathbf{X}}$ appartiennent. Aussi, le produit $\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}^{1/2}$ fournit les coordonnées des individus dans cette base. Ces coordonnées sont appelées *composantes principales*. Notons que $\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}^{1/2}$ et $\widehat{\mathbf{V}}$ sont une estimation des composantes principales et des vecteurs propres de la matrice $\tilde{\mathbf{X}}$ respectivement. Ces paramètres sont ceux d'une méthode d'analyse factorielle. Ainsi, on peut représenter $\widehat{\mathbf{X}}$ en visualisant les coordonnées des individus sur chacun des $S$ axes.

L'objectif dans notre cadre n'est pas de représenter $\widehat{\mathbf{X}}$ graphiquement afin de comprendre l'information portée dans le jeu de données, mais d'utiliser cette information pour prédire des valeurs du tableau. La matrice $\widehat{\mathbf{X}}$ obtenue en reconstituant les données à partir des composantes principales et des vecteurs propres estimés fournit une telle prédiction. Il est important de noter que cette prédiction est obtenue en estimant un nombre réduit de paramètres indépendants : $p$ pour le centrage de $\mathbf{X}$, $(n-1)\,S$ pour estimer les composantes principales car celles-ci sont centrées et $pS$ pour l'estimation des vecteurs propres, auxquels on retranche $S^2$ paramètres du fait des contraintes d'orthogonalité, soit un total de $p+S(n-1+p-S)$ paramètres. Pour $S$ fixé, ce nombre évolue linéairement en fonction du nombre de lignes et linéairement en fonction du nombre de colonnes. Ainsi, le modèle n'est pas surparamétré même si le nombre d'individus ou le nombre de variables est grand. De plus, la présence de corrélations fortes entre les variables ne constitue pas non plus un problème car aucune inversion n'est utilisée pour effectuer l'estimation des paramètres.

L'ACM et l'AFDM peuvent être vues comme des extensions de l'ACP utilisant d'autres métriques. 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. La Figure 1 illustre un tel recodage. Le tableau $\mathbf{X}$ dans le cadre de l'ACM correspond au tableau recodé centré. Il possède $n$ lignes et $J$ colonnes où $J$ correspond au nombre total de modalités de réponses. 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 :

$$d^2_{i,i'} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{x_{ij} - x_{i'j}}{\sqrt{n_J}}\right)^2 \tag{4}$$

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 ($\mathbf{D} = \frac{1}{n}\mathbb{I}_n$). On obtient alors les composantes principales en effectuant la SVD du triplet $(\mathbf{X}, \mathbf{M}, \mathbf{D})$.

| $X_1$ | ... | $X_p$ |
|---|---|---|
| $A$ | ... | $A$ |
| $A$ | ... | $A$ |
| $A$ | ... | $A$ |
| $B$ | ... | $B$ |
| | | |
| | | |
| | | |
| $B$ | ... | $C$ |
| $B$ | ... | $B$ |

$\rightarrow$

| $X_1^a$ | $X_1^b$ | ... | $X_p^a$ | $X_p^b$ | $X_p^c$ |
|---|---|---|---|---|---|
| 1 | 0 | ... | 1 | 0 | 0 |
| 1 | 0 | ... | 1 | 0 | 0 |
| 1 | 0 | ... | 1 | 0 | 0 |
| 0 | 1 | ... | 0 | 1 | 0 |
| | | | | | |
| | | | | | |
| | | | | | |
| 0 | 1 | ... | 0 | 0 | 1 |
| 0 | 1 | ... | 0 | 1 | 0 |

FIGURE 1 – 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.

Pour l'AFDM, la matrice **X** correspond au tableau de données où les variables qualitatives ont été recodées de la même façon qu'en ACM, tandis que les variables quantitatives elles ne sont pas modifiées. La métrique sur l'espace des individus est ici $M = diag\left(s^2_{X_1}, \ldots, s^2_{X_{J_1}} n_{J_1+1}/n, \ldots, n_{J_2}/n\right)$ la matrice diagonale telle que les $J_1$ premiers éléments diagonaux soient les variances des variables quantitatives et les $J_2$ derniers soient les proportions de chaque modalités de réponses. La distance entre deux individus s'écrit :

$$d^2_{i,i'} = \sum_{j=1}^{J_1} \left(\frac{x_{ij} - x_{i'j}}{s_{X_j}}\right)^2 + \sum_{j=J_1+1}^{J_2} \left(\frac{x_{ij} - x_{i'j}}{\sqrt{n_j}}\right)^2 \quad (5)$$

Cette métrique permet notamment d'équilibrer l'influence des variables quantitatives et qualitatives dans la définition des distances entre individus. La pondération associée reste ici aussi $\mathbf{D} = \frac{1}{n}\mathbb{I}_n$.

Les méthodes d'analyse factorielle se distinguent par les métriques utilisées dans l'espace des individus. La métrique sur l'espace des variables définit le poids des individus et reste identique quelque soit la nature des variables. L'estimation des paramètres de ces méthodes est effectuée à l'aide d'une décomposition en valeurs singulières. Pour pouvoir imputer selon ces méthodes, il est nécessaire de savoir estimer ces paramètres en présence de données manquantes et donc d'effectuer une décomposition en valeurs singulières avec données manquantes.

# 2 Estimation des paramètres sur un jeu incomplet

Différentes méthodes ont été proposées pour estimer les paramètres des méthodes d'analyse factorielle en présence de données manquantes dans le cadre de l'ACP (*e.g.* Christoffersson (1970); Josse *et al.* (2009); Wasito et Mirkin (2005, 2006)). Parmi elles, Josse *et al.* (2009) ont proposé un algorithme appelé ACP itérative. Celui-ci consiste à estimer

les paramètres en alternant des étapes d'imputation du jeu de données et d'estimation des composantes principales et vecteurs propres. Le tableau incomplet est d'abord complété par des valeurs initiales, puis les composantes principales et les vecteurs propres sont estimés sur le jeu rendu complet. Les données manquantes sont alors mises à jour en utilisant les données reconstituées. Ces étapes sont répétées jusqu'à convergence. Plus précisément, l'algorithme s'écrit comme suit :

1. Initialisation : la matrice $\mathbf{X}$ incomplète est centrée, ses valeurs manquantes remplacées par la valeur 0 (ce qui correspond à une imputation par la moyenne). On obtient une matrice $\mathbf{X}^0$. La variance des variables est calculée de façon à initialiser la métrique $\mathbf{M}$ sur l'espace des individus. On note cette métrique $\mathbf{M}^0$.

2. Pour $\ell$ de 1 à $L$ :

   (a) centrage de $\mathbf{X}^{\ell-1}$

   (b) estimation des paramètres : on effectue la SVD de $\left(\mathbf{X}^{\ell-1}, \mathbf{M}^{\ell-1}, \mathbf{D}\right)$. On obtient les composantes principales $\widehat{\mathbf{U}}^{\ell}\widehat{\mathbf{\Lambda}}^{1/2}$ et les vecteurs propres $\widehat{\mathbf{V}}^{\ell\top}$

   (c) reconstitution : les valeurs manquantes de $\mathbf{X}$ sont remplacées par celles de $\widehat{\mathbf{X}}^{\ell} = \widehat{\mathbf{U}}^{\ell}\widehat{\mathbf{\Lambda}}^{1/2\ell}\widehat{\mathbf{V}}^{\top\ell}$ et le jeu de données est décentré. On obtient un nouveau tableau $\mathbf{X}^{\ell}$

   (d) la métrique $\mathbf{M}^{\ell}$ est mise à jour en estimant la variance des variables de $\mathbf{X}^{\ell}$

Cet algorithme permet d'estimer les paramètres de l'ACP en présence de données manquantes. Notons qu'il fournit également à la suite de chaque étape de reconstitution un nouveau jeu de données imputé $\mathbf{X}^{\ell}$. La procédure a été étendue à l'ACM par (Josse *et al.*, 2012) en recodant les variables et en adaptant la métrique de l'espace des individus. Cet algorithme pourrait aussi être étendu à l'AFDM.

Disposant d'un algorithme pour estimer les paramètres d'une méthode d'analyse factorielle en présence de données manquantes, il est maintenant possible de proposer une méthode d'imputation basées sur ces méthodes.

# 3 Imputation simple par analyse factorielle

Les travaux précédents, permettant d'estimer les paramètres de l'ACP (Josse *et al.*, 2009) et de l'ACM (Josse *et al.*, 2012) en présence de données manquantes, laissaient déjà entrevoir un moyen d'effectuer de l'imputation simple pour des données quantitatives ou qualitatives. Néanmoins, les algorithmes itératifs proposés n'ont jamais été étudiés en termes de qualité de prédiction des données manquantes. Ainsi, cette section présente l'imputation simple par composantes principales à travers l'imputation par AFDM, l'imputation par ACP ou ACM pouvant être vues comme des cas particuliers. L'objectif ici n'est pas d'appliquer une méthode statistique sur un tableau incomplet mais d'évaluer les propriétés des

méthodes d'imputation par méthodes en composantes principales. Par conséquent, on s'intéressera ici à la qualité de prédiction des données manquantes. Cette section correspond à l'article Audigier *et al.* (2014).

# A principal component method to impute missing values for mixed data

Vincent Audigier, François Husson, Julie Josse
Agrocampus Ouest, 65 rue de St-Brieuc, F-35042 Rennes
Tel.: +33-223485874
Fax: +33-223485871
audigier@agrocampus-ouest.fr
husson@agrocampus-ouest.fr
josse@agrocampus-ouest.fr|

Dec, 2014

**Abstract**

We propose a new method to impute missing values in mixed data sets. It is based on a principal component method, the factorial analysis for mixed data, which balances the influence of all the variables that are continuous and categorical in the construction of the principal components. Because the imputation uses the principal axes and components, the prediction of the missing values is based on the similarity between individuals and on the relationships between variables. The properties of the method are illustrated via simulations and the quality of the imputation is assessed using real data sets. The method is compared to a recent method (Stekhoven and Bühlmann, 2011) based on random forest and shows better performance especially for the imputation of categorical variables and situations with highly linear relationships between continuous variables.

*Keywords* missing values, mixed data, imputation, principal component method, factorial analysis of mixed data.

## 1 Introduction

Missing data are a key problem in statistical practice. Indeed, they are never welcome, because most statistical methods cannot be applied directly on an incomplete data set. One of the common approaches to deal with missing values consists of imputing missing values by plausible values. This leads to a complete data set that can be analyzed by any statistical method. However, results must be interpreted cautiously, since there is necessarily uncertainty associated with the prediction of values.

Several imputation methods are available for continuous data such as K-nearest neighbors imputation (Troyanskaya et al, 2001), imputation based on multivariate normal model (Schafer, 1997) or multivariate imputation by chained equations (van Buuren et al, 1999; van Buuren, 2007). The multivariate normal model defines a joint distribution for the data which in practice can be restrictive. Imputation by chained equations consists of defining a model for each variable with missing data, which can afford a finer modeling but requires defining many models. It is also possible to impute continuous data with principal component analysis (PCA). The idea is to use an algorithm that performs PCA despite the missingness of some data (Kiers, 1997; Josse et al, 2009; Ilin and Raiko, 2010). Principal components and axes obtained by this algorithm are then used to reconstruct the data, which provides an imputation of the missing entries. This method has the particular advantage of simultaneously imputing any missing data by taking into account the similarities between individuals and the relationships between variables.

The imputation of categorical data can also be done with non-parametric methods such as the K-nearest neighbours or with parametric methods based on different models. The most common model is the log-linear one (Schafer, 1997). It has a major drawback: the number of parameters increases rapidly with the number of levels of categorical variables. Therefore, in practice, it becomes unusable in some cases. Other models have been proposed to overcome this problem, such as the latent class model (Vermunt et al, 2008), or the log-linear hierarchical model (Schafer, 1997; Little and Rubin, 1987, 2002).

Finally, for mixed data, *i.e.* both continuous and categorical, literature is less abundant. Indeed, dealing with mixed data is difficult because it requires to take into account the relationships between the variables that are of different types. One possible solution consists of coding the categorical variables using the indicator matrix of dummy variables, and using an imputation method dedicated to continuous variables on the concatenated matrix of continuous variables and the indicator matrix. However, this method is not satisfactory since the usual assumptions for continuous variables do not hold for dummy variables. Schafer (1997) proposed an imputation based on the general location model which can be seen as combination of the log-linear model and multivariate normal model; this imputation has the benefits and drawbacks of such models. Imputation by chained equations (van Buuren et al, 1999; van Buuren, 2007) is one of the only approaches that can be easily extended to the mixed case by defining a specific model for each continuous and each categorical variable. However, as mentioned for the continuous case, many models have to be defined. Recently, Stekhoven and Bühlmann (2011) proposed an imputation method based on random forest (Breiman, 2001). The imputation is done using the following iterative algorithm: after replacing the missing data with initial values, the missing values of the variable with the fewest missing values are predicted by random forest. This operation is performed for each variable in the data set and the procedure is repeated until the predictions stabilize. For mixed data, this method was compared to the imputation by chained equations and to a version of the K-nearest neighbours method adapted for mixed data. Their approach clearly

outperforms the competing methods across many simulations and real data sets. It provides a good quality of imputation regardless of the number of observations and variables and the type of relationship between variables. Further, it is largely insensitive to the tuning parameters. Thus this non-parametric method based on random forest can serve as a reference among the existing methods.

We propose a new imputation method for mixed data based on a principal component method dedicated to mixed data: *the factorial analysis for mixed data* (FAMD) presented in Escofier (1979), also known as PCAMIX (Kiers, 1991). Because the method is based on a principal component method, imputation using FAMD allows predicting missing values using the similarities between individual and the relationships between variables simultaneously, i.e., between continuous variables, between categorical variables and between variables of different types. The specificity of FAMD lies in weighting each variable in order to balance the influence of each one, while taking into account the type of each variable in the weighting. Thus the imputation method based on this principal component method uses well the structure of the data set to impute it. We begin by presenting the imputation method (Section 2) and then illustrate its properties using simulations (Section 3). Finally, the method is assessed on real data sets (Section 4). The competitiveness of the method is highlighted by comparing its performances to the ones of Stekhoven and Bühlmann (2011) method.

# 2 Imputation for mixed type-data using the factorial analysis for mixed data

## 2.1 FAMD in the complete case

FAMD is a principal component method to describe, summarize and visualize a matrix with mixed data. As with any principal component method, its aim is to study the similarities between individuals, the relationships between variables (here continuous and categorical variables) and to link the study of the individuals with that of the variables. Such methods reduce the dimensionality of the data and provide the subspace that best represents the data.

The principle of FAMD is to balance the influence of the continuous and the categorical variables in the analysis. The rationale is to weight the variables in such a way that each variable of either types contributes equally to the construction of the principal components. It is the same idea as scaling for continuous variables in PCA. As mentioned by Benzécri (1973): "Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize".

Let us denote $I$ as the number of individuals, $K_1$ as the number of continuous variables, $K_2$ as the number of categorical variables and $K = K_1 + K_2$ as the total number of variables. Suppose the first $K_1$ variables are the continuous ones and the last $K_2$ variables are the categorical ones. The first step of FAMD consists of coding the categorical variables using the indicator matrix of dummy

variables. We denote $\mathbf{X}_{I\times J}$ as the matrix where $(\mathbf{x}_j)_{1\leq j\leq K_1}$ are continuous variables and $(\mathbf{x}_j)_{K_1+1\leq j\leq J}$ are dummy variables. The total number of columns is $J = K_1 + \sum_{k=K_1+1}^{K} q_k$ where $q_k$ is the number of categories of the variable $k$.

FAMD can be represented as the PCA of $\left((\mathbf{X}-\mathbf{M})\mathbf{D}_\Sigma^{-1/2}\right)$ where $\mathbf{M}_{I\times J}$ is the matrix with each row being the vector of the means of each column of $\mathbf{X}$ and $\mathbf{D}_\Sigma$ is the diagonal matrix $diag\left(\hat{\sigma}^2_{x_1},...,\hat{\sigma}^2_{x_{K_1}},p_{K_1+1},\ldots,p_j,\ldots,p_J\right)$ with $\hat{\sigma}_{x_j}$ being the standard deviation of the continuous variable $\mathbf{x}_j$ and $p_j$ being the proportion of individuals in the category $j$ ($j = K_1+1,\ldots,J$). The matrix $\mathbf{D}_\Sigma$ is the metric used to compute distances between rows. The loss function (known as the reconstruction error) which is minimized in the PCA of matrix $\mathbf{X}$ is:

$$\|\mathbf{X}-\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\|^2 \tag{1}$$

Thus, FAMD can be defined as minimizing:

$$\|(\mathbf{X}-\mathbf{M})\mathbf{D}_\Sigma^{-1/2}-\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\|^2 \tag{2}$$

FAMD provides the best low rank ($S < (J-K_2)$) approximation of the matrix $(\mathbf{X}-\mathbf{M})\mathbf{D}_\Sigma^{-1/2}$ in the least square sense. The solution is given by the singular value decomposition (SVD) of the matrix $(\mathbf{X}-\mathbf{M})\mathbf{D}_\Sigma^{-1/2}$ with $\hat{\mathbf{U}}_{I\times S}$ being the left singular-vectors and $\hat{\mathbf{V}}_{J\times S}$ being the right-singular vectors associated with the $S$ largest singular values gathered in the matrix $\left(\hat{\mathbf{\Lambda}}_{\boldsymbol{S\times S}}\right)^{1/2} = diag\left(\sqrt{\hat{\lambda}_1},...,\sqrt{\hat{\lambda}_S}\right)$. Notice that the maximum number of non-null eigenvalues is $(J-K_2)$ because of the linear restrictions on the columns for the categorical variables (the row sum for each variable equals 1).

The specific weighting implies that the distances between two individuals $i$ and $i'$ in the initial space (before approximating the distances by keeping the first $S$ dimensions obtained from the SVD) is:

$$d^2(i,i') = \sum_{k=1}^{K_1}\frac{(x_{ik}-x_{i'k})^2}{\hat{\sigma}^2_{\mathbf{x}_k}} + \sum_{j=K_1+1}^{J}\frac{1}{p_j}(x_{ij}-x_{i'j})^2$$

Weighting by $\frac{1}{\hat{\sigma}^2_{\mathbf{x}_k}}$ ensures that units of continuous variables do not influence the (square) distance between individuals. Furthermore, weighting by $\frac{1}{p_j}$ implies that two individuals in different categories for the same variable are more distant when one of them is in a rare category than when both of them are in frequent categories. The marginal frequencies of the categorical variables play an important role in this method. The frequencies are related to the variance in the initial space, also called inertia, of the category $j$ (Escofier, 1979): $\text{Inertia}(\mathbf{x}_j) = 1-p_j$. Categories with a small frequency have a greater inertia

than the others and consequently rare categories have a greater influence on the construction of the principal components.

The specific weighting implies also that, in FAMD, the principal components maximize the associations with both continuous and categorical variables. More precisely, the first principal component $\mathbf{f}_1$ maximizes

$$\sum_{k=1}^{K_1} R^2\left(\mathbf{x}_k, \mathbf{f}_1\right) + \sum_{k=K_1+1}^{K} \eta^2\left(\mathbf{z}_k, \mathbf{f}_1\right) \tag{3}$$

with $(\mathbf{z}_k)_{k=K+1,\ldots,K}$ the categorical variables. The first principal component is the synthetic variable the most correlated with both the continuous variables in terms of the coefficient of determination ($R^2$), and the categorical variables in terms of the squared correlation ratio ($\eta^2$). The second principal component is the synthetic variable which maximizes the criterion among variables orthogonal to the first principal component, etc.

Regarding criterion (3), we can note that FAMD reduces to PCA when there are only continuous variables and reduces to multiple correspondence analysis (Lebart et al, 1984; Greenacre and Blasius, 2006) when there are only categorical variables.

## 2.2 The iterative FAMD algorithm

An approach commonly used to deal with missing values in exploratory data analysis methods (such as PCA) consists of ignoring the missing values by minimizing the reconstruction error over all non-missing elements. For PCA, this can be achieved by introducing a weight matrix $\mathbf{W}$ (with $w_{ij} = 0$ if $x_{ij}$ is missing and $w_{ij} = 1$ otherwise) in the least square criterion (1):

$$\left\|\mathbf{W} * \left(\mathbf{X} - \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right)\right\|^2 \tag{4}$$

with $*$ being the Hadamard product. Different algorithms can be used to minimize this criterion such as an algorithm of iterative imputation of the missing entries during the estimation of the parameters (Kiers, 1997). The algorithm essentially sets the missing elements at initial values, performs the PCA on the completed data set, imputes the missing values with values predicted by the reconstruction formula (defined by the fitted matrix obtained with the axes and components) using a predefined number of dimensions, and repeats the procedure on the newly obtained matrix until the total change in the matrix falls below an empirically determined threshold. This type of algorithm can thus be seen as a single imputation method and consequently it is a method to impute continuous data based on PCA.

Since FAMD has been presented as the PCA of the matrix $\left((\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2}\right)$, criterion (2) becomes, with the addition of missing values:

$$\left\|\mathbf{W} * \left((\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2} - \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right)\right\|^2 \tag{5}$$

The methodology used to impute data with PCA can be extended to FAMD, but the algorithm must be adapted to take into account the specificities of FAMD. More precisely, the same algorithm can be used but the matrix $\mathbf{D}_\Sigma$ as well as the mean matrix $\mathbf{M}$ must be updated during the estimation process because they depend on all the data. Indeed, after imputing data with the reconstruction formula, the variance of the continuous variables as well as the column margins of the categorical variables of the new data table change. It is not recommended to fix $\mathbf{M}$ and $\mathbf{D}_\Sigma$ from the observed data as a first step in the analysis. The estimates could be seriously biased if the mechanism generating missing values is not completely at random (Rubin, 1976), meaning that the probability that a value is missing is unrelated to the value itself and any values in the data set, missing or observed. Since $\mathbf{M}$ and $\mathbf{D}_\mathbf{\Sigma}$ are estimated during the algorithm, it is not guaranteed that the criterion (5) is minimized with the iterative algorithm which is then merely a heuristic.

The algorithm to impute missing values for mixed data based on FAMD begins as follow. There is initially a table of mixed data with missing values (first table in Figure 1). This table is then transformed to obtain the matrix $\mathbf{X}$ coding categorical variables using an indicator matrix of dummy variables. A missing value on a categorical variable then leads to a row of missing values in the indicator matrix (second table in Figure 1). Then this data table is imputed according to the following algorithm. The imputation algorithm, based on FAMD and called iterative FAMD, then proceeds as follows:

1. initialization $\ell = 0$: substitute missing values by initial values and calculate $\mathbf{M}^0$ and $\mathbf{D}_\Sigma^0$ on this completed data set.
2. step $\ell$:
   (a) perform the FAMD, in other words the PCA of $\left(\mathbf{X}^{\ell-1} - \mathbf{M}^{\ell-1}\right)\left(\mathbf{D}_\Sigma^{\ell-1}\right)^{-1/2}$ to obtain $\hat{\mathbf{U}}^\ell$, $\hat{\mathbf{V}}^\ell$ and $\left(\hat{\mathbf{\Lambda}}^\ell\right)^{1/2}$;
   (b) keep the first $S$ dimensions and use the reconstruction formula to compute the fitted matrix:
   $$\hat{\mathbf{X}}^\ell_{I\times J} = \left(\hat{\mathbf{U}}^\ell_{I\times S}\left(\hat{\mathbf{\Lambda}}^\ell_{S\times S}\right)^{1/2}\left(\hat{\mathbf{V}}^\ell_{J\times S}\right)'\right)\left(\left(\mathbf{D}_\Sigma^{\ell-1}\right)_{I\times J}\right)^{1/2} + \mathbf{M}^{\ell-1}_{I\times J}$$
   and the new imputed data set becomes $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (\mathbf{1} - \mathbf{W}) * \hat{\mathbf{X}}^\ell$ with $\mathbf{1}_{I\times J}$ being a matrix with only ones. The observed values are the same but the missing ones are replaced by the fitted values;
   (c) from the new completed matrix $\mathbf{X}^\ell$, $\mathbf{D}_\Sigma^\ell$ and $\mathbf{M}^\ell$ are updated.
3. steps (2.a), (2.b) and (2.c) are repeated until the change in the imputed matrix falls below a predefined threshold $\sum_{ij}(\hat{x}_{ij}^{\ell-1} - \hat{x}_{ij}^\ell)^2 \leq \varepsilon$, with $\varepsilon$ equals to $10^{-6}$ for example.

In the initialization step, missing values are replaced, for instance, by the mean of the variable for the continuous variables and the marginal proportion of the category for each category using the non-missing entries. Note that for the categorical variables, the sum of the initial entries corresponding to one individual and one categorical variable must equal one. At the end of the algorithm, imputed values for the missing entries for the categories are not equal to 0 and 1 but are real numbers (third table in Figure 1). However, the constraint in the initialization step ensures that the sum of the entries for one individual and one categorical variable is equal to 1. This property comes from the specific weighting and is demonstrated in the framework of multiple correspondence analysis (Tenenhaus and Young, 1985; Josse et al, 2012). The same proof hold in the framework of FAMD. Consequently, the imputed values can be considered as degrees of membership to the associated category and it is possible to impute the categorical variable with the most plausible value (last table in Figure 1).

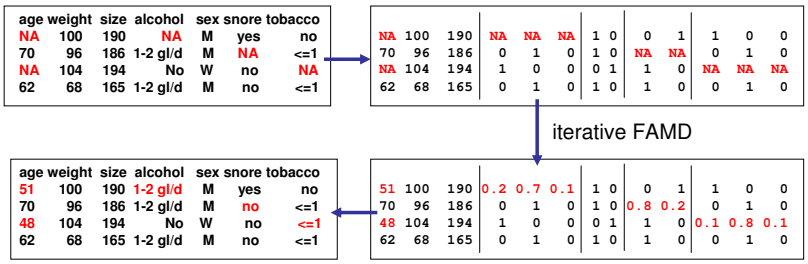


**Figure 1:** Diagram for the iterative FAMD algorithm: the raw mixed data, the matrix $\mathbf{X}$, the imputed data obtained by iterative FAMD and the imputed mixed data.

Such algorithms which alternate a step of estimation of the parameters via a singular value decomposition and a step of imputation of the missing values are known to suffer from overfitting problems. These problems occur even if these methods reduce the dimensionality of the data. A fortiori, when there are missing values, the overfitting is more marked which means that imputed values are over-dispersed. Yet, the more missing values there are, the less trusted the relationships between variables are. Consequently, it would be more satisfactory to impute using the mean imputation than using a model taking into account the relationships between variables. Geometrically, mean imputation is equivalent to bring the individuals closer from the center of gravity of the point cloud, which limits the dispersion. In order to avoid these major problems of overfitting, the iterative singular value decomposition algorithm has been replaced by an iterative thresholded singular value decomposition algorithm in the framework of PCA (Josse and Husson, 2012; Mazumder et al, 2010). It is the same rationale as in ridge regression where the regularized version stabilizes the prediction in comparison with the ordinary least squares. We follow the approach proposed

by Josse and Husson (2012) and define a regularized iterative FAMD algorithm by replacing the singular values $\left(\sqrt{\hat{\lambda}_s^\ell}\right)_{s=1,\ldots,S}$ of step (2.b) by $\left(\frac{\hat{\lambda}_s^\ell - \hat{\sigma}^2}{\sqrt{\hat{\lambda}_s^\ell}}\right)_{s=1,\ldots,S}$ with $\hat{\sigma}^2 = \sum_{s=S+1}^{J-K_2} \frac{\lambda_s}{J-K_2-S}$. The rationale is to remove the noise in order to avoid instabilities in the prediction. Implicitly, it is assumed that the first $S$ dimensions contain both information and noise whereas the last ones contain only noise; hence, the variance of the noise is estimated by the mean of the last eigenvalues. The regularized method shrinks the first $S$ singular values with a greater amount of shrinkage for the smallest ones, which is acceptable since these small singular values are more responsible for instability. When the noise is small, the regularized algorithm is quite similar to the non-regularized one. When the variance of the noise is large, $(\hat{\lambda}_s - \hat{\sigma}^2)$ tends to zero, and the regularized algorithm simply imputes continuous variables with their means and imputes categorical variables with their marginal proportions. This is acceptable because when the data are overwhelmed by noise, the structure of the data (the relationship between variables) is very weak and imputing with the mean of the variables is an effective strategy. Thus the regularized method provides imputed values less dispersed than for the non-regularized one. The regularized iterative FAMD algorithm is also a heuristic since no explicit penalized loss function is optimized.

Note that FAMD is close to another extension of the PCA for mixed data, called nonlinear PCA. When there are only categorical data, nonlinear PCA without restrictions (also known as homogeneity analysis) is equivalent to Multiple Correspondence Analysis. Nonlinear PCA allows the analysis of variables of different nature via for instance the rank-1 restrictions (see, Michailidis and de Leeuw (1998)). There are solutions to deal with missing values in the Gifi system such as the 'missing data passive' approach (Gifi, 1990). The rationale for categorical variables (van der Heijden and Escofier, 2003) is based on the following assumption: if an individual $i$ is missing on the variable $j$, one considers that the individual has not chosen any category for the variable. Consequently, in the indicator matrix, the entries in the row corresponding to individual $i$ and variable $j$ are marked 0. Josse et al (2012) compared thoroughly this approach to the equivalent of iterative FAMD for categorical variables and showed that the "missing data passive" method is equivalent to adding a new category for the missing values and then to putting it as supplementary element. Consequently the rationale of both approaches is different: in the former, when there is a missing value for a categorical variable, one considers that the individual has not chosen any category for the variable whereas in the latter, one considers that a missing value represents an underlying category among the available categories. In the Gifi system, another method is available to deal with missing values named "missing multiple" which consists in adding a new category for each missing values. van der Heijden and Escofier (2003) described the issues raised by this approach. Note that in the Gifi system, no attempts have been made in order to use the approach as a possible imputation method for mixed data even if the extension could be imagined.

# 3 Properties of the imputation method

We discuss the main properties of the new imputation method and illustrate those properties on different toy data sets. We focus on the regularized version of the algorithm in order to avoid overfitting problem. In addition this version converges faster, which is more convenient to perform simulations. We compare the imputation results obtained with regularized iterative FAMD algorithm to the ones obtained with the method based on random forest (Stekhoven and Bühlmann, 2011) to highlight some important properties. In the latter method, some parameters such as the number of trees per forest, the number of variables selected for each forest as well as minimum size of terminal nodes can be tuned. However, Stekhoven and Bühlmann (2011) modify these parameters mainly to minimize computational time, finding that modifying these parameters offers little improvement of the imputations themselves. Consequently, the parameters suggested by default in their implementation of the method were used in the simulations.

### Simulation process

We simulate toy data sets which differ with respect to the number of continuous variables, the number of categorical variables, the number of categories per variable, the number of individuals per category, the number of underlying dimensions and the strength of the relationship between variables through different signal to noise ratios (Mazumder et al, 2010). The signal to noise ratios (SNR) is the square root of the ratio between the variance of the signal and the variance of the noise. In the particular case of continuous data with variance equal to 1, the variance of the signal corresponds to the number of variables and the variance of the noise corresponds to $\sigma^2$ times this one. Consequently the SNR is simply the inverse of $\sigma$. Thus, a high SNR implies that the variables are very correlated, whereas a low SNR implies that the data are very noisy. More precisely, the toy data sets are almost all simulated according to the following procedure:

- $S'$ independent variables are drawn from a standard Gaussian distribution;
- each variable $s'$ (for $s' = 1, \ldots, S'$) is replicated $K^{s'}$ times which guarantees (in expectation) $S'$ orthogonal groups of correlated variables. $S'$ will be called the number of underlying dimensions;
- Gaussian noise is added with different levels of variance to obtain different signal to noise ratios;
- categorical variables are obtained by splitting continuous variables in equal count categories.

Then, we insert different percentages of missing values (10%, 20% and 30%) completely at random (Rubin, 1976). The code to reproduce all the simulations is available on the webpage of the first author. For each set of parameters, 200 simulations are performed.

### Criteria

Two criteria are used to assess the quality of the imputation, the proportion of falsely classified (PFC) entries for categorical variables and the normalized root mean squared error (NRMSE) for continuous data:

$$NRMSE = \sqrt{\frac{\sum_{k=1}^{K_1}\sum_{i=1}^{I}(1-w_{ik})\left(\frac{x_{ik}-\hat{x}_{ik}}{\hat{\sigma}_{\mathbf{x}_k}}\right)^2}{\sum_{k=1}^{K_1}\sum_{i=1}^{I}(1-w_{ik})}}$$

$NRMSE$ allows the consideration of variables with different variances. Moreover, when $NRMSE$ equals zero, the imputation is perfect, whereas when it is close to 1, the imputation yields results similar to those obtained using the mean imputation.

## 3.1 Relationships between continuous and categorical variables

The fundamental objective of FAMD is to take into account relationships between continuous and categorical variables. Taking into account both types of variables improves the imputation as is illustrated with a data set that has two underlying dimensions ($S' = 2$). Each dimension is composed of two continuous variables and two categorical variables with four categories. Missing data are then added completely at random for the three selected percentages and finally the imputation algorithm is performed according to three strategies:

(1) using only continuous variables, which leads to an imputation of continuous variables with only those variables;

(2) using only categorical variables, which leads to an imputation of categorical variables with only those variables;

(3) using and imputing variables of both kinds.

Figure 2 compares the distributions of the NRMSE for the three percentages of missing data according to the strategies (1) and (3). As expected, when the proportion of missing data increases, the imputation error is larger. When the SNR decreases, the quality of the imputation (not shown here) decreases which is also expected. It can be noted that even with 30% missing data the imputation with iterative FAMD greatly outperforms the mean imputation (NRMSE less than 1). This is due to the relationships between variables, which improve the mean imputation that forms the first step of iterative FAMD. The imputation error is lower when considering both types of variables (boxplots in grey) than when considering only continuous variables (boxplots in dark grey). Taking into account categorical variables thus improves the imputation of continuous variables. This behavior is the same regardless of the proportion of missing data.

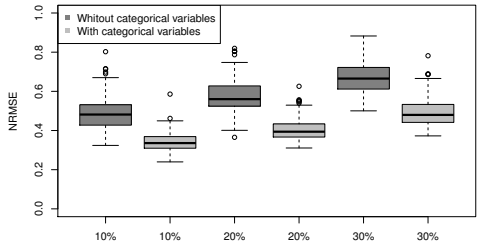


**Figure 2:** Distribution of the $NRMSE$ for different amounts of missing values (10%, 20%, 30%). Dark grey boxplots correspond to the error of imputation for continuous variables when only continuous variables are used whereas grey boxplots correspond to the error when categorical and continuous variables are used.

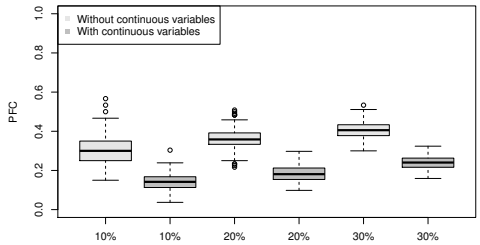


**Figure 3:** Distribution of the $PFC$ for different amounts of missing values (10%, 20%, 30%). Light grey boxplots correspond to the error of imputation for categorical variables when only categorical variables are used, whereas grey boxplots correspond to the error when categorical and continuous variables are used.

Figure 3 compares the distributions of rates of misclassification according to the strategies (2) and (3). The results of the categorical variables are similar to those obtained for continuous variables: when the rate of missing data increases, the proportion of misclassification increases, but even for 30% missing data the imputation by FAMD yields better results than a random imputation (the latter having error 0.66). Regardless of the rate of missing data, taking into account the continuous variables for the imputation of categorical variables (light grey boxplots) reduce the proportion of misclassification.
*Remark*: It is possible in theory to perfectly impute a categorical variable using a continuous variable. On the contrary, it is difficult to impute the continuous variables with only categorical variables. For example, using $K_2$ categorical variables with $q$ categories each can produce only $q^{K_2}$ distinct imputations. Consequently, this imputation cannot reflect all possible values that the continuous variable can take. However, in practice it can be a reasonable imputation.

## 3.2 Influence of the relationships between variables

### 3.2.1 Linear and nonlinear relationships

FAMD can be seen as a variation of PCA, and PCA is based on linear relationships between variables. When there are strong linear relationships between continuous variables, the imputation of these variables with iterative FAMD will thus be accurate. To illustrate this behavior, a data set is generated according to the simulation process with $S' = 1$ initial variable from which 2 continuous variables and 3 categorical variables with 4 categories are built. The imputation by FAMD (Figure 4, boxplots in grey) is compared to the imputation based on random forest (Figure 4, boxplots in white). The error for continuous variables as well as the error for categorical variables with iterative FAMD is very small and is much smaller than the error of the algorithm based on random forest. Moreover, when the percentage of missing values increases, the error of the iterative FAMD algorithm increases slightly, whereas the error of the algorithm based on random forest increases more. Such results are representative of all the results obtained with different data sets.

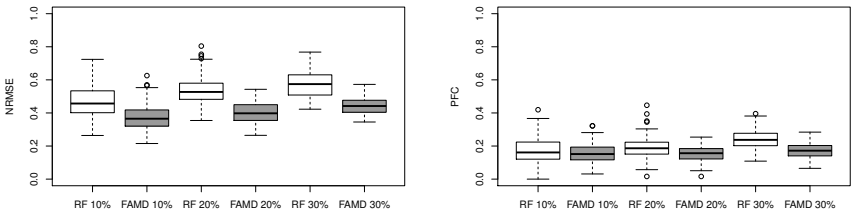


**Figure 4:** Distribution of the NRMSE (left) and of the PFC (right) when the relationships between variables are linear for different amounts of missing values (10%, 20%, 30%). White boxplots correspond to the imputation error for the algorithm based on random forest (RF) and grey boxplots to the imputation error for iterative FAMD.

What about nonlinear relationships in FAMD? Thanks to the presence of categorical variables, FAMD may impute missing values when there are nonlinear relationships between the continuous variables. Indeed, the principal components of FAMD are linear combinations of the continuous variables and of the columns of the indicator matrix (equation 3). A linear combination of dummy variables may approximate a nonlinear function of a variable by a piecewise constant function. To illustrate this behavior, a data set is generated with $S' = 1$ initial variable from which 3 continuous variables and 1 categorical variable with 10 categories are built. The results of applying FAMD to this data, illustrated in the left panel of Figure 5 are in accordance with the previous ones: the imputation with FAMD is very accurate when there are linear relationships. Then we take the same data set but the second continuous variable is squared and the cosine function is applied to the third variable. In this case, the results obtained by iterative FAMD are worse than those obtained by the algorithm

based on random forest (Figure 5, graph on the right), which is known to deal well with nonlinear relationships. However, the difference in performance is modest.

Remark: in practice, if nonlinear relationships between continuous variables are suspected, a solution can be to create new categorical variables by discretizing the continuous variables into categories.

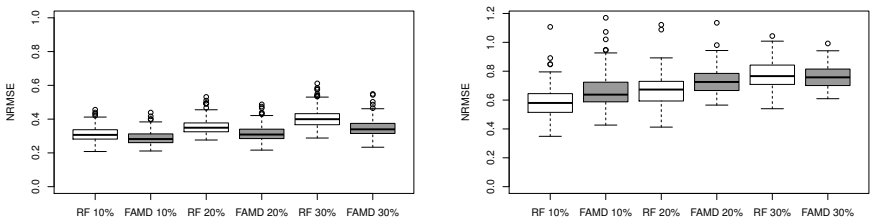


**Figure 5:** Distribution of the NRMSE when the relationships between variables are linear (left) and nonlinear (right) for different amounts of missing values (10%, 20%, 30%). White boxplots correspond to the imputation error for the algorithm based on random forest (RF), and grey boxplots to the imputation error for iterative FAMD.

### 3.2.2 Taking into account interactions between categorical variables

Other relationships between variables can be challenging. FAMD is based on relationships between pairs of variables. Consequently data including complex interactions could make the imputation difficult. To illustrate this behavior, a data set with 3 variables (1 continuous and 2 categorical) illustrated in Table 1 is constructed. It consists of a fractional factorial design $3^{3-1}$ that is replicated several times (with replications vertically stacked). Variables are pairwise independent but there are interactions between them.

The quality of the imputation of the continuous and categorical variables is poor with iterative FAMD (Figure 6). It is similar to the mean imputation for the continuous variable and to the imputation by the proportion of the categories for the categorical variables. The imputation based on random forest takes into account the interactions between variables and provides better results. This relatively worse performance of FAMD-based imputation can be seen as a drawback. However, we can address this problem by introducing an additional variable in the data set that corresponds to the interaction, for example, by creating a variable $x_4$ which has 9 levels "aa", "ba", "ca", etc. The imputation problem thus reduces to the case without interaction and the quality of the imputation will be very good.

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ |
|---|---|---|---|
| 1 | $a$ | $a$ | 1 |
| 2 | $b$ | $a$ | 2 |
| 3 | $c$ | $a$ | 3 |
| 4 | $a$ | $b$ | 2 |
| 5 | $b$ | $b$ | 3 |
| 6 | $c$ | $b$ | 1 |
| 7 | $a$ | $c$ | 3 |
| 8 | $b$ | $c$ | 1 |
| 9 | $c$ | $c$ | 2 |
| 10 | $a$ | $a$ | 1 |
| 11 | $b$ | $a$ | 2 |
| 12 | $c$ | $a$ | 3 |
| 13 | $a$ | $b$ | 2 |
| 14 | $b$ | $b$ | 3 |
| 15 | $c$ | $b$ | 1 |
| 16 | $a$ | $c$ | 3 |
| 17 | $b$ | $c$ | 1 |
| 18 | $c$ | $c$ | 2 |
| ... | | | |

**Table 1:** data set with interaction generated with a fractional factorial design $3^{3-1}$; the defining relation of the fractional design is $I = 123$.

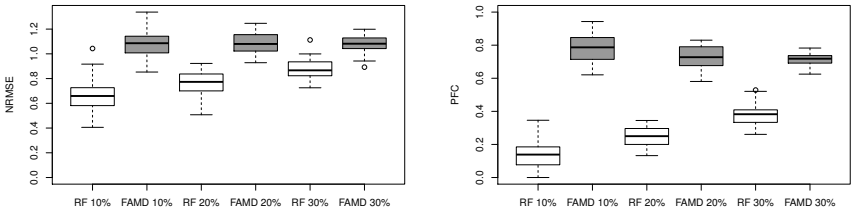


**Figure 6:** Distribution of the NRMSE (left) and of the PFC (right) when there are interactions between variables. Results are given for different amounts of missing values (10%, 20%, 30%). White boxplots correspond to the imputation error for the algorithm based on random forest (RF) and grey boxplots to the imputation error for iterative FAMD.

## 3.3 Imputation of rare categories

FAMD performed on a complete data set weights each category by the inverse of the number of individuals taking this category (section 2.1). Thus FAMD assigns more variance to rare categories both in the cloud of categories as well as in the cloud of individuals in the initial space. Consequently, rare categories are privileged when constructing the principal components and because the algorithm uses the principal components to impute the data, rare categories

may be well predicted.

In order to illustrate the ability of the method to impute rare categories, simulations have been conducted with 2 continuous variables, 3 categorical variables and 3 categories per categorical variables. The frequencies of the categories are respectively equal to $(1/3, 1/3, 1/3)$ for one categorical variable and to $(f, (1-f)/2, (1-f)/2)$ for the 2 other categorical variables, with $f$ being a frequency that varies between 0.004 and 0.1. The rare categories (each having frequency $f$) of these 2 categorical variables are related in the sense that they are taken by the same individuals. Then a rare value is suppressed for one individual on one of the two categorical variables and the imputation algorithms are performed. This strategy allows us only to focus on the prediction of a rare category. Simulations are performed for different numbers of individuals and, for a given frequency $f$, we expect that it would be easier to recover the true category when the number of individuals is large because the category contains more individuals. The results (Table 2) show that the algorithm based on FAMD successfully recovers the rare category. The advantage of FAMD over the algorithm based on random forest is especially large when imputing a rare category.

| Number of individuals | $f$ | FAMD | Random forest |
|---|---|---|---|
| 100 | 10% | 0.060 | 0.096 |
| 100 | 4% | 0.082 | 0.173 |
| 1000 | 10% | 0.042 | 0.041 |
| 1000 | 4% | 0.060 | 0.071 |
| 1000 | 1% | 0.074 | 0.167 |
| 1000 | 0.4% | 0.107 | 0.241 |

**Table 2:** Percentage of error (PFC) over 1000 simulations when recovering a rare category for data sets with different numbers of individuals and different frequencies for the rare category ($f$). Results are given for the imputation with FAMD and with the algorithm based on random forest.

## 3.4 Extensive study

Here, we assess the influence of other parameters like size, the balance between number of categorical and continuous variables and the influence of the number of categories of the categorical variables. We generate data sets using 2 underlying dimensions and vary the number of individuals (50 or 200), the number of variables (12 or 24), the proportion of continuous variables (1/3 or 2/3), the number of categories per variable (3 or 6), as well as the signal to noise ratio (2 or 4). On each data set, 10% of missing values are added, then the FAMD algorithm and the one based on random forests are performed. Results over 200 replications are gathered in Table 3.

| case | $I$ | $K$ | $K1/K$ | $q$ | SNR | FAMD PFC | RF PFC | FAMD NRSME | RF NRMSE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 12 | 0.667 | 3 | 2 | **0.171** | **0.14** | **0.325** | **0.437** |
| 2 | 50 | 12 | 0.667 | 3 | 4 | **0.061** | **0.079** | **0.247** | **0.318** |
| 3 | 50 | 12 | 0.667 | 6 | 2 | 0.498 | 0.451 | 0.408 | 0.425 |
| 4 | 50 | 12 | 0.667 | 6 | 4 | 0.125 | 0.192 | 0.229 | 0.318 |
| 5 | 50 | 12 | 0.333 | 3 | 2 | 0.107 | 0.168 | 0.402 | 0.471 |
| 6 | 50 | 12 | 0.333 | 3 | 4 | 0.035 | 0.045 | 0.398 | 0.404 |
| 7 | 50 | 12 | 0.333 | 6 | 2 | **0.267** | **0.314** | **0.403** | **0.454** |
| 8 | 50 | 12 | 0.333 | 6 | 4 | 0.076 | 0.124 | 0.294 | 0.368 |
| 9 | 50 | 24 | 0.667 | 3 | 2 | 0.149 | 0.16 | 0.285 | 0.373 |
| 10 | 50 | 24 | 0.667 | 3 | 4 | 0.023 | 0.064 | 0.204 | 0.254 |
| 11 | 50 | 24 | 0.667 | 6 | 2 | 0.258 | 0.296 | 0.364 | 0.367 |
| 12 | 50 | 24 | 0.667 | 6 | 4 | 0.039 | 0.071 | 0.211 | 0.276 |
| 13 | 50 | 24 | 0.333 | 3 | 2 | 0.093 | 0.136 | 0.371 | 0.394 |
| 14 | 50 | 24 | 0.333 | 3 | 4 | 0.015 | 0.021 | 0.379 | 0.334 |
| 15 | 50 | 24 | 0.333 | 6 | 2 | **0.08** | **0.133** | **0.4** | **0.412** |
| 16 | 50 | 24 | 0.333 | 6 | 4 | 0.018 | 0.027 | 0.296 | 0.324 |
| 17 | 200 | 12 | 0.667 | 3 | 2 | 0.165 | 0.167 | 0.26 | 0.309 |
| 18 | 200 | 12 | 0.667 | 3 | 4 | 0.074 | 0.066 | 0.203 | 0.19 |
| 19 | 200 | 12 | 0.667 | 6 | 2 | **0.311** | **0.332** | **0.246** | **0.287** |
| 20 | 200 | 12 | 0.667 | 6 | 4 | 0.122 | 0.109 | 0.154 | 0.185 |
| 21 | 200 | 12 | 0.333 | 3 | 2 | 0.094 | 0.126 | 0.363 | 0.371 |
| 22 | 200 | 12 | 0.333 | 3 | 4 | 0.027 | 0.039 | 0.358 | 0.295 |
| 23 | 200 | 12 | 0.333 | 6 | 2 | **0.16** | **0.224** | **0.267** | **0.307** |
| 24 | 200 | 12 | 0.333 | 6 | 4 | 0.047 | 0.059 | 0.232 | 0.238 |
| 25 | 200 | 24 | 0.667 | 3 | 2 | 0.088 | 0.092 | 0.22 | 0.259 |
| 26 | 200 | 24 | 0.667 | 3 | 4 | 0.039 | 0.043 | 0.181 | 0.18 |
| 27 | 200 | 24 | 0.667 | 6 | 2 | 0.16 | 0.195 | 0.212 | 0.247 |
| 28 | 200 | 24 | 0.667 | 6 | 4 | 0.053 | 0.067 | 0.142 | 0.179 |
| 29 | 200 | 24 | 0.333 | 3 | 2 | 0.07 | 0.075 | 0.335 | 0.327 |
| 30 | 200 | 24 | 0.333 | 3 | 4 | 0.022 | 0.024 | **0.347** | **0.25** |
| 31 | 200 | 24 | 0.333 | 6 | 2 | 0.085 | 0.103 | 0.266 | 0.299 |
| 32 | 200 | 24 | 0.333 | 6 | 4 | 0.038 | 0.04 | **0.234** | **0.236** |

**Table 3:** Mean of NRMSE and PFC for the FAMD algorithm and for the method based on random forests. Results are obtained over 200 simulations for 10% of missing values, varying the number of individuals ($I$), the number of variables ($K$), the proportion of continuous variables ($K1/K$), the number of categories per variable ($q$), as well as the signal to noise ratio (SNR). The results in bold are detailed in the text and are characteristic of the general trends of the imputation methods.

As expected, when the number of data increases (*i.e.* when $I$ or $K$ increases), the imputation error decreases for the two algorithms (compare for example the mean errors on the cases 7 and 23, as well as the mean error on the cases 7 and 15). In addition, when the proportion of continuous variables increases, the NRMSE decreases for continuous variables and the PFC increases for categorical variables (see for example the cases 19 and 23). This common behaviour for the two algorithms illustrates the difficulty in taking into account the links between variables of different types: an imputation between variables of a same type is easier than between variables of different types. Then the NRMSE decreases when the number of categories increases (compare for example the cases 30 and 32). This was expected since the information carried by the categorical variable is finer when the number of categories increases, which allows to better impute continuous variables. Note that the PFC increases mechanically because the probability to make a mistake is higher. Finally, the error decreases when the SNR decreases (see the cases 1 and 2 for example). Comparing to the algorithm based on random forests, most of the time, the FAMD algorithm provides better imputations for the two types of variables.

## 3.5 Choice of the number of dimensions

At each iteration of the iterative FAMD algorithm, data are reconstructed using only the $S$ first dimensions (step 2.b). If $S$ is too small then relevant information is lost and cannot be used for the imputation. On the other hand, if $S$ is too large then noise is considered as signal, which may lead to instability of the imputations. The number $S$ is thus an important parameter of the algorithm, and has to be chosen *a priori*. In this section we focus on choosing this number from an incomplete mixed data set.

First of all, we can note that a categorical variable with $q_k$ categories evolves within a space with $(q_k-1)$ dimensions. Therefore, it is impossible to predict the values of the categories with a choice of $S$ less than $(q_k-1)$. This may be a clue that guiding the choice of $S$. However, some of these $(q_k-1)$ dimensions may be unrelated to all the other variables, especially when they are many categories. In this case, even if $S$ is large, it is impossible to impute this variable correctly and choosing many dimensions may lead to instability.

Many strategies are available in the literature to select a number of dimensions from a complete data set in PCA. Cross-validation (Bro et al, 2008) or an approximation of cross-validation such as generalized cross-validation (Josse and Husson, 2011), for example, perform well. These methods have the advantage that they can be directly extended to incomplete data. Since their extension is straightforward for FAMD, in practice we use cross-validation to select the number of dimensions. However, this topic may deserve more research. Consequently, it is important to assess the impact of a poor choice for the number of dimensions on the results.

We consider a data set generated from $S' = 2$ groups of orthogonal variables ($K^{1'} = 8$ variables, 4 continuous and 4 categorical variables, and $K^{2'} = 4$ variables, 2 continuous and 2 categorical variables). Each categorical variable

has 3 categories. Consequently the underlying number of dimensions of this data set is $S = 4$. Figure 7 shows the evolution of the average of the errors over 200 simulations according to $S$ for the continuous variables (graphs on the left) and for the categorical variables (graphs on the right). Structured data sets (SNR=3) are used in the graphs on the top whereas noisy data sets (SNR=1) are used in the graphs on the bottom.

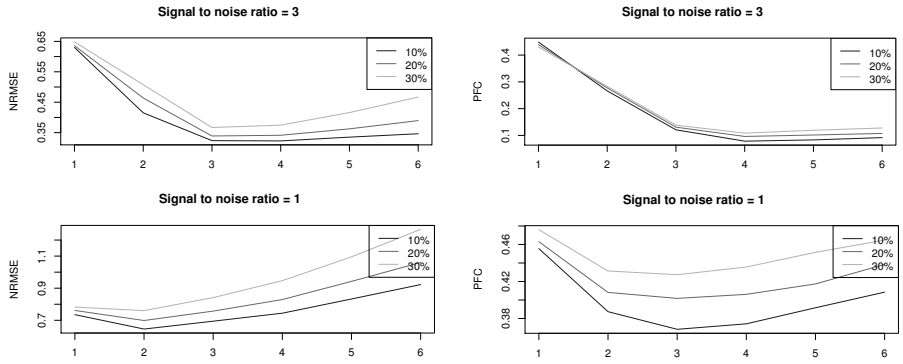


**Figure 7:** Average error of imputation over 200 simulations according to the number of dimensions used in the algorithm and for 3 amounts of missing values (10%, 20%, 30%): error for the continuous variables on the left and for the categorical variables on the right. The signal to noise ratio equals 3 for the simulations represented on the top, and 1 for the simulations represented on the bottom.

When the signal to noise ratio is high, the error for the categorical variables decreases until it reaches the optimal value 4 and then it increases slowly, regardless of the percentage of missing values. The same comment can be made for the continuous variables even if the minimum is reached for $S = 3$, since the continuous variables are only linked to the first 3 dimensions. These results were expected. However, the behavior is very different when the signal to noise ratio is small. Indeed, even for the categorical variables, it is preferable to choose fewer than the true underlying number of dimensions. This is especially true as the percentage of missing data increases. It arises because selecting a smaller number of dimensions can be regarded as performing stronger regularization which is a good strategy when the data are very noisy.
These simulations show that it is never preferable to consider more dimensions than the true number but it may be preferable to consider fewer. A good strategy is thus to use cross-validation which, in practice, produces satisfactory results in the sense that it finds the "best" number of dimensions to impute the data set: it favors the true number of dimensions when there are strong relationships between the variables and a smaller number when the data are very noisy.

# 4 Comparison on real data sets

The imputation method is evaluated on real data sets that cover many situations.The regularised imputation method is evaluated on real data sets that cover many situations. They differ in terms of number of individuals, number of variables, number of categories for the categorical variables, and they represent different areas of application. Missing values are added at random to these complete sets and then the imputation is performed with iterative FAMD and with Stekhoven and Bühlmann (2011) algorithm. Each configuration is simulated 200 times for three different percentages of missing data. The number of dimensions for the reconstruction step of the iterative FAMD algorithm is determined by cross-validation. This number is held constant for the 200 simulations in order to save computational time. The evaluation is based on the following mixed data sets.

**Tips** This data set, from the package `rggobi` (Lang et al, 2012) of the `R` software (R Development Core Team, 2011), concerns the tips given to a waiter in a restaurant in the U.S. in the early 1990s. The $K = 8$ variables of the data set concern the price of the meal for $I = 244$ customers, on the tip amount and the conditions of the restaurant meal (number of guests, time of day, etc.). There are $K_1 = 3$ continuous variables and $K_2 = 5$ categorical variables with between 2 and 6 levels.

**BMI** This data set (Lafaye de Micheaux et al, 2011) concerns body mass index of $I = 152$ French children aged 3 to 4 years. The $K = 6$ variables concern their morphology and the characteristics of their kindergarten ($K_1 = 4$, $K_2 = 2$). All the categorical variables have two levels.

**Ozone** This data set (Cornillon et al, 2012) contains $I = 112$ daily measurements of meteorological variables (wind speed, temperature, rainfall, etc.) and ozone concentration recorded in Rennes (France) during summer 2001. There are $K_1 = 11$ continuous variables and $K_2 = 2$ categorical variables with 2 or 4 levels.

**German Breast Cancer Study Group (GBSG)** This data set, from the package `ipred` (Peters and Hothorn, 2012) of the `R` software, described $I = 686$ women with breast cancer using variables concerning the status of the tumours and the hormonal system of the patient ($K_1 = 7$, $K_2 = 3$). Categorical variables have between 2 or 3 levels.

Imputation results for all the data sets are presented in Figure 8. The graphs on the left evaluate the quality of the imputation for the continuous variables (NRMSE) whereas the graphs on the right evaluate the quality of the imputation for the categorical variables. In general, the iterative FAMD provides a slightly better imputation than the one obtained by the algorithm based on

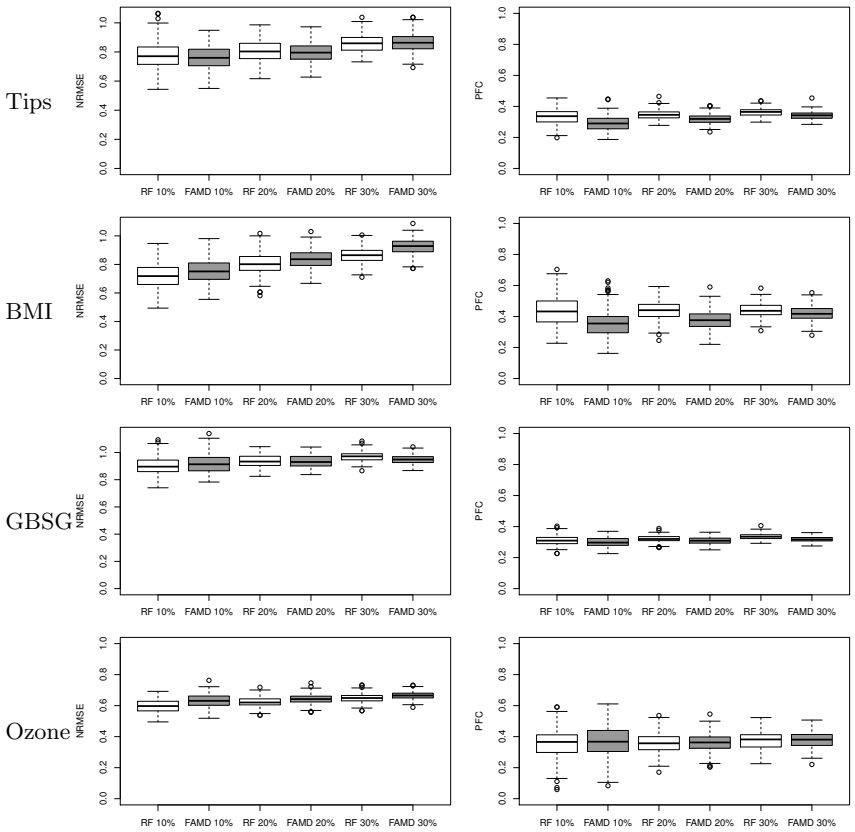


**Figure 8:** Distribution of the NRMSE (left) and of the PFC (right) for different amounts of missing values (10%, 20%, 30%) and for different data sets (Tips, BMI, GBSG, Ozone). White boxplots correspond to the imputation error of the algorithm based on random forest (RF) and grey boxplots correspond to the imputation error of iterative FAMD.

random forest. However, the results obtained by the latter algorithm are often better for continuous variables. On the Ozone and BMI data sets, the difference between the errors reached 5%. However, imputation with the iterative FAMD is more efficient on categorical variables. For the data sets on tips and BMI, the difference between the two methods is 5%. Note that the NRMSE errors may be close to 1. This means that the imputation methods improve on the mean imputation but the gain is small.

These conclusions extend to the case of non-mixed data sets. We now consider two of them: one continuous (Parkinson) and one categorical (Credit) data set.

**Parkinson** This data set (Stekhoven and Bühlmann, 2011) contains $K = 22$ measurements on the voice of $I = 195$ patients with or without Parkinson's disease. The response categorical variable sick/healthy is excluded for these simulations.

**Credit** This data set (Cornillon et al, 2012) concerns $I = 66$ customers profiles of subscribers to consumer credit in a bank. The $K = 11$ variables include the financial conditions under which the customer subscribes to the credit as well as some socio-demographic characteristics. The number of levels for these variables is between 2 and 5.

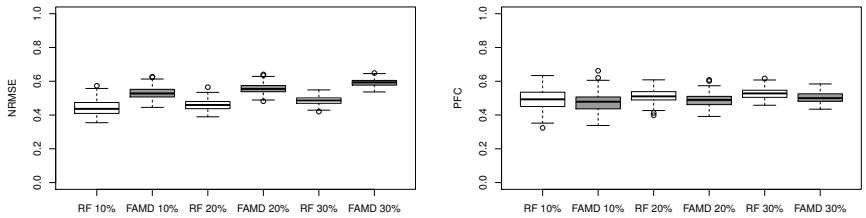


**Figure 9:** Distribution of the error for 10%, 20%, 30% of missing values for the data sets Parkinson (left) and Credit (right). White boxplots correspond to the imputation error of the algorithm based on random forest (RF) and grey boxplots correspond to the error of iterative FAMD.

For imputation of the continuous data set (Parkinson), the random forest algorithm outperforms FAMD, with a difference of about 10 % in the NRMSE (Figure 9 on the left). We thus find the same results about continuous variables as those for mixed-type data. This is accentuated by the fact that relationships between variables are here nonlinear. The lack of categorical variables implies that it is not possible to impute these variables correctly with iterative FAMD. For the categorical data set (Credit), the errors (Figure 9 on the right) are smaller with the iterative FAMD algorithm, regardless of the percentage of

missing values. These results are also similar to those observed in the mixed-type data.

As mentioned in section 2.2, an approach close to FAMD is available. Consequently, to get hints of the potential of an imputation approach using the missing passive method of Gifi in the framework of nonlinear PCA, we performed the simulations on the real data sets. We used the function homals from the package Homals (de Leeuw and Mair, 2009) of the free R software. Note that this function performs nonlinear PCA with missing values using the missing data passive approach but does not give as an output a completed dataset. Consequently, we had to modify the code to obtain an imputed mixed data matrix. We performed this method taking first the parameters by default (rank 1 restriction) on all the simulations of the real data sets. These first attempts show very bad results compared to the two other methods especially on the continuous variables. Different combinations of the parameters (restrictions, transformation) could be considered. It can be investigated for future research, but it is out of the scope of this paper.

# Conclusion

Imputing mixed data is very challenging and very few methods are proposed in the literature. The new imputation method proposed in this paper is based on a principal component method, the factorial analysis for mixed data, and allows imputation of missing data that simultaneously takes into account similarities between individuals and the associations between all variables, continuous variables and categorical variables. This method produces particularly good predictions for the missing entries of categorical variables and when there are linear relationships between variables. A strong point is that rare categories are well imputed. In addition, the method provides good results both in terms of quality of the imputation and computational time compared to the best method available based on random forests.

The iterative FAMD algorithm requires a tuning parameter which is the number of dimensions used to reconstruct the data. Cross-validation, while time-consuming, can in practice be used to select this parameter. Approximation of cross-validation such as generalized cross-validation could be proposed to select this number without resorting to an intensive computational method.

The imputation method based on FAMD is implemented in the package `missMDA` (Husson and Josse, 2012) of the `R` software. The function `imputeFAMD` takes as input the incomplete data set and the number of dimensions used to reconstruct the data at each step of the algorithm. The function returns a table with the imputed mixed data as well as the table concatenating the imputed continuous variables and the imputed indicator matrix.

As with all methods of imputation, imputation quality deteriorates with increasing percentage of missing data. However, this deterioration depends on the structure of the data set. Indeed, if the variables are uncorrelated, even

a single missing value is problematic. On the other hand, if the variables are highly correlated, very little data per individual are sufficient to impute the data set. For this reason, it is of course not possible to offer a percentage of missing data below which imputation is acceptable and above which the imputation is no longer satisfactory. It would therefore be desirable to provide confidence intervals around the imputed values. We expect narrow confidence intervals when the data are highly correlated, indicating higher confidence in the imputed values.

The proposed method is a method of single imputation. Like any single imputation method, it is limited because it does not take into account the uncertainty associated with the prediction of missing values based upon observed values. Thus, if we apply a statistical method on the completed data table, the variability of the estimators will be underestimated. To avoid this problem, a solution is to perform multiple imputation (Little and Rubin, 1987, 2002). In this case, different values are predicted for each missing value, which leads to several imputed data sets; the variability across the imputations reflects the variance of the prediction of each missing entry. The second step of multiple imputation consists of performing the statistical analysis on each completed data set. The third step combines the results to obtain the estimators of the parameters and of their variability taking into account uncertainty due to missing data. The proposed iterative FAMD imputation algorithm could be a first step in a multiple imputation method for mixed data.

## References

Benzécri JP (1973) L'analyse des données. Tome II: L'analyse des correspondances. Dunod

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Bro R, Kjeldahl K, Smilde AK, Kiers HAL (2008) Cross-validation of component model: a critical look at current methods. Anal Bioanal Chem 390:1241–1251

Cornillon PA, Guyader A, Husson F, Jégou N, Josse J, Kloareg M, Matzner-Løber E, Rouvière L (2012) R for Statistics. Chapman & Hall/CRC Computer Science & Data Analysis, Boca Raton, FL.

de Leeuw J, Mair P (2009) Gifi methods for optimal scaling in R: The package homals. Journal of Statistical Software 31(4):1–20, URL http://www.jstatsoft.org/v31/i04/

Escofier B (1979) Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. Les cahiers de l'analyse des données 4(2):137–146

Gifi A (1990) Nonlinear Multivariate Analysis. Wiley, Chichester, England

Greenacre M, Blasius J (2006) Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC

Husson F, Josse J (2012) missMDA: Handling missing values with/in multivariate data analysis (principal component methods). URL http://www.agrocampus-ouest.fr/math/husson, r package version 1.4

Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 99:1957–2000, URL http://dl.acm.org/citation.cfm?id=1859890.1859917

Josse J, Husson F (2011) Selecting the number of components in PCA using cross-validation approximations. Computational Statististics and Data Analysis 56(6):1869–1879

Josse J, Husson F (2012) Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique 153 (2):1–21

Josse J, Pagès J, Husson F (2009) Gestion des données manquantes en analyse en composantes principales. Journal de la Société Française de Statistique 150:28–51

Josse J, Chavent M, Liquet B, Husson F (2012) Handling missing values with regularized iterative multiple correspondence analysis. Journal of Classification 29:91–116

Kiers HAL (1991) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika 56:197–212

Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 62:251–266

Lafaye de Micheaux P, Drouilhet R, Liquet B (2011) Le logiciel R. Springer, Paris

Lang DT, Swayne D, Wickham H, Lawrence M (2012) rggobi: Interface between R and GGobi. URL http://CRAN.R-project.org/package=rggobi, r package version 2.1.19

Lebart L, Morineau A, Werwick KM (1984) Multivariate Descriptive Statistical Analysis. Wiley, New-York

Little RJA, Rubin DB (1987, 2002) Statistical Analysis with Missing Data. Wiley series in probability and statistics, New-York

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11:2287–2322

Michailidis G, de Leeuw J (1998) The Gifi system of descriptive multivariate analysis. Statistical Science 13(4):307–336

Peters A, Hothorn T (2012) ipred: Improved Predictors. URL http://CRAN.R-project.org/package=ipred, R package version 0.9-1

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Schafer JL (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, London

Stekhoven D, Bühlmann P (2011) Missforest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28:113–118

Tenenhaus M, Young FW (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(62001):520–525

van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16:219–242

van Buuren S, Boshuizen H, Knook D (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18:681–694

van der Heijden P, Escofier B (2003) Multiple correspondence analysis with missing data. In: Analyse des correspondances, Presse universitaire de Rennes, pp 153–170

Vermunt JK, van Ginkel JR, van der Ark LA, Sijtsma K (2008) Multiple imputation of incomplete categorical data using latent class analysis. Sociological Methodology 33:369–397

## 3.7 Compléments : focus sur les données MAR

Les propriétés des méthodes d'analyse factorielle en font des méthodes d'imputation simple performantes. La capacité à imputer les modalités rares est notamment particulièrement intéressante pour imputer des données MAR. En effet, en présence d'un mécanisme MAR, les individus complètement observés ne constituent plus un échantillon représentatif des données. Ainsi, certaines données deviennent rares du fait qu'elles sont particulièrement sujet à la présence de données manquantes. Notons que pour des variables quantitatives, cette rareté se traduit par des intervalles où la densité des données observée est faible. Il a déjà été mis en avant que l'imputation par les méthodes d'analyse factorielle était particulièrement pertinente pour prédire des modalités rares, mais ceci n'a pas été évoqué pour des variables quantitatives.

Pour illustrer les propriétés de l'imputation par analyse factorielle en présence d'un mécanisme MAR, nous reprenons l'étude 3.4 de l'article précédent (Audigier *et al.*, 2014) en générant des données manquantes uniquement sur une seule variable, quantitative ou qualitative, tel que la probabilité d'apparition d'une donnée manquante soit fonction de la première variable $X_1$ comme suit :

$$\mathbb{P}(R = 0) = \frac{exp(\beta_0 + \beta_1 X_1)}{1 + exp(\beta_0 + \beta_1 X_1)}. \tag{6}$$

Ainsi, si $\beta_1 = 0$ la présence de données manquantes est indépendante des données, le mécanisme est MCAR. Aussi, plus $\beta_1$ est grand, plus la probabilité d'observer une donnée manquante est forte si la valeur absolue de $X_1$ est grande et $X_1$ positif, et faible si la valeur absolue de $X_1$ est grande et $X_1$ est négatif. Par conséquent, plus $\beta_1$ est grand, plus le caractère MAR du mécanisme est prononcé.

La Figure 2 représente l'erreur de prédiction moyenne en fonction de la valeur de $\beta_1$ pour une des configurations de l'étude 3.4. Que ce soit pour des données quantitatives ou qualitatives, l'erreur commise à la suite de l'imputation par AFDM est plutôt stable par rapport à la valeur de $\beta_1$ ce qui indique que l'imputation par analyse factorielle est robuste à un mécanisme MAR. En revanche, l'imputation par forêts aléatoires ne l'est pas. En effet, la prédiction par la forêt est le résultat d'une agrégation de données observées exclusivement : les données observées de la variable réponse quantitative (resp. qualitative) sont réparties dans les feuilles des arbres et la prédiction fournie par la forêt est la moyenne (resp. la modalité majoritaire) des valeurs de certaines de ces feuilles. Pour des données quantitatives, cela implique que si les valeurs dans les queues de distribution sont manquantes, alors les feuilles ne contiennent que des valeurs proches de la moyenne, et la prédiction est biaisée. Pour des variables qualitatives, les modalités devenues rares deviennent minoritaires au sein de leur feuille, ou isolées et sont donc situées dans des feuilles instables. Dans les deux cas, ceci favorise la prédiction par une modalité commune.

Ce comportement se vérifie également sur les autres configurations de l'étude 3.4. Ainsi les méthodes d'imputation par analyse factorielle s'adaptent assez bien à des données MAR.

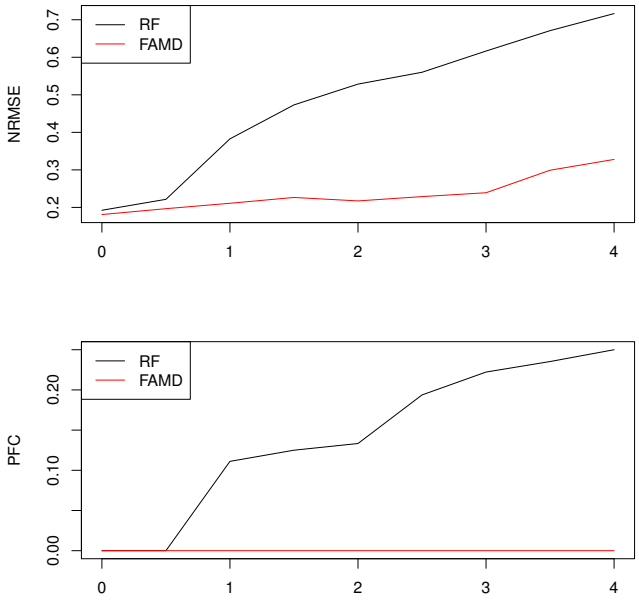


FIGURE 2 – Evolution de l'erreur médiane de prédiction sur 200 simulations pour l'imputation simple par AFDM et forêts aléatoires en fonction du paramètre $\beta_1$ contrôlant le mécanisme. L'erreur pour les variables qualitatives correspond au graphique du haut et celle pour la variable qualitative au graphique du bas. Le jeu de données comporte 50 individus, 12 variables dont 4 quantitatives à 6 modalités, et correspond à un SNR de 4.

CHAPITRE 4

# IMPUTATION MULTIPLE DE DONNÉES QUANTITATIVES

Dans ce chapitre, une méthode d'imputation multiple par ACP pour des données quantitatives est proposée. L'incertitude sur les paramètres du modèle d'imputation est reflétée à l'aide de l'approche Bayésienne du modèle d'ACP proposée par Verbanck *et al.* (2013). Une étude approfondie de la méthode sur jeux réels et simulés illustrent son caractère proper pour des quantités d'intérêt telles que des moyennes, des coefficients de correlation ou des coefficient de régression. Par rapport à l'imputation par modèle Gaussien ou par l'imputation par équations enchaînées, l'imputation par ACP offre une solution à la présence de corrélations importantes ou à un nombre de variables trop grand devant le nombre d'individus.

## Contents



L'imputation multiple de données quantitatives quand les données sont continues est généralement effectuée via un modèle joint selon le modèle Gaussien (Schafer, 1997; King *et al.*, 2001) ou, via les équations enchaînées selon le modèle linéaire Gaussien. Ces deux méthodes nécessitent des inversions de matrices de variance-covariance ce qui les met en défaut quand le nombre d'individus est inférieur au nombre de variables ou que les corrélations entre variables sont fortes. D'autre part, le nombre de paramètres estimés par ces méthodes croît rapidement avec le nombre de variables ce qui amène rapidement à des problèmes de surajustement. L'imputation simple par AFDM présentée au Chapitre 3 a montré que les méthodes d'analyse factorielle gère correctement ce type de configurations.

L'extension de l'imputation simple, non stochastique, à une méthode d'imputation multiple nécessite, d'une part, d'ajouter un aléa sur la prédiction par ACP de façon à mieux respecter la structure des données et, d'autre part, à refléter l'incertitude sur les composantes et vecteurs propres. Une méthode d'imputation multiple par ACP utilisant une approche bootstrap avait déjà été proposée par Josse et Husson (2011) afin de construire des ellipses de confiance en ACP. Ce chapitre propose une méthode d'imputation multiple utilisant la modélisation Bayésienne de l'ACP proposée par Verbanck *et al.* (2013). Il correspond à l'article Audigier *et al.* (2015b).

# Multiple imputation for continuous variables using a Bayesian principal component analysis

Vincent Audigier[1], François Husson[2] and Julie Josse[2]

Applied Mathematics Department, Agrocampus Ouest, 65 rue de Saint-Brieuc, F-35042 RENNES Cedex, France
audigier@agrocampus-ouest.fr
husson@agrocampus-ouest.fr
josse@agrocampus-ouest.fr

**Abstract**

We propose a multiple imputation method based on principal component analysis (PCA) to deal with incomplete continuous data. To reflect the uncertainty of the parameters from one imputation to the next, we use a Bayesian treatment of the PCA model. Using a simulation study and real data sets, the method is compared to two classical approaches: multiple imputation based on joint modelling and on fully conditional modelling. Contrary to the others, the proposed method can be easily used on data sets where the number of individuals is less than the number of variables and when the variables are highly correlated. In addition, it provides unbiased point estimates of quantities of interest, such as an expectation, a regression coefficient or a correlation coefficient, with a smaller mean squared error. Furthermore, the widths of the confidence intervals built for the quantities of interest are often smaller whilst ensuring a valid coverage.

*Keywords*: missing values, continuous data, multiple imputation, Bayesian principal component analysis, data augmentation

## 1 Introduction

Data with continuous variables are ubiquitous in many fields. For instance in biology, samples are described by the expression of the genes, in chemometrics, components can be described by physico-chemical measurements, in ecology, plants are characterized by traits, etc. Whatever the field, missing values occur frequently and are a key problem in statistical practice. Indeed most statistical methods cannot be applied directly on an incomplete data set. To deal with this issue, one of the common approaches is to perform single imputation. This consists in imputing missing values by plausible values. It leads to a complete data set that can be analysed by any standard statistical method.

However, single imputation is limited because it does not take into account the uncertainty associated with the prediction of missing values based on observed values. Thus, if we apply a statistical method on the completed data

[1]Principal corresponding author
[2]Corresponding author

table, the variability of the estimators will be underestimated. To avoid this problem, a first solution is to adapt the procedure to be applied on an incomplete data set. To do this, an Expectation-Maximization (EM) algorithm [1] combined, for instance, with a Supplemented Expectation-Maximization algorithm [2] could be used to get the maximum likelihood estimates as well as their variance from incomplete data. Note that, the maximum likelihood estimate using these algorithms obviates the necessity for imputation. However it is not always easy to establish these algorithms. Another solution is to perform multiple imputation [3, 4] which consists in predicting different values for each missing value, which leads to several imputed data sets. The variability across the imputations reflects the variance of the prediction of each missing entry. Then, multiple imputation consists in performing the statistical analysis on each completed data set. Finally, the results are combined using Rubin's rules [3] to obtain an estimate of parameters and an estimate of their variability taking into account uncertainty due to missing data.

Therefore, a multiple imputation method is based on a single imputation method. Denoting $\theta$ the parameters of the imputation model, a multiple imputation method requires generating a set of $M$ parameters $(\widehat{\theta}_1, \ldots, \widehat{\theta}_M)$ to reflect the uncertainty in the estimate of the model's parameters. Multiple imputation methods are distinguished in the way the uncertainty is spread using either a bootstrap or a Bayesian approach. The bootstrap approach consists in producing $M$ new incomplete data sets and estimating $\theta$ on each bootstrap replication. The Bayesian approach consists of determining a posterior distribution for the model's parameters using a prior distribution and the observed entries. Then the set of parameters $(\widehat{\theta}_1, \ldots, \widehat{\theta}_M)$ is drawn from the posterior distribution. There are also two classical ways of performing multiple imputation. The first one is to use an explicit joint model to all variables [5]. A normal distribution is often assumed on variables which may seem restrictive but is known to be fairly robust with respect to the assumption of normality [5, p.211-218]. The second way to perform multiple imputation is to use chained equations [6]: a model is defined for each variable with missing data and variables are successively imputed using these models. Typically, imputation is done using the regression model or by predictive mean matching. The chained equations approach is more flexible than the joint modelling, however it requires specifying a model for each variable with missing values, which is quite tedious with a lot of incomplete variables. In addition it may not converge to a stationary distribution if the separate models are not compatible [7], that is to say that there is no joint distribution for variables with the conditional distributions chosen. More generally, the theoretical properties of chained equations are not well understood and they are a current topic of research [8]. Both the joint and conditional methods have their own advantages and drawbacks as investigated recently in [9]. However, both approaches share the drawback that regression models are rapidly ineffective for data sets where the number of individuals is too low compared to the number of variables or when the variables are highly correlated. Even if some solutions using regularization are available to handle such situations, it is not straightforward to deal with such cases.

Recently, [10] proposed a method of single imputation based on a PCA model. This method gives good results in terms of quality of the imputation when there are linear relationships between variables and also has the advantage of being able to be performed on a data set where the number of individuals is smaller than the number of variables.

We propose to extend it to multiple imputation and we spread the uncertainty of parameters of the PCA imputation model using a Bayesian approach. In Section 2, we describe the procedure called BayesMIPCA for multiple imputation based on a Bayesian treatment of the PCA model. Then, in Section 3, we present a simulation study in which we compare this method to other multiple imputation methods and demonstrate that multiple imputation by the BayesMIPCA method produces little bias and valid confidence intervals under a variety of conditions. Finally, we apply the methods on real data sets.

# 2 Method

## 2.1 PCA model

PCA can be expressed using a fixed effect model [11] where the data matrix $\mathbf{X}_{n\times p}$ can be decomposed as a signal, denoted $\tilde{\mathbf{X}}_{n\times p}$, of low rank $S$ considered as known, plus noise denoted $\mathbf{E}_{n\times p}$:

$$\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \mathbf{E}_{n\times p} \tag{1}$$

where $\mathbf{E} = (\varepsilon_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The parameters of this model are the elements of $\tilde{\mathbf{X}}$ and $\sigma$.

Imputation under the PCA model requires estimating these parameters from the incomplete data set. The method which achieves this is closely related to the one applied on a complete data set.

### 2.1.1 PCA on complete data

PCA consists in finding the matrix $\hat{\mathbf{X}}$ with rank $S$ which minimizes the least squares criterion $\parallel \hat{\mathbf{X}} - \mathbf{X} \parallel^2$ with $||\cdot||$ the Frobenius norm. Therefore, $\hat{\mathbf{X}}$ corresponds to the least squares estimator of $\tilde{\mathbf{X}}$. The solution is obtained using the singular value decomposition (SVD) of the matrix $\mathbf{X}$: $\widehat{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ where columns of $\mathbf{U}_{n\times S}$ are the left singular vectors, $\mathbf{\Lambda}_{S\times S} = diag(\lambda_1, \ldots, \lambda_S)$ is the matrix of the singular values of $\mathbf{X}$ and columns of $\mathbf{V}_{p\times S}$ are the right singular vectors. The principal components are given by $\mathbf{U}\mathbf{\Lambda}$ and the loadings are given by $\mathbf{V}$. This solution also corresponds to the maximum likelihood estimate of model (1). The expression of the general term of $\hat{\mathbf{X}}$ is given by

$$\hat{x}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}. \tag{2}$$

Then $\sigma^2$ is estimated by

$$\hat{\sigma}^2 = \frac{\sum_{ij} (x_{ij} - \hat{x}_{ij})^2}{np - (p + S(n - 1 + p - S))} \tag{3}$$

which corresponds to dividing the sum of the squared residuals by the number of entries minus the number of independent model parameters [12].

The classical PCA estimator (2), while providing the best low rank approximation of the data matrix, does not ensure the best recovery of the underlying signal $\tilde{\mathbf{X}}$. Thus, other estimators, obtained from regularized versions of PCA, have been suggested in the literature [13–15]. The rationale is exactly the same as in ordinary regression analysis where the maximum likelihood estimates are not necessarily the best ones in terms of mean squared error (MSE), whereas regularized estimators, although more biased have less variability, which lead to a smaller MSE. By redefining the problem as finding the best approximation of the unknown signal $\tilde{\mathbf{X}}$ in terms of MSE, instead of finding the best low rank approximation of the data matrix $\mathbf{X}$, [14] suggested a ridge version of the PCA estimator. We focus on this estimator, since, as we will see later in Section 2.1.3, it has a straightforward Bayesian interpretation. This better estimator of $\tilde{\mathbf{X}}$ in the sense of the mean squared error criterion is defined as follows. Denoting

$$\hat{x}_{ij}^{(s)} = \sqrt{\lambda_s} u_{is} v_{js}$$

the $s^{th}$ term of the sum (2), this better estimator is determined by searching $(\phi_s)_{1 \leq s \leq S}$ in order to minimize

$$\mathbb{E}\left[\sum_{i,j}\left(\left(\sum_{s=1}^{S}\phi_s \hat{x}_{ij}^{(s)}\right) - \tilde{x}_{ij}\right)^2\right].$$

Note that a parallel with regression analysis and ridge regression can be drawn. [14] showed that $\phi_s$ is given by

$$\phi_s = \frac{\sum_{i,j} \mathbb{E}\left[\hat{x}_{ij}^{(s)}\right] \tilde{x}_{ij}}{\sum_{i,j}\left(\mathbb{V}\left[\hat{x}_{ij}^{(s)}\right] + \mathbb{E}\left[\hat{x}_{ij}^{(s)}\right]^2\right)}.$$

In the asymptotic framework where $\sigma^2$ tends to 0, [16] showed that the expectation of $\hat{x}_{ij}^{(s)}$ is equal to $\tilde{x}_{ij}^{(s)}$. [14] approximated the variance of $\hat{x}_{ij}^{(s)}$ by the noise variance $\frac{1}{\min(n-1, p)}\sigma^2$. Using these assumptions, [14] showed that the shrinkage terms can be written as the ratio between the variance of the signal and the total variance for the $s$ dimension, that they estimated using a plug-in estimator:

$$\hat{\phi}_s = \frac{\lambda_s - \frac{np}{\min(n-1, p)}\hat{\sigma}^2}{\lambda_s} \text{ for all } s \text{ from 1 to } S. \tag{4}$$

Although the theoretical properties of this estimator have not been exhibited, the simulation study conducted indicates that retaining this estimate for the shrinkage terms substantially reduces the mean squared error.

Thus, the regularized PCA solution $\hat{\mathbf{X}}^{rPCA}$ is defined by [14] as follows:

$$\hat{x}_{ij}^{rPCA} = \sum_{s=1}^{S} \hat{\phi}_s \sqrt{\lambda_s} u_{is} v_{js}. \tag{5}$$

### 2.1.2 PCA on incomplete data

With missing values, the classical solution to perform PCA is determined by minimizing the criterion $\| \hat{\mathbf{X}} - \mathbf{X} \|^2$ on the observed data only. This is equivalent to introducing a weight matrix $\mathbf{W}$, where $w_{ij} = 0$ if $x_{ij}$ is missing and $w_{ij} = 1$ otherwise, in the criterion which becomes $\| \mathbf{W} * \left( \hat{\mathbf{X}} - \mathbf{X} \right) \|^2$ where $*$ is the Hadamard product. To minimize this criterion, it is possible to use an EM algorithm called iterative PCA [17]. The algorithm essentially sets the missing elements at initial values, performs the PCA on the completed data set, imputes the missing values with values predicted by the model (2) using a predefined number of dimensions ($S$), and repeats the procedure on the newly obtained matrix until the total change in the matrix falls below an empirically determined threshold. However such algorithms which alternate a step of estimation of the parameters using a singular value decomposition and a step of imputation of the missing values are known to suffer from overfitting problems. This means that the observed values are well fitted but the quality of prediction is poor. This occurs especially when the relationships between variables are low and/or when the number of missing values is high. To avoid these problems of overfitting, [10] proposed to alternate the imputation and estimation steps by regularized PCA (5). The new algorithm is then called regularized iterative PCA.

Thus, the regularized iterative PCA algorithm can be used as a single imputation method since it produces a completed data set from the incomplete one. As stated in the introduction, performing multiple imputation requires taking into account the uncertainty of the estimation of the imputation model's parameters. In this aim, we suggest a Bayesian approach to get $M$ matrices $(\hat{\mathbf{X}}_m)_{1 \leq m \leq M}$ which will be obtained using draws from the posterior distribution of $\tilde{\mathbf{X}}$. Before describing the Bayesian approach on a data set with missing values, we present it on a complete data set.

### 2.1.3 Bayesian PCA on complete data

[14] proposed a Bayesian treatment of the PCA model using the following prior distribution for $\tilde{x}_{ij}^{(s)}$:

$$\tilde{x}_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2) \quad \text{for all } 1 \leq s \leq S.$$

Combining this prior distribution with the PCA model (1), the posterior distribution has an explicit form: it is a normal distribution whose parameters

depend on $\tau_s$ and $\sigma$:

$$p\left(\tilde{x}_{ij}^{(s)}|x_{ij}^{(s)}\right) = \mathcal{N}\left(\frac{\tau_s^2}{\tau_s^2 + \frac{1}{\min(n-1,\,p)}\sigma^2}x_{ij}^{(s)}, \frac{\tau_s^2 \frac{\sigma^2}{\min(n-1,\,p)}}{\tau_s^2 + \frac{\sigma^2}{\min(n-1,\,p)}}\right).$$

Using an empirical Bayesian approach, $\tau_s$ and $\sigma$ are fixed from their estimates from the data as:

$$\hat{\tau}_s^2 \quad = \quad \frac{1}{np}\lambda_s - \frac{\hat{\sigma}^2}{\min(n-1,\,p)}$$

and $\hat{\sigma}^2$ defined in (3). Thus, [14] showed that the posterior distribution of $\tilde{x}_{ij}^{(s)}$ is a normal distribution which has for expectation $\hat{x}_{ij}^{(s)\,rPCA}$ (5) and for variance $\frac{\hat{\sigma}^2\hat{\phi}_s}{\min(n-1,\,p)}$ where $\phi_s$ given by $\frac{\tau_s^2}{\tau_s^2+\frac{\sigma^2}{\min(n-1,\,p)}}$ is estimated by plug-in which corresponds to the estimate given in (4).

Note that this modelling is in line with the one of [18] for a matrix $\tilde{\mathbf{X}}$ of full rank, and can be seen as a truncated version.

### 2.1.4 Bayesian PCA on incomplete data

Generally, when a data set contains missing values, the posterior distribution of model parameters is often intractable. An algorithm which can be used in this context is the data augmentation (DA) algorithm [19]. It consists in 'augmenting' the observed data by predictions on missing data. The posterior becomes easier to calculate because the data set has become complete. DA simulates alternatively imputed values and parameters using a Markov chain which converges in probability to the observed posterior distribution. The algorithm consists of two steps:

(I) imputing from the current parameters and the observed data,

(P) drawing of new parameters from the posterior given the new imputation and a prior distribution on the model's parameters.

Steps (I) and (P) are repeated a predefined number of times. At the end of the algorithm draws from the posterior distribution are obtained from an incomplete data set.

Inspired by the data augmentation algorithm to perform draws of $\tilde{x}$ in its posterior distribution, we essentially perform the two following steps:

(I) given $\tilde{\mathbf{X}}$ and $\hat{\sigma}^2$, imputing the missing values $x_{ij}$ by a draw from the predictive distribution $\mathcal{N}\left(\tilde{x}_{ij}, \hat{\sigma}^2\right)$

(P) drawing $\tilde{x}_{ij}$ from its posterior distribution $\mathcal{N}\left(\hat{x}_{ij}^{rPCA}, \frac{\hat{\sigma}^2\sum_s \hat{\phi}_s}{\min(n-1,\,p)}\right)$ where $\hat{x}_{ij}^{rPCA}, \hat{\sigma}^2$ and $(\hat{\phi}_s)_{1\leq s\leq S}$ are calculated from the completed data set obtained from step (I).

Note that the estimates of $\phi$ and $\sigma$, that appear in the posterior distributions of $\tilde{\mathbf{X}}$, are updated by their maximum likelihood estimates in step (P), and are not fixed. Thus, it can be viewed as a marriage between a DA algorithm and an EM algorithm with unknown convergence properties.

## 2.2 Multiple imputation with the BayesMIPCA algorithm

### 2.2.1 Presentation of the algorithm

In addition providing a posterior distribution of the parameters from an incomplete data set, the data augmentation algorithm can also be straightforwardly used to get multiple imputed data sets. To do so, after a burn-in step, we simply keep $M$ approximately independent draws leading to $M$ imputed data sets. Thus, an imputed data set is saved at regular intervals.

This procedure of multiple imputation with Bayesian PCA is thus called the BayesMIPCA method. The details of the algorithm are as follows:

1. Initialization:
   - calculate the matrix of means $\mathbf{M}^{[0]}$ which is the matrix of size $n \times p$ with each row being the vector of the means of each column of the incomplete data set $\mathbf{X}$. The means are computed on the observed values.
   - centre $\mathbf{X}$: $\mathbf{X}^{[0]} \leftarrow \mathbf{X} - \mathbf{M}^{[0]}$. Since $\mathbf{X}$ is incomplete, $\mathbf{X}^{[0]}$ is also incomplete.
   - estimate the initial parameters $\tilde{\mathbf{X}}^{[0]}, \sigma^{2\,[0]}$ using, for instance, the regularized iterative PCA algorithm on $\mathbf{X}^{[0]}$
2. Burn in: for $\ell$ from 1 to `Lstart`
   (I)
   - perform a random imputation according to the current parameters (drawn from the predictive distribution): $\mathbf{X}^{[\ell]} \leftarrow \mathbf{W} * \mathbf{X}^{[\ell-1]} + (\mathbf{1} - \mathbf{W}) * (\tilde{\mathbf{X}}^{[\ell-\mathbf{1}]} + \mathbf{E})$ where $\mathbf{1}_{I\times J}$ being a matrix with only ones and $\mathbf{E}_{n\times p} = (\varepsilon_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ being a matrix of independent residuals so that $\varepsilon_{ij} \sim \mathcal{N}(0, \hat{\sigma}^{2[\ell-1]})$; therefore $\mathbf{X}^{[\ell]}$ contains no missing values
   - add the matrix of means $\mathbf{X}^{[\ell]} \leftarrow \mathbf{X}^{[\ell]} + \mathbf{M}^{[\ell-1]}$

   (P)
   - calculate $\mathbf{M}^{[\ell]}$, the matrix of means of $\mathbf{X}^{[\ell]}$
   - centre the imputed data $\mathbf{X}^{[\ell]} \leftarrow \mathbf{X}^{[\ell]} - \mathbf{M}^{[\ell]}$
   - evaluate posterior parameters: calculate $\hat{\mathbf{X}}^{[\ell]}$, $\hat{\sigma}^{2\,[\ell]}$ and $\hat{\phi}^{[\ell]}$ from which we can deduce $\hat{\mathbf{X}}^{rPCA[\ell]}$
   - draw new parameters from the posterior: draw $\tilde{x}_{ij}^{[\ell]}$ from $\mathcal{N}\left(\hat{x}_{ij}^{rPCA[\ell]}, \frac{\hat{\sigma}^{2[\ell]} \sum_s \hat{\phi}_s^{[\ell]}}{\min(n-1, p)}\right)$.
3. Create $M$ imputed data sets: for $m$ from 1 to $M$ alternate steps (I) and (P) `L` times. `L` is fixed and should be large enough to obtain independent imputations from one data set to another.

### 2.2.2 Modelling and analysis considerations

The parameter $S$ is supposed to be known *a priori*. Many strategies are available in the literature to select a number of dimensions from a complete data set in PCA [20]. Cross-validation [21] or an approximation of cross-validation such as generalized cross-validation [22] perform well. We suggest these approaches since they can be directly extended to incomplete data [10].

A simple chain is used to perform multiple imputation by data augmentation: `Lstart` iterations are passed in order to forget the dependence between the current settings and the initial parameters. `Lstart` is equal to 1000 in our case. The $M$ imputed data sets are obtained after `Lstart+L`, `Lstart+2*L`, `Lstart+3*L`,..., `Lstart+`$M$`*L` iterations with `L` equal to 100.

Assessing the convergence of this kind of algorithm is still an open area of research. In practice, we investigate the values of some summaries, as sample moments or quantiles, through several iterations of the algorithm [5]. The number of iterations required to observe stationarity for the summaries defines `Lstart`, the number of iterations for the burn in step. Then, the autocorrelation of the summaries is investigated to determine a minimum value for `L`.

Concerning the choice of $M$, generating three to five data sets is usually enough in multiple imputation [3]. However, due to increasing computational power, it is possible to generate a greater number of imputed data sets [23, p.49]. We use $M = 20$.

## 2.3 Combining results from multiple imputed data sets

As mentioned in the introduction, the aim of a multiple imputation procedure is to estimate a parameter and its variance from incomplete data. We detail hereafter the methodology described in [3, 24] to combine the results from multiple imputed data sets under the assumption of an estimator normally distributed and evaluated on a large sample. Note that this methodology is the same whatever the multiple imputation method used. Let $\psi$ denote a quantity of interest that we want to estimate from an incomplete data set. To estimate this quantity and a confidence interval from $M$ imputed data sets obtained from a multiple imputation method, the following steps are performed:

- for $m = 1, ...M$, $\hat{\psi}_m$ is computed on the imputed data set $m$ as well as its variance $\widehat{Var}(\hat{\psi}_m)$;
- the results are pooled as:

$$\begin{aligned} \hat{\psi} &= \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_m, \\ \widehat{Var}(\hat{\psi}) &= \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\psi}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\psi}_m - \hat{\psi}\right)^2. \end{aligned}$$

  The estimate of the variability of $\hat{\psi}$ is composed of two terms: the within-imputation variance corresponding to the sampling variability and the

between-imputation variance corresponding to the variability due to missing values. The factor $(1 + \frac{1}{M})$ corrects the fact that $\widehat{\psi}$ is an estimate for a finite number of imputed tables;

- the 95% confidence interval is calculated as:

$$\hat{\psi} \pm t_{\nu,.975} \sqrt{\widehat{Var}(\hat{\psi})}$$

where $t_{\nu,.975}$ is the quantile corresponding to probability .975 of the Student's $t-$distribution with $\nu$ degrees of freedom estimated as suggested by [25].

# 3 Evaluation of the methodology

To assess the multiple imputation method based on PCA, we conducted an extensive simulation study. We generated data sets drawn from normal distributions. These data sets differ with respect to the number of variables, the number of individuals and the strength of relationships between variables. We also considered real data sets. The code to reproduce all the simulations with the R software [26] is available on the webpage of the first author.

## 3.1 Competing algorithms

The BayesMIPCA method is compared to the two following multiple imputation methods: a first one based on joint modelling implemented in the R-package Amelia [27, 28] and a second one based on chained equations implemented in the R-package mice [29, 30].

- **Amelia** imputes missing values by assuming a multivariate normal distribution for the variables. The uncertainty on the parameters is spread using a bootstrap approach [31]. More precisely, $M$ bootstrap incomplete data sets are generated and on each incomplete data set, the covariance matrix is estimated using an expectation-maximization algorithm. Then, the $M$ covariance matrices are used to produce $M$ imputed data sets. The algorithm is implemented in the function `amelia`. In the presence of high collinearity between variables, or a number of individuals too low compared to the number of variables, the variance-covariance matrix is not computationally invertible and therefore imputation under the normal distribution is not possible. In order to perform imputation in such conditions, it would be possible to introduce a ridge term to improve the conditioning of the regression problem.

- **Mice (BayesMI method)** requires specifying a model for each variable with missing data. The BayesMI method provides an imputation by regression for continuous variables where uncertainty on regression parameters is spread using a Bayesian approach. This method is implemented

in the function `mice.impute.norm` in the mice package. In the same way as the Amelia package, a ridge term could be introduced to overcome collinearity problems or lack of observations. It is also possible to specify a conditional model where only a subset of variables is used as explanatory variables in each regression model.

- **Listwise deletion** deletes individuals with missing values. This is not a multiple imputation method, but it is a benchmark for the variability of estimates. Because listwise deletion is equivalent to performing a statistical method on a sub-sample, variability should be greater than for a multiple imputation method.

## 3.2 Simulation study with a block diagonal structure for the covariance matrix

### 3.2.1 Simulation design

A data set $\mathbf{X}$ with $n$ rows and $p$ columns is drawn from a normal distribution with null expectation and variance-covariance matrix of the form:

$$\begin{pmatrix} 1 & \rho & \dots & \rho & \rho & & & \\ \rho & 1 & \dots & \rho & \rho & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & & 0 & \\ \rho & \rho & \dots & 1 & \rho & & & \\ \rho & \rho & \dots & \rho & 1 & & & \\ & & & & & 1 & \dots & \rho \\ & & 0 & & & \vdots & \ddots & \vdots \\ & & & & & \rho & \dots & 1 \end{pmatrix}$$

with $0 < \rho < 1$. The variables are divided into two groups of size 2/3 and 1/3. Within each group, the pairwise correlation between variables is equal to $\rho$ and the two groups of correlated variables are independent. Thus, the number of underlying dimensions $S$ is equal to 2. The coefficient $\rho$ takes the values 0.9 or 0.3 to obtain strong or weak relationships between variables. The number of variables is $p = 6$ or $p = 60$ and the number of individuals $n = 30$ or $n = 200$. Then, we insert missing values (10% or 30%) completely at random, meaning that the probability that a value is missing is unrelated to the value itself and any values in the data set, missing or observed. Each simulation is repeated $K = 1000$ times.

Note that this simulation design is also suited for the competing algorithms, which are dedicated to normal data: the one in the Amelia package assumes multivariate normal distribution and the one in the mice package assumes a regression model for each variable.

### 3.2.2 Criteria

We consider three quantities of interest $\psi$ to be estimated from incomplete data: the expectation of a variable $\mathbb{E}[X_1]$, the correlation coefficient $\rho(X_{p-1}, X_p)$ between two variables and the regression coefficient $\beta_{X_2}$, which corresponds to the coefficient of the first explanatory variable in the regression model where $X_1$ is the response and $(X_2, \ldots, X_p)$ the explanatory variables. The first quantity of interest is an indicator on a distribution of one variable and others on the relationships between variables.

The criteria of interest are the bias $\frac{1}{K}\sum_{k=1}^{K}\hat{\psi}_k - \psi$, the root mean squared error (RMSE) $\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\psi}_k - \psi)^2}$, the median (over the $K$ simulations) of the confidence intervals width as well as the 95% coverage. This latter is calculated as the percentage of cases where the true value $\psi$ is within the 95% confidence interval. "The 95% coverage should be 95% or higher. Coverages below 90% are considered undesirable" [23, p.47].

As a benchmark, we also calculated the confidence intervals for the data sets without missing values which we call "Full data". The confidence interval obtained by multiple imputation should be greater.

**Remark.** Confidence intervals are based on the assumption that $\hat{\psi}$ is normally distributed. This is not true for the correlation coefficient $\rho$. Therefore a Fisher $z$ transformation is needed [5]:

$$z(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

### 3.2.3 Results

For the point estimate of the expectation of a variable ($\psi = \mathbb{E}[X_1]$), all methods give good results: they produce unbiased estimates (results not shown here). In addition, the root mean squared errors are of the same order of magnitude. Thus, the simulations do not highlight differences between the methods in terms of point estimate. Concerning the estimate of the variability of the estimator, Table 1 gives the median of the confidence intervals width and the 95% coverage over the 1000 simulations for different simulations' configurations. In addition, when an algorithm fails on a configuration, no result is given. With the current version of Amelia [27], it is impossible to get results for the cases where $n < p$ for our simulations. These problems may be a pitfall of the implementation of the method since in theory using regularization may be able to handle such situations. Nevertheless, it would still be difficult to run the simulations since only expertise allows the selection of the tuning parameter in a missing data framework. For these reasons no results are provided for cases 5, 6, 7, 8. In addition, the algorithm regularly fails when there are many missing values. This problem is exacerbated when the number of variables is high or when the number of individuals is low (cases 2, 4, 13, 14, 15, 16). Since the imputation by chained equations using the BayesMI method requires estimating the parameters of a regression model for each variable to be imputed, it suffers from the same kind

Table 1: Results for the mean. Median confidence intervals width and 95% coverage for $\psi = \mathbb{E}[X_1]$ estimated by several methods (Listwise deletion, Amelia, BayesMI and BayesMIPCA) for different configurations varying the number of individuals ($n = 30$ or 200), the number of variables ($p = 6$ or 60), the strength of the relationships between variables ($\rho = 0.3$ or 0.9) and the percentage of missing values (10% or 30%). For each configuration, 1000 data sets with missing values are generated. Some values are not available because of failures of the algorithms.

| | parameters | | | | confidence interval width | | | | coverage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | $\rho$ | % | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* |
| 1 | 30 | 6 | 0.3 | 0.1 | 1.034 | 0.803 | 0.805 | 0.781 | 0.936 | 0.955 | 0.953 | 0.950 |
| 2 | 30 | 6 | 0.3 | 0.3 | | | 1.010 | 0.898 | | | 0.971 | 0.949 |
| 3 | 30 | 6 | 0.9 | 0.1 | 1.048 | 0.763 | 0.759 | 0.756 | 0.951 | 0.952 | 0.95 | 0.949 |
| 4 | 30 | 6 | 0.9 | 0.3 | | | 0.818 | 0.783 | | | 0.965 | 0.953 |
| 5 | 30 | 60 | 0.3 | 0.1 | | | | 0.775 | | | | 0.955 |
| 6 | 30 | 60 | 0.3 | 0.3 | | | | 0.864 | | | | 0.952 |
| 7 | 30 | 60 | 0.9 | 0.1 | | | | 0.742 | | | | 0.953 |
| 8 | 30 | 60 | 0.9 | 0.3 | | | | 0.759 | | | | 0.954 |
| 9 | 200 | 6 | 0.3 | 0.1 | 0.383 | 0.291 | 0.294 | 0.292 | 0.938 | 0.947 | 0.947 | 0.946 |
| 10 | 200 | 6 | 0.3 | 0.3 | 0.864 | 0.328 | 0.334 | 0.325 | 0.942 | 0.954 | 0.959 | 0.952 |
| 11 | 200 | 6 | 0.9 | 0.1 | 0.385 | 0.281 | 0.281 | 0.281 | 0.945 | 0.953 | 0.95 | 0.952 |
| 12 | 200 | 6 | 0.9 | 0.3 | 0.862 | 0.288 | 0.289 | 0.288 | 0.942 | 0.948 | 0.951 | 0.951 |
| 13 | 200 | 60 | 0.3 | 0.1 | | | 0.304 | 0.289 | | | 0.957 | 0.945 |
| 14 | 200 | 60 | 0.3 | 0.3 | | | 0.384 | 0.313 | | | 0.981 | 0.958 |
| 15 | 200 | 60 | 0.9 | 0.1 | | | 0.282 | 0.279 | | | 0.951 | 0.948 |
| 16 | 200 | 60 | 0.9 | 0.3 | | | 0.296 | 0.283 | | | 0.958 | 0.952 |

of problems as the Amelia's algorithm. The solution to this problem consists in selecting a subset of explanatory variables for each conditional model. But it is difficult to make an appropriate selection of the predictors and there is no fully automatic default solution for the BayesMI method. For this reason no output is provided in the case where $n < p$. Finally, the listwise deletion cannot be performed on data sets where the rate of missing data is too high compared to the number of entries. On the contrary, multiple imputation using the BayesMIPCA method can be applied on data sets of various kinds: when the collinearity between variables is weak or strong, when the rate of missing data is large or small, the number of individuals less than or greater than the number of variables.

All the algorithms give valid coverage, close to 95% in all conditions where they perform. As expected, the confidence intervals for the multiple imputation methods are larger than those obtained from a complete dataset (0.734 for $n = 30$ and 0.278 for $n = 200$) and smaller than those obtained by listwise deletion. However the width of the confidence interval is often shorter for the BayesMIPCA method than for the other multiple imputation algorithms

(particularly on the cases 1, 2, 4, 13, 14, 16).

Concerning the correlation coefficient, as for the expectation, the main differences between the algorithms are highlighted using the criteria that assess the variability of the estimator. Results are gathered in Table 2. Note that according to the true value of $\rho$, the width of the confidence interval is not the same, because $\rho$ lies in the interval $[-1, 1]$. If $\rho = 0.9$, then $\rho$ is close to a bound and the interval is necessarily shorter than if $\rho = 0.3$. For this reason, the widths of the confidence intervals have to be compared to those obtained from a complete data set. Thus, the median width of the confidence intervals obtained from a complete data set is considered as the reference and the increase from this width is given in Table 2. The BayesMI and Amelia methods produce confidence in-

Table 2: Results for the correlation coefficient. Increase of the median of the widths of the confidence intervals obtained by the imputation method and the one obtained by full data as well as 95% coverage for $\psi = \rho(X_{p-1}, X_p)$. Results are given for several methods (Listwise deletion, Amelia, BayesMI and BayesMIPCA) on different configurations varying the number of individuals ($n = 30$ or 200), the number of variables ($p = 6$ or 60), the strength of the relationships between variables ($\rho = 0.3$ or 0.9) and the percentage of missing values (10% or 30%). For each set of parameters, 1000 data sets with missing values are generated. Some values are not available because of failures of the algorithms.

| | parameters | | | | confidence interval width | | | | coverage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | $\rho$ | % | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* |
| 1 | 30 | 6 | 0.3 | 0.1 | +36% | +16% | +17% | +14% | 0.938 | 0.957 | 0.964 | 0.963 |
| 2 | 30 | 6 | 0.3 | 0.3 | | | +56% | +36% | | | 0.976 | 0.956 |
| 3 | 30 | 6 | 0.9 | 0.1 | +49% | +32% | +31% | +14% | 0.935 | 0.968 | 0.962 | 0.968 |
| 4 | 30 | 6 | 0.9 | 0.3 | | | +221% | +40% | | | 0.974 | 0.983 |
| 5 | 30 | 60 | 0.3 | 0.1 | | | | +13% | | | | 0.971 |
| 6 | 30 | 60 | 0.3 | 0.3 | | | | +27% | | | | 0.989 |
| 7 | 30 | 60 | 0.9 | 0.1 | | | | +13% | | | | 0.976 |
| 8 | 30 | 60 | 0.9 | 0.3 | | | | +26% | | | | 0.99 |
| 9 | 200 | 6 | 0.3 | 0.1 | +38% | +11% | +12% | +10% | 0.959 | 0.947 | 0.952 | 0.967 |
| 10 | 200 | 6 | 0.3 | 0.3 | +202% | +45% | +47% | +27% | 0.939 | 0.942 | 0.949 | 0.974 |
| 11 | 200 | 6 | 0.9 | 0.1 | +40% | +8% | +9% | +6% | 0.958 | 0.953 | 0.956 | 0.967 |
| 12 | 200 | 6 | 0.9 | 0.3 | +247% | +30% | +43% | +23% | 0.940 | 0.948 | 0.943 | 0.973 |
| 13 | 200 | 60 | 0.3 | 0.1 | | | +15% | +8% | | | 0.964 | 0.981 |
| 14 | 200 | 60 | 0.3 | 0.3 | | | +55% | +21% | | | 0.945 | 0.989 |
| 15 | 200 | 60 | 0.9 | 0.1 | | | +23% | +6% | | | 0.914 | 0.969 |
| 16 | 200 | 60 | 0.9 | 0.3 | | | +83% | +13% | | | 0.683 | 0.985 |

tervals of similar widths while they are shorter with the BayesMIPCA method which moreover has a better coverage. This good behaviour of the BayesMIPCA method can be explained by the properties of the imputation model. Indeed, PCA is a dimensionality reduction method used to isolate the relevant information of a data set. This makes it very stable and implies that the imputation

from a table to another does not change much: the between-variability is lower than for the other methods, which explains that the confidence intervals are shorter. When the strength of the relationships between variables is low (cases 2, 10, 14), the difference between the width of the confidence intervals obtained from the BayesMIPCA method and the width of those obtained from the two other methods is moderate. At the most the increase between the median of the widths of the confidence intervals and the median of the widths obtained from a complete data set attempts +55% for the BayesMI method versus +21% for the BayesMIPCA one. However, when the relationships between variables are strong (cases 4, 12, 16), the BayesMI and Amelia algorithms encounter great difficulties. The width of the confidence interval obtained with BayesMI is up to 3 times larger than the one obtained from a complete set (case 4) versus 1.4 for the BayesMIPCA method. For the $16^{th}$ case, it even leads to very bad results with a coverage close to 68%.

The results on the estimate of the regression coefficient lead to the same conclusions as those already mentioned for the expectation and for the correlation coefficient: with BayesMIPCA, confidence intervals are shorter and coverages are accurate. In addition, the BayesMIPCA method systematically gives the smallest mean squared error. The results for this quantity are presented in the appendix.

## 3.3 Simulation study with a fuzzy principal component structure

As a complement to the previous simulations in Section 3.2, we assess the BayesMIPCA algorithm when the low dimensional structure of the data is less obvious. Instead of generating the data sets using covariance matrices with a two block diagonal structure, we generate covariance matrices at random as in [32]. More precisely, the draw is uniform over the space of positive definite correlation matrices. The method is implemented in the R package clusterGeneration [33]. We generated two covariance matrices, one for $p = 6$ variables and another one for $p = 60$ variables. From each matrix, $K = 1000$ data sets are drawn varying the number of individuals ($n = 30$ or $n = 200$) and the percentage of missing values (10% or 30%). Multiple imputation (using $M = 20$ imputed data sets) is performed on each of them to estimate the quantities of interest (an expectation, a regression coefficient and a correlation coefficient). The quality of the imputation is assessed using the same quantities of interest and the same criteria as those used in Section 3.2. The results for the mean are gathered in Table 3 and the ones for the correlation coefficient are gathered in Table 4.

Since the dimensional structure of the data is less obvious, the potential number of underlying dimensions is unknown *a priori*. Thus, we are in a setting close to what happens with real data and we use cross-validation [21] to select $S$, the number of underlying dimensions used in the BayesMIPCA algorithm. However, cross-validation is time consuming, consequently we cannot perform it for each configuration (*i.e.* for a number of individuals, a number of variables

and a percentage of missing values) and for each of the $K = 1000$ incomplete data sets. For this reason, for each configuration, the choice of $S$ is based on cross-validation performed on 20 incomplete data sets only. This is sufficiently large because of the relative stability of the results. The most frequent number of underlying dimensions over the 20 simulations is retained.

Table 3: Results for the mean. Median confidence intervals width and 95% coverage for $\psi = \mathbb{E}[X_1]$ estimated by several methods (Listwise deletion, Amelia, BayesMI and BayesMIPCA) for different configurations varying the number of individuals ($n = 30$ or 200), the number of variables ($p = 6$ or 60) and the percentage of missing values (10% or 30%). The data sets are drawn from a random covariance matrix. The number of underlying dimensions $S$ is estimated by cross-validation. For each configuration, 1000 data sets with missing values are generated. Some values are not available because of failures of the algorithms.

| | parameters | | | | confidence interval width | | | | coverage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | % | $S$ | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* |
| 1 | 30 | 6 | 0.1 | 4 | 1.026 | 0.777 | 0.777 | 0.765 | 0.949 | 0.948 | 0.947 | 0.947 |
| 2 | 30 | 6 | 0.3 | 2 | | | 0.945 | 0.839 | | | 0.965 | 0.948 |
| 3 | 30 | 60 | 0.1 | 5 | | | | 0.786 | | | | 0.956 |
| 4 | 30 | 60 | 0.3 | 5 | | | | 0.92 | | | | 0.956 |
| 5 | 200 | 6 | 0.1 | 4 | 0.391 | 0.285 | 0.286 | 0.284 | 0.947 | 0.94 | 0.942 | 0.937 |
| 6 | 200 | 6 | 0.3 | 4 | | 0.312 | 0.315 | 0.303 | | 0.945 | 0.954 | 0.937 |
| 7 | 200 | 60 | 0.1 | 5 | | | 0.284 | 0.291 | | | 0.941 | 0.943 |
| 8 | 200 | 60 | 0.3 | 5 | | | 0.359 | 0.321 | | | 0.971 | 0.941 |

The results for the mean are very similar to the ones obtained in Section 3.2.3: the estimator is unbiased for all cases (results not shown here), the coverages are valid and the confidence intervals are shorter for the BayesMIPCA algorithm than for the others. On the contrary, the results for the correlation coefficient highlight the difficulties encountered by BayesMIPCA for data sets with a fuzzy principal component structure. In the cases 3, 4, 7 and 8, where the number of variables is high compared to the number of underlying dimensions estimated (*cf.* Table 4), the coverages are very good and the confidence interval widths are close to the ones obtained by the BayesMI method when it provides results. The hypothesis of an underlying signal in a lower dimensional space is likely in these cases, and consequently, the results are similar to those obtained with a two block structure for the covariance matrix. In the other cases, where the number of variables is small compared to the number of underlying dimensions estimated, the coverages remain satisfactory (greater than 90%) but sometimes worse than previously: in cases 2 and 6 the coverage is close to 92% instead of 95%. Thus, the BayesMIPCA method is all the more efficient in the

Table 4: Results for the correlation coefficient. Increase of the median of the widths of the confidence intervals obtained by the imputation method and the one obtained by the full data, as well as 95% coverage for $\psi = \rho(X_{p-1}, X_p)$. Results are given for several methods (Listwise deletion, Amelia, BayesMI and BayesMIPCA) on different configurations varying the number of individuals ($n$ = 30 or 200), the number of variables ($p$ = 6 or 60) and the percentage of missing values (10% or 30%). The data sets are drawn from a random covariance matrix. The number of underlying dimensions $S$ is estimated by cross-validation. For each configuration, 1000 data sets with missing values are generated. Some values are not available because of failures of the algorithms.

| | parameters | | | | confidence interval width | | | | coverage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | % | $S$ | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* |
| 1 | 30 | 6 | 0.1 | 4 | +47% | +10% | +11% | +30% | 0.944 | 0.959 | 0.962 | 0.947 |
| 2 | 30 | 6 | 0.3 | 2 | | | +66% | +70% | | | 0.977 | 0.911 |
| 3 | 30 | 60 | 0.1 | 5 | | | | +10% | | | | 0.975 |
| 4 | 30 | 60 | 0.3 | 5 | | | | +26% | | | | 0.991 |
| 5 | 200 | 6 | 0.1 | 4 | +41% | +3% | +3% | +9% | 0.946 | 0.953 | 0.953 | 0.954 |
| 6 | 200 | 6 | 0.3 | 4 | | +14% | +21% | +31% | | 0.954 | 0.961 | 0.92 |
| 7 | 200 | 60 | 0.1 | 5 | | | +4% | +8% | | | 0.959 | 0.958 |
| 8 | 200 | 60 | 0.3 | 5 | | | +39% | +23% | | | 0.988 | 0.96 |

case of a low dimensional structure.

In order to go deeper and to deal with larger data, another configuration with 1000 individuals, 200 variables and 10% of missing values is considered. The covariance matrix of size 200×200 is drawn at random [32]. In this configuration, the cross-validation method does not provide a reliable number of dimensions (it gives as a solution the number of variables). Consequently, $S = 17$ dimensions are kept using an ad hoc strategy (by looking at the barplot of the eigenvalues). Because dealing with a big data set is time consuming, multiple imputation using only $M = 5$ imputed data sets is performed. The results for the BayesMI and the BayesMIPCA methods are gathered in Table 5 (the Amelia's algorithm failed on these simulations).

As previously, the coverages are greater than 90% for the BayesMIPCA method, but nevertheless below 95% for the correlation coefficient. The number of underlying dimensions is crudely approximated and we can suppose that in reality it is not sufficiently small compared to the number of variables to reach a coverage of 95%. BayesMIPCA performs better in the case of a low dimensional structure.

Finally, some simulations are performed based on a real large data set. Therefore the low dimensional structure of the data set is again unclear. This

Table 5: Results for the mean and the correlation coefficient. Bias, root mean squared error, median confidence intervals width and 95% coverage for $\psi = \mathbb{E}[X_1]$ and $\psi = \rho(X_{p-1}, X_p)$ estimated by BayesMI and BayesMIPCA for a configuration with $n = 1000$ individuals, $p = 200$ variables and 10% of missing values. The data sets are drawn from a random covariance matrix. 1000 data sets with missing values are generated. Results for the full data are also provided.

| | mean | | | correlation coefficient | |
|---|---|---|---|---|---|
| | BayesMI | BayesMIPCA | Full data | BayesMI | BayesMIPCA |
| bias | -0.001 | 0 | 0 | 0 | 0.011 |
| rmse | 0.032 | 0.034 | 0.032 | 0.032 | 0.035 |
| confidence interval width | 0.127 | 0.131 | 0.124 | +3.31% | +9.09% |
| coverage | 0.955 | 0.958 | 0.96 | 0.949 | 0.933 |

data set is a subset of the million song dataset (MSD)[34]. It contains 463715 songs (rows) and 90 acoustic features (variables) dealing with the timbre of the song. Each feature corresponds to a particular "segment", which is generally delimited by note onsets, or other discontinuities in the signal. It contains also a variable corresponding to the year of the song. The aim is to predict the year of a song using its features. In fact, listeners often have particular affection for music from certain periods of their lives, thus the predicted year could be useful as a basis for recommendation [34]. This subset is available on the web page `http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD`.

To perform simulations from a real data set, the data set is preliminarily scaled to be more likely in lines with the assumption of a homoscedastic noise as stated in 1. We consider that this data set defines the population. Thus, the true value of the quantity of interest is known. Here, we are interested in the regression coefficient corresponding to the first explanatory variable in the regression model predicting the year of the song. To assess the multiple imputation methods, $K = 1000$ samples of size $n = 300$ are drawn from the population, 10% of missing values are added and multiple imputation is performed using $M = 20$ imputed data sets. The cross-validation procedure indicates that 8 dimensions should be retained. The results for the BayesMI method and the BayesMIPCA one are gathered in Table 6 (the Amelia's algorithm fails again which seems to be strongly related to the current version of their implementation).

The BayesMIPCA method provides results close to the ones obtained from the complete data set. The under-coverage observed on the full data could be explained by the small size of the samples compared to the size of the population (300 vs 463715), also by the heterogeneity of the population. The sample size was selected in order to perform simulations in a reasonable time. BayesMIPCA provides results that are more convincing than those of BayesMI (smaller size of the confidence interval). We can suppose that on this real data set, the hypothesis of an underlying signal of lower dimension is likely, and the BayesMIPCA method is well suited.

Table 6: Results for the regression coefficient. Bias, root mean squared error, median confidence intervals width and 95% coverage for $\psi = \beta_{X_2}$ estimated by BayesMI and BayesMIPCA on a subset of size $463715 \times 90$ of the million song dataset. Multiple imputation is performed on 1000 samples of size $n = 300$, drawn from the population and become incomplete with 10% of missing values.

| | BayesMI | BayesMIPCA | Full data |
|---|---|---|---|
| bias | 0.112 | -0.121 | 0.071 |
| rmse | 0.216 | 0.152 | 0.148 |
| confidence interval width | 0.754 | 0.479 | 0.438 |
| coverage | 0.911 | 0.887 | 0.859 |

## 3.4 Simulations from real data

Finally, in order to evaluate the method in practical situations, we perform simulations using four real data sets. In comparison to the previous ones (Section 3.3), here we do not sample from these data sets but consider them as real data sets: it means that each data set is a sample from an unknown population. The first data set refers to $n = 41$ athletes' performances during a decathlon event [35]. It contains $p = 11$ variables, the trials plus the score obtained by the athletes which is strongly related to the 10 other variables. The second data set concerns an isoprenoid gene network in A. Thaliana [36]. This gene network includes $p = 39$ genes each with $n = 118$ gene expression profiles corresponding to different experimental conditions. The genetic data are known to present complex relationships. The third data set deals with $n = 112$ daily measurements of $p = 11$ meteorological variables and ozone concentration recorded in Rennes (France) during summer 2001 [37]. The last data set comes from a sensory study [35] where $n = 21$ wines of Val de Loire were evaluated on $p = 29$ descriptors. The number of individuals is less than the number of variables for this data.

On each data set, 30% of missing values is randomly added and the three multiple imputation methods (Sections 2.2.1 and 3.1) are performed. The listwise deletion method cannot be used for this percentage of missing values. We repeat this process 1000 times. As for the simulations (Section 3.2), we focus on the following quantities: a mean $\mu$, a regression coefficient $\beta$, as well as a correlation coefficient $\rho$. Because we deal with true data sets, the true values for the quantities of interest are unknown. Indeed, these real data sets are samples from a larger unknown population. In Table 7, we report the point estimate and the confidence interval for each quantity, as well as the ones obtained from the completed data sets.

The behaviour of the BayesMIPCA method is quite similar to the one observed on simulations: the method can be applied whatever the data set, and gives the smallest confidence interval. For many cases, the three multiple imputation methods provide similar results close to the ones obtained from the completed data sets. However, the BayesMI method seems very unstable on the

data set Decathlon. For example, the median confidence interval width for the $\beta$ coefficient is equal to 3.363. This could be explained by the collinearity in the data set combined with a small number of individuals.

Table 7: Mean of the point estimates and median confidence intervals width (or relative increase compared to the complete case) for $\mu$, $\rho$, $\beta$ over 1000 simulations. Results are given for several methods (Amelia, BayesMI and BayesMIPCA) on different real datasets (Decathlon, Isoprenoid, Ozone, Wine) with 30% of missing values. Results for the full data are also provided. Some values are not available because of failures of the algorithms.

| | | estimate | | | | confidence interval width | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Amelia* | *BayesMI* | *BayesMIPCA* | *Full data* | *Amelia* | *BayesMI* | *BayesMIPCA* | *Full data* |
| $\mu$ | Decathlon | | 0 | 0 | 0 | | 0.704 | 0.717 | 0.631 |
| | Isoprenoid | | 0.003 | 0.004 | 0 | | 0.448 | 0.406 | 0.365 |
| | Ozone | 0.002 | 0.002 | 0.001 | 0 | 0.403 | 0.409 | 0.402 | 0.374 |
| | Wine | | | 0.014 | 0 | | | 0.998 | 0.91 |
| $\rho$ | Decathlon | | 0.491 | 0.545 | 0.616 | | +92% | +47% | 0.396 |
| | Isoprenoid | | 0.609 | 0.637 | 0.705 | | +82% | +44% | 0.185 |
| | Ozone | 0.65 | 0.66 | 0.654 | 0.685 | +38% | +43% | +30% | 0.2 |
| | Wine | | | 0.536 | 0.607 | | | +35% | 0.585 |
| $\beta$ | Decathlon | | −0.149 | −0.16 | −0.175 | | 3.363 | 0.793 | 0.01 |
| | Isoprenoid | | 0.134 | 0.076 | 0.203 | | 0.584 | 0.44 | 0.382 |
| | Ozone | 0.423 | 0.42 | 0.408 | 0.409 | 0.4 | 0.43 | 0.412 | 0.273 |
| | Wine | | | 0.841 | 0.949 | | | 0.746 | 0.302 |

# 4 Conclusion

Multiple imputation by Bayesian PCA provides valid confidence intervals for both quantities related to the marginal distribution of a variable as well as for quantities related to the relationships between variables from an incomplete continuous data set. Compared to its competitors, it often gives confidence intervals with a smaller width. This is due to the imputation based on PCA. Indeed, PCA is a dimensionality reduction method which isolate the relevant information from the noise. This makes the imputation stable and consequently decreases the variability of the estimator. In addition, the multiple imputation by Bayesian PCA can be easily performed on any kind of data where for instance the number of individuals is less than the number of variables, which is a configuration where other methods encounter difficulties. We have shown that the method is well suited when the hypothesis of an underlying signal of low dimension is verified. In practice, this hypothesis is often true for many data sets. Nevertheless, when the hypothesis of a structure of low dimension is not met, the BayesMIPCA method remains competitive. Note also that since the

imputation is based on PCA, it is particularly well fitted to situations where the relationships between variables are linear, and more generally when the data can be considered as being generated from a PCA model. Thus, the multiple imputation method BayesMIPCA has many advantages and is a flexible alternative to the classical multiple imputation procedures suggested in the literature. However, this method requires tuning a parameter which is the number of dimensions $S$. We suggest the use of cross-validation, or of an approximation of cross-validation, such as generalized cross-validation described in [22] to choose $S$. Simulations not presented here indicated that the method is fairly robust to a misspecified choice for $S$, as long as $S$ is not too small (to be able to capture the relevant information). The BayesMIPCA method is available as an R function on the webpage of the first author.

Future research includes the assessment of the suggested method in cases where there are complex interactions or relationships between variables or cases where for instance a variable $X_1$ and its squared $X_1^2$ are of interest. [38] compared different strategies to handle this latter situation such as the JAV (just another variable) approach which considers the squared version as a new variable in itself without taking into account its link with $X_1$. [39] suggested another MI method to handle such situations better but it does not give the possibility to deal with missing values in all the variables in its current form.

The encouraging results of the Bayesian PCA for continuous variables prompt the extension of the method to perform multiple imputation for categorical variables using multiple correspondence analysis [40] and using factorial analysis for mixed data [41, 42]. [43] suggested single imputation methods based on principal component methods for data with continuous, categorical and mixed variables showing good results to predict the missing entries. However, the extension to multiple imputation is not straightforward, because the method presented for continuous variables is based on a Bayesian treatment of a joint model for all variables. The model is well known for PCA, but the model is yet unknown for multiple correspondence analysis and *a fortiori* for the factor analysis of mixed data. Further research would be required if a Bayesian approach of these principal component methods, and therefore multiple imputation based on these methods, was being considered.

# References

[1] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society B. 1977; 39:1–38.

[2] Meng XL, Rubin DB. Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm. Journal of the American Statistical Association. 1991 Dec;86(416):899–909.

[3] Rubin DB. Multiple imputation for non-response in survey. Wiley; 1987.

[4] Little RJA, Rubin DB. Statistical analysis with missing data. New-York: Wiley series in probability and statistics; 1987, 2002.

[5] Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall/CRC; 1997.

[6] Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006;76:1049–1064.

[7] Besag J. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society Series B (Methodological). 1974; 36(2).

[8] Liu J, Gelman A, Hill J, Su YS, Kropko J. On the stationary distribution of iterative imputations. Biometrika. 2014 Mar;:155–173.

[9] Kropko J, Goodrich B, Gelman A, Hill J. Multiple imputation for continuous and categorical data: Comparing joint and conditional approaches. Political Analysis. 2014;.

[10] Josse J, Husson F. Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique. 2012;153 (2):1–21.

[11] Caussinus H. Models and uses of principal component analysis (with discussion). In: Multidimensional data analysis. DSWO Press; 1986. p. 149–178.

[12] Candès EJ, Tao T. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans Inf Theor. 2009 May;56(5):2053–2080.

[13] Shabalin A, Nobel B. Reconstruction of a low-rank matrix in the presence of gaussian noise. Journal of Multivariate Analysis. 2013;118(0):67 – 76.

[14] Verbanck M, Josse J, Husson F. Regularised PCA to denoise and visualise data. Statistics and Computing. 2013;:1–16.

[15] Josse J, Sardy S. Adaptive shrinkage of singular values. Statistics and Computing. 2015;:1–10.

[16] Huet S, Denis J, Adamczyk K. Bootstrap confidence intervals in nonlinear regression models when the number of observations is fixed and the variance tends to o. application to biadditive models. Statistics. 1999;32:203–227.

[17] Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. Psychometrika. 1997;62:251–266.

[18] Efron B, Morris C. Empirical Bayes on Vector Observations: An Extension of Stein's Method. Biometrika. 1972;59(2):335–347.

[19] Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association. 1987; 82:805–811.

[20] Jolliffe IT. Principal component analysis. Springer; 2002.

[21] Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component model: a critical look at current methods. Anal Bioanal Chem. 2008; 390:1241–1251.

[22] Josse J, Husson F. Selecting the number of components in PCA using cross-validation approximations. Computational Statististics and Data Analysis. 2011;56(6):1869–1879.

[23] Van Buuren S. Flexible imputation of missing data (chapman & hall/crc interdisciplinary statistics). 1st ed. Chapman and Hall/CRC; 2012.

[24] Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. Bmc Medical Research Methodology. 2009;9(5):57.

[25] Barnard J, Rubin DB. Small Sample Degrees of Freedom with Multiple Imputation. Biometrika. 1999;86:948–955.

[26] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria. 2014; Available from: http://www.R-project.org.

[27] Honaker J, King G, Blackwell M. Amelia ii: A program for missing data. 2014; r package version 1.7.2.

[28] Honaker J, King G, Blackwell M. Amelia II: A program for missing data. Journal of Statistical Software. 2011;45(7):1–47.

[29] Van Buuren S. mice. 2014; r package version 2.18.

[30] Van Buuren S, Groothuis-Oudshoorn CGM. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67.

[31] Honaker J, King G. What to do about missing values in time series cross-section data. American Journal of Political Science. 2010;54:561–581.

[32] Joe H. Generating random correlation matrices based on partial correlations. J Multivar Anal. 2006 Nov;97(10):2177–2189.

[33] Qiu W, Joe H. clustergeneration: random cluster generation (with specified degree of separation). 2013; r package version 1.3.1.

[34] Bertin-Mahieux T, Ellis D, Whitman B, Lamere P. The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011); 2011.

[35] Husson F, Josse J, Le S, Mazet J. Factominer: Multivariate exploratory data analysis and data mining with r. 2013; r package version 1.25; Available from: http://CRAN.R-project.org/package=FactoMineR.

[36] Wille A, Zimmermann P, Vranova E, Furholz A, Laule O, Bleurer S, Henning L, Prelic A, Von Rohr P, Thiele L, Zitzler E, Gruissem W, Buhlmann P. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. Genome Biology. 2004;5(11):R92+.

[37] Cornillon PA, Guyader A, Husson F, Jégou N, Josse J, Kloareg M, Matzner-Løber E, Rouvière L. R for statistics. Rennes: Chapman & Hall/CRC Computer Science & Data Analysis; 2012.

[38] Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC medical research methodology. 2012;12(1):46.

[39] Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. ArXiv e-prints. 2013;In revision.

[40] Greenacre M, Blasius J. Multiple correspondence analysis and related methods. Chapman & Hall/CRC; 2006.

[41] Kiers HAL. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika. 1991;56:197–212.

[42] Pagès J. Multiple factor analysis by example using r. Chapman & Hall/CRC The R Series; Taylor & Francis; 2014; Available from: http://books.google.fr/books?id=EOxZngEACAAJ.

[43] Audigier V, Husson F, Josse J. A principal components method to impute missing values for mixed data. ArXiv e-prints. 2013;In revision.

# A Simulation study with a block diagonal structure for the covariance matrix - Results for the regression coefficient

Table 8: Results for the regression coefficient. Root mean squared error for the parameter $\psi = \beta_{X_2}$ estimated by Listwise deletion, Amelia, BayesMI and BayesMIPCA on different configurations varying the number $n$ of individuals, the number $p$ of variables, the correlation $\rho$ between variables and the percentage of missing values. The median confidence interval width for the full data are also provided. For each configuration, 1000 incomplete data sets are generated. Note that $\beta_{X_2}$ can not be estimated if $n < p$. Some values are not available because of failures of the algorithms

| | parameters | | | | root mean square error | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | $\rho$ | % | LD | Amelia | BayesMI | BayesMIPCA |
| 1 | 30 | 6 | 0.3 | 0.1 | 0.352 | 0.269 | 0.249 | 0.194 |
| 2 | 30 | 6 | 0.3 | 0.3 | | | 0.391 | 0.183 |
| 3 | 30 | 6 | 0.9 | 0.1 | 0.335 | 0.277 | 0.242 | 0.171 |
| 4 | 30 | 6 | 0.9 | 0.3 | | | 0.362 | 0.127 |
| 9 | 200 | 6 | 0.3 | 0.1 | 0.099 | 0.078 | 0.078 | 0.066 |
| 10 | 200 | 6 | 0.3 | 0.3 | 0.266 | 0.115 | 0.109 | 0.062 |
| 11 | 200 | 6 | 0.9 | 0.1 | 0.093 | 0.075 | 0.074 | 0.058 |
| 12 | 200 | 6 | 0.9 | 0.3 | 0.265 | 0.118 | 0.11 | 0.046 |
| 13 | 200 | 60 | 0.3 | 0.1 | | | 0.113 | 0.072 |
| 14 | 200 | 60 | 0.3 | 0.3 | | | 0.171 | 0.054 |
| 15 | 200 | 60 | 0.9 | 0.1 | | | 0.113 | 0.072 |
| 16 | 200 | 60 | 0.9 | 0.3 | | | 0.11 | 0.053 |

Table 9: Results for the regression coefficient. 95% coverage and median confidence interval width for the parameter $\psi = \beta_{X_2}$ estimated by Listwise deletion, Amelia, BayesMI and BayesMIPCA on different configurations varying the number $n$ of individuals, the number $p$ of variables, the correlation $\rho$ between variables and the percentage of missing values. The median confidence interval width for the full data are also provided. For each configuration, 1000 incomplete data sets are generated. Note that $\beta_{X_2}$ can not be estimated if $n < p$. Some values are not available because of fails of the algorithms

| | parameters | | | | confidence interval width | | | | | coverage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $p$ | $\rho$ | % | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* | *Full data* | *LD* | *Amelia* | *BayesMI* | *BayesMIPCA* |
| 1 | 30 | 6 | 0.3 | 0.1 | 1.332 | 1.058 | 0.989 | 0.936 | 0.818 | 0.945 | 0.94 | 0.953 | 0.974 |
| 2 | 30 | 6 | 0.3 | 0.3 | | | 2.492 | 1.147 | 0.818 | | | 0.981 | 0.997 |
| 3 | 30 | 6 | 0.9 | 0.1 | 1.286 | 1.051 | 0.991 | 0.915 | 0.791 | 0.952 | 0.951 | 0.957 | 0.994 |
| 4 | 30 | 6 | 0.9 | 0.3 | | | 2.972 | 1.108 | 0.791 | | | 0.992 | 1 |
| 9 | 200 | 6 | 0.3 | 0.1 | 0.389 | 0.313 | 0.313 | 0.307 | 0.278 | 0.954 | 0.955 | 0.96 | 0.98 |
| 10 | 200 | 6 | 0.3 | 0.3 | 1.011 | 0.444 | 0.432 | 0.359 | 0.278 | 0.953 | 0.945 | 0.94 | 0.995 |
| 11 | 200 | 6 | 0.9 | 0.1 | 0.374 | 0.307 | 0.306 | 0.3 | 0.267 | 0.956 | 0.958 | 0.971 | 0.99 |
| 12 | 200 | 6 | 0.9 | 0.3 | 0.966 | 0.465 | 0.442 | 0.349 | 0.267 | 0.956 | 0.944 | 0.949 | 0.999 |
| 13 | 200 | 60 | 0.3 | 0.1 | | | 0.467 | 0.373 | 0.332 | | | 0.955 | 0.989 |
| 14 | 200 | 60 | 0.3 | 0.3 | | | 2.716 | 0.428 | 0.332 | | | 1 | 1 |
| 15 | 200 | 60 | 0.9 | 0.1 | | | 0.465 | 0.373 | 0.332 | | | 0.956 | 0.993 |
| 16 | 200 | 60 | 0.9 | 0.3 | | | 1.012 | 0.431 | 0.332 | | | 1 | 1 |

# CHAPITRE 5

# IMPUTATION MULTIPLE DE DONNÉES QUALITATIVES

Dans ce chapitre, une méthode d'imputation multiple par ACM pour des variables qualitatives nominales est proposée. La méthode est comparée, sur la base de jeux réels, à cinq autres méthodes d'imputation multiple : imputation par modèle Gaussien (King *et al.*, 2001), imputation par le modèle log-linéaire (Schafer, 1997), imputation par le modèle à classes latentes (Si et Reiter, 2013), imputation par équations enchaînées utilisant des modèles de régression logistiques (Van Buuren, 2012), ou des forêts aléatoires (Shah *et al.*, 2014). Le caractère proper de la méthode pour des coefficients de régression logistique est illustrée. L'imputation multiple par ACM a l'avantage de permettre l'inférence sur des jeux où le nombre de variables, ou le nombre de modalités est grand. La méthode est également rapide d'exécution.

## Contents



L'imputation des variables qualitatives présente une difficulté supplémentaire par rapport à l'imputation des variables quantitative du fait du phénomène d'explosion combinatoire dès lors que le nombre de modalités ou le nombre de variables est élevé. L'imputation classique sous le modèle log-linéaire ou Gaussien se retrouve ainsi rapidement surparamétrée quand le nombre de variables est grand. L'approche par équations enchaînées est une alternative intéressante, mais le temps de calcul nécessaire est fonction du nombre de modèles conditionnels, ce qui reste un inconvénient sur un grand jeu de données. Ce chapitre a pour but de présenter une méthode d'imputation multiple pour des données nominales utilisant l'ACM afin d'apporter une méthode d'imputation multiple par modèle joint applicable quelque soit le nombre de variables ou le nombre de modalités. A nouveau, l'extension de l'imputation simple par ACM, présentée au Chapitre 3, à l'imputation multiple nécessite d'ajouter un aléa sur la prédiction par ACM afin de mieux respecter la structure du jeu de données et également de refléter l'incertitude sur les composantes principales et les vecteurs propres. Cette méthode est détaillée dans l'article Audigier *et al.* (2015a) qui constitue la suite de ce chapitre.

# MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis

Vincent Audigier, François Husson, Julie Josse

Agrocampus Ouest, 65 rue de St-Brieuc, F-35042 Rennes
Tel.: +33-223485874 Fax: +33-223485871
audigier@agrocampus-ouest.fr
husson@agrocampus-ouest.fr
josse@agrocampus-ouest.fr

Nov, 2015

**Abstract**

We propose a multiple imputation method to deal with incomplete categorical data. This method imputes the missing entries using the principal components method dedicated to categorical data: multiple correspondence analysis (MCA). The uncertainty concerning the parameters of the imputation model is reflected using a non-parametric bootstrap. Multiple imputation using MCA (MIMCA) requires estimating a small number of parameters due to the dimensionality reduction property of MCA. It allows the user to impute a large range of data sets. In particular, a high number of categories per variable, a high number of variables or a small the number of individuals are not an issue for MIMCA. Through a simulation study based on real data sets, the method is assessed and compared to the reference methods (multiple imputation using the log-linear model, multiple imputation by logistic regressions) as well to the latest works on the topic (multiple imputation by random forests or by the Dirichlet process mixture of products of multinomial distributions model). The proposed method provides a good point estimate of the parameters of the analysis model considered, such as the coefficients of a main effects logistic regression model, and a reliable estimate of the variability of the estimators. In addition, MIMCA has the great advantage that it is substantially less time consuming on data sets of high dimensions than the other multiple imputation methods.

*Keywords* missing values, categorical data, multiple imputation, multiple correspondence analysis, bootstrap.

# 1 Introduction

Data sets with categorical variables are ubiquitous in many fields such in social sciences, where surveys are conducted through multiple-choice questions. Whatever the field, missing values frequently occur and are a key problem in statistical practice since most of statistical methods cannot be applied directly on incomplete data.

To deal with missing values one solution consists in adapting the statistical method so that it can be applied on an incomplete data set. For instance, the maximum likelihood (ML) estimators can be derived from incomplete data using an Expectation-Maximization (EM) algorithm [18] and their standard error can be estimated using a Supplemented Expectation-Maximization algorithm [35]. The ML approach is suitable, but not always easy to establish [5].

Another way consists in replacing missing values by plausible values according to an *imputation model*. This is called *single imputation*. Thus, the data set is complete and any statistical method can be applied on this one. Figure 1 illustrates three simple single imputation methods. The data set used con-

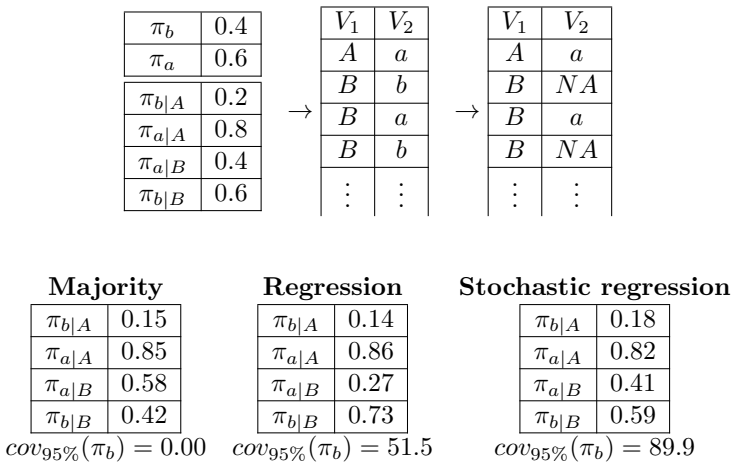


**Figure 1:** Illustration of three imputation methods for two categorical variables: the top part described how the data are built (marginal and conditional proportions, associated complete data, incomplete data set generated where NA denotes a missing value) and the bottom part sums up the observed conditional proportions after an imputation by several methods (majority, regression, stochastic regression). The last line indicates the coverage for the confidence interval for the proportion of $b$ over 1000 simulations.

tains 1000 individuals and two variables with two categories: $A$ and $B$ for the first variable, $a$ and $b$ for the second one. The data set is built so that 40%

of the individuals take the category $a$ and 60% the category $b$. In addition, the variables are linked, that is to say, the probability to observe $a$ or $b$ on the second variable depends on the category taken on the first variable. Then, 30% of missing values are generated completely at random on the second variable. A first method could be to impute according to the most taken category of the variable. In this case, all missing values are imputed by $a$. Consequently, marginal proportions are modified, as well as conditional proportions (see the bottom part of Figure 1). This method is clearly not suitable. A more convenient solution consists in taking into account the relationship between the two variables, following the rationale of the imputation by regression for continuous data. To achieve this goal, the parameters of a logistic regression are estimated from the complete cases, providing fitted conditional proportions. Then, each individual is imputed according to the highest conditional proportion given the first variable. This method respects the conditional proportions better, but the relationship between variables is strengthened which is not satisfactory. In order to obtain an imputed data set with a structure as close as possible to the generated data set, a suitable single imputation method is to perform stochastic regression: instead of imputing according to the most likely category, the imputation is performed randomly according to the fitted probabilities.

An imputation model used to perform single imputation has to be sufficiently complex compared to the statistical method desired (the *analysis model*). For instance, if the aim is to apply a logistic regression from an incomplete data set, it requires using an imputation model taking into account the relationships between variables. Thus, a suitable single imputation method, such as the stochastic regression strategy, leads to unbiased estimates of the parameters of the statistical method (see Figure 1). However, although the single imputation method respects the structure of the data, it still has the drawback that the uncertainty on the imputed values is not taken into account in the estimate of the variability of the estimators. Thus, this variability remains underestimated. For instance, in Figure 1, the level of the confidence interval of $\pi_b$, the proportion of $b$, is 89.9% and does not reach the nominal rate of 95%.

Multiple imputation (MI) [34, 38] has been developped to avoid this issue. The principle of multiple imputation consists in creating $M$ imputed data sets to reflect the uncertainty on imputed values. Then, the parameters of the statistical method, denoted $\psi$, are estimated from each imputed data set, leading to $M$ sets of parameters $(\widehat{\psi}_m)_{1\leq m\leq M}$. Lastly, these sets of parameters are pooled to provide a unique estimation for $\psi$ and for its associated variability using Rubin's rules [38].

MI is based on the *ignorability* assumption, that is to say ignoring the mechanism that generated missing values. This assumption is equivalent to: first, the parameters that govern the missing data mechanism and the parameters of the analysis model are independent; then, missing values are generated *at random*, that is to say, the probability that a missing value occurs on a cell is independent from the value of the cell itself. In practice, ignorability and value missing at random (MAR), are used interchangeably. This assumption is more plausible when the number of variables is high [39, 49], but remains difficult to

verify.

Thus, under the ignorability assumption, the main challenge in multiple imputation is to reflect the uncertainty of the imputed values by reflecting *properly* [38, p. 118-128] the uncertainty on the parameters of the model used to perform imputation to get imputed data sets yielding to valid statistical inferences. To do so, two classical approaches can be considered. The first one is the Bayesian approach: a prior distribution is assumed on the parameters $\theta$ of the imputation model, it is combined with the observed entries, providing a posterior distribution from which $M$ sets of parameters $\left(\tilde{\theta}_m\right)_{1\leq m\leq M}$ are drawn. Then, the incomplete data set is imputed $M$ times using each set of parameters. The second one is a bootstrap approach: $M$ samples with replacement are drawn leading to $M$ incomplete data sets from which the parameters of the imputation model are obtained. The $M$ sets of parameters $(\theta_m)_{1\leq m\leq M}$ are then used to perform $M$ imputations of the original incomplete data set.

In this paper, we detail in Section 2 the main available MI methods to deal with categorical data. Two general modelling strategies can be distinguished for imputing multivariate data: joint modelling (JM) [39] and fully conditional specification (FCS)[46]. JM is based on the assumption that the data can be described by a multivariate distribution. Concerning FCS, the multivariate distribution is not defined explicitly, but implicitly through the conditional distributions of each variable only. Among the presented methods, three are JM methods: MI using the loglinear model, MI using the latent class model and MI using the normal distribution; the two others are FCS strategies: the FCS using logistic regressions and FCS using random forests [15]. In Section 3, a novel JM method based on a principal components method dedicated to categorical data, namely multiple correspondence analysis (MCA), is proposed. Principal components methods are commonly used to highlight the similarities between individuals and the relationships between variables, using a small number of principal components and loadings. MI based on this family of methods uses these similarities and these relationships to perform imputation, while using a restricted number of parameters. The performances of the imputation are very promising from continuous data [7, 30] which motivates the consideration of a method for categorical data. In Section 4, a simulation study based on real data sets, evaluates the novel method and compares its performances to other main multiple imputation methods. Lastly, conclusions about MI for categorical data and possible extensions for the novel method are detailed.

## 2 Multiple imputation methods for categorical data

The imputation of categorical variables is rather complex. Indeed, contrary to continuous data, the variables follow a distribution on a discrete support defined by the combinations of categories observed for each individual. Because of the explosion of the number of combinations when the number of categories

increases, the number of parameters defining the multivariate distribution could be extremely large. Consequently, defining an imputation model is not straightforward for categorical data. In this section we review the most popular approaches commonly used to deal with categorical data: JM using the loglinear model, JM using the latent class model, JM using the normal distribution and FCS using multinomial logistic regression or random forests.

Hereinafter, matrices and vectors will be in bold text, whereas sets of random variables or single random variables will not. Matrices will be in capital letters, whereas vectors will be in lower case letters. We denote $\mathbf{X}_{I\times K}$ a data set with $I$ individuals and $K$ variables and $\mathbf{T}$ the corresponding contingency table. We note the observed part of $\mathbf{X}$ by $\mathbf{X}_{obs}$ and the missing part by $\mathbf{X}_{miss}$, so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$. Let $q_k$ denote the number of categories for the variable $\mathbf{X}_k$, $J = \sum_{k=1}^{K} q_k$ the total number of categories. We note $\mathbb{P}(X, \theta)$ the distribution of the variables $X = (X_1, \ldots, X_K)$, where $\theta$ is the corresponding set of parameters.

## 2.1 Multiple imputation using a loglinear model

The saturated loglinear model (or multinomial model) [1] consists in assuming a multinomial distribution $\mathcal{M}(\theta, I)$ as joint distribution for $\mathbf{T}$, where $\theta = (\theta_{x_1\ldots x_K})_{x_1\ldots x_K}$ is a vector indicating the probability to observe each event $(X_1 = x_1, \ldots, X_K = x_K)$. Like any MI method, MI with the loglinear model [39] requires to reflect, through the imputed values, the uncertainty on the parameters of the imputation model. It can be achieved by using a Bayesian approach. More precisely, a Bayesian treatment of this model can be specified as follows:

$$T|\theta \sim \mathcal{M}(\theta, I) \tag{1}$$

$$\theta \sim \mathcal{D}(\alpha) \tag{2}$$

$$\theta|T \sim \mathcal{D}(\alpha + T) \tag{3}$$

where $\mathcal{D}(\alpha)$ denotes the Dirichlet distribution with parameter $\alpha$, a vector with the same dimension as $\theta$. A classical choice for $\alpha$ is $\alpha = (1/2, \ldots, 1/2)$ corresponding to the non-informative Jeffreys prior [13]. Combining the prior distribution and the observed entries, a posterior distribution for the model's parameters is obtained (Equation (3)).

Because missing values occur in the data set, the posterior distribution is not tractable, therefore, drawing a set of model's parameters in it is not straightforward. Thus, a data-augmentation algorithm [43] is used. In the first step of the algorithm, missing values are imputed by random values. Then, because the data set is now completed, a draw of $\theta$ in the posterior distribution (3) can easily be obtained. Next, missing values are imputed from the predictive distribution (1) using the previously drawn parameter and the observed values. These steps of imputation and draw from the posterior distribution are repeated until convergence. At the end, one set of parameters $\tilde{\theta}_m$, drawn from the observed posterior distribution, is obtained. Repeating the procedure $M$ times in parallel, $M$ sets of parameters are obtained from which multiple imputation

can be done. In this way, the uncertainty on the parameters of the imputation model is reflected, insuring a proper imputation.

The loglinear model is considered as the gold standard for MI of categorical data [50]. Indeed, this imputation model reflects all kind of relationships between variables, which enables applying any analysis model. However, this method is dedicated to data sets with a small number of categories because it requires a number of independent parameters equal to the number of combinations of categories minus 1. For example, it corresponds to 9 765 624 independent parameters for a data set with $K = 10$ variables with $q_k = 5$ categories for each of them. This involves two issues: the storage of $\theta$ and overfitting. To overcome these issues, the model can be simplified by adding constraints on $\theta$. The principle is to write $log(\theta)$ as a linear combination of a restricted set of parameters $\lambda = \left[\lambda_0, \lambda_{x_1}, \ldots, \lambda_{x_K}, \ldots, \lambda_{x_1 x_2}, \ldots, \lambda_{x_1 x_K}, \ldots, \lambda_{x_{K-1} x_K}\right]$, where each element is indexed by a category or a couple of categories. More precisely, the constraints on $\theta$ are given by the following equation:

$$log(\theta_{x_1 \ldots x_K}) = \lambda_0 + \sum_k \lambda_{x_k} + \sum_{k>k'} \lambda_{x_k x_{k'}} \quad \text{for all } (X_1 = x_1, \ldots, X_K = x_K) \quad (4)$$

where the second sum is the sum over all the couples of categories possible from the set of categories $(x_1, \ldots, x_K)$. Thus, the imputation model reflects only the simple (two-way) associations between variables, which is generally sufficient. Equation (4) leads to 760 independent parameters for the previous example. However, although it requires a smaller number of parameters, the imputation under the loglinear model still remains difficult in this case, because the data-augmentation algorithm used [39, p.320] is based on a modification of $\theta$ at each iteration and not of $\lambda$. Thus the storage issue remains.

## 2.2 Multiple imputation using a latent class model

To overcome the limitation of MI using the loglinear model, another JM method based on the latent class model can be used. The latent class model [1, p.535] is a mixture model based on the assumption that each individual belongs to a latent class from which all variables can be considered as independent. More precisely, let $Z$ denote the latent categorical variable whose values are in $\{1, \ldots, L\}$. Let $\theta_Z = (\theta_\ell)_{1 \leq \ell \leq L}$ denote the proportion of the mixture and $\theta_X = \left(\theta_x^{(\ell)}\right)_{1 \leq \ell \leq L}$ the parameters of the $L$ components of the mixture. Thus, let $\theta = (\theta_Z, \theta_X)$ denote the parameters of the mixture, the joint distribution of the data is written as follows:

$$\mathbb{P}\left(X = (x_1, \ldots, x_K); \theta\right) = \sum_{\ell=1}^{L} \left( \mathbb{P}\left(Z = \ell, \theta_Z\right) \prod_{k=1}^{K} \mathbb{P}\left(X_k = x_k | Z = \ell; \theta_x^{(\ell)}\right) \right) \quad (5)$$

Assuming a multinomial distribution for $Z$ and $X_k|Z$, Equation (5), can be rewritten as follows:

$$\mathbb{P}\left(X = (x_1, \ldots, x_K); \theta\right) = \sum_{\ell=1}^{L} \left( \theta_\ell \prod_{k=1}^{K} \theta_{x_k}^{(\ell)} \right) \tag{6}$$

The latent class model requires $L \times (J - K) + (K - 1)$ independent parameters, *i.e.* a number that linearly increases with the number of categories.

[51] reviews in detail different multiple imputation methods using a latent class model. These methods can be distinguished by the way used to reflect the uncertainty on the parameters of the imputation model and by the way that the number of components of the mixture is chosen: automatically or *a priori*. The quality of the imputation is quite similar from one method to another, the main differences remain in computation time. One of the latest contributions in this family of methods uses a non-parametric extension of the model namely the Dirichlet process mixture of products of multinomial distributions model (DPMPM) [20, 42]. This method uses a fully Bayesian approach in which the number of classes is defined automatically and is not too computationally intensive. DPMPM assumes a prior distribution on $\theta_Z = (\theta_\ell)_{1 \leq \ell \leq L}$ without fixing the number of classes which is supposed to be infinite. More precisely, the prior distribution for $\theta_Z$ is defined as follows:

$$\theta_\ell = \zeta_\ell \prod_{g<\ell} (1 - \zeta_g) \text{ for } \ell \text{ in } 1, \ldots, \infty \tag{7}$$

$$\zeta_\ell \sim \mathcal{B}(1, \alpha) \tag{8}$$

$$\alpha \sim \mathcal{G}(.25, .25) \tag{9}$$

where $\mathcal{G}$ refers to the gamma distribution, $\alpha$ is a positive real number, $\mathcal{B}$ refers to the beta distribution; the prior distribution for $\theta_X$ is defined by:

$$\theta_x^{(\ell)} \sim \mathcal{D}\left(1, \ldots, 1\right) \tag{10}$$

corresponding to a uniform distribution over the simplex defined by the constraint of sum to one. The posterior distribution of $\theta$ is not analytically tractable, even when no missing value occur. However, the distribution of each parameter is known if the others are given. For this reason, a Gibbs sampler is used to obtain a draw from the posterior distribution. The principle of this is to draw each parameter while fixing the others. From an incomplete data set, missing values require to be preliminarily imputed. More precisely, a draw from the posterior distribution is obtained as follows: first, the parameters and missing values are initialized; then, given the current parameters, particularly $\theta_Z$ and $\theta_X$, each individual is randomly affected to one class according to its categories; next, each parameter ($\theta_Z$, $\theta_X$, $\alpha$) is drawn conditionally to the others; finally, missing values are imputed according to the mixture model. These steps are then repeated until convergence (for more details, see [42]).

Despite the infinite number of classes, the prior on $\theta_\ell$ typically implies that the posterior distribution for $\theta_\ell$ is non negligible for a finite number of classes

only. Moreover, for computational reasons, the number of classes has to be bounded. Thus, [42] recommends to fix the maximum number of latent classes to twenty. Consequently, the simulated values of $\theta$ are some realisations of an approximated posterior distribution only.

Multiple imputation using the latent class model has the advantages and drawbacks of this model: because the latent class model approximates quite well any kind of relationships between variables, MI using this model enables the use of complex analysis models such as logistic regression with some interaction terms and provides good estimates of the parameters of the analysis model. However, the imputation model implies that given a class, each individual is imputed in the same way, whatever the categories taken. If the class is very homogeneous, all the individuals have the same observed values, and this behaviour makes sense. However, when the number of missing values is high and when the number of variables is high, it is not straightforward to obtain homogeneous classes. It can explain why [51] observed that the multiple imputation using the latent class model can lead to biased estimates for the analysis model in such cases.

## 2.3 Multiple imputation using a multivariate normal distribution

Another popular strategy to perform MI for categorical data is to adapt the methods developed for continuous data. Because multiple imputation using the normal multivariate distribution is a robust method for imputing continuous non-normal data [39], imputation using the multivariate normal model is an attractive method for this. The principle consists in recoding the categorical variables as dummy variables and applying the multiple imputation under the normal multivariate distribution on the recoded data. The imputed dummy variables are seen as a set of latent continuous variables from which categories can be independently derived. More precisely, let $\mathbf{Z}_{I\times J}$ denote the disjunctive table coding for $\mathbf{X}_{I\times K}$, *i.e.*, the set of dummy variables corresponding to the incomplete matrix. Note that one missing value on $\mathbf{x}_k$ implies $q_k$ missing values for $\mathbf{z}_k$. The following procedure implemented in [27, 28] enables the multiple imputation of a categorical data set using the normal distribution:

- perform a non-parametric bootstrap on $\mathbf{Z}$: sample the rows of $\mathbf{Z}$ with replacement $M$ times. $M$ incomplete disjunctive tables $\left(\mathbf{Z}_m^{boot}\right)_{1\leq m\leq M}$ are obtained;
- estimate the parameters of the normal distribution on each bootstrap replicate: calculate the ML estimators of $(\mu_m, \Sigma_m)$, the mean and the variance of the normal distribution for the $m^{th}$ bootstrap incomplete replicate, using an EM algorithm. Note that the set of $M$ parameters reflects the uncertainty required for a proper multiple imputation method;
- create $M$ imputed disjunctive tables: impute $\mathbf{Z}$ from the normal distribution using $(\mu_m, \Sigma_m)_{1\leq m\leq M}$ and the observed values of $\mathbf{Z}$. $M$ imputed

disjunctive tables $(\mathbf{Z}_m)_{1\leq m\leq M}$ are obtained. In $\mathbf{Z}_m$, the observed values are still zeros and ones, whereas the missing values have been replaced by real numbers;

- create $M$ imputed categorical data sets: from the latent continuous variables contained in $(\mathbf{Z}_m)_{1\leq m\leq M}$, derive categories for each incomplete individual.

Several ways have been proposed to get the imputed categories from the imputed continuous values. For example [4] recommends to attribute the category corresponding to the highest imputed value, while [11, 17, 52] propose some rounding strategies. However, *"A single best rounding rule for categorical data has yet to be identified."* [45, p. 107]. A common one proposed by [11] is called *Coin flipping*. Coin flipping consists in considering the set of imputed values of the $q_k$ dummy variables $\mathbf{z}_k$ as an expectation given the observed values $\theta_k = \mathbb{E}\left[(z_1,\ldots,z_{q_k})|Z_{obs};\hat{\mu},\hat{\Sigma}\right]$. Thus, randomly drawing one category according to a multinomial distribution $\mathcal{M}(\theta_k,1)$, suitably modified so that $\theta_k$ remains between 0 and 1, imputes plausible values. The values lower than 0 are replaced by 0 and the imputed values higher than 1 are replaced by 1. In this case, the imputed values are scaled to respect the constraint of sum to one.

Because imputation under the normal multivariate distribution is based on the estimate of a covariance matrix, the imputation under the normal distribution can detect only two-way associations between categorical variables. In addition, this method assumes independence between categories conditionally to the latent continuous variables. This implies that if two variables are linked, and if an individual has missing values on these ones, then the categories derived from the imputed disjunctive table will be drawn independently. Consequently, the two-way associations can not be perfectly reflected in the imputed data set. Note that, contrary to the MI using the latent class, the parameter of the multinomial distribution $\theta_k$ is specific to each individual, because the imputation of the disjunctive table is performed given the observed values. This behaviour makes sense if the variables on which missing values occur are linked with the others. The main drawback of the MI using the normal distribution is the number of independent parameters estimated. This number is equal to $\frac{(J-K)\times(J-K+1)}{2}+(J-K)$, representing 860 parameters for a data set with 10 variables with 5 categories. It increases rapidly when the total number of categories $(J)$ increases, leading quickly to overfitting. Moreover, the covariance matrix is not invertible when the number of individuals is lower than $(J-K)$. To overcome these issues, it is possible to add a ridge term on its diagonal to improve the conditioning of the regression problem.

## 2.4 Fully conditional specification

Categorical data can be imputed using a FCS approach instead of a JM approach: for each variable with missing values, an imputation model is defined,

(*i.e.* a conditional distribution), and each incomplete variable is sequentially imputed according to this, while reflecting the uncertainty on the model's parameters. Typically, the models used for each incomplete variable are some multinomial logistic regressions and the variability of the models parameter is reflected using a Bayesian point of view. More precisely, we denote by $\theta_k = (\theta_{k_\ell})_{1 \leq \ell \leq q_k}$ the set of parameters for the multinomial distribution of the variable to impute $X_k$ (the set of the other variables is denoted $X_{-k}$). We also denote by $\boldsymbol{\beta_k} = (\boldsymbol{\beta_{k1}}, \ldots, \boldsymbol{\beta_{kL}})$ the set of regression parameters that defines $\theta_k$, such as $\boldsymbol{\beta}_{k\ell}$ is the regression parameter vector associated with the category $\ell$ of the response variable $X_k$ and $\mathbf{Z}_k$ is the design matrix associated. Note that identifiability constraints are required on $\boldsymbol{\beta}_k$, that is why $\boldsymbol{\beta}_{kL}$ is fixed to the null vector. Thus, the imputation is built on the following assumptions:

$$X_k|\theta_k \sim \mathcal{M}(\theta_k, 1) \tag{11}$$

$$\begin{aligned} \theta_{k\ell} &= \mathbb{P}(X_k = \ell | X_{-k}, \beta) \\ &= \frac{exp(\mathbf{Z}_k \boldsymbol{\beta}_{k\ell})}{1 + \sum_{\ell=1}^{L-1} exp(\mathbf{Z}_k \boldsymbol{\beta}_{k\ell})} \end{aligned} \tag{12}$$

$$\beta|X \sim \mathcal{N}(\hat{\beta}, \hat{V}) \tag{13}$$

where $\hat{\beta}, \hat{V}$ are the estimators of $\beta$ and of its associated variance. For simplicity, suppose that the data set contains 2 binary variables $\mathbf{x}_1$ and $\mathbf{x}_2$, with $\mathbf{x}_2$ as incomplete and $\mathbf{x}_1$ as complete. To impute $\mathbf{x}_2$ given $\mathbf{x}_1$ the first step is to estimate $\beta$ and its associated variance using complete cases by iteratively reweighted least squares. Then, a new parameter $\tilde{\beta}_k$ is drawn from a normal distribution centred in the previous estimate with the covariance matrix previously obtained. Lastly, the fitted probability $\theta_k$ are obtained from the logistic regression model with parameter $\tilde{\beta}_k$ and $\mathbf{x}_2$ is imputed according to a multinomial distribution with parameters $\theta_k$ [45, p.76]. Note that $\beta$ is drawn in an approximated posterior distribution. Indeed, as explained by [38, p.169-170], the posterior distribution has not a neat form for reasonable prior distributions. However, on a large sample, assuming a weak prior on $\beta$, the posterior distribution can be approximated by a normal distribution. Thus, draw $\beta$ in a normal distribution with $\hat{\beta}$ and $\hat{V}$ as parameters makes sense.

In the general case, where the data set contains $K$ variables with missing values, each variable is imputed according to a multinomial logistic regression given all the others. More precisely, the incomplete data set is firstly randomly imputed. Then, the missing values of the variable $\mathbf{x}_k$ are imputed as explained previously: a value of $\beta_k$ is drawn from the approximated posterior distribution and an imputation according to $\mathbb{P}(X_k|X_{-k}; \theta_k)$ is performed. The next incomplete variable is imputed in the same way given the other variables, and particularly from the new imputed values of $\mathbf{x}_k$. We proceed in this way for all variables and repeat it until convergence, this provides one imputed data set. The procedure is performed $M$ times in parallel to provide $M$ imputed data sets.

Implicitly, the choices of the conditional distributions $\mathbb{P}(X_k|X_{-k}; \theta_k)$ deter-

mine a joint distribution $\mathbb{P}(X_k;\theta)$, in so far as a joint distribution is compatible with these choices [12]. The convergence to the joint distribution is often obtained for a low number of iterations (5 can be sufficient), but [45, p.113] underlines that this number can be higher in some cases. In addition, FCS is more computationally intensive than JM [45, 50]. This is not a practical issue when the data set is small, but it becomes so on a data set of high dimensions. In particular, checking the convergence becomes very difficult.

The imputation using logistic regressions on each variable performs quite well, that is why this method is often used as a benchmark to perform comparative studies [19, 41, 42, 49]. However, the lack of multinomial regression can affect the multiple imputation procedure using this model. Indeed, when separability problems occur [3], or when the number of individuals is smaller than the number of categories [1, p.195], it is not possible to get the estimates of the parameters. In addition, the number of parameters is very large when the number of categories per variable is high, implying overfitting when the number of individuals is small. When the number of categories becomes too large, [45, 47] advise to use a method dedicated to continuous data: the predictive mean matching (PMM). PMM treats each variable as continuous variables, predicts them using linear regression, and draws one individual among those the nearest to the predicted value. However, PMM often yields to biased estimates [49].

Typically, the default models selected for each logistic regression are main effects models. Thus, the imputation model captures the two-way associations between variables well [1, 49], which is generally sufficient for the analysis model. However, models taking into account interactions can be used but the choice of these models requires a certain effort by the user. To overcome this effort, in particular when the variables are numerous, conditional imputations using random forests instead of logistic regression have been proposed [19, 41]. According to [19], an imputation of one variable $X_k$ given the others is obtained as follows:

- build a forest of 10 trees :
  - draw 10 bootstrap samples from the individuals without missing value on $X_k$;
  - fit one tree on each bootstrap sample: draw randomly a subset of $\sqrt{K-1}$ variables among the $K-1$ explanatory variables. Build one tree from this subset of explanatory variables and this bootstrap sample. Note that the uncertainty due to missing values is reflected by the use of one random forest instead of a unique tree;
- impute missing values:
  - for an individual $i$ with a missing value on $X_k$, gather all the donors from the 10 predictive leaves from each tree and draw randomly one donor from it.
  - repeat for all individuals with missing values on $X_k$

Then, the procedure is performed for each incomplete variable and repeated until convergence. Using random forests as conditional models allows capturing complex relationships between variables. In addition, the method is very robust to the number of trees used, as well as to the number of explanatory variables retained. Thus, the default choices for these parameters (10 trees, $\sqrt{K-1}$ explanatory variables) are very suitable in most of the cases. However, the method is more computationally intensive than the one based on logistic regressions.

# 3 Multiple Imputation using multiple correspondence analysis

This section deals with a novel MI method for categorical data based on multiple correspondence analysis (MCA) [25, 32], *i.e.* the principal components method dedicated for categorical data. Like the imputation using the normal distribution, it is a JM method based on the imputation of the disjunctive table. We first introduce MCA as a specific singular value decomposition on specific matrices. Then, we present how to perform this SVD with missing values and how it is used to perform single imputation. We explain how to introduce uncertainty to obtain a proper MI method. Finally, the properties of the method are discussed and the differences with MI using the normal distribution highlighted.

## 3.1 MCA for complete data

MCA is a principal components method to describe, summarise and visualise multidimensional matrices with categorical data. This powerful method allows us to understand the two-way associations between variables as well as the similarities between individuals. Like any principal components method, MCA is a method of dimensionality reduction consisting in searching for a subspace of dimension $S$ providing the best representation of the data in the sense that it maximises the variability of the projected points (*i.e.* the individuals or the variables according to the space considered). The subspace can be obtained by performing a specific singular value decomposition (SVD) on the disjunctive table.

More precisely, let $\mathbf{Z}_{I\times J}$ denote the disjunctive table corresponding to $\mathbf{X}_{I\times K}$. We define a metric between individuals through the diagonal matrix $\frac{1}{K}\mathbf{D}_\Sigma^{-1}$ where
$\mathbf{D}_\Sigma = \mathbf{diag}\left(\mathbf{p_1^{x_1}}, \ldots, \mathbf{p_{q_1}^{x_1}}, \ldots, \mathbf{p_1^{x_K}}, \ldots, \mathbf{p_{q_K}^{x_K}}\right)$ is a diagonal matrix with dimensions $J \times J$, $p_\ell^{\mathbf{x}_k}$ is the proportion of observations taking the category $\ell$ on the variable $\mathbf{x}_k$. In this way, two individuals taking different categories for the same variable are more distant from the others when one of them takes a rare category than when both of them take frequent categories. We also define a uniform weighting for the individuals through the diagonal matrix $\frac{1}{I}\mathbb{1}_I$ with $\mathbb{1}_I$ the identity matrix of dimensions $I$. By duality, the matrices $\frac{1}{K}\mathbf{D}_\Sigma^{-1}$ and $\frac{1}{I}\mathbb{1}_I$ define also a weighting and a metric for the space of the categories respec-

tively. MCA consists in searching a matrix $\widehat{\mathbf{Z}}$ with a lower rank $S$ as close as possible to the disjunctive table $\mathbf{Z}$ in the sense defined by these metrics. Let $\mathbf{M}_{I\times J}$ denote the matrix where each row is equal to the vector of the means of each column of $\mathbf{Z}$. MCA consists in performing the SVD of the matrix triplet $\left(\mathbf{Z}-\mathbf{M}, \frac{1}{K}\mathbf{D}_\Sigma^{-1}, \frac{1}{I}\mathbb{1}_I\right)$ [24] which is equivalent to writing $(\mathbf{Z}-\mathbf{M})$ as

$$\mathbf{Z}-\mathbf{M}=\mathbf{U}\mathbf{\Lambda}^{\mathbf{1/2}}\mathbf{V}^\top \tag{14}$$

where the columns of $\mathbf{U}_{I\times J}$ are the left singular vectors satisfying the relationship
$\mathbf{U}^\top\mathbf{diag}(\mathbf{1/I},\ldots,\mathbf{1/I})\mathbf{U}=\mathbb{1}_\mathbf{J}$; columns of $\mathbf{V}_{J\times J}$ are the right singular vectors satisfying the relationship $\mathbf{V}^\top\frac{1}{K}\mathbf{D}_\Sigma^{-1}\mathbf{V}=\mathbb{1}_J$ and
$\mathbf{\Lambda}^{\mathbf{1/2}}_{\mathbf{J}\times\mathbf{J}}=\mathbf{diag}\left(\lambda_\mathbf{1}^{\mathbf{1/2}},\ldots,\lambda_\mathbf{J}^{\mathbf{1/2}}\right)$ is the diagonal matrix of the singular values.

The $S$ first principal components are given by $\widehat{\mathbf{U}}_{I\times S}\widehat{\mathbf{\Lambda}}^{1/2}_{S\times S}$, the product between the first columns of $\mathbf{U}$ and the diagonal matrix $\mathbf{\Lambda}^{\mathbf{1/2}}$ restricted to its $S$ first elements. In the same way, the $S$ first loadings are given by $\widehat{\mathbf{V}}_{J\times S}$. $\widehat{\mathbf{Z}}$ defined by:

$$\widehat{\mathbf{Z}}=\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}^\top+\mathbf{M} \tag{15}$$

is the best approximation of $\mathbf{Z}$, in the sense of the metrics, with the constraint of rank $S$ (Eckart-Young theorem [21]). Equation (15) is called *reconstruction formula*.

Note that, contrary to $\mathbf{Z}$, $\widehat{\mathbf{Z}}$ is a fuzzy disjunctive table in the sense that its cells are real numbers and not only zeros and ones as in a classic disjunctive table. However, the sum per variable is still equal to one [44]. Most of the values are contained in the interval $[0,1]$ or close to it because $\widehat{\mathbf{Z}}$ is as close as possible to $\mathbf{Z}$ which contains only zeros and ones, but values out of this interval can occur.

Performing MCA requires $J-K$ parameters corresponding to the terms useful for the centering and the weighting of the categories, $IS-S-\frac{S(S+1)}{2}$ for the centered and orthonormal left singular vectors and $(J-K)S-S-\frac{S(S+1)}{2}$ for the orthonormal right singular vectors, for a total of $J-K+S\left(I-1+(J-K)-S\right)$ independent parameters. This number of parameters increases linearly with the number of cells in the data set.

## 3.2 Single imputation using MCA

[29] proposed an iterative algorithm called "iterative MCA" to perform single imputation using MCA. The main steps of the algorithm are as follows:

1. initialization $\ell=0$: recode $\mathbf{X}$ as disjunctive table $\mathbf{Z}$, substitute missing values by initial values (the proportions) and calculate $\mathbf{M}^0$ and $\mathbf{D}_\Sigma^0$ on this completed data set.

2. step $\ell$:

(a) perform the MCA, in other words the SVD of $\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$ to obtain $\hat{\mathbf{U}}^{\ell}$, $\hat{\mathbf{V}}^{\ell}$ and $\left(\hat{\mathbf{\Lambda}}^{\ell}\right)^{1/2}$;

(b) keep the $S$ first dimensions and use the reconstruction formula (15) to compute the fitted matrix:

$$\hat{\mathbf{Z}}^{\ell}_{I\times J} = \left(\hat{\mathbf{U}}^{\ell}_{I\times S}\left(\hat{\mathbf{\Lambda}}^{\ell}_{S\times S}\right)^{1/2}\left(\hat{\mathbf{V}}^{\ell}_{J\times S}\right)^{\top}\right) + \mathbf{M}^{\ell-1}_{I\times J} \tag{16}$$

and the new imputed data set becomes $\mathbf{Z}^{\ell} = \mathbf{W} * \mathbf{Z} + (\mathbb{1} - \mathbf{W}) * \hat{\mathbf{Z}}^{\ell}$ with $*$ being the Hadamard product, $\mathbb{1}_{I\times J}$ being a matrix with only ones and $\mathbf{W}$ a weighting matrix where $w_{ij} = 0$ if $z_{ij}$ is missing and $w_{ij} = 1$ otherwise. The observed values are the same but the missing ones are replaced by the fitted values;

(c) from the new completed matrix $\mathbf{Z}^{\ell}$, $\mathbf{D}^{\ell}_{\Sigma}$ and $\mathbf{M}^{\ell}$ are updated.

3. steps (2.a), (2.b) and (2.c) are repeated until the change in the imputed matrix falls below a predefined threshold $\sum_{ij}(\hat{z}^{\ell-1}_{ij} - \hat{z}^{\ell}_{ij})^2 \leq \varepsilon$, with $\varepsilon$ equals to $10^{-6}$ for example.

The iterative MCA algorithm consists in recoding the incomplete data set as an incomplete disjunctive table, randomly imputing the missing values, estimating the principal components and loadings from the completed matrix and then, using these estimates to impute missing values according to the reconstruction formula (15). The steps of estimation and imputation are repeated until convergence, leading to an imputation of the disjunctive table, as well as to an estimate of the MCA parameters.

The algorithm can suffer from overfitting issues, when missing values are numerous, when the relationships between variables are weak, or when the number of observations is low. To overcome these issues, a regularized version of it has been proposed [29]. The rationale is to remove the noise in order to avoid instabilities in the prediction by replacing the singular values $\left(\sqrt{\hat{\lambda}^{\ell}_s}\right)_{1\leq s\leq S}$ of step (2.b) by *shrunk* singular values $\left(\frac{\hat{\lambda}^{\ell}_s - \sum_{s=S+1}^{J-K}\frac{\lambda_s}{J-K-S}}{\sqrt{\hat{\lambda}^{\ell}_s}}\right)_{1\leq s\leq S}$. In this way, singular values are shrunk with a greater amount of shrinkage for the smallest ones. Thus, the first dimensions of variability take a more significant part in the reconstruction of the data than the others. Assuming that the first dimensions of variability are made of information and noise, whereas the last ones are made of noise only, this behaviour is then satisfactory. Geometrically, the regularization makes the individual closer to the center of gravity. Concerning the cells of $\widehat{\mathbf{Z}}$, the regularization makes the values closer to the mean proportions and consequently, these values are more often in the interval $[0, 1]$.

The regularized iterative MCA algorithm enables us to impute an incomplete disjunctive table but not an initial incomplete data set. A strategy to go from the imputed disjunctive table to an imputed categorical data set is required. We also suggest the use of the coin flipping approach. Let us note that for each set of dummy variables coding for one categorical variable, the sum per row is equal to one, even if it contains imputed values. Moreover, most of the imputed cells are in the interval $[0, 1]$ or are close to it. Consequently, modifications of these cells are not often required.

## 3.3 MI using MCA

To perform MI using MCA, we need to reflect the uncertainty concerning the principal components and loadings. To do so, we use a non-parametric bootstrap approach based on the specificities of MCA. Indeed, as seen in Section 3.1, MCA enables us to assign a weight to each individual. This possibility to include a weight for the individual is very useful when the same lines of the data set occur several times. Instead of storing each replicate, a weight proportional to the number of occurrences of each line can be used, allowing the storage only of the lines that are different. Thus, a non-parametric bootstrap, such as the one used for the MI using the normal distribution, can easily be performed simply by modifying the weight of the individuals: if an individual does not belong to the bootstrap replicate, then its weight is null, otherwise, its weight is proportional to the number of times the observation occurs in the replicate. Note that individuals with a weight equal to zero are classically called *supplementary individuals* in the MCA framework [24].

Thus, we define a MI method called multiple imputation using multiple correspondence analysis (MIMCA). First, the algorithm consists in drawing $M$ sets of weights for the individuals. Then, $M$ single imputations are performed: at first, the regularized iterative MCA algorithm is used to impute the incomplete disjunctive table using the previous weighting for the individuals; Next, coin flipping is used to obtain categorical data and mimic the distribution of the categorical data. At the end, $M$ imputed data sets are obtained and any statistical method can be applied on each one. In detail, the MIMCA algorithm is written as follows:

1. Reflect the variability on the set of parameters of the imputation model: draw $I$ values with replacement in $\{1, .., I\}$ and define a weight $r_i$ for each individual proportional to the number of times the individual $i$ is drawn.

2. Impute the disjunctive table according to the previous weighting:

   (a) initialization $\ell = 0$: recode $\mathbf{X}$ as a disjunctive table $\mathbf{Z}$, substitute missing values by initial values (the proportions) and calculate $\mathbf{M}^0$ and $\mathbf{D}_\Sigma^0$ on this completed data set.

   (b) step $\ell$:

i. perform the SVD of

$$\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1}, \mathbf{diag}\left(r_1, \ldots, r_I\right)\right)$$

to obtain $\hat{\mathbf{U}}^{\ell}$, $\hat{\mathbf{V}}^{\ell}$ and $\left(\hat{\mathbf{\Lambda}}^{\ell}\right)^{1/2}$;

ii. keep the $S$ first dimensions and compute the fitted matrix:

$$\hat{\mathbf{Z}}^{\ell} = \left(\hat{\mathbf{U}}^{\ell}\left(\hat{\mathbf{\Lambda}}^{\ell}_{shrunk}\right)^{1/2}\left(\hat{\mathbf{V}}^{\ell}\right)^{\top}\right) + \mathbf{M}^{\ell-1}$$

where $\left(\hat{\mathbf{\Lambda}}^{\ell}_{shrunk}\right)^{1/2}$ is the diagonal matrix containing the shrunk singular values and derive the new imputed data set $\mathbf{Z}^{\ell} = \mathbf{W} * \mathbf{Z} + (\mathbf{1} - \mathbf{W}) * \hat{\mathbf{Z}}^{\ell}$

iii. from the new completed matrix $\mathbf{Z}^{\ell}$, $\mathbf{D}_{\Sigma}^{\ell}$ and $\mathbf{M}^{\ell}$ are updated.

(c) step (2.b) is repeated until convergence.

3. Mimic the distribution of the categorical data set using coin flipping on $\mathbf{Z}^{\ell}$ :

   (a) if necessary, modify suitably the values of $\mathbf{Z}^{\ell}$: negative values are replaced by zero, and values higher than one are replaced by one. Then, for each set of dummy variables coding for one categorical variable, scale in order to verify the constraint that the sum is equal to one.

   (b) for imputed cells coding for one missing value, draw one category according to a multinomial distribution.

4. Create $M$ imputed data sets: for $m$ from 1 to $M$ alternate steps 1, 2 and 3.

## 3.4 Properties of the imputation method

MI using MCA is part of the family of joint modelling MI methods, which means that it avoids the runtime issues of conditional modelling. Most of the properties of the MIMCA method are directly linked to MCA properties. MCA provides an efficient summary of the two-way associations between variables, as well as the similarities between individuals. The imputation benefits from these properties and provides an imputation model sufficiently complex to apply then an analysis model focusing on two-way associations between variables, such as a main effects logistic regression model. In addition, like the MI using the normal distribution, MIMCA uses draws from a multinomial distribution with parameter $\theta$ (obtained by the disjunctive table) specific to each individual and depending on the observed values of the other variables. Lastly, because of the relatively small number of parameters required to perform MCA, the

imputation method works well even if the number of individuals is small. These properties have been highlighted in previous works on imputation using principal components methods [7, 8].

Since these two methods, MIMCA and the multiple imputation with the normal distribution, provide several imputations of the disjunctive table, and then use the same strategy to go from the disjunctive table to the categorical data set, they seem very close. However, they differ on many other points.

The first one is that the imputation of the disjunctive table by MCA is a deterministic imputation, replacing a missing value by the most plausible value given by the estimate of the principal components and the estimate of the loadings. Then, coin flipping is used to mimic the distribution of the categorical data. On the contrary, the multiple imputation based on the normal distribution uses stochastic regressions to impute the disjunctive table, that is to say, a Gaussian noise is added to the conditional expectation given by the observed values. Then, coin flipping is used, adding uncertainty a second time.

The second difference between the two methods is the covariance of the imputed values. Indeed, the matrix $\widehat{\mathbf{Z}}^{\ell}$ contains the reconstructed data by the iterative MCA algorithm and the product $\widehat{\mathbf{Z}}^{\ell^\top}\widehat{\mathbf{Z}}^{\ell}$ provides the covariance matrix of this data. The rank of it is $S$. On the contrary, the rank of the covariance matrix used to perform imputation using the normal distribution is $J - K$ (because of the constraint of the sum equal to one per variable). Consequently, the relationships between imputed variables are different.

The third difference is the number of estimated parameters. Indeed, although the imputation by the normal distribution requires a extremely large number of parameters when the number of categories increases, the imputation using MCA requires a number of parameters linearly dependent to the number of cells. This property is essential from a practical point of view because it makes it very easy to impute data sets with a small number of individuals.

# 4 Simulation study

As mentioned in the introduction, the aim of MI methods is to obtain an inference on a quantity of interest $\psi$. Here, we focus on the parameters of a logistic regression without interaction, which is a statistical method frequently used for categorical data. At first, we present how to make inference for the parameters from multiple imputed data sets. Then, we explain how we assess the quality of the inference built, that is to say, the quality of the MI methods. Finally, the MI methods presented in Sections 2 and 3 are compared through a simulation study based on real data sets. It thus provides more realistic performances from a practical point of view. The code to reproduce all the simulations with the R software [36], as well as the data sets used, are available on the webpage of the first author.

## 4.1 Inference from imputed data sets

Each MI method gives $M$ imputed data sets as outputs. Then, the parameters of the analysis model (for instance the logistic regression) as well as their associated variance are estimated from each one. We denote $\left(\widehat{\psi}_m\right)_{1\leq m\leq M}$ the set of the $M$ estimates of the model's parameters and we denote $\left(\widehat{Var}\left(\widehat{\psi}_m\right)\right)_{1\leq m\leq M}$ the set of the $M$ associated variances. These estimates have to be pooled to provide a unique estimate of $\psi$ and of its variance using Rubin's rules [38].

This methodology is explained for a scalar quantity of interest $\psi$. The extension to a vector is straightforward, proceeding in the same way element by element. The estimate of $\psi$ is simply given by the mean over the $M$ estimates obtained from each imputed data set:

$$\hat{\psi} \quad = \quad \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_m, \tag{17}$$

while the estimate of the variance of $\hat{\psi}$ is the sum of two terms:

$$\begin{aligned}\widehat{Var}(\hat{\psi}) =& \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\psi}_m\right) \\ &+ \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\psi}_m-\hat{\psi}\right)^2.\end{aligned} \tag{18}$$

The first term is the within-imputation variance, corresponding to the sampling variance. The second one is the between-imputation variance, corresponding to the variance due to missing values. The factor $(1+\frac{1}{M})$ is due to the fact that $\widehat{\psi}$ is estimated from a finite number of imputed tables.

Then, the 95% confidence interval is calculated as:

$$\hat{\psi} \pm t_{\nu,.975}\sqrt{\widehat{Var}(\hat{\psi})}$$

where $t_{\nu,.975}$ is the .975 critical value of the Student's $t-$distribution with $\nu$ degrees of freedom estimated as suggested by [9].

## 4.2 Simulation design from real data sets

The validity of MI methods are often assessed by simulation [45, p.47]. We design a simulation study using real data sets to assess the quality of the MIMCA method. Each data set is considered as a population data and denoted $\mathbf{X}_{pop}$. The parameters of the logistic regression model are estimated from this population data and they are considered as the true coefficients $\psi$. Then, a sample $\mathbf{X}$ is drawn from the population. This step reflects the sampling variance. The values of the response variable of the logistic model are drawn according to the probabilities defined by $\psi$. Then, incomplete data are generated completely at

random to reflect the variance due to missing values [14]. The MI methods are applied and the inferences are performed. This procedure is repeated $T$ times.

The performances of a MI method are measured according to three criteria: the bias given by $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\psi}_t - \psi\right)$, the median (over the $T$ simulations) of the confidence intervals width as well as the coverage. This latter is calculated as the percentage of cases where the true value $\psi$ is within its 95% confidence interval.

A coverage sufficiently close to the nominal level is required to consider that the inference is correct, but it is not sufficient, the confidence interval width should be as small as possible.

To appreciate the value of the bias and of the width of the confidence interval, it is useful to compare them to those obtained from two other methods. The first one consists in calculating the criteria for the data sets without missing values, which we named the "Full data" method. The second one is the listwise deletion. This consists in deleting the individuals with missing values. Because the estimates of the parameters of the model are obtained from a subsample, the confidence intervals obtained should be larger than those obtained from multiple imputation.

A single imputation method (named *Sample*) is added as a benchmark to understand better how MI methods benefit from using the relationships between variables to impute the data. This single imputation method consists in drawing each category according to a multinomial distribution $\mathcal{M}(\theta, 1)$, with $\theta$ defined according to the proportion of each category of the current variable.

## 4.3 Results

The methods described in this paper are performed using the following R packages: *cat* [26] for MI using the saturated loglinear model, *Amelia* [27, 28] for MI using a normal distribution, *mi* [22] for MI using the DPMPM method, *mice* [47, 48] for the FCS approach using iterated logistic regressions and random forests. This package will also be used to pool the results from the imputed data sets. The tuning parameters of each MI method are chosen according to their default values implemented in the R packages. Firstly, the tuning parameter of the MIMCA method, that is to say, the number of components, is chosen to provide accurate inferences. Its choice will be discussed later in Section 4.3.3.

The MI methods are assessed in terms of the quality of the inference as well as the time consumed from data sets covering many situations. The data sets differ in terms of the number of individuals, the number of variables, the number of categories per variable, the relationships between variables.

The evaluation is based on the following categorical data sets. For each data set a categorical response variable is available.

- *Saheart*: This data set [37] provides clinical attributes of $I_{pop} = 462$ males of the Western Cape in South Africa. These attributes can explain the presence of a coronary heart disease. The data set contains $K = 10$ variables with a number of categories between 2 and 4.

- *Galetas*: This data set [6] refers to the preferences of $I_{pop} = 1192$ judges regarding 11 cakes in terms of global appreciation and in terms of color aspect. The data set contains $K = 4$ variables with two that have 11 categories.

- *Sbp*: The $I_{pop} = 500$ subjects of this data set are described by clinical covariates explaining their blood pressure [23]. The data set contains $K = 18$ variables that have 2 to 4 categories.

- *Income*: This data set, from the R package *kernlab* [31], contains $I_{pop} = 6876$ individuals described by several demographic attributes that could explain the annual income of an household. The data set contains $K = 14$ variables with a number of categories between 2 and 9.

- *Titanic*: This data set [16] provides information on $I_{pop} = 2201$ passengers on the ocean liner *Titanic*. The $K = 4$ variables deal with the economic status, the sex, the age and the survival of the passengers. The first variable has four categories, while the other ones have two categories. The data set is available in the R software.

- *Credit*: German Credit Data from the UCI Repository of Machine Learning Database [33] contains $I_{pop} = 982$ clients described by several attributes which enable the bank to classify themselves as good or bad credit risk. The data set contains $K = 20$ variables with a number of categories between 2 and 4.

The simulation design is performed for $T = 200$ simulations and 20% of missing values generated completely at random.The MI methods are performed with $M = 5$ imputed data sets which is usually enough [38].

### 4.3.1 Assessment of the inferences

First of all, we can note that some methods cannot be applied on all the data sets. As explained previously, MI using the loglinear model can be applied only on data sets with a small number of categories such as *Titanic* or *Galetas*. MI using the normal distribution encounters inversion issues when the number of individuals is small compared to the number of variables. That is why no results are provided for MI using the normal distribution on the data sets *Credit* and *Sbp*. The others MI methods can be applied on all the data sets.

For each data set and each method, the coverages of all the confidence intervals of the parameters of the model are calculated from $T$ simulations (see Table 2 in Appendix A for more details on these models). All the coverages are summarized with a boxplot (see Figure 2). The results for the bias and the confidence interval width are presented in Figure 4 and 5 in Appendix B.

As expected, MI using the loglinear model performs well on the two data sets where it can be applied. The coverages are close to the nominal levels, the biases are close to zero, and the confidence interval widths are small.

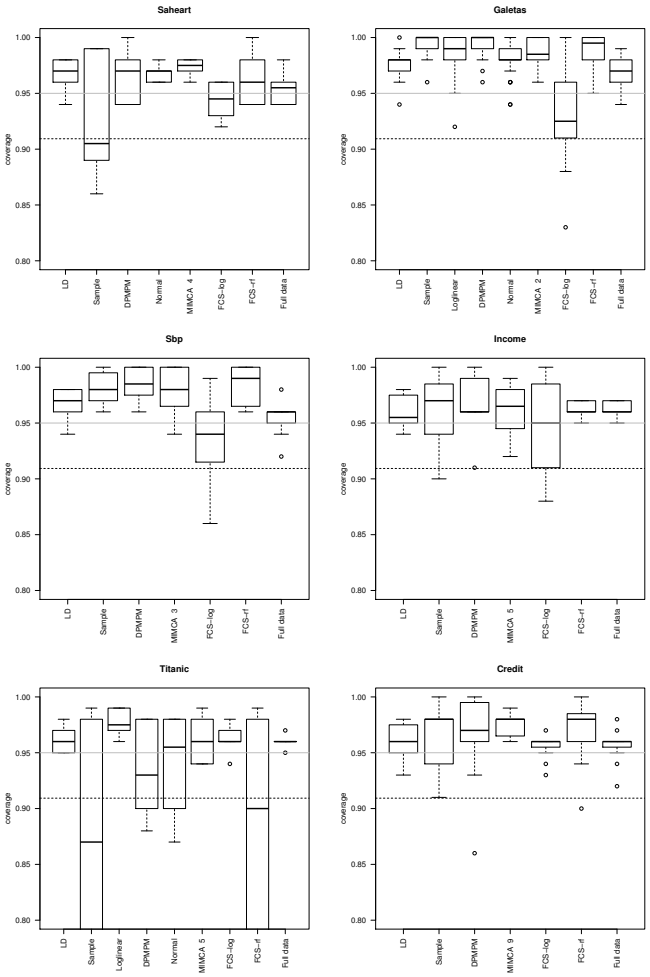


**Figure 2:** Distribution of the coverages of the confidence intervals for all the parameters, for several methods (Listwise deletion, Sample, Loglinear model, Normal distribution, DPMPM, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The horizontal dashed line corresponds to the lower bound of the 95% confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method [2]. Coverages under this line are considered as undesirable.

MI using the non-parametric version of the latent class model performs quite well since most of the quantities of interest have a coverage close to 95%. However, some inferences are incorrect from time to time such as on the data set *Credit* or *Titanic*. This behaviour is in agreement with the study of [42] which also presents some unsatisfactory coverages. [51] note that this MI model can have some difficulties in capturing the associations among the variables, particularly when the number of variables is high or the relationships between variables are complex, that can explain the poor coverages observed. Indeed, on the data set *Credit*, the number of variables is the highest among the data sets considered, while on the data set *Titanic*, the relationships between variables can be described as complex, in the sense that the survival status of the passengers is linked to all the other variables, but these are not closely connected. Moreover, the very poor coverages for the method Sample indicates that the imputation model has to take into account these relationships to provide confidence intervals that reach the nominal rate.

MI using the normal distribution can be applied on three data sets only. On these data sets, the coverages can be too small (see *Titanic* in Figure 2). This highlights that despite the fact that this method is still often used in practice to deal with incomplete categorical data, it is not suitable and we do not recommend using such a strategy. However, [39] showed that this method could be used to impute mixed data (*i.e.* with continuous and categorical data) but only continuous variables contain missing values.

The FCS using logistic regressions encounters difficulties on the data sets with a high number of categories such as *Galetas* and *Income*. This high number of categories implies a high number of parameters for each conditional model that may explain the undercoverage on several quantities.

The FCS using random forests performs well and the method encounters difficulties only on the *Titanic* data set. This behaviour can be explained by the step of subsampling variables in the imputation algorithm (Section 2.4), *i.e.*, each tree is built with potentially different variables and with a smaller number than $(K-1)$. In the *Titanic* data set, the number of variables is very small and the relationships between the variables are weak and all the variables are important to predict the survival response. Thus, it introduces too much bias in the individual tree prediction which may explain the poor inference. Even if, in the most practical cases, MI using random forests is very robust to the misspecification of the parameters, on this data set, the inference could be improved in increasing the number of explanatory variables retained for each tree.

Concerning MI using MCA, all the coverages observed are satisfying. The confidence interval width is of the same order of magnitude than the other MI methods. In addition, the method can be applied whatever the number of categories per variables, the number of variables or the number of individuals. Thus, it appears to be the easiest method to use to impute categorical data.

| | Saheart | Galetas | Sbp | Income | Titanic | Credit |
|---|---|---|---|---|---|---|
| Loglinear | NA | 4.597 | NA | NA | 0.740 | NA |
| DPMPM | 20.050 | 17.414 | 56.302 | 143.652 | 10.854 | 24.289 |
| Normal | 0.920 | 0.822 | NA | 26.989 | 0.483 | NA |
| MIMCA | 5.014 | 8.972 | 7.181 | 58.729 | 2.750 | 8.507 |
| FCS log | 20.429 | 38.016 | 53.109 | 881.188 | 4.781 | 56.178 |
| FCS forests | 91.474 | 112.987 | 193.156 | 6329.514 | 265.771 | 461.248 |

**Table 1:** Time consumed (in seconds) to impute data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit), for different methods (Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests). The imputation is done for $M = 5$ data sets. Calculation has been performed on an Intel®Core™2 Duo CPU E7500, running Ubuntu 12.04 LTS equipped with 3 GB ram. Some values are not provided because all methods cannot be performed on each data set.

#### 4.3.2 Computational efficiency

MI methods can be time consuming and the running time of the algorithms could be considered as an important property of a MI method from a practical point of view. Table 1 gathers the times required to impute $M = 5$ times the data sets with 20% of missing values.

First of all, as expected, the FCS method is more time consuming than the others based on a joint model. In particular, for the data set *Income*, where the number of individuals and variables is high, the FCS using random forests requires 6,329 seconds (*i.e.* 1.75 hours), illustrating substantial running time issues. FCS using logistic regressions requires 881 seconds, a time 6 times higher than MI using the latent class model, and 15 times higher than MI method using MCA. Indeed, the number of incomplete variables increases the number of conditional models required, as well as the number of parameters in each of them because more covariates are used. In addition, the time required to estimate its parameters is non-negligible, particularly when the number of individuals is high. Then, MI using the latent class model can be performed in a reasonable time, but this is at least two times higher than the one required for MI using MCA. Thus, the MIMCA method should be particularly recommended to impute data sets of high dimensions.

Having a method which is not too expensive enables the user to produce more than the classical $M = 5$ imputed data sets. This could lead to a more accurate inference.

#### 4.3.3 Choice of the number of dimensions

MCA requires a predefined number of dimensions $S$ which can be chosen by cross-validation [29]. Cross-validation consists in searching the number of dimensions $S$ minimizing an error of prediction. More precisely, missing values

are added completely at random to the data set $\mathbf{X}$. Then, the missing values of the incomplete disjunctive table $\mathbf{Z}$ are predicted using the regularized iterative MCA algorithm. The mean error of prediction is calculated according to $\frac{1}{Card(\mathcal{U})} \sum_{(i,j)\in\mathcal{U}} (z_{ij} - \hat{z}_{ij})^2$, where $\mathcal{U}$ denotes the set of the added missing values. The procedure is repeated $k$ times for a predefined number of dimensions. The number of dimensions retained is the one minimizing the mean of the $k$ mean errors of prediction. This procedure can be used whether the data set contains missing values or not.

To evaluate how the choice of $S$ impacts on the quality of the inferences, we perform the MIMCA algorithm varying the number of dimensions around the one provided by cross-validation. Figure 3 presents how this tuning parameter influences the coverages in the previous study. The impacts on the width of the confidence intervals are reported in Figure 6 and the ones on the bias in Figure 7 in Appendix B.

Except for the data set *Titanic*, the coverages are stable according to the number of dimensions retained. In particular, the number of dimensions suggested by cross-validation provides coverages close to the nominal level of the confidence interval. In the case of the data set *Titanic*, the cross-validation suggests retaining 5 dimensions, which is the choice giving the smallest confidence intervals, while giving coverages close to 95%. But retaining less dimensions leads to worse performances since the covariates are not closely related (Section 4.3.1). Indeed, these covariates can not be well represented within a space of low dimensions. Consequently, a high number of dimensions is required to reflect the useful associations to impute the data. *Titanic* illustrates that underfitting can be problematic. The same comment is made by [50] who advise choosing a number of classes sufficiently high in the case of MI using the latent class model. However, overfitting is less problematic because it increases the variance, but it does not skip the useful information.

# 5 Conclusion

This paper proposes an original MI method to deal with categorical data based on MCA. The principal components and the loadings that are the parameters of the MCA enables the imputation of data. To perform MI, the uncertainty on these parameters is reflected using a non-parametric bootstrap, which results in a specific weighting for the individuals.

From a simulation study based on real data sets, this MI method has been compared to the other main available MI methods for categorical variables. We highlighted the competitiveness of the MIMCA method to provide valid inferences for an analysis model requiring two-way associations (such as logistic regression without interaction, or a homogeneous loglinear model, proportion, odds ratios, etc).

We showed that MIMCA can be applied to various configurations of data. In particular, the method is accurate for a large number of variables, for a large number of categories per variables and when the number of individuals is small.

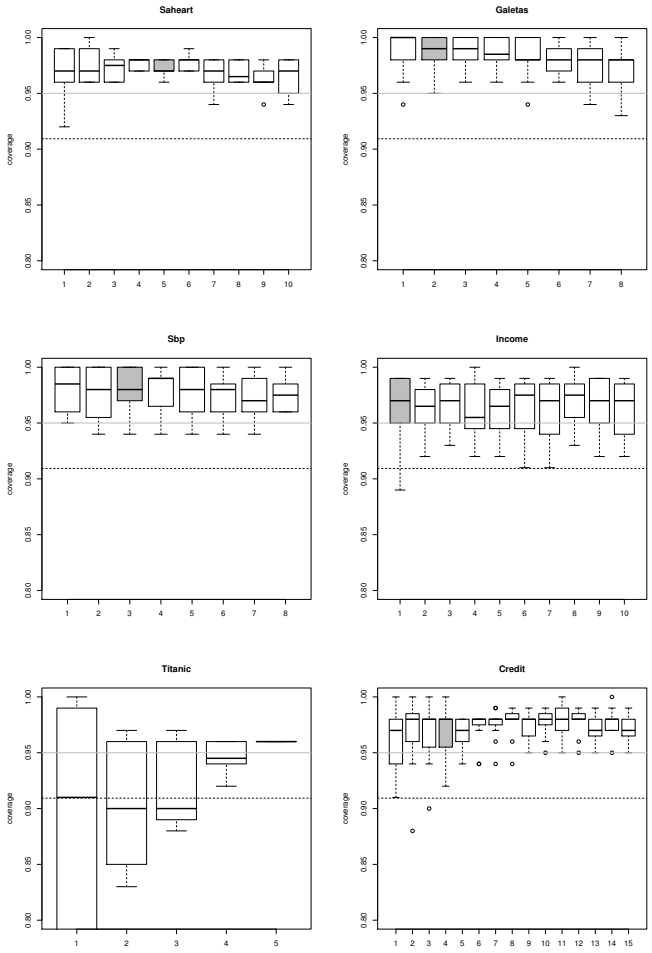


**Figure 3:** Distribution of the coverages of the confidence intervals for all the parameters for the MIMCA algorithm for several numbers of dimensions and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The results for the number of dimensions provided by cross-validation are in grey. The horizontal dashed line corresponds to the lower bound of the 95% confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method [2]. Coverages under this line are considered as undesirable.

Moreover, the MIMCA algorithm performs fairly quickly, allowing the user to generate more imputed data sets and therefore to obtain more accurate inferences ($M$ between 20 and 100 can be beneficial [45, p.49]). Thus, MIMCA is very suitable to impute data sets of high dimensions that require more computation. Note that MIMCA depends on a tuning parameter (the number of components), but we highlighted that the performances of the MI method are robust to a misspecification of it.

Because of the intrinsic properties of MCA, MI using MCA is appropriate when the analysis model contains two-way associations between variables such as logistic regression without interaction. To consider the case with interactions, one solution could be to introduce to the data set additional variables corresponding to the interactions. However, the new variable "interaction" is considered as a variable in itself without taking into account its explicit link with the associated variables. It may lead to imputed values which are not in agreement with each others. This topic is a subject of intensive research for continuous variables [10, 40].

In addition, the encouraging results of the MIMCA to impute categorical data prompt the extension of the method to impute mixed data. The first research in this direction [8] has shown that the principal components method dedicated to mixed data (called Factorial Analysis for Mixed Data) is efficient to perform single imputation, but the extension to a MI method requires further research.

# References

[1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.

[2] A. Agresti and B. A. Coull. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, May 1998.

[3] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[4] P. D. Allison. *Missing Data*. Thousand Oaks, CA: Sage, 2002.

[5] P. D. Allison. Handling missing data by maximum likelihood. In *SAS global forum*, pages 1–21, 2012.

[6] Applied Mathematics Department, Agrocampus Rennes, France. galetas data set. Available on http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/74258_galetas.txt.

[7] V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a Bayesian principal component analysis. *ArXiv e-prints*, January 2014.

[8] V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, pages 1–22, 2014. in press.

[9] J. Barnard and D. B. Rubin. Small Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86:948–955, 1999.

[10] J. W. Bartlett, S. R. Seaman, I. R. White, and J. R. Carpenter. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, 2014.

[11] C. A. Bernaards, T. R. Belin, and J. L. Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6):1368–1382, mar 2007.

[12] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 1974.

[13] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. A Wiley-Interscience publication. Wiley, New York, 1992.

[14] J. P. L. Brand, S. van Buuren, K. Groothuis-Oudshoorn, and E. S. Gelsema. A toolkit in sas for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36–45, 2003.

[15] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[16] R. Dawson and MacG. J. The 'unusual episode' data revisited. *Journal of Statistics Education*, 3, 1995.

[17] H. Demirtas. Rounding strategies for multiply imputed binary data. *Biometrical journal*, 51(4):677–88, 2009.

[18] 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 B*, 39:1–38, 1977.

[19] L. L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014.

[20] D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. 104(487):1042–1051, September 2009.

[21] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

[22] A. Gelman, J. Hill, Y. Su, M. Yajima, M. Grazia Pittau, B. Goodrich, and Y. Si. *mi: Missing Data Imputation and Model Checking*, 2013. R package version 0.9-93.

[23] GlaxoSmithKline, Toronto, Ontario, Canada. Blood pressure data set. Available on http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMainF.htm.

[24] M. J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, London, 1984.

[25] M. J. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, 2006.

[26] T. Harding, F. Tusell, and J. L. Schafer. *cat: Analysis of categorical-variable datasets with missing values*, 2012. R package version 0.0-6.5.

[27] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[28] J. Honaker, G. King, and M. Blackwell. *Amelia II: A Program for Missing Data*, 2014. R package version 1.7.2.

[29] J. Josse, M. Chavent, B. Liquet, and F. Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29:91–116, 2012.

[30] J. Josse and F. Husson. Multiple imputation in PCA. *Advances in data analysis and classification*, 5:231–246, 2011.

[31] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[32] L. Lebart, A. Morineau, and K. M. Werwick. *Multivariate Descriptive Statistical Analysis*. Wiley, New-York, 1984.

[33] M. Lichman. UCI machine learning repository, 2013.

[34] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New-York, 1987, 2002.

[35] X. L. Meng and D. B. Rubin. Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm. *Journal of the American Statistical Association*, 86(416):899–909, December 1991.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[37] J. Rousseauw, J. du Plessis, A. Benade, P. Jordann, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.

[38] D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, 1987.

[39] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.

[40] S. R. Seaman, J. W. Bartlett, and I. R. White. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1):46, 2012.

[41] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6):764–774, March 2014.

[42] Y. Si and J.P. Reiter. Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38:499–521, 2013.

[43] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:805–811, 1987.

[44] M. Tenenhaus and F. W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119, 1985.

[45] S. Van Buuren. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 1 edition, 2012.

[46] S. Van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76:1049–1064, 2006.

[47] S. Van Buuren and C. G. M. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

[48] S. Van Buuren and K. Groothuis-Oudshoorn. *mice*, 2014. R package version 2.22.

[49] D.W. van der Palm, L.A. van der Ark, and J.K. Vermunt. A comparison of incomplete-data methods for categorical data. *Statistical methods in medical research*, 2014. in press.

[50] J. K. Vermunt, J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology, Vol 38*, 38:369–397, 2008.

[51] D. Vidotto, M. C. Kapteijn, and Vermunt J.K. Multiple imputation of missing categorical data using latent class models: State of art. *Psychological Test and Assessment Modeling*, 2014. in press.

[52] R. M. Yucel, Y. He, and A. M. Zaslavsky. Using calibration to improve rounding in imputation. *The American Statistician*, 62:125–129, 2008.

# A Simulation design: analysis models and sample characteristics

| Data set | number of individuals | number of variables | sample size | logistic regression model | number of quantities of interest |
|---|---|---|---|---|---|
| Saheart | 462 | 10 | 300 | CHD = FAMHIST + TOBACCO + ALCOHOL | 30 |
| Galetas | 1192 | 4 | 300 | GALLE = GRUPO | 6 |
| Sbp | 500 | 18 | 200 | SBP = SMOKE + EXERCISE + ALCOHOL | 12 |
| Income | 6876 | 14 | 1500 | INCOME = SEX | 8 |
| Titanic | 2201 | 4 | 300 | SURV = CLASS+AGE+SEX | 6 |
| Credit | 982 | 20 | 300 | CLASS = CHECKING_STATUS + DURATION + CREDIT_HISTORY + PURPOSE | 11 |

**Table 2:** Set of the sample characteristics and of the analysis models used to perform the simulation study (Section 4.2) for the several data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit).

# B Simluation study: complementary results

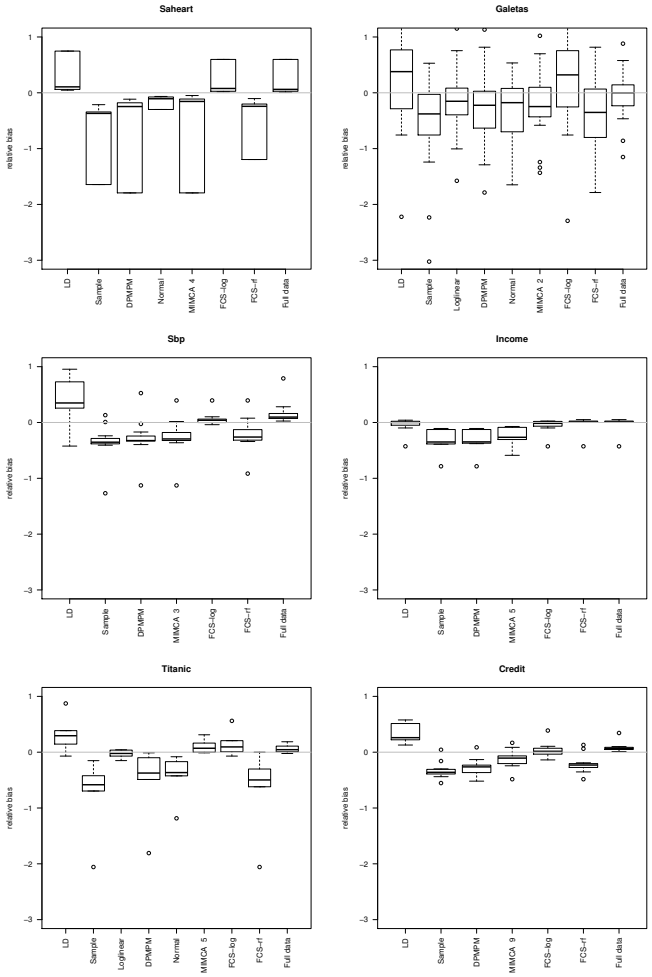


**Figure 4:** Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for several methods (Listwise deletion, Sample, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient.

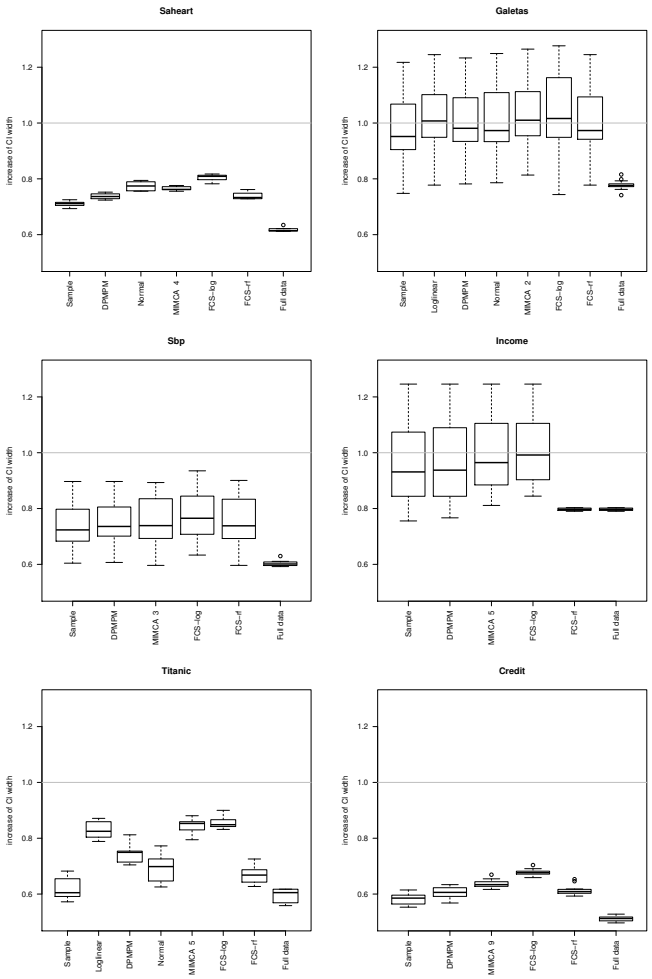


**Figure 5:** Distribution of the median of the confidence interval for the several quantities of interest for several methods (Sample, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median of the confidence interval observed for one coefficient divided by the one obtained by Listwise deletion. The horizontal dashed line corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion.

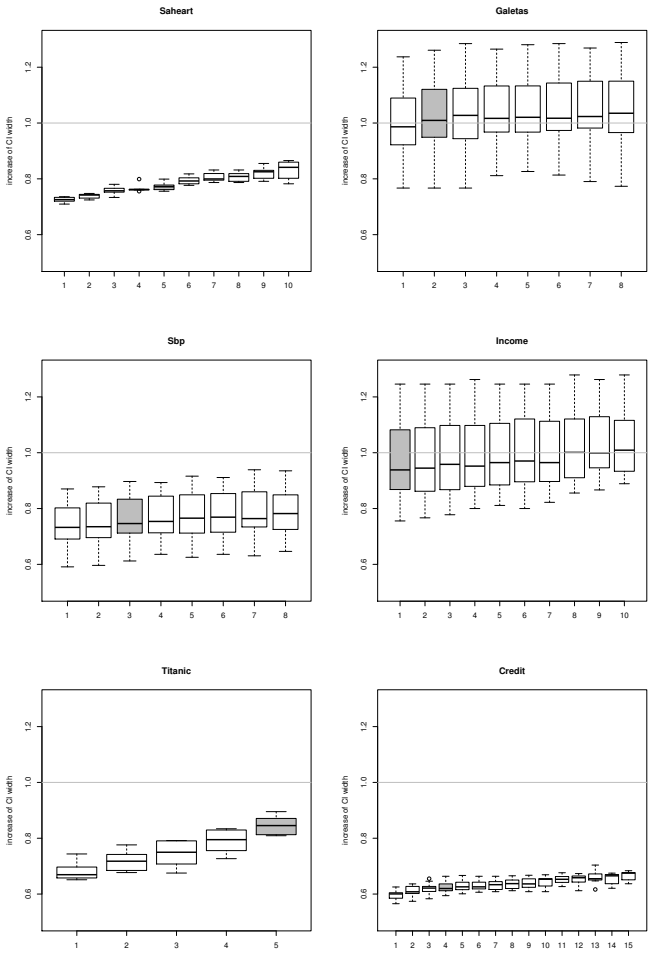


**Figure 6:** Distribution of the median of the confidence interval for the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median of the confidence interval observed for one coefficient divided by the one obtained by Listwise deletion. The horizontal dashed line corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion. The results for the number of dimensions provided by cross-validation are in grey.

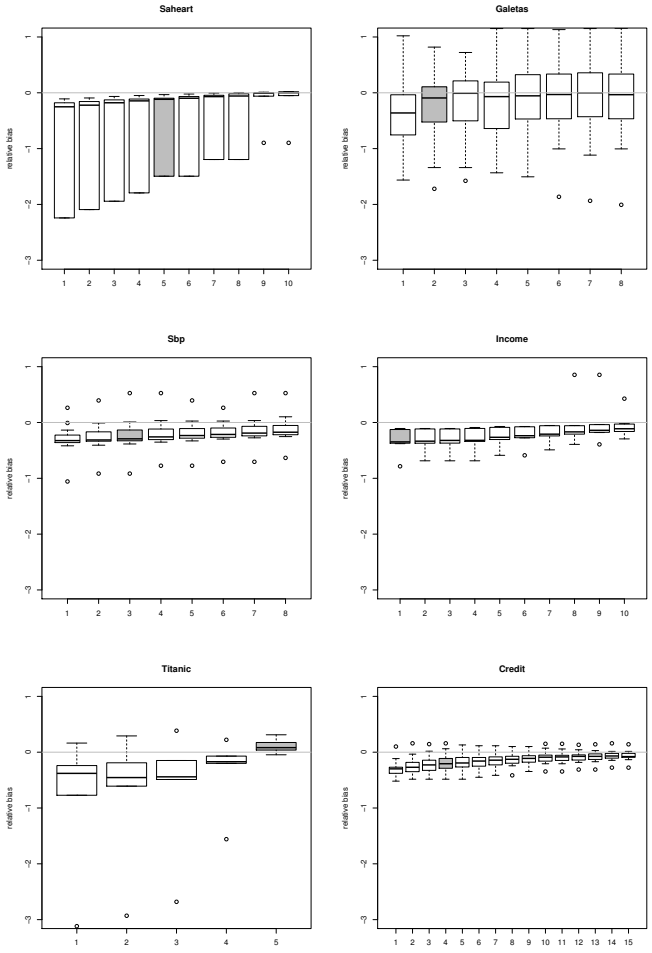


**Figure 7:** Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient. The results for the number of dimensions provided by cross-validation are in grey.

# 7 Compléments : focus sur les interactions

L'imputation multiple par ACM offre une nouvelle méthode d'imputation pour des données qualitatives. L'article précédent illustre bien son intérêt par rapport aux autres méthodes pour la mise en œuvre d'un modèle d'analyse sans interaction. Toutefois, l'imputation par forêts aléatoires, par le modèle à classes latentes, par le modèle log-linéaire et dans une moindre mesure l'imputation par régressions logistiques, sont des méthodes d'imputation qui sont présentées comme apportant une solution à l'inférence en présence de termes d'interaction.

Les propriétés des méthodes d'analyse factorielle exhibées en Chapitre 3 indiquent que ces méthodes ne sont pas adaptées pour effectuer ce type d'inférence.

Afin de mieux comprendre l'incidence de termes d'interaction dans le modèle d'analyse, nous reprenons un des jeux réels de l'étude précédente où un effet d'interaction a été identifié (`Credit`). Nous étudions, selon le protocole de la Section 3.3.2 de l'article, le comportement des méthodes d'imputation pour le modèle de régression incluant cette interaction. Les résultats pour les 15 quantités d'intérêt sont résumés en Figure 1. Notons

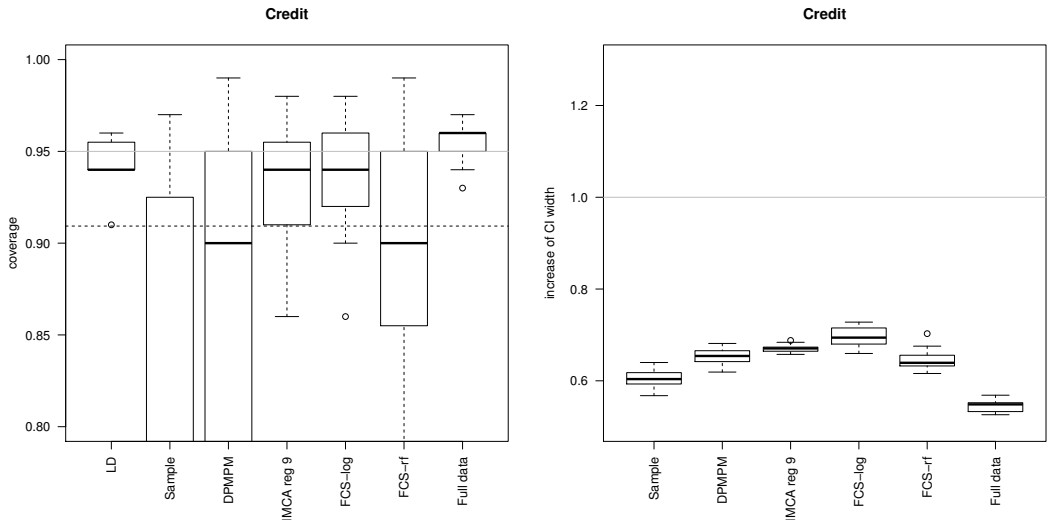


FIGURE 1 – Performance en présence d'effet d'interaction pour 15 quantités d'intérêt pour plusieurs méthodes d'imputation (Suppression par liste, Imputation selon les marginales, modèle log-lineaire, modèle Gaussien, modèle DPMPM, MIMCA, FCS par régression logistique, FCS par forêts aléatoires, Données complète) pour le jeu de données Credit. A gauche : distribution des taux de couverture des intervalles de confiance pour les 200 simulations ; à droite : augmentation de la taille de l'intervalle de confiance par rapport à la suppression par listes.

qu'afin de garantir une stabilité de l'estimateur des coefficients du modèle de régression logistique, la taille de l'échantillon a été augmentée. L'augmentation de la taille de l'échantillon rend l'estimation des intervalles de confiance très sensible à la méthode d'imputation mettant en exergue les difficultés des méthodes. Ainsi, il n'est pas possible de comparer

directement les valeurs des critères avec ceux de l'étude 3.3.2. Les méthodes peuvent seulement être comparées les unes par rapport aux autres.

Le taux de couverture des 15 quantités d'intérêt estimées par la méthode MIMCA est satisfaisant pour les trois quarts d'entre eux, mais un quart indique des taux de couverture non satisfaisant. Les tailles des intervalles de confiance restent bien plus petites que celles d'un intervalle de confiance obtenu par l'analyse du cas complet. Ces résultats sont très corrects par rapport à ceux des autres méthodes, seule l'imputation par régressions logistiques mènent à des taux de couverture plus proches de leur niveau nominal.

Bien que l'ACM ne permettent pas de prendre en compte les effets d'interaction pour imputer, ces performances restent donc globalement raisonnables sur ce jeu réel, et mêmes supérieures à l'imputation par le modèle à classe latentes ou par les forêts aléatoires. En effet, les difficultés observées pour prendre en compte les associations d'ordre 2 par ces méthodes (Section 3.3.2) restent valables ici en présence d'associations d'ordre 2 et 3.

Notons que si les interactions entre variables étaient très marquées, il pourrait être envisagé d'intégrer une variable recodant l'interaction comme suggéré dans l'article. Une autre solution serait de découper le jeu de données (Carpenter et Kenward, 2013) selon une ou plusieurs variables et d'imputer chaque partie indépendamment comme expliqué en introduction. L'ACM semble bien se prêter à un tel découpage car ses paramètres seront toujours estimables si le nombre d'individus est plus faible. Toutefois, cette procédure nécessite que la variable qui sert au découpage soit complète.

# CHAPITRE 6

# CONCLUSION ET PERSPECTIVES

Dans cette thèse nous nous sommes intéressés à l'inférence en présence de données manquantes. Plus précisément, nous nous sommes focalisés sur les méthodes d'imputation multiple. En utilisant les méthodes d'analyse factorielle à des fins d'imputation, nous avons proposé de nouvelles méthodes d'imputation multiple pour des données quantitatives ou qualitatives. Ces méthodes ont été positionnées à la fois par rapport aux méthodes d'imputation de référence et par rapport aux méthodes d'imputation les plus récentes. Elles répondent en partie aux problèmes auxquels sont confrontées les méthodes d'imputation actuelles.

**Des modèles joints** En partant des modèles joints que constituent l'ACP, l'ACM et l'AFDM nous avons proposé des méthodes d'imputation multiple par modèles joints pour des données quantitatives et pour des données qualitatives. Le modèle d'ACP est bien formalisé. La modélisation de Caussinus (1986) exprime la matrice des données comme un signal perturbé par un bruit Gaussien. Ce modèle permet d'établir une méthode d'imputation simple par ACP via l'ajout d'un aléa Gaussien sur la prédiction par ACP. La version Bayésienne de ce même modèle proposée par Verbanck *et al.* (2013) permet de disposer en plus d'une distribution sur les paramètres de l'ACP. L'utilisation de cette dernière modélisation permet de formuler une méthode d'imputation multiple pour des variables quantitatives. Le choix de la loi a priori sur les paramètres est justifié par le fait qu'elle conduit à une loi a posteriori dont l'espérance est donnée par l'estimateur déjà connu d'ACP régularisée (Josse *et al.*, 2009), mais d'autres choix pourraient être envisagés.

L'ACM ne bénéficie pas pour le moment d'une telle modélisation. Ainsi, une stratégie de tirage a été utilisée pour proposer une version stochastique de l'imputation par ACM. La variabilité sur les paramètres a ensuite été obtenue en utilisant une procédure de bootstrap non-paramétrique. Notons que cette stratégie est perfectible. En effet, le tableau disjonctif est bien imputé à partir des relations entre variables, mais le tirage des modalités est effectué de façon indépendante pour un même individu. En procédant ainsi, les relations entre variables peuvent ne pas être parfaitement préservées. Toutefois, ceci n'est vrai que pour plusieurs modalités manquantes simultanément sur des variables liées entre elles.

Pour améliorer la prise en compte des liaisons entre variables, il pourrait être envisagé de transformer le tableau disjonctif imputé en un tableau de Burt, c'est-à-dire en le tableau définissant les effectifs des couples de modalités. De cette façon il est possible de définir des probabilités sur les couples de modalités plutôt que sur les modalités elles-mêmes et donc de tirer conjointement deux modalités pour deux variables incomplètes. Ceci peut aussi être étendu à des $k$-uplets de modalités, mais l'explosion combinatoire limite cette approche.

A fortiori, l'AFDM ne jouit pas non plus d'une modélisation à l'image de l'ACP. L'imputation simple par AFDM ouvre la voie à une méthode d'imputation multiple pour les données mixtes, mais l'extension à une méthode d'imputation multiple n'est pas immédiate. En l'absence de modèle explicite, une telle méthode semblerait nécessiter l'emploi d'une procédure de bootstrap non-paramétrique.

La détermination de modèles explicites pour l'ACM et l'AFDM permettraient d'exploiter davantage les propriétés des méthodes d'analyse factorielle dans une optique d'imputation multiple. L'ACM est assez proche d'un modèle log-linéaire homogène dans la mesure où elle met en évidence les relations entre variables deux à deux tandis que les statistiques suffisantes du modèle log-linéaire sont les effectifs de chaque couple de modalités. Ces liens ont déjà été mis en évidence dans le cadre de l'analyse des correspondances (Greenacre, 1984). L'analyse des correspondances est la méthode d'analyse factorielle dédiée à l'étude d'une paire de variables qualitatives et constitue, en ce sens, un cas particulier de l'ACM. Escoufier (1982) a montré que, quand les effets d'interaction sont faibles, l'analyse des correspondances fournit une estimation des termes d'interaction d'un modèle log-linéaire. Ceci constitue un début de modélisation pour l'ACM qui mérite d'être poursuivi. Dans la même ligne, l'AFDM peut être rapprochée du general location model (Olkin et Tate, 1961). En effet, celui-ci fait l'hypothèse d'un modèle log-lineaire pour les variables qualitatives et d'un modèle Gaussien pour les variables quantitatives conditionnellement aux variables qualitatives. La proximité entre l'ACP et le modèle Gaussien d'une part et entre l'ACM avec le modèle log-lineaire amène assez naturellement à ce rapprochement. L'utilisation des méthodes d'imputation par les méthodes d'analyse factorielle permettrait de gérer l'explosion combinatoire et un nombre de variables élevé qui actuellement limitent le general location model.

**Gestion d'un nombre de variables important** En réduisant la dimension, les méthodes d'imputation multiple proposées sont particulièrement adaptées à des jeux de données où le nombre de variables est important. En contrepartie elles nécessitent de spécifier la dimension du sous-espace considéré. Il s'agit là de l'unique paramètre de réglage de ces méthodes. Une remarque générale concernant ce choix est dictée par le problème de la congénialité : le modèle d'imputation doit reposer sur une distribution au moins aussi générale que celle du modèle d'analyse. Retenir un nombre de dimensions faible revient à faire une hypothèse forte sur les données ce qui peut conduire à modifier les relations entre les variables imputées et donc induire des biais. Au contraire, choisir un nombre de dimensions élevé conduit à capter, en plus de la structure des données, de la variabilité propre à l'échantillon. Cependant, les données imputées étant elles-mêmes sensées refléter des valeurs issues d'un échantillon, cette variabilité supplémentaire a pour principale

conséquence d'augmenter la variabilité des estimateurs, ce qui est moins préjudiciable que l'introduction de biais. Ainsi, si la dimension de la structure n'est pas claire, il est préférable de retenir un nombre de dimensions assez grand. Notons que cette remarque est en phase avec les conseils sur le choix du nombre de classes dans le modèle à classes latentes (Vidotto *et al.*, 2014).

Un nouvel objectif serait de proposer des méthodes d'imputation pour lesquelles le choix du nombre de dimensions serait géré de façon automatique à l'intérieur de l'algorithme, sans avoir à parcourir une grille comme ce qui est fait dans la validation croisée. Josse et Sardy (2015) ont récemment proposé un estimateur pour approcher une matrice de rang inférieur qui ne nécessite pas de fixer ce rang. Cet estimateur a été développé dans le cadre standard sans données manquantes. Il est basé sur une régularisation de l'ensemble des valeurs propres de la SVD. Cette régularisation annule par construction les dernières valeurs propres ce qui implicitement correspond à un choix d'un nombre de dimensions particulier. L'utilisation de cet estimateur pour déterminer les composantes principales et les vecteurs propres constitue une piste de recherche pour un choix automatique du nombre de dimensions en imputation multiple.

Enfin, les méthodes d'imputation multiple par ACP et ACM, en fixant le nombre de dimensions à l'avance, ne reflètent pas l'incertitude sur ce nombre de paramètres. Il serait préférable de prendre celle-ci en compte. Dans le cadre de l'ACP Bayésienne, une possibilité serait de mettre un a priori sur ce nombre. Hoff (2007) a par exemple proposé une modélisation Bayésienne de l'ACP où l'a priori porte sur les matrices de la décomposition en valeurs singulières $\mathbf{U}, \mathbf{\Lambda}, \mathbf{V}$. Cet a priori annule certaines des colonnes de $\mathbf{U}, \mathbf{\Lambda}, \mathbf{V}$ ce qui définit un rang pour le signal. Dans notre cas, une possibilité serait de conserver la loi a priori des paramètres $\tilde{x}_{ij}^{(s)}$, définie conditionnellement au nombre de dimensions, et de définir un a priori supplémentaire pour le nombre de dimensions. Au contraire du modèle de Verbanck *et al.* (2013), la loi a posteriori n'aurait certainement plus de forme explicite. Sous réserve que la loi a posteriori sur le nombre d'axes soit néanmoins connue et puisse être facilement simulée, il pourrait être envisagé d'effectuer des tirages dans cette distribution via l'utilisation d'un algorithme de Gibbs en tirant le nombre de dimensions dans sa loi a posteriori, puis les autres paramètres dans leur loi a posteriori conditionnellement au nombre de dimensions. Toutefois, une telle modélisation reste à définir. En l'absence de modèle pour l'ACM cette incertitude paraît pour le moment difficile à considérer pour des variables qualitatives.

**Prise en compte des interactions** Les méthodes d'imputation multiple par analyse factorielle offrent une certaine réponse à l'inférence en présence de termes d'interaction. Les modèles d'imputation ne sont certes pas adaptés pour gérer ce type de liaisons, mais les conséquences d'un point de vue pratique sont discutables. Considérer les interactions dans l'imputation reste toutefois une piste d'amélioration de ces méthodes d'imputation multiple. Leur faible nombre de paramètres leur permet d'être appliquées sur des jeux de données où le nombre d'individus est faible devant le nombre de variables. Ceci offre la possibilité de découper le jeu de données selon des variables complètement observées. L'extension à des variables incomplètes est également possible dans la mesure où

celles-ci possèdent suffisamment d'individus complets et que la suppression des individus incomplets pour ces variables n'engendre pas de biais dans l'analyse. Si ces conditions sont vérifiées, cette approche est préférable à un recodage des variables, plus complexe à mettre en œuvre (Carpenter et Kenward, 2013).

**Gestion des contraintes** D'autres problèmes restent ouverts, notamment la gestion des contraintes au sein du jeu de données imputé. Un premier type de contraintes est la contrainte de somme. Par exemple, la première variable du jeu de données peut être la somme de deux autres. Supposons que ces deux autres variables soient incomplètes simultanément, alors on attend des valeurs imputées que leur somme soit égale à la donnée observée sur la première variable. Ce type de configuration est fréquent dans les études sur les ménages ou sur les habitudes de consommation. Par exemple, la première variable peut correspondre à une question du type "combien d'argent accordez-vous à l'achat de nourriture ?" et les suivantes à "combien d'argent dépensez-vous pour les fruits et légumes ?", "pour la viande ?", etc. Il est alors fréquent qu'une personne sache répondre à la première question mais pas aux suivantes. Cette problématique suscite un intérêt depuis peu (de Waal *et al.*, 2011; Hron *et al.*, 2010) et reste un sujet de recherche à l'heure actuelle. Les contraintes de type inégalité sont plus facilement gérables, via l'utilisation des lois tronquées (Templeman, 2007; Kim *et al.*, 2014). La simulation d'une loi tronquée peut se faire en itérant les tirages de façon à satisfaire les contraintes. L'imputation multiple par ACP pourrait gérer ces contraintes de cette façon. D'autres contraintes qu'il n'est pas simple à prendre en compte sont les zéros structurels, c'est-à-dire des combinaisons impossibles de certaines valeurs, par exemple être retraité et avoir moins de 25 ans. Le modèle à classes latentes proposé par Si et Reiter (2013) a notamment été récemment adapté à ce type de données (Manrique-Vallier et Reiter, 2014). La gestion des zéros structurels reste un problème à explorer.

**Une limite en terme de pourcentage de données manquantes ?** Ce travail a montré que les méthodes proposées sont applicables sur des jeux de données incomplets très variés. Il n'en demeurent pas moins qu'on ne peut pas attendre d'inférence sans biais et avec une erreur quadratique faible dans certaines configurations. Ceci est lié en partie à la quantité de données manquantes, mais en partie seulement. Prenons l'exemple d'un jeu de données constitué de variables quantitatives, toutes extrêmement corrélées. Dans pareille situation, l'imputation multiple par ACP sera robuste à un taux de données manquantes très élevé car l'information observée est suffisante pour imputer de façon précise les données incomplètes, même si celles-ci sont très nombreuses. En conséquence, l'inférence menée sera voisine de celle que l'on aurait obtenue en l'absence de données manquantes. Au contraire, si les variables sont indépendantes, l'imputation est très aléatoire et l'inférence qui en résulte peut être autant variable que celle obtenue par l'analyse du cas complet. La difficulté à imputer est donc également liée à la structure du jeu de données. Le bénéfice de l'emploi de ces méthodes d'imputation multiple est fonction de la fraction d'information manquante, *i.e.* la part de variance attribuable aux données manquantes. Si l'information manquante est élevée, alors peu d'information sur le paramètre

d'intérêt peut être obtenue des individus incomplets, l'imputation multiple n'améliore pas beaucoup l'inférence par rapport à la méthode du cas complet. Au contraire si elle est faible, ceci indique qu'une information importante peut être exploitée et l'imputation multiple améliore sensiblement l'inférence par rapport à l'analyse du cas complet. Notons que la fraction d'information manquante peut être estimée à partir des variabilités intra-imputation et inter-imputation obtenue à la suite de l'imputation multiple (*cf.* Equation 34 Chapitre 2).

Une structure forte (et donc une information manquante potentiellement faible) est bien prise en compte par les méthodes d'analyse factorielle, alors qu'au contraire elle engendre une instabilité pour les méthodes de régression. Ainsi, l'imputation multiple par ACP et ACM est plus susceptible de tolérer des taux de données manquantes élevés dans ces conditions, mais en l'absence de structure (et donc d'une information manquante potentiellement forte) les valeurs imputées seront nécessairement très variables et les estimateurs des quantités d'intérêt également.

**Au-delà de l'inférence en présence de données manquantes** L'imputation multiple, et ce travail en particulier, ont des utilisations qui vont au-delà de l'inférence sur une quantité d'intérêt en présence de données manquantes. Une première utilisation est la construction d'intervalles de prédiction pour des cellules incomplètes d'un tableau de données. L'imputation multiple par ACP est en particulier une alternative intéressante à la régression en présence de fortes corrélations entre variables. En imputant les données un grand nombre de fois, on peut refléter la variabilité de prédiction des données manquantes et déterminer des intervalles de prédiction pour celles-ci. Ceci serait plus compliqué en utilisant simplement des modèles de régression du fait de la corrélation. Notons que cette approche est aussi un outil de diagnostic pour l'imputation multiple dans son utilisation classique (Honaker *et al.*, 2011) : en ôtant une cellule du tableau, on peut vérifier qu'un intervalle de confiance pour cette cellule a un taux de couverture adéquat, indiquant un ajustement correct du modèle d'imputation.

D'autres utilisations de l'imputation multiple n'ont pas trait aux données manquantes (Reiter et Raghunathan, 2007). L'une d'entre elles est la gestion des outliers définis comme des données entâchées d'une erreur de mesure. En effet, les données manquantes peuvent être vues comme des données dont l'erreur de mesure est tellement grande qu'aucune information ne peut en être extraite. Ces données sont alors remplacées par des valeurs issues de la distribution prédictive des données manquantes, c'est-à-dire remplacées par des données dépourvues d'erreur de mesure. Dans le cas où certaines données sont des outliers (ou sont éventuellement manquantes), celles-ci apportent une information qu'il convient d'exploiter. Dans ce but, les données sont pondérées selon l'erreur de mesure dont elles sont affectées. Le jeu de paramètres du modèle d'imputation est alors déterminé selon ces poids et les outliers sont imputés selon ces paramètres. La méthode statistique souhaitée est ensuite appliquée sur chacun des tableaux et les résultats agrégés. L'imputation multiple par ACP semble bien se prêter à ce type d'utilisation dans la mesure où la modélisation Bayésienne permet d'intégrer des poids via l'introduction de loi a priori (Blackwell *et al.*, 2015). L'imputation multiple par ACM semble également pertinente pour ce type d'utilisation. En effet, l'algorithme d'ACM

itérative utilise une matrice de pondération des observations de façon à affecter un poids nul aux données non observées. On pourrait envisager de modifier cette matrice de façon à attribuer aux données un poids fonction de l'erreur dont elles sont affectées.

Une autre application de ce travail est la protection de la confidentialité (Rubin, 1993). Certaines données mises à la disposition du public ne doivent en effet pas révéler l'identité des répondants ou certaines caractéristiques sensibles comme les revenus. Pour protéger la confidentialité, certaines variables peuvent être supprimées, certaines données agrégées ou modifiées, mais cela altère potentiellement la structure initiale des données et donc les analyses faites par la suite. Dans ce contexte, l'imputation multiple s'avère être un outil particulièrement intéressant (Rubin, 1993) : les données collectées sont échantillonnées de façon à créer le nombre de jeux de données souhaité. Pour chacun de ces jeux, une partie ou la totalité des données sont remplacées par des données générées selon le modèle calibré à partir des données collectées. Cette opération est répétée pour chaque jeu de façon à fournir les versions multiples correspondantes. Disposant de ces versions multiples, le chercheur peut alors appliquer sa méthode statistique sur chacun des tableaux et agréger les estimations obtenues (Reiter et Raghunathan, 2007). Ainsi, dans la mesure où la congénialité est respectée, l'inférence menée correspond bien aux caractéristiques des données collectées. L'imputation multiple par ACP ou par ACM pourraient être également utilisées dans ce but.

CHAPITRE 7

# LISTE DES TRAVAUX

## PUBLICATIONS

Vincent Audigier, François Husson et Julie Josse. *A principal component method to impute missing values for mixed data*, Advances in Data Analysis and Classification, p. 1–22, (2014). A paraître.

Vincent Audigier, François Husson et Julie Josse. *Multiple imputation for continuous variables using a Bayesian principal component analysis*, Journal of Statistical Computation and Simulation (accepté) (2015).

Vincent Audigier, François Husson et Julie Josse. *MIMCA : Multiple imputation for categorical variables with multiple correspondence analysis*, Statistics and Computing (révision mineure) (2015).

## COMMUNICATIONS ORALES

*Le nom de l'orateur est en gras*

Vincent Audigier, Julie Josse and **François Husson**. Missing values imputation for mixed data based on principal component methods. Cyprus, August 27-31th 2012 COMPSTAT.

**Vincent Audigier**, Julie Josse and François Husson. Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA. Bordeaux, France July 2-3th 2012. Premières rencontres R.

**Vincent Audigier**, Julie Josse and François Husson. Imputation multiple à l'aide des méthode d'analyse factorielle. Toulouse, France, Mai 27-31th 2013. 45èmes Journées de Statistique.

Vincent Audigier, **Julie Josse** and François Husson. Imputation of mixed data : Random Forests versus PCA. London, United-Kingdom, December 14-16th 2013. ERCIM.

Vincent Audigier, **Julie Josse** and François Husson. Multiple imputation with Bayesian PCA. Bordeaux, France, March 6th 2014. Séminaire de l'institut de mathématiques de Bordeaux.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation with Bayesian PCA. Rennes, France, June 2-6th 2014. 46èmes Journées de Statistique.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation with MCA. Rennes, France, October 23-24th 2014. 46èmes Journées de Statistique de Rennes.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation with MCA. Lille, France, June 1-5th 2015. 47èmes Journées de Statistique.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation for categorical data using MCA. Rennes, France, June 17-19th 2015. missData2015.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation for categorical data using MCA. Naples, Italy, September 21-23th 2015. CARME2015.

**Vincent Audigier**, Julie Josse and François Husson. Multiple imputation with MCA. Nantes, France, October 28-30th 2015. Rencontres doctorales Lebesgue.

# APPENDIX A

# MULTIPLE IMPUTATION WITH PRINCIPAL COMPONENT METHODS: A USER GUIDE

In this chapter, we present how to use the multiple imputation methods described previously: the BayesMIPCA method, allowing multiple imputation of continuous data using PCA and MIMCA, allowing multiple imputation for categorical data using MCA. Both are available in the R package missMDA, which contains also functions to tune the parameters of the MI methods and functions to make diagnostics. These functions are presented through real incomplete data sets. First, we present how to investigate an incomplete data set and point out the utility of principal component methods to achieve this goal. Then, we explain how to perform MI using PCA if the data set is continuous and MI using MCA when data are categorical. Finally, we show how to perform analysis and how to pool the analysis results.

## Contents

# 1 Exploratory analysis of incomplete data

Like in the standard statistical framework without missing values, the exploration of data before performing a statistical method is required. However, exploring incomplete data does not only relate to the observed values, but also to the missing data pattern, and to the mechanism that connects values and pattern as well. Principal component methods offer interesting ways to perform it.

In this part, we will principally use the data set *sleep* from the R package VIM (Templ et al., 2015) as an illustrative example where the missing data mechanism is unknown. From Allison and Chichetti (1976), this data set deals with 62 mammal species on the interrelationship between sleep, ecological, and constitutional variables. The data set contains missing values on five variables. The three last variables are five-point scales.

```
> library(VIM)
> data(sleep, package = "VIM")
> don<-sleep
> summary(don,digits=5)
```

```
    BodyWgt            BrainWgt            NonD             Dream
 Min.   :   0.005   Min.   :   0.14   Min.   : 2.1000   Min.   :0.000
 1st Qu.:   0.600   1st Qu.:   4.25   1st Qu.: 6.2500   1st Qu.:0.900
 Median :   3.342   Median :  17.25   Median : 8.3500   Median :1.800
 Mean   : 198.790   Mean   : 283.13   Mean   : 8.6729   Mean   :1.972
 3rd Qu.:  48.203   3rd Qu.: 166.00   3rd Qu.:11.0000   3rd Qu.:2.550
 Max.   :6654.000   Max.   :5712.00   Max.   :17.9000   Max.   :6.600
                                      NA's   :14        NA's   :12
     Sleep             Span              Gest             Pred
 Min.   : 2.600   Min.   :  2.000   Min.   : 12.00   Min.   :1.000
 1st Qu.: 8.050   1st Qu.:  6.625   1st Qu.: 35.75   1st Qu.:2.000
 Median :10.450   Median : 15.100   Median : 79.00   Median :3.000
 Mean   :10.533   Mean   : 19.878   Mean   :142.35   Mean   :2.871
 3rd Qu.:13.200   3rd Qu.: 27.750   3rd Qu.:207.50   3rd Qu.:4.000
 Max.   :19.900   Max.   :100.000   Max.   :645.00   Max.   :5.000
 NA's   :4        NA's   :4         NA's   :4
      Exp              Danger
 Min.   :1.0000   Min.   :1.0000
 1st Qu.:1.0000   1st Qu.:1.0000
 Median :2.0000   Median :2.0000
 Mean   :2.4194   Mean   :2.6129
 3rd Qu.:4.0000   3rd Qu.:4.0000
 Max.   :5.0000   Max.   :5.0000
```

## 1.1 Missing data pattern

The exploration of incomplete data requires the investigation of the missing data pattern. First, this investigation is important because it can define the method to use to deal with

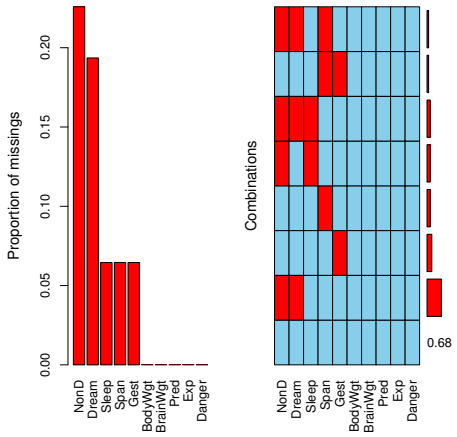


Figure 1: Visualisation of the missing data pattern: Aggregation graphic of the incomplete data set *sleep*

the missing values. Indeed, if the number of missing values is very small compared to the number of individuals, MI is not necessarily required and listwise deletion, which is a simpler method, could be preferred. In addition, although the BayesMIPCA and MIMCA methods are suitable for monotone or non-monotone pattern, it could be interesting for other MI method to identify such a pattern. For instance, chained equations can be tuned to deal with a monotone one (Rubin, 1987). Secondly, the proportion of missing values generally affects the convergence of iterative algorithms used in MI, such as Expectation-Maximisation (EM) algorithm or Data-Augmentation (DA) algorithm. If the proportion of missing values is large, these algorithms can require a larger number of iterations than usual. In addition, more imputed data sets can be required if a large part of the values of the analysis model are missing. Thirdly, the study of the frequencies of combinations of missing values on several variables can highlight a MCAR mechanism. Indeed, if the mechanism is MCAR, missing values occur independently on each incomplete variable, which implies that each combination of missing values is equally probable.

The R package VIM provides many tools to visualise a missing data pattern. The function `aggr` plots the amount of missing values in each variable and the amount of missing values in certain combinations of variables (Figure 1). Missing values occurs on 5 variables through a non-monotone pattern (right-hand side of Figure 1). The percentage of missing values is moderate, but only 68% of the individuals are complete, which justify the use of multiple imputation. The plot of the combinations (Figure 1) indicates that missing values often occur simultaneously on the variables Dream and NonD for instance. Nevertheless, this plot quickly becomes unreadable when the number of incomplete variables is large.

MCA can be straightforwardly used to visualise the missing data pattern even if the number of variable is large. Indeed, the missing data pattern can be viewed as a data set

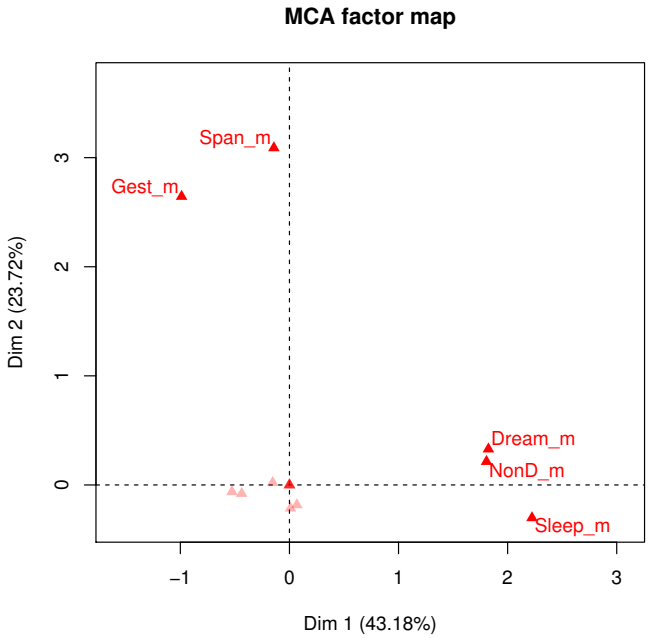


Figure 2: Visualisation of the missing data pattern: Graph of the categories from the MCA on the missing data pattern for the data set *sleep*

where all variables have two categories: *mis*, indicating a missing value and *obs* for an observed one. MCA will highlight associations between pairs of categories by searching the common dimension of variability between the corresponding variables. Thus, it will show if missing values simultaneously occur in several variables or if missing values occur when some other variables are observed. To perform MCA on the missing data pattern and visualise the associations between categories, the following lines can be run:

```
> library(FactoMineR)
> # Creation of a categorical data set with "o" when observed and "m" when missing
> pattern <- matrix("o",nrow=nrow(don),ncol=ncol(don))
> pattern[is.na(don)] <- "m"
> pattern<-as.data.frame(pattern)
> dimnames(pattern) <- dimnames(don)
> # MCA
> res.mca<-MCA(pattern,graph=F)
> plot(res.mca,selectMod=grep("_m",rownames(res.mca$var$coord)),invisible="ind")
```

The graph of the categories is represented in Figure 2. It highlights two groups of categories:

- group 1: Sleep_m, NonD_m, Dream_m
- group 2: Gest_m, Span_m

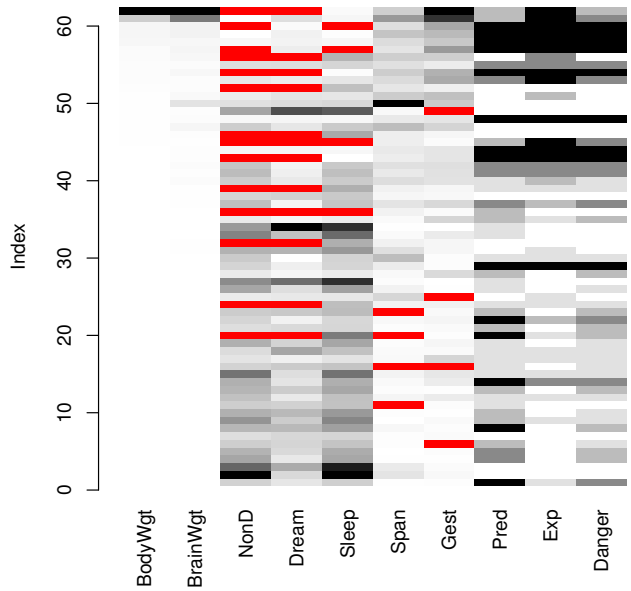


Figure 3: Visualisation of the missing data mechanism: Matrix plot of the incomplete *sleep* data set

The categories of the first group have large coordinates (positive) on the first axis (horizontal) whereas the categories of the second group have large coordinates on the second axis (vertical) only. It means that missing values tend to occur simultaneously on the variables Span and Gest on the one hand, and on the variables Dream, NonD and Sleep on the other hand. Consequently, MCA suggests that the missing data are not missing completely at random. To go further in the understanding of the mechanism, it can be useful to make the link between missing values and observed values.

## 1.2 Missing data mechanism

Even if the MAR assumption cannot be checked, the simultaneous investigation of the missing data pattern and the observed values can allow a better understanding of the missing data mechanism. To obtain some clues about its nature, first, it can be useful to visualise the data matrix (see Figure 3). This representation can be obtained by using the function `matrixplot` from the package VIM: all cells of a data matrix are visualised by rectangles. Available data are coded according to a grey scale scheme, while missing data are in red. By ordering the lines according to one variable, the link between missing values and this variable can be highlighted.

However, as previously, it can be difficult to highlight in this way relationships between missing data pattern and observed values when the number of variables is large. MCA can also be used to point out these relationships, but contrary to the previous case, the

principal component method need to be performed on a data set containing information on observed data and the missing data pattern as well. To take both into account, we consider observed data and include missing values as a category of the variables. Thus, if there is a link between the missing values and the observed values, then MCA can highlight it. Note that MCA is not dedicated to continuous variables. If the data set contains such variables, they need to be split into categories. Continuous variables of the *sleep* data set can be recoded as follows:

```
> don.cat<-as.data.frame(don)
> for(i in 1:7){
+   breaks<-c(-Inf,quantile(don.cat[[i]],na.rm=T)[-1])
+   don.cat[[i]]<-cut(don.cat[[i]],breaks=breaks,labels=F)
+   don.cat[[i]]<-addNA(don.cat[[i]],ifany=T)
+ }
> for(i in 8:10){
+   don.cat[[i]]<-as.factor(don.cat[[i]])
+ }
> summary(don.cat)
```

```
 BodyWgt BrainWgt NonD    Dream   Sleep   Span    Gest    Pred   Exp    Danger
 1:16    1:16     1 :12   1 :14   1 :15   1 :15   1 :15   1:14   1:27   1:19
 2:15    2:15     2 :12   2 :13   2 :14   2 :14   2 :14   2:15   2:13   2:14
 3:15    3:15     3 :14   3 :10   3 :15   3 :14   3 :14   3:12   3: 4   3:10
 4:16    4:16     4 :10   4 :13   4 :14   4 :15   4 :15   4: 7   4: 5   4:10
                  NA:14   NA:12   NA: 4   NA: 4   NA: 4   5:14   5:13   5: 9
```

Note that the first variables are split into 4 categories with the same number of individuals per category, whereas the last ones are only recoded since they are five-point scales.

Then, MCA can be applied on the recoded data set:

```
> res.MCA<-MCA(don.cat,graph=F)
> plot(res.MCA,choix="ind",invisible="ind")
```

The Figure 4 sums up the relationships between observed categories and missing categories. For instance, the category SleepNA has high coordinates on the first axis. Missing values on the variables Sleep can be associated with the observed categories having the largest coordinates on the first axis, such as Exp_5, Danger_5 BodyWgt_4 and BrainWgt_4. Thus, MCA shows a link between missing values on the variables Sleep and the values of the variables Exp, Danger, BodyWgt and BrainWgt. Many other comments could be done from this graph. It would be also interesting to analyse the following dimensions by specifying the argument `axes=c(3,4)`:

```
> plot(res.MCA,choix="ind",invisible="ind",axes=c(3,4))
```

Note that the graph can be sensitive to the splitting. If the number of individuals is large, it can be interesting to increase the number of categories to obtain a finer understanding

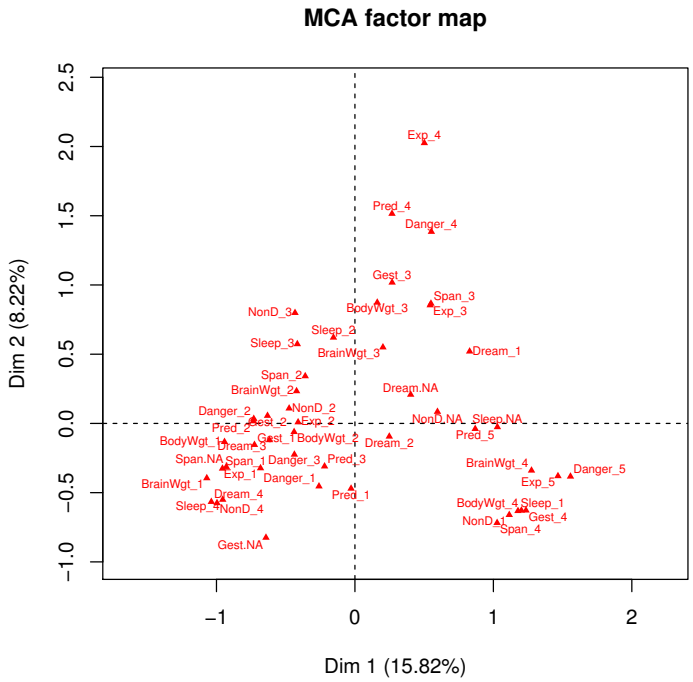


Figure 4: Visualisation of the missing data mechanism: Graph of the categories from MCA performed on the data set *sleep* where continuous data are split into categories and missing values are recoded as a category

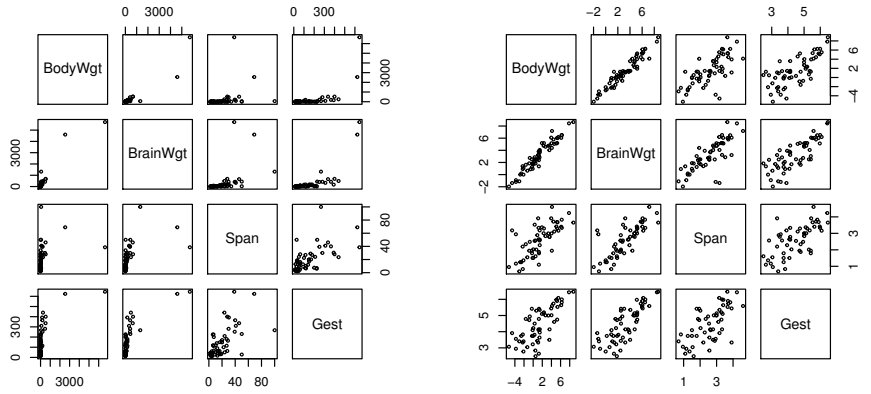


Figure 5: Bivariate plots for four variables of the *sleep* data set. The graph on the left corresponds to the original variables, and the graph on the right to the variables after logarithm transformations.

of the relationships between categories, and therefore of the mechanism. In addition, the splitting is performed according to the quantiles, but other ways could be used. The advantage of quantiles is to avoid rare categories that can have a great influence on the representation.

## 1.3 Observed data

The analysis of the observed values is also important for understanding the data. Principal component methods are powerful tool for this purpose. However, applying principal component methods on an incomplete data set is not straightforward since the most of statistical methods, they cannot be used directly on incomplete data. The package missMDA (Husson and Josse, 2015; Josse and Husson, 2015) enables the use of principal components on an incomplete data set.

### 1.3.1 Preliminary transformations

Principal component methods provide better representation when relationships between continuous variables are linear. To apply PCA or FAMD to incomplete data (continuous or mixed respectively), it can be first useful to check that a linear trend occurs between variables. If linearity between them is not verified, then transformations of the variables (such as logarithm, square root, logistic) can be performed. For instance, to apply PCA on the incomplete data set *sleep*, we begin by generating the bivariate plots. The Figure 5 illustrates that a log transformation of some variables of the data set *sleep* can improve linearity.

```
> don.log<-sleep
> don.log[,c(1:2,6,7)]<-log(don.log[,c(1:2,6,7)])
```

Logarithm transformation can be useful for skew variables, while square root transformation will be generally suitable for count data, and logistic transformation for proportions. Note that these transformations will be also useful to perform MI. To apply BayesMIPCA on an incomplete data set, it will be more suitable to work with the transformed data set, which could be eventually back-transformed.

### 1.3.2 Principal component methods with missing values

Principal component methods are particularly useful to better understand the structure of the data. This allows the understanding of the relationships between variables as well as the similarities between individuals. In particular, This can be useful to detect outliers. For a continuous data set, PCA is the suitable method to use, whereas MCA will be appropriate for a categorical data set and FAMD for a mixed one. However, missing values make it difficult to be applied. The package missMDA allows the use of principal component methods for an incomplete data set. To achieve this goal in the case of PCA, the missing values are predicted using the iterative PCA algorithm (Josse and Husson, 2012) for a predefined number of dimensions. Then, PCA is performed on the imputed data set. The rationale is the same for MCA and FAMD.

The single imputation step requires tuning the number of dimensions used to impute the data. Through the argument `method.cv`, the function `estim_ncpPCA` proposes several cross-validation procedures to choose this number. The rationale of these methods is to search the number of components minimising the prediction error for observed values. The default method is the generalised cross-validation method (`method.cv="gcv"`). It consists in searching the number of dimensions which minimises the generalised cross-validation criterion, which can be seen as an approximation of the leave-one-out cross-validation criterion (Josse and Husson, 2011). The procedure is very fast, because it does not require adding explicitly missing values and predicting them for each cell of the data set. However, the number of dimensions minimising the criterion can sometimes be unobvious when several local minimum occur. In such a case, more computationally intensive methods, those performing explicit cross-validation, can be used, such as Kfold (`method.cv="Kfold"`) or leave-one-out (`method.cv="loo"`). More precisely, the number of dimensions can be estimated as follows on the *sleep* data set:

```
> library(missMDA)
> res.ncp<-estim_ncpPCA(don.log,method.cv="Kfold")
> plot(res.ncp$criterion~names(res.ncp$criterion),xlab="number of dimensions")
> ncp<-4
```

The Kfold cross-validation suggests to retain 4 dimensions for the imputation of the data set *sleep* (Figure 7). Thus, the incomplete data set can be imputed using the function `imputePCA`, specifying the number of dimensions through the argument `ncp=4`. The function returns the imputed data set through the output `completeObs` on which PCA can be performed to summarise the relationships between variables and similarities between individuals:

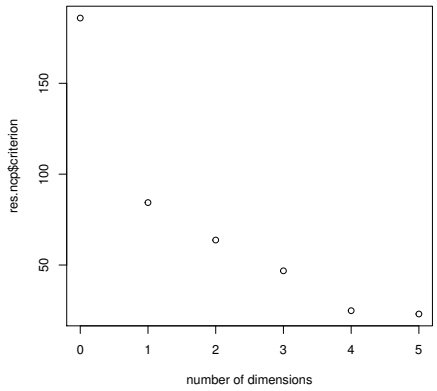


Figure 6: Cross-validation error according to the number of dimensions used for the data set *sleep*

```
> #single imputation
> res.imp<-imputePCA(don.log,ncp = ncp)
> #PCA on the imputed data set
> res.PCA<-PCA(res.imp$completeObs,graph=F)
> #Graph of the variables and graph of the individuals
> plot(res.PCA,choix="var")
> plot(res.PCA,choix="ind")
```

The correlation circle summarises the relationships between variables (left-hand side of Figure 7). The coordinates of the variable on one axis is given by its correlation with the corresponding component. This graph shows several groups of linked variables. For instance, the variables Dream, Sleep and NonD are linked and quite negatively correlated with the others. The graph of the individuals summarises similarities between individuals. The individuals are mapped so that individuals having similar values within the data set of all variables are also close on the map. For instance, Figure 7 shows similarities between some individuals, such as the individuals 37 and 38, and shows oppositions between others, such as between the individuals 5 and 61. It does not highlight outliers.

Note that for a categorical data set, MCA can be performed in the same way: first, the number of components is estimated, next single imputation with iterative MCA algorithm is used (Josse et al., 2012), and then MCA is performed on the imputed disjunctive table. It will be also the same for FAMD on mixed data. For instance, MCA can be applied on the incomplete categorical data set *TitanicNA* from the package missMDA.

```
> data(TitanicNA)
> summary(TitanicNA)

  CLASS        AGE          SEX          SURV
 0   :733   0   :  85   0   : 388   0   :1183
```

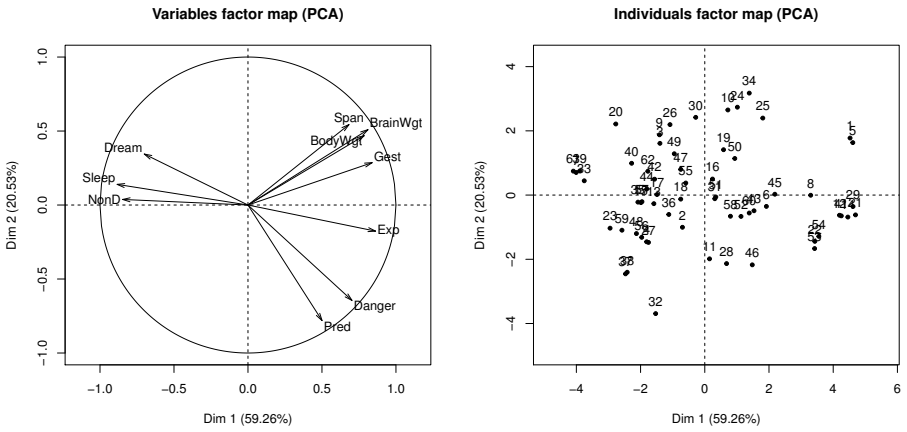


Figure 7: Correlation circle and graph of individuals from the principal component analysis on the data set *sleep*

```
 1   :256   1   :1662   1   :1393   1   : 566
 2   :225   NA's: 454   NA's: 420   NA's: 452
 3   :553
NA's:434
```

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarised according to economic status (class), sex, age and survival. Twenty percent of values are missing completely at random on each variable. To perform MCA on such an incomplete categorical data set, we can run:

```
> #number of components
> estim_ncpMCA(TitanicNA)
> #single imputation
> res.imp<-imputeMCA(TitanicNA,ncp = 5)
> #MCA on the imputed data set
> res.MCA<-MCA(TitanicNA,res.imp$tab.disj,graph=F)
> #Graph of the categories and individuals
> plot(res.MCA,choix="ind",invisible="ind")
> plot(res.MCA,choix="ind")
```

To sum up, the exploration of an incomplete data set allows the user to understand the missing data pattern, the mechanism, as well as the observed values. The principal component methods are very suitable tool in this goal. On one hand, MCA can be used to analyse the missing data pattern. On the other hand, the mechanism can be studied by analysing simultaneously data and missing data pattern with MCA: if data are continuous, then variables must be split and a new category for the missing values added, otherwise, if data are categorical, MCA can be used adding a category for missing values. Lastly, PCA, MCA, or even FAMD, can be used to understand the structure of the observed values, continuous, categorical, or mixed, respectively.

# 2 Multiple imputation for continuous data

## 2.1 Multiple Imputation

The BayesMIPCA method can be used to perform multiple imputation of continuous variables through PCA (Audigier et al., 2015a). The algorithm is based on the Bayesian treatment of the PCA model proposed by Verbanck et al. (2013). When missing values occur in the data set, a classical way to draw parameters from a posterior distribution consists in using a DA algorithm. Then, the imputation of the data set is performed from each parameter obtained by the DA algorithm. This requires alternating `Lstart` times imputation of the missing values and draw of the parameters in the posterior distribution (burn-in period). Note that the imputation step is called *step I* and the step of draw in the posterior is called *step P*. After convergence (to the posterior distribution), imputed data sets are kept each `L` iterations to ensure independence between successive imputations. The function `MIPCA` performs multiple imputation according to PCA from a DA algorithm. The BayesMIPCA method can be used by specifying the argument `method.mi="Bayes"`. Otherwise, a bootstrap version is performed (Josse and Husson, 2011). Note that this bootstrap method has been assessed to evaluate uncertainty in PCA, rather than to apply a statistical method on an incomplete data set. The function `MIPCA` requires as inputs the incomplete data set `X`, the number of imputed data sets `nboot` and the number of dimensions `ncp`. The number of dimensions can be chosen by cross-validation as described previously (Section 1.3). The values of `Lstart` and `L` are fixed to 1000 and 100, respectively. The function returns the multiply imputed data set through the value `res.MI`.

```
> #estimate the number of dimensions
> res.ncp<-estim_ncpPCA(don.log,method.cv="Kfold")
> plot(res.ncp$criterion~names(res.ncp$criterion),xlab="number of dimensions")
> ncp<-4

> #Multiple imputation
> res.BayesMIPCA<-MIPCA(X=don.log,nboot=100,ncp=ncp,method.mi="Bayes")

> #Extract the first data set
> imp1<-res.BayesMIPCA$res.MI[[1]]
> summary(imp1)
```

## 2.2 Diagnostics

### 2.2.1 BayesMIPCA algorithm

The number of iterations for the burn-in period (`Lstart`), and the number of iterations between each retained imputed data set (`L`) play a role in the BayesMIPCA algorithm. Typically, their values are empirically chosen through graphical investigations.

To check if the number of iterations for the burn-in period is sufficiently high, the successive values of some summaries (*e.g.* mean or correlation coefficients) are investigated

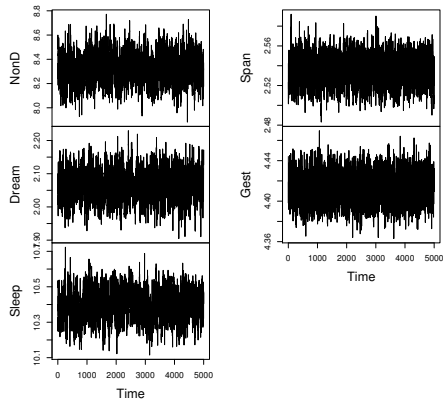


Figure 8: Successive values of the mean through the imputed data set for the imputed variables of the data set *sleep*

through the successive iterations of the algorithm. However, the imputed data sets given by the BayesMIPCA algorithm (with default parameters) are not the imputed data sets from each step (I), but only a sample of them because a data set is kept each `L` iterations from the `Lstart`$^{th}$. To obtain the successive imputed data sets, the algorithm need to be run by specifying the arguments `Lstart=1` and `L=1` for a large number of imputed data sets `nboot`. The successive values for the mean of each incomplete variables can be plotted with the time series function `ts`:

```
> res.conv<-MIPCA(X=don.log,nboot=5000,ncp=ncp, method.mi="Bayes",Lstart=1,L=1)
> res.mean<-lapply(res.conv$res.MI,colMeans)
> X<-ts(do.call(rbind,res.mean))
> plot(X[,3:7],main="")
```

If stationarity of the successive values seems to be verified after 1000 iterations, then `Lstart` can be preserved. Otherwise, it could be increased. Figure 8 shows the successive values for the means. The convergence seems to have been reached quickly, meaning that the default parameter `Lstart=1000` is suitable.

Concerning the number of iterations between each retained data set, it can be checked by visualising the autocorrelograms of the summaries (Figure 9). An autocorrelogram represents the correlation between the vector of the successive statistics and its shifted version for several lags. We aim finding a lag `L` sufficiently large to avoid correlation between the vectors of statistics. The autocorrelograms for the means can be obtained as follows:

```
> par(mfrow=c(3,2))
> Lstart<-1000
> X<-ts(X[Lstart:nrow(X),],start=Lstart)
> apply(X[,3:7],2,acf)
```

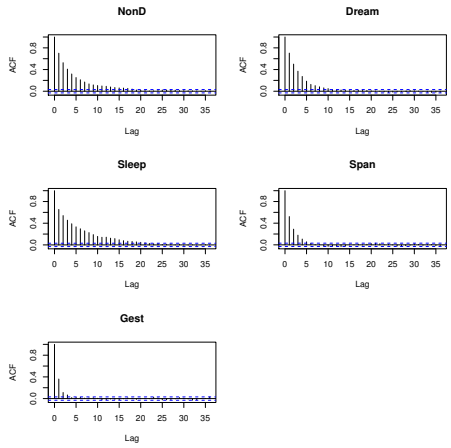


Figure 9: Autocorrelograms for the mean the imputed variables of the data set *sleep*

The Figure 9 shows the correlation between the mean of each incomplete variable, and its shift version for a lag smaller than 35. We see that a lag larger than 25 is sufficient to have a correlation close to zero. This indicates that `L` should be over 25 to expect independence between the imputed values from one data set to another. Therefore, the default parameter `L=100` is suitable.

These investigations are not foolproof and it can be useful to focus on others summaries, such as correlation coefficients that deal with the relationships between variables, whereas the means deal with marginal distribution only.

### 2.2.2 Fit of the model

The fit of the imputation model is interesting to evaluate the accuracy of the imputed values. A first way to assess accuracy consists in comparing the distribution of the imputed values and the distribution of the observed ones. The package VIM can be used for this purpose. Note that a difference between these distributions does not mean that the model is unsuitable. Indeed, when the missing data mechanism is not MCAR, it could make sense to observe differences between the distribution of imputed values and the distribution of observed values. However, if differences occur, more investigations would be required to try to explain them. The principal component methods can be very useful in this aim (Section 1).

Another way consists in comparing imputed values to their 'true' values. However, the values that are missing are not available, by definition. To assess the fit of the Bayesian PCA model, we propose to use *overimputation* (Blackwell et al., 2015). It consists in imputing each observed values from the parameters of the imputation model obtained from the MI procedure. The comparison between the "overimputed" values and the observed values is made by building a confidence interval for each observed value. If the model fits well the data, then the 90% confidence interval should contain the observed value

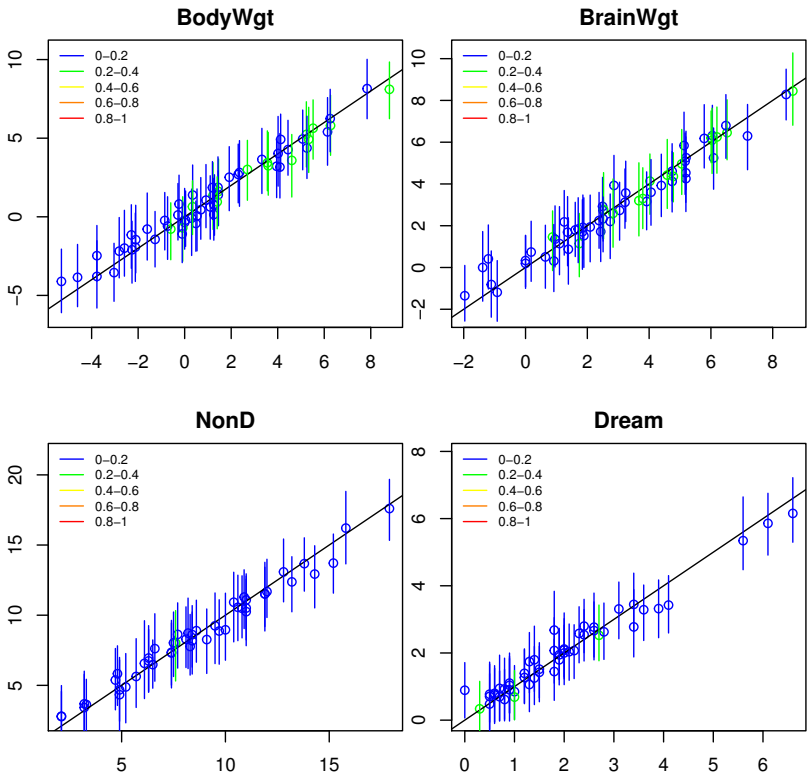


Figure 10: Assessment of the Bayesian PCA model. The dots represent the mean imputation and the vertical segments the confidence intervals for the missing value. Around 90 percent of these confidence intervals should contain the first bisector which corresponds to the true values of the missing values. The color of the line represents the proportion of missing observations in the missing data pattern for that observation.

in 90% of the cases. The fit can be assessed by the function `Overimpute` which takes as an input the output of the MIPCA function (`output`) and the variables that are plotted (`plotvars`). Note that this function is similar to the function `overimpute` from the Amelia package (Honaker et al., 2014).

For instance, to check the fit of the method on the first variables we can run

```
> Overimpute(res.BayesMIPCA,plotvars=1:4)
```

The function plots the predicted values and the confidence intervals as shown in Figure 10. The x-axis corresponds to the values that are suppressed, whereas the y-axis corresponds to the predicted values. Thus, the first bisector represents the line where the prediction is perfect. For each observation, the 90% confidence intervals is plotted with a color depending on the percentage of missing values on the covariates. The predicted values with a high number of missing covariates tend to have larger confidence intervals. If the multiple imputation performs well, then 90% of the confidence intervals should cross the first bisector. Note that the confidence intervals are construct according to the quantiles of

the overimputed values, therefore, a large number of imputed values is required, meaning that the output of MIPCA should be obtained for a parameter `nboot` greater than 100.

# 3 Multiple imputation for categorical data

The MIMCA method performs MI for categorical data using MCA (Audigier et al., 2015b). The rationale of the method consists in performing a non-parametric bootstrap of the individuals to reflect the uncertainty on the parameters of the imputation model; next in imputing the incomplete disjunctive table according to the MCA parameters estimated from each bootstrap replicate; and finally, in proposing categories from it by making a random draw for each missing entry. More precisely, several weightings are first defined for the individuals. Then, the iterative regularized MCA algorithm (Josse et al., 2012) is applied according to each weighting for a predefined number of dimensions. It leads to several imputations of the disjunctive table. These imputed tables are scaled to verify the constraint that the sum is equal to one per variable and per individual. Lastly, missing categories are drawn from the probabilities given by the imputed tables. Thus, the multiply imputed categorical data set is obtained.

The function `MIMCA` takes as an input the incomplete data set (`X`), the number of imputed data sets (`nboot`), and the number of dimensions (`ncp`). The imputed data sets are available in the output `res.MI`. The function can be applied on the incomplete data set *TitanicNA* as follows:

```
> ## Number of components
> res.ncp <- estim_ncpMCA(TitanicNA,method.cv="loo")
> plot(res.ncp$criterion~names(res.ncp$criterion),xlab="number of dimensions")
> ## Multiple Imputation
> res.MIMCA <- MIMCA(TitanicNA, ncp=5)
> ## First completed data matrix
> res.MIMCA$res.MI[[1]]
```

To assess the imputation of missing categorical values according to the MIMCA method, it is also possible to compare the distributions of the imputed values and the observed ones. However, the bivariate clouds cannot be represented for categorical variables. A way to achieve this goal consists in visualising the contingency table of one imputed data set, while distinguish the counts according to the fact whether missing data occurs on a variable (or a set of variables) or not. For a MCAR mechanism, equal counts for the imputed data set are expected if missing values occur or not on other variables. The function `mosaicMiss` from the R package VIM proposes such a diagnostic. Due to the fact that the output could be non-intuitive, it is rather recommended for experienced user.

# 4 Applying a statistical method

MI aims to apply a statistical method on an incomplete data set. To apply such a method on the multiply imputed data set obtained from the function MIPCA or MIMCA, one simple way is to use the R package Zelig (Owen et al., 2013), which enables a large range of statistical models. The function `zelig` performs analysis and pools the results. It can be simply applied on the imputed data sets provided in the output `res.MI`. For instance, to predict the survival status of the passengers of the ocean liner Titanic according to their age, sex and economic status, a logistic regression model can be applied as follows:

```
> library(Zelig)

> z.out <- zelig(SURV~ AGE+SEX+CLASS, model = "logit", data = res.MIMCA$res.MI,cite=F)
> summary(z.out,digits=5)

  Model: logit
  Number of multiply imputed data sets: 100

Combined results:

Call:
glm(formula = formula, weights = w, family = binomial(link = "logit"),
    model = F, data = data)

Coefficients:
                 Value Std. Error      t-stat      p-value
(Intercept)  2.2753580  0.4710785   4.8301033 2.236290e-06
AGE1        -0.9607067  0.4010577  -2.3954325 1.734181e-02
SEX1        -2.6139338  0.1729783 -15.1113411 2.931294e-46
CLASS1       0.8411288  0.1944820   4.3249701 1.678661e-05
CLASS2      -0.1220363  0.2249342  -0.5425424 5.876332e-01
CLASS3      -0.9267939  0.1978353  -4.6846749 3.511780e-06

For combined results from datasets i to j, use summary(x, subset = i:j).
For separate results, use print(summary(x), subset = i:j).
```

The range of the models proposed by Zelig is very large, we refer to Owen et al. (2013) for more details on the use of these models. Note that others packages than Zelig can be used to pool the analysis results. In particular, the package `mice` (Van Buuren and Groothuis-Oudshoorn, 2014) provides interesting outputs such as the bounds of the 95% confidence intervals or the fraction of missing information (fmi). This last quantity can be interpreted as the part of variability due to missing values. Large values (*e.g.* fmi$> .5$) indicate that the results are sensitive to the MI method used.

```
> library(mice)
> z.mira <- as.mira(z.out)
> summary(pool(z.mira))
```

```
                 est        se           t        df     Pr(>|t|)      lo 95
(Intercept)  2.2940979 0.4745173   4.8345929  280.4428 2.201674e-06  1.3600301
AGE1        -0.9903224 0.3905133  -2.5359508  266.5059 1.178593e-02 -1.7592060
SEX1        -2.6026064 0.1861629 -13.9802631  561.9537 0.000000e+00 -2.9682665
CLASS1       0.8495560 0.1922040   4.4200757 1098.3381 1.084472e-05  0.4724276
CLASS2      -0.1150298 0.2451277  -0.4692649  417.3012 6.391254e-01 -0.5968688
CLASS3      -0.9120911 0.1978792  -4.6093329  560.5303 5.006856e-06 -1.3007664
                hi 95 nmis       fmi    lambda
(Intercept)  3.2281658   NA 0.5968091 0.5939439
AGE1        -0.2214388   NA 0.6121783 0.6092788
SEX1        -2.2369463   NA 0.4215791 0.4195242
CLASS1       1.2266844   NA 0.3012625 0.2999913
CLASS2       0.3668091   NA 0.4893152 0.4868735
CLASS3      -0.5234157   NA 0.4221151 0.4200568
```

In such a case, the comparison of the obtained results with the ones obtained by listwise deletion is recommended. A smaller variance for estimator obtained by MI is expected. If differences occur in the point estimates, it makes sense to attach more trust to the MI results since MI is theoretically superior to Listwise deletion. However, it remains important to try to explain differences. Listwise deletion on the data set *TitanicNA* could be performed with the function `zelig` applied on the incomplete data set:

```
> z.out.ld <- zelig(SURV~ AGE+SEX+CLASS, model = "logit", data = TitanicNA, cite=F)
> summary(z.out.ld,digits=5)

Call:
glm(formula = formula, weights = w, family = binomial(link = "logit"),
    model = F, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1622  -0.6887  -0.6695   0.4506   2.1979

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.34395    0.52070   4.502 6.75e-06 ***
AGE1        -0.94021    0.42788  -2.197 0.027993 *
SEX1        -2.72175    0.23895 -11.391  < 2e-16 ***
CLASS1       0.83232    0.23980   3.471 0.000519 ***
CLASS2       0.03716    0.27416   0.136 0.892176
CLASS3      -1.00372    0.24802  -4.047 5.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1120.73  on 897  degrees of freedom
Residual deviance:  859.11  on 892  degrees of freedom
  (1303 observations deleted due to missingness)
AIC: 871.11

Number of Fisher Scoring iterations: 4
```

As expected, the variability of the estimators is larger with listwise deletion, while the point estimates are close to the ones obtained by MI since the mechanism is MCAR.

# BIBLIOGRAPHY

T. Allison and D. Chichetti. Sleep in mammals: ecological and constitutional correlates. *Science*, 194(4266):732–734, 1976. (page 1)

V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 2015a. doi: 10.1080/00949655.2015.1104683. (page 11)

V. Audigier, F. Husson, and J. Josse. MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *ArXiv e-prints*, 2015b. (page 15)

M. Blackwell, J. Honaker, and G. King. A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods and Research*, pages 1–39, 2015. (page 13)

J. Honaker, G. King, and M. Blackwell. *Amelia II: A Program for Missing Data*, 2014. R package version 1.7.2. (page 14)

F. Husson and J. Josse. *missMDA: Handling Missing Values with Multivariate Data Analysis*, 2015. URL `http://www.agrocampus-ouest.fr/math/husson,http://juliejosse.com/`. R package version 1.9. (page 7)

J. Josse and F. Husson. Multiple imputation in PCA. *Advances in data analysis and classification*, 5:231–246, 2011. (pages 8 et 11)

J. Josse and F. Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153 (2):1–21, 2012. (page 8)

J. Josse and F. Husson. missmda a package to handle missing values in principal component methods. *Journal of Statistical Software*, 2015. (page 7)

J. Josse, M. Chavent, B. Liquet, and F. Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29:91–116, 2012. (pages 9 et 15)

M. Owen, K. Imai, G. King, and O. Lau. *Zelig: Everyone's Statistical Software*, 2013. URL http://CRAN.R-project.org/package=Zelig. R package version 4.2-1. (page 16)

D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987. (page 2)

M. Templ, A. Alfons, A. Kowarik, and B. Prantner. *VIM: Visualization and Imputation of Missing Values*, 2015. URL http://CRAN.R-project.org/package=VIM. R package version 4.3.0. (page 1)

S. Van Buuren and K. Groothuis-Oudshoorn. *mice*, 2014. R package version 2.22. (page 16)

M. Verbanck, J. Josse, and F. Husson. Regularised PCA to denoise and visualise data. *Statistics and Computing*, pages 1–16, 2013. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-013-9444-y. (page 11)

# BIBLIOGRAPHIE


P. D. ALLISON : *Missing Data*. Thousand Oaks, CA : Sage, 2002. (page 14)

P. D. ALLISON : Imputation of categorical variables with PROC MI. *In SAS Users Groups International (SUGI)*, vol. 30, p. 1–14, 2005. (page 4)

T. W. ANDERSON : Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203, 1957. (page 19)

V. AUDIGIER, F. HUSSON et J. JOSSE : A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, p. 1–22, 2014. ISSN 1862-5347. In press. (pages 7, 37 et 63)

V. AUDIGIER, F. HUSSON et J. JOSSE : MIMCA : Multiple imputation for categorical variables with multiple correspondence analysis. *ArXiv e-prints*, 2015a. (pages 8 et 94)

V. AUDIGIER, F. HUSSON et J. JOSSE : Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 2015b. (pages 8 et 66)

J. BARNARD et D. B. RUBIN : Small Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86:948–955, 1999. (page 25)

D. BARTHOLOMEW, M. KNOTT et I. MOUSTAKI : *Latent Variable Models and Factor Analysis : A Unified Approach*. Wiley Series in Probability and Statistics. Wiley, 2011. (page 27)

C. A. BERNAARDS, T. R. BELIN et J. L. SCHAFER : Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6): 1368–1382, mar 2007. (page 4)

J. BESAG : Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 1974. (page 4)

C. BIERNACKI, G. CELEUX et G. GOVAERT : Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Comput. Stat. Data Anal.*, 41(3-4):561–575, 2003. (page 20)

M. BLACKWELL, J. HONAKER et G. KING : A unified approach to measurement error and missing data : Overview and applications. *Sociological Methods and Research*, p. 1–39, 2015. (page 137)

W. BOSCARDIN et X. ZHANG : *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, chap. Modeling the covariance and correlation matrix of repeated measures. Wiley Series in Probability and Statistics. Wiley, 2004. (page 5)

W. BOSCARDIN, X. ZHANG et T. BELIN : Modeling a mixture of ordinal and continuous repeated measures. *Journal of Statistical Computation and Simulation*, 78(10):873–886, 2008. (pages 4 et 5)

L. BREIMAN : Random forests. *Machine Learning*, 45(1):5–32, 2001. (pages 6 et 29)

J. M. BRICK : Unit nonresponse and weighting adjustments : A critical review. *Journal of Official Statistics*, 2013. (page 15)

J. CARPENTER et M. KENWARD : *Multiple imputation and its application*. Wiley, 1st edition édn, 2013. (pages 6, 131 et 136)

H. CAUSSINUS : Models and uses of principal component analysis (with discussion). *In Multidimensional Data Analysis*, p. 149–178. DSWO Press, 1986. (page 133)

A. CHRISTOFFERSSON : *The one component model with incomplete data*. Wilkinson, 1970. (page 35)

T. de WAAL, J. PANNEKOEK et S. SCHOLTUS : *Handbook of data editing and imputation*. Wiley, 2011. (page 136)

H. DEMIRTAS : Rounding strategies for multiply imputed binary data. *Biometrical journal*, 51(4):677–88, 2009. ISSN 1521-4036. (page 4)

H. DEMIRTAS : A distance-based rounding strategy for post-imputation ordinal data. *Journal of Applied Statistics*, 37(3):489–500, 2010. (page 4)

A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977. (page 19)

F. DI LASCIO, S. GIANNERINI et A. REALE : Exploring copulas for the imputation of complex dependent data. *Statistical Methods & Applications*, 24(1):159–175, 2015. ISSN 1618-2510. (page 5)

L. L. DOOVE, S. VAN BUUREN et E. DUSSELDORP : Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014. URL http://dx.doi.org/10.1016/j.csda.2013.10.025. (pages 6 et 29)

C. ECKART et G. YOUNG : The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, sept. 1936. ISSN 0033-3123. (page 33)

C. K. ENDERS : *Applied Missing Data Analysis*. Methodology in the social sciences. Guilford Press, 2010. ISBN 9781606236390. (pages 22, 24 et 27)

E. ESCOFIER : Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *Les cahiers de l'analyse des données*, 4(2):137–146, 1979. (page 6)

Y. ESCOUFIER : The analysis of simple and multiple contingency tables. *In* R. COPPI, éd. : *Proceedings of the international meeting of the analysis of multidimensional contingency tables*, p. 53–77, 1982. (page 134)

R. FAY : A design-based perspective on missing data variance. *In Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 1991. (page 3)

R. FAY : When are inferences from multiple imputation valid ? *In the Section on Survey Research Methods*. American Statistical Association, 1992. (page 3)

R. FAY : Valid inferences from imputed survey data. *In the Section on Survey Research Methods*. American Statistical Association, 1993. (page 3)

R. FAY : Analyzing imputed survey data sets with model-assisted estimators. *In the Section on Survey Research Methods*. American Statistical Association, 1994. (page 3)

R. FAY : Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91:490–498, 1996. (page 3)

G. FITZMAURICE, M. KENWARD, G. MOLENBERGHS, G. VERBEKE et A. TSIATIS : *Missing data : Introduction and Statistical Preliminaries*, chap. 1, p. 3–22. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, New York, 2014. (pages 13 et 23)

A. E. GELFAND et A. F. M. SMITH : Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. ISSN 01621459. (pages 3 et 21)

A. GELMAN, J. B. CARLIN, H. S. STERN et D. B. RUBIN : *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 édn, 2003. (page 22)

S. GEMAN et D. GEMAN : Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984. (pages 3 et 21)

J. GRAHAM : Adding missing-data relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, 2003. (page 23)

M. J. GREENACRE : *Theory and applications of correspondence analysis*. Academic Press, London, 1984. ISBN 0-12-299050-1. (page 134)

M. J. GREENACRE et J. BLASIUS : *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, 2006. (page 6)

D. HAZIZA, V. DONGMO JIONGO et P. DUCHESNE : Triple robustesse en présence de données imputées dans les enquêtes. *In Proceedings of the JMS*, 2012. (page 16)

R. HE : *Multiple Imputation of High-dimensional Mixed Incomplete Data*. Thèse de doctorat, University of California, 2012. (pages 4 et 5)

P. HOFF : Model averaging and dimension selection for the singular value decomposition. *J. Amer. Statist. Assoc.*, 102(478):674–685, 2007. ISSN 0162-1459. (page 135)

J. HONAKER, G. KING et M. BLACKWELL : Amelia II : A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. (page 137)

N. J. HORTON, S. R. LIPSITZ et M. PARZEN : A potential for bias when rounding in multiple imputation. *The American Statistician*, 57:229–232, 2003. (page 4)

D. HORVITZ et D. THOMPSON : A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. (page 14)

K. HRON, M. TEMPL et P. FILZMOSER : Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54 (12):3095 – 3107, 2010. ISSN 0167-9473. (page 136)

R. HUGHES, I. WHITE, S. SEAMAN, J. CARPENTER, K. TILLING et J. STERNE : Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1), 2014. URL `http://dx.doi.org/10.1186/1471-2288-14-28`. (page 4)

H. JEFFREY : An invariant form for the prior probability in estimation problems. *In Proceedings of the Royal Society of London, Series A*, vol. 186, p. 453–461, 1946. (page 22)

H. JOE : *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1997. (page 4)

I. JOLLIFFE : *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002. ISBN 9780387954424. (page 6)

J. JOSSE, M. CHAVENT, B. LIQUET et F. HUSSON : Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29:91–116, 2012. (page 36)

J. JOSSE et F. HUSSON : Multiple imputation in PCA. *Advances in data analysis and classification*, 5:231–246, 2011. (page 66)

J. JOSSE, J. PAGÈS et F. HUSSON : Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique*, 150:28–51, 2009. (pages 35, 36 et 133)

J. JOSSE et S. SARDY : Adaptive shrinkage of singular values. *Statistics and Computing*, p. 1–10, 2015. ISSN 0960-3174. URL http://dx.doi.org/10.1007/s11222-015-9554-9. (page 135)

J. D. Y. KANG et J. L. SCHAFER : Demystifying double robustness : A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4):523–539, 2007. (page 15)

H. A. L. KIERS : Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56:197–212, 1991. (page 6)

H. J. KIM, J. P. REITER, Q. WANG, L. H. COX et A. F. KARR : Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32 (3):375–386, 2014. (page 136)

G. KING, J. HONAKER, A. JOSEPH et K. SCHEVE : Analyzing incomplete political science data : An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69, 2001. (pages 4, 7, 66 et 93)

J. KROPKO, B. GOODRICH, A. GELMAN et J. HILL : Multiple imputation for continuous and categorical data : Comparing joint and conditional approaches. *Political Analysis*, 2014. (page 4)

E. KÄÄRIK et M. KÄÄRIK : Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference*, 139(11):3830–3835, 2009. (page 5)

L. LEBART, A. MORINEAU et K. M. WERWICK : *Multivariate Descriptive Statistical Analysis*. Wiley, New-York, 1984. (page 6)

K. J. LEE et J. CARLIN : Multiple Imputation for Missing Data : Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171 (5):624–632, 2010. (page 4)

R. J. A. LITTLE : Modelling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995. (page 12)

R. J. A. LITTLE et D. B. RUBIN : *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New-York, 2002. (pages 1, 16 et 17)

R. J. A. LITTLE et S. ZANGANEH : Missing at random and ignorability for inferences about subsets of parameters with missing data, 2013. (page 18)

R. LIU : *Multiple imputation for missing items in multi-themed questionnaires*. Thèse de doctorat, The Pennsylvania State University, 2010. (page 5)

D. MANRIQUE-VALLIER et J. REITER : Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, 40:125–134, 2014. (page 136)

M. MARBAC-LOURDELLE : *Model-based clustering for categorical and mixed data sets*. Thèse de doctorat, Université lille 1, 2014. (page 6)

X. L. MENG : Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10:538–573, 1994. (page 27)

X. L. MENG et D. B. RUBIN : Maximum Likelihood Estimation via the ECM Algorithm : A General Framework. *Biometrika*, 80(2):267–278, 1993. ISSN 00063444. (page 20)

J. S. MURRAY et J. P. REITER : Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence. *ArXiv e-prints*, 2014. (page 6)

R. B. NELSEN : *An Introduction to Copulas*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. (page 4)

I. OLKIN et R. F. TATE : Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.*, 32(2):448–465, 1961. (pages 4 et 134)

J. PAGÈS : *Multiple Factor Analysis by Example Using R*. Chapman & Hall/CRC The R Series. Taylor & Francis, 2015. ISBN 9781482205473. (page 6)

M. PRAGUE, R. WANG, A. STEPHENS, E. TCHETGEN TCHETGEN et V. DEGRUTTOLA : Accounting for interactions and complex inter-subject dependency for estimating treatment effect in cluster randomized trials with missing at random outcomes. *Biometrics*, 2015. In review. (page 16)

J. P. REITER et T. E. RAGHUNATHAN : The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102:1462–1471, 2007. (pages 137 et 138)

G. ROBERTS : *Markov Chain Concepts related to sampling algorithms*, chap. 3, p. 45–58. Chapman & Hall, Londres, 1996. (page 4)

A. ROTNITZKY : *Inverse probability weighted methods*, p. 453,476. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, New York, 2009. (page 15)

A. ROTNITZKY et S. VANSTEELANDT : *Double-robust methods*, chap. 9, p. 185–212. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, New York, 2014. (page 15)

D. B. RUBIN : Inference and missing data. *Biometrika*, 63:581–592, 1976. (page 12)

D. B. RUBIN : *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987. (pages 3, 17, 25, 26 et 29)

D. B. RUBIN : Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9 (2):461–468, 1993. (page 138)

D. B. RUBIN et N. SCHENKER : Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, 81:366–374, 1986. (page 25)

D. RUBIN : Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. (page 27)

V. SAVALEI et P. M. BENTLER : A two-stage approach to missing data : Theory and application to auxiliary variables. *Structural Equation Modeling : A Multidisciplinary Journal*, 16(3):477–497, 2009. (page 23)

J. L. SCHAFER : *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997. (pages 4, 5, 7, 13, 17, 22, 28, 29, 66 et 93)

J. L. SCHAFER : Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35, 2003. (page 27)

J. L. SCHAFER et J. W. GRAHAM : Missing data : our view of the state of the art. *Psychological Methods*, 7:147–177, 2002. (page 16)

D. O. SCHARFSTEIN, A. ROTNITZKY et J. M. ROBINS : Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1146, 1999. (page 15)

A. D. SHAH, J. W. BARTLETT, J. CARPENTER, O. NICHOLAS et H. HEMINGWAY : Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study. *American Journal of Epidemiology*, 179(6):764–774, mars 2014. ISSN 1476-6256. (pages 6, 7, 29 et 93)

Y. SI et J. REITER : Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38:499–521, 2013. (pages 6, 7, 93 et 136)

A. SKLAR : *Fonctions de répartition à n dimensions et leurs marges*. 1959. (page 4)

J. SONG et T. BELIN : Analysis of incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in medicine*, 23:2827–2843, 2004. (page 5)

D. J. STEKHOVEN et P. BÜHLMANN : Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. (pages 7 et 29)

M. A. TANNER et W. H. WONG : The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:805–811, 1987. (page 21)

C. TEMPLEMAN : *Imputation of restricted data*. Thèse de doctorat, University of Groningen, 2007. (page 136)

A. TSIATIS : *Semiparametric Theory and Missing Data*. Springer, 2006. (page 15)

A. TSIATIS et M. DAVIDIAN : *Missing data methods : A semi-parametric perspective*, chap. 8, p. 149–184. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, New York, 2014. (page 16)

S. VAN BUUREN : *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 1 édn, 2012. (pages 3, 4, 7, 26, 29 et 93)

S. VAN BUUREN : *Fully Conditional Specification*, chap. 13, p. 267–294. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, New York, 2014. (page 3)

S. VAN BUUREN, J. P. L. BRAND, C. G. M. GROOTHUIS-OUDSHOORN et D. B. RUBIN : Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76:1049–1064, 2006. (page 3)

D. van der PALM, L. van der ARK et J. VERMUNT : A comparison of incomplete-data methods for categorical data. *Statistical methods in medical research*, 2014. ISSN 1477-0334. in press. (page 13)

M. VERBANCK, J. JOSSE et F. HUSSON : Regularised PCA to denoise and visualise data. *Statistics and Computing*, p. 1–16, 2013. ISSN 0960-3174, 1573-1375. (pages 7, 65, 66, 133 et 135)

J. K. VERMUNT, J. R. van GINKEL, L. A. van der ARK et K. SIJTSMA : Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology, Vol 38*, 38:369–397, 2008. (page 6)

D. VIDOTTO, M. C. KAPTEIJN et J. VERMUNT : Multiple imputation of missing categorical data using latent class models : State of art. *Psychological Test and Assessment Modeling*, 2014. in press. (pages 6 et 135)

I. WASITO et B. MIRKIN : Nearest neighbour approach in the least-squares data imputation algorithms. *Information Sciences*, 169(1-2):1–25, 2005. (page 35)

I. WASITO et B. MIRKIN : Nearest neighbours in least-squares data imputation algorithms with different missing patterns. *Computational Statistics & Data Analysis*, 50(4):926–949, 2006. (page 35)

K. H. YUAN et P. M. BENTLER : Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, 30:165–200, 2000. (page 23)

R. M. YUCEL, Y. HE et A. M. ZASLAVSKY : Using calibration to improve rounding in imputation. *The American Statistician*, 62:125–129, 2008. (page 4)

X. ZHANG, W. J. BOSCARDIN et T. R. BELIN : Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational Statistics & Data Analysis*, 52(7):3697–3708, 2008. (page 5)