Journal de la Société Française de Statistique
Volume 150, numéro 2, 2009

Gestion des données manquantes en Analyse en Composantes Principales

Julie Josse, François Husson & Jérôme Pagès[1]

Title

Handling missing values in Principal Component Analysis

Résumé

Une solution classique pour réaliser une Analyse en Composante Principale (ACP) sur données incomplètes consiste à chercher les axes et les composantes qui minimisent l'erreur de reconstitution sur les données présentes. Plusieurs algorithmes ont été proposés dans la littérature comme NIPALS, une approche par moindres carrés alternés pondérés et une approche par ACP itérative. Cette dernière consiste en une imputation itérative des données au cours du processus d'estimation et s'apparente à un algorithme EM d'un modèle particulier. Ces algorithmes sont décrits dans le cadre commun de la minimisation du critère. Cette présentation unifiée permet de mieux comprendre leurs propriétés et les difficultés qu'ils rencontrent. Nous nous focalisons ensuite sur le problème principal du surajustement et montrons comment la formulation probabiliste de l'ACP (Tipping & Bishop, 1997) offre un terme de régularisation adapté pour pallier à ce problème. Les performances de l'algorithme finalement proposé sont comparées à celles des autres algorithmes à partir de simulations.

Mots-clés : *ACP, données manquantes, moindres carrés alternés pondérés, algorithme EM, ACP-GEM, surajustement, ACP probabiliste*

Abstract

An approach commonly used to handle missing values in Principal Component Analysis (PCA) consists in ignoring the missing values by optimizing the loss function over all non-missing elements. This can be achieved by several methods, including the use of NIPALS, weighted regression or iterative PCA. The latter is based on iterative imputation of the missing elements during the estimation of the parameters, and can be seen as a particular EM algorithm. First, we review theses approaches with respect to the criterion minimization. This presentation gives a good understanding of their properties and the difficulties encountered. Then, we point out the problem of overfitting and we show how the probabilistic formulation of PCA (Tipping & Bishop, 1997) offers a proper and convenient regularization term to overcome this problem. Finally, the performances of the new algorithm are compared to those of the other algorithms from simulations.

Keywords : *PCA, missing values, alternating weighted least squares, EM algorithm, GEM-PCA, overfitting, probabilistic PCA*

Mathematics Subject Classification: (62H25)

[1]Laboratoire de mathématiques, Agrocampus — 65 rue de Saint-Brieuc, F-35042 Rennes Cedex
julie.josse@agrocampus-ouest.fr

1 Introduction

La gestion des données manquantes est un problème incontournable dans la pratique statistique. Deux ouvrages synthétisent les recherches et font référence : Little & Rubin [26] et Schafer [35]. Jusqu'aux années 1970, les seules méthodes disponibles pour analyser un tableau de données (individus $\times$ variables) incomplet consistaient à supprimer les individus ayant des valeurs manquantes ou à imputer ces valeurs, c'est-à-dire les compléter par des valeurs plausibles. Ces méthodes conduisent à l'obtention d'un tableau de données "complet" sur lequel toute analyse statistique peut être mise en œuvre. Les méthodes d'imputation sont préférées à la suppression car toute l'information observée est conservée. Dans le cas de variables quantitatives, l'imputation la plus connue consiste à remplacer les valeurs manquantes par la moyenne de chaque variable. La prise en compte des liaisons entre variables, par régression par exemple, améliore l'imputation. Cependant, ces méthodes peuvent déformer les distributions et les relations entre variables [36] et souffrent de ne pas prendre en compte l'incertitude due aux données manquantes. En effet, les valeurs imputées ne sont pas des "vraies" valeurs mais des "suppositions" et l'incertitude sur ces estimations n'est pas prise en compte dans la suite de l'analyse. Ainsi les variances des estimateurs calculées sur les jeux de données imputés sont sous-estimées même si le modèle qui a généré l'imputation est correct.

Deux alternatives ont alors été proposées pour estimer les paramètres d'intérêts et leur variance associée : l'approche par maximum de vraisemblance et l'imputation multiple [32]. Elles sont fondées sur la théorie des données manquantes de Rubin [26] qui considère un mécanisme probabiliste à l'origine des données manquantes. Quand ce processus est ignorable (selon la définition de [26]), l'estimateur du maximum de vraisemblance $\hat{\theta}$ du paramètre inconnu θ est obtenu à partir de la vraisemblance observée. Très souvent, $\hat{\theta}$ ne peut pas être calculé de façon explicite et il est possible de recourir à l'algorithme itératif EM [9].

Il est important de souligner que [36] et [26] se placent dans un cadre inférentiel, cadre bien adapté à la modélisation statistique mais peu utilisé en analyse de données. En effet, l'analyse de données est souvent caractérisée par un objectif exploratoire et présentée sans référence à des hypothèses de nature probabiliste au profit de critères géométriques. Elle est fondée sur la dualité et insiste sur les représentations graphiques, celle des individus comme celle des variables. Il est donc nécessaire de prendre en compte ces caractéristiques particulières afin de proposer des méthodes pour traiter les données manquantes en analyse des données. Peu de littérature est disponible sur ce sujet et de ce fait, de nombreux logiciels de statistique, qui laissent une place importante à l'analyse des données, gèrent les données manquantes en les remplaçant par la moyenne de chaque variable. Des stratégies plus élaborées ont pourtant été proposées [45, 28, 16]. Ces approches se situent souvent à la frontière entre modélisation et analyse exploratoire.

Dans cet article, nous nous focalisons sur la gestion des données manquantes en Analyse en Composantes Principales (ACP). Après avoir présenté des algorithmes itératifs permettant de réaliser une ACP sur données complètes (section 2), nous les étendons au cas incomplet en les présentant dans un cadre commun (section 3). L'analyse des propriétés des algorithmes (section 4), nous conduit à nous intéresser à l'ACP probabiliste (section 5) afin de remédier au problème du surajustement. Enfin, les performances de ces algorithmes sont évaluées à partir de simulations (section 6).

2 L'Analyse en Composantes Principales (ACP)

2.1 Formulation de l'ACP

L'ACP est une technique de réduction de la dimension qui permet l'exploration et la visualisation d'un tableau de données individus par variables quantitatives. Classiquement, l'ACP est présentée comme la recherche du sous-espace qui maximise la variance des points projetés, autrement dit le sous-espace qui représente au mieux la diversité des individus. De façon équivalente, elle peut être présentée comme la recherche du sous-espace qui minimise l'erreur de reconstitution [29], c'est-à-dire la distance entre les individus et leur projection.

Soit une matrice X de dimension $I \times K$, supposée centrée sans perte de généralité, $x_{i.}$ la ligne i, $x_{.k}$ la colonne k et $\|A\| = \sqrt{tr(AA')}$ la norme de Frobenius. Minimiser l'erreur de reconstitution revient à chercher une matrice de rang inférieur S ($S < K$) qui approche au mieux la matrice X au sens des moindres carrés. Ceci équivaut à chercher deux matrices $F_{I\times S}$ et $u_{K\times S}$ qui minimisent le critère suivant :

$$\mathcal{C} \quad = \quad \|X - Fu'\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}(x_{ik} - \sum_{s=1}^{S} F_{is}u_{ks})^2. \tag{1}$$

Avec la contrainte d'axes orthogonaux et de norme unité, la solution unique est fournie par les composantes principales notées $\hat{F}$ (normées à la valeur propre) et les axes principaux notés $\hat{u}$ de l'ACP, vecteurs propres respectivement de la matrice de produit-scalaire et de variance-covariance. Classiquement, l'ACP est réalisée par diagonalisation de l'une ou l'autre de ces matrices selon que le nombre d'individus est supérieur ou non au nombre de variables. Il est également possible d'obtenir les axes et composantes sans calculer explicitement ces matrices par des algorithmes itératifs. Ces méthodes sont moins coûteuses car seuls les S premiers vecteurs propres et valeurs propres sont calculés. Les solutions de l'ACP étant emboîtées, le choix de la dimension S du sous-espace est sans impact sur l'interprétation des axes. Nous présentons deux de ces algorithmes dans le cas complet, l'algorithme NIPALS et l'algorithme que nous appellerons de "recherche directe du sous-espace".

2.2 ACP via les moindres carrés alternés

2.2.1 NIPALS

Présentation de l'algorithme NIPALS En ACP, l'algorithme NIPALS, proposé par Wold en 1966 [46], détermine les axes et composantes, dimension par dimension, en alternant deux étapes de régressions simples : une pour calculer les axes et l'autre les composantes. Cette méthode consiste, dans un premier temps, à rechercher la meilleure approximation de rang un de la matrice X, c'est-à-dire F_1 et u_1 qui minimisent le critère :

$$\mathcal{C}_1 = \sum_{i=1}^{I}\sum_{k=1}^{K}(x_{ik} - F_{i1}u_{k1})^2. \tag{2}$$

Pour cela, F_1 (resp. u_1) est estimé conditionnellement à u_1 (resp. F_1) par moindres carrés (estimations notées respectivement $\hat{F}_1$ et $\hat{u}_1$). NIPALS revient ainsi à itérer les étapes suivantes :

$$\begin{cases} \frac{\partial \mathcal{C}_1}{\partial u_{k1}} = 0 \Rightarrow \hat{u}_{k1} = \frac{\sum_i (x_{ik} \times F_{i1})}{\sum_i F_{i1}^2}, \text{ pour } k = 1, ..., K, \\ \frac{\partial \mathcal{C}_1}{\partial F_{i1}} = 0 \Rightarrow \hat{F}_{i1} = \frac{\sum_k (x_{ik} \times u_{k1})}{\sum_k u_{k1}^2}, \text{ pour } i = 1, ..., I. \end{cases}$$

A chaque étape, le critère (2) diminue et le minimum global est atteint à la convergence. L'étape de normalisation de l'axe doit être effectuée en fin d'algorithme pour retrouver la solution de l'ACP. Cependant, il est usuel d'intégrer cette étape à l'intérieur de l'algorithme pour éviter des problèmes d'instabilité numérique liés à l'augmentation des normes à chaque pas.

Une fois la première dimension $(\hat{F}_1, \hat{u}_1)$ obtenue, les données reconstituées $\hat{F}_1 \hat{u}_1'$ sont soustraites de la matrice X. Sur cette nouvelle matrice $\tilde{X} = X - \hat{F}_1 \hat{u}_1'$, souvent appelée matrice des résidus, la même procédure de recherche de la première dimension est appliquée ($\hat{F}_2$ et $\hat{u}_2$ sont obtenus). Les axes et composantes sont ainsi calculés séquentiellement.

Remarque 2.1. Comme les $\hat{F}_k$ et $\hat{u}_k$ successifs sont orthogonaux, le critère global (1) est minimisé. On retrouve une propriété bien connue en régression : une régression multiple effectuée sur deux variables explicatives orthogonales X_1 et X_2 revient à effectuer une régression simple des résidus de la première régression (sur X_1) sur X_2.

Remarque 2.2. Les deux étapes de l'algorithme NIPALS pour obtenir une dimension correspondent à une itération de la méthode de la puissance itérée [15], algorithme de calcul d'un vecteur propre dominant par multiplications successives de la matrice à diagonaliser. En effet, les étapes $u_1 = X'F_1$ et $F_1 = Xu_1$ sont alternées, ce qui revient à calculer $u_1^\ell = X'Xu_1^{\ell-1}$ à l'itération ℓ. A convergence, les vecteurs u_1 et F_1 sont donc respectivement les vecteurs propres de la matrice de variance-covariance et de la matrice de produit-scalaire associés à la plus grande valeur propre λ_1. La convergence de l'algorithme est d'autant plus rapide que le rapport $\frac{\lambda_2}{\lambda_1}$ est petit.

Interprétation géométrique L'obtention d'une dimension par l'algorithme NIPALS est illustrée figure 1. L'algorithme consiste à projeter les K variables (de $\mathbb{R}^I$) sur une composante

- Soit F_1 une variable au hasard de $\mathbb{R}^I$
- Répéter jusqu'à la convergence :
  - pour $k = 1, ..., K$:
    $\hat{u}_{k1} = \langle x_{.k}, \frac{F_1}{\|F_1\|^2} \rangle$
  - normalisation de u_1 :
    $\hat{u}_1 = \frac{\hat{u}_1}{\|\hat{u}_1\|}$
  - pour $i = 1, ..., I$:
    $\hat{F}_{i1} = \langle x_{i.}, \frac{u_1}{\|u_1\|^2} \rangle$

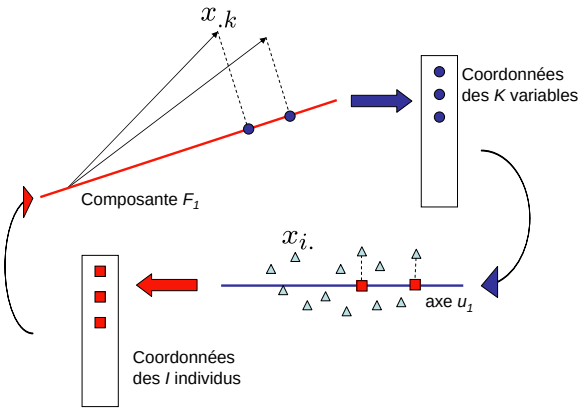

FIGURE 1 – *Fonctionnement de l'algorithme NIPALS.*

principale F_1 (au départ un vecteur quelconque de $\mathbb{R}^I$) et à récupérer leurs coordonnées qui

constituent un nouvel axe u_1 (de $\mathbb{R}^K$). Les I individus sont ensuite projetés sur cet axe et leurs coordonnées engendrent une nouvelle composante (de $\mathbb{R}^I$). Ces étapes sont alternées jusqu'à la convergence. L'algorithme NIPALS est décrit ici pour chaque individu et pour chaque variable pour faciliter l'extension au cas incomplet. De la même manière, la division par $\|u_1\|^2$ dans $\hat{F}_{i1} = \langle x_{i.}, \frac{u_1}{\|u_1\|^2} \rangle$ est inutile dans le cas complet (car $\|u_1\|^2 = 1$) mais est nécessaire quand des données sont manquantes.

2.2.2 Algorithme de "recherche directe du sous-espace"

Le second algorithme dit de "recherche directe du sous-espace" consiste à minimiser le critère (1) en alternant deux étapes de régressions multiples en lieu et place des régressions simples. Le sous-espace de dimension $S \geq 1$ est alors obtenu directement plutôt que séquentiellement. Cet algorithme consiste à itérer les deux étapes suivantes :

$$\begin{cases} \frac{\partial \mathcal{C}}{\partial u_{k.}} = 0 & \Rightarrow \hat{u}_{k.} = (F'F)^{-1}F'x_{.k}, \text{ pour } k = 1, ..., K, \\ \frac{\partial \mathcal{C}}{\partial F_{i.}} = 0 & \Rightarrow \hat{F}_{i.} = x_{i.}u(u'u)^{-1}, \text{ pour } i = 1, ..., I. \end{cases}$$

A chaque étape, le critère (1) diminue et le minimum global est atteint à la convergence. Le sous-espace obtenu est orthonormalisé pour retrouver la solution de l'ACP. L'interprétation géométrique de l'algorithme est similaire à celle de NIPALS, les variables et les individus étant cette fois projetés sur des sous-espaces engendrés par des vecteurs de dimension supérieure à 1.

3 Minimisation du critère des moindres carrés pondérés

Une solution classique pour gérer les données manquantes en ACP consiste à introduire, dans le critère à minimiser (1), une matrice de poids W telle que $w_{ik} = 0$ si x_{ik} est manquant et $w_{ik} = 1$ sinon :

$$\mathcal{C} = \|W * (X - Fu')\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}(w_{ik}x_{ik} - \sum_{s=1}^{S} F_{is}w_{ik}u_{ks})^2, \tag{3}$$

avec $*$ le produit d'Hadamard. La matrice X est supposée dans un premier temps centrée par colonne sur les éléments présents. Cette écriture du critère met en évidence que les données manquantes ont un poids nul. Les axes et composantes sont obtenus en minimisant l'erreur de reconstitution sur les données observées uniquement (les données manquantes sont "sautées"). Contrairement au cas complet, le critère des moindres carrés pondérés (3) n'a pas de solution explicite et il est nécessaire de recourir à des algorithmes itératifs.

Les algorithmes, présentés dans la section précédente, sont étendus au cas incomplet : deux étapes de régressions simples pondérées sont alternées dans l'algorithme NIPALS et deux étapes de régressions multiples pondérées sont alternées dans l'algorithme de "recherche directe du sous-espace". Un algorithme d'ACP itérative, minimisant aussi le critère (3), et qui apporte un éclairage différent sur la gestion des données manquantes, est ensuite détaillé.

3.1 NIPALS avec données manquantes

Le problème des données manquantes en ACP a été discuté pour la première fois par Wold en 1966 [45] et étudié par Christofferson en 1969 [8], élève de Wold, qui a détaillé l'algorithme NIPALS sur données incomplètes. Le premier axe et la première composante sont ainsi obtenus en alternant les étapes :

$$\hat{u}_{k1} = \frac{\sum_i (w_{ik} x_{ik} F_{i1})}{\sum_i w_{ik} F_{i1}^2}, \text{ pour } k = 1, ..., K, \tag{4}$$

$$\hat{u}_1 = \frac{\hat{u}_1}{\|\hat{u}_1\|},$$

$$\hat{F}_{i1} = \frac{\sum_k (w_{ik} x_{ik} u_{k1})}{\sum_k w_{ik} u_{k1}^2}, \text{ pour } i = 1, ..., I. \tag{5}$$

A chaque étape, le critère $\sum_{i=1}^{I} \sum_{k=1}^{K} w_{ik}(x_{ik} - F_{i1}u_{k1})^2$ diminue et à la convergence un minimum, éventuellement local, est atteint [25]. Comme dans le cas complet, les dimensions suivantes sont obtenues par déflation, ce qui assure l'emboîtement des solutions.

Les dénominateurs des équations (4) et (5) ne sont pas les mêmes par individu et par variable car les projections réalisées sont spécifiques. En effet, la matrice W implique que le poids des variables est différent d'un individu à l'autre et que le poids des individus est différent d'une variable à l'autre (cf. matrice W de la figure 2). Ainsi, on associe à l'individu i une métrique $M_i = \text{diag}(W_{i.})$ et à la variable k une métrique $D_k = \text{diag}(W_{.k})$, comme illustré figure 2. La variable k (resp. l'individu i) est donc projetée sur F_1 (resp. u_1) avec sa métrique propre :

$$\hat{u}_{k1} = (F_1' D_k F_1)^{-1} F_1' D_k x_{.k},$$
$$\hat{F}_{i1} = (u_1' M_i u_1)^{-1} u_1' M_i x_{i.}.$$

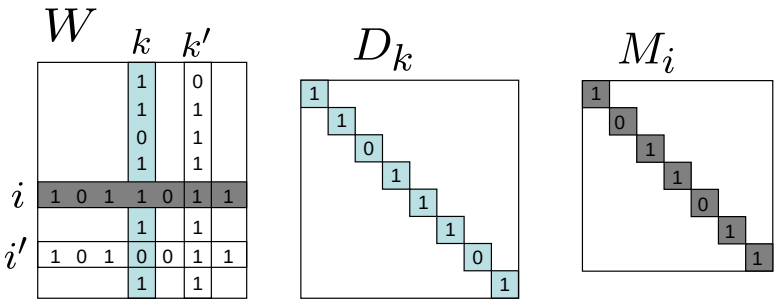

FIGURE 2 – *Lien entre W et les métriques associées.*

NIPALS fournit des solutions acceptables en pratique avec peu de données manquantes [4, 39]. Cependant, quand le nombre de données manquantes augmente, la procédure devient instable car les erreurs se propagent axe après axe. De plus, contrairement au cas complet, les dimensions obtenues ne sont pas orthogonales et par suite le critère global (3) n'est pas minimisé par la procédure séquentielle. Pour cette raison, NIPALS fait l'objet de nombreuses critiques [14, 17, 24] et les auteurs suggèrent de s'orienter vers la méthode optimale de "recherche directe du sous-espace". Par la suite, nous privilégierons des méthodes qui minimisent le critère (3).

3.2 "Recherche directe du sous-espace" avec données manquantes

La méthode de recherche directe du sous-espace peut aussi être étendue au cas incomplet. Pour cela, le critère (3) est décomposé par ligne et par colonne :

$$\mathcal{C} = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(w_{ik}x_{ik} - \sum_{s=1}^{S} F_{is}w_{ik}u_{ks}\right)^2 = \sum_{i=1}^{I}\mathcal{C}_i = \sum_{k=1}^{K}\mathcal{C}_k,$$

avec

$$\mathcal{C}_i = \sum_{k=1}^{K}\left(w_{ik}x_{ik} - \sum_{s=1}^{S} F_{is}w_{ik}u_{ks}\right)^2,$$

et

$$\mathcal{C}_k = \sum_{i=1}^{I}\left(w_{ik}x_{ik} - \sum_{s=1}^{S} F_{is}w_{ik}u_{ks}\right)^2.$$

Il est clair que minimiser $\mathcal{C}$ revient à minimiser chacun des $\mathcal{C}_i$ ou chacun des $\mathcal{C}_k$. Ces deux décompositions permettent de définir les deux étapes de l'algorithme itératif. Pour une ligne i fixée et pour $u_{k.}$ connu, $F_{i.}$ est obtenu en minimisant $\mathcal{C}_i$. La solution est alors :

$$\hat{F}_{i.} = x_{i.}A^i(A^{i\prime}A^i)^{-1},$$

avec A^i la matrice de terme général $A^i_{ks} = w_{ik}u_{ks}$. Chaque ligne de la matrice F correspond donc aux coefficients de la régression multiple pondérée des lignes de X sur les colonnes de u. De la même manière, pour une colonne k donnée et pour $F_{i.}$ connu, $u_{k.}$ est obtenu en minimisant $\mathcal{C}_k$. La solution est alors :

$$\hat{u}_{k.} = (B^{k\prime}B^k)^{-1}B^{k\prime}x_{.k},$$

avec B^k la matrice de terme général $B^k_{is} = F_{is}w_{ik}$. Chaque ligne de la matrice u correspond donc aux coefficients de la régression multiple pondérée des colonnes de X sur celles de F.

A chaque étape, le critère (3) diminue et à la convergence un minimum, éventuellement local, est atteint [10, 25, 37]. Le sous-espace obtenu à la convergence est, comme dans le cas complet, orthonormalisé.

L'interprétation géométrique de cette méthode est identique à celle de NIPALS : les projections des individus et des variables sont effectuées avec des métriques spécifiques sur des sous-espaces de dimension supérieure à 1.

Remarque 3.1. Une fois les axes et composantes estimés ($\hat{u}$ et $\hat{F}$), il est possible de compléter les données manquantes par la formule de reconstitution $\hat{X} = \hat{F}\hat{u}'$. L'ACP réalisée sur le tableau de données ainsi complété fournit le même sous-espace que celui obtenu à la convergence de l'algorithme.

Remarque 3.2. Cet algorithme est également connu sous le nom de "criss cross multiple regression" [14] quand W est une matrice de poids quelconques.

3.3 ACP itérative

3.3.1 Présentation de la méthode

Kiers [23] s'est intéressé à la minimisation du critère général de moindres carrées pondérés $||W * (X - \mathcal{M})||^2$ avec W une matrice de poids quelconques et $\mathcal{M}$ un modèle général pour les données. Inspiré par les travaux d'Heiser [20], il a montré qu'il était possible de minimiser ce critère dès lors qu'un algorithme existe pour minimiser le critère de moindres carrées ordinaires $||X - \mathcal{M}||^2$. Son approche est basée sur l'utilisation itérative de l'algorithme issu de ces moindres carrées ordinaires. Dans le cas particulier du modèle bilinéaire de l'ACP où $\mathcal{M} = Fu'$ et d'une matrice de poids constituée uniquement de 0 et de 1, minimiser le critère (3) peut se faire en minimisant le critère (1) à chaque pas de l'algorithme :

1. initialisation $\ell = 0$: X^0 est obtenu en remplaçant les valeurs manquantes par une valeur initiale, comme par exemple la moyenne de chaque variable
2. itération ℓ :
   (a) recherche de $(\hat{F}^\ell, \hat{u}^\ell)$ comme les paramètres (F, u) qui minimisent le critère $\|X^{\ell-1} - Fu'\|^2$, S dimensions sont retenues
   (b) X^ℓ est obtenu en remplaçant les valeurs manquantes de X par les valeurs reconstituées $\hat{X}^\ell = \hat{F}^\ell \hat{u}^{\ell\prime}$. Le nouveau tableau complété peut s'écrire $X^\ell = W * X + (1 - W) * \hat{X}^\ell$
3. les étapes (a) et (b) sont répétées jusqu'à la convergence

Cette procédure consiste finalement à effectuer des ACP de façon itérative sur des tableaux de données complets. En effet, les données manquantes sont initialement remplacées par des valeurs quelconques. L'ACP est ensuite réalisée sur le tableau de données complété. Les S premières dimensions sont retenues et à partir de ces axes et composantes, les données manquantes sont estimées par la formule de reconstitution. La procédure est répétée jusqu'à la convergence.

Notons que dans l'étape 2.a, il est possible de rechercher $(\hat{F}^\ell, \hat{u}^\ell)$ comme les paramètres (F, u) qui diminuent le critère $\|X^{\ell-1} - Fu'\|^2$ et non plus comme les paramètres qui le minimisent. Ainsi, il est possible d'effectuer une seule étape des moindres carrés alternés à la place de l'ACP, ce qui diminue le coût de calcul. L'étape 2.a s'écrit alors :

$$\begin{aligned} \hat{u}^\ell &= X^{\ell-1\prime} \hat{F}^{\ell-1} (\hat{F}^{\ell-1\prime} \hat{F}^{\ell-1})^{-1}, \\ \hat{F}^\ell &= X^{\ell-1} \hat{u}^\ell (\hat{u}^{\ell\prime} \hat{u}^\ell)^{-1}. \end{aligned}$$

Kiers [23] a montré que la procédure converge (le critère 2.a diminue à chaque itération et est borné par 0) vers un minimum éventuellement local.

3.3.2 Estimation - imputation

Dans cette méthode, les paramètres inconnus (axes et composantes) et les données manquantes sont estimés simultanément. Les données manquantes sont imputées sans que cela influe sur la construction des axes et composantes au sens suivant : la contribution des données

imputées dans l'ACP diminue à chaque étape et est nulle à la convergence. Montrons que le critère minimisé à partir de la matrice de données imputées peut, à la convergence, se réécrire en fonction des valeurs observées uniquement. A l'itération ℓ, le minimum du critère est atteint pour $(\hat{F}^\ell, \hat{u}^\ell)$ et vaut :

$$\|\hat{\varepsilon}^\ell\|^2 = \|X^{\ell-1} - \hat{F}^\ell \hat{u}^{\ell\prime}\|^2.$$

Ce minimum peut se réécrire :

$$\|\hat{\varepsilon}^\ell\|^2 = \|W * X + (1-W) * \hat{F}^{\ell-1}\hat{u}^{\ell-1\prime} - \hat{F}^\ell \hat{u}^{\ell\prime}\|^2.$$

A convergence, $\hat{F}^{\ell-1} = \hat{F}^\ell$ et $\hat{u}^{\ell-1} = \hat{u}^\ell$, et donc :

$$\|\hat{\varepsilon}^\ell\|^2 = \|W * (X - \hat{F}^\ell \hat{u}^{\ell\prime})\|^2.$$

L'erreur de reconstitution est bien minimisée sur les données présentes. Les valeurs imputées peuvent être considérées comme des intermédiaires de calculs qui n'interviennent pas dans le critère final. "Sauter" les données manquantes ou imputer selon le modèle est ici équivalent.

Remarque 3.3. Notons que le principe général qui consiste à alterner imputation et estimation jusqu'à la convergence est ancien et remonte à [19]. Dans le cadre de l'analyse des données, cette procédure d'imputation itérative a été initialement proposée en Analyse Factorielle des Correspondances par Nora-Chouteau [28] et par Greenacre [16, p.238].

L'algorithme d'ACP itérative est souvent appelé ACP-EM pour Expectation-Maximisation en référence aux idées d'alterner des étapes d'estimation et d'imputation présentes dans l'algorithme EM [9]. Cette formulation peut tout de même paraître surprenante car aucune référence à un modèle n'est spécifiée. Nous formalisons cette idée intuitive que l'ACP itérative est un algorithme EM dans le paragraphe suivant.

3.3.3 Un algorithme EM

L'ACP peut aussi s'écrire comme un modèle à effet fixe [7] signal plus bruit :

$$x_{ik} = \sum_{s=1}^{S} F_{is} u_{ks} + \varepsilon_{ik}, \text{ avec } \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2). \tag{6}$$

Dans le cas complet, les estimateurs du maximum de vraisemblance de u et de F (sous la contrainte $u'u = \mathbb{I}_S$) sont respectivement les vecteurs propres de la matrice de variance-covariance et de la matrice de produit-scalaire. Ainsi, comme en régression, les estimateurs du maximum de vraisemblance correspondent aux estimateurs des moindres carrés. Pour estimer les paramètres sur données incomplètes, il faut maximiser la vraisemblance observée (on suppose le processus à l'origine des données manquantes ignorable [26]).

En général, dans le cas incomplet, la vraisemblance observée est difficile à calculer. L'algorithme EM détourne cette difficulté selon le principe suivant : la vraisemblance des données observée X_{obs} est maximisée indirectement par des maximisations successives de la vraisemblance des données complètes $X = (X_{obs}, X_{miss})$. Les calculs sont simplifiés par cette "augmentation" de données car la vraisemblance des données complètes est plus facile à maximiser.

Cependant, comme les éléments de cette dernière ne sont pas tous observés, elle est remplacée par son espérance conditionnelle sachant les données observées et les paramètres de l'itération en cours. Les deux étapes E de calcul d'Espérance conditionnelle et M de Maximisation de l'espérance de la vraisemblance complète sont alternées jusqu'à la convergence.

Pour le modèle (6), si toutes les données sont observées, la log-vraisemblance complète s'écrit :

$$L_c(F,u,\sigma^2) = -\frac{IK}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}||X - Fu'||^2.$$

L'étape E correspond au calcul de l'espérance de x_{ik} et de x_{ik}^2 sachant les observations et les valeurs courantes des paramètres à l'itération ℓ :

$$\mathbb{E}(x_{ik}|X_{obs},\hat{F}^\ell,\hat{u}^\ell,\hat{\sigma}^{2\,(\ell)}) = \begin{cases} x_{ik} & \text{si } x_{ik} \text{ est observé,} \\ \hat{x}_{ik} = \sum_{s=1}^{S} \hat{F}_{is}^\ell \hat{u}_{ks}^\ell & \text{sinon.} \end{cases}$$

$$\mathbb{E}(x_{ik}^2|X_{obs},\hat{F}^\ell,\hat{u}^\ell,\hat{\sigma}^{2\,(\ell)}) = \begin{cases} x_{ik}^2 & \text{si } x_{ik} \text{ est observé,} \\ (\sum_{s=1}^{S} \hat{F}_{is}^\ell \hat{u}_{ks}^\ell)^2 + \hat{\sigma}^{2\,(\ell)} & \text{sinon.} \end{cases}$$

L'étape M correspond à la maximisation de l'espérance de la log-vraisemblance complète

$$\begin{aligned} L_c(F,u,\sigma^2) \;=\; & -\frac{IK}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{ik\in obs}(x_{ik} - \sum_{s=1}^{S} F_{is}u_{ks})^2 \\ & -\frac{1}{2\sigma^2}\sum_{ik\in miss}\left((\hat{x}_{ik} - \sum_{s=1}^{S} F_{is}u_{ks})^2 + \hat{\sigma}^{2\,(\ell)}\right), \end{aligned}$$

et fournit $\hat{F}^{\ell+1}$, $\hat{u}^{\ell+1}$ et $\hat{\sigma}^{2\,(\ell+1)}$. Les estimateurs de F et de u ne nécessitent pas la connaissance de $\hat{\sigma}^{2\,(\ell)}$. L'algorithme EM peut ainsi être simplifié en ne calculant pas $\mathbb{E}(x_{ik}^2)$ à l'étape E et $\hat{\sigma}^2$ à l'étape M. Les nouvelles estimations des paramètres $\hat{F}^{\ell+1}$ et $\hat{u}^{\ell+1}$ sont alors obtenues en réalisant l'ACP sur le tableau de données complété à l'étape E. Ainsi, cet algorithme EM correspond exactement à l'algorithme d'ACP itérative, d'où le nom ACP-EM. Cette équivalence positionne l'ACP itérative dans la théorie des données manquantes. Elle bénéficie ainsi des propriétés du maximum de vraisemblance et des caractéristiques des algorithmes EM. Nous pouvons noter que l'algorithme d'ACP itérative avec étape des moindres carrés correspond à un algorithme GEM, pour Generalized EM [9], où la vraisemblance est augmentée à chaque étape et non plus maximisée. Par la suite, nous appellerons cet algorithme ACP-GEM.

Remarque 3.4. L'équivalence entre une procédure d'imputation itérative et un algorithme EM se retrouve aussi en régression lors de l'estimation de β [26, p.238]. Cependant, l'algorithme EM n'est pas toujours équivalent à imputer les données, réaliser l'analyse, imputer et itérer. Ce n'est généralement pas un algorithme d'imputation.

L'algorithme EM présenté ici est particulier car il impute les données manquantes et estime simultanément les axes et composantes. Cet algorithme préserve ainsi la dualité, caractéristique intrinsèque de l'analyse de données. L'algorithme de "recherche directe du sous-espace" possède aussi cette caractéristique.

Remarque 3.5. Wentzell *et al.* [44] ont étudié un modèle similaire au modèle (6) mais avec une variance des résidus non constante. Les données manquantes peuvent être gérées en attribuant aux cellules correspondantes une variance infinie. Cela revient à leur attribuer un poids nul et donc à minimiser le critère de moindres carrés pondérés.

4 Propriétés

L'algorithme d'ACP itérative et celui de "recherche directe du sous-espace" apportent un éclairage différent sur la gestion des données manquantes en ACP qui permet de mieux comprendre les propriétés de la méthode. L'introduction d'une matrice de poids dans le critère (1) rend l'optimisation complexe ce qui engendre plusieurs problèmes, d'où l'intérêt de s'appuyer sur plusieurs points de vue.

4.1 Imputation implicite

L'algorithme de "recherche directe du sous-espace" peut être perçu comme un algorithme qui "saute" les données manquantes via l'introduction d'une matrice de poids dans le critère. L'ACP itérative peut être vue comme une méthode d'imputation simple. Du fait de la similarité entre ces deux algorithmes, il apparaît donc que l'ACP itérative "saute" les données manquantes et que l'algorithme de "recherche directe du sous-espace" est un algorithme d'imputation simple dans lequel l'imputation est implicite. De cette imputation découle plusieurs propriétés : la nécessité d'un recentrage et la réduction de la variabilité.

Il est important de remarquer que l'imputation par ACP est intéressante en tant que telle pour compléter des tableaux avant d'éventuellement réaliser l'analyse statistique de son choix. En effet, elle est fondée sur la dualité et tient compte simultanément des ressemblances entre individus et des liaisons entre variables. De plus, l'imputation est globale car toutes les valeurs sont imputées simultanément.

4.2 Centrage, réduction

En ACP, les variables sont toujours centrées. Or, les réestimations successives des données manquantes décentrent le jeu de données. Il est donc important d'intégrer une étape de recentrage à l'intérieur de la procédure d'estimation. L'algorithme d'ACP-GEM devient ainsi :

1. initialisation $\ell = 0$: X^0 est obtenu en remplaçant les valeurs manquantes par une valeur initiale, comme par exemple la moyenne de chaque variable. Le tableau de données est centré par la moyenne de X^0 définie comme $\hat{m}^0$ et est noté X_c^0
2. itération ℓ :
   (a) à partir du tableau de données complété centré $X_c^{\ell-1}$:
      i. $\hat{u}^\ell = X_c^{(\ell-1)\prime}\hat{F}^{\ell-1}(\hat{F}^{(\ell-1)\prime}\hat{F}^{\ell-1})^{-1}$
      ii. $\hat{F}^\ell = X_c^{\ell-1}\hat{u}^\ell(\hat{u}^{\ell\prime}\hat{u}^\ell)^{-1}$
   (b) imputation des données manquantes (ce qui décentre les données)

i. ajout de la moyenne $\hat{m}^{(\ell-1)}$ et calcul de la nouvelle moyenne $\hat{m}^{\ell}$

ii. recentrage des données par $\hat{m}^{\ell}$

3. les étapes sont répétées jusqu'à la convergence. Les données imputées sont obtenues en rajoutant $\hat{m}^{L}$

Les étapes $b.i$ et $b.ii$ permettent de conserver les changements de moyenne en vue de l'imputation finale.

L'exemple fictif ci-dessous illustre l'intérêt du recentrage à partir d'un jeu de données en deux dimensions. Les variables X_1 et X_2, respectivement X_3 et X_4 sont identiques. Deux données sont manquantes pour x_{62} et x_{54} :

	X_1	X_2	X_3	X_4
i1	-2	-2	-1	-1
i2	0	0	-1	-1
i3	2	2	-1	-1
i4	-2	-2	1	1
i5	0	0	1	NA
i6	2	NA	1	1

L'algorithme ACP-GEM est utilisé avec et sans centrage. Avec centrage, les valeurs sont parfaitement reconstituées ($\hat{x}_{62} = 2$ et $\hat{x}_{54} = 1$), sans centrage elles valent respectivement $\hat{x}_{62} = 1.41$ et $\hat{x}_{54} = 0.76$. Par suite, les axes et composantes sont différents de ceux obtenus sur le tableau complet.

Dans l'algorithme de "recherche directe du sous-espace", cette étape de recentrage est moins naturelle mais il est nécessaire de l'incorporer à cause de l'imputation implicite. Il faut alors introduire, à chaque itération, une étape de reconstitution dédiée exclusivement au recentrage.

Selon le même principe, en ACP normée, l'écart-type des variables doit être recalculé à chaque itération.

Remarque 4.1. Cette étape de recentrage est aussi nécessaire dans l'algorithme NIPALS. Par contre il n'est pas possible d'incorporer l'étape de réduction en raison de la procédure de déflation. Il est possible de réduire les variables en amont de l'analyse et de réaliser une ACP non-normée. Il subsiste cependant un problème : les variables n'ont plus le même poids au fur et à mesure de l'algorithme.

Remarque 4.2. Cette étape de recentrage est indispensable et entraîne finalement une modification du critère à minimiser. Tous les algorithmes présentés avec étape de recentrage minimisent en fait le critère :

$$\sum_{i=1}^{I}\sum_{k=1}^{K} w_{ik}(x_{ik} - m_k - \sum_{s=1}^{S} F_{is}u_{ks})^2,$$

avec $m = (m_k)_{k=1,\ldots,K}$ un vecteur de $\mathbb{R}^K$. Cette présentation implique que m est un paramètre à estimer au même titre que F et u et que son estimation est mise à jour à chaque étape des algorithmes.

4.3 Minima locaux

Dans le cas complet, tous les algorithmes convergent vers le minimum global. Dans le cas incomplet, les algorithmes peuvent converger vers des minima locaux car la surface étudiée est très chahutée au sens où elle présente de nombreux modes. Le choix de la position initiale est donc important et plusieurs solutions ont été envisagées [14]. La plus simple consiste à initialiser les algorithmes par les axes et composantes de l'ACP réalisée sur le tableau de données dans lequel les données manquantes ont été complétées par la moyenne de chaque variable. Une autre initialisation envisageable consiste à utiliser différentes méthodes pour estimer la matrice de variance-covariance et ensuite obtenir une première estimation des axes principaux. Il est par exemple possible d'estimer la matrice de variance-covariance à partir des observations disponibles (calcul par paire de variables) ou par maximum de vraisemblance [26]. Cependant, l'utilisation d'initialisations aléatoires est en pratique l'approche la plus satisfaisante pour explorer au mieux l'ensemble des solutions possibles.

Pour pallier ce problème de dépendance aux points de départ, d'autres algorithmes d'optimisation avancés ont été proposés dans la littérature. Srebro [37] évoque par exemple l'utilisation d'une variante de la méthode de descente de gradient qui permet d'éviter les minima locaux et de converger plus rapidement. Buchanan & Fitzigibbon [5] proposent, quant à eux, un algorithme "hybride" combinant des moindres carrés alternés et la méthode de Newton pour explorer l'ensemble des solutions. Des méthodes de recuit simulé peuvent aussi être envisagées.

4.4 Choix du nombre d'axes

Les solutions des algorithmes qui minimisent le critère (3) ne sont pas emboîtées : la solution à $S-1$ dimensions n'est pas incluse dans la solution à S dimensions. Dès lors, le choix du nombre d'axes apparaît comme important. Ce choix, effectué *a priori*, a un impact sur la qualité de la reconstitution des données et sur les solutions de l'ACP (axes et composantes). Il faut sélectionner suffisamment de dimensions pour obtenir une estimation satisfaisante des paramètres sans pour autant avoir une valeur de S trop élevée. En effet, les dernières dimensions ne contiennent pas d'information sur la structure des données et sont souvent considérées comme du bruit. Dès lors, leur prise en compte peut rendre la procédure instable. De plus, des problèmes de surajustements peuvent apparaître en raison du grand nombre de paramètres à estimer (cf. §4.5).

Différentes stratégies sont envisageables pour choisir la dimension du sous-espace S. Une première possibilité, empirique, consiste à représenter le diagramme en bâtons des valeurs propres obtenu après imputation des données manquantes par la moyenne de chaque variable. L'inconvénient est que l'imputation peut déformer le nuage de points et conduire à une estimation erronée des valeurs propres. Une alternative, utilisée par Stacklies *et al.* [38] et Walczak & Massart [43], consiste à chercher la valeur S qui minimise l'erreur de reconstitution des données. Pour cela, la procédure de validation croisée suivante est employée : des données sont supprimées, reconstruites avec s dimensions, s variant de 1 à $K-1$, et l'erreur de reconstitution est calculée. Cette opération est répétée plusieurs fois et une valeur de s optimale est déterminée. Bien entendu, cette procédure est peu satisfaisante sur des tableaux de données déjà incomplets et est de plus coûteuse en temps de calcul. En pratique, les deux méthodes sont utilisées et le

choix reste très empirique.

4.5 Surajustement

Des problèmes de surajustement peuvent survenir : le critère (3) peut être faible sur les données d'apprentissage (les données présentes) mais la qualité de prédiction (l'estimation des paramètres) très mauvaise. Cette situation peut apparaître dès lors que le nombre de paramètres à estimer, ici $(I + K) \times S$, est important par rapport au nombre de données disponibles qui vaut au maximum $I \times K$. Ainsi, dès que le nombre de données manquantes est grand, ou que la dimension S du sous-espace est élevée, les algorithmes peuvent être instables et converger vers des solutions loin de l'optimum global.

Raiko *et al.* [30] illustrent ce problème de surajustement sur un exemple fictif d'un tableau de données avec deux variables X_1 et X_2. Seuls deux individus ont des enregistrements complets. Les autres individus ont une valeur manquante soit sur X_1 ou sur X_2. La solution qui minimise le critère (3) est définie par la droite qui passe par les deux individus complets. Raiko *et al.* précisent alors qu'il n'y a aucune raison de "croire" à l'existence d'une liaison significative entre les deux variables car cette information n'est portée que par deux individus. Une solution plus raisonnable consisterait à considérer la corrélation observée comme illusoire et à reconstituer les valeurs manquantes par la moyenne de chaque variable. Cet exemple très simple et caricatural en deux dimensions illustre parfaitement les problèmes qui peuvent se rencontrer en dimension supérieure.

Pour éviter des problèmes de surajustement, une première stratégie consiste à rechercher un sous-espace de dimension inférieure permettant d'estimer moins de paramètres. Cependant, les autres dimensions peuvent porter une information importante. Une autre solution consiste à contraindre la norme des paramètres en ajoutant un terme de pénalité au critère (3) :

$$\|W * (X - Fu')\|^2 + \tau_1 \|F\|^2 + \tau_2 \|u\|^2. \tag{7}$$

Bien entendu, cela nécessite de choisir des paramètres de régularisation, ce qui est souvent délicat.

Dans l'algorithme ACP-GEM, le surajustement peut se traduire par des matrices $(F'F)$ et $(u'u)$ mal conditionnées. Il est alors possible de réaliser des régressions ridges à la place des régressions. Nous montrons dans la section 5 que la formulation probabiliste de l'ACP due à Tipping & Bishop [41] offre des paramètres ridges adaptés à la gestion des données manquantes.

4.6 Réduction de la variabilité

A convergence de l'algorithme ACP-GEM, les données manquantes sont complétées par la formule de reconstitution $\hat{F}\hat{u}'$. Ainsi, la variabilité du jeu de données final est sous-estimée. En effet, le terme d'erreur est absent pour les données reconstituées, ces dernières ont été "créées" par le modèle c'est-à-dire imputées sur le sous-espace. Les autres algorithmes de moindres carrées alternés, qui imputent implicitement, souffrent aussi de cette diminution de la variabilité. Cela se traduit donc par des pourcentages d'inertie optimistes qu'il faut interpréter avec prudence.

5 L'ACP Probabiliste

Plusieurs auteurs ont travaillé sur des modèles pour l'ACP [11]. Récemment, Tipping & Bishop [41] et indépendamment Roweis [31] ont proposé l'ACP Probabiliste (ACPP) qui leur a permis entre autres d'étendre l'ACP aux modèles de mélanges [40], de proposer une formulation bayésienne de l'ACP [2], etc.

L'objet de cette section est de montrer comment l'ACPP peut être utilisée pour limiter le problème de surajustement de l'algorithme ACP-GEM.

5.1 Présentation du modèle

L'ACPP est un modèle d'Analyse en Facteurs Communs et Spécifiques (AFCS) particulier. En AFCS [1], l'objectif est d'expliquer les liaisons entre les K variables (de la matrice $X_{I\times K}$ complète, supposée centrée) à l'aide d'un petit nombre (S) de variables latentes regroupées dans la matrice $Z_{I\times S}$. Le modèle est le suivant :

$$x_{i.} = \Gamma z_{i.} + e_{i.},$$

avec $\Gamma_{K\times S}$ une matrice de coefficients et les hypothèses classiques sur les facteurs communs et le terme d'erreur : $z_{i.} \sim \mathcal{N}(0, \mathbb{I}_S)$, $e_{i.} \sim \mathcal{N}(0, \Psi)$ et les variables aléatoires $z_{i.}$ et $e_{i.}$ indépendantes. La matrice Ψ est supposée diagonale ce qui assure la propriété fondamentale d'indépendance conditionnelle : $x_{i.}|z_{i.} \sim \mathcal{N}(\Gamma z_{i.}, \Psi)$. L'ACPP est un modèle d'AFCS dans lequel le bruit est isotrope : $\Psi = \sigma^2\mathbb{I}_K$. Contrairement au modèle présenté §3.3.3, le modèle d'ACPP est un modèle à effets aléatoires où les paramètres sont estimés à partir d'un échantillon de I vecteurs aléatoires indépendants et identiquement distribués :

$$x_{i.} \sim \mathcal{N}(0, \Sigma) \text{ avec } \Sigma = \Gamma\Gamma' + \sigma^2\mathbb{I}_K, \text{ pour } i = 1, ..., I.$$

Les estimateurs du maximum de vraisemblance sont :

$$\hat{\Gamma} = u(\Lambda - \sigma^2\mathbb{I}_S)^{1/2}R \text{ et } \hat{\sigma}^2 = \frac{1}{K-S}\sum_{k=S+1}^{K}\lambda_k,$$

avec u les S premiers vecteurs propres de la matrice de variance-covariance empirique, Λ la matrice diagonale des valeurs propres associées, et $R_{S\times S}$ une matrice de rotation quelconque (classiquement $R = \mathbb{I}_S$). L'estimation de σ^2 correspond à la moyenne des variances des dernières dimensions.

5.2 Utilisation de l'ACPP pour limiter le surajustement

5.2.1 Algorithme EM pour l'ACPP (cas complet)

Même s'il existe une solution explicite pour les estimateurs du maximum de vraisemblance, un algorithme EM [34], où les variables latentes Z sont considérées comme manquantes, peut aussi être utilisé pour maximiser la vraisemblance. Les données complètes (si $z_{i.}$ était observé)

sont définies comme le couple $(x_{i.}, z_{i.})$. La loi du couple est égale au produit de la loi conditionnelle par la marginale : $(x_{i.}, z_{i.}) = (x_{i.}|z_{i.})(z_{i.})$. La log-vraisemblance des données complètes s'écrit alors :

$$L_c(\Gamma, \sigma^2) = -\frac{-IK}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}(x_{i.} - \Gamma z_{i.})'(x_{i.} - \Gamma z_{i.}) - \frac{1}{2}\sum_{i=1}^{I} z_{i.}z_{i.}'.$$

L'étape E consiste à calculer la distribution des variables latentes sachant les observations et les paramètres de l'itération en cours. Cette distribution obtenue par le théorème de Bayes est la suivante :

$$z_{i.}|x_{i.} \sim \mathcal{N}(M^{-1}\Gamma' x_{i.}, \sigma^2 M^{-1}), \text{ avec } M = \Gamma'\Gamma + \sigma^2 \mathbb{I}_S.$$

L'espérance conditionnelle correspond à une régression régularisée :

$$\mathbb{E}(z_{i.}|x_{i.}) = (\Gamma'\Gamma + \sigma^2\mathbb{I}_S)^{-1}\Gamma' x_{i.}. \tag{8}$$

Remarque 5.1. La régression bayésienne fournit une interprétation probabiliste de la régression ridge [18, p.60]. Soit le modèle de régression linéaire classique $Y = X\beta + \varepsilon$, avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. La régression bayésienne suppose une distribution *a priori* pour β : $\beta \sim \mathcal{N}(0, \tau^2)$. L'estimateur MAP (Maximum A Posteriori) de β vaut alors $\hat{\beta}_r = (X'X + \kappa)^{-1}X'Y$, avec $\kappa = \sigma^2/\tau^2$. Ici, $z_{i.} \sim \mathcal{N}(0, \mathbb{I}_S)$ d'où la régularisation par σ^2.

L'étape M consiste à maximiser l'espérance conditionnelle de la log-vraisemblance complète et fournit :

$$\hat{\Gamma} = \left[\sum_{i=1}^{I} x_{i.}\mathbb{E}(z_{i.}|x_{i.})'\right]\left[\sum_{i=1}^{I}\mathbb{E}(z_{i.}z_{i.}'|x_{i.})\right]^{-1},$$

$$\hat{\sigma}^2 = \frac{1}{IK}\sum_{i=1}^{I}\left(x_{i.}x_{i.}' - 2\mathbb{E}(z_{i.}|x_{i.})'\Gamma' x_{i.} + \text{trace}(\mathbb{E}(z_{i.}z_{i.}'|x_{i.})\Gamma'\hat{\Gamma})\right),$$

avec $\mathbb{E}(z_{i.}z_{i.}'|x_{i.}) = \sigma^2 M^{-1} + \mathbb{E}(z_{i.}|x_{i.})\mathbb{E}(z_{i.}|x_{i.})'$. Ces nouvelles valeurs sont alors utilisées dans l'étape E et les étapes sont répétées jusqu'à la convergence. Matriciellement (en notant $\hat{Z}$ l'estimation de Z par (8)) les étapes qui sont alternées dans l'algorithme EM à chaque itération peuvent se réécrirent de la manière suivante :

$$\text{Etape E} : \hat{Z}' = (\hat{\Gamma}'\hat{\Gamma} + \hat{\sigma}^2\mathbb{I}_S)^{-1}\hat{\Gamma}'X', \tag{9}$$

$$\text{Etape M} : \hat{\Gamma}' = (\hat{Z}'\hat{Z} + I\hat{\sigma}^2\hat{M}^{-1})^{-1}\hat{Z}'X. \tag{10}$$

Les étapes E et M correspondent aux étapes de moindres carrés alternés de l'ACP mais régularisées. Quand le bruit tend vers 0, les étapes de l'ACP sont retrouvées :

$$\text{Etape E} : \hat{Z}' = (\hat{\Gamma}'\hat{\Gamma})^{-1}\hat{\Gamma}'X',$$
$$\text{Etape M} : \hat{\Gamma}' = (\hat{Z}'\hat{Z})^{-1}\hat{Z}'X.$$

Remarque 5.2. A convergence, quand σ^2 tend vers 0, $\hat{\Gamma} = u\Lambda^{1/2}$ et $\hat{Z} = \Lambda^{-1/2}Xu$. $\hat{Z}$ correspond bien aux composantes principales normées à 1 (*i.e.* $\frac{F}{\|F\|}$ avec les notations utilisées en ACP).

5.2.2 Reconstitution par ACPP

Il est possible de reconstituer les données, comme en ACP classique, à partir des S premières dimensions. La matrice reconstituée s'écrit alors

$$\hat{X}_{acpp} = \hat{Z}\hat{\Gamma}' = (X\hat{\Gamma}(\hat{\Gamma}'\hat{\Gamma} + \hat{\sigma}^2\mathbb{I}_S)^{-1})\hat{\Gamma}'.$$

En remplaçant $\hat{\Gamma}$ par son estimation $\hat{u}(\Lambda - \hat{\sigma}^2\mathbb{I}_S)^{1/2}$, $\hat{X}_{acpp} = \sum_{s=1}^{S} \frac{\lambda_s - \hat{\sigma}^2}{\lambda_s}\hat{X}^s_{acp}$, avec $\hat{X}^s_{acp}$ la reconstitution de la matrice X par la s^e dimension de l'ACP. Cette écriture montre que les coordonnées des individus sur l'axe s de l'ACP sont "diminuées" par $\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s}$. La reconstitution par ACPP accorde donc plus d'importance à la première dimension relativement aux autres et les coordonnées des individus sont plus "diminuées" sur les dernières dimensions.

Remarque 5.3. La présentation de la régression ridge via une décomposition en valeurs singulières de X, $X = UDV'$, [18, p.60] montre que la valeur de Y prédite vaut $\hat{Y}_r = X\hat{\beta}_r = \sum_k u_k \frac{d_k^2}{d_k^2+\kappa} u_k' Y$. Ainsi, comme en régression, la régression ridge calcule les coordonnées de Y dans la base orthonormée engendrée par les vecteurs $u_1, ..., u_K$. Cependant, ces coordonnées sont "diminuées" d'un facteur $\frac{d_j^2}{d_j^2+\kappa}$. La régression sur composantes principales supprime quant à elle les dernières dimensions. Ici, la décomposition en valeurs singulières de Γ est immédiate et permet de retrouver les ratios $\frac{\lambda_s - \sigma^2}{\lambda_s - \sigma^2 + \sigma^2} = \frac{\lambda_s - \sigma^2}{\lambda_s}$.

5.2.3 Gestion des données manquantes

L'algorithme EM de l'ACPP peut être étendu au cas incomplet [3, 6, 27] en calculant à l'étape E la distribution des variables latentes et des données manquantes sachant les données observées et les paramètres de l'itération en cours. Ce nouvel algorithme revient à réaliser l'algorithme d'ACP-GEM avec des régressions régularisées (étapes (9) et (10)) à la place des régressions classiques [42]. La procédure consiste ainsi, après une initialisation aléatoire, à calculer $\hat{Z}$, $\hat{\Gamma}$ et $\hat{\sigma}^2$, reconstruire les données manquantes par la formule de reconstitution $\hat{Z}\hat{\Gamma}'$ et itérer jusqu'à la convergence.

Cet algorithme fournit automatiquement un terme de régularisation pour chaque régression ($\hat{\sigma}^2$ et $I\hat{\sigma}^2\hat{M}^{-1}$) adapté au cas des données manquantes. Quand il y a beaucoup de bruit et/ou beaucoup de données manquantes, la valeur de σ^2 est élevée et les coordonnées des individus "diminuent" et se rapprochent du centre de gravité. Ainsi, l'algorithme est "prudent" et les valeurs imputées se rapprochent d'une imputation par la moyenne de chaque variable. Ce comportement est satisfaisant car imputer par la moyenne est acceptable quand il y a peu d'information dans les données. Quand la valeur de $\hat{\sigma}^2$ est faible, la solution obtenue se rapproche de la solution de l'algorithme non régularisé.

Ainsi, l'algorithme ACP-GEM régularisé avec les coefficients précédents permet de limiter le problème important du surajustement de façon automatique et judicieuse.

Remarque 5.4. Cet algorithme peut aussi s'écrire comme un algorithme d'ACP itérative où l'étape de reconstitution par ACP est remplacée par la reconstitution de l'ACPP :

1. initialisation $\ell = 0$: X^0 est obtenu en remplaçant les valeurs manquantes par une valeur initiale. Le tableau de données est centré par $\hat{m}^0$ et est noté X_c^0

2. itération ℓ :
   (a) décomposition en valeurs singulières de $X_c^{\ell-1}$,
   $X_c^{\ell-1} = \frac{\hat{F}^\ell}{\|\hat{F}^\ell\|} \text{ diag } \left(\sqrt{\lambda_s}\right)_{s=1,\ldots,K} \hat{u}^{\ell\prime}$: S axes, composantes et valeurs propres sont retenus. σ^2 est estimée par la moyenne des dernières valeurs propres (de $S+1$ à K)
   (b) X^ℓ est obtenu en remplaçant les valeurs manquantes de X par les valeurs reconstituées $\hat{X}^\ell = \frac{\hat{F}^\ell}{\|\hat{F}^\ell\|} \text{ diag } \left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)_{s=1,\ldots,S} \hat{u}^{\ell\prime} + \hat{m}^{\ell-1}$
   (c) calcul de $\hat{m}^\ell$ et recentrage des données par $\hat{m}^\ell$
3. les étapes (a), (b) et (c) sont répétées jusqu'à la convergence

L'étape (2.a) de l'algorithme d'ACP itérative (page 9) est conservée. L'étape (2.b) est simplement modifiée en remplaçant la valeur singulière $\sqrt{\lambda_s}$ par $\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)$ ce qui correspond à $\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s} \times \sqrt{\lambda_s}$. Ce nouvel algorithme s'apparente ainsi à un algorithme de décomposition en valeurs singulières "rétrécies" itérative. Il permet de mieux comprendre le comportement de la méthode et de faire un nouveau parallèle avec la régression régularisée. La régularisation permet d'améliorer la prévision et d'inclure plus de paramètres dans le modèle. Dans notre cas, la diminution des valeurs singulières permet ainsi de mieux reconstituer les données manquantes et de conserver plus de dimensions. Ceci explique que l'algorithme avec régularisation est assez robuste au choix du nombre d'axes : une reconstitution avec trop d'axes aura un impact limité sur les valeurs reconstituées et par suite sur les résultats de l'ACP.

6 Simulations

Des simulations systématiques ont été réalisées afin de comparer les résultats des algorithmes NIPALS, ACP-GEM, et ACP-GEM avec régularisation. Ces algorithmes sont également comparés à la méthode qui consiste à imputer les valeurs manquantes par la moyenne de chaque variable. Les données sont générées selon un modèle signal (en deux dimensions) plus bruit : $x_{ik} = \tilde{x}_{ik} + \varepsilon_{ik}$, avec $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$. Les performances des différents algorithmes sont évaluées à partir de différents jeux de données (en faisant varier le nombre d'individus et le nombre de variables), de différentes structure de données (en faisant varier σ^2) et différentes configurations de données manquantes (en faisant varier le pourcentage et le dispositif à l'origine des données manquantes). Pour chaque ensemble de paramètres, 250 simulations sont générées.

Pour évaluer la qualité des algorithmes, nous comparons les résultats de l'ACP sur le jeu de données complet (X) et ceux des différents algorithmes sur le jeu de données incomplet selon deux critères :

- le coefficient RV entre les coordonnées des individus (des deux premières dimensions) sur les données complètes et celles obtenues par les algorithmes. Le coefficient RV [12, 22] est un coefficient de corrélation permettant de comparer les positions relatives des individus d'un espace à l'autre ;
- l'erreur de reconstitution entre les valeurs initiales et les valeurs prévues par les algorithmes $\sum_{i,k}(x_{ik} - \hat{x}_{ik})^2$.

Les graphiques de la figure 3 représentent l'évolution de la médiane des coefficients RV en fonction de la valeur de σ^2 pour 10 %, 20% et 40% de données manquantes sur un jeu de

données comprenant 21 individus et 7 variables normées. Le dispositif à l'origine des données manquantes est généré de façon complètement aléatoire. Pour les algorithmes, ACP-GEM et ACP-GEM régularisée, différentes initialisations aléatoires sont utilisées et le résultat optimal est conservé. Ces algorithmes ont été lancés avec des choix *a priori* de 2 puis 3 dimensions ce qui permet d'évaluer l'impact d'un mauvais choix de la dimensionalité. Les résultats présentés sont caractéristiques de l'ensemble des résultats observés en faisant varier les différents paramètres.

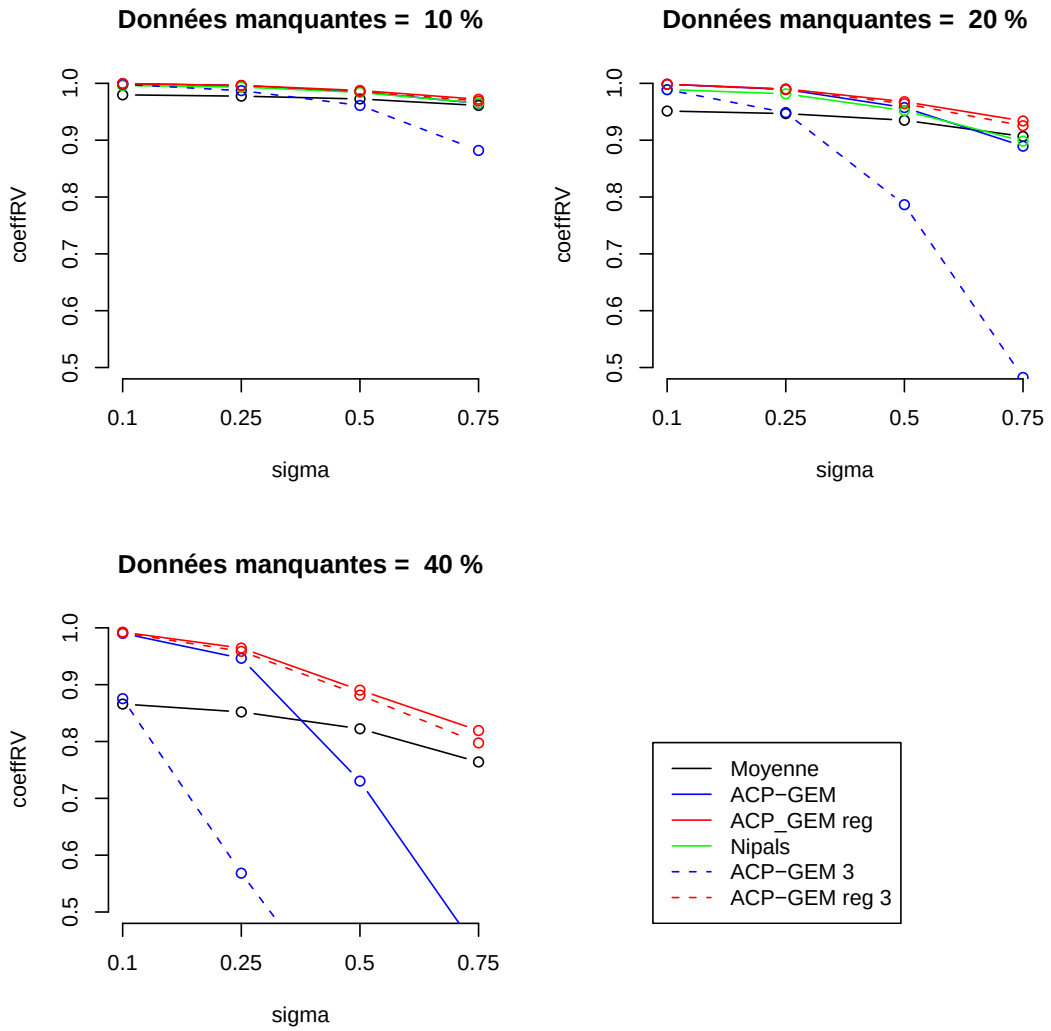

FIGURE 3 – *Evolution du coefficient RV en fonction du bruit et du pourcentage de données manquantes.*

Comme attendu, tous les algorithmes se comportent bien quand peu de données sont manquantes et que la structure est forte (σ^2 petit). Les résultats se dégradent quand le bruit et/ou le nombre de données manquantes augmentent. L'imputation par la moyenne fournit des valeurs de coefficient RV satisfaisantes dans beaucoup de situations ce qui peut s'expliquer par le dispositif à l'origine des données manquantes complètement au hasard. NIPALS fournit des

résultats satisfaisants mais globalement moins bons que les deux autres algorithmes et rencontre des problèmes de convergence quand le nombre de données manquantes est grand (NIPALS ne converge pas toujours pour 40% de données manquantes ou bien vers une solution aberrante). Les performances de l'ACP-GEM et de l'ACP-GEM régularisée sont très satisfaisantes (pour un choix *a priori* de $S = 2$). L'ACP-GEM régularisée se comporte beaucoup mieux quand le nombre de données manquantes et le bruit augmentent. Comme attendu, les résultats de l'algorithme ACP-GEM régularisé se rapproche de ceux obtenus par l'imputation par la moyenne dans les situations extrêmes.

Les performances de l'algorithme ACP-GEM se dégradent très fortement avec $S = 3$ dimensions. Sa version régularisée est par contre très peu sensible à un mauvais choix de la dimensionalité. En effet, elle limite les problèmes de surajustement dus au nombre plus important de paramètres à estimer et à la modélisation du "bruit". L'algorithme est plus stable car les coordonnées des individus correspondant à la troisième dimension sont "diminuées". Cette caractéristique est particulièrement utile en pratique car le choix de la dimension est délicat avec des données manquantes.

La figure 4 montre que les résultats sur l'erreur de reconstitution des données sont très similaires à ceux du coefficient RV. La version régularisée de l'algorithme ACP-GEM fournit toujours la meilleure reconstitution des données manquantes que l'algorithme soit utilisé avec deux ou trois dimensions choisies *a priori*. Quand il y a beaucoup de bruit et 40% de données manquantes les résultats de cette méthode se rapprochent de ceux obtenus par la méthode d'imputation par la moyenne.

7 Conclusion

L'approche retenue pour gérer les données manquantes en ACP consiste à modifier la procédure statistique via l'introduction d'une matrice de poids pour prendre en compte le caractère incomplet des données. Cette approche est conceptuellement simple mais engendre beaucoup de problèmes. La version régularisée (issue de l'ACPP) de l'algorithme ACP-GEM permet de réaliser une ACP sur données incomplètes en évitant le problème important du surajustement. Cet algorithme qui impute itérativement les données au cours du processus d'estimation montre de bonnes performances en terme de reconstitution des données et d'estimation des paramètres pour différentes configurations de données manquantes. De plus, il permet de limiter le problème du choix de la dimension. Il a été implémenté dans FactoMineR [21].

Des algorithmes du type ACP-GEM qui permettent d'approcher une matrice incomplète par une matrice de rang inférieur se retrouvent dans la littérature dans différents domaines comme en "machine learning" [37] ou en "computer vision" et sous différents noms comme "matrix completion", "matrix factorization", etc. Ils sont principalement utilisés pour compléter des matrices de très grandes dimensions avec beaucoup de données manquantes. Par exemple, dans le projet de la compagnie de location de DVD Netflix [13], 100 millions de notes ont été données par 480 189 utilisateurs sur 17770 films. Le tableau de données contient près de 98.8% de données manquantes. L'objectif est alors de prédire les notes des films que les utilisateurs n'ont pas vus afin de les guider dans leur choix. Pour pouvoir utiliser l'algorithme ACP-GEM régularisé sur des données de type Netflix, il faudrait utiliser des techniques d'optimisation

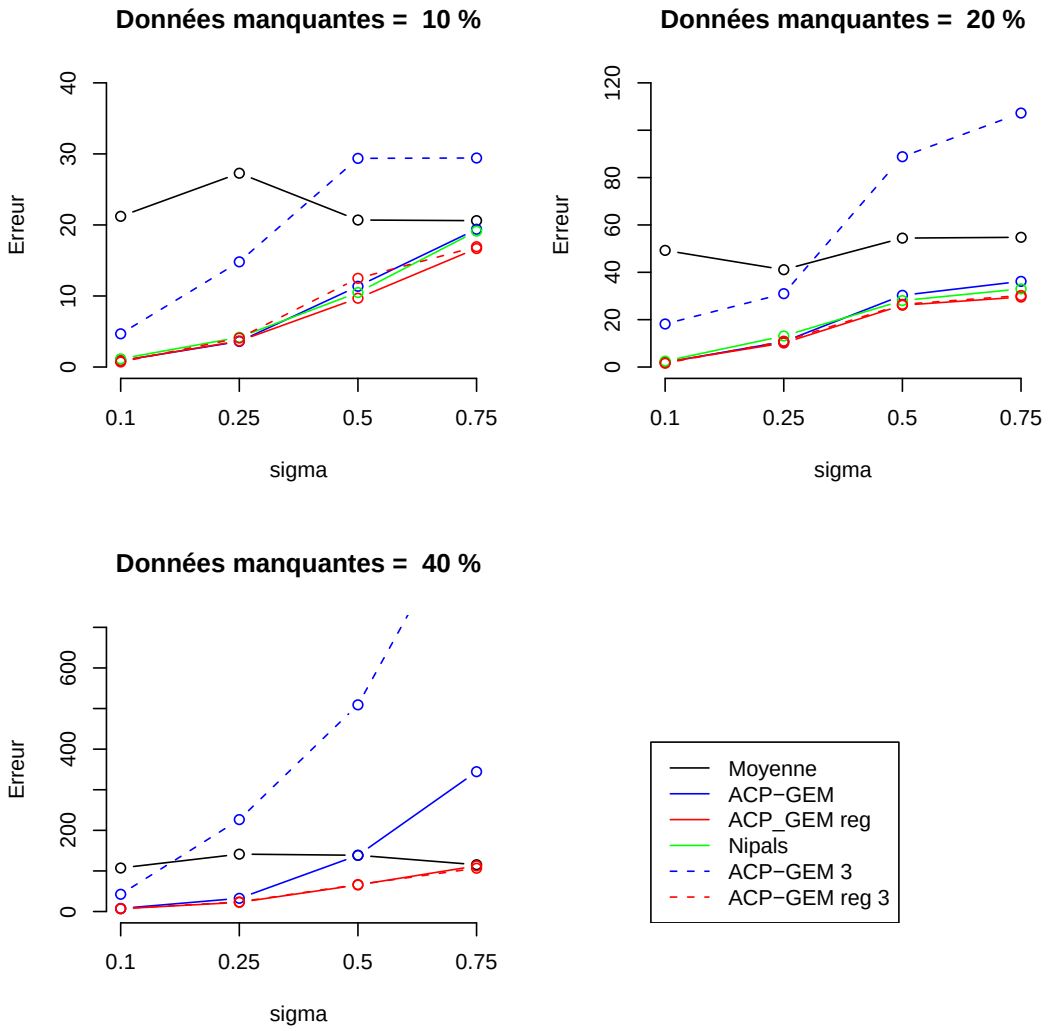

FIGURE 4 – *Evolution de l'erreur de reconstitution en fonction du bruit et du pourcentage de données manquantes.*

avancées pour améliorer le temps de calcul et le stockage des données.

Les techniques présentées dans cet article fournissent une estimation ponctuelle des données et des axes et composantes de l'ACP. Cette estimation ponctuelle est souvent suffisante pour interpréter les proximités entre individus et les liaisons entre variables. Cependant, il peut être utile d'associer une zone de confiance autour de la position de ces points. Des techniques bootstrap ou d'imputation multiple [32, 33], adaptées au cadre de l'ACP, peuvent être envisagées pour affiner les interprétations et visualiser l'incertitude due aux données manquantes.

Références

[1] D. BARTHOLOMEW – *Latent variable models and factor analysis*, Griffin, 1987.

[2] C. M. BISHOP – « Bayesian pca », in *Proceedings of the 1998 conference on Advances in neural information processing systems II* (Cambridge, MA, USA), MIT Press, 1999, p. 382–388.

[3] — , *Pattern recognition and machine learning*, Springer, 2006.

[4] R. BRO – « Multi-way analysis in the food industry. models, algorithms and applications », Thèse, 1998.

[5] A. M. BUCHANAN & A. W. FITZIGIBBON – « Damped newton algorithms for matrix factorization with missing data », *Computer vision and pattern recognition* **2** (2005), p. 316–322.

[6] J. F. CANNY – « Collaborative filtering with privacy via factor analysis », in *Proceding IEEE Symposium on Security and Privacy*, 2002, p. 45–57.

[7] H. CAUSSINUS – « Models and uses of principal component analysis (with discussion) », in *Multidimensional Data Analysis* (J. de Leeuw, W. Heiser, J. Meulman & F. Critchley, éds.), DSWO Press, 1986, p. 149–178.

[8] A. CHRISTOFFERSON – « The one-component model with incomplete data », Thèse, Uppsala University, Institute of statistics, 1969.

[9] A. P. DEMPSTER, N. M. LAIRD & D. B. RUBIN – « Maximum likelihood from incomplete data via the em algorithm », *Journal of the Royal Statistical Society B* **39** (1977), p. 1–38.

[10] J.-B. DENIS – « Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes », *Revue de statistique appliquée* **39** (1991), p. 5–24.

[11] J.-J. DROESBEKE, B. FICHET & P. TASSI – *Modèles pour l'analyse des données multidimensionelles*, Economica, 1992.

[12] Y. ESCOUFIER – « Le traitement des variables vectorielles », *Biometrics* **29** (1973), p. 751–760.

[13] S. FUNK – « Netflix challenge `http://sifter.org/~simon/journal/20061211.html` », 2008.

[14] K. R. GABRIEL & S. ZAMIR – « Lower rank approximation of matrices by least squares with any choice of weights », *Technometrics* **21** (1979), p. 236–246.

[15] G. H. GOLUB & C. F. VAN LOAN – *Matrix computations*, Johns Hopkins University Press, 1996.

[16] M. GREENACRE – *Theory and applications of correspondence analysis*, Acadamic Press, 1984.

[17] B. GRUNG & R. MANNE – « Missing values in principal component analysis », *Chemiometrics and intelligent laboratory systems* **42** (1998), p. 125–139.

[18] T. HASTIE, R. TIBSHIRANI & J. FRIEDMAN – *The elements of statistical learning. data mining, inference and prediction*, Springer series in statistics, 2001.

[19] M. J. R. HEALY & M. WESMACOTT – « Missing values in experiments analyzed on automatic computers », *Applied statistics* **5** (1956), p. 203–206.

[20] W. J. HEISER – « Convergent computation by iterative majorization : theory and applications in multidimensional data analysis », in *Recent Advances in Descriptive Multivariate Analysis* (W. J. Krzanowski, éd.), Oxford University Press, 1995, p. 157–189.

[21] F. HUSSON, J. JOSSE & S. LÊ – « Factominer *R package version 1.11*, url = `http://factominer.free.fr` », 2009.

[22] J. JOSSE, J. PAGÈS & F. HUSSON – « Testing the significance of the rv coefficient », *Computational Statististics and Data Analysis* **53** (2008), p. 82–91.

[23] H. A. L. KIERS – « Weighted least squares fitting using ordinary least squares algorithms », *Psychometrica* **62** (1997), p. 251–266.

[24] P. M. KROONENBERG – *Applied multiway data analysis (chap.7)*, Wiley series in probability and statistics, 2008.

[25] C. L. DE LIGNY, G. H. E. NIEWDORP, W. K. BTRDERODE, W. E. HAMMERS & J. C. VAN HOUWELING – « An application of factor analysis with missing data », *Technometrics* **23** (1981), p. 91–95.

[26] R. J. A. LITTLE & D. B. RUBIN – *Statistical analysis with missing data*, Wiley series in probability and statistics, New-York, 1987, 2002.

[27] B. M. MARLIN – « Missing data problems in machine learning », Thèse, University of Toronto, 2008.

[28] C. NORA-CHOUTEAU – « Une méthode de reconstitution et d'analyse de données incomplètes », Thèse, Université Pierre et Marie Curie, 1974.

[29] K. PEARSON – « On lines and plane of closest fit to systems of points in space », *Phil. Mag.* **2** (1901), p. 559–572.

[30] T. RAIKO, A. ILIN & J. KARHUNEN – « Principal component analysis for sparse high-dimensional data », in *Neural Information Processing*, 2007, p. 566–575.

[31] S. ROWEIS – « Em algorithms for pca and sensible pca », *Advances in Neural Information Processing Systems* **10** (2008), p. 626–632.

[32] D. B. RUBIN – *Multiple imputation for non-response in survey*, Wiley, 1987.

[33] — , « Multiple imputation after 18+ years », *Journal of the American Statistical Association* **91** (1996), p. 473–489.

[34] D. B. RUBIN & D. T. THAYER – « Em algorithms for ml factor analysis », *Psychometrica* **47** (1982), p. 69–76.

[35] J. L. SCHAFER – *Analysis of incomplete multivariate data*, Chapman & Hall/CRC, 1997.

[36] J. L. SCHAFER & J. W. GRAHAM – « Missing data : Our view of the state of the art », *Psychological Methods* **7** (2002), p. 147–177.

[37] N. SREBRO – « Learning with matrix factorizations », Thèse, Massachusetts institute of technology, 2004.

[38] W. STACKLIES, H. REDESTIG, M. SCHOLZ, D. WALTHER & J. SELBIG – « pcamethods-a bioconductor package providing pca methods for incomplete data », *Bioinformatics* **23** (2007), p. 1164–1167.

[39] M. TENENHAUS – *La régression pls théorie et pratique*, Technip, 1998.

[40] M. TIPPING & C. M. BISHOP – « Mixture of probabilistic principal component analysers », *Neural Computation* **11** (1999), p. 443–482.

[41] — , « Probabilistic principal component analysis », *Journal of the Royal Statistical Society B* **61** (1999), p. 611–622.

[42] J. J. VERBEEK – « Notes on probabilistic pca with missing values », Tech. report, 2009.

[43] B. WALCZAK & D. L. MASSART – « Dealing with missing data part i », *Chemiometrics and Intelligent Laboratory System* **58** (2001), p. 15–27.

[44] P. D. WENTZELL, D. T. ANDREWS, D. C. HAMILTON, K. FABER & B. R. KOWALSKI – « Maximum likelihood principal component analysis », *J. Chemom.* **11** (2002), p. 339–366.

[45] H. WOLD – « Estimation of principal components and related methods by iterative least squares », in *Multivariate Analysis* (P. R. Krishnaiah, éd.), Academic Press, 1966, p. 391–420.

[46] — , « Nonlinear estimation by iterative least squares procedures », in *Research Papers in Statistics : Festschrift for Jerzy Neyman* (F. N. David, éd.), Wiley, 1966, p. 411–444.