

# APPORT DE L'ACP PROBABILISTE POUR LA GESTION DES DONNÉES MANQUANTES EN ACP

Julie Josse, François Husson & Jérôme Pagès

*Laboratoire de mathématiques appliquées, Agrocampus Ouest  
65 rue de St-Brieuc, 35042 Rennes Cedex  
julie.josse@agrocampus-ouest.fr*

## Résumé

Dans cette présentation, nous nous intéressons à la gestion des données manquantes en Analyse en Composantes Principales (ACP). Une solution classique pour réaliser une ACP sur données incomplètes consiste à chercher les axes et les composantes qui minimisent l'erreur de reconstitution sur les données présentes. Pour ce faire, différents algorithmes ont été proposés dans la littérature dont 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. Nous détaillons dans un premier temps ces deux algorithmes et donnons leurs propriétés. Nous évoquons ensuite les difficultés rencontrées par ces algorithmes pour nous focaliser sur le problème de surajustement. Puis nous montrons comment la formulation probabiliste de l'ACP (Bishop et Tipping, 1997) offre un terme de régularisation adapté pour pallier au surajustement. Enfin, les performances de ces algorithmes sont évaluées à partir de simulations et d'exemples réels.

*mots clés* : ACP, données manquantes, algorithme EM, moindres carrés pondérés, moindres carrés alternés, ACP probabiliste, surajustement.

## Abstract

In this presentation, we focus on handling missing values in Principal Component Analysis (PCA). An approach commonly used 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 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 quickly review these two approaches and give their properties. Then, we point out the problem of overfitting and we show how the probabilistic formulation of PCA (Bishop and Tipping, 1997) offers a proper and convenient regularization term to overcome this problem. Then we propose an algorithm to perform a PCA on incomplete data set. Finally, performances of the algorithm are

evaluated for both simulations and real data sets.

*key words* : PCA, missing values, EM algorithm, weighted least squares, alternating least squares, Probabilistic PCA, overfitting.

## 1 L'analyse en composantes principales

Soit une matrice  $X$  de dimension  $I \times J$  et  $\|A\| = \sqrt{\text{tr}(AA')}$  la norme euclidienne. L'ACP peut être présentée comme la recherche du sous-espace qui minimise l'erreur de reconstitution, c'est-à-dire la distance entre les individus et leur projection. Minimiser l'erreur de reconstitution revient à chercher une matrice de rang inférieur  $K$  ( $K < J$ ) qui approche au mieux la matrice  $X$  au sens des moindres carrés. Ceci équivaut à chercher deux matrices  $F_{I \times K}$  et  $u_{J \times K}$  qui minimisent le critère suivant :

$$\mathcal{F} = \|X - Fu'\|^2 = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \sum_{k=1}^K F_{ik}u_{jk})^2. \quad (1)$$

En ajoutant les contraintes classiques (orthogonalité des axes et norme unité), la solution unique est fournie par les composantes principales  $F$  et les axes principaux  $u$  de l'ACP. Classiquement, l'ACP est réalisée par décomposition en valeurs singulières de la matrice  $X$  mais il existe d'autres alternatives fondées sur des algorithmes itératifs. Par exemple, l'algorithme dit de "recherche directe du sous-espace" consiste à minimiser le critère (1) en alternant deux étapes de régressions multiples. Pour cela, l'ensemble des paramètres est séparé en deux et  $F$  (resp.  $u$ ) est estimé par la méthode des moindres carrés conditionnellement à  $u$  (resp.  $F$ ) :

$$\begin{cases} \frac{\partial \mathcal{F}}{\partial u} = 0 & \Rightarrow u = X'F(F'F)^{-1}, \\ \frac{\partial \mathcal{F}}{\partial F} = 0 & \Rightarrow F = Xu(u'u)^{-1}. \end{cases}$$

A chaque étape, le critère (1) diminue et il est minimum à convergence. Cet algorithme permet ainsi d'effectuer une ACP à moindre coût, sans calculer explicitement, ni diagonaliser la matrice de variance-covariance.

## 2 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  $w_{ij} = 0$  si  $x_{ij}$  est manquant et  $w_{ij} = 1$  sinon :

$$\mathcal{F}_w = \|W * (X - Fu')\|^2 = \sum_{i=1}^I \sum_{j=1}^J (w_{ij}x_{ij} - \sum_{k=1}^K F_{ik}w_{ij}u_{jk})^2, \quad (2)$$

avec  $*$  le produit d'Hadamard. Contrairement au cas complet, il n'existe pas de solution explicite et il est nécessaire de recourir à des algorithmes itératifs. Gabriel et Zamir (1979) ont ainsi proposé l'algorithme "Criss-Cross Multiple Regression" qui est l'extension de l'algorithme de "recherche directe du sous-espace" et consiste à alterner des étapes de régressions multiples pondérées. Kiers (1997) a montré qu'il était possible de minimiser le critère (2) par la procédure suivante :

1. initialisation  $l = 0$  :  $X^0$  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  $(F^l, u^l)$  qui diminuent ou minimisent le critère  $\|X^{l-1} - Fu^l\|^2$ ;  $K$  dimensions sont retenues
  - (b)  $X^l$  est obtenu en remplaçant les valeurs manquantes de  $X$  par les valeurs reconstituées  $\hat{X}^l = \hat{F}^l \hat{u}^l$ . Le nouveau tableau complété peut ainsi s'écrire  $X^l = W * X + (1 - W) * \hat{X}^l$
3. les étapes (a) et (b) sont répétées jusqu'à convergence

Cette procédure consiste ainsi à effectuer des ACP de façon itérative sur des tableaux de données complétées. L'étape 2.a peut être résolue soit en effectuant l'ACP de  $X^l$  classiquement (le critère (1) est minimisé), soit en effectuant une seule étape des moindres carrés alternés  $\hat{u}^l = X_c^{l-1} F^{l-1} (F^{l-1} F^{l-1})^{-1}$ ,  $\hat{F}^l = X_c^{l-1} u^l (u^l u^l)^{-1}$  (le critère (1) est diminué). En analyse des données, cette procédure d'imputation itérative a été initialement proposée en Analyse Factorielle des Correspondances par Nora-Chouteau (1974). Il est également intéressant de noter que l'ACP itérative correspond à un algorithme EM (Dempster, Laird & Rubin, 1977) du modèle simple :

$$X = Fu' + \varepsilon, \text{ avec } \varepsilon \sim \mathcal{N}(0, \sigma^2 I_J).$$

Si toutes les données étaient observées, la log-vraisemblance complète serait :

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

Soit l'itération  $l$  de l'algorithme d'ACP itérative et  $\theta = (F, u, \sigma^2)$ . L'étape E correspond au calcul de l'espérance de la distribution des données manquantes sachant les observations et les valeurs courantes des paramètres :

$$\mathbb{E}(x_{ij}|X_{obs}, \theta^l) = \begin{cases} x_{ij} & \text{si } x_{ij} \text{ est observé,} \\ \hat{x}_{ij} = F^l u^l & \text{si } x_{ij} \text{ est manquant.} \end{cases}$$

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

$$L_c(F, u, \sigma^2) = -\frac{IJ}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left( \sum_{ij \in \text{obs}} (x_{ij} - F_{ik}u_{jk})^2 + \sum_{ij \in \text{miss}} ((\hat{x}_{ij} - F_{ik}u_{jk})^2 + \sigma^2) \right).$$

Ainsi, la nouvelle estimation des paramètres  $F$  et  $u$  est obtenue en réalisant l'ACP sur le tableau de données complété à l'étape E.

### 3 Propriétés

L'introduction d'une matrice de poids dans le critère (1) rend le problème d'optimisation associé très complexe ce qui engendre plusieurs problèmes.

**Re-centrage.** La continuelle réestimation des données manquantes décentre le jeu de données, il est alors nécessaire d'intégrer une étape de recentrage à l'intérieur de la procédure d'estimation.

**Minima locaux.** Les algorithmes peuvent converger vers des minima locaux car la surface étudiée est très chahutée. Le choix de la position initiale est donc important. Plusieurs solutions ont été envisagées dans la littérature (Srebro, 2004). L'utilisation d'initialisations aléatoires semble être l'approche la plus satisfaisante pour explorer au mieux l'ensemble des solutions possibles.

**Choix du nombre d'axes.** Les solutions des algorithmes qui minimisent le critère (2) ne sont pas emboîtées : la solution à  $K - 1$  dimensions n'est pas incluse dans la solution à  $K$  dimensions. Dès lors, le choix du nombre d'axes apparaît comme important. Ce choix, qui est 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 en obtenir une estimation satisfaisante sans pour autant avoir une valeur de  $K$  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. Différentes stratégies sont envisageables et seront présentées pour choisir la dimension du sous-espace  $K$ .

**Réduction de la variabilité.** A convergence, les données manquantes sont complétées par la formule de reconstitution  $\hat{F}\hat{u}'$ . Ainsi, la variabilité du jeu de données imputé est sous-estimée car le terme d'erreur est absent. De plus, les "vraies" valeurs étant inconnues, il est impossible de savoir comment ces données auraient influencé l'estimation des axes et composantes. Cette instabilité sur les axes et composantes n'étant pas prise en compte, la précision des estimations est surestimée. Ce phénomène de sous-estimation de

la variabilité se retrouve dans toute méthode d'imputation simple. L'imputation multiple (Rubin, 1987) permet de répondre à ces problèmes. Dans le cadre de l'ACP, peu d'auteurs se sont intéressés à l'imputation multiple. On peut citer les travaux de d'Aubigny (2004) et de Kroonenberg (2008).

**Surajustement.** Des problèmes de surajustement peuvent survenir : le critère (2) est faible sur les données d'apprentissage (les données présentes), mais la qualité de prédiction (l'estimation des données manquantes) est très mauvaise. Cette situation peut apparaître dès lors que le nombre de paramètres à estimer est grand par rapport au nombre de données disponibles. Ainsi, dès que le nombre de données manquantes est grand, ou que la dimension  $K$  du sous-espace est élevée, les algorithmes peuvent converger vers des solutions peu plausibles. Pour éviter des solutions surajustées, 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 pénaliser le critère (2) (Raiko, 2008).

## 4 L'ACP probabiliste

L'Analyse en Composantes Principales Probabiliste (ACPP) a été présentée pour la première fois par Tipping et Bishop (1999) et indépendamment par Roweis (1998). Ils ont étendu des travaux précédents sur l'Analyse en Facteurs et ont montré comment l'ACP pouvait être formulée comme une solution du maximum de vraisemblance d'un modèle à variables latentes particulier :

$$x_{J \times 1} = \mu_{J \times 1} + \Gamma_{J \times Q} z_{Q \times 1} + e_{J \times 1},$$

avec  $\mu$  le vecteur de constante,  $\Gamma$  la matrice de "loadings",  $p(e) \sim \mathcal{N}(0, \sigma^2 I_J)$  et  $p(z) \sim \mathcal{N}(0, I_Q)$ , avec  $Q < J$ . La distribution des observations est alors :

$$p(x) \sim \mathcal{N}(\mu, \Sigma) \text{ avec } \Sigma = \Gamma \Gamma' + \sigma^2 I_J.$$

La restriction du modèle d'Analyse en Facteurs sur la variance du bruit implique l'existence d'une solution analytique pour les estimateurs du maximum de vraisemblance :

$$\hat{\Gamma} = U_Q (\Lambda_Q - \sigma^2 I_Q)^{1/2} R \text{ et } \hat{\sigma}^2 = \frac{1}{J - Q} \sum_{i=Q+1}^J \lambda_i,$$

avec  $U_Q$  les  $Q$  premiers vecteurs propres de  $S$  la matrice de variance covariance empirique,  $\Lambda_Q$  la matrice diagonale des valeurs propres associées, et  $R_{Q \times Q}$  une matrice de rotation quelconque. Un algorithme EM (Rubin & Thayer, 1982), où les variables latentes  $z$  sont considérées comme manquantes, peut aussi être utilisé pour maximiser la vraisemblance :

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

$$\text{Etape M : } \hat{\Gamma}' = (Z' Z + I \sigma^2 (\Gamma' \Gamma + \sigma^2 I_Q)^{-1})^{-1} Z' X \text{ et } \hat{\sigma} = \|P_{\hat{z}}^\perp e\|^2.$$

Les étapes EM correspondent aux étapes de moindres carrés alternés de l'ACP mais régularisées. Ces équations peuvent être utilisées dans l'algorithme d'ACP itérative. La régularisation rend le problème bien conditionné et permet d'éviter les problèmes de surajustement. De plus, les termes de régularisation sont bien adaptés au cas des données manquantes. En effet, quand il y a peu d'information, l'algorithme est "prudent" et les individus vont être rapprochés du centre de gravité.

Nous comparons les différents algorithmes entre eux et montrons la supériorité de l'algorithme d'ACP itérative régularisée à partir de simulations et d'exemples réels. Nous comparons également les algorithmes à l'algorithme NIPALS (Wold, 1966) fréquemment utilisé en pratique pour réaliser une ACP sur données incomplètes.

## Bibliographie

- [1] D'Aubigny, G. (2004). Une méthode d'imputation multiple en ACP. Papier présenté aux XXXVI Journées de la Société Française de Statistique. Montpellier, France.
- [2] Dempster, I.A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (B)*, **Vol. 39 (1)** : 1-38.
- [3] Gabriel, K.R. & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, **Vol. 21 (4)** : 236-246.
- [4] Kiers, H.A.L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, **Vol. 62 (2)** : 251-266.
- [5] Kroonenberg, P.M. (2008). Applied Multiway data analysis. Wiley series in probability and statistics. Chapter 7: Missing data in multiway analysis.
- [6] Nora-Chouteau, C. (1974). Une méthode de reconstitution et d'analyse de données incomplètes. Thèse, Université Pierre et Marie Curie.
- [7] Raiko, T., Ilin, A. & Karhunen, J. (2008). Principal Component Analysis for sparse High-Dimensional Data. *Neural Information Processing*. **Vol. 4984**.
- [8] Roweis, S. (1998). EM algorithms for PCA and Sensible PCA. *Advances in Neural Information Processing Systems*. **Vol. 10** : 626-632.
- [9] Rubin, D.B. (1987). Multiple imputation for nonresponse in Survey. Wiley, New York.
- [10] Rubin, D.B. & Thayer, D.T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, **Vol. 47 (1)**.
- [11] Srebro, N. (2004). Learning with Matrix Factorizations. Doctoral Thesis. Massachusetts institute of technology.
- [12] Tipping, M.E & Bishop, C.M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society (B)*, **Vol. 61 (3)** : 611-622.
- [13] Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. Research papers in statistics, Wiley, New York, 411-444.