

VARIABILITÉ DES DIMENSIONS EN ACP : CAS COMPLET ET INCOMPLET

Julie Josse & François Husson

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

Résumé

Dans cette présentation, nous nous intéressons à évaluer la stabilité des dimensions en Analyse en Composantes Principales (ACP). La présentation de l'ACP via un modèle à effets fixes permet de proposer une technique de bootstrap des résidus et d'associer des zones de confiance autour de la position des individus et des variables. Cet algorithme est étendu au cas incomplet et permet de prendre en compte l'incertitude supplémentaire due aux données manquantes. Une méthode d'imputation multiple adaptée au cadre de l'ACP est ensuite proposée pour évaluer la variabilité des dimensions due aux données manquantes.

Mots-clés : ACP, stabilité des dimensions, bootstrap, données manquantes, imputation multiple

Abstract

This paper focus on assessing the stability of the dimensions in Principal Component Analysis (PCA). PCA is presented via a fixed effect model and a technique of residuals bootstrap is detailed to assess the stability of the individuals and variables coordinates. Then, we describe two ways to take into account the supplement uncertainty due to missing values. The technique of residuals bootstrap is extended to the incomplete case. And a multiple imputation method adapted to the framework of PCA is proposed.

Keywords: PCA, Stability of the dimensions, Bootstrap, Missing Values, Multiple Imputation

1 Analyse en Composantes Principales (ACP)

1.1 Estimation ponctuelle des axes et composantes

L'ACP est souvent utilisée comme une méthode exploratoire d'analyse des données c'est à dire comme un outil descriptif multidimensionnel, pour explorer, résumer et visualiser un

jeu de données. Les représentations graphiques des individus et des variables sont au coeur de l'analyse. L'ACP, comme la classification et de nombreuses méthodes exploratoires, est ainsi souvent présentée algorithmiquement ou géométriquement et sans référence à des hypothèses de nature probabiliste. L'ACP est alors définie comme la recherche d'un sous espace qui minimise la distance entre les points et leur projection. Soit X une matrice de taille $I \times K$ (supposée centrée sans perte de généralité) et $\|A\| = \sqrt{\text{tr}(AA')}$ la norme de Frobenius. Plus formellement, l'ACP peut être définie comme la recherche d'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 l'erreur de reconstitution :

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

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.

Depuis quelques années, les méthodes exploratoires sont fréquemment repositionnées dans un cadre probabiliste. Ce cadre permet, entre autres, d'enrichir ces méthodes avec des notions de variance et de tests d'hypothèses (Droesbeke *et. al.*, 1992). L'ACP est alors présentée comme un modèle où les observations sont décomposées en une partie signal z_{ik} plus un bruit ε_{ik} :

$$x_{ik} = z_{ik} + \varepsilon_{ik}. \quad (2)$$

En ACP Probabiliste (ACPP) présentée par Tipping & Bishop (1999), cas particulier de l'Analyse en Facteur, les effets z_{ik} sont considérés comme aléatoires. Ce modèle dit structurel est plus adapté à une situation où les individus étudiés sont un échantillon issu d'une population. Dans le modèle présenté par Caussinus (1986), les effets sont considérés comme fixes. Ce modèle dit fonctionnel est plus adapté au cas où les individus représentent une population entière, c'est à dire qu'ils sont intéressants en tant que tel et ne sont pas considérés comme interchangeables.

Le but de ce travail est d'étudier la stabilité des dimensions en ACP dans le cas d'un tableau de données représentant une population. Nous considérons donc le modèle à effets fixes que nous réécrivons comme un modèle bilinéaire :

$$x_{ik} = \sum_{s=1}^S F_{is}u_{ks} + \varepsilon_{ik}, \text{ avec } \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2). \quad (3)$$

Les estimateurs du maximum de vraisemblance de u et de F correspondent, comme en régression, aux estimateurs des moindres carrés.

1.2 Etude de la stabilité des dimensions

Timmerman *et al.* (2007) ont comparé différentes approches pour évaluer la stabilité des dimensions en ACP. Ils ont montré que les méthodes bootstrap sont plus flexibles et fournissent des résultats plus satisfaisants que les méthodes asymptotiques. Ils ont utilisé un bootstrap non-paramétrique, c'est-à-dire le bootstrap usuel qui consiste à rééchantillonner les individus avec remise. Ce bootstrap est adapté au cas où les individus sont considérés comme *iid* et permet de répondre à la question : quelle est la variabilité des dimensions due aux fluctuations d'échantillonnage? Cette approche est détaillée dans Chateau *et al.* (1996) et permet d'associer des zones de confiance autour de la position des variables. Dans le cadre du modèle à effets fixes, où les individus ont des espérances différentes et l'aléas ne provient que du terme résiduel, le bootstrap des résidus, dit semi-paramétrique, est plus approprié. Il permet d'obtenir des zones de confiance autour de la position des individus et des variables et de répondre à la question : quel est l'impact du bruit sur les paramètres? Même si seules les dernières dimensions sont bootstrappées, tous les paramètres changent car "le bruit va partout". Le bootstrap des résidus repose sur la validité du modèle et dépend fortement du choix de la dimension S . L'algorithme est le suivant :

- réaliser l'ACP de X pour estimer \hat{F} et \hat{u} ;
- reconstituer les données $\hat{X} = \hat{F}\hat{u}'$ et calculer la matrice des résidus $\hat{\varepsilon} = X - \hat{X}$;
- répéter B fois les étapes :
 1. bootstrapper (par cellule) les résidus estimés pour obtenir ε^*
 2. construire un nouveau jeu de données $X^* = \hat{F}\hat{u}' + \varepsilon^*$
 3. réaliser l'ACP de X^* pour estimer F^* et u^*
- représenter les zones de confiance autour des individus (resp. des variables).

Remarque 1 : cette procédure est très similaire à celle utilisée en régression pour obtenir des intervalles de confiance bootstrappés de $\hat{\beta}$ quand la matrice du plan d'expérience est supposée fixe.

Remarque 2 : l'algorithme précédent peut être amélioré en corrigeant la non-homogénéité des variances des résidus estimés. En effet, les résidus estimés peuvent sous-estimer fortement les vraies erreurs. Il faut alors définir des résidus modifiés à la façon des résidus standardisés en régression.

Remarque 3 : comme les erreurs sont supposées distribuées selon une loi normale, l'étape 1 peut être remplacée par un tirage dans une loi normale d'espérance nulle et de variance $\hat{\sigma}^2$. L'estimateur du maximum de vraisemblance de σ^2 est biaisé et peut être corrigé en

utilisant le nombre de degrés de liberté proposé par Denis (1991) :

$$\hat{\sigma}^2 = \frac{\|X - \hat{F}\hat{u}'\|^2}{IK - (IS + KS + K - S - S^2)}.$$

2 ACP sur données incomplètes

2.1 Estimation ponctuelle des axes et composantes

Pour gérer les données manquantes en ACP, une approche fréquemment utilisée consiste à “sauter” les données manquantes en minimisant l’erreur de reconstitution sur les données présentes. Soit W une matrice de poids avec ($w_{ik} = 0$ si x_{ik} est manquant et $w_{ik} = 1$ sinon), le critère devient :

$$\mathcal{C} = \|W * (X - m - Fu')\|^2 = \sum_{i=1}^I \sum_{k=1}^K w_{ik} (x_{ik} - m_k - \sum_{s=1}^S F_{is} u_{ks})^2, \quad (4)$$

avec $*$ le produit de Hadamard . Contrairement au cas complet, il n’existe pas de solution explicite et il est nécessaire de recourir à des algorithmes itératifs. L’algorithme d’ACP itérative (Kiers, 1997) minimise le critère (4) et consiste à :

1. imputer les valeurs manquantes par des valeurs initiales;
2. réaliser l’ACP pour estimer les axes et les composantes;
3. imputer les valeurs manquantes en utilisant la formule de reconstitution à l’ordre S ;
4. répéter les étapes 2 et 3 jusqu’à convergence.

Cet algorithme correspond aussi à un algorithme EM du modèle (3), d’où le nom EM-PCA. Les propriétés de cet algorithme sont étudiées dans Josse *et al.* (2009).

2.2 Incertitude due aux données manquantes

Les techniques disponibles pour réaliser une ACP avec données manquantes fournissent une estimation ponctuelle des paramètres (axes et composantes) mais aussi une estimation des données manquantes. Ces approches s’apparentent donc à des techniques d’imputation simple et ne restituent pas l’incertitude associée à la prédiction des valeurs manquantes. La variance des dimensions obtenues à partir d’une ACP réalisée sur le tableau de données complété est sous-estimée.

Nous présentons deux méthodes pour prendre en compte la variance supplémentaire due aux données manquantes : le bootstrap et l’imputation multiple.

2.2.1 Bootstrap des résidus

Le but est d'étudier la variabilité des dimensions due au bruit et aux données manquantes. L'algorithme proposé dans le cas complet est ainsi étendu au cas incomplet; le bootstrap est utilisé directement (Efron, 1994) :

- réaliser l'algorithme EM-PCA sur la matrice incomplète X pour obtenir \hat{F} et \hat{u} ;
- reconstituer les données $\hat{X} = \hat{F}\hat{u}'$ et calculer les résidus $\hat{\varepsilon} = X - \hat{X}$. La matrice $\hat{\varepsilon}$ est incomplète;
 1. bootstrapper les résidus (par cellule) pour obtenir une nouvelle matrice des résidus ε^* . Les données manquantes sont aussi bootstrappées;
 2. $X^* = \hat{F}\hat{u}' + \varepsilon^*$;
 3. réaliser l'algorithme EM-PCA sur X^* pour obtenir des nouveaux (\hat{F}^*, \hat{u}^*) ;
- représenter les zones de confiance autour des individus et des variables.

2.2.2 Imputation multiple

L'imputation multiple (Little et Rubin, 2002) est une technique qui a été proposée pour obtenir des estimations ponctuelles et des écarts-types des paramètres en prenant en compte le caractère incomplet des données. Il y a trois étapes pour faire de l'imputation multiple. La première consiste à générer D tableaux de données complétés. Chaque valeur manquante est remplacée par D valeurs simulées qui tendent à reproduire l'incertitude associée à la prévision d'une valeur manquante. La deuxième étape consiste à réaliser l'analyse statistique sur chaque tableau de données et estimer le paramètre d'intérêt θ . Enfin, les résultats sont combinés : le paramètre θ est estimé comme la moyenne des $\hat{\theta}$ sur chaque tableau et sa variance totale se décompose en une variance inter-imputation et une variance intra-imputation. La génération de tableaux de données complétés nécessite un modèle d'imputation qui doit être "proper" (Little et Rubin, 2002), c'est-à-dire que l'incertitude sur les paramètres du modèle doit être propagée. L'imputation multiple est souvent réalisée en se positionnant dans un cadre Bayésien mais il est également possible d'utiliser une approche non-paramétrique à l'aide de rééchantillonnages bootstrap.

Nous proposons ici une version de l'imputation multiple adaptée au cadre de l'ACP qui n'étudie que la variance inter-imputation. La première étape consiste à générer des tableaux de données imputés à partir du modèle d'ACP. L'algorithme est le suivant :

- (a) calculer D valeurs pour les paramètres, $(\hat{F}\hat{u}')^1, \dots, (\hat{F}\hat{u}')^D$
- (b) pour $d = 1, \dots, D$, imputer les valeurs manquantes x_{ik}^d en tirant dans la distribution prédictive des données manquantes sachant les données observées et les paramètres : $\sum_{s=1}^S (\hat{F}_{is}\hat{u}'_{ks})^d + \tilde{\varepsilon}$, avec $\tilde{\varepsilon}$ un résidu choisi aléatoirement dans la distribution empirique des résidus (ou dans une loi normale).

Cet algorithme reflète la variance de prévision d'une donnée à partir d'un modèle qui se décompose en une variance d'estimation des paramètres (étape a) et une variance due au bruit (étape b). L'étape (a) utilise les F^* et u^* obtenus par l'algorithme bootstrap présenté précédemment. La deuxième étape de l'imputation multiple consiste à réaliser une ACP sur chacun des D tableaux de données imputés. La troisième étape est ici modifiée par rapport à l'imputation multiple classique car seule la variance inter-imputation est examinée. Elle correspond à la variabilité des dimensions obtenue entre chaque ACP. Seul l'impact des données manquantes est évalué sur les dimensions. Cette méthode permet de répondre à la question : qu'elles auraient été les dimensions avec d'autres prévisions pour les données manquantes?

La méthode bootstrap permet de s'intéresser à la variance totale (due au bruit et aux données manquantes) tandis que la méthode d'imputation multiple adaptée au cadre de l'ACP permet ici de se focaliser sur la variance due uniquement aux données manquantes. Nous présenterons des résultats de simulation pour évaluer les méthodologies proposées et nous présenterons comment visualiser ces différentes sources de variabilité.

Bibliographie

- Caussinus, H. (1986). Models and uses of principal component analysis (with discussion). *Multidimensional Data Analysis*, DSWO Press, 149–178.
- Chateau, F. & Lebart, L. (1996). Assessing sample variability in the visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. *COMPSTAT*, 205–210.
- Denis, J.B. (1991). Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. *Revue de statistique appliquée*, **39**, 5–24.
- Droesbeke, J.-J., Fichet, B. & Tassi, P. (1992). Modèles pour l'analyse des données multidimensionnelles. *Economica*.
- Efron, B. (1994). Missing data, Imputation and the Bootstrap. *Journal of the American Statistical Association*, **426**, 463–475.
- 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.
- Kiers, H A L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, **62**, 251–266.
- Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley, New-York.
- Timmerman, M. E., Kiers, H. A. L. & Smilde, A. K. (2007). Estimating confidence intervals for principal component loadings: a comparaison between the bootstrap and asymptotic results. *British Journal of Mathematica and Statistical Psychology*, **60**, 295–314.
- Tipping, M. & Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society (B)*, **61**, 611–622.