

ACP Parcimonieuse

Stéphane Caron/Sofia Harrouch

2018-03-13

Contents

Introduction	1
Exemple de motivation	2
Description de la méthodologie	2
Justification de la méthodologie	2
Application de la méthodologie	2
Autres éléments pertinents	2
Bibliographie	2

Introduction

Les méthodes statistiques de réduction de la dimensionnalité ont généralement comme objectif de réduire la dimension d'un jeu de données dans le but de simplifier l'interprétation des données, de permettre la visualisation des données ou même de d'améliorer la performance de certaines méthodes appliquées sur ces données réduites. En termes simples, réduire la dimensionnalité revient à réduire le nombre de variables (p) mesurées.

L'analyse en composante principale est une méthode classique de réduction de la dimensionnalité. Cette méthode permet de créer des combinaisons linéaires des différentes variables du jeu de données tout en conservant le plus de variabilité possible. Chacune des nouvelles composantes principales créées possèdent un vecteur de coefficients de saturation (loadings) de dimension $p \times 1$, correspondant en quelque sorte à l'importance attribuée à chacune des différentes variables originales du jeu de données. Il est donc possible d'interpréter ces coefficients de saturation et d'obtenir une interprétation plus généralisée de certaines composantes principales calculées.

Cependant, cette interprétation peut se révéler assez complexe dans le cas où une composante principale est expliquée (coefficients de saturation élevés) par plusieurs variables originales du jeu de données. De plus, il peut être difficile de définir à partir de quelle valeur exactement un coefficient de saturation est considéré comme étant "non important" pour une composante principale. Pour palier à ce problème d'interprétation, il existe différentes méthodes connues. Par exemple, les rotations (I. T. Jolliffe 1989) cherche à simplifier l'interprétation des composantes principales. Il pourrait également être possible d'écarter les coefficients de saturation inférieurs à une certaine valeur ou simplement restreindre les valeurs possibles que ces coefficients peuvent prendre (ex: -1, 0 ou 1). Ces méthodes sont des exemples de stratégie pour faciliter l'interprétation des composantes principales, mais elles ont tous certains désavantages.

La méthodologie introduite dans le présent document (N. T. T. Jolliffe Ian T. and Uddin 2003) est en quelque sorte une alternative à ces méthodes. En bref, elle consiste à ajouter certaines contraintes au modèle d'analyse en composante principale qui auront comme objectif d'améliorer l'interprétabilité des composantes dérivées. Cela permettra notamment d'obtenir des coefficients de saturation exactement égale à zero. La section 2 fera l'illustration du genre de problème qu'on peut éprouver avec l'analyse en composante principale et les rotations en terme d'interprétabilité. La section 3 a comme objectif de décrire la méthodologie. Dans la section 4, nous verrons plus en détails la justification théorique et les résultats de simulation de la méthodologie. La section 5 permettra d'illustrer avec un exemple complet les résultats de la méthodologie. Finalement, la section 6 aura comme but de conclure brièvement en plus de mentionner d'autres éléments à savoir à propos de la méthodologie.

Exemple de motivation

Description de la méthodologie

Justification de la méthodologie

Application de la méthodologie

Autres éléments pertinents

Bibliographie

Jolliffe, Ian T. 1989. “Rotation of Ill-de Ned Principal Components.” *Applied Statistics*, 139–47. http://www.jstor.org/stable/2347688?seq=1#page_scan_tab_contents.

Jolliffe, Nickolay T. Trendafilov, Ian T., and Mudassir Uddin. 2003. “A Modified Principal Component Technique Based on the Lasso.” *Journal of Computational and Graphical Statistics*, 531–47. <https://pdfs.semanticscholar.org/debd/04f4b87a7f7b15bde7efdb2cd57b3603e2cc.pdf>.