ACP Parcimonieuse

Stéphane Caron/Sofia Harrouch

2018-03-13

Contents

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 donnnées 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 crées possèdent un vecteur de coefficients de saturation (loadings) de dimension p x 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 révèler 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 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 permetttant de 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. On pourrait donc dire que cette méthode permet de combiner l'aspect réduction de la dimensionnalité apportée par l'ACP et l'aspect simplification de l'interprétabilité apporté par les exemples décrits plus haut.

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'interprétatibilité. 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

Pour illustrer la motivation derrère la méthodologie, supposons qu'on cherche à simplifier un jeu de données provenant d'un échantillon de 180 coupes de bois de pin afin d'avoir une meilleure compéhension des différentes mesures (variables) impliquées. Les différentes variables du jeu de données en question sont présentées dans le tableau 1. À partir de la matrice de corrélation, il est possible de commencer par faire l'analyse en composante principale et analyser les différentes composantes calculées.

Variables	Description
x1	Diamètre dans le haut de l'arbre (en pouces)
x2	Longeur (en pouces)
x3	Humidité (% poids sec)
x4	Gravité au moment du test
x5	Nombre d'anneaux dans le haut de l'arbre
x6	Nombre d'anneaux dans le haut de l'arbre
x7	Branche principale (en pouces)
x8	Distance du bout de la branche principale au haut de l'arbre
x9	Nombre de spires
x10	Longueur de l'hélice transparente du haut (en pouces)
x11	Nombre moyen de nœuds par verticille
x12	Diamètre moyen des noeuds (en pouces)
x13	Diamètre dans le haut de l'arbre (en pouces)

Comme mentionné dans l'introduction, l'ACP consiste essentiellement à trouver des combinaisons linéaires des variables originales du jeu données (disons la matrice X) tout en maximisant la variance. En termes plus théoriques, la première composante principale est calculée en maximisant la fonction

$$F(\alpha_1) = \alpha_1' \Sigma \alpha_1$$

avec la contrainte que $\alpha_1' \alpha_1 = 1$. La matrice Σ correspond à la matrice de covariance ou à la matrice de corrélation (dépend situation).

Le vecteur α_1 correspond au vecteur de coefficients de saturation de la première composante principale. On refait la même chose pour la deuxième composante principale en ajoutant la contrainte que

$$cov(\alpha_1' X, \alpha_2' X) = 0$$

Il est également possible de démontrer que si on fait la décomposition en valeurs singulières de la matrice de corrélation (ou covariance), on trouve que les vecteurs $\alpha 1, ..., \alpha_p$ correspondent aux vecteurs propres de la matrice Σ alors que la variance de chacune des composantes principales correspond aux p valeurs propres de la même matrice Σ.

	PC1	PC2	PC3	PC4	PC5	PC6
x1	-0.404	-0.218	0.207	-0.091	0.083	0.120
x2	-0.406	-0.186	0.235	-0.103	0.113	0.163
x3	-0.124	-0.541	-0.141	0.078	-0.350	-0.276
x4	-0.173	-0.456	-0.352	0.055	-0.356	-0.054
x5	-0.057	0.170	-0.481	0.049	-0.176	0.626
x6	-0.284	0.014	-0.475	-0.063	0.316	0.052

	PC1	PC2	PC3	PC4	PC5	PC6
x7	-0.400	0.190	-0.253	-0.065	0.215	0.003
x8	-0.294	0.189	0.243	0.286	-0.185	-0.055
x9	-0.357	-0.017	0.208	0.097	0.106	0.034
x10	-0.379	0.248	0.119	-0.205	-0.156	-0.173
x11	0.011	-0.205	0.070	0.804	0.343	0.175
x12	0.115	-0.343	-0.092	-0.301	0.600	-0.170
x13	0.113	-0.309	0.326	-0.303	-0.080	0.626
Simplicity	0.000	0.000	0.000	0.000	0.000	0.000
Variance (%)	0.325	0.183	0.144	0.085	0.070	0.063
Cumulative variance (%)	0.325	0.507	0.652	0.737	0.807	0.870

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.