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 créées 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 de 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 calculées. 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.

Table 1: Présentation des différentes variables du jeu de données pitprops.

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 normés de la matrice Σ alors que la variance de chacune des composantes principales correspond aux p valeurs propres de la même matrice Σ. Ainsi, après avoir fait la décomposition en valeurs et vecteurs propres de la matrice de corrélation, il est possible d'analyser essentiellement deux choses:

1. L'interprétabilité de chacune des composantes principales
2. L'information conservée à chacune des composantes principales

La première peut être analysée en tentant d'interpréter les coefficients de saturation (vecteurs propres). Plus il y a de coefficients similaires, plus la composante est difficile à interpréter. À l'extrême, le cas le plus

simple serait le cas où seulement un seul coefficient ne serait pas égale à 0. Dans ce cas-ci, la composante serait effectivement facile à interpréter, mais il y aurait probablement beaucoup de perte d'information (peu de variance exepliquée), ce qui n'est pas nécésssairement désiré. L'information conservée à chacune des composantes peut quant à elle être quantifiée avec la variance expliquée par la composante principale.

Le tableau 2 montre les résultats des 6 premières composantes principales calculés par l'ACP. On garde seulement les 6 premières composantes étant donné qu'elles expliquent plus de 87% de la variabilité totale du jeu de données.

Table 2: Tableau 2: Coefficients de saturation de l'analyse en composante principale effectuée sur la matrice de corrélation du jeu de données pitprops.

	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
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
Variance (%)	32.451	18.293	14.448	8.534	7.000	6.272
Cumulative variance (%)	32.451	50.744	65.192	73.726	80.726	86.999

Dans le tableau 2, on remarque que les premières composantes principales ont beaucoup de coefficients qui se ressemblent, ce qui rend difficile l'interprétation de celles-ci. Pour palier à ce problème, nous pouvons effectuer une rotation de ces composantes principales. Une rotation classique dans ce genre de situation serait la rotation varimax. Cette rotation est de type orthogonale, c'est donc dire que le système de coordonnées actuel ne subit seulement qu'une rotation. La rotation est faite dans le but de rapprocher le plus possible les coefficients de saturation vers 0 ou 1. Le tableau 3 montre les résultats obtenus après avoir effectué la rotation varimax.

En analysant de plus près le tableau 3, on remarque que la rotation effectuée à permit d'améliorer légèrement l'interprétabilité des premières composantes. Désormais, on remarque que les variables x1 et x2 se démarquent davantage des autres dans la première composante, même chose pour x3 et x4 dans la deuxième composante. Bien que la rotation permet d'améliorer l'interprétabilité, on remarque qu'on perd de la variabilité dans les premières composantes après la rotation. Dans l'ACP classique, les 3 premières composantes expliquaient environ 65% de la variabilité alors que dans le cas de l'ACP avec rotation, les 3 mêmes composantes expliquent environ 56% de la variabilité. De plus, on remarque que ce ne sont plus nécessairement les mêmes composantes qui expliquent successivement le maximum de variabilité. Par exemple, la 4ème composante principale dans l'ACP classique est celle qui explique le moins de variabilité après la rotation. Ces constats sont en quelque sorte les inconvéniants pouvant être rattachés aux rotations et sont également la motivation derrière l'ACP parcimonieuse.

Table 3: Coefficients de saturation de l'analyse en composante principale effectuée sur la matrice de corrélation du jeu de données pitprops après avoir effectué une rotation orthogonale 'varimax'.

	PC1	PC2	PC3	PC6	PC5	PC4
x1	0.912	0.255	0.003	-0.106	0.031	0.011
x2	0.935	0.183	0.013	-0.127	0.023	0.009
x3	0.134	0.961	-0.134	-0.041	0.108	0.077
x4	0.125	0.944	0.246	0.020	0.081	0.031
x5	-0.149	0.024	0.897	0.015	-0.199	-0.026
x6	0.353	0.177	0.638	0.472	0.276	-0.040
x7	0.625	-0.025	0.496	0.517	-0.015	-0.150
x8	0.558	-0.086	-0.090	0.206	-0.553	0.089
x9	0.773	0.028	-0.017	0.094	-0.145	0.103
x10	0.687	-0.090	0.035	0.312	-0.342	-0.420
x11	0.051	0.076	-0.045	0.006	-0.001	0.974
x12	-0.066	0.121	-0.134	-0.069	0.872	0.046
x13	0.075	0.034	-0.082	-0.930	0.169	-0.012
Variance (%)	0.280	0.150	0.120	0.120	0.110	0.090
Cumulative variance (%)	0.280	0.430	0.560	0.670	0.780	0.870

Description de la méthodologie

Comme mentionn? pr?c?demment, L'ACP souffre parfois du probl?me d'interpr?tation des axes. Pour cela on a recours ? une nouvelle m?thode **ACP parcimonieuse (SPARSE)** qui nous donne des axes "Sparses" expliqu?s par un petit nombre des variables; une m?thode qui ignore l'effet de certaines variables sur les axes principales. Dans cette section, on expliquera les diff?rentes m?thodes propos?es pour l'estimation de ces axes. La premi?re m?thode est bas?e sur la propri?t? d'obtention d'une variance maximale des composantes principale(SCoTLASS), la deuxi?me m?thode est construit en se basant sur la propri?t? de l'erreur de recontruction/ de regression(SPCA) et la derni?re m?thode est celle obtenue en utilisant la d?composition PMD (PCA).

L'approche du Lasso en ACP: SCoTLASS

Une technique propos?e par I. T. JOLLIFFE (2003), empruntant l'id?e de Tibshiran (1996) du lasso "the least absolute shrinkage and selection operation" qu'on applique d'habitude dans la r?gresion multiple quand le nombre d'?quation est ?lev?e et le probl?me de l'interpr?tation se pose. Cette approche est nomm? "Simplifed Component Technique LASSO"" permet d'introduire une borne sur la somme des valeurs absolues des coefficients, ces derniers deviennent nul s'ils sont inf?rieurs ? cette borne.

Soit X le jeu de donn?es $X = (X_1, .., X_p)^T$, sa matrice de corr?lation $R = corr(X)$ Soit une ACP sur la matrice de corr?lation qui donne les composantes principales qui sont des combinaisions lin?aires des p variables mesur?es de soit $Y_k = a'_k X = \sum_{i=1}^{p} a_{ki} X_i$, $(k = 1, ..., p)$. On note ensuite la variance de l'axe principale Y_k par $var(Y_k) = a'_k * R * a_k$. le probl?me de maximisation de l'ACP pour conserver la plus grande quantit? d'information possible est:

$$\max a'_k * R * a_k$$

$$s/c \begin{cases} aa'_k \times a_k = 1, & (k \geq 2) \\ a'_h \times a_k = 0, & (h \geq k) \end{cases}$$

La m?thode du lasso appliqu? ? l'asso'ACP, rajoute une troisi?me contrainte sur les coefficients des variables sur les axes.

$$\sum_{j=1}^{p} |a_{kj}| \leq t$$

Dans cette m?thode, il n y a pas de r?gle ni d'orientation pour le choix de t, ce t qu' ? partir duquel on ignore la significativit? des variables initiales sur les axes. ? titre indicatif, on a: 1. pour t $\geq \sqrt{p}$, on a l'ACP. 2. pour t ≤ 1, il n'existe pas de solution. 3. pour t=1, on a exactement une valeur non nulle de a_{kj} pour chaque k. Donc on choisit des valeurs diff?rentes de t ce qui a comme cons?quence un co?t ?lev? de calcul, une probl?me d'optimisation non convexe. Sans oublier aussi le fait que cette m?thode ne nous fournisse pas vraiment des vecteurs propres assez sparses quand le besoin d'un grand pourcentage de la variance expliqu?e est exprim?, on se trouve donc dans une situation de compromis entre le pourcentage de variance expliqu?e et l'interpr?tation des variables.

SPCA: Sparse Principal Component Analysis

Dans cette partie, on introduit une autre approche estimant les composantes principales tout en rendant les vecteurs parcimonieux (sparse). L'ACP peut ?tre r?ecrite comme un probl?me d'optimisation d'une r?gression en imposant une p?nalit? quadratique: la p?nalit? du Lasso via l'Elastic net. En fait, chaque composante principale est ?crite comme une combinaison lin?aire des p variables, donc les coefficients des variables sur les CPs peuvent ?tre obtenus en regressant la CPs sur ces p variables. Apr?s l'application de l'ACP sur notre jeu de donn?e, on reconstruit les facteurs (loadings) par une regression ridge, cette m?thode d?pend des r?sultats obtenus de l'ACP(post ACP).

On note Y_i la composante principale, λ positive, l'estimateur de ridge est:

$$\hat{\beta}_{ridge} = argmin_{\beta} ||Y_i - X\beta||^2 + \lambda ||\beta||^2$$

Soit $\hat{v} = \frac{\hat{\beta}_{ridge}}{||\hat{\beta}_{ridge}||}$, donc $\hat{v} = V_i$ Cette p?nalit? de ridge n'en n'est pas vraiment une, elle sert simplement ? reconstruire les composantes.

Ensuite, la m?thode rajoute une nouvelle p?nalit? L_1;, la p?nalit? de Lasso, ? l'equation pr?cedante, ce qui donne un nouveau probl?me d'optimisation:

$$\hat{\beta} = argmin_{\beta} ||Y_i - X\beta||^2 + \lambda ||\beta||^2 + \lambda_1 ||\beta||_1$$

telle que: $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ est la norme 1 de β. On appelle $\hat{V}_i = \frac{\hat{\beta}}{||\hat{\beta}||}$ l'approximation de V_i, et la $X\hat{V}_i$ la i?me composante principale approxim?e.

Dans notre cas, on consid?re les k premiers CPs. Soit $A_{p*k} = [\alpha_1, \alpha_2, ..., \alpha_k]$ et $B_{p*k} = [\beta_1, \beta_2, ..., \beta_k]$

Le probl?me d'optimisation g?narlis? obtenu est:

$$(\hat{A}, \hat{B}) = argmin \sum_{i=1}^{n} ||X_i - AB^T X_i||^2 + \sum_{j=1}^{k} |\beta_j||^2 + \sum_{j=1}^{k} \lambda_{1,j} ||\beta_j||_1$$

sous la contrainte

$$A^T A = I_{k*k}$$

Justification de la méthodologie

Justification de la méthodologie

DAns cette partie on va comparer la performance des deux m?thodes ScotLASS et SPCA. On consid?re trois facteurs:

$$V_1 \sim N(0, 290) V_2 \sim N(0, 300) V_3 = -0.3 * V_1 + 0.925 * V_2 + \epsilon, \epsilon \sim N(0, 1)$$

tel que $V_{1}, V_{2}, $et $\epsilon \sim N(0, 1)$ On construit 10 variable tel que:

$X_i = V_1 + \epsilon_i^1, \epsilon_i^1 \sim N(0, 1)$pour i=1, 2, 3, $4 X_i = V_2 + \epsilon_i^2, \epsilon_i^2 \sim N(0, 1)$pour i=5, 6, 7, $8 X_i = V_3 + \epsilon_i^3, \epsilon_i^3 \sim N(0, 1)$pour i=1,2 ,3, 4

On remarque que les deux premiers axes de l'ACP explique plus de 80% de la variance totale, et que

```
##              [,1]        [,2]        [,3]        [,4]        [,5]
##  [1,]  1.00000000  1.00000000  1.00000000  1.00000000 -0.04196468
##  [2,]  1.00000000  1.00000000  1.00000000  1.00000000 -0.04196468
##  [3,]  1.00000000  1.00000000  1.00000000  1.00000000 -0.04196468
##  [4,]  1.00000000  1.00000000  1.00000000  1.00000000 -0.04196468
##  [5,] -0.04196468 -0.04196468 -0.04196468 -0.04196468  1.00000000
##  [6,] -0.04196468 -0.04196468 -0.04196468 -0.04196468  1.00000000
##  [7,] -0.04196468 -0.04196468 -0.04196468 -0.04196468  1.00000000
##  [8,] -0.04196468 -0.04196468 -0.04196468 -0.04196468  1.00000000
##  [9,] -0.33108128 -0.33108128 -0.33108128 -0.33108128  0.95478830
## [10,] -0.33108128 -0.33108128 -0.33108128 -0.33108128  0.95478830
##              [,6]        [,7]        [,8]       [,9]       [,10]
##  [1,] -0.04196468 -0.04196468 -0.04196468 -0.3310813 -0.3310813
##  [2,] -0.04196468 -0.04196468 -0.04196468 -0.3310813 -0.3310813
##  [3,] -0.04196468 -0.04196468 -0.04196468 -0.3310813 -0.3310813
##  [4,] -0.04196468 -0.04196468 -0.04196468 -0.3310813 -0.3310813
##  [5,]  1.00000000  1.00000000  1.00000000  0.9547883  0.9547883
##  [6,]  1.00000000  1.00000000  1.00000000  0.9547883  0.9547883
##  [7,]  1.00000000  1.00000000  1.00000000  0.9547883  0.9547883
##  [8,]  1.00000000  1.00000000  1.00000000  0.9547883  0.9547883
##  [9,]  0.95478830  0.95478830  0.95478830  1.0000000  1.0000000
## [10,]  0.95478830  0.95478830  0.95478830  1.0000000  1.0000000
```

```
##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
##  [1,]  0.140 -0.472 -0.086  0.000  0.000  0.000  0.000  0.277  0.000
##  [2,]  0.140 -0.472 -0.086  0.064  0.129  0.221  0.221  0.225 -0.660
##  [3,]  0.140 -0.472 -0.086 -0.324  0.224  0.178 -0.010  0.082  0.668
##  [4,]  0.140 -0.472 -0.086  0.260 -0.353 -0.399 -0.212 -0.585 -0.008
##  [5,] -0.388 -0.164  0.269  0.121 -0.587  0.617 -0.041  0.000  0.090
##  [6,] -0.388 -0.164  0.269 -0.548 -0.254 -0.506  0.309  0.155 -0.086
##  [7,] -0.388 -0.164  0.269  0.266  0.296 -0.215 -0.645  0.341 -0.022
##  [8,] -0.388 -0.164  0.269  0.161  0.545  0.104  0.377 -0.496  0.018
##  [9,] -0.399 -0.012 -0.583 -0.454  0.086  0.181 -0.355 -0.262 -0.226
## [10,] -0.399 -0.012 -0.583  0.454 -0.086 -0.181  0.355  0.262  0.226
##       Comp.10
##  [1,]  0.821
##  [2,] -0.381
##  [3,] -0.332
##  [4,] -0.107
```

```
##  [5,]  0.000
##  [6,] -0.052
##  [7,] -0.115
##  [8,]  0.168
##  [9,]  0.088
## [10,] -0.088
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
## Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8
##                Comp.9 Comp.10
## SS loadings       1.0     1.0
## Proportion Var    0.1     0.1
## Cumulative Var    0.9     1.0
```

```
## Importance of components:
##                             Comp.1     Comp.2       Comp.3
## Standard deviation     42.2380068 33.0759439 1.1728922251
## Proportion of Variance  0.6195814  0.3799409 0.0004777577
## Cumulative Proportion   0.6195814  0.9995222 1.0000000000
##                                          Comp.4                      Comp.5
## Standard deviation     0.000000187925629418325555 0.000000180631368654366500
## Proportion of Variance 0.000000000000000012264890 0.000000000000000011331250
## Cumulative Proportion  1.000000000000000000000000 1.000000000000000000000000
##                                          Comp.6
## Standard deviation     0.00000015618988205565691
## Proportion of Variance 0.00000000000000008472222
## Cumulative Proportion  1.00000000000000000000000
##                                          Comp.7 Comp.8 Comp.9 Comp.10
## Standard deviation     0.000000146713219733587335      0      0       0
## Proportion of Variance 0.000000000000000007475324      0      0       0
## Cumulative Proportion  1.000000000000000000000000      1      1       1
```

```
## You may wish to restart and use a more efficient way
## let the argument x be the sample covariance/correlation matrix and set type=Gram
```

```
##
## Call:
## spca(x = x, K = 2, para = c(4, 4), type = "predictor", sparse = "varnum")
##
## 2 sparse PCs
## Pct. of exp. var. : 41.4 38.7
## Num. of non-zero loadings :  4 4
## Sparse loadings
##       PC1 PC2
##  [1,] 0.0 0.5
##  [2,] 0.0 0.5
##  [3,] 0.0 0.5
##  [4,] 0.0 0.5
##  [5,] 0.5 0.0
##  [6,] 0.5 0.0
##  [7,] 0.5 0.0
##  [8,] 0.5 0.0
##  [9,] 0.0 0.0
```

```
## [10,] 0.0 0.0
```

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.