ACP Parcimonieuse

Stéphane Caron/Sofia Harrouch

2018-04-04

Contents

1. 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 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 composantes principales 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 possède 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 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) cherchent 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 permettant de faciliter l'interprétation des composantes principales, mais elles ont toutes certains désavantages.

La méthodologie introduite dans le présent document est en quelque sorte une alternative à ces méthodes. En bref, elle consiste à ajouter certaines contraintes au modèle d'analyse en composantes principales qui auront comme objectif d'améliorer l'interprétabilité des composantes calculées. Cela permettra notamment d'obtenir des coefficients de saturation exactement égales à zéro. On pourrait donc dire que cette méthode permet de combiner l'aspect de réduction de la dimensionnalité apportée par l'ACP et l'aspect de 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.

2. Exemple de motivation

Pour illustrer la motivation derrière 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 compréhension 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.

```
library(elasticnet)

# Ce jeu de données correspond à la matrice de corrélation
data(pitprops)
```

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 bas 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 de la 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 expliquée), ce qui n'est pas nécessairement 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 calculées 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.

```
library(stats)

# Faire l'ACP classique sur la matrice de corrélation
acp <- prcomp(x = pitprops, center = FALSE)
```

Table 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.

```
library(psych)

# Faire l'ACP, mais en faisant une rotation varimax des CP
acp_rotated <- principal(pitprops, 6, rotate = "varimax", eps = 1e-14)
```

En analysant de plus près le tableau 3, on remarque que la rotation effectuée a permis 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

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

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 inconvénients pouvant être rattachés aux rotations et sont également la motivation derrière l'ACP parcimonieuse.

3. Description de la méthodologie

Comme mentionné précédemment, l'ACP souffre parfois du problème d'interprétation des axes, ou des vecteurs de coefficients de saturation. Pour palier à ce problème, il est possible d'avoir recours à une méthode spécifique, soit l'ACP parcimonieuse. Cette méthode permet de trouver des axes "éparses" qui sont expliqués par un petit nombre des variables seulement. On pourrait donc dire que la méthode fait en quelque sorte une régularisation et ignore l'effet de certaines variables sur les axes principales. Une connaissance sur les méthodes de régularisation peut donc être fort utile pour bien comprendre la mécanique derrière les méthodes d'ACP parcimonieuse. Ainsi, nous avons ajouté en annexe une section faisait une brève introduction sur les fondements derrière les méthodes de régularisation.

Dans cette section, on expliquera les différentes méthodes proposées pour l'estimation de vecteurs de coefficients de saturation. La première méthode est basée sur la propriété d'obtention d'une variance maximale des composantes principales (SCoTLASS) alors que la la deuxième méthode est construite en se basant sur la propriété de l'erreur de reconstruction (SPCA).

Méthode 1: SCoTLASS

Cette technique (N. T. T. Jolliffe Ian T. and Uddin 2003), emprunte l'idée du LASSO (Least Absolute Shrinkage and Selection Operation) introduite par Tibshirani (1996). Le LASSO est généralement appliqué en régression multiple quand le nombre de variables est élevé en faisant une régularisation sur celles-ci. Dans le problème de l'interprétabilité de l'ACP, cette méthode peut se révéler fort utile pour garder uniquement les variables importantes et ainsi faciliter l'interprétation. La méthode SCoTLASS (Simplifed Component

Technique LASSO) permet d'introduire une borne sur la somme des valeurs absolues des coefficients, ces derniers devenant nul s'ils sont inférieurs à cette borne.

Soit X le jeu de données $X = (X_1, .., X_p)^T$ et $R = corr(X)$ sa matrice de corrélation. En faisant l'analyse en composante principale sur la matrice de corrélation, on obtient les composantes qui sont des combinaisons linéaires des p variables mesurées, soit:

$$Y_k = \alpha'_k X = \sum_{i=1}^{p} \alpha_{ki} X_i$$

pour $(k = 1, ..., p)$. On note ensuite la variance de l'axe principale Y_k par $var(Y_k) = \alpha'_k * R * \alpha_k$.

Le problème de maximisation de l'ACP pour conserver la plus grande quantité d'information possible est donc donné par le paramètre α qui maximise cette fonction:

$$F(\alpha_1) = \alpha'_1 R \alpha_1$$

avec les contraintes suivantes:

$$\begin{aligned} \alpha'_k \alpha_k &= 1 \qquad \forall k \\ \text{cov}(\alpha'_1 X, \alpha'_2 X) &= 0 \qquad \forall k \geq 2 \text{ et } h \neq k \end{aligned}$$

La méthode du LASSO appliquée à l'ACP (SCoTLASS) rajoute une troisième contrainte sur les coefficients des variables sur les différents axes des composantes:

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

Dans cette méthode, il faut donc définir un hyperparamètre (t) qui permettra de controler la régularisation effectuée sur les coefficients de saturation. Il n'y a pas d'orientation particulière pour le choix de t, mais on remarque que le choix de celui-ci peut être décortiqué de cette manière:

1. pour $t \geq \sqrt{p}$, on a l'ACP.
2. pour $t \leq 1$, il n'existe pas de solution.
3. pour $t = 1$, on a exactement une valeur non nulle de a_{kj} pour chaque k.

Le point (2) est expliqué par la contrainte que le vecteur de coefficients de saturation doit être normé. Si on veut effectuer une régularisation sur les coefficients, il faut donc choisir une valeur de t entre 1 et $\sqrt{p}$. Une stratégie possible est d'essayer plusieurs valeurs différentes, mais cela a comme conséquence un coût élevé par rapport au temps de calcul.

Dans le choix du paramètre t, il faut aussi considérer le fait que plus la régularisation est importante, plus les composantes seront facilement interprétables, mais cela au coût de perdre de la variabilité. Il faut donc trouver un équilibre entre la simplicité des composantes et la quantité d'information conservée. Ce compromis dépend effectivement du jeu de données et également du contexte.

En bref, la contrainte ajoutée par la méthode SCoTLASS permet de régulariser les valeurs des coefficients ce qui permet d'obtenir des coefficients de saturation exactement égals à 0, ce qui facilite donc l'interprétation des composantes.

SPCA: Sparse Principal Component Analysis

Dans cette partie, on montre une deuxième approche pour l'obtention d'une ACP modifiée et *sparse*. Pour cette méthode, on peut faire l'analogie entre l'ACP et la regression de Ridge. Tout d'abord, on montrera que les composantes principales d'une ACP peuvent être écrites comme un problème d'optimisation d'une regression de Ridge. Finalement, on ajoute la pénalité du Lasso afin d'obtenir un problème d'optimisation sous la forme d'une regression sous la pénalité d'Elastic Net ce qui va nous permettre d'obtenir des composantes principales parcimonieuses.

Approche simple de regression pour l'ACP

On considère $Y_k = \alpha_k' X$ la kième composante principale obtenu à partir d'une ACP, qui est une combinaison linéaire des p variables intiales. En fait, ces coefficients de saturation α_k peuvent aussi être obtenu en faisant une regression multiple de la composante principale sur les p variables initiales ($Y_k = X\beta + \epsilon$). Ensuite, on peut étendre cette regression à une regression de Ridge en ajoutant la pénalité L_2 dans le but de manipuler toute sorte de données et obtenir une solution unique comme l'ACP.

Pour résumer, cette pénalité nous permet juste de reconstruire les composantes principales et non pas de les "pénaliser". Enfin, on rajoute la pénalité du Lasso (L_1) pour pénaliser les composantes et de les rendre parcimonieuses. Le problème d'optimisation devient:

$$\hat{\beta} = argmin_\beta ||Y_k - X\beta|| + \lambda||\beta||^2 + \lambda_1||\beta||_1 \quad (1.1)$$

et la kième composante principale approximée est $X\hat{V}_k = X\frac{\hat{\beta}}{||\hat{\beta}||}$.

Cette technique dépend essentiellement du résultat de l'ACP, car on applique l'ACP et on utilise l'équation (1.1) pour trouver une autre apprixmation appropriée des composantes principales pour qu'elles soient parcimonieuses. Dans la prochaine section, nous verrons comment il est possible d'utiliser l'ACP uniquement comme point de départ et ensuite d'itérativement améliorer les estimations des coefficients parcimonieux.

Approche complexe de regression pour l'ACP

Soient $\mathbf{X}$ la matrice d'observations, x_i la ième ligne de cette matrice, $A_{p\times k} = [\alpha_1, \alpha_2, ..., \alpha_k]$ les k premières composantes principales et $B_{p\times k} = [\beta_1, \beta_2, ..., \beta_k]$ les k premières composantes principales estimées de tel sorte que les composantes principales soient parcimonieuses.

Le théorème 1 suivant montre que le problème de l'ACP peut se transformer en un problème de regression.

Théorème 1 Pour tout $\lambda > 0$:

$$(\hat{A}, \hat{B}) = argmin_{A,B} \sum_{i=1}^{n} ||x_i - AB^T x_i||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2$$

Sous la contrainte

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

Alors $\hat{\beta}_j \infty V_j$ pour $j = 1, 2, ..., k$.

Ce théorème nous permet de bien voir la transformation du problème d'optimisation.

En fait,

$$\sum_{i=1}^{n} ||x_i - AB^T x_i||^2 = ||X - XAB^T||^2$$

$$= ||XA_\perp||^2 + ||XA - XB^T||^2$$

Comme A est orthonormale, $A_\perp$ est une matrice quelconque orthonormale telle que $[A; A_\perp]$ est $p \times p$ orthonormal.

Pour A fixé, le problème de minimisation devient:

$$(A, \hat{B}) = argmin_B C_\lambda(A, \hat{B}) = argmin_{A,B} \sum_{j=1}^{k} ||X\alpha_j - X\beta_j||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2$$

Ce qui est équivalent à resoudre k regressions de Ridge indépendantes. Cela nous donne $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T X A$. En fait, si A correspond aux composantes principales classiques, alors on sait effectivement que B est proportionnelle à V (ce qu'on a montré dans la sous-section précédente).

Après avoir montré l'analogie entre les deux méthodes, on rajoute la pénalité de Lasso pour avoir la "sparsité" des coefficients de regression (ou de saturation). Le problème devient:

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

Sous la contrainte:

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

Outils pour appliquer la méthode

Voici les différents éléments qui seront nécéssaires pour appliquer la méthode et passer au travers de l'algorithme décrit dans la prochaine section.

1. On suppose que A est connu et on cherche B

$$\hat{\beta}_j = argmin_{\beta_j} ||Y_j - X\beta_j||^2 + \lambda ||\beta_j||^2 + \lambda_{1,j} ||\beta_j||_1$$

$$\hat{\beta}_j = argmin_{\beta_j} (\alpha_j - \beta_j)^T X^T X (\alpha_j - \beta_j) + \lambda ||\beta_j||^2 + \lambda_1, j ||\beta_j||_1$$

2. On suppose que B est connu et on cherche A (on ignore la pénalité qui est liée aux termes de B)

$$min_\beta \sum_{i=1}^{k} ||x_i - AB^T x_i||^2 = ||X - XBA^T||$$

Sous la contrainte:

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

La solution est obtenu en utilisant le théorème 2, soit en calculant la décomposition en valeurs singulières (SVD) $(X^T X)B = UDV^T$ et posant $\hat{A} = UV^T$.

Théorème 2 Soit $M_{n\times p}$ et $N_{n\times k}$ deux matrices. On considère le problème de minimisation suivant

$$\hat{A} = argmin_A ||M - NA^T||$$

Sous la contrainte

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

La décomposition en valeurs singulières de $M^T N$ est UDV^T, donc $\hat{A} = UV^T$.

Algorithme pour appliquer la méthode

1. On commence par $A = [\alpha_1, \alpha_2, ..., \alpha_K]$ les coefficients de saturation des K premières composantes principales.
2. Sachant $A = [\alpha_1, ..., \alpha_k]$, on résout le problème

$$\beta_j = argmin_\beta(\alpha_j - \beta)^T X^T X(\alpha_j - \beta) + \lambda||\beta||^2 + \lambda_1, j||\beta||_1$$

   pour $j = 1, 2, ..., k$. Cela nous permet d'obtenir une estimation de B.
3. Pour une matrice $B = [\beta_1, \beta_2, ..., \beta_k]$ fixée, on calcule la décomposition de la matrice en valeur singulière (SVD) de $X^T XB = UDV^T$. Par la suite on met A à jour, ce qui donne $A = UV^T$.
4. On répète les étapes 2-3 jusqu'à la convergence.
5. On normalise V, $\hat{V}_j = \frac{\beta_j}{||\beta_j||}$

Les V_j sont les nouvelles composantes principales parcimonieuses.

4. Justification de la méthodologie

Pour justifier la méthodologie, il est possible de réaliser des simulations et voir comment performe l'analyse en composante principale parcimonieuse versus d'autres méthodes comme la méthode par rotation. Dans cet exemple, nous allons définir 3 structures de vecteurs et valeurs propres différentes que nous cherchons à retrouver au travers de simulations. En effet, disons que nous avons une matrice et vecteurs propres A et un vecteur de valeurs propres I, il est possible de retrouver la structure de corrélation de cette manière:

$$\Sigma = ADA^{-1}$$

où D est une matrice diagonale construite avec les valeurs propres(I). De cette manière, on refait en quelque sorte le chemin inverse de la décomposition en valeurs et vecteurs propres de la structure de corrélation (ou covariance). Une fois que la structure de corrélation est retrouvée, il est possible de simuler des lois normales multivariées à partir de cette même structure.

Les 3 structures différentes de vecteurs et valeurs propres sont définis dans les tableaux 4, 5 et 6. La première est construite de manière à ce qu'il y aille 3 variables plus importantes et 3 variables moins importantes dans chacune des composantes. La troisième structure est construite de manière à avoir des coefficients de saturation proches les uns des autres pour chacune des composantes. Pour finir, la deuxième structure est une intermédiaire entre les deux structures décrites plus tôt.

La figure 1 présente les matrices de corrélation pour les données simulées à partir de ces 3 différentes structures. Cela n'est pas si évident à voir, mais si on regarde de près la figure 1 on remarque que la structure en blocks contient des blocks de variables très corrélées (carré de couleur très uniforme), alors que la structure uniforme contient des variables avec des corrélations plus uniformément distribuées (carré de couleur moins uniforme).

Table 4: Coefficients de saturation théoriques de la structure en blocks.

variable	PC1	PC2	PC3	PC4	PC5	PC6
x1	0.096	-0.537	0.759	-0.120	0.335	-0.021
x2	0.082	-0.565	-0.599	0.231	0.511	-0.013
x3	0.080	-0.608	-0.119	-0.119	-0.771	0.016
x4	0.594	0.085	-0.074	-0.308	0.069	0.731
x5	0.584	0.096	-0.114	-0.418	0.052	-0.678
x6	0.533	0.074	0.180	0.805	-0.157	-0.069

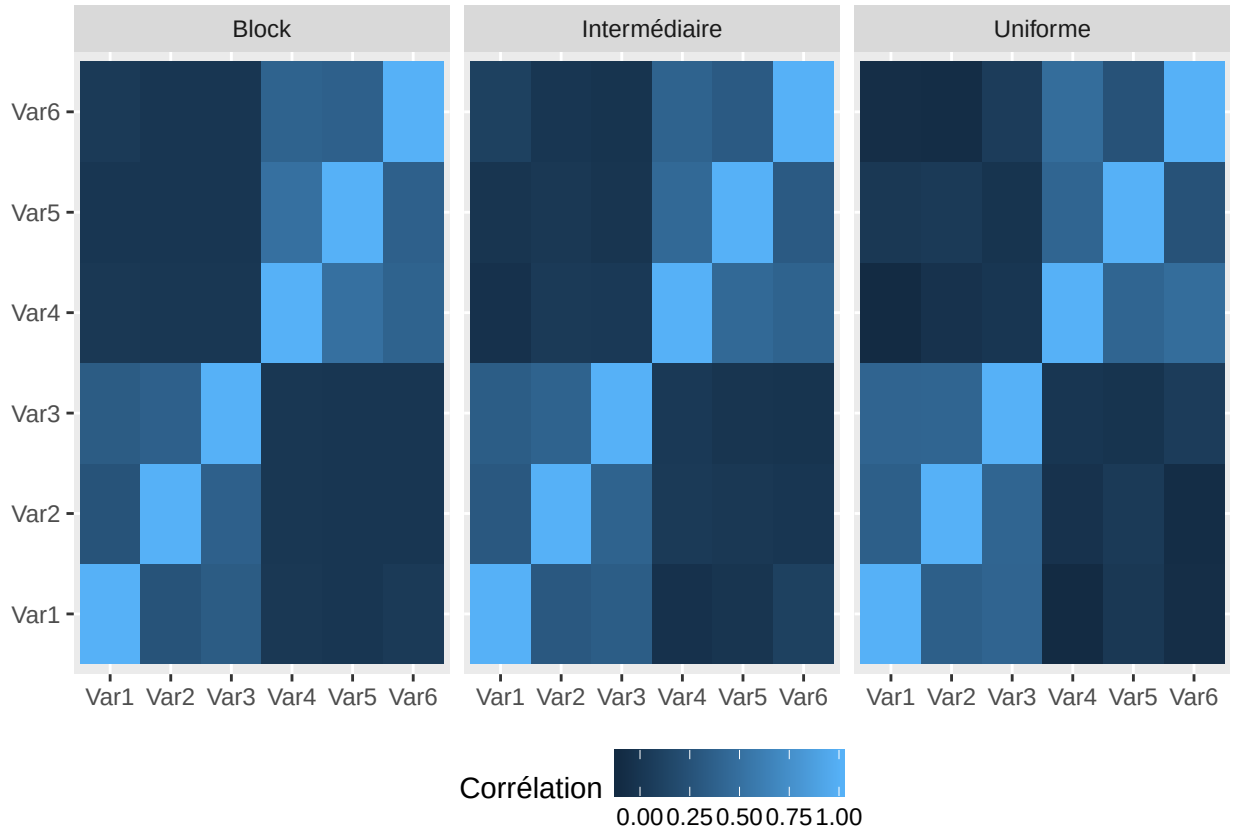

Figure 1: Représentations des différentes matrices de corrélation pour les données simulées à partir des 3 différentes structures.

Table 5: Coefficients de saturation théoriques de la structure intermédiaire.

variable	PC1	PC2	PC3	PC4	PC5	PC6
x1	0.224	-0.509	0.604	0.297	-0.327	0.361
x2	0.253	-0.519	-0.361	-0.644	-0.341	-0.064
x3	0.227	-0.553	-0.246	0.377	0.608	-0.267
x4	0.553	0.249	-0.249	-0.052	0.262	0.706
x5	0.521	0.254	-0.258	0.451	-0.509	-0.367
x6	0.507	0.199	0.561	-0.384	0.281	-0.402

Table 6: Coefficients de saturation théoriques de la structure uniforme.

variable	PC1	PC2	PC3	PC4	PC5	PC6
x1	-0.455	0.336	-0.087	0.741	-0.328	0.125
x2	-0.439	0.370	-0.212	-0.630	-0.445	-0.175
x3	-0.415	0.422	0.378	-0.110	0.697	0.099
x4	0.434	0.458	0.040	-0.136	-0.167	0.744
x5	0.301	0.435	-0.697	0.114	0.356	-0.306
x6	0.385	0.416	0.563	0.104	-0.234	-0.545

À partir des données simulées, il est possible d'appliquer l'ACP suivi de la rotation varimax ainsi que l'ACP parcimonieuse et voir quelle des deux méthodes permet le mieux de retrouver les structures définit initialement. Les tableaux 7, 8 et 9 montrent les distances entre les vecteurs de coefficients de saturation trouvés par les deux méthodes avec les "vrais" vecteurs de coefficients définit dans les 3 structures différentes.

Dans les tableaux 7, 8 et 9, on remarque que la méthode d'analyse en composante principale parcimonieuse (SPCA) retrouve des coefficients de corrélation qui sont plus près de ceux définis au départ, et ce, pour les 3 types de structure.

Table 7: Distances entre les vecteurs de coefficients de saturation reconstruits d'après les 2 méthodes avec les vecteurs définis dans la structure en blocks.

Méthode	PC1	PC2	PC3	PC4	PC5	PC6
RPCA	0.6843231	0.7235368	0.8555593	0.9405739	0.7977633	0.8543321
SPCA	0.1589661	0.1583390	0.2323479	0.2935097	0.1815585	0.0457828

Table 8: Distances entre les vecteurs de coefficients de saturation reconstruits d'après les 2 méthodes avec les vecteurs définis dans la structure en intermédiaire.

Méthode	PC1	PC2	PC3	PC4	PC5	PC6
RPCA	0.7322748	0.9439513	0.8417853	0.8612562	0.8800388	0.8614375
SPCA	0.5796275	0.5177138	0.6546063	0.7886618	0.6901659	0.5971975

5. Application de la méthodologie

Cette partie a comme but de revenir au jeu de données présenté dans la section "Exemple de motivation" et voir comment se comporte les méthodes introduites dans ce document. On se rappelle que l'ACP classique

Table 9: Distances entre les vecteurs de coefficients de saturation reconstruits d'après les 2 méthodes avec les vecteurs définis dans la structure uniforme.

Méthode	PC1	PC2	PC3	PC4	PC5	PC6
RPCA	0.9095058	0.8363265	0.9970825	0.8611997	0.8636592	0.8542949
SPCA	0.5364418	0.5743705	0.5857645	0.3183985	0.6630822	0.4184422

appliquée sur le jeu de données pitprops permettait de définir des composantes principales qui étaient difficiles à interpréter. Pour palier à ce problème, il était possible de faire une rotation de ces composantes. Toutefois, cela amenait d'autres problèmes comme la perte d'information dans les premières composantes principales ainsi que d'autres problèmes d'interprétabilité.

Dans cette section, nous allons donc comparer les résultats obtenus par les trois méthodes suivantes:

1. L'ACP classique (PCA)
2. L'ACP suivi d'une rotation varimax (RPCA)
3. L'analyse en composante principale parcimonieuse (SPCA)

La première méthode (ACP) est implémentée dans plusieurs librairies *R*, mais pour l'exemple nous avions utilisé la fonction *prcomp* de la librairie **stats**. Pour la deuxième méthode (RPCA), nous avons utilisé la fonction *principal* de la librairie **psych**. Pour la dernière méthode, nous avons utilisé la librairie **elasticnet** et la fonction *spca*.

Les hyperparamètres de la méthode SPCA ont été définis dans le but d'obtenir une variance expliquée similaire à celle de l'ACP pour les 6 premières composantes principales. Voici les paramètres:

$$\lambda = 0.000001$$
$$\lambda_1 = [0.06, 0.16, 0.1, 0.5, 0.5, 0.5]$$

On commence par définir les hyperparamètres définits ci-dessus et on applique la fonction sur le jeu de données pitpros.

```
library(elasticnet)
data("pitprops")

lambda <- 0.000001
lambda_1 <- c(0.06, 0.16, 0.1, 0.5, 0.5, 0.5)

spca_pitprops <- spca(pitprops, K = 6, type = "Gram", sparse = "penalty",
                      para = lambda_1, lambda = lambda)
```

En appliquant la méthode SPCA sur le jeu de données, on obtient des coefficients de saturation qui sont en fonction d'un plus petit nombre de variables. Ces différents vecteurs de coefficients de saturation sont présentés dans le tableau 4.

En analysant le tableau 10, on remarque que 7 variables sur 13 sont non nulles pour la première composante principale. Pour la deuxième et la troisième, 4 sur 13 sont non nulles. Cela permet donc de faciliter l'interprétation des composantes, car on peut immédiatement écarter les coefficients nuls, comparativement à la méthode de la rotation, qui ne permettait pas cela. Ainsi, on peut conclure que la méthode d'analyse en composante principale parcimonieuse donne les résultats les plus faciles à interpréter, suivi dans l'ordre par la méthode de la rotation et de l'ACP classique. Il est également possible de voir, pour chacune des composantes principales, si le nombre de coefficients nuls de la méthode SPCA sont enlignés avec le nombre coefficients "faibles" (<= 0.2) des autres méthodes.

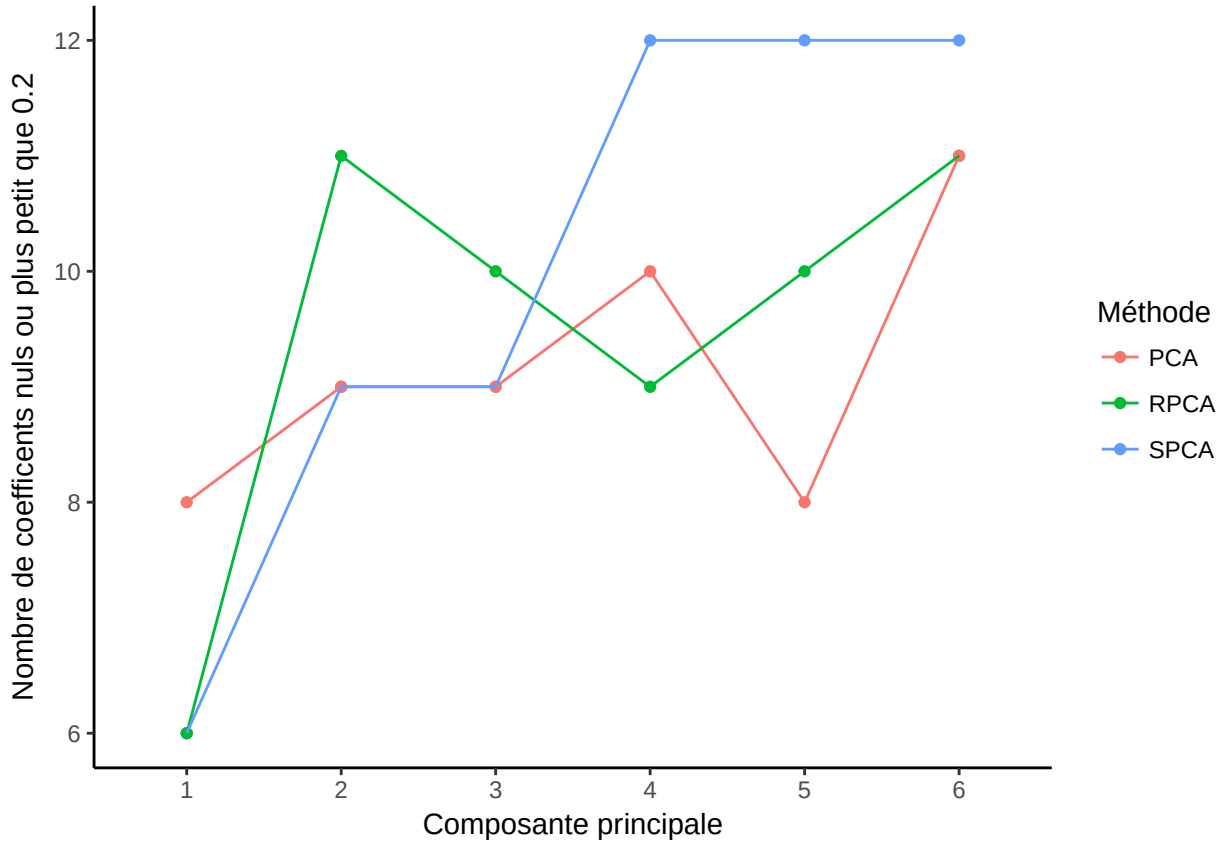

Figure 2: Nombre de coefficients de saturation nuls ou faible (<=0.2) selon les différentes méthodes.

Table 10: Coefficients de saturation de l'analyse en composante principale parcimonieuse (SPCA) effectuée sur la matrice de corrélation du jeu de données pitprops.

	PC1	PC2	PC3	PC4	PC5	PC6
x1	-0.477	0.000	0.000	0.000	0.000	0.000
x2	-0.476	0.000	0.000	0.000	0.000	0.000
x3	0.000	0.785	0.000	0.000	0.000	0.000
x4	0.000	0.619	0.000	0.000	0.000	0.000
x5	0.177	0.000	0.641	0.000	0.000	0.000
x6	0.000	0.000	0.589	0.000	0.000	0.000
x7	-0.250	0.000	0.492	0.000	0.000	0.000
x8	-0.344	-0.021	0.000	0.000	0.000	0.000
x9	-0.416	0.000	0.000	0.000	0.000	0.000
x10	-0.400	0.000	0.000	0.000	0.000	0.000
x11	0.000	0.000	0.000	-1.000	0.000	0.000
x12	0.000	0.013	0.000	0.000	-1.000	0.000
x13	0.000	0.000	-0.016	0.000	0.000	1.000
Variance (%)	0.280	0.140	0.133	0.074	0.068	0.062
Cumulative variance (%)	0.280	0.420	0.553	0.627	0.695	0.758

On remarque à la figure 2 que la régularisation faite par la méthode SPCA est forte sur les composantes 4 à 6. Cela est effectivement contrôlé par l'hyperparamètre λ_1 définit un peu plus tôt. Pour les premières composantes, on voit qu'il y a un lien entre le nombres de coefficients nuls de la méthode SPCA et le nombre de coefficients faibles des autres méthodes. Cela veut donc dire que les coefficients qui ont été tiré vers 0 sont probablement des coefficients "peu importants" et sont ceux que nous aurions exclut intuitivement de notre interprétation.

Comme mentionné précédemment, l'amélioration de l'interprétabilité se fait souvent au prix de perte de l'information, ou de la variabilité dans les composantes principales. En regardant de plus près les tableaux 2, 3 et 4, on remarque que la rotation varimax et la méthode SPCA ont le même genre de pertes au niveau de la variabilité. Toutefois, étant donné que nous avons un gain au niveau de l'interprétabilité avec la méthode d'analyse en composante principale parcimonieuse, il est possible de conclure que celle-ci constitue une amélioration par rapport aux autres méthodes comme la rotation.

6. Autres considérations de la méthode

Dans ce document, nous avons décrits deux méthodes en lien avec l'analyse en composantes principales parcimonieuse, soit la méthode SCoTLASS et la méthode SPCA. Il est pertinent de mentionner qu'une autre méthode a été développée, soit la méthode PMD (Witten and Hastie 2009). Cette dernière est basée sur une décomposition pénalisée de la matrice de covariance (ou corrélation). Il existe d'ailleurs une implémentation de cette méthode en **R** qu'on peut retrouver dans la librairie *PMA*.

7. Annexe

Les méthodes de régularisation (Ridge, Lasso et Elastic net)

Dans le domaine des mathématiques et de la statistique, plus particulièrement dans le domaine de l'apprentissage automatique, la régularisation fait référence à un processus consistant à ajouter de l'information à un problème pour éviter le **surapprentissage**. Cette information prend généralement la forme d'une pénalité.

Une méthode généralement utilisée est de pénaliser les valeurs extrêmes des paramètres, qui sont souvent associées à un surapprentissage. Pour cela, on va utiliser une norme sur ces paramètres que l'on va ajouter à la fonction qu'on cherche à minimiser. Les normes les plus couramment employées pour cela sont les normes L_1 et L_2. La norme L_1 offre l'avantage de faire une sélection de paramètres, mais elle n'est pas différentiable, ce qui peut être un inconvénient pour les algorithmes utilisant un calcul de gradient pour l'optimisation. Cette régularisation permet alors d'avoir des coefficients qui peuvent être estimés exactement à zéro. Par conséquent, ces méthodes peuvent effectuer une sélection des variables importantes pour la variable réponse.

Dans le cadre d'une régression multiple $Y = X\beta + \epsilon$ tel que $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ et $X = (1, X_1, ..., X_p)$, on se trouve parfois dans la situation où le nombre de paramètres (p) est supérieur au nombre d'observations (n). En présence de ce nombre élevé de prédicteurs, il sera nécessaire de réduire le nombre de paramètres, car la solution donnée par la méthode des moindres carrés classique ne sera pas unique et la variance aura tendance à être élevée avec un petit biais. Ainsi, les méthodes de régularisation offrirons une réduction importante de variance et une petite augmentation du biais.

La méthode de Ridge

La méthode de Ridge est une technique de régularisation qui se base sur la norme L_2. Cette méthode a comme pénalité $p(\beta) = ||\beta||_2^2$ avec $||\beta||_2^2 = \sum_{i=1}^{p} \beta_i^2$.

L'estimateur de Ridge de β est défini par

$$\hat{\beta}^{Ridge} = argmin_\beta \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

sous la contrainte de $\sum_{j=1}^{p} \beta_j^2 \leq t$.

L'estimateur de Ridge est donc donné par $\hat{\beta}^{Ridge} = (X^T X + \lambda I)^{-1} X^T y$.

Cette méthode est utilisée dans le cas où il y a une corrélation entre les variables explicatives. Dans ce cas-ci, $X^T X$ a des valeurs proche de zéro et la MMCO n'est pas satisfaisante. Cette méthode permet donc d'ajouter un terme λ pour augmenter la valeur de $X^T X$ et ainsi les rendre plus stables. De ce fait, la méthode contrôle donc la variance des estimateurs en pénalisant les grandes valeurs de $\hat{\beta}$ ce qui a comme avantage l'obtention d'une erreur de prédiction moins faible. Par contre, elle ne permet pas d'avoir un modèle parcimonieux, car on ne pénalise pas les variables nuisibles par des coefficients nuls. Au final, on introduit toutes les variables explicatives dans le modèle ce qui peut compliquer l'interprétation du modèle.

La méthode du Lasso

Cette méthode est basée sur la norme L_1. La pénalité de cette méthode est quant à elle donnée par $p(\beta) = ||\beta||_1$ avec $||\beta||_1 = \sum_{i=1}^{p} |\beta_i|$.

L'estimateur de β par la méthode du Lasso est défini par

$$\hat{\beta}^{Lasso} = argmin_\beta \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

avec $\sum_{j=1}^{p} |\beta|_j \leq t$.

Par un calcul analytique, la solution de ce problème est donnée par

$$\hat{\beta}_j^{lasso} = signe\{\hat{\beta}_j^{MCO}\}(|\hat{\beta}_j^{MCO}| - \lambda)_+$$

Cette méthode permet de créer une parcimonie, car elle élimine les variables nuisibles dans le modèle en estimant leur coefficients par des zéros. Cela permet de réduire le nombre de coefficients (β) en conservant les

variables qui contribuent le plus dans le modèle. Par contre, elle est inappropriée dans le cas où un ensemble de prédicteurs est fortement corrélé, car elle choisit un parmi ceux-ci et annule les autres. Sans oublié le fait qu'elle choisit un maximum de n variables dans le cas où $p > n$. Ces problèmes sont traités par la méthode qui suit.

La méthode Elastic Net

Cette méthode de régularisation combine les deux normes (L_1 et L_2). On peut donc la voir comme un compromis entre la méthode du Lasso et la méthode de Ridge. Sa pénalité est donnée par $p(\beta) = \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$.

L'estimateur de β par la méthode de l'Elastic Net est défini par

$$\hat{\beta}^{EN} = argmin_\beta\{||y - X\beta||^2 \ + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2\}$$

En introduisant la pénalité L_1, on ajoute la possibilité d'obtenir un modèle parcimonieux. L'ajout de la pénalité quadratique L_2 assume la suppression de la limitation du nombre des variables sélectionnées, encourage l'effet du groupe (group effect), car elle permet de sélectionner un nombre de paramètres (p) supérieur à n. De plus, elle permet de tenir en compte la corrélation entre les prédicteurs.

8. 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.

Witten, Robert Tibshirani, Daniela M., and Trevor Hastie. 2009. "A Penalized Matrix Decomposi- Tion, with Applications to Sparse Principal Components and Canonical Correlation Analysis." *Biostatistics*, 515–34.