Analyse en Composante Principal

Sujet : Etude d'une méthode statistique pour l'évaluation des risques de défaillance d'une entreprise

Définition de la méthode ACP

La méthode d'Analyse en Composantes Principales (ou PCA en anglais) est un outil composé de plusieurs calculs statistiques. Elle permet la synthèse d'une grande somme de données quantitatives [1] tout en gardant un maximum d'information. Elle réduit et centre les valeurs autour d'une moyenne et met à l'échelle toutes les données (exemple : mise à l'échelle entre un âge et un chiffre d'affaires).

Autrement dit, elle permet de représenter des résultats complexes sous forme de graphiques et de tableaux simplifiés (en deux dimensions). La lecture du résultat est donc plus simple et facile à interpréter par les data scientistes ou les analystes.

Pour réaliser la méthode ACP, il est nécessaire de suivre un processus précis.

Etape de la méthode ACP

1- Construction d'un tableau

Ce tableau rassemble toutes les données et les variables que l'on souhaite étudier. Il se présente sous la forme d'un tableau individu par variable. Le but de la méthode ACP sera de résumer ce dernier et de visualiser le positionnement des individus les uns par rapport aux autres et de mettre en évidence la corrélation entre les variables.

Exemple simple d'un tableau individu par variable :

	Sexe	Age	Profession
Jean	M	38	Ingénieur d'études
Marie	F	43	Chef de projet
Bernard	M	70	Retraité

Dans notre tableau, Jean, Marie et Bernard sont associés à des variables de poids, d'âge et de revenus.

Mises dans un graphique, ces données représentent deux sortes de nuages de points.

— Un nuage de points individus (n individus)

— Un nuage de points variables (m dimensions)

Quand la dimension des variables est supérieure à 3, il devient compliqué de visualiser et donc de mettre en évidence les relations entre les variables et les individus. C'est pour cela que l'on transforme les données pour les représenter dans un espace en deux dimensions.

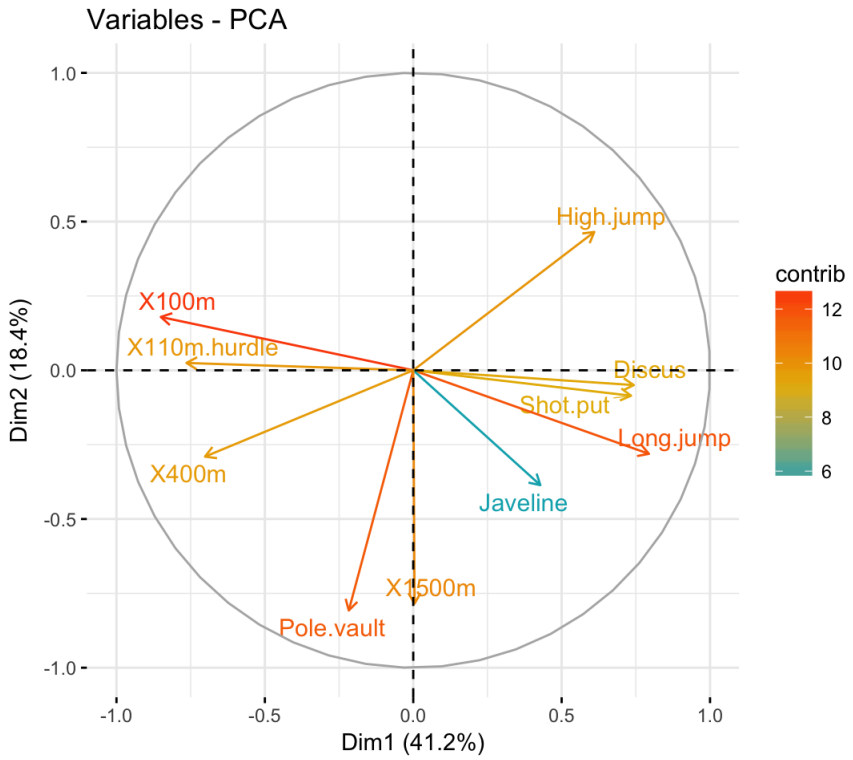
1. données numérique mesurable

2- Analyse directe

L'analyse directe est la construction d'un espace associé au nuage de points-individus. Avec cette analyse, on peut regrouper les individus adjacents. Ainsi nous pouvons construire des groupes d'individus (techniques de clustering [2]). Cela peut être utile pour faire de la segmentation et à la suite réaliser un scoring sur ces groupes.

3- Analyse duale

L'analyse duale est la construction de l'espace associé au nuage de points-variables. m axes (m mesures en entrée) peuvent être représentés par des vecteurs. Prenons l'exemple du décathlon. Nous avons 10 épreuves (variables) où les individus (athlètes) ont chacun leur résultat par épreuve. Nous pouvons observer ci-dessous les différents vecteurs des variables épreuves et leur direction dans l'espace créé par le PCA.



On peut aussi observer les variables qui sont corrélées. Dans cet exemple, il est donc logique de retrouver les courses de sprint vers le même axe et assez proche les uns des autres (100m,110m haies,400m). La couleur contribution sert à démontrer l'importance de l'écart-type. Ainsi plus la contribution est importante (rouge) plus les valeurs sont disparates. Ainsi au javelot il y a moins de disparité au niveau des points attribué qu'au saut à la perche (pole vault).

4- Interprétation des résultats

Lors de cette étape, il faut choisir un certain nombre d'axes pouvant former un espace en 2 ou 3 dimensions selon le modèle. La construction des nuages de points est projetée sur ces différents axes. S'en suit une interprétation des axes principaux parmi le nombre d'axes une étude de proximité entre les points (proximité entre individu et variable).

5- Synthèse des résultats

Il est possible de créer un tableau réduit appelé tableau des composantes principales. Il permet de visualiser un nuage de points simplifié avec les axes des variables. On peut alors faire du clustering et comprendre le lien entre les individus et les variables.

2. Méthode d'analyse statistique permettant de partitionner des données

Pourquoi utiliser cette méthode ?

Le data mining, minage/exploration/fouille de données en français, défini tout un ensemble d'outils facilitant l'exploration et l'analyse de données. Le but est de déterminer des relations entre les variables ou de repérer des modèles pour des études décisionnelles. C'est une technique utilisée dans le big data pour traiter les larges volumes et variétés de données.

Les résultats peuvent ensuite être réutilisés par différentes entités (des entreprises par exemple) pour augmenter leur productivité (comme le chiffre d'affaire) ou réduire des coûts. Ils peuvent aussi servir à déterminer le comportement de clients. Par exemple grâce à une technique de data mining il a été prouvé que les couches étaient souvent achetées en même temps qu'un pack de bière. En réponse à cette observation certains magasins ont mis ces deux produits à proximité pour maximiser leurs profits.

Pour obtenir ces résultats, les organismes déterminent des comportements ou des schémas grâce à des algorithmes plus ou moins complexes.
L'ACP est une méthode statistique sous-jacente au data mining et qui permet de répondre à ce genre de problématique. Elle offre aux entreprises un outil permettant l'interprétation de données, la synthèse d'informations, et permet d'identifer la corrélation entre plusieurs facteurs.

Pour faire suite à la définition de la méthode ACP détaillée dans la première partie du document, nous remarquons qu'elle peut apporter de nombreuses réponses aux problématiques auxquelles certains organismes sont confrontés. Elle permet de prendre en compte autant de variables que l'on souhaite tout en gardant un maximum d'information. Elle permet ainsi d'interpréter les données, de synthétiser l'information et de détecter des corrélations entre plusieurs facteurs.

Cette méthode permet également un gain de temps considérable. Son usage ne nécessite qu'une série de calculs pour traiter toutes les variables. Le résultat présentera directement une vue avec toutes ces variables. Avec une autre méthode, il faudrait choisir un plus petit nombre de variables à croiser, effectuer les calculs, visualiser les résultats et finalement faire une interprétation générale. Cela reviendrait à avancer à tâtons afin de trouver la corrélation intéressante à exploiter. Bien sûr, plus le nombre de variable est grand, plus le temps pour obtenir le résultat final sera long.

Exemple d'implémentation de l'ACP

Notre but avec la technique ACP va être d'extraire l'information d'un grand espace dimensionnel pour le projeter sur un plus petit sous espace. Avec cette technique nous pourrons obtenir des informations essentielles entre les variables et les individus. Nous allons pouvoir faire du clustering afin de classer les données.

Pour cela créons un exemple fictif afin de démontrer un cas potentiel d'utilisation. Voici l'exemple d'une base de données fictive avec les labels correspondant aux attentes de TradeIn.

Client	Chiffre d'Affaire	Capital	Année d'expérience	Vente	Croissance des bénéfices
A	100000	125000	5	600	-40000
B	250000	1000000	30	1500	-50000
C	600	15000	1	150	600
D	750000	2500000	15	2000	50000
E	15000	100000	3	1000	9000
F	160000	120000	8	1350	7500

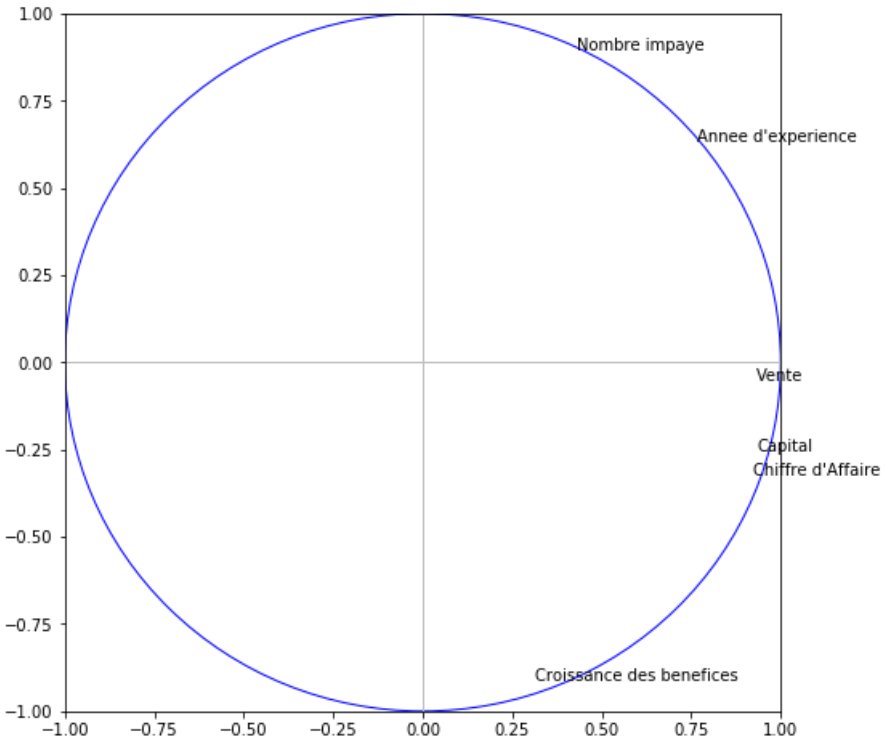
Avec le langage informatique Python nous allons utiliser l'outil de statistique inclus dans la librairie Sklearn pour réaliser l'analyse.
Nous utilisons pour cet exercice Python 3.7 avec Spyder comme IDE2. Premièrement nous récupérons les données stockées dans le fichier Excel.

Client	Chiffre d'Affaire	Capital	...	Croissance des benefices	Nombre impaye
			...		
A	100000.0	125000.0	...	-40000.0	5.0
B	250000.0	1000000.0	...	-50000.0	20.0
C	600.0	15000.0	...	600.0	1.0
D	750000.0	2500000.0	...	50000.0	3.0
E	15000.0	100000.0	...	9000.0	2.0
F	160000.0	120000.0	...	7500.0	6.0

Ensuite nous allons passer à la partie technique et statistique. Pour cela nous importons la librairie Sklearn. Nous allons standardiser et transformer les données. L'objectif sera de centrer et mettre à l'échelle les variables afin de préparer à réaliser un ACP normé.

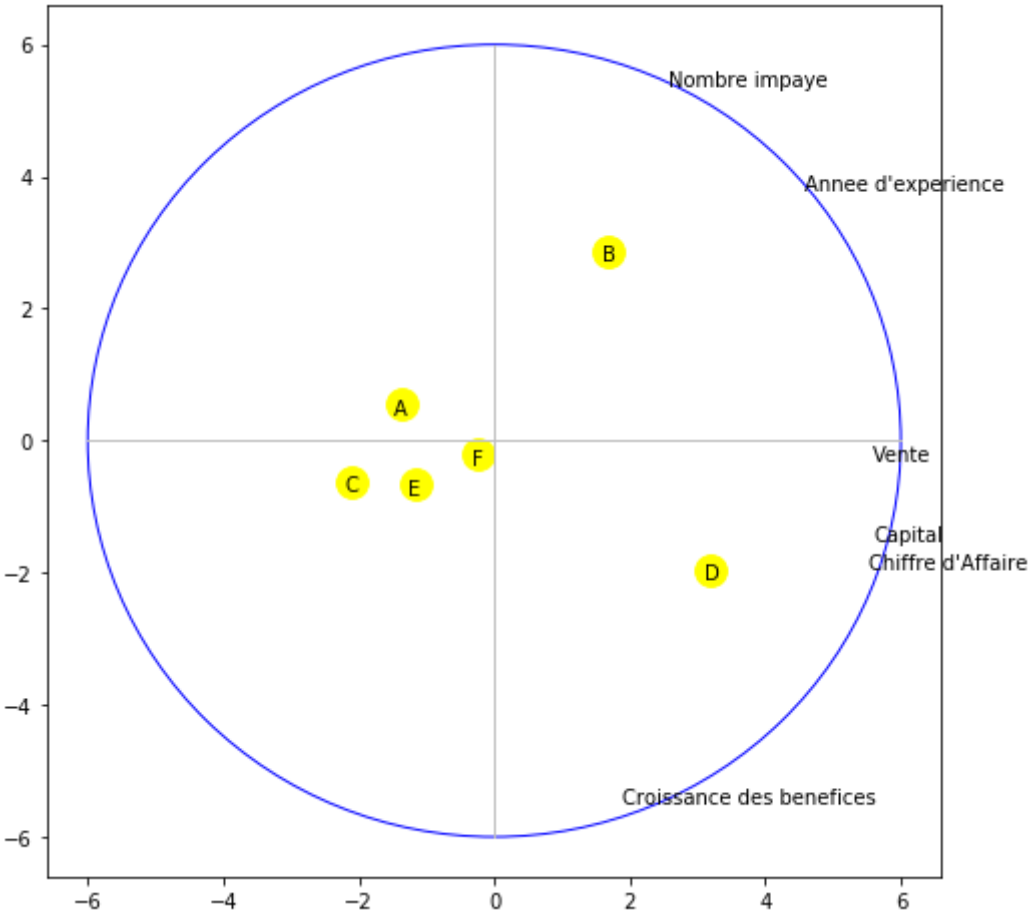
```
[[-0.4418616  -0.57909319 -0.54089872 -0.82666275 -1.0868424  -0.1818949 ]
 [ 0.14676398  0.3984757   1.99456404  0.6613302  -1.3872134   2.15675376]
 [-0.83192415 -0.70198756 -0.94657277 -1.57065922  0.13266386 -0.80553454]
 [ 2.10884924  2.07430807  0.47328638  1.48799295  1.6164966  -0.49371472]
 [-0.77541609 -0.60702373 -0.74373574 -0.16533255  0.3849755  -0.64962463]
 [-0.20641137 -0.5846793  -0.23664319  0.41333137  0.33991985 -0.02598499]]
```

Désormais les données sont réduites et centrées et les écarts types sont unitaires (égaux à 1). Tout est en place pour démarrer l'étude ACP.
Nous construisons un cercle de corrélation afin d'observer la corrélation entre les différentes variables.



Dans cet exemple fictif, nous observons une certaine corrélation entre les années d'expériences et le nombre d'impayé. Il y a aussi une forte corrélation entre le capital et le chiffre d'affaire. Ces différents points sont très proches les uns des autres. Plus un point est proche d'un autre plus la corrélation est forte.

Nous ajoutons les individus au cercle des corrélations. Nous pouvons ainsi voir comment les individus interagissent avec les variables. Sur ce schéma, nous pouvons voir d'intéressantes relations.



Nous pouvons mettre dans une même classe les entreprises A,C,E,F. Ces points sont très proches les uns des autres et forment un groupe. Ils se rapprochent beaucoup de l'origine (0,0) qui représente la moyenne. L'entreprise B quant à elle a de plus d'impayés que les autres entreprises. C'est donc normal de la retrouver dans cette position proche de l'axe du nombre d'impayé.
L'entreprise D a peu d'impayé et possède un plus grand chiffre d'affaire et capital. Il est donc logique de le retrouver dans cette position, proche de ces différentes variable cité précédemment.

Cet exemple permet de mieux comprendre ce qu'est une ACP. Elle donne une meilleur vision des données et de leur relation grâce au graphique résultant de l'analyse.

Ce programme a été réalisé en Python mais il peut également être programmé en langage R ou avec XL-Stat. Ces outils possèdent également des librairies et packages appropriés pour permettre des études avec la méthode ACP. Chacun possède des spécificités sur la présentation des données. Le choix des outils est subjectif et dépend des compétences des personnes souhaitant réaliser une ACP.

La méthode ACP pour le projet TradeIn : Limites et alternatives

L'ACP est peu robuste pour les valeurs aberrantes. En effet, ce genre de valeurs vont fausser les résultats en augmentant considérablement la moyenne et/ou l'écart type. Il faut un traitement préalable pour éviter que ce type de données parasitent le modèle.

D'autre part cette technique est limitée aux variables quantitatives. Il est aussi regrettable que chaque composant soit pris en compte de la même manière. Les variables n'ont pas toujours la même importance et un tel système ne prend pas en compte cet aspect.

Pour répondre à cette problématique, nous proposons une alternative : le Machine Learning. Cette stratégie est très flexible car elle comporte plusieurs types d'algorithme. On peut appliquer ces techniques sur différents types de variable comme les nombres ou les chaines de caractères par exemple. De plus, l'algorithme va classer les variables les plus intéressantes en fonction de leur impact sur le résultat (le nombre d'impayés par exemple). On peut aussi interpréter les variables texte, faire du clustering plus adapté et plus flexible en fonction des paramètres accordés pour gérer la notation de ces groupes (scoring). Dans un exemple ce type nous pourrions regrouper les entreprises en fonction de leurs caractéristiques pour trouver les nombres d'impayés. Les techniques de clustering de Machines Learning tel que SVM, KNN ou Sklearn Cluster seraient intéressante à analyser pour approfondir cette option.

Conclusion

Pour conclure, la méthode de l'ACP présente de nombreux avantages comme l'analyse multivariée qui permet une visualisation simple et efficace des liens entre les données. Cependant, elle présente certaines limites comme sa non-robustesse aux valeurs aberrantes. Elle ne permet pas non plus d'intégrer l'importance de certaines variables ou de réaliser un scoring automatique. Ces dernières problématiques peuvent être résolues en utilisant d'autres méthodes. Le machine learning est une alternative intéressante à étudier pour trouver une solution pertinente au problème posé initialement. Il est aussi possible d'allier les deux méthodes. Le machine learning et la méthode de l'ACP peuvent être complémentaires pour apporter une solution complète au problème.

Bibliographie

L'intérêt de l'analyse en composantes principales (ACP) pour la recherche en sciences sociales
https ://journals.openedition.org/cal/7364

Data Mining : Analyse factorielle
https ://www.slideshare.net/mohamedhenyselmi/data-mining-acp-analyse-en-composantes-principales

Data Mining : qu'est ce que l'exploration de données ?
https ://www.lebigdata.fr/data-mining-definition-exemples

Soyez attentif aux spécificités de l'ACP
https ://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees/5280463-soyez-attentif-aux-specificites-de-lacp

Analyse en Composantes Principales (avec SPAD) et Classification Ascendante Hiérarchique
http ://irma.math.unistra.fr/~fbertran/enseignement/DataMining_2010/ACP_spad.pdf

Techniques Exploratoires Multivariées : Analyse en Composantes Principales
http ://statsoft.fr/concepts-statistiques/analyse-en-composantes-principales/analyse-en-composantes-principales.php#.Xfji6(

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)
https ://www.xlstat.com/fr/solutions/fonctionnalites/analyse-en-composantes-principales-acp

Principal Component Analysis (PCA) in Python
https ://www.datacamp.com/community/tutorials/principal-component-analysis-in-python

Comment évaluer la santé d'une entreprise ?
https ://www.l-expert-comptable.com/a/532120-comment-evaluer-la-sante-d-une-entreprise.html

Sept clignotants à surveiller pour savoir si votre boîte est en bonne santé
https ://lentreprise.lexpress.fr/gestion-fiscalite/budget-financement/sept-clignotants-a-surveiller-pour-savoir-si-votre-boite-est-en-bonne-sante_1524227.html

Tanagra Data Mining
http ://eric.univ-lyon2.fr/ricco/tanagra/fichiers/fr_Tanagra_ACP_Python.pdf

Articles - Méthodes des Composantes Principales dans R : Guide Pratique
http ://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/79-acp-dans-r-prcomp-vs-princomp/