COURS DE DATA MINING
6 : MODELISATION NON-SUPERVISEE
LES ANALYSES FACTORIELLES

EPF – 4/ 5ème année - Option Ingénierie d'Affaires et de Projets - Finance
Bertrand LIAUDET

6 : Modélisation non-supervisée - 2 : les analyses factorielles 2
Généralités sur les analyses factorielles 2
Principes 2
Techniques 2
Techniques de l'analyse en composantes principales : ACP 2
Principes 2
Distance et métrique 3
Matrices des covariances et des corrélations 4
Calcul matriciel 5
Qualité et choix d'une composante principale : la valeur propre 5
Qualité et choix d'une composante principale : la valeur propre 6
Exemple 6
Usages de l'ACP : le plan factoriel et le cercle des corrélations 9
Audit et histogramme des composantes principales 9
Le plan factoriel : l'espace des individus 10
Contribution des individus 11
Le cercle des corrélations : l'espace des variables 12
Les rotations de l'ACP 16
Conclusion sur les usages 16
Présentation de l'analyse factorielle des correspondances : AFC et ACM 17
Principes 17
Technique 17
Usages de l'AFC et de l'ACM 20
Plan factoriel et cercle des corrélations 20
L'offre Clementine : exemple 06 20

6 : MODELISATION NON-SUPERVISEE - 2 : LES ANALYSES FACTORIELLES

Généralités sur les analyses factorielles

Principes

Les analyses factorielles permettent de :

- Représenter graphiquement (en 2 ou 3 dimensions) les individus d'une population (donc de façon intuitive).
- Détecter les liaisons entre les variables.
- Détecter les variables séparant le mieux les individus.

Les analyses factorielles produisent de nouvelles variables, les « axes factoriels » qui sont des combinaisons linéaires des variables initiales. Les axes factoriels sont plus ou moins corrélés aux variables initiales.

Techniques

Les techniques factorielles regroupent :

- Les ACP : analyses en composantes principales.
- Les AFC : analyses factorielles des correspondances.

Ces analyses sont très appréciées des statisticiens.

C'est une analyse peu répandue dans les pays anglo-saxons : on l'appelle aussi « analyse des données à la française ».

Techniques de l'analyse en composantes principales : ACP

Principes

Créer une ou plusieurs variables qui soient une combinaison linéaire de n variables

Un individu décrit par n variables appartenant à R peut être représenté par un point dans un espace R^n à n dimensions.

Malheureusement, il n'y a pas de représentation graphique possible d'un espace de dimension supérieure à 3 !

L'ACP permet, à partir de n variables appartenant à R, de construire m (<=n) autres variables, appelées **composantes principales (ou axes factoriels),** qui sont des **combinaisons linéaires des n variables initiales**.

L'analyse en composantes principales est donc une sorte de régression linéaire à N dimension qui crée une variable. Chaque composante principale sera donc défini par une équation linéaire mettant en jeu les variables qui ont participé à sa construction.

Caractéristiques des variables composantes principales

Ces nouvelles variables ont les caractéristiques suivantes :

- Elles sont caractérisées par la **mesure de l'information qu'elle restitue** des n variables initiales.
- Elles sont **ordonnées selon l'information qu'elle restitue** des n variables initiales.
- Elles ne sont **pas corrélées linéairement entre elles**.
- Les premières, au moins, sont **moins sensibles aux variations de la population** que les variables initiales.

Distance et métrique

L'objectif de la présentation technique est de comprendre les grandes lignes du fonctionnement de l'analyse en composantes principales et les paramétrages possibles.

Distance entre les individus

Dans un espace à 2 dimensions, la distance entre deux individus (deux points), i1=(x1, y1) et i2=(x2, y2) est donnée par la formule :

$$d(i1, i2) = \sqrt{ (x2-x1)^2 + (y2-y1)^2 }$$

Dans un espace à N dimensions, le principe est le même. La distance entre deux individus est donnée par la formule :

$$d(i1, i2) = \sqrt{ (x2-x1)^2 + (y2-y1)^2 + (z2-z1)^2 + \ldots (n2-n1)^2 }$$

x, y, z, …, n : représentent les N variables et donc les N axes de l'espace à N dimensions.

Métrique euclidienne

La métrique euclidienne c'est celle du calcul d'une distance dans un espace de dimension N.

Cette métrique est adaptée à des variables dont les valeurs sont semblables. C'est le cas pour des données du monde physique (par exemple les taux des différents composants dans un produit alimentaire).

Métrique économétrique : « inverse des variances »

Les données économétriques présentent des variables très dissemblables (l'âge, le revenu, le chiffre d'affaire, etc.)

On utilise donc une autre métrique

$$d(i1, i2) = \sqrt{ ((x2-x1) / ecx)^2 + ((y2-y1) / exy)^2 + \ldots ((n2-n1) / ecn)^2 }$$

x, y, …, n : représentent les N variables et donc les N axes de l'espace à N dimensions.

ecx, ecy, …, ecn : représentent les écarts types (ec) des N variables x (ecx), y (ecy), …, n (ecn)

L'écart-type étant la distance typique entre un individu et la moyenne des individus, chaque distance pour une variable (x2-x1) est **ramenée à une proportion** par rapport à l'écart type.

Rappelons que la variance est moyenne des déviations au carré de chaque variable par rapport à la moyenne de l'ensemble des variables (somme($(xi - moy(x))^2$) / n) et que l'écart-type est la racine carré de la variance.

Cette technique permet de ramener les ordres de grandeurs de chaque variable à des valeurs homogènes.

Matrices des covariances et des corrélations

Calcul matriciel : matrice des covariances et matrice des corrélations

Pour déterminer les composantes principales, on part d'une matrice carrée des variables.

Le choix du type la matrice est un paramètre de l'ACP.

Il est directement fonction de la métrique adaptée au problème.

Matrice des covariances

- ***La covariance***

La covariance de deux variables v1 et v2 est un indicateur de la variation simultanée.

La covariance est positive quand v2 croît chaque fois que v1 croit. Elle est négative quand v2 décroît chaque fois que v1 croit. Elle est nulle si v1 et v2 sont indépendantes.

Covariance et corrélation sont de même signe.

La covariance est fonction du coefficient de corrélation :

cov (v1, v2) = écart-type(v1) * écart-type(v2) * corrélation(v1, v2)

- ***Choix de la matrice des covariances***

Si on a des **données homogènes avec des ordres de grandeurs identiques** (typiquement dans le cas de **données physiques**), alors on a une métrique euclidienne et on travaille avec une matrice des covariances.

A chaque couple de variables (v1, v2), la valeur de la case de la matrice est celle de la covariance du couple (v1, v2).

Matrice des corrélations

- ***La corrélation***

corrélation(v1, v2) = cov (v1, v2) / (écart-type(v1) * écart-type(v2))

A chaque couple de variables (v1, v2), la valeur de la case de la matrice est celle du coefficient de corrélation entre v1 et v2.

- ***Choix de la matrice des corrélations***

Si on a des **données hétérogènes avec des ordres de grandeurs différents** (typiquement dans le cas de **données économétriques**), alors on a une métrique « inverse des variances » et on travaille avec une matrice des corrélations.

Calcul matriciel

Les composantes principales

A partir d'une matrice, on peut calculer les composantes principales par calcul matriciel (diagonalisation de la matrice).

Chaque composante principale est une combinaison linéaire des variables impliquées dans sa détermination. C'est cette fonction qui permet de calculer la valeur de la composante pour chaque individu.

Le calcul de la composante est fait de telle sorte que :

- la somme et la moyenne de la composante valent 0 ;
- la variance et l'écart-type de la composante valent 1.

Qualité et choix d'une composante principale : la valeur propre

Valeur propre des composantes principales

- Chaque composante principale a une **valeur propre**.
- La valeur propre d'une composante principale est égale à la **somme des coefficients de corrélation au carré de chaque variable** d'entrée avec la composante.
- La notion de valeur propre est assez abstraite. On peut dire que **chaque valeur propre mesure la part de variance totale des variables impliquées pour la composante principale correspondante** (rappelons que la variance : 1) caractérise une variable et les valeurs de sa population ; 2) mesure une dispersion qui est la moyenne des déviations au carré de chaque individu par rapport à la moyenne de la variable).
- La somme des valeurs propres correspond à la variance totale
- Les composantes principales sont **classées par valeur propre décroissante**. Pour mieux situer le niveau de l'information restituée par chaque composante, on donne aussi la **proportion de la valeur propre de chaque composante** par rapport à la somme des valeurs propres de toutes les composantes.

Combien de composantes principales faut-il garder ?

Il y a 3 critères empiriques pour savoir combien de composantes principales garder :

- ***Le critère de Kaiser***

Si on a utilisé une matrice des corrélations (cas le plus courant), on ne garde que les composantes principales dont la valeur propre est > 1.

Ce n'est pas un critère absolu.

- ***Valeur du pourcentage***

La valeur propre est aussi donnée en %age. On peut garder les %ages significatifs.

En regardant la courbe des %ages cumulé, on peut faire apparaître un moment de flexion significatif qui montre qu'à partir de là, il y a peu d'information restituée.

- ***Valeurs des coefficients de corrélation avec les variables d'origine***

On peut aussi ne garder que les composantes principales qui ont un coefficient de corrélation élevé avec au moins une variable d'origine.

Qualité et choix d'une composante principale : la valeur propre

Valeur propre des composantes principales

- Chaque composante principale a une **valeur propre**.

Exemple

On part d'un fichier de voitures.

L'audit des données montre des valeurs aberrantes l'origine, le poids et la consommation : on corrige les origines et on retire les individus concernés de l'analyse.

Le nœud ACP-facteur élimine automatiquement les enregistrements incomplets. Il travaille uniquement sur les attributs numériques (type intervalle).

Résultats généraux

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	5.042	72.023	72.023	5.042	72.023	72.023
2	.908	12.977	85.000	.908	12.977	85.000
3	.643	9.184	94.184	.643	9.184	94.184
4	.196	2.796	96.979	.196	2.796	96.979
5	.125	1.785	98.764	.125	1.785	98.764
6	5.27E-002	.753	99.517			
7	3.38E-002	.483	100.000			
Méthode d'extraction : Analyse en composantes principales.						

Tableau des variances

Le calcul produit le nombre maximum de composantes principales : ici 7, qui correspond au nombre d'attributs en jeu.

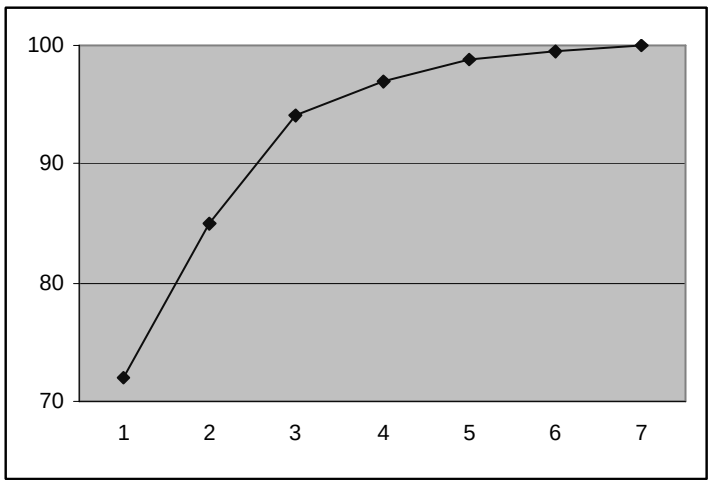
Le logiciel ne garde que 5 composantes : c'est un paramètre fixé et qu'on peut modifier.

Choix du nombre de composantes à exploiter

- ***D'après le critère de Kaiser***

On ne devrait garder que la 1, la 2 étant limite.

- ***D'après le critère des pourcentages***



Courbe des %ages cumulés par composante principale

On voit un coude à partir de la 3ème composante : on pourrait donc garder les 3 premières composantes.

- ***D'après le critère des valeurs du coefficient de corrélation avec la variable d'origine***

	Composante				
	1	2	3	4	5
Consommation en miles par gallon	-.888	.201	-.201	.347	8.84E-002
Poids	.924	.204	.250	-3.49E-002	9.97E-002
Cylindres	.933	.167	.123	.189	-.202
cm3	.963	.149	9.93E-002	.132	-1.42E-002
hp	.947	8.56E-002	-.145	2.04E-002	.244
time-to-60	-.702	-1.95E-002	.703	6.01E-002	7.69E-002
Année	-.461	.877	-2.25E-002	-.131	-2.46E-002
Méthode d'extraction : Analyse en composantes principales.					
a. 5 composantes extraites.					

Tableau des coefficients de corrélation
entre variables d'origine et composantes principales

Seules les trois premières composantes présentent au moins un coefficient de corrélation élevé (0,877 pour CP2, 0,703 pour CP3 tandis que pour CP4 on passe à 0,347).

- ***Equations des composantes***

L'équation de la première composante est la suivante :

Composante 1 =

-0,02246 * Consommation en miles par gallon +
0,000216 * Poids +
0,1075 * Cylindres +
0,001779 * cm3 +
0,004688 * hp +
-0,04853 * time-to-60 +
-0,02526 * Année +
+ 49,12

Usages de l'ACP : le plan factoriel et le cercle des corrélations

Audit et histogramme des composantes principales

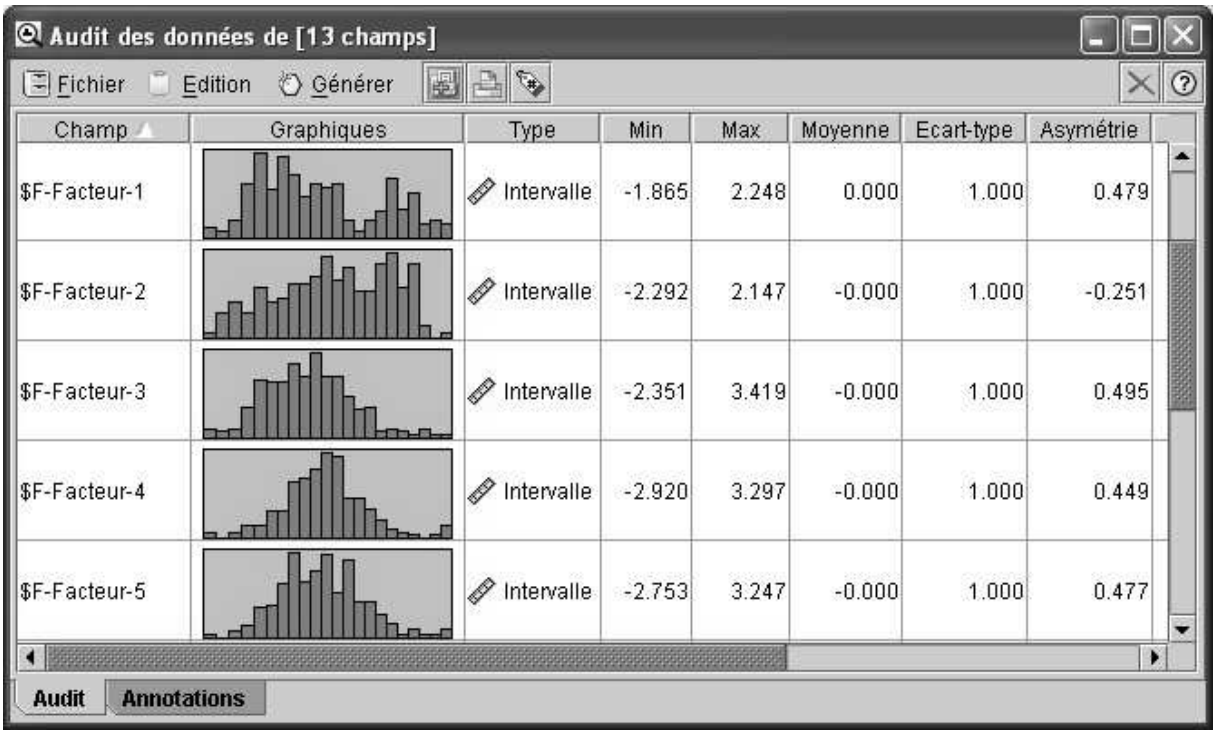
L'audit des composantes principales

L'audit des composantes principales montre les caractéristiques suivantes :

La moyenne vaut 0 et l'écart type vaut 1 pour toutes les composantes principales. Cela vient du principe même de la construction des composantes principales.

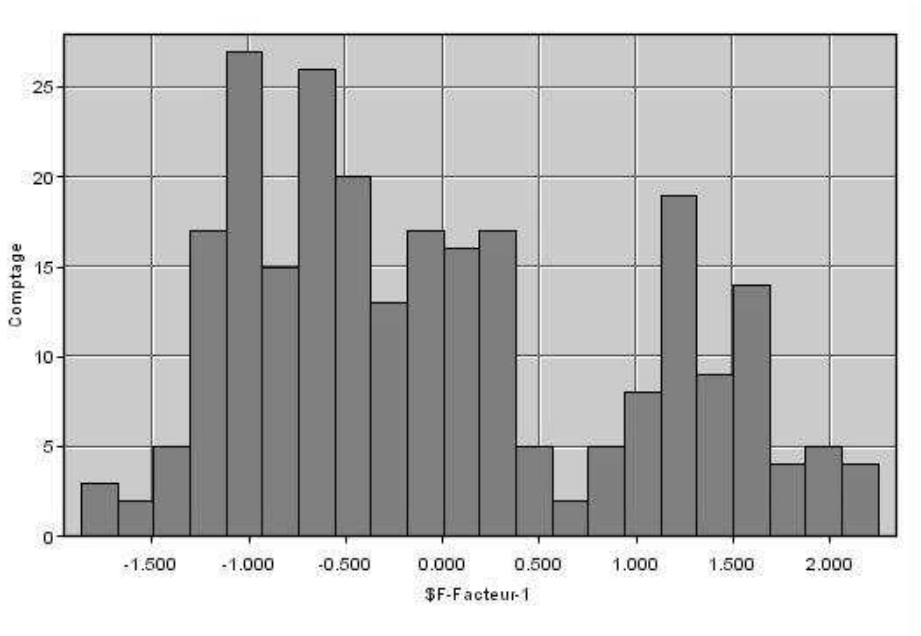
L'analyse de chaque histogramme des composantes principales permet de montrer des individus hors nomes (outliers) : individus isolés aux extrémités le plus souvent. On aura intérêt à retirer ces individus dans la suite de l'analyse pour qu'ils ne faussent pas les résultats.

L'analyse des histogrammes peut aussi permettre de mettre au jour des classes d'individus différents.



Histogramme des composantes principales

Dans l'audit, on voit déjà que la première composante a un histogramme qui présente deux courbes de Gauss et donc sépare la population en deux ensembles : c'est un bon critère pour distinguer deux classes.



On peut ensuite analyser les caractéristiques des deux sous-ensembles :

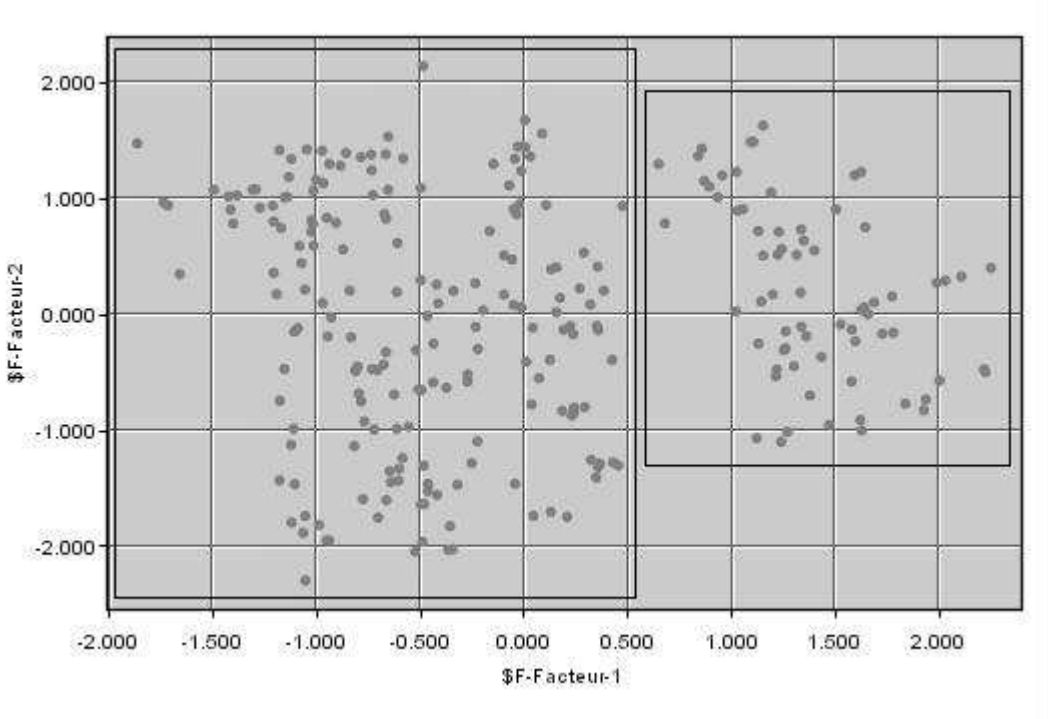
Les deux audit vont mettre au jour d'un côté les américaines à forte cylindré et de l'autre un mélange d'américaines, européennes et japonaises à cylindrés moyennes et faibles.

Le plan factoriel : l'espace des individus

Présentation

Le plan factoriel est le plan dont les axes sont constitués par des composantes principales (la première et la deuxième prioritairement).

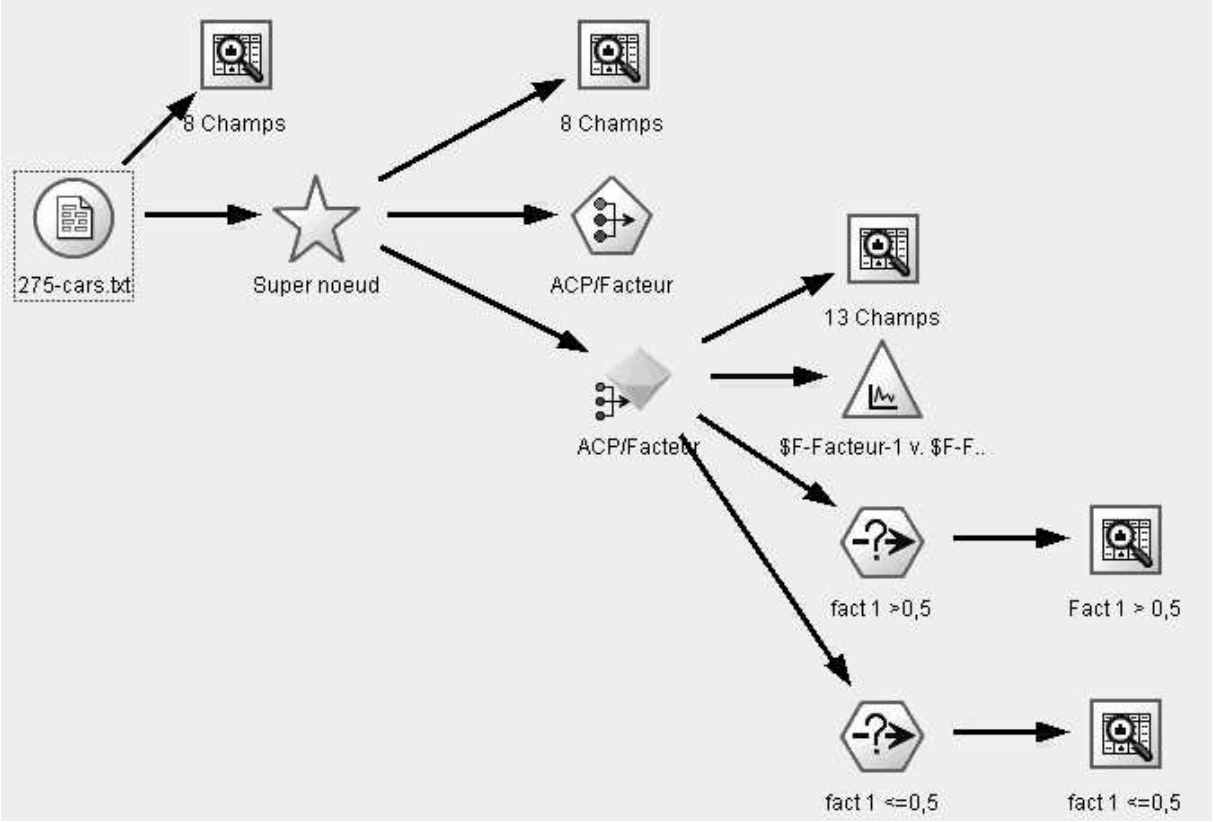
L'analyse du plan factorielle est une analyse du nuage des points dans ce plan. Il va permettre, éventuellement, de définir des classes dans la population.



L'exemple ci-dessus montre deux classes dans la population des voitures. La limite se trouve autour de la valeur 0,5 pour le facteur 1. Ce qui rejoint l'analyse de l'histogramme du facteur 1.

Clémentine permet de générer le nœud de sélection correspondant à chacune de ces classes.

Bilan des flux des exemples n° 1 et 2 :



Contribution des individus

Principes

Etant donné que :

$$\text{variance}(x) = \text{somme}((xi - \text{moy}(x))^2) * pi$$

avec

x : la variable considérée

pi : poids de l'individu i, en général 1/n

xi : valeur de l'individu i

et que, pour chaque composante principale, moy(cp)=0

on obtient :

$$\text{variance}(x) = \text{somme}(xi^2) * pi$$

de qui est équivalent à :

$$\text{variance}(x) = \text{somme}(pi * xi^2)$$

La contribution d'un individu i est donc donnée par la formule suivante :

contribution (i) = pi * xi^2 / variance(x)

Une règle empirique dit que :

Une contribution qui dépasse le poids de l'individu est importante.

Une contribution qui dépasse 0,25 est très importante.

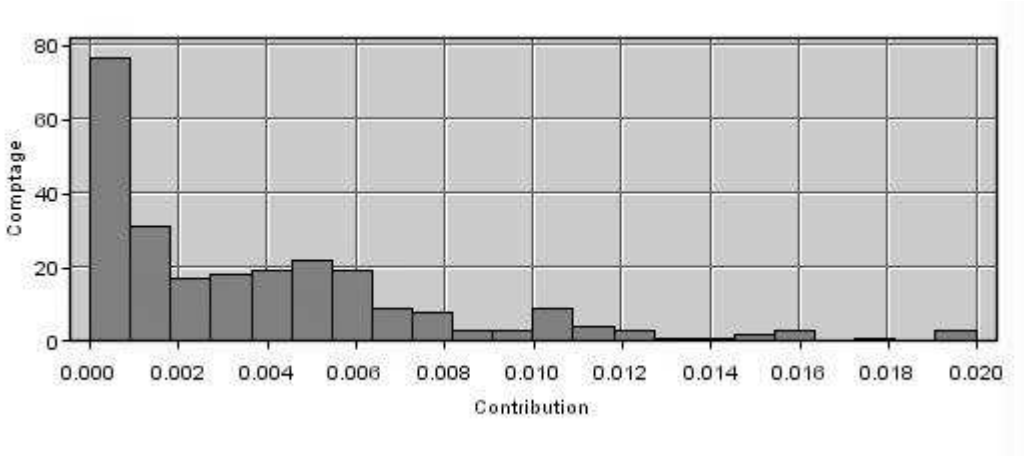
Dans les deux cas, on peut être amené à retirer les individus de l'analyse.

Exemple

On repart de l'exemple 2. On fait un « super nœud » avec les « remplacer » et les « sélection ».

Derrière le nœud « ACP facteur », on crée un champ « contribution ».

L'audit de données montre que la contribution est comprise entre 0 et 0,02.



Aucune contribution n'est supérieure à 0,25.

Par contre, le poids des individus vaut 1 /253 soit 0,0039. On a donc une centaine d'individus à contribution importante. Il paraît difficile de s'en séparer pour l'analyse.

Le cercle des corrélations : l'espace des variables

Présentation

On peut calculer le coefficient de corrélation de chaque variable d'origine avec toutes les composantes principales.

Le coefficient de corrélation est une valeur comprise entre –1 et 1.

Le cercle des corrélations est le plan dont les axes sont constitués par des composantes principales (la première et la deuxième prioritairement).

Chaque variable d'origine a des coordonnées dans ce plan.

On a un cercle, car les corrélations c1 et c2 d'une variable d'origine avec deux composantes principales sont telles que :

$$c1^2 + c2^2 <= 1$$

La projection des variables dans ce plan (nuage de points) permet visuellement de :

- Détecter les variables d'origine liées entre elles.
- Interpréter chaque composante principale d'après ses corrélations avec les variables d'origine.

Interprétation

- Des projections proches entre elles et proches du cercle des corrélations correspondent à des variables corrélées.
- Des projections proches de l'horizontale montrent une corrélation avec la composante principale horizontale.
- Des projections proches de la verticale montrent une corrélation avec la composante principale verticale.

Exemple 04

On repart de la situation de l'exemple 3.

A partir du résultat de l'ACP, on arrive au tableau suivant :

	Composante				
	1	2	3	4	5
Consommation en miles par gallon	-.888	.201	-.201	.347	8.84E-002
Poids	.924	.204	.250	-3.49E-002	9.97E-002
Cylindres	.933	.167	.123	.189	-.202
cm3	.963	.149	9.93E-002	.132	-1.42E-002
hp	.947	8.56E-002	-.145	2.04E-002	.244
time-to-60	-.702	-1.95E-002	.703	6.01E-002	7.69E-002
Année	-.461	.877	-2.25E-002	-.131	-2.46E-002
Méthode d'extraction : Analyse en composantes principales.					
a. 5 composantes extraites.					

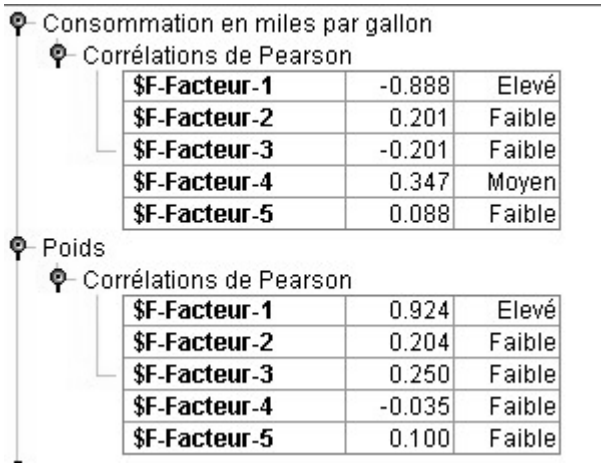
Tableau des coefficients de corrélation

entre variables d'origine et composantes principales

Ce tableau fournit toutes les corrélations entre les composantes et les variables en entrée.

➢ ***Vérification 1***

Si on calcule les corrélations entre toutes les variables d'origine et les composantes principales par un nœud « statistiques », on retrouve les mêmes valeurs.



Consommation en miles par gallon

Corrélations de Pearson

$F-Facteur-1	-0.888	Elevé
$F-Facteur-2	0.201	Faible
$F-Facteur-3	-0.201	Faible
$F-Facteur-4	0.347	Moyen
$F-Facteur-5	0.088	Faible

Poids

Corrélations de Pearson

$F-Facteur-1	0.924	Elevé
$F-Facteur-2	0.204	Faible
$F-Facteur-3	0.250	Faible
$F-Facteur-4	-0.035	Faible
$F-Facteur-5	0.100	Faible

➢ ***Vérification 2***

Si on fait la somme des carrés des coefficients de corrélation pour la composante 1, on trouve 5,042, c'est-à-dire la valeur propre de la composante 1.

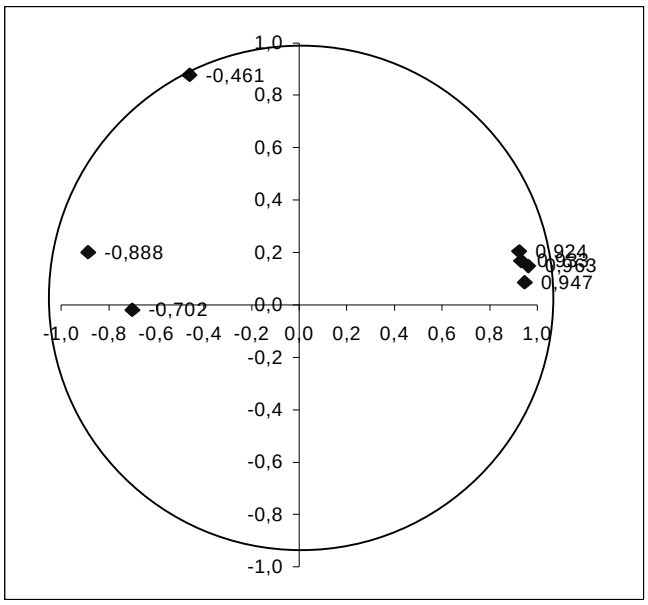
Exemple 05 : cercle des corrélations sous excel

Ces valeurs de corrélations peuvent servir de coordonnées des variables sur les axes factoriels. Mais Clementine ne permet pas de produire un cercle des corrélations.

Version sous Excel du cercle des corrélations :

	Conso	Poids	Cylindres	cm3	hp	time-to-60	Année
CP1	-0,888	0,924	0,933	0,963	0,947	-0,702	-0,461
CP2	0,201	0,204	0,167	0,149	0,086	-0,020	0,877
CP3	-0,201	0,250	0,123	0,099	-0,145	0,703	-0,023
CP4	0,347	-0,035	0,189	0,132	0,020	0,060	-0,131
CP5	0,088	0,100	-0,202	-0,014	0,244	0,077	-0,025

Tableau des coefficients de corrélations



Cercle des corrélations pour CP1 (72%) et CP2 (13%)

On constate la corrélation entre :

- Poids, Cylindres, cm3 et hp
- Conso et Time-to-60

On constate que l'année est à part.

On constate que la CP 1 concerne :

- Poids, Cylindres, cm3, hp, Conso et Time-to-60

et ne concerne pas :

- Time-to-60

On constate que la CP 2 concerne uniquement :

- Année

On peut continuer l'analyse avec les autres composantes conservées. D'après le critère de Kaiser, on en gardait que 2. D'après le critères des pourcentage, on pouvait en garder 3.

En regardant le tableau des coefficients de corrélations, on voit que seule la variable time-to-60 a un fort coefficient de corrélation pour la CP 3 ce qui ne nous apporte aucune information supplémentaires.

Les rotations de l'ACP

La force de l'ACP, c'est de restituer le maximum de variance (d'information) dans une seule nouvelle variable (la première composante principale).

La faiblesse de l'ACP, c'est qu'elle ne permet pas de repérer des groupes de variables car les variables ont tendance à être toutes orientées dans la direction du premier axe (la première composante).

La solution technique consiste à faire pivoter les axes de l'ACP : c'est une rotation.

Il existe deux grands types de rotation :

- **Les rotations orthogonales :** elles préservent la non-corrélation des composantes principales, ce qui facilite l'interprétation. Les principales ACP orthogonales sont les ACP varimax, quartimax et equamax. **L'ACP varimax est la plus répandue des variantes d'ACP**. L'ACP equamax est un compromis entre la varimax et la quartimax.
- **Les rotations obliques :** elles ne préservent pas la corrélation des composantes principales, ce qui rend l'interprétation difficile. Les principales ACP obliques sont les ACP **oblimin** et **promax**. L'ACP promax est plus rapide et adaptée au gros volume de données.

Le **principe de la rotation varimax** c'est, pour chaque composante principale, de maximiser la variance des coefficients de corrélation de la composante avec l'ensemble des variables, et non plus la somme des carrés de ces coefficients de corrélation.

Ainsi chaque composante principale est fortement corrélée à quelques variables et faiblement corrélée aux autres.

Conclusion sur les usages

Les composantes principales s'utilisent principalement pour déterminer les variables qui vont ensemble.

On peut aussi les utiliser pour faire de la classification. Toutefois, les autres méthodes de classification proposées par le data-mining sont toujours plus pertinentes.

Présentation de l'analyse factorielle des correspondances : AFC et ACM

Principes

Les principes sont les mêmes que pour l'ACP.

- L'ACP s'appliquait à n variables continues numériques. L'**AFC** s'applique à **2 variables catégorielles.**
- Une AFC s'appliquant à **plus de 2 variables** est appelée : **ACM** : analyse des correspondances multiples.
- L'AFC et l'ACM peuvent aussi s'appliquer à des variables numériques : il suffit pour cela qu'elles soient discrétisées.

Technique

Le test du Khi-2

Le test du Khi-2 est utilisé pour comparer une distribution expérimentale à une distribution théorique (loi normale, loi binomiale, loi de Poisson, etc.).

Il permet de détecter une dépendance entre deux variables.

L'AFC peut être vue comme une ACP avec une métrique particulière : la métrique du Khi-2.

Le tableau de contingence

Le tableau de contingence de deux variables catégorielles donne le nombre d'individus pour chaque couple constitué à partir des catégories des deux variables.

Il correspond à la commande SQL suivante :

```
select V1, V2, count(*) from table
group by V1, V2.
```

Le tableau de contingence est représenté sous la forme d'une matrice avec les catégories de V1 en ligne et les catégories de V2 en colonne.

Dans chaque case de la matrice, on donne **l'effectif pour le couple de catégories.**

Les effectifs et la contribution au Khi-2 de chaque cellule montrent les liaisons entre les catégories des deux variables. Un sureffectif traduit une forte liaison positive, un sous-effectif traduit une forte liaison négative, un équilibre traduit une liaison faible.

La lecture du tableau de contingence éclaire bien les rapports entre les variables, mais sa lecture peut être fastidieuse s'il y a beaucoup de catégories. S'il y avait plus de deux variables à croiser, la lecture serait encore plus difficile.

L'AFC offre une visualisation en 2 dimensions des tableaux de contingence.

Le tableau de contingence multiple : tableau de Burt

Avec plusieurs variables catégorielles (ACM), on ne peut plus se contenter d'un tableau de contingence.

Ce tableau de contingence simple est remplacé par un tableau de contingence multiple : le tableau de Burt.

Le tableau de Burt est une matrice carrée symétrique. Les lignes et les colonnes correspondent à toutes les catégories des N variables.

La diagonale du tableau donne le nombre d'individus pour cette catégorie.

Le croisement de deux catégories différentes donne le nombre d'individus ayant à la fois la catégorie de la ligne et celle de la colonne : c'est la même chose que le tableau de contingence simple (à noter que si les catégories croisées appartiennent à la même variable, l'effectif vaut 0).

Le tableau de Burt est donc constitué N^2 tableaux : N matrices carrées (la « diagonale ») et N^2-N tableaux de contingence simple. Les N matrices carrés de la diagonales sont remplies de 0 à l'exception de la diagonale qui donne l'effectif pour la catégorie.

Catégorie	US	Europe	Japon	3 cyl.	4 cyl.	5 cyl.	6 cyl.	8 cyl.
US	161	0	0	0	37	0	48	72
Europe	0	47	0	0	42	3	2	0
Japon	0	0	51	2	44	0	5	0
3 cyl.	0	0	2	2	0	0	0	0
4 cyl.	37	42	44	0	123	0	0	0
5 cyl	0	3	0	0	0	3	0	0
6 cyl	48	2	5	0	0	0	55	0
8 cyl.	72	0	0	0	0	0	0	72

Exemple de tableau de contingence

Tableau de Burt à 2 variables

Le tableau disjonctif complet

Le tableau disjonctif complet a une ligne par individu et une colonne par catégorie, en prenant en compte toutes les catégories des N variables.

C'est une matrice creuse : elle ne contient que des 1 et des 0. 1 si l'individu « tombe sous » la catégorie de la colonne, 0 sinon.

Résultats

Une AFC fournit les mêmes axes factoriels, qu'elle soit calculée sur le tableau de contingence ou sur le tableau disjonctif complet.

Combien d'axes factoriels faut-il garder ?

On retrouve les 3 méthodes employées en ACP, avec une variante pour le critère de Kaiser :

- ***Variante du critère de Kaiser***

On ne garde que les axes factoriels dont la valeur propre est supérieure à 1 / nombre de variables.

$$vp > 1 / nbvar$$

Ce n'est pas un critère absolu.

Qualité des résultats obtenus

Les résultats obtenus atteignent un maximum de fiabilité lorsque :

1) les variables ont toutes à peu près le même nombre de catégories ;
2) les effectifs de ces catégories ne sont jamais trop faibles.

Usages de l'AFC et de l'ACM

Plan factoriel et cercle des corrélations

On peut reprendre les usages de l'ACP avec les AFC.

L'offre Clementine : exemple 06

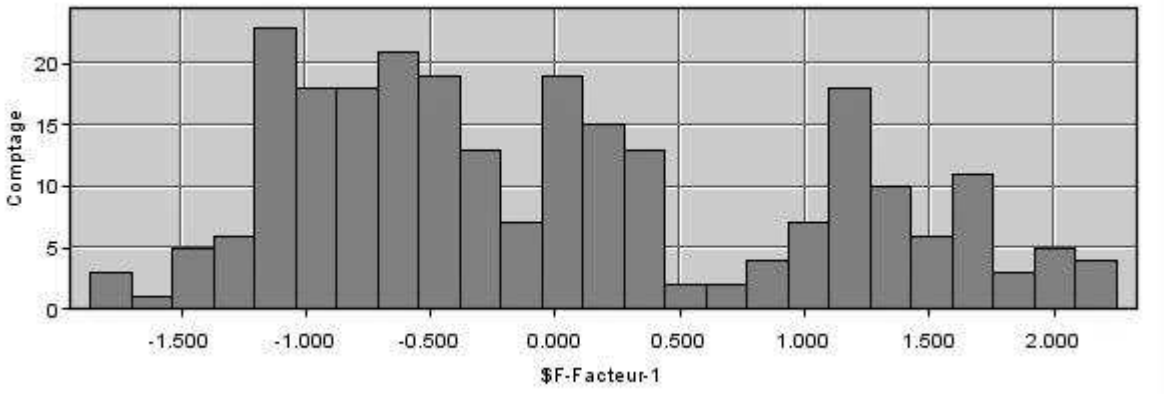
Le nœud modélisation / ACP-Facteur pemet de faire des analyses factorielles.

On reprend le fichier des voitures.

L'analyse en composantes principales : ACP

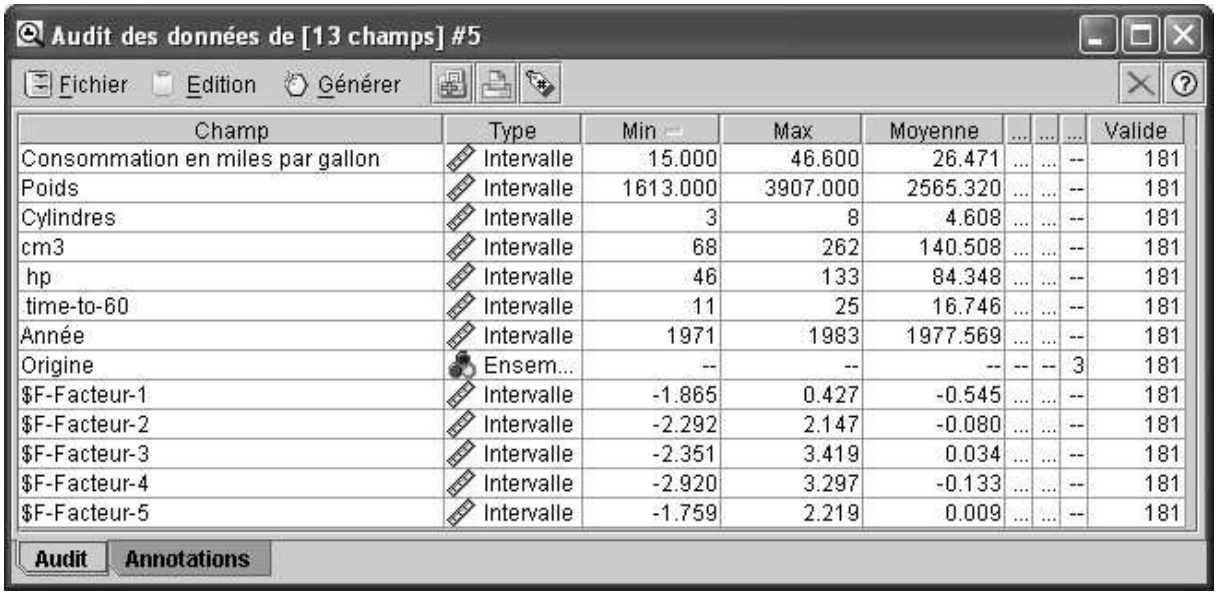
Il s'agit de la méthode par défaut. Par défaut, Clementine utilise une matrice des corrélations, ce qui assure une normalisation des valeurs des variables. En choisissant une matrice des covariances, l'analyse n'aboutira que si les données sont homogènes.

La méthode produit N nouvelles variables.



L'histogramme du facteur 1 montre deux courbes de Gauss, c'est-à-dire deux populations distinctes : la population de la droite de l'histogramme (>0,44), et la population de la gauche de l'histogramme (<=0,44).

Les caractéristiques de ces deux populations sont les suivantes :



Audit des données de [13 champs] #5

Fichier Edition Générer

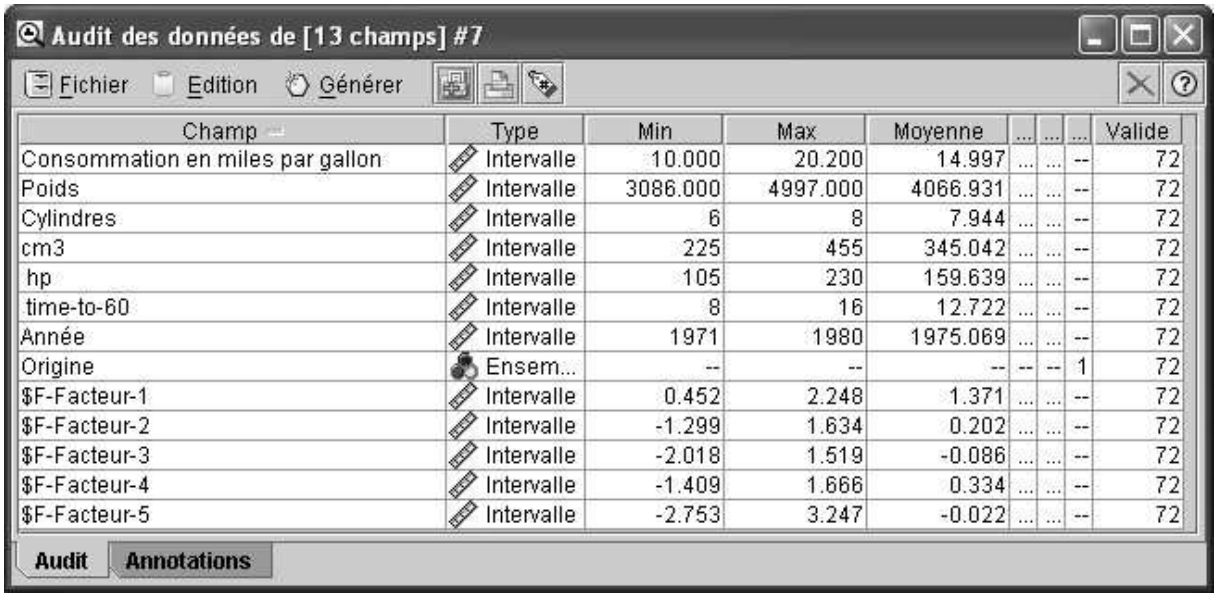
Champ	Type	Min	Max	Moyenne	...	...	...	Valide
Consommation en miles par gallon	Intervalle	15.000	46.600	26.471	...	...	--	181
Poids	Intervalle	1613.000	3907.000	2565.320	...	...	--	181
Cylindres	Intervalle	3	8	4.608	...	...	--	181
cm3	Intervalle	68	262	140.508	...	...	--	181
hp	Intervalle	46	133	84.348	...	...	--	181
time-to-60	Intervalle	11	25	16.746	...	...	--	181
Année	Intervalle	1971	1983	1977.569	...	...	--	181
Origine	Ensem...	--	--	--	--	--	3	181
$F-Facteur-1	Intervalle	-1.865	0.427	-0.545	...	...	--	181
$F-Facteur-2	Intervalle	-2.292	2.147	-0.080	...	...	--	181
$F-Facteur-3	Intervalle	-2.351	3.419	0.034	...	...	--	181
$F-Facteur-4	Intervalle	-2.920	3.297	-0.133	...	...	--	181
$F-Facteur-5	Intervalle	-1.759	2.219	0.009	...	...	--	181

Audit Annotations

Population de la droite de l'histogramme

Cette première population concerne le plus grand nombre d'individus.

L'origine reste répartie sur les trois valeurs possibles, mais avec une sur-représentation américaine (ce qui n'apparaît pas dans la figure ci-dessus).



Audit des données de [13 champs] #7

Fichier Edition Générer

Champ	Type	Min	Max	Moyenne	...	...	...	Valide
Consommation en miles par gallon	Intervalle	10.000	20.200	14.997	...	...	--	72
Poids	Intervalle	3086.000	4997.000	4066.931	...	...	--	72
Cylindres	Intervalle	6	8	7.944	...	...	--	72
cm3	Intervalle	225	455	345.042	...	...	--	72
hp	Intervalle	105	230	159.639	...	...	--	72
time-to-60	Intervalle	8	16	12.722	...	...	--	72
Année	Intervalle	1971	1980	1975.069	...	...	--	72
Origine	Ensem...	--	--	--	--	--	1	72
$F-Facteur-1	Intervalle	0.452	2.248	1.371	...	...	--	72
$F-Facteur-2	Intervalle	-1.299	1.634	0.202	...	...	--	72
$F-Facteur-3	Intervalle	-2.018	1.519	-0.086	...	...	--	72
$F-Facteur-4	Intervalle	-1.409	1.666	0.334	...	...	--	72
$F-Facteur-5	Intervalle	-2.753	3.247	-0.022	...	...	--	72

Audit Annotations

Population de la gauche de l'histogramme

Cette deuxième population est moins nombreuse que la précédente.

Elle ne concerne que les voitures américaines (ce qui n'apparaît pas dans la figure ci-dessus).

Elle concerne les voitures de forte cylindré (6 à 8), lourdes, de forte puissance, de fort volume, qui consomme beaucoup et peu « nerveuse » (time to 60 faible) Autrement dit, l'analyse a mis au jour les « grosses voitures familiales américaines ».

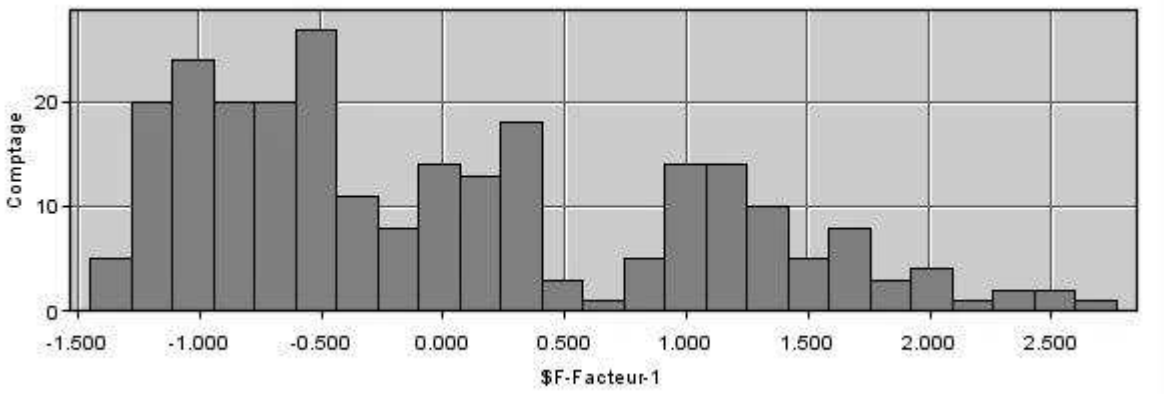
Moindres carrés non pondérés : ACM

L'analyse factorielle des moindres carrés recherche tous les facteurs pouvant reproduire le modèle des corrélations entre les champs d'entrée.

Pour fonctionner, **la matrice factorielle doit être triée.**

On peut aussi préciser le nombre de facteurs souhaités : toutefois, ce nombre est inférieur à celui qu'on peut atteindre avec une ACP. En général, on se contente d'un seul facteur.

Elle peut aussi subir des rotations dont l'usage est assez empirique.



Dans l'exemple des voitures, cette analyse donne à peu près les mêmes résultats que la précédente. Il y a deux populations. En comparant les individus, on constate que ce sont bien les mêmes dans les deux analyses.

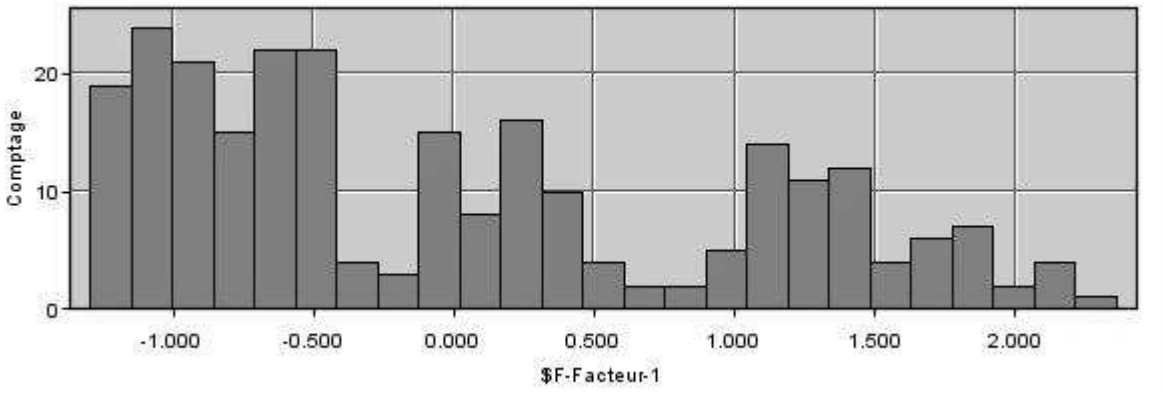
Moindres carrés généralisés : ACM

Similaire à la méthode précédente, cette méthode utilise une technique de pondération afin de donner moins de valeur aux champs présentant une variance unique (non partagée).

Pour fonctionner, **la matrice factorielle doit être triée.**

On peut aussi préciser le nombre de facteurs souhaités : toutefois, ce nombre est inférieur à celui qu'on peut atteindre avec une ACP. En général, on se contente d'un seul facteur.

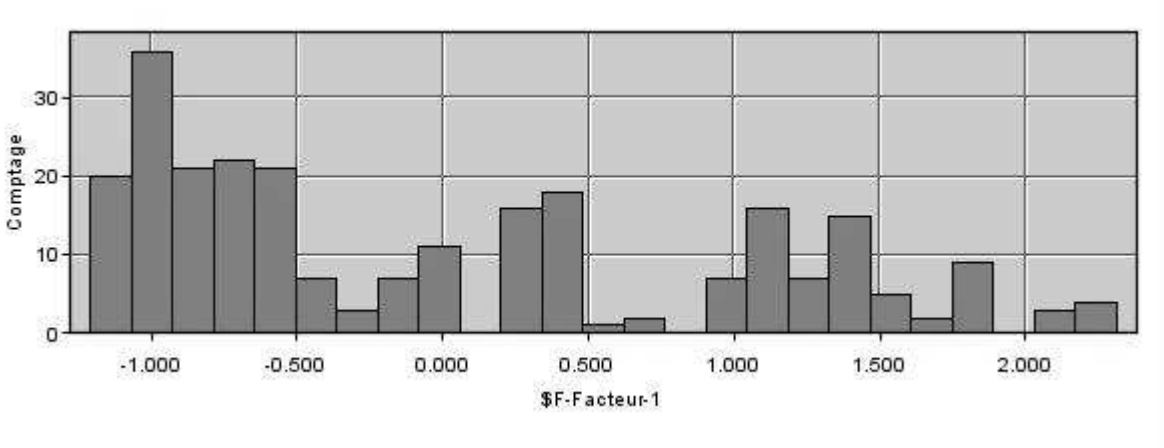
Elle peut aussi subir des rotations dont l'usage est assez empirique.



Ici, il semble qu'on mettre au jour 3 sous-ensembles.

Maximum de vraisemblance : ACM

Cette méthode, du même type que les deux précédentes, génère les équations factorielles qui sont les plus susceptibles d'avoir produit le modèle des corrélations observé dans les variables d'entrée.



Dans l'exemple des voitures, l'histogramme du facteur 1 montre non plus deux mais trois deux populations distinctes : la population de la droite de l'histogramme (>0,8), la population du centre de l'histogramme (>0,2 et <=0,8) et la population de la gauche de l'histogramme (<=0,2).

Les caractéristiques de ces populations sont les suivantes :

Audit des données de [9 champs] #12

Fichier Edition Générer

Champ	Type	Min	Max	Moyenne	...	...	...	Valide
Consommation en miles par gallon	Intervalle	10.000	20.200	14.809	...	...	--	68
Poids	Intervalle	3086.000	4997.000	4102.426	...	...	--	68
Cylindres	Intervalle	8	8	8.000	...	...	--	68
cm3	Intervalle	302	455	350.882	...	...	--	68
hp	Intervalle	129	230	161.603	...	...	--	68
time-to-60	Intervalle	8	16	12.588	...	...	--	68
Année	Intervalle	1971	1980	1974.926	...	...	--	68
Origine	Ensem...	--	--	--	--	--	1	68
$F-Facteur-1	Intervalle	0.959	2.310	1.421	...	...	--	68

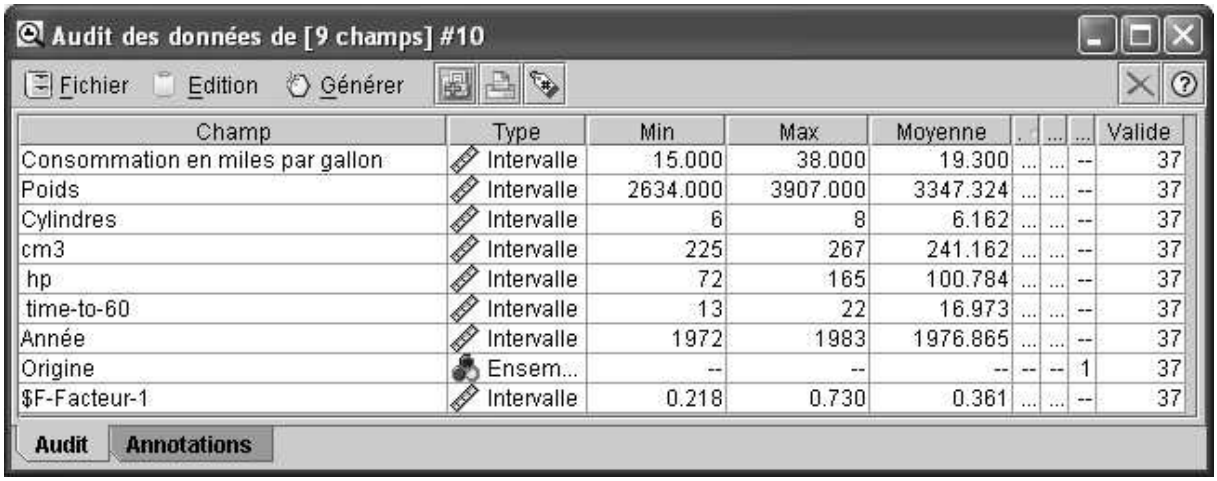
Audit Annotations

Population de la droite de l'histogramme

Elle ne concerne que les voitures américaines (ce qui n'apparaît pas dans la figure ci-dessus).

Cette population est un sous-ensemble de la population des « grosses voitures familiales américaines ». C'est le sous-ensemble des plus grosses voitures : les min de cylindres, cm3 et hp sont plus élevés. On passe de 72 individus à 68

.

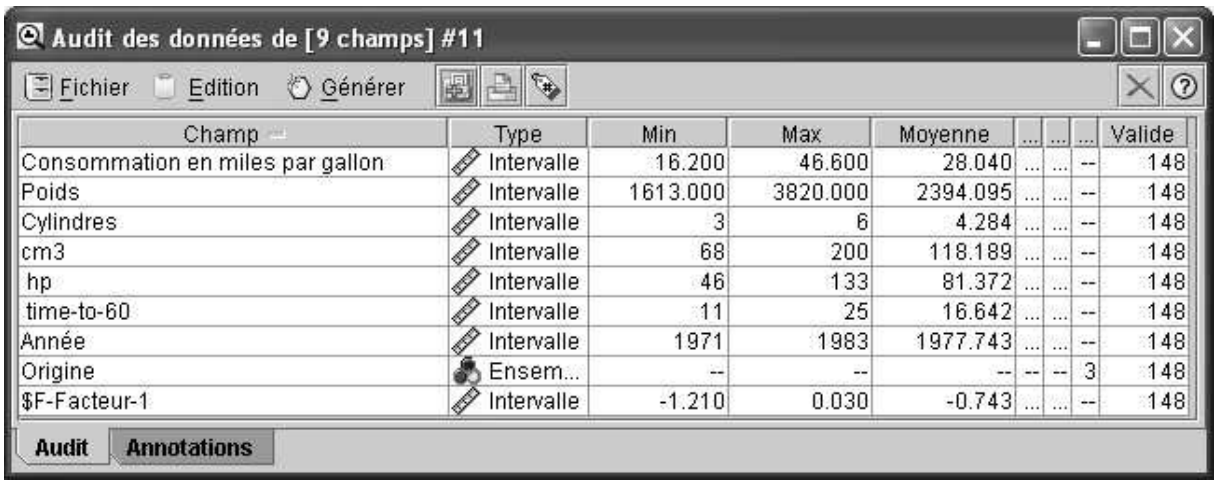


Population du centre de l'histogramme

C'est une nouvelle population.

Elle ne concerne que les voitures américaines (ce qui n'apparaît pas dans la figure ci-dessus).

Ce sont des voitures de puissance moyenne.



Population de la gauche de l'histogramme

Cette population est un sous-ensemble de la population mise au jour par l'ACP (non-« grosses voitures familiales américaines »).

Ces voitures sont moins puissantes que dans la population ACP, et surtout, la répartition géographique est équilibrée : il n'y a plus de sur-représentation américaine.

Factorisation en axes principaux

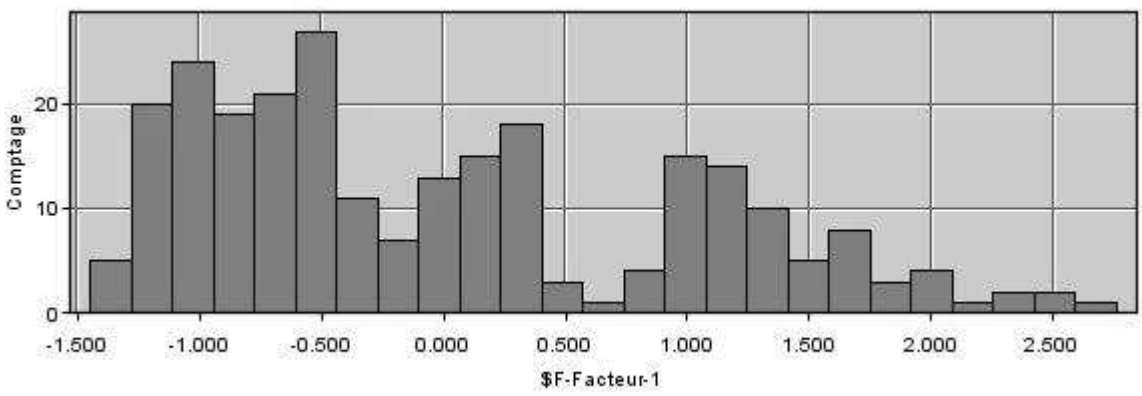
Cette méthode d'analyse factorielle est très proche de l'ACP : elle donne aussi des résultats proches de l'ACP. Elle se concentre exclusivement sur la variance partagée.

Le paramétrage est le même que pour les méthodes précédentes :

La matrice factorielle doit être triée.

On peut préciser le nombre de facteurs souhaités : toutefois, ce nombre est inférieur à celui qu'on peut atteindre avec une ACP. En général, on se contente d'un seul facteur.

Elle peut aussi subir des rotations dont l'usage est assez empirique.

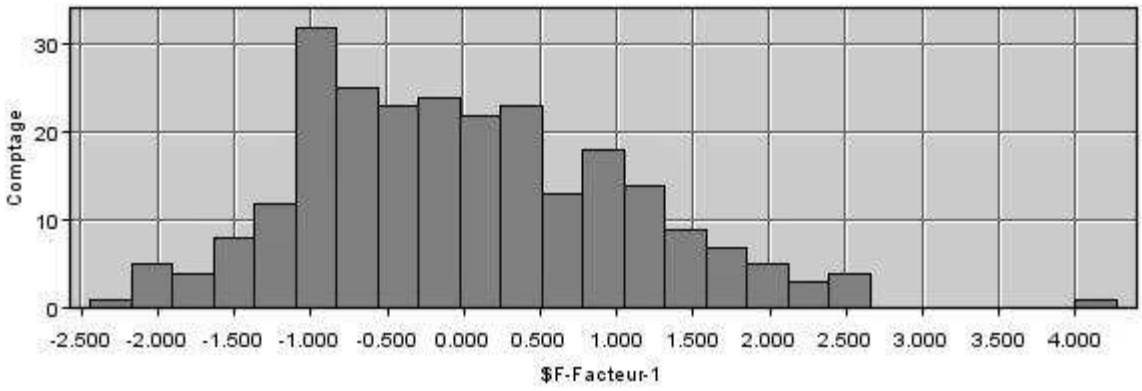


On retrouve 2 populations (ou 3).

Alpha-maximisation

Cette méthode d'analyse factorielle considère les champs à analyser comme un échantillon des champs d'entrée potentiels. Elle maximise la fiabilité statistique des facteurs.

Le paramétrage est le même que pour les méthodes précédentes.



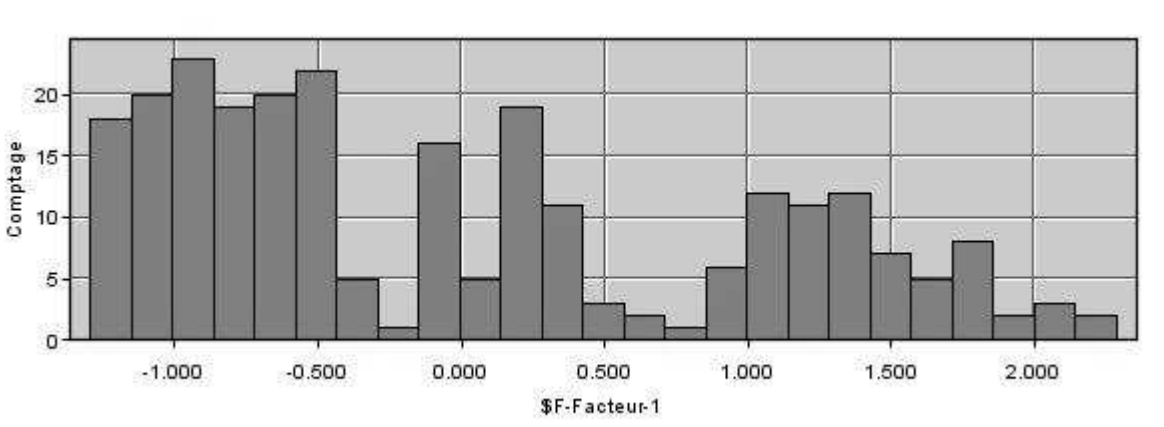
Ici, la généralisation fait perdre toute classification.

Factorisation en projections

Cette méthode d'analyse factorielle utilise la technique d'estimation des données afin d'isoler la variance commune, ainsi que ses facteurs descriptifs.

On peut laisser le paramétrage par défaut.

La méthode produit N nouvelles variables.



On retrouve 3 populations.