Les méthodes statistiques de régression et de classification sont nombreuses. Nous abordons ici les méthodes linéaires qui appartiennent dans à une même catégorie de méthodes.

De nombreux logiciels sont accessibles pour effectuer ces calculs, le plus populaire étant aujourd'hui R. Nous utilisons ici Excel et le calcul matriciel, afin de mieux comprendre les détails de ces méthodes.

A partir de trois fonctions matricielles : le produit, l'inversion et le calcul des valeurs et vecteurs propres, nous pouvons effectuer tous ces calculs.

Nous passons en revue la régression linéaire simple et multiple, et l'analyse en composante principale (ACP). L'analyse des correspondances (AF) et l'Analyse discriminante linéaire (ADL) seront dans le prochain article. Le lecteur peut s'aider des 2 exemples numériques afin de suivre le déroulement des calculs théoriques.

1–Régression simple.

La régression linéaire simple consiste à identifier la droite qui 'linéarise' au plus près, n points (r,s) dans un plan d'abscisse X et d'ordonnée Y (nous gardons la notation (x,y) pour les variables centrées).

On démontre que cette droite passe par le centre de gravité (CDG) du nuage de points.

Notations :

- $\sum_{i=1}^{n} r_i = \bar{R}$
- $\sum_{i=1}^{n} s_i = \bar{S}$

Le centre de gravité devient : $(\frac{\bar{R}}{n}, \frac{\bar{S}}{n})$.

Il est plus facile de travailler avec des données centrées, car cela simplifie les formules tout en gardant une même interprétation. En effet, lorsque la moyenne est nulle, la variance est égale à la somme des carrés. Le centrage consiste en une translation des données du point origine au CDG.

Nouvelles notations :

- $x_i = r_i - \bar{R}$
- $y_i = s_i - \bar{S}$
- $\sum_{i=1}^{n}(x_i)^2 = \overline{XX}$
- $\sum_{i=1}^{n}(y_i)^2 = \overline{YY}$
- $\sum_{i=1}^{n} x_i \,.\, y_i = \overline{XY}$

Il ne reste alors à définir que son coefficient directeur qui est égal au rapport de la covariance par la variance:

$$a = \frac{Cov(X,Y)}{Var(X,X)} = \frac{\sum_{i=1}^{n} x.y}{\sum_{i=1}^{n} x.x} = \frac{\overline{XY}}{\overline{XX}} \qquad (1.1)$$

L'équation de la droite devient :

$$y = a.x \qquad (1.2)$$

La formule précédente minimise les écarts sur les y, donc selon l'axe des ordonnées. Les plus courtes distances sont des droites verticales. Il est alors possible d'inverser les x et y, et d'obtenir une autre régression, selon des distances prises horizontalement. Le coefficient directeur de la droite de régression devient a' :

$$a' = \frac{Cov(X,Y)}{Var(Y,Y)} = \frac{\sum_{i=1}^{n} x.y}{\sum_{i=1}^{n} y.y} = \frac{\overline{XY}}{\overline{YY}} \qquad (1.3)$$

Il existe une troisième voie : prendre la plus courte distance entre les points (x, y) et la droite de régression. Dans ce cas, on projette perpendiculairement les points sur la droite de régression. La solution devient géométrique : cette projection orthogonale est à la base de toutes les méthodes factorielles que nous allons voir ensuite.

Le calcul s'effectue en deux phases :

- Projection orthogonale des points P(x,y) sur une droite (D) passant par l'origine, pour obtenir le point H,
- Rotation de cette droite afin d'obtenir une projection de variance maximale.

La projection orthogonale fait appel au nuage de points (x,y), et à un vecteur directeur v de la droite (D), qu'on prendra égal à v(α,β), avec la condition $\alpha^2+\beta^2=1$, afin d'avoir un vecteur unitaire (et de simplifier les formules).

Au lieu de calculer les nouvelles coordonnées des points projetés sur (D), nous calculons la coordonnée du point projeté sur cette droite, à savoir, la distance du point projeté à l'origine : au lieu d'avoir un couple (x,y), nous avons une distance di.

La distance di est égale au produit scalaire entre v et OP :

$$d_i = \alpha.x + \beta.y \qquad (1.4)$$

On obtient alors : $OH=d_i.V$, qui permet d'obtenir les coordonnées du point projeté H, dans le plan X,Y.

La variance des points projetés, est obtenue en sommant toutes ces distances d_i au carré:

$$V = \sum_{i=1}^{n} {d_i}^2 = \sum (\alpha.x_i + \beta y_i)^2$$

Un développement rapide nous donne :

$$n.V = \alpha^2.\overline{XX} + 2.\alpha.\beta.\overline{XY} + \beta^2.\overline{YY}$$

Ou sous forme matricielle :

$$V = {}^T v * COV * v \qquad (1.5)$$

Avec, COV la matrice de covariance :

$$COV = \frac{1}{n}.\begin{bmatrix}\overline{XX} & \overline{XY}\\ \overline{XY} & \overline{YY}\end{bmatrix} \qquad (1.6)$$

La minimisation de cette variance fait appel à la résolution d'une optimisation sous contrainte. La variance maximale est égale à la valeur propre maximale de la matrice de covariance COV ; le premier vecteur propre étant le vecteur directeur de la droite de régression : v.

2–Régression multiple.

La régression multi linéaire consiste à expliquer un vecteur Y à partir de q vecteurs X. L'échantillon statistique comportant n données (n lignes ou n individus), Y sera un vecteur (nx1) et X une matrice (nxq).

Dans la pratique, l'utilisateur a le choix d'introduire dans ses variables explicatives une constante ; dans ce cas, la première colonne de X est composée de la valeur constante 1. Les colonnes de X, vont générer un hyperplan (P), dont tous les points seront combinaison linéaire des X : si X possède 2 colonnes, donc 2 variables explicatives, l'hyperplan devient un plan.

L'objectif de la régression est d'obtenir un vecteur Z de (P) (souvent noté $\hat{Y}$, appelé Y chapeau), le plus proche possible de Y de départ, au sens des moindres carrés. Les écarts sont ici mesurés par la distante au carré (distance euclidienne), toujours selon la direction des y, et non pas en projection orthogonale, qui sera résolu par l'ACP.

La solution s'obtient par la formule matricielle suivante :

$$Z = X * (X' * X)^{-1} * X' * Y \qquad (2.1)$$

Soit sous Excel :

```
PRODUITMAT(PRODUITMAT(PRODUITMAT(MX ;MINVERSE(PRODUITMAT(TRANSPOSE(MX) ;MX)));TRANSPOSE(MX));MY)
```

Cette formule s'interprète de deux façons différentes :

- Projection de Y qu'on cherche à expliquer, sur l'espace (P). On applique alors la transformation Px, application linéaire, de matrice Px, au vecteur Y. La matrice Px, est de taille nxn, et ne dépend que de X :

$$Px = X * (X' * X)^{-1} * X' \qquad (2.2)$$

$$Z = Px * Y \qquad (2.3)$$

- Recherche du point Z, comme combinaison linéaire des X. Ainsi Z appartient bien à (P). Le vecteur β et constant, et servira ensuite à effectuer des prévisions à partir de tout nouveau vecteur X :

$$\beta = (X' * X)^{-1} * X' * Y \qquad (2.4)$$

$$Z = X * \beta \quad (2.5)$$

La seconde approche est plus facile à mettre en œuvre, car au lieu d'avoir nxn coefficients pour Px, nous n'avons que n coefficient β.
La formule (2.1) peut sembler complexe, mais nous retrouvons le cas de la régression simple, en identifiant le numérateur XY' et le dénominateur XX' du coefficient de régression a dans (1.1).
Il est aussi toujours possible de changer la variable à expliquer Y, par une des colonnes de X, ce qui revient à changer les directions de projection comme en régression simple.

Une fois Z obtenu, l'efficacité de la régression passe par le calcul des écarts : $(Y\text{-}Z)^2$, qui sera traité comme une variance, car les écarts ont aussi une moyenne nulle (ceci n'est pas du à notre centrage initial, mais est vrai dans tous les cas).
Notations :

- $e_{i=}(y_i - z_i), les\ écarts$
- $\sum_{i=1}^{n} {y_i}^2 = \overline{YY} = SCT$
- $\sum_{i=1}^{n} e^2 = \overline{EE} = SCR$
- $\sum_{i=1}^{n} {z_i}^2 = \overline{ZZ} = SCE$
- $R^2 = SCE/SCT$

Avec SCT, variance totale, SCE, variance estimée, SCR variance résiduelle.
L'objectif est d'obtenir le plus grand R^2.

3–Analyse en Composantes Principales (ACP).

L'ACP effectue une régression multi dimensionnelle avec des projections orthogonales. Dans un premier temps, l'objectif n'est plus la prédiction, mais l'analyse directe de la géométrie du nuage de points, défini dans la table (D), avec n lignes et q colonnes.
On cherche les axes principaux qui minimisent les variances des projections orthogonales.

Dans le cas d'une matrice (X) centrée, obtenues à partir de (D), nous rappelons que la matrice de covariance (COV) est égale au produit matriciel :

$$n.COV = {}^T X * X \quad (3.1)$$

Et la variance des projections sur le vecteur v(α,β) est toujours donné par (1.5) :

$$V = {}^T v * COV * v = {}^T (X * v) * (X * v) \quad (3.2)$$

Le problème à résoudre est de faire varier v, de façon à obtenir une variance V maximum.
Si la matrice X est non singulière (de déterminant non nul), la solution est :

- Les axes sont les vecteurs propres de la matrice de covariance (COV),
- Les variances sur chaque axe, sont les valeurs propres de la matrice de covariance (COV).

Dans ce nouveau repère :

- la matrice (COV) devient diagonale,
- la diagonale est composée des valeurs propres, homogènes à des variances,
- les corrélations entre les axes sont nulles, car les axes sont orthogonaux.

Le calcul des valeurs propres et des vecteurs propres, utilisent deux fonctions développées sous VBA utilisant l'algorithme d'Hermite qui ne s'applique qu'aux matrices symétriques. C'est notre cas, car les matrice de covariance sont toujours symétriques.

- Fonction valeurs propres MVALP(V), pour obtenir un vecteur ligne (MVALP)
- Fonction valeurs propres MVECP(V), pour obtenir une matrice (MVECP)

Les vecteurs propres permettent d'obtenir les coordonnées des projections sur les nouveaux axes. Ces coordonnées sont différentes des coordonnées des projections dans l'ancien repère.
Les projections (PJ) sont obtenues par le produit matriciel :

PJ = MCENTRE x MVECP

La somme de chaque colonne de (PJ) est nulle, car les nouvelles variables sont aussi centrées.
La somme des carrés de chaque colonne de (PJ) est égale à la variance, donc à la valeur propre correspondant à la colonne.

Les **contributions** (CTR) de chaque colonne, cherche à identifier les individus participant le plus à la nouvelle variable (en colonne). Comme la somme des carrés est égale à n fois la valeur propre, on divise le carré de (PJ) par n fois MVALP, afin d'obtenir un pourcentage dans la colonne, et donc la contribution.

Idem pour les lignes, appelé **Cosinus Carrés** (COS2). On divise chaque ligne par la somme des carrés de la ligne de (PJ). On remarque que cette somme est identique à celle obtenu en prenant les carrés de la matrice centrée d'origine.

Enfin, on calcule les **corrélations** (CO) entre les nouveaux et les anciens axes. La somme des carrés des colonnes est la valeur propre de la colonne, la somme des carrés des lignes est la variance initiale, et la somme totale des carrés est la variance totale.
L'ACP peut aussi s'effectuer sur une matrice de Corrélation, ou même sur les données brutes (D).

4-Exemple de régression

Exemple numérique de régression avec n=6 lignes :

X	y	xx	yy	xy
-4,50	-7,50	20,25	56,25	33,75
-7,50	22,50	56,25	506,25	-168,75
-4,50	0,50	20,25	0,25	-2,25
4,50	3,50	20,25	12,25	15,75
3,50	-11,50	12,25	132,25	-40,25
8,50	-7,50	72,25	56,25	-63,75
0,00	**0,00**	**201,50**	**763,50**	**-225,50**

Tab1- données et sommes.

Ces données sont déjà centrées. Afin de privilégier l'interprétation des résultats, nous évitons de diviser par n=6.
Les premiers calculs portent sur les variances, comme somme de carrés et la covariance, à partir des produits x.y :
XX=201,5
YY=763,5
XY=-225,50
a=225,5/201,5=-1,1191
a'=225,5/763,5=-0,2953
En projetant selon y, les y deviennent : - $\hat{y}$=1,1191 . x :

y	$\hat{y}$	$\hat{y}\hat{y}$	e=y-$\hat{y}$	ee
-7,50	5,04	25,36	-12,54	157,15
22,50	8,39	70,45	14,11	199,00
0,50	5,04	25,36	-4,54	20,58
3,50	-5,04	25,36	8,54	72,86
-11,50	-3,92	15,34	-7,58	57,50
-7,50	-9,51	90,49	2,01	4,05
0,00	**0,00**	**252,36**	**0,00**	**511,14**

Tab2- estimations et écarts.

SCT	SCE	SCR	R2
763,5000	252,3586	511,1414	0,33053

Tab3- variances et R2.

On vérifie bien la relation :

SCT = SCE + SCR

5-Exemple d'ACP

Soit les données quantitatives suivantes, avec n=10 individus, et q=4 variables explicatives.

poids	taille	age	note
45	1,50	13	14
50	1,60	13	16
50	1,65	13	15
60	1,75	15	9
60	1,70	14	10
60	1,70	14	7
70	1,60	14	8
65	1,60	13	13
60	1,55	15	17
65	1,70	14	11

Tab4- données pour ACP : D

A partir des données nous obtenons la matrice de Covariance suivante: COV.

```
COV=PRODUITMAT(TRANSPOSE(D-MOY) ;(D-MOY)) /10
```

Avec MOY, le vecteur moyenne :

58,5000	1,6350	13,8000	12,0000

Tab5- moyennes.

55,2500	0,2025	2,7000	-14,0000
0,2025	0,0055	0,0220	-0,1550
2,7000	0,0220	0,5600	-0,8000
-14,0000	-0,1550	-0,8000	11,0000

Tab6- matrice de covariance.

La variance totale est égale à la somme des éléments diagonaux ci-dessus : 66,8155.

59,44298	6,943924	0,425541	0,003079

0,959367	0,278523	-0,045148	0,000975
0,004011	-0,013333	0,024583	0,999601
0,04777	-0,002592	0,998548	-0,024783
-0,278058	0,960333	0,016131	0,013528

Tab7- valeurs propres et vecteurs propres.

On vérifie que la somme des valeurs propres est égale à la variance totale : 66,8155.

PROJECTIONS			
-13,5463	-1,8355	-0,1604	-0,1012
-9,3052	1,4764	-0,3514	0,0307
-9,0270	0,5154	-0,3663	0,0671
2,3310	-2,4679	1,0850	0,0461
2,0050	-1,5043	0,1013	0,0344
2,8392	-4,3853	0,0529	-0,0062
12,1544	-0,6384	-0,3849	-0,0828
5,9195	2,7733	-1,0770	0,0047
0,1057	5,2175	1,2091	-0,0456
6,5238	0,8487	-0,1083	0,0528
0	**0**	**0**	**0**
594,43	**69,44**	**4,26**	**0,03**

Tab8- projections.

La somme des colonnes est nulle et la somme des carrés est égale à la valeur propre (multipliée par n=10).

CONTRIBUTIONS			
0,3087	0,0485	0,0060	0,3328
0,1457	0,0314	0,0290	0,0305
0,1371	0,0038	0,0315	0,1463
0,0091	0,0877	0,2766	0,0690
0,0068	0,0326	0,0024	0,0385
0,0136	0,2769	0,0007	0,0012
0,2485	0,0059	0,0348	0,2228
0,0589	0,1108	0,2726	0,0007
0,0000	0,3920	0,3435	0,0675
0,0716	0,0104	0,0028	0,0906
1,00	**1,00**	**1,00**	**1,00**

Tab9-(CTR): contributions (par colonne).

La somme de chaque colonne est égale à 1.

COSINUS CARRES				
0,9818	0,0180	0,0001	0,0001	**1**
0,9741	0,0245	0,0014	0,0000	**1**
0,9951	0,0032	0,0016	0,0001	**1**
0,4277	0,4794	0,0927	0,0002	**1**
0,6387	0,3595	0,0016	0,0002	**1**
0,2953	0,7046	0,0001	0,0000	**1**
0,9962	0,0027	0,0010	0,0000	**1**
0,7983	0,1752	0,0264	0,0000	**1**
0,0004	0,9486	0,0509	0,0001	**1**
0,9830	0,0166	0,0003	0,0001	**1**

Tab10-(COS2) : cosinus carrés.

La somme de chaque ligne est égale à 1.

CORRELATIONS				
7,3967	0,7339	-0,0295	0,0001	**55,25**
0,0309	-0,0351	0,0160	0,0555	**0,01**
0,3683	-0,0068	0,6514	-0,0014	**0,56**
-2,1438	2,5306	0,0105	0,0008	**11,00**
59,44	**6,94**	**0,43**	**0,00**	**66,82**

Tab11-(CO) : corrélations entre axes anciens et axes nouveaux.

La somme des carrés des colonnes sont les valeurs propres.

La somme des carrés des lignes sont les variances initiales.

La somme totales des carrées, est la variance totale : 66,8155.

6-Exemple financier.

La méthode ACP est très utile pour réduire le nombre de facteurs de risque d'un portefeuille, tout particulièrement en ce qui concerne les taux d'intérêt qui sont fortement corrélés le long de la courbe de taux.

Nous effectuons une ACP sur les taux Euribor officiels (www.euribor-ebf.eu), pris entre 1 semaine et 12 mois sur 8 facteurs de risque, et sur 3 années : 2015 à 2017 complètes.

ACP centrée, avec matrice de covariance suivante :

1w	2w	1m	2m	3m	6m	9m	12m
0,0158	0,0160	0,0176	0,0172	0,0173	0,0167	0,0158	0,0163
0,0160	0,0162	0,0179	0,0175	0,0176	0,0170	0,0161	0,0166
0,0176	0,0179	0,0199	0,0194	0,0195	0,0188	0,0178	0,0184
0,0172	0,0175	0,0194	0,0190	0,0192	0,0185	0,0175	0,0181
0,0173	0,0176	0,0195	0,0192	0,0195	0,0190	0,0180	0,0186
0,0167	0,0170	0,0188	0,0185	0,0190	0,0188	0,0179	0,0187
0,0158	0,0161	0,0178	0,0175	0,0180	0,0179	0,0172	0,0180
0,0163	0,0166	0,0184	0,0181	0,0186	0,0187	0,0180	0,0189

Tab12- Covariance sur Euribor.

0,14268	0,00224	0,00019	4,9E-05	1,4E-05
98,3%	1,5%	0,1%	0,0%	0,0%

Tab13- valeurs propres sur Euribor et %.

La première valeur propre explique 98% du risque, c'est le parallèle shift, la seconde valeur propre n'explique plus que 1,5% - on peut facilement se limiter alors à 3 facteurs de risque.

0,3290	-0,3419	0,4610	-0,4570	-0,3056
0,3342	-0,3361	0,4032	-0,0895	0,2616
0,3705	-0,3408	-0,0687	0,5541	0,3479
0,3631	-0,2356	-0,3547	0,2768	-0,3062
0,3689	-0,0567	-0,5143	-0,2206	-0,2942
0,3607	0,3057	-0,3135	-0,5015	0,3566
0,3432	0,4183	0,1585	0,0829	0,4249
0,3563	0,5717	0,3263	0,3040	-0,4786

Tab14- vecteurs propres Euribor.

La table des projections montre bien la nature des nouveaux axes principaux :

Le premier axe ne possède pas de valeur négative, toutes les valeurs sont très proche, environ $1/\sqrt{8} = 0{,}3535$, car ce sont des vecteurs unitaires.

Le second axe est partagé, entre du positif et du négatif, c'est le risque de rotation.

Les suivants ont à chaque fois un changement de signe supplémentaire – ceci est visible sur le graphe suivant (avec de l'entrainement).

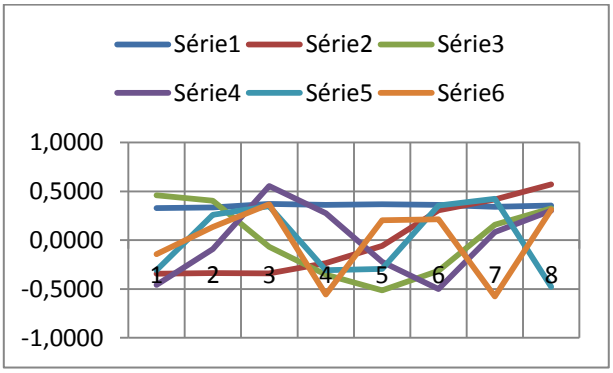

Fig1- Vecteurs propres sur Euribor

7-Conclusion

Les méthodes présentées font largement appel à des calculs algébriques, facilement réalisable avec les formules matricielles d'Excel. Toutes ces méthodes sont très similaires – une bonne connaissance de l'algèbre linéaire est très utile. La suite de cet article abordera les méthodes d'analyse factorielle des correspondances et l'analyse factorielle discriminante.

-/-/-