Contrôle des connaissances "Machine learning"

PIRES - KESADRI- ASLIMI

2 juin 2014

Chapitre 1

Description des données.

1.1 Introduction

Tout d'abord, nous avons utiliser pour ce projet, deux logiciels qui sont R et Rstudio. En effet, les graphiques issues de l'ACP sont difficilement interprétable avec RStudio, tandis que les graphiques d'évolution des cours le sont. Il a donc fallut alterner entre ces deux interfaces. Concernant les packages, nous avons utiliser ici deux packages à savoir FactoMineR et Tseries. FactoMineR permet de réaliser l'analyse des composantes principale et Tseries permet de travailler sur la matrice contenant tout les cours.

1.2 Description des données

Sur les 100 entreprises composant le FTSE (Financial Times Stock Exchange), nous avons choisis de prendre 10 capitalisations parmis les 15 plus importante de cette place financière. Il s'agit ici, de HSBC, Vodafone, British Américan Tobacco, Diageo, Imperial Tobacco, British Petroleum, GSK, AstraZeneca, BS Group et Tesco. On remarque que ces cours concernent des domaines divers et variés comme l'industrie pétrolière, bancaire, agro-alimentaire, pharmaceutique ou encore celle du tabac. L'objectif étant de mesurer l'impact des ces valeurs décisives sur le reste du marché.

Au départ, nous voulions prendre les dix plus grosses capitalisations du FTSE. Cependant, nous avons rencontré des problèmes sur Yahoo Finance quant à l'interprétation des données. En effet, il arrive que certains titres ne soient pas coter en bourse certains jours (Krash, Fusion acquisition etc...), d'où un nombre inégale de valeurs télécharger. Il a donc fallut trouver une solution, nous avons alors pris dix des quinze plus grosses capitalisations du FTSE et nous avons également changer la période d ?étude des cours. Nous avons pris des cours cotant de 2005 à 2009 pour

étudier l'impact de la crise sur ces valeurs. Notre but d'étudier l ?intérêt de l'ACP pour ce domaine d'étude.

Chapitre 2

Description de la méthode

2.1 Les données

1er Etape On importe les cours d'ouvertures (SERIESO) et de fermetures (SERIESC), puis on fait une moyenne de ces cours (SERIESmean). En effet, on ne se contente pas de d'utiliser les cours d'ouverture ou de fermeture car en début de séance le marché est extrêmement volatile alors qu'en fin de séance, tous les acteurs cherchent à se couvrir (Hedge). C'est pourquoi il nous a sembler judicieux de prendre une moyenne de ces deux cours pour qu'elle reflète au mieux la réalité du marché.

2nd Etape - Summary On utilise la fonction Summary, celle-ci nous donne un détail des cours et permet une analyse statistique de base sur chacun des cours. On trace alors, l'ensemble des cours sur un même graphique.

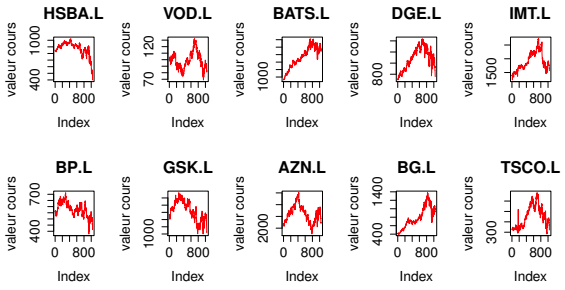

FIGURE 2.1 – 10cours

3ème Etape - Calcul des log rendement On calcul les rendements sur les cours moyens à l'aide de la fonction diff(log(SERIESmean). Puis on utilise à nouveau la fonction Summary et on trace les fonctions associées aux Logrendements.

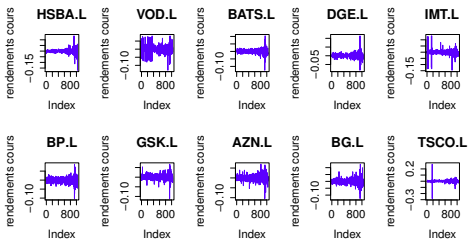

FIGURE 2.2 – 10rendement

2.2 Analyse des Composantes Principales (ACP)

Méthode On utilise la fonction PCA(rSERIES) pour réaliser une ACP. On obtient alors deux graphiques. Le premier graphique Variable Factor Map (ci-dessous) qui permet de voir comment les individus (cours) sont repartit par rapport au deux premières dimensions.

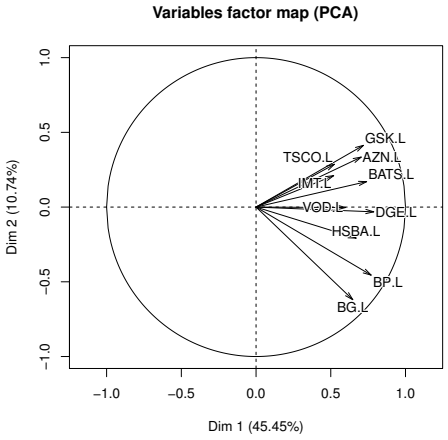

FIGURE 2.3 – camem

Quant au deuxième (ci-dessous), il sert a voir comment les variables (les entreprises) sont repartit par rapport aux deux premières dimensions.

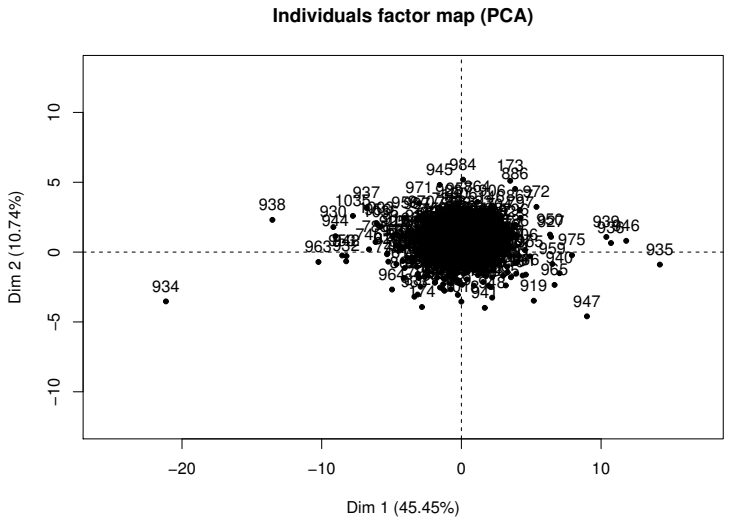

FIGURE 2.4 – nuagePTS

2.3 III-Reconstitution des données

Méthode On reconstitue alors, les cours de bourse à partir de la première composante principal. K correspond, ici, à la contribution de chaque cours à la composante principales.

On se sert alors des rendements obtenues pour tracer les rendements reconstitués.

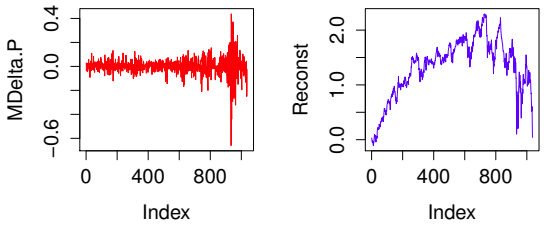

FIGURE 2.5 – coursRdtRecon

2.4 Commentaire sur le code

Vous trouverez, en pièce jointe, le code R associé à ce rapport.

Chapitre 3

La description des résultats

3.1 Interprétation des données

On utilise la fonction Summary pour déterminier la corrélation de la première variable.

Summary :

HSBA.L 0.6698199

3.2 Interprétation ACP

Individual factor Map (PCA) :

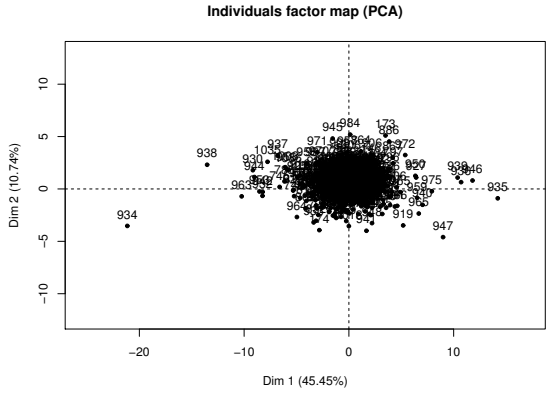

FIGURE 3.1 – nuagePTS

Concernant les points, il faut savoir que ces derniers servent à calculer les valeurs propres, soit le système solution permettant d'annuler chaque équation. On calcul des composantes qui sont censées réduire au maximum X, ce qui explique que la plupart des points soient concentrés autour de 0. Quand un point est éloigner du nuage global, et en fait plus particulièrement loin de l'origine de ton graphe,il est considéré comme atypique, car ses coordonnées ne font pas partie de la masse d'individus considérée comme normal et qui a permis de réduire le système et donc de sortir les valeurs propres. On constate que globalement, il n'y a pas d'individus atypique mis à part 934 et 938.

Variables factor map (PCA) :

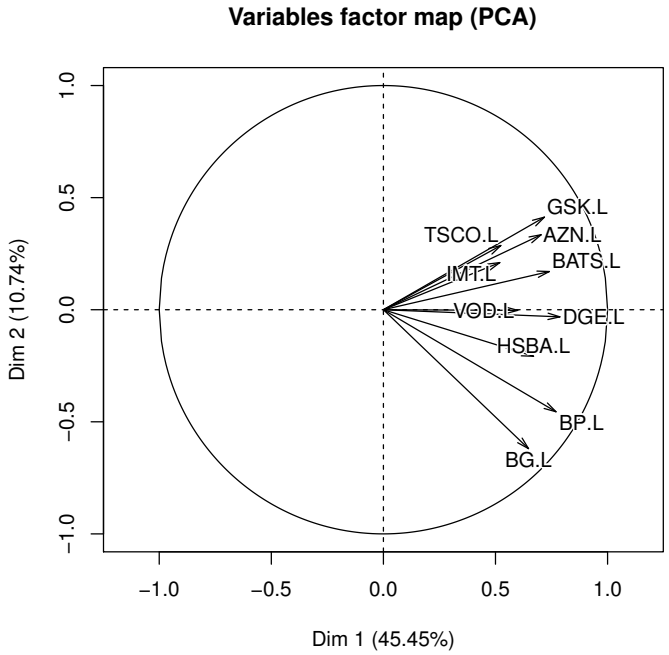

FIGURE 3.2 – camem

On rappelle qu'en ACP, plus une variable, voir un groupe de variables contribue à la formation d'un axe plus elles sont proches en terme de corrélation, l'ACP projetant les variables selon une métrique basé sur la matrice de corrélations.

Selon la dimension 1, Tous les cours sont corrélés, ici entre 0,5 et 1. Or selon la dimension 2, on distingue clairement sur se graphique 4 groupes : le premier (Tesco et IMT), le deuxième (GSK,AZN et BATS), le troisième (DGE et HSBC) et enfin

le quatrième (BP et BGL). On peut conclure ici que les variables de ces quatre groupes sont très liées entre elles . Il n'y a pas de variables corrélées négativement. qu'avec la variable expo qui va (en fait qui se comporte de façon inverse aux premières). Ainsi, pour les variable de chaque groupe , quand l'une des valeur est positive l'autre le sera aussi

Toujours selon la deuxième dimension , on explique 57% des données choisies. En effet, en utilisant la fonction Summary sur note ACP on obtient :

> summary(D.PCA)

Call :

PCA(rSERIES)

Eigenvalues Dim.1 Dim.2 Variance 4.545 1.074 Cumulative

3.3 Reconstitution de l'indice domestique

On reconstitue ici, l'indice FTSE à partir des rendement des indices. On constate que globalement on peut observer les mêmes tendances sur les indices et sur les rendements. L'impact de la crise de 2008 2009 est ici très visibles. Par ailleurs, on peut expliquer le premier pic si l'on sait qu'il correspond à un moment ou le barils de Brent était au plus hauts. Ce qui implique une extrême volatilité des cours, notamment pour les valeur pétrolières.

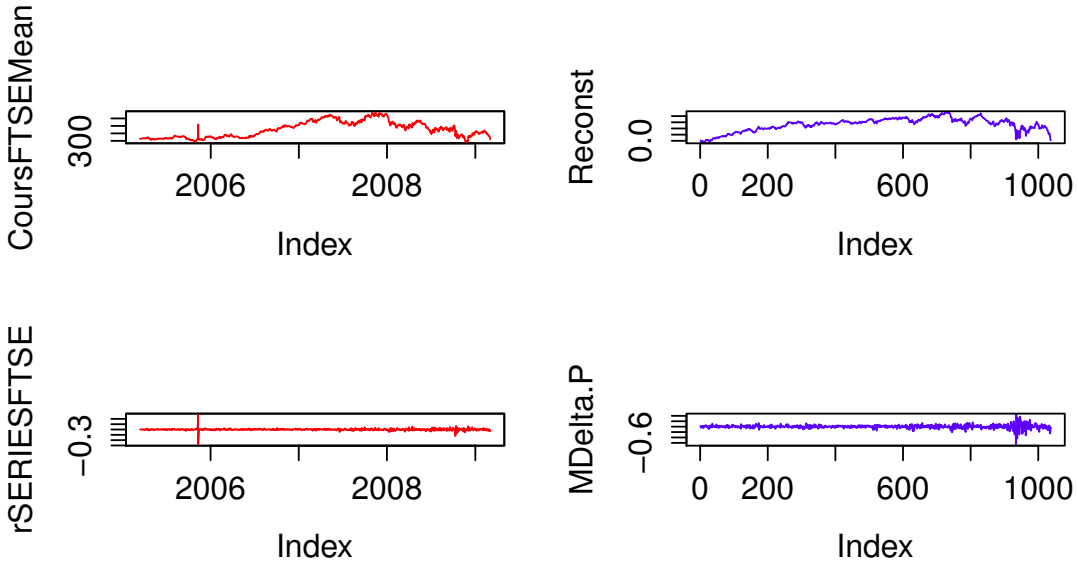

FIGURE 3.3 – PlotFinal

Chapitre 4

Conclusion

L'analyse des composantes principales s'est révélé très efficace. En effet, avec 57 % des données, nous avons pu expliquer l'indice domestique. En effet, avec seulement 10 cours, correspondant à 10 des 15 plus grosses capitalisations du FTSE, nous avons dégager la même tendance générale que celle des 100 cours de cette indice. Nous avons ainsi, mis en évidence l'intérêt de l'ACP pour l'étude des cours financiers.

Chapitre 5

Sources

Cour de Monsieur Berra - ESILV Statistical learning 2014