Méthodes Statistiques pour la Finance

Victor MOINE

Contents

Abstract

13 pages (en enlevant les pages de garde, les espaces dûs à l'inclusion de graphiques sur Latex) / Les liens des données sont en commentaires des scripts Matlab

Part I
Analyse en composantes principales.

1 Présentation du problème.

Le but de l'analyse en composantes principales est de représenter de manière parcimonieuse et pertinente un nuage de points vivant dans un espace de grande dimension. C'est une méthode de projection linéaire pour diminuer le nombre de paramètres. En effet, l'ACP va projeter les données sur un espace de dimension moins élevée par rapport à l'espace d'origine.

2 Données du problème.

On dispose d'un tableau de n lignes et p colonnes.

- Une ligne i : $(x_i^1, ..., x_i^p)$ =1 individu i
- Une colonne j : $(x_1^j, ..., x_n^j)$ =1 individu j

Dans notre exemple:

- les individus seront les jours de trading: du 02/11/2015 au 22/01/2016
- les variales seront les rendements des actifs : 36 actifs du CAC40 (et non 40, certaines données par exemple pour Alstom n'étant pas disponible sur toute la période)

On va considérer des versions centrées et réduites des variables: on va effectuer les opérations suivantes sur les variables :

$$\bar{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

$$x_i^j \leftarrow \frac{x_i^j - \bar{x}^j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^j - \bar{x^j})^2}}$$

En effet ici, bien que toutes les données soient "homogènes à des rendements" (ie: elles sont à peu près de même grandeur), on peut dans certain cas plus généraux réaliser une ACP sur des des variables hétérogènes. Normer les variables, permet ainsi de réduire l'effet de l'ordre de grandeur de variables de natures différentes.

3 Résolution.

Comme dans le cours, les 3 étapes d'une ACP sont :

- Etape 1 : construction de la matrice de corrélation empirique entre les variables V
- Etape 2 : recherche des plus grandes valeurs propres de V et des vecteurs propres associés
- Etape 3 : représentation par projection et interprétation

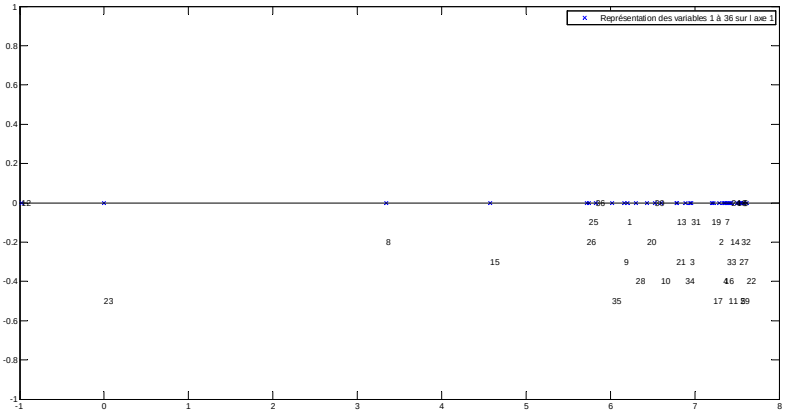
Algorithm 1 extrait de acp.m - Suite : cf annexe

```
% Centrage et normalisation des points variables
for j=1:p
    mu = (1/n)*sum(X(:,j)); % moyenne du point variable j
    Y = X(:,j)-mu;
    va = (1/n)*(Y')*Y;
    X(:,j) = (X(:,j)-mu)/sqrt(va);
end
```

4 Interprétation.

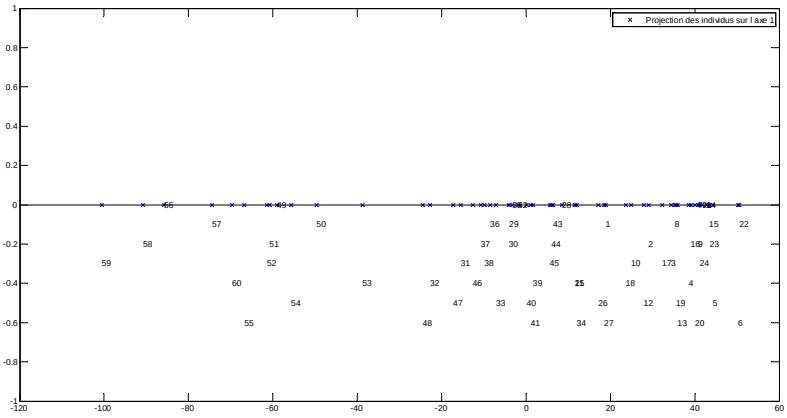
4.1 Représentation des variables sur l'axe 1 :

Les variables sont numérotées de 1 à 36 (le numéro est placé en dessous ou au dessus de la projection de la variable sur l'axe 1).



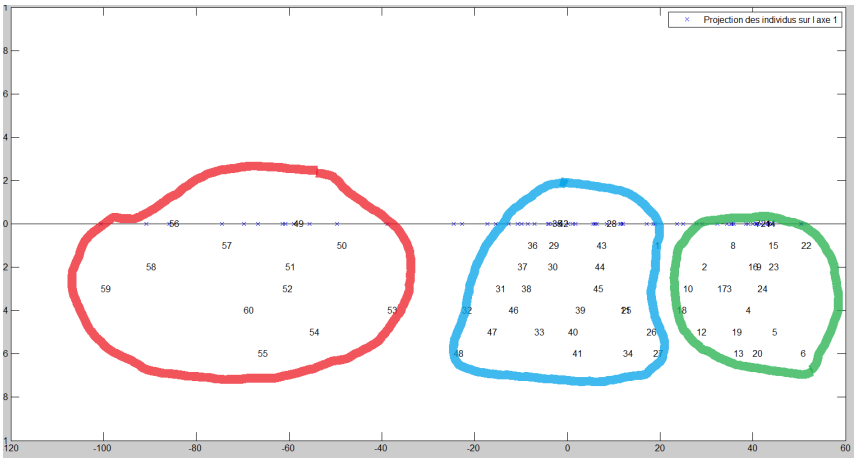
L'axe 1 est le vecteur propre $v_1 = [-0.0577, 0.0638, -0.0527, ...]$ **qui est bien de la forme** $[\alpha +/- \epsilon, \alpha +/- \epsilon, ...]$ **d'après Matlab.**

4.2 Projection des individus sur l'axe 1 :



On remarque, comme le confirme la théorie, que l'axe 1 est la meilleure représentation en dimension 1 du nuage des individus, au sens où il maximise la dispersion, car les projections des "jours de trading" sont en effet bien étalés sur l'axe 1.

On remarque par exemple que puisque la composante Accor (variable 1) est positive sur l'axe 1, alors pour tous les jours qui se projettent positivement, la corrélation entre ces jours et Accor est positive. Ou réciproquement, pour tous les jours tels que leur corrélation avec Accor est positive, alors s'ils se projettent négativement sur l'axe 1 cela implique que Accor était ≤ 0 ces jours-là.

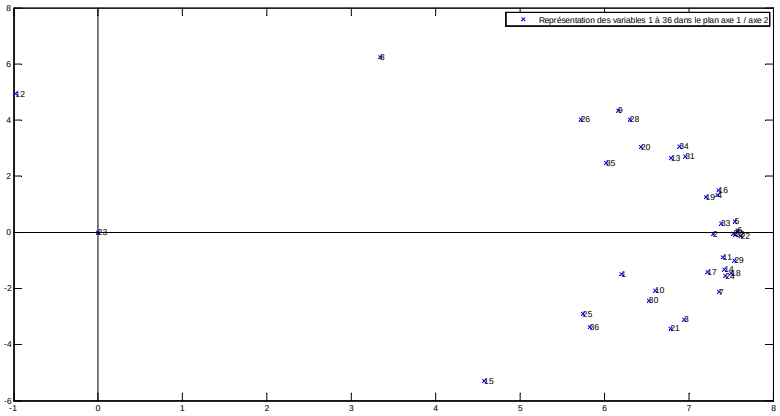


On peut aussi distinguer 3 groupes : les jours de trading (numérotés de 1 à 60 à partir de la date initiale du 02/11/2015) où le marché est

- bull : pour les jours entourés en vert (se projettent très positivement sur l'axe 1)
- bear : pour les jours entourés en rouge
- flat : pour les jours entoués en bleu

On constate aussi un effet de clustering : des jours consécutifs appartiennent au même groupe.

4.3 Représentation des variables dans le plan axe 1 / axe 2 :



On constate que l'axe 2 oppose 2 types de variables: les variables

- 8/26/35/9/28/20/13/34/31/19/4/12
- 25/36/1/30/10/21/7/17

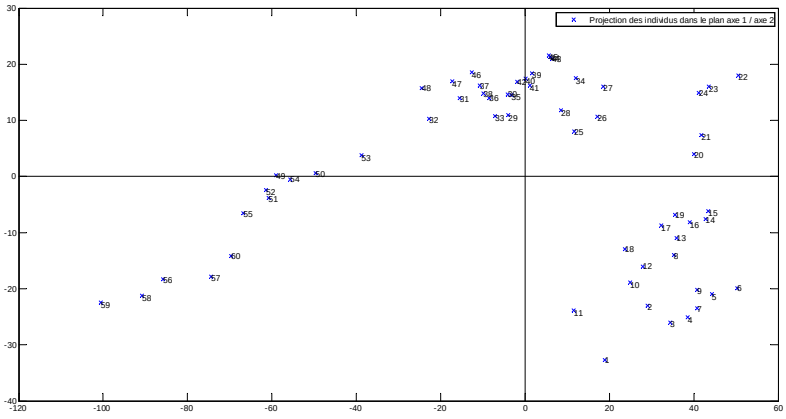
correspondant respectivement aux actifs :

- CapGemini/Publicis/Veolia/Axa/Renault/Legrand/Engie/Peugeot/St Gobain/Klepierre/Airbus/Bouygues
- Orange/Vivendi/Accor/Sanofi/Vinci/LVMH/Carrefour/Kering

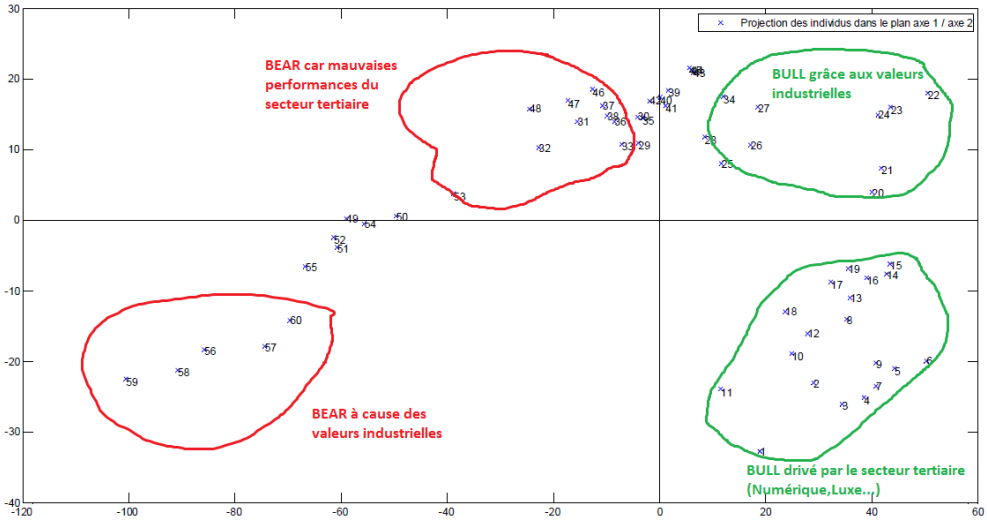
On constate donc que l'axe 2 distingue globalement les valeurs industrielles (Veolia,Renault,Legrand,Peugeot,St Gobain,Airbus,Bouygues) aux valeurs du secteur tertiaire des services (du numérique : Orange, Vivendi / du luxe : LVMH, Kering): l'axe 2 est un axe sectoriel.

4.4 Projection des individus dans le plan axe 1 / axe 2 :

Voici la meilleure représentation de notre nuage de points en dimension 2 : là encore on remarque un effet de clustering (les jours 56,57,58,59,60 sont proches, de même que 46,47,48,...) ce qui est logique.



On peut observer ceci :



En effet, par exemple pour les valeurs du secteur tertiaire telles que LVMH (variable 21), celles-ci sont très négatives sur l'axe 2, donc la corrélation entre les jours 4/5/6/7 et ces valeurs du secteur tertiaire sont négatives: les jours 4/5/6/7 (correspondant aux 6,7,8,9 novembre 2015) sont donc des journées BULL drivées par les valeurs du secteur tertiaire (à 1 jour qui se projette négativement correspond une valeur du secteur tertiaire positive car la corrélation est négative).

En résumé, il y a 4 types de journées: BULL ou BEAR et ceci A CAUSE(ou GRACE) aux valeurs INDUSTRIELLES ou du SECTEUR TERTIAIRE.

Part II

Régression linéaire : régression MCO et Ridge.

5 MCO : présentation du problème & hypothèses

La régression linéaire a pour but d'expliquer une variable Y à partir d'une combinaison linaire de variables explicatives X , de sorte que

$$Y = \phi\theta + \varepsilon$$

où :

- Y est le vecteur des n observations
- ϕ est la matrice (n, p) des régresseurs
- θ est le vecteur inconnu de dimension p
- ε est un bruit

On fait les 2 hypothèses suivantes :

- H1 : $rang(X) = p$
- H2 : $\varepsilon \sim N(0, \sigma^2 I_p)$

L'hypothèse H2 n'est pas nécessaire pour trouver l'expression de l'estimateur des moindres carrés mais elle est importante pour la détermination des intervalles de confiance et des p-valeurs.

6 Données du problème.

Ici, on veut expliquer les rendements de l'entreprise Total sur la période du 01/01/2015 au 31/12/2015 en fonction du :

1. cours du CAC40
2. cours du Dow Jones
3. cours d'Airfrance
4. cours de BP
5. cours de Danone
6. cours de L'Oréal
7. cours de Michelin
8. cours d'Orange
9. cours de Vinci
10. prix du baril de brut en Europe

Commentaire sur les régresseurs:

1. on peut *a priori* penser que le cours du CAC40 explique bien le cours de Total puisque les 2 sont très liés (Total étant la 1ère capitalisation boursière du CAC40)
2. on peut aussi aussi se poser la question de l'influence des places boursières étrangères telles que l'indice Dow Jones sur le cours de Total
3. les cours d'Airfrance et Total sont sûrement anticorrélés. En effet, plus le prix du barril augmente, plus Total fait des bénéfices mais plus Airfrance doit essuyer des pertes (sa marge diminuant à cause du prix du carburant)
4. Le cours de British Petroleum doit expliquer le cours de Total (et *vice verca*)
5. *a priori* aucun lien entre Danone et Total
6. *a priori* aucun lien entre L'Oréal et Total
7. les cours de Total et Michelin doivent eu aussi être anticorrélés pour la même raison que pour Airfrance
8. *a priori* aucun lien entre Orange et Total
9. *a priori* aucun lien entre Vinci et Total
10. le cours du baril doit très bien expliquer le cours de Total (plus le premier est élévé et plus Total fait des bénéfices)

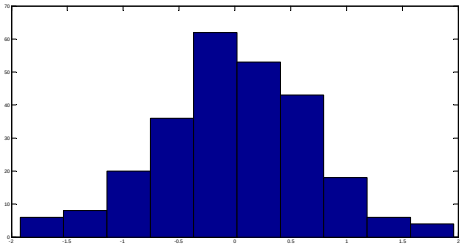
Remarque: il aurait été intéressant d'y inclure le régresseur "température" mais il est compliqué d'extraire les données d'Internet.

7 Résolution et interprétation.

D'après le cours, on sait que l'estimateur des MCO est :

$$\widehat{\theta} = (\phi^T \phi)^{-1} \phi^T Y$$

Voici ci-dessous l'histogramme des résidus $e = Y - \hat{Y}$:



On constate que les résidus ont à peu près une distribution normale, ce qui est cohérent avec les conséquences des hypothèses faites au début.

7.1 Intervalles de confiance sur θ_j.

D'après le cours, on sait que

$$\frac{\hat{\theta_j} - \theta_j}{\sqrt{\hat{\sigma}^2(\phi^T\phi)^{-1}_{j,j}}} \sim Student(n-p)$$

où $\hat{\sigma}$ est l'estimateur de la variance :

$$\hat{\sigma} = \frac{||e||^2}{n-p}$$

d'où :

$$\mathbb{P}[-a \leq \frac{\hat{\theta_j} - \theta_j}{\sqrt{\hat{\sigma}^2(\phi^T\phi)^{-1}_{j,j}}} \leq a] = F_{Student}(a) - F_{Student}(-a)$$

ie:

$$\mathbb{P}[\theta_j \in [\hat{\theta_j} - a\sigma\sqrt{\alpha_j}, \hat{\theta_j} + a\sigma\sqrt{\alpha_j}] = F_{Student}(a) - F_{Student}(-a), \alpha_j = (\phi^T\phi)^{-1}_{j,j}$$

On choisist a pour que le membre de droite soit égal à $1-\alpha$ et alors $[\hat{\theta_j} - a\sigma sqrt(\alpha_j), \hat{\theta_j} + a\sigma sqrt(\alpha_j)]$ est un intervalle de confiance à $100(1-\alpha)\%$ sur θ_j. Puis en utilisant le fait que la distribution de la loi de Student est paire, on déduit qu'il faut prendre $a = -F^{-1}_{Student}(-\frac{\alpha}{2})$

Voici l'estimateur des moindres carrés et les intervalles de confiance à 95% sur θ_j:

```
theta_chapeau =

   -0.0036
   -0.0001
    0.0771
    0.0617
   -0.0829
    0.1083
   -0.0193
    0.8435
    0.2562
   -0.0063


intervalle =

   -0.0049   -0.0022
   -0.0002    0.0001
   -0.1071    0.2614
    0.0528    0.0706
   -0.1737    0.0079
    0.0738    0.1428
   -0.0503    0.0118
    0.6608    1.0262
    0.2000    0.3124
   -0.0402    0.0276
```

7.2 Sélection de variables.

Testons maintenant l'hypothèse de significativité d'un coefficient θ_j (test de Student) : on sait que

$$\frac{\hat{\theta_j} - \theta_j}{\sqrt{\hat{\sigma}^2(\phi^T\phi)^{-1}_{j,j}}} \sim Student(n-p)$$

d'où sous H0:"$\theta_j = 0$", on a : $t = \frac{\hat{\theta_j}}{\sqrt{\hat{\sigma}^2(\phi^T\phi)^{-1}_{j,j}}} \sim Student(n-p)$.

On calcule alors la p-valeur associé au test t pour chaque coefficient (qui n'est autre que $1 - F_{Student(n-p)}(t)$):

```
t =

   -5.0793
   -0.6891
    0.8247
   13.6274
   -1.7993
    6.1808
   -1.2227
    9.0936
    8.9799
   -0.3665


p_valeur =

    1.0000
    0.7543
    0.2052
         0
    0.9634
    0.0000
    0.8887
         0
         0
    0.6429
```

Or, plus la p-valeur est faible et plus on rejette H0 l'hypothèse de nullité du coefficient. L'analyse des p-valeurs nous indique alors que :

- θ_4,θ_6,θ_8, θ_9 sont significatifs
- θ_1,θ_2, θ_5, $\theta_7$ne sont pas significatifs

Autrement dit, le cours de Total est bien expliqué :

- par BP (ce qui est cohérent car BP et Total sont toutes les deux des entreprises pétrolières)
- par L'Oréal (ce qui est surprenant, mais une recherche sur Internet montre qu'il y a du pétrole et du plastique dans les cosmétiques)
- le cours d'Orange et de Vinci

En revanche, le cours du CAC40, du Dow Jones, de Danone n'expliquent pas le cours de Total.

Enfin, le cours du baril de brut n'explique que moyennement le cours de Total: ceci est sûrement dû au fait que l'impact d'une montée du prix du brut n'est répercutée que plus tard dans l'année.

8 Régression Ridge.

8.1 Expression de l'estimateur Ridge.

La régression Ridge permet de corriger les 2 inconvénients suivants de l'estimateur des MCO :

- l'estimateur des MCO privilégie le biais à la variance (c'est d'ailleurs le meilleur estimateur sans biais au sens où c'est celui de variance minimale)
- il faut que $\phi^T\phi$ inversible et donc $n \geq k$, ce qui n'est pas toujours le cas

L'estimateur Ridge est alors défini comme :

$$\hat{\theta}_{Ridge} = argmin(||Y - \phi\theta||_2^2 + \lambda||\theta||_2^2)$$

où λ est un paramètre de pénalisation à fixer par l'utilisateur. Nous verrons dans la suite comment le fixer de façon optimale.

D'après le cours, l'expression de $\hat{\theta}_{Ridge}$ est :

$$\hat{\theta}_{Ridge} = (\phi^T\phi + \lambda I_p)^{-1}\phi^T Y$$

8.2 Comment fixer λ : la cross-validation.

La cross-validation est une méthode pratique pour fixer le λ optimal (qui dépend de θ inconnu). On procède comme suit :

Etape 1 : on partitionne les données en K ensembles de taille égales

Etape 2 : pour chaque $k = 1, ..., K$, on calcule $\hat{\theta}_{-k}(\lambda)$ l'estimateur Ridge sur l'ensemble des données exceptée celles du k-ième ensemble. On calcule l'erreur de prévision empirique de cet estimateur sur l'ensemble numéro k défini comme suit :

$$CVerreur_k(\lambda) = \frac{1}{card(ensemble_k)} \sum_{Y_j, X_j \in \{ensemble_k\}} (Y_j - \phi_j\theta_{-k}(\lambda))^2$$

Algorithm 2 extrait de regression.m - Suite : cf annexe

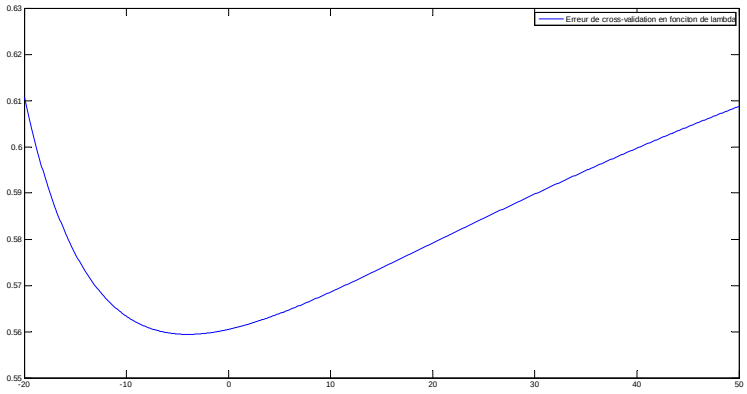
```
for u=1:Nu
    for k = 1:K-1
        Y_moins_k = Y;
        phi_moins_k = phi;
        %Pour construire theta_ridge sur toutes les données moins celles de l'esemble k
        for i=1:K
            Y_moins_k(K*(k-1)+i)=[];
            phi_moins_k(K*(k-1)+i,:)=[];
        end
        %Yj,phij dans k - on ne garde que l'ensemble numéro k
        for j=1:K
            phi_que_k(j,:) = phi(K*(k-1)+j,:);
            Y_que_k(j)=Y(K*(k-1)+j);
        end

        theta_ridge_moins_k =  pinv((phi_moins_k'*phi_moins_k)+lambda(u)*diag(ones(p,1)))*phi_moins_k'*Y_moins_k;

        V = Y_que_k'-phi_que_k*theta_ridge_moins_k;
        CV_erreur_sur_k(k) = (1/K)*(V'*V);
    end
    CV_totale(u) = (1/K)*sum(CV_erreur_sur_k);
end
```

Etape 3 : on calcule l'erreur de cross-validation totale pour un λ donné :

$$CVerreur(\lambda) = \frac{1}{K}\sum_{k=1}^{K}(CVerreur_k(\lambda))$$

Enfin on choisit le λ qui minimise cette dernière quantité. Voici le script dans *regression.m* qui programme la procédure de cross-validation :

Voici le graphique représentation l'erreur de cross-validation totale en fonction de λ(le graphique en cloche est cohérent avec ce qu'on doit atendre):



On choisit donc $\lambda \approx -3.9$. L'estimateur Ridge correspondant est :

```
theta_ridge_star =

   -0.0037
   -0.0001
    0.0887
    0.0612
   -0.0979
    0.1132
   -0.0142
    0.9047
    0.2549
   -0.0032
```

On peut alors constater que l'estimateur Ridge et l'estimateur MCO sont très proches, ce qui est cohérent (excepté pour le dernier coefficient):

```
theta_ridge_star =
                    theta_chapeau =
   -0.0037
   -0.0001             -0.0036
    0.0887             -0.0001
    0.0612              0.0771
   -0.0979              0.0617
    0.1132             -0.0829
   -0.0142              0.1083
    0.9047             -0.0193
    0.2549              0.8435
   -0.0032              0.2562
                       -0.0063
```

Part III
Portefeuille de Markowitz

9 Présentation du problème.

Le but d'un portefeuille de Markowitz est de trouver des combinaisons de titres (les portefeuilles) les plus intéressants statistiquement ie fournissant la meilleure performance pour un niveau de risque donné. La performance sera mesurée par la moyenne des rendements et le risque par leur variance. Autrement dit, on cherche les portefeuilles d'espérance de rendement maximal pour une variance inférieure à un niveau donné.

10 Données du problème.

Dans les 2 cas (avec ou en présence d'actif risqué), on considère un marché avec:

- une date initiale $t = 0 \overset{ici}{=} 02/11/2015$
- un horizon $t = 1 \overset{ici}{=} 22/01/2015$

Le marché est constitué de 36 actifs du CAC40 (et non 40 car certaines cotations n'étaient disponibles qu'en partie comme par exemple Alstom).

11 Cas 1 : en présence d'actif sans risque.

11.1 Mise en équation.

En plus des actifs risqués mentionnés ci-dessus, on considère un actif sans risque de taux sans risque $r = 0.03$ arbitraire (cf script $markowitz.m$)

La résolution du problème de minimisation sous contrainte :

$$max_{a_0,a}\mathbb{E}[V_1(a_0,a)]$$

sous les contraintes :

$$Var[V_1(a_0,a)] \leq \sigma^2, V_0(a_0,a) = v$$

où :

- V : valeur du portefeuille, dépendant de
- a_0: le nombre d'unités d'actif sans risque
- a : le nombre d'unités d'actifs risqué

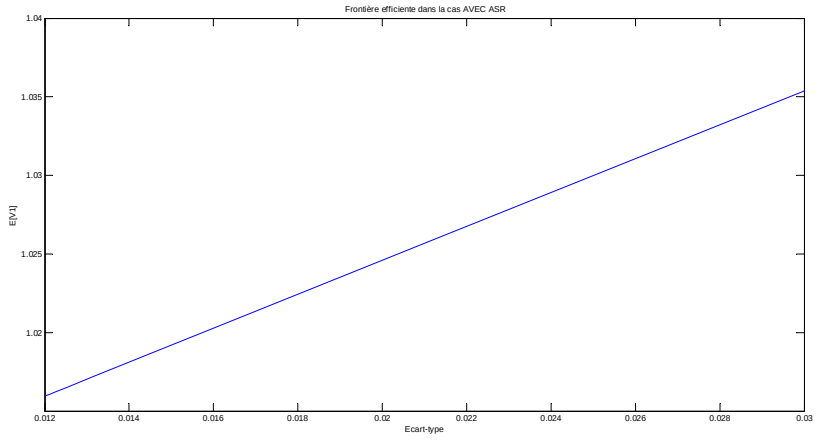
11.2 Solution.

D'après le cours, pour les portefeuilles de valeur v en $t = 0$, la frontière efficience est constitué des portefeuilles tels que :

- $a = \frac{1}{\lambda}(diag(p_0))^{-1}\Omega^{-1}\tilde{\mu}$
- $a_0 = v - \frac{1}{\lambda}\tilde{\mu}^T\Omega^{-1}e$
- $\lambda = \frac{1}{\sigma}(\tilde{\mu}^T\Omega^{-1}\tilde{\mu})^{1/2}$

La frontière efficiente est (en paramétrisant par λ) dans le plan espérance/écart-type : $\mathbb{E}[V_1(\lambda)] = v(1+r) + \sqrt{Var(V_1(\lambda)}(\tilde{\mu}^T\Omega^{-1}\tilde{\mu})^{1/2}$

Voici le graphique généré par *markowitz.m* **pour le cas avec ASR:**



Voici un extrait du vecteur des composition :

```
a =

   -0.0248
   -0.2694
   -0.0590
    0.0540
   -0.0645
   -0.1229
    0.0472
    0.0118
    0.0537
```

Il faut dont détenir 0.0540 unités de l'actif 4 et vendre 0.0248 unités de l'actif 1 par exemple.

12 Cas 2 : sans ASR.

12.1 Mise en équation.

Avec les mêmes notations, le problème s'écrit :

$$max_a \mathbb{E}[V_1(a)]$$

sous les contraintes :

$$Var[V_1(a)] \leq \sigma^2, V_0(a) = v$$

12.2 Solution.

Pour déterminer la frontière efficiente, nous allons procéder différemment. En effet, on ne dispose pas (dans le cours) de l'expression analytique de la frontière efficiente Nous allons raisonner en % pour la composition du portefeuille ce qui sera beaucoup plus pratique. La procédure consistera alors à "ploter" un grand nombre de portefeuille (avec des poids aléatoires en % pour chacun des actifs). La frontière efficiente se révélera d'autant mieux que le nombre de portefeuilles tirés sera grand. Puis grâce à une autre procédure, on déterminera le meilleur portefeuille pour un écart-type donné.

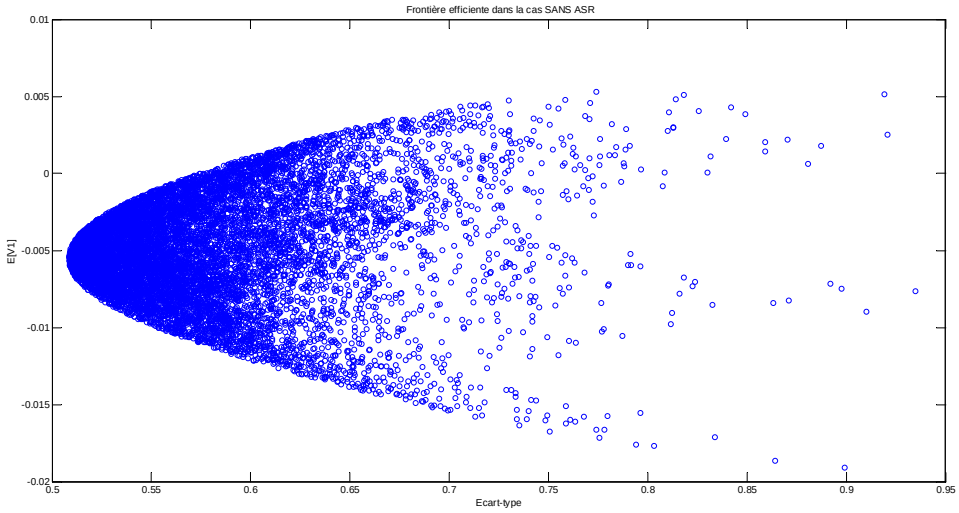
12.3 Résultat.

12.3.1 Constatation.

En considérant les rendements des 36 actifs du CAC40, on constate qu'on obtient une $1/2$ **parabole.**

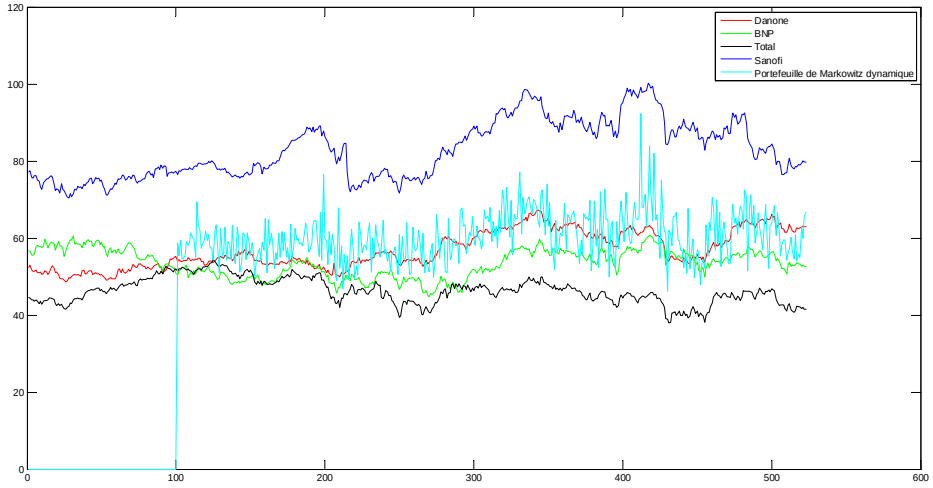
Ceci est dû au fait que la période considérée du $02/11/2015$ **au** $22/11/2015$ **est trop petite, si bien que les rendements ne sont pas normaux. En effet, quand on veut tracer la frontière efficente avec comme jeu de**

données des rendements normaux, on obtient une parabole entière.



12.3.2 Exemple appliqué du trader.

Tout ceci n'est pas très appliqué. Mettons-nous alors à la place d'un trader voulant déterminer à chaque jour de trading, quel est sont portfefeuille de Markowitz. Considérons un jeu de donnée plus restreint : on va prendre seulement 4 actifs risqués (Danone, BNP, Total, Sanofi) mais sur une période plus grande du 01/01/2014 au 01/01/2016. Un des problèmes du trader va alors consister à estimer entre autre la moyenne de ses rendements: une stratégie possible est de prendre une fenêtre dynamique de 100 observations pour l'estimation des moyennes . Ces 100 observations constituent l'historique des données pour le trader. Puis, à chaque nouvelle journée de trading, cette fenêtre se décalera en intégrant le nouveau rendement (c'est une fenêtre glissante). C'est une sorte de portefeuille de Markowitz dynamique. Voici le résultat :



La procédure est assez longue (environ 4mn, mais immédiate pour le trader si il l'exécute chaque jour) car pour les 423 journées de trading, il faut à chaque fois:

- recalculer les moyennes et les covariances en intégrant le nouveau rendement
- simuler environ 1000 portefeuilles aléatoires basés sur ces rendements pour déterminer la frontière efficiente
- trouver les poids optimaux pour un niveau de risque donné

Le portfeuille performe assez bien, car il résiste bien aux périodes de baisse de 3 actifs sur 4 (période 100-250) et augmente significativement lorsque ceux-ci augmentent ensemble (période 400-425). On constate aussi que la valeur du portefeuille oscille beaucoup, ce qui est sans doute dû au fait que la moyenne des rendements estimée au fur et à mesure oscille aussi. Une manière d'améliorer l'estimation de la moyenne de ces rendements aurait été d'utiliser un filtre de Kalman.

Part IV

Annexe : code des 3 parties (ACP - Régression MCO - Régression Ridge - Markowitz) - cf pages suivantes

Algorithm 3 *acp.m*

```
clear all;
close all;
clc;

load cac40.csv;

X = cac40;
[n,p] = size(cac40);
X = X(:,2:p);
[n,p] = size(X);


% Centrage et normalisation des points variables
for j=1:p
    mu = (1/n)*sum(X(:,j)); % moyenne du point variable j
    Y = X(:,j)-mu;
    va = (1/n)*(Y')*Y;
    X(:,j) = (X(:,j)-mu)/sqrt(va);
end

% Matrice des corrélations (matrice nxn)
V = (1/n)*X*X';

% Recherche des plus grandes valeurs propres et vecteurs propres
[vep,vap] = eig(V); % les vecteurs propres sont normalisés


%%
% Représentation des variables sur l'axe 1
RepVariableAxe1 = X'*vep(:,1); % vep(:,n) : + grand vecteur propre de V.
SurAxe = zeros(1,p);
figure;
plot(RepVariableAxe1,SurAxe,'x');
hold on;
xL = xlim;
yL = ylim;
plot(xL, [0 0], 'k-');
legend('Représentation des variables 1 à 36 sur l axe 1');
for i=1:p
    text(RepVariableAxe1(i),SurAxe(i)+(-0.1)*mod(i,6),int2str(i))
end



% Projection des individus sur l'axe 1

% un individu = 1 jour de trading = 1 vecteur de taille p = 36
% l'axe 1 = le vecteur propre de XX' de taille (n,p)x(p,n)=(n,n) =
1 vecteur de taille n = 60
% La j-ième composante principale fournit les coordonnées des n individus sur le j-ième axe principal
% C_{j} = X*vecteur_propre_{j} de V (n,p)*(n,1)
figure;
RepIndividusAxe1 = X*(1/sqrt(vap(1,1)))*(X')*vep(:,1);
plot(RepIndividusAxe1,zeros(1,n),'x');
legend('Projection des individus sur l axe 1');
U = zeros(n,1);
for i=1:n
    text(RepIndividusAxe1(i),U(i)+(-0.1)*mod(i,7),int2str(i))
end
hold on;
xL = xlim;
yL = ylim;
plot(xL, [0 0], 'k-');

% Représentation des variables dans le plan (axe1,axe2)
RepVariableAxe2 = X'*vep(:,2);
figure;
plot(RepVariableAxe1,RepVariableAxe2,'x');
legend('Représentation des variables 1 à 36 dans le plan axe 1 / axe 2');
for i=1:p
    text(RepVariableAxe1(i),RepVariableAxe2(i),int2str(i))
end
hold on;
xL = xlim;
yL = ylim;
plot(xL, [0 0], 'k-');
plot([0 0],yL, 'k-');


% Projection des individus dans le plan (axe1,axe2)
figure;
RepIndividusAxe2 = X*(1/sqrt(vap(2,2)))*(X')*vep(:,2);
plot(RepIndividusAxe1,RepIndividusAxe2,'x');
legend('Projection des individus dans le plan axe 1 / axe 2');
for i=1:n
    text(RepIndividusAxe1(i),RepIndividusAxe2(i)+(-0.1)*mod(i,6),int2str(i))
end
hold on;
xL = xlim;
yL = ylim;
plot(xL, [0 0], 'k-');
plot([0 0],yL, 'k-');

%Dans l'espace ...
RepVariableAxe3 = X'*vep(:,3);
figure;
plot3(RepVariableAxe1,RepVariableAxe2,RepVariableAxe3,'x');
grid on;
figure;
RepIndividusAxe3 = X*(1/sqrt(vap(3,3)))*(X')*vep(:,3);
plot3(RepIndividusAxe1,RepIndividusAxe2,RepIndividusAxe3,'x');
legend('Projection des individus dans l espace axe1/axe2/axe3');
for i=1:n
    text(RepIndividusAxe1(i),RepIndividusAxe2(i),RepIndividusAxe3(i),int2str(i))
end
hold on;
grid on;
```

Algorithm 4 *regression.m* MCO et RIDGE

```
clear all;
close all;
clc;

% DATES : 01/01/2015 ----- 31/12/2015 (1 an)
% Action Total (ou EDF) expliquée par :
% (100 jours de cotations par exemple = 100 obs yi)
% Rq : avec les cours actuels, large spectre d'obeservations
% Suivre démarche pdf régression

% ATTENTION: comparer ques les jours de cotations coincident (trous le week
% end)

% 1/ cours du cac40 (https://fr.finance.yahoo.com/q/hp?s=^FCHI)
% 2/ cours du dow jones https://research.stlouisfed.org/fred2/series/DJIA/downloaddata
% 3/ cours d'Airfrance (qd pétrole augmente, moins de bénéf (https://fr.finance.yahoo.com/q/hp?s=AF.PA)
% 4/ cours d'une entreprise concurrente - BP (https://fr.finance.yahoo.com/q/hp?s=BP.L)
% 5/ cours d'entreprises qui n'ont (vraiment) rien à voir, par exemple
DANONE (https://fr.finance.yahoo.com/q/hp?s=BN.PA)
% 6/ cours d'entreprises qui n'ont (apparemment) rien à voir, par exemple
L'OREAL (https://fr.finance.yahoo.com/q/hp?s=OR.PA&b=1&a=00&c=2015&e=31&d=11&f=2015&g=d)
% 7/ cours de Michelin (pneus<- pétrole) (https://fr.finance.yahoo.com/q/hp?s=OR.PA)
% 8 cours d'Orange https://fr.finance.yahoo.com/q/hp?s=ORA.PA&b=1&a=00&c=2015&e=31&d=11&f=2015&g=d
% 9/cours Vinci https://fr.finance.yahoo.com/q/hp?s=DG.PA&b=1&a=00&c=2015&e=31&d=11&f=2015&g=d
% 10/prix du baril de brut Europe (taper : https://research.stlouisfed.org/fred2/series/DCOILBRENTEU/downloaddata)


% de 01/01/2010 à 01/01/2015, chaque jour

%Régression linéaire simple
% Variable observée (cours d'ouverture de total (2eme colonne))
load total.csv

% Variables explicatives
load cac40.csv;load dowjones.csv;load af.csv;load bp.csv;load danone.csv;load loreal.csv;load michelin.csv;load or

n=size(total(:,2));
n=n(1)
p = 10; % nombre de régresseurs

% Estimateurs MC
phi = [cac40(:,2),dowjones(:,2),af(:,2),bp(:,2),danone(:,2),loreal(:,2),michelin(:,2),orange(:,2),vinci(:,2),brut
theta_chapeau = pinv((phi'*phi))*phi'*total(:,2)
Ychapeau= phi*theta_chapeau;

% Histogramme des résidus
Y = total(:,2);
e= Y-Ychapeau;
hist(e);

% Intervalles de confiance à 95% sur theta_j
e = Y-Ychapeau;
MSE = e'*e/(n-p); %non n-p+1
sigma_hat = sqrt(MSE);
alpha = 0.05;
quantile = -qt(alpha/2,n-p); %qt: quantile de la loi de Student à n-(p+1) degrés de liberté
intervalle = zeros(p,2);
M = pinv(phi'*phi);
for j=1:p
    a = quantile*sigma_hat*sqrt(M(j,j));
    intervalle(j,:) = theta_chapeau(j)+[-a,a];
end
intervalle

% Sélection de variables
% p-valeurs
std = sqrt(diag(MSE*inv(phi'*phi)));
t = theta_chapeau./std % fournit un test de student pour chaque coefficient
p_valeur = 1-tcdf(t,n-p)
% Plus la p-valeur est faible, plus on a confiance dans H1 ie plus on
% rejette H0: l'hypothèse de nullié du coefficient

% Régression Ridge
Y = total(:,2);
phi = [cac40(:,2),dowjones(:,2),af(:,2),bp(:,2),danone(:,2),loreal(:,2),michelin(:,2),orange(:,2),vinci(:,2),brut
lam = 2;
theta_ridge = pinv((phi'*phi)+lam*diag(ones(p,1)))*phi'*Y;
% Choix de lambda - Cross validation: ici n = 256: on partitionne en K = 16
% ensembles de 16 éléments chacun
K = 16;
lambda = [-20:0.1:50];
Nu = size(lambda);
Nu = Nu(2);

for u=1:Nu
    for k = 1:K-1
        Y_moins_k = Y;
        phi_moins_k = phi;
        %Pour construire theta_ridge sur toutes les données moins celles de l'esemble k
        for i=1:K
            Y_moins_k(K*(k-1)+i)=[];
            phi_moins_k(K*(k-1)+i,:)=[];
        end
        %Yj,phij dans k - on ne garde que l'ensemble numéro k
        for j=1:K
            phi_que_k(j,:) = phi(K*(k-1)+j,:);
            Y_que_k(j)=Y(K*(k-1)+j);
        end

        theta_ridge_moins_k =  pinv((phi_moins_k'*phi_moins_k)+lambda(u)*diag(ones(p,1)))*phi_moins_k'*Y_moins_k;

        V = Y_que_k'-phi_que_k*theta_ridge_moins_k;
        CV_erreur_sur_k(k) = (1/K)*(V'*V);
    end
    CV_totale(u) = (1/K)*sum(CV_erreur_sur_k);
end
figure;
plot(lambda,CV_totale)
legend('Erreur de cross-validation en fonciton de lambda');
[CV_totale_min,u_cvmin] = min(CV_totale)
lambda_star = lambda(1) +0.1*u_cvmin;
theta_ridge_star = pinv((phi'*phi)+lambda_star*diag(ones(p,1)))*phi'*Y
```

Algorithm 5 *markowitz.m* pour la frontière efficiente avec ASR

```
clear all;
close all;
clc;

% Source : Yahoo finance
% 36 actifs (et non 40 car certaines cotations n'étant disponibles qu'en partie comme par exemple Alstom
% du 02/11/2015 au 22/01/2016

load cac40.csv;

% n = nombre d'obs / p = nb d'actifs + 1
c = cac40;
[n0,p0] = size(c);
prix = cac40(:,2:p0);
[n,p] = size(prix);

% Vecteur des "rendements" entre  t0 = 21/01/2016 et t = 22/01/2016
yi = prix(n,:)./prix(n-1,:);

% Vecteur des moyennes de rendements
y = (prix(2:n,:)./prix(1:n-1,:)); % Attention hypothèse mu_i > 1+r
mu_hat = mean(y) ;

% Vecteur des moyennes des excédents de rendement
r = 0.003;
mu_tilde = (mu_hat -(1+r)*ones(1,p))';


% Matrice de corrélation empirique

% Vecteur colonne contenant les rendements de chaque actif i = 1,...,p
yi = yi';
mu_hat = mu_hat';

% Estimateur empirique de la matrice de corrélation
%omega = (1/n)*(yi-mu_hat)*(yi-mu_hat)'; <- MAUVAIS ESTIMATEUR
omega = cov(y); % <- MEILLEUR ESTIMATEUR (MATLAB)


% Matrice dont la diagonale est p_i,0 (actif i, date t = 0 = 21/01/2016 ie n-1-ième ligne du tableau
diag_po =  diag(prix(n-1,:));

% CAS 1 : AVEC ASR
% Portefeuille optimal pour sigma et v donné arbitrairement...

sigma = 0.1; % Niveau de risque
v = 1; % Valeur initiale du portefeuille

la = sqrt(mu_tilde'*pinv(omega)*mu_tilde)/sigma
ao = v -(mu_tilde'*pinv(omega)*ones(p,1)/la)
a = pinv(diag_po)*pinv(omega)*mu_tilde/la %

% Frontière optimale

%%%for k = 1:10
    %sig = sqrt(mu_tilde'*pinv(omega)*mu_tilde)/lambda
  %%%  EV1lambda(k) = v*(1+r)+(k-1)*sqrt(mu_tilde'*pinv(omega)*mu_tilde);
%%%end
ma_fonction = @(ecart_type_portefeuille) v*(1+r)+ ecart_type_portefeuille*sqrt(mu_tilde'*pinv(omega)*mu_tilde);
fplot(ma_fonction,[0.012,0.03]);
%plot([0:9],EV1lambda);
title('Frontière efficiente dans la cas AVEC ASR');
xlabel('Ecart-type');
ylabel('E[V1]');

% CAS 1 : SANS ASR
% Frontière optimale


hold on;
figure
M = 10000;


% Valeur poretfeuille = Pt = SUM(ponderations_i x Valeurs_actif_i)
% On raisonne en pourcentage

%rend = [0.05 0.08]';
%m_o = [0.03 0.3 ; 0.3 0.05];
%omega(1:2,1:2)
%omega = [0.04 0.001 0.002 ; 0.001 0.5 0.06; 0.002 0.06  0.001];
%omega = [0.04 0.01 ; 0.01 0.5];

for i = 1:M
    %ponderation = randn(p,1);  ATTENTION : ce sont les rendements qui sont
    %normaux, pas les poids. Il y a symétrie dans la parabole si les
    %rendements sont normaux. Donc prendre plus de données, pas que 60
    %journées mais sur 1 ans et avec quelques actifs seulement
     ponderation = rand(p,1);
     w = ponderation./sum(ponderation); %poids en %
     esperance = (w')*mu_hat;
     %omega = single(omega);
     variance = (w')*omega*w; % normal si pas Sn++ -> actifs redondants ?
     % Pour éviter les points extrêmes (pour le plot)
     if (sqrt(variance)<0.5)
         plot(sqrt(variance),esperance,'o');
     end
     hold on;
end
title('Frontière efficiente dans la cas SANS ASR');
xlabel('Ecart-type');
ylabel('E[V1]');
```

Algorithm 6 *markowitzbis.m* frontrière efficiente dans le cas sans ASR et avec des rendements normaux

```
clear all;
close all;
clc;

p = 4 ; %nombres d'actifs
n = 1000 ; %nombre de jours de trading
M = 10000; % nombre de portefeuilles aléatoires
%ATTENTION : il y a symétrie dans la parabole si les rendements sont normaux. Donc prendre plus de données ...

rendements = randn(n,p);
mu_hat = (mean(rendements))';
omega = cov(rendements)

risque = 0.6;
epsilon = 0.01;
res = -15;
% Tracer la frontière efficiente
% Et pour un niveau de risque donné, donne l'espérance de rendement
% correspondante ainsi que la ponderation associée
for i = 1:M
    aux = 1;
    ponderation = rand(p,1);
    w = ponderation./sum(ponderation); %poids en %
    esperance = (w')*mu_hat;
    variance = (w')*omega*w; % il faut omega dans Sn++
    if (sqrt(variance)<1) % pour éviter les points extrèmes (pour le plot)
        plot(sqrt(variance),esperance,'o');
    end
    if (abs(sqrt(variance)-risque)<epsilon) && (esperance>res)
        res = esperance;
        w_star = w;
    end
    hold on;
end
res
title('Frontière efficiente dans la cas SANS ASR');
xlabel('Ecart-type');
ylabel('E[V1]');
```

Algorithm 7 *markody.m* fonction appelée à chaque nouveau jour de trading

```
function [w_star,res,bool] = marko_dy(p,M,rendements,risque,epsilon)
    mu_hat = (mean(rendements))';
    omega = cov(rendements);

    ponderation = rand(p,1);
    w_star = ponderation./sum(ponderation);
    w_star_init = w_star;
    res = -15;
    % Tracer la frontière efficiente
    % Et pour un niveau de risque donné, donne l'espérance de rendement
    % correspondante ainsi que la ponderation associée
    %w_star = zeros(p,1); % au pire tt faire ds même fenêtre
    for i = 1:M
        ponderation = rand(p,1);
        w = ponderation./sum(ponderation); %poids en %
        esperance = (w')*mu_hat;
        variance = (w')*omega*w; % il faut omega dans Sn++
        if (sqrt(variance)<1) % pour éviter les points extrèmes (pour le plot)
            %plot(sqrt(variance),esperance,'o'); <- pas besoin de ploter
            %ici
        end
        if (abs(sqrt(variance)-risque)<epsilon) && (esperance>res) %Pour trouver w_star pour un niveau de risque d
            res = esperance;
            w_star = w;
        end
        hold on;
    end
    %title('Frontière efficiente dans la cas SANS ASR');
    %xlabel('Ecart-type');
    %ylabel('E[V1]');
    bool = (w_star_init==w_star); % Pour vérifier que ça a bien optimisé...
end
```

Algorithm 8 *markodynamique.m* script plotant la valeur du portefeuille de Markowitz au cours du temps et appelant la fonction *markody.m*

```
clear all;
close all;
clc;

% 01/01/2014 --- 01/01/2016
load actifs.csv;

c = actifs;
[n0,p0] = size(c); % n0 = nombre d'obs / p0 = nb d'actifs + 1
prix = actifs(:,2:p0);
[n,p] = size(prix); % n = nombre d'observatiobs TOTAL, p = nombre d'actifs

figure;
plot(1:n,prix(1:end,1),'r');
hold on;
plot(prix(1:end,2),'g');
hold on;
plot(prix(1:end,3),'k');
hold on;
plot(prix(1:end,4),'b');
hold on;

% rendements = prix(n,:)./prix(n-1,:); % vecteur des "rendements" entre
t0 = 21/01/2016 et t = 22/01/2016
% le fait que la frontière ne soit pas belle est dû au fait que les
% rendements ne sont pas normaux. En effet, cf figure où la frontière est
% "belle" dans le cas de rendements normaux...

% On se place dans le rôle du trader qui voit arriver les cotations au fur
% et à mesure.
% On prend une fenêtre dynamique de 100 observations pour l'estimation de mu
% Ces 100 obserrvations constituent l'historique des données pour le trader
%for i = 101:n

M = 100;
risque = 0.012;
epsilon = 0.00001;
fen = 100;
for i = 101:523
    rendements = (prix(2+(i-101):fen+(i-101),:)./prix(1+(i-101):fen+(i-101)-1,:));
    [w_star,res,bool] = marko_dy(p,M,rendements,risque,epsilon);
    cotations = (prix(i,:))';
    portefeuille(i) = (w_star')*cotations;
end
%size([101:300])
%size(portefeuille)
plot([1:523],portefeuille,'-c');
legend('Danone','BNP','Total','Sanofi','Portefeuille de Markowitz dynamique');
```