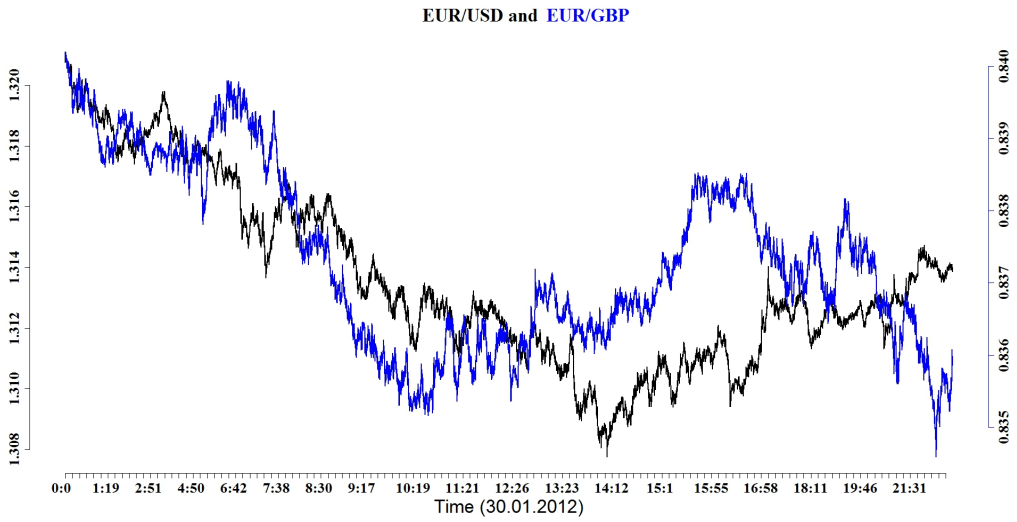
Trading Haute Fréquence
Modélisation et Arbitrage Statistique

Alexis Fauth†

Université Lille I
Master 2 Mathématiques et Finance
Mathématiques du Risque & Finance Computationnelle

2014/2015



† Membre associé de l'université Paris I, Pantheon-Sorbonne, laboratoire SAMM. alexis.fauth@gmail.com, http://samm.univ-paris1.fr/-Alexis-Fauth-

Table des matières

Présentation des Marchés Financiers

Le premier rôle des marchés financiers est de faciliter le transfert de liquidité entre un agent ayant des capacités de placement vers un autre agent ayant un besoin de financement. A cette fonction, on peut conférer aux marchés financiers un outil de couverture de risque. Le premier groupe mondial de places boursières est la bourse New-York Stock Exchange - Euronext (NYSE-Euronext), créé le 4 avril 2007 suite à la fusion entre Euronext et le NYSE.

0.1 Marché Financier

0.1.1 Organisé

Un marché *organisé*, ou *bourse*, est une plateforme permettant aux acheteurs et aux vendeurs de produits financiers d'effectuer des transactions, d'exécuter des ordres de bourses, via un intermédiaire qui transmet l'ordre à un membre officiel de la bourse. La *chambre de compensation* est un pilier central dans la bourse, elle intervient comme contrepartie entre acheteurs et vendeurs en garantissant la bonne fin des opérations. Il ne s'agit ni plus ni moins que de l'acheteur de tous les vendeurs et du vendeur de tous les acheteurs. L'établissement Clearnet s'occupe du bon fonctionnement de l'ensemble des bourses européennes (Euronext).

0.1.2 De Gré à Gré

Un marché de *gré à gré* ou *Other-The-Counter* (OTC) est un marché sur lequel les transactions se font directement entre un acheteur et un vendeur. La principale différence avec la bourse est donc qu'il n'existe pas de chambre de compensation permettant d'éviter le risque de contrepartie. Les raisons de l'existence de cette place sont multiples. Notons déjà qu'une partie des sociétés qui y sont présentes le sont car elles sont trop petites, ne remplissent pas les critères pour être cotées sur les marchés organisés. De plus, comme les contrats sont établis directement entre les deux parties, cela permet d'acheter (ou vendre) un produit correspondant bien plus à ses besoins, comme par exemple pour couvrir un risque de change. Par construction, ce marché est moins transparent que la bourse.

0.1.3 Acteurs

(Presque) n'importe qui peut intervenir sur les marchés financiers, nous pouvons tout de même classifier les intervenants en 3 grandes catégories, les *hedgers*,

les *spéculateurs* et les *market makers*. Pour les deux premières catégories, sans grandes surprises, il s'agit de gestionnaires intervenant respectivement pour se couvrir d'un certain risque et, pour tirer un profit suite à une transaction. Les market makers sont présents pour assurer la liquidité du marché. Ils offrent une contrepartie à tout acheteur ou vendeur sur certains titres souvent illiquides. Ces agents sont en général rémunérés par les bourses qui souhaitent offrir à leur client le plus de liquidité possible.

0.1.4 Coût de Transaction

Si l'on achète ou vend un produit financier, quel qu'il soit, il faut rémunérer les intermédiaires, personne ne travaille gratuitement ! On ne peut pas donner ici de pourcentage précis de coût, cela dépend du volume, du montant du produit financier que l'on traite et également de votre broker, plus on achète, plus le pourcentage va être avantageux. Il faut donc bien se renseigner sur les montants demandés par votre broker et les prendre en compte dans votre stratégie.

0.2 Produit Financier

0.2.1 Action

Une action (*share*, ou *stock*) est un *titre de propriété* d'une entreprise confèrant à son détenteur (l'actionnaire) un droit de vote dans la politique de gestion de l'entreprise ainsi que des dividendes. Il s'agit donc d'une fraction du capital social de l'entreprise émise afin de se financer sur les marchés. La valeur initiale de l'action est déterminée par la société, ensuite, elle évolue sur les marchés en fonction de l'offre et de la demande.

Il existe deux types d'action, *ordinaire* ou *préférentielle*. Le détenteur d'une action préférentielle a un droit de priorité par rapport au détenteur d'une action ordinaire, une société n'a pas le droit de verser des dividendes aux actionnaires ordinaires tant que les privilégiés n'ont pas été rémunérés. La deuxième priorité intervient en cas de liquidation de l'entreprise, le détenteur bénéficiera d'un remboursement prioritaire pour la valeur de chacune de ses actions.

Enormément de facteurs rentrent en compte dans la formation des prix, en premier lieu les fondamentaux de l'entreprise, notons les variations de taux de changes, les annonces économiques/macroéconomiques, les décisions gouvernementales ainsi que les mouvements purement spéculatifs. Un simple ordre passé sur le marché,

disons de vente, quelle qu'en soit la raison, s'il est relativement important, peut entraîner à lui seul un effet de panique et conduire à une vente de la part des autres détenteurs de la même action, on parle alors de *price impact*.

0.2.2 Indice

Un *indice* est un indicateur proposé par une bourse pour évaluer les performances d'un pays, d'un secteur.

Citons quelques exemples parisiens. Le CAC 40 est le principal indice de la bourse de Paris (Euronext Paris), initialement, l'acronyme CAC signifiait 'Compagnie des Agents de Change', aujourd'hui 'Cotation Assistée en Continu'. La valeur de l'indice est une pondération de chacun des 40 titres de sociétés en fonction de leur capitalisation flottante. Sa composition est mise à jour tous les trimestres. Quand une société n'est plus cotée, elle est remplacée par une des valeurs du CAC Next 20 en fonction de la liquidité de son titre, capitalisation boursière, volumes échangés, etc. Enfin, notons que sa cotation est réactualisée toutes les 15 secondes. Le CAC Next 20 regroupe les vingt valeurs dont l'importance suit celle des valeurs composant le CAC 40. Le SBF 120 (Société des Bourses Françaises) est composé du CAC 40 et des 80 valeurs les plus liquides à Paris parmi les 200 premières capitalisations boursières françaises. Il existe également le SBF 250 dont on comprend facilement la composition. Notons tout de même que le SBF 80 est quant à lui les 80 valeurs suivant le CAC 40, donc hors CAC 40.

Parmi les indices internationaux nous pouvons également noter le S&P 500 (Standard & Poors, filiale de McGraw-Hill), composé des 500 plus grandes capitalisations boursières des bourses américaines. Comme précédemment, il existe des variantes, S&P 100, 400 et 600. Et toujours comme précédemment, il est pondéré par la capitalisation boursière de chacune des ses composantes. Pour une plus grande diversité sur les marchés américains on pourra regarder le Russell 1000, 2000, 3000 ou le Wilshire 5000.

Pour voyager un peu dans différents pays, notons que l'indice principal en Allemagne est le DAX (30 valeurs), à Londres le FTSE 100 ('Footsie'), Hong-Kong le Hang Seng Index, Taiwan le TSEC et le SSE Composite Index pour Shangai.

Les indices sont donc donnés par,

$$I(t) = C \sum_{i=1}^{N} \alpha_i x_{i,t} \tag{0.2.1}$$

où N est le nombre de valeurs composant l'indice, $x_{i,t}$ la valeur de l'action i à l'instant t, α_i le nombre d'actions i en circulation et C est une constante de normalisation à une valeur de référence (par exemple 100) à une date bien précise.

On peut conclure cette courte présentation en notant que le DJIA (Dow Jones Industrial Average), indice des 30 plus grosses valeurs du NYSE et du NASDAQ, tout comme le Nikkéi 225 à Tokyo ne sont pas pondérés par la capitalisation boursière de leurs composantes, mais par la valeur des actions. Le DJIA est le plus vieil indice boursier au monde, 1896, il était calculé à la main toutes les heures, on peut donc comprendre en partie pourquoi il est pondéré ainsi.

0.2.3 Obligation

Une *obligation* est un *titre de créance négociable* émis par une société ou une collectivité publique, principalement échangée sur les marchés de gré à gré. Les paramètres rentrant dans le calcul d'une obligation sont la date d'émission, la date d'échéance, le taux d'intérêt, la devise dans laquelle elle est émise et la périodicité du coupon. Les modalités de remboursement et le mode de rémunération des prêteurs sont fixés contractuellement, la rémunération pouvant être fixe ou variable, indexée alors sur un taux d'intérêt et non sur le résultat de l'entreprise. La valeur nominale permet le calcul des intérêts (coupon) que l'emprunteur s'engage à verser à des échéances fixées à l'avance. Le marché obligataire est une des principales raisons justifiant l'importance des agences de rating, dont les trois principales sont S&P, Moody's et Fitch. Le système de notation de S&P va du fameux triple A : AAA pour une forte capacité à rembourser au C pour peu d'espoir de recouvrement et D pour défaut de paiement. On peut classifier ces produits en 4 catégories bien distinctes :

- Un Etat dans sa propre devise, *emprunt d'État*.
- Un Etat dans une autre devise que la sienne, *obligation souveraine*.
- Une entreprise du secteur public, un organisme public, une collectivité locale, *obligation du secteur public*.
- Une entreprise privée, une association, ou tout autre personne morale, *obligation corporate*.

Nous présentons maintenant un peu plus en détail le cas des obligations d'Etat Américaines. Les obligations à plus courte échéance sont les Treasury Bill (T-Bill) avec une maturité allant de 1 mois à 1 an. De la même manière que les obligations zéro-coupons, ils ne versent pas d'intérêts avant l'échéance. Les maturités habi-

tuelles sont 1 mois, 3 mois ou 6 mois. Les T-Bills sont reconnus comme constituant les obligations du trésor les moins risquées aux États-Unis par les investisseurs.

Les Treasury Notes (T-Notes) sont à échéances moyennes 2, 5 et 10 ans et leur rémunération est assurée par un coupon payé tous les six mois au souscripteur. La T-Note à échéance de 10 ans est devenue la valeur la plus fréquemment citée dans les études des performances du marché obligataire américain.

Enfin, les obligations avec la plus grande maturité sont les Treasury Bond (T-Bond), variant entre 10 et 30 ans ils sont également rémunérés par un coupon payé tous les six mois. Le revenu que les détenteurs perçoivent possède en outre l'avantage de n'être imposé qu'au niveau fédéral. Les équivalents Français des obligations à courte, moyenne et longue maturité sont respectivement les BTF (Bons du Trésor à taux Fixe et à intérêt précompté), BTAN (Bons du Trésor à intérêts ANnuels) et les OAT (Obligations Assimilables du Trésor).

0.2.4 Matière Première

Les *commodities*, c'est à dire les matières premières, prennent elles aussi une importante place dans les marchés organisés, tant pour les besoins de gestion des risques que pour leur qualité de valeur refuge. Pour le secteur de l'énergie on pourra citer le baril de pétrole brut, le gaz naturel ou l'électricité ; pour les matériaux précieux ou industriels, l'or, l'argent, le platinium, le cuivre ou le palladium ; pour les grains, le maïs, le soja, l'avoine, le riz ou le blé ; les viandes comme les bovins vivants, bovins d'engraissement, le porc maigre, ou la poitrine de porc ; et autres comme le bois, le coton, le caoutchouc, le jus d'orange, le sucre, le café, le cacao, le lait, etc.

0.2.5 Taux de Change

Nous aurions pu mettre cette section dans le chapitre sur les marchés financiers puisque l'ensemble des *devises* qui sont échangées les unes contre les autres à un certain *taux de change* sont cotées sur le *Forex*, contraction de Foreign Exchange. Ce marché mondial est le deuxième marché financier en terme de volume échangé derrière celui des taux d'intérêts, il est ouvert 24h/24h du fait des décalages horaires entre les bourses européenne, américaine, australienne et japonaise, tout de même fermées le week-end ; fermant à 22h GMT le vendredi à New-York et ouvrant à 22h GMT avec la bourse australienne.

Deux types de devises se distinguent, celle fixée par un Etat, on parle alors d'exotique comme le yuan chinois, et celle flottante, valorisée par l'offre et la

demande sur les marchés, comme l'euro ou le dollar américain.

Un taux de change est toujours exprimé de la manière suivante X/Y=z, en d'autres termes, 1 X vaut z Y. La devise de gauche est la *devise de base* et celle de droite la *devise de contrepartie*. Les taux de changes varient jusqu'à la quatrième décimale, ce que l'on appelle le *pip*.

0.2.6 Tracker

Un *tracker* a pour but de *répliquer* un produit financier bien précis. Il peut porter sur un indice actions, obligataire ou encore de matières premières. Si l'on prend l'indice CAC 40, il n'est bien sûr pas possible d'acheter du CAC 40, on pourrait quand même avoir en portefeuille toutes les composantes du CAC 40 en prenant garde de bien réactualiser son portefeuille suite aux sorties et rentrées dans le CAC, d'avoir le même montant que la pondération, cela peut s'avérer bien compliqué. L'achat d'un tracker sur le CAC 40, par exemple le Lyxor ETF CAC 40 permet de simplifier tout cela en achetant un seul produit, au passage cela diminue forcément les coûts de commissions, les *fees*. Le SPDR S&P 500(Standard & Poor's Depositary Receipts), communément appelé Spyder S&P 500, est naturellement le tracker sur le S&P 500 et le SPDR Gold Trust qui réplique la performance de l'or physique sont les deux trackers les plus échangés au monde avec des encours respectifs de 76.5 et 77.5 milliards de dollars au 08/11. Les gérants d'un fond indiciel (Exchange traded funds) sont tenus d'avoir en portefeuille toutes les composantes de l'indice en les pondérant, et pour le cas de l'or, de stocker l'équivalent en or dans leurs coffres.

0.2.7 Forward et Future

Les contrats *forward* et *future* sont des *contrats à termes*, ils s'engagent fermement à acheter ou à vendre à une échéance définie $t = T$, à un prix fixé, une certaine quantité d'un actif financier le *sous-jacent*, à la date $t = 0$. La différence entre un contrat forward et un contrat future est que le premier est OTC tandis que le second s'échange sur les marchés organisés. Le 'sûr-mesure' proposé par le marché de gré à gré s'expose naturellement au risque de liquidité. Le sous-jacent du contrat peut être physique, or, pétrole, etc., une action, un indice, taux d'intérêt, devises, obligations. Il s'agit initialement d'un produit destiné à la gestion des risques.

0.2.8 Option

Les options les plus connues sont le *call* et le *put européens*, donnant respectivement le droit d'acheter et le droit de vendre un actif sous-jacent à une date déterminée à l'avance, moyennant le versement d'une prime. La différence entre ces deux types de produits financiers avec les forward ou future est qu'ils confèrent le droit, et non l'obligation d'exercer son contrat à maturité. Ces deux premiers exemples sont dit *vanille*. Les options *exotiques* sont plus complexes et la liste complète serait trop longue à citer ici. Rapidement, les options *américaines* qui donnent le droit d'exercer son option à n'importe quelle date entre le moment d'achat et la maturité ; l'option *bermuda* qui donne le droit d'exercer son contrat à plusieurs dates bien fixées ; une option asiatique paye la valeur moyenne du sous jacent durant un certain intervalle de temps ; option *lookback* paye n'importe quel prix atteint par l'action au cours de la période, on choisira naturellement celui qui maximise le gain ; option *as-you-like-this*, donne le droit de choisir si elle deviendra un call ou un put.

Première partie

Analyse Statistique des Données Financières

Chapitre 1

Comportement Empirique des Marchés

1.1 Faits Stylisés

Pour pouvoir construire des modèles adaptés à la réalité des marchés financiers, il faut étudier comment les différents actifs évoluent au cour du temps pour ne pas prendre des positions issues de formules inadaptées. Nous allons donc présenter dans cette partie ce que l'on appelle les faits stylisés des marchés financiers, il s'agit des caractéristiques propres aux séries temporelles financières.

Rappelons déjà que l'étude appronfondie des marchés financiers et des premiers modèles proposés a été initiée par Louis Bachelier en 1900 dans sa thèse *Théorie de la Spéculation* sous la direction d'Henri Poincaré. Cette thèse, pourtant pilier fondateur de la théorie moderne des mathématiques financières est passée inaperçue en son temps avant d'être dépoussiérée par de grand noms, Samuelson, Black et Scholes dans les années 60. Aujourd'hui très critiquée par ses hypothèses non réalistes, c'est à dire que les prix boursiers suivent une loi gaussienne. Rappelons que l'étude de L. Bachelier portait sur des données antérieures aux années 1900. Nous allons voir quelles sont les caractéristiques des séries de prix à différentes fréquences pour voir quelles hypothèses sont vérifiées ou non.

Pendant toute la suite du cours nous noterons $(p(t); t \geq 0)$ le cours d'un actif et $p(t) = \ln X(t)$ le log-prix. Les variations financières, i.e. les rendements, s'écrivent $\frac{p(t+1)-p(t)}{p(t)}$, et les log-rendements, $\delta_1 X(t) = X(t+1) - X(t)$ et de manière plus

générale,

$$\delta_\tau X(t) = X(t+\tau) - X(t) = \log\left(\frac{p(t+\tau)}{p(t)}\right) \tag{1.1.1}$$

La donnée la plus 'grossière' que l'on peut avoir sur les marchés financiers est la 'barre', c'est à dire l'information de l'open, high, low, close, où,

$$O(t) \equiv p(t-\delta t) \quad H(t) \equiv \max_t p(t-\delta t, t], \quad L(t) \equiv \min_t p(t-\delta t, t], \quad C(t) \equiv p(t)$$

typiquement, δt est une journée (9h30-16h00 aux Etats Unis), un mois, une heure, une seconde, etc. Nous pouvons la réprésenter comme sur la figure (1.1).

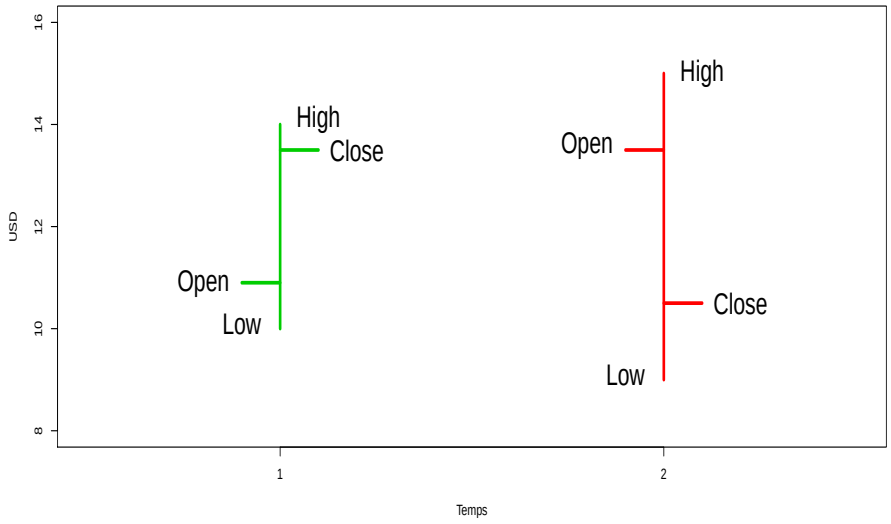


FIGURE 1.1 – Représentation des variations d'une série financière sur deux pas de temps.

Avant de pouvoir rentrer dans le vif du sujet, c'est-à-dire les données hautes fréquences, il convient de s'intéresser aux données basses fréquences, que nous considérons d'observation supérieure à un jour. Pour toute cette partie nous allons nous intéresser au cours de l'indice S&P 500 en données journalières ou inférieures, (1.2, 1.3, 1.4). Même si les propriétés que nous allons mettre en valeurs sont vérifiées par la plupart des actifs financiers, il ne faudra pas non plus généraliser.

FIGURE 1.2 – Cours du S&P 500 du 13/07/1998 au 16/03/2012, 3444 points.

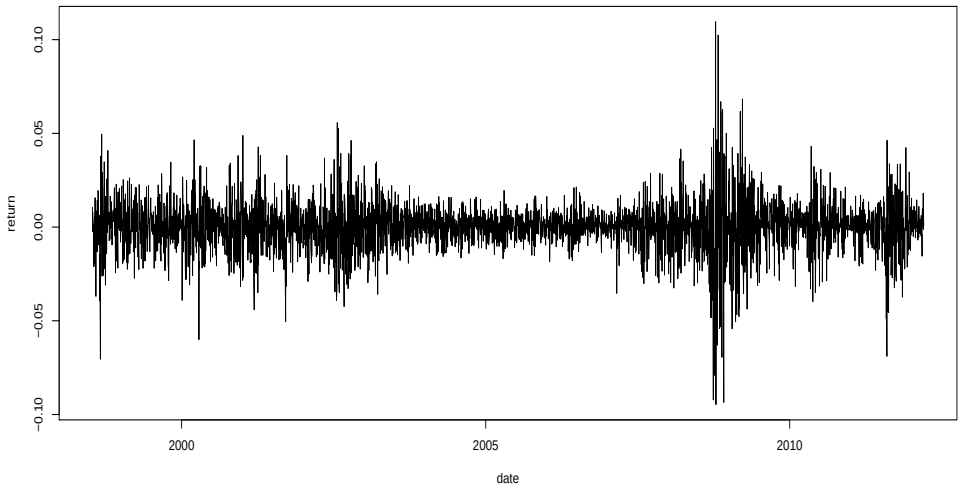


FIGURE 1.3 – Log-rendements du S&P 500 du 13/07/1998 au 16/03/2012.

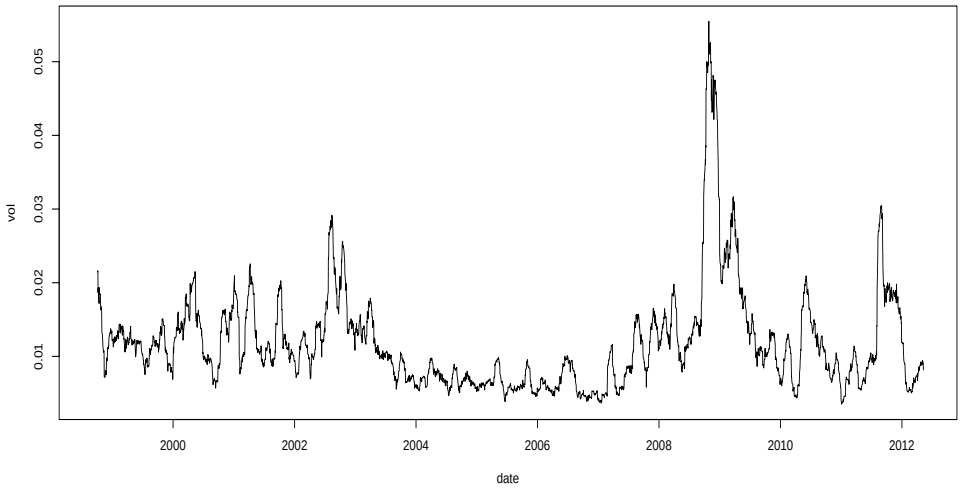


FIGURE 1.4 – Volatilité du S&P 500 du 13/07/1998 au 16/03/2012.

Une caractéristique (plus ou moins) simple à vérifier est la distribution des rendements, suivent-ils une loi gaussienne par exemple ? Pour pouvoir identifier si tel ou tel produit financier se rapproche au mieux de telle ou telle loi de probabilité, on applique un *test d'adéquation.*

Proposition 1.1.1 (Test de Kolmogorov-Smirnov)**.** *Le test de Kolmogorov-Smirnov s'appuie sur la distance D définie par,*

$$D_n = \sqrt{n} \max_x |F_n(x) - F(x)| \tag{1.1.2}$$

où $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x_i \leq x}$ est la fonction de répartition empirique de la série $(x_1, \cdots, x_n)$ et F est la fonction de répartition à laquelle on souhaite comparer notre échantillon. Sous l'hypothèse $H_0 : F_n = F$, la statistique D_n converge vers la loi de Kolmogorov-Smirnov,

$$\mathbb{P}(D_n > c) = 2\sum_{n=1}^{\infty}(-1)^{n-1}e^{-2n^2c^2} \tag{1.1.3}$$

Par la construction de D_n, on voit bien que le test est happé par les valeurs centrales et que les queues de distributions ne sont pas bien prises en compte par ce test.

Proposition 1.1.2 (qq-plot)**.** *Soit* $(Y_t, 0 < t \leq n)$ *une série d'observations. On souhaite vérifier si elle suit une certaine loi* $\mathcal{L}$ *de paramètres* $\theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d$*. On génère un échantillon* $(x_1, x_2, \cdots, x_n)$ *de loi* $\mathcal{L}$ *et de même paramètres* θ *que ceux trouvés sur l'échantillon* $(y_1, y_2, \cdots, y_n)$*. Le qq-plot est le graphique formé par les couples* $Q_j = (q_{x_j}, q_{y_j})$ *où* q_{u_j} *est le quantile à l'ordre* j *de la variable* u*. Si les deux séries ont la même loi sous jacente, alors les* Q_j *décrivent une droite. Pour faire simple, il s'agit simplement du graphe des quantiles de l'échantillon observé sur un axe et des quantiles d'une loi de probabilité bien choisie sur l'autre axe. Naturellement, si les deux lois sous jacentes aux observations sont les mêmes, nous devrions avoir une droite comme graphe.*

On présente le qq-plot des log-rendements du S&P 500 sur la figure (1.1) en prenant la loi normale comme loi de référence.

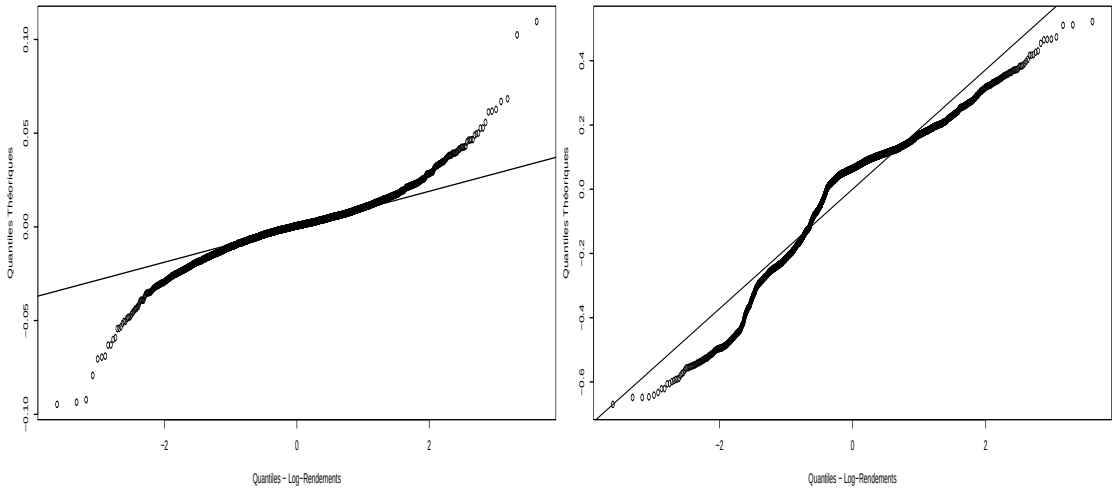


FIGURE 1.5 – qq-plot des log-rendements du S&P 500 du 13/07/1998 au 16/03/2012, à gauche avec $\tau = 1$ et à droite avec $\tau = 251$, soit les rendements sur des lags d'une année. Le trait plein est la droite d'Henry, si les observations suivaient une loi gaussienne, les quantiles devraient être par dessus cette droite. On constate donc que plus la fréquence choisie est basse, plus la distribution des rendements se rapproche de la loi normale, un 'bon' modèle devra donc avoir cette propriété.

Un autre moyen de regarder la distribution des rendements financiers est sim-

plement de tracer la distribution empirique (1.1),

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i<x} \tag{1.1.4}$$

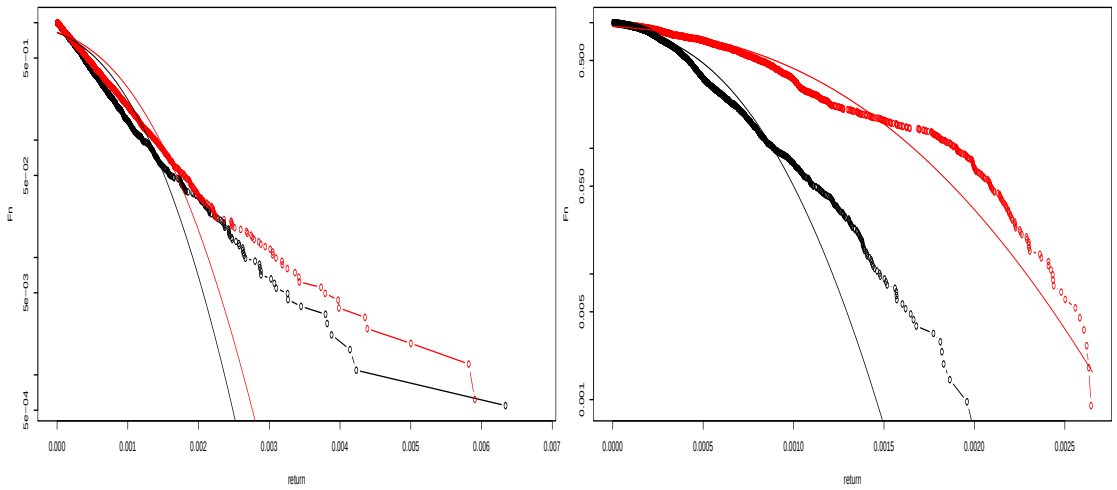


FIGURE 1.6 – Distribution cumulée empirique, $\mathbb{P}(\delta_\tau X \geq x)$, des log-rendements du S&P 500, du du 13/07/1998 au 16/03/2012, échelle log, à gauche $\tau = 1$, à droite $\tau = 251$. En noir les rendements positifs, en rouge les rendements négatifs en valeurs absolues. Trait plein pour le fit gaussien.

Avant d'introduire les notions de skewness et de kurtosis, faisons quelques rappels de probabilité.

Les moments d'ordre n, i.e. la moyenne de X exposant n sont données par,

$$m_n \equiv \mathbb{E}(X^n) = \int_\Omega x^n f(x)\mathrm{d}x. \tag{1.1.5}$$

La connaissance des moments d'une série nous permet d'en déduire sa densité de probabilité (nous faisons un 'gros' raccourci car pour certaines lois, les moments peuvent coïncider). Néanmoins, l'estimation des moments pour des ordres n grand n'est faisable que si nous avons des échantillons suffisamment grands, ce qui n'est

pas forcément le cas pour des historiques de données financières.

La fonction caractéristique $\phi(t)$, définie comme la transformée de Fourier de $f(t)$,

$$\phi(t) \equiv \mathbb{E}(e^{itx}) = \int_\Omega f(x)e^{itx}\mathrm{d}x. \tag{1.1.6}$$

Inversement, connaissant la fonction caractéristique de f nous pouvons retrouver la densité à l'aide de la transformée de Fourier inverse,

$$f(x) = \frac{1}{2\pi}\int_\Omega \phi(x)e^{-itx}\mathrm{d}x. \tag{1.1.7}$$

Une propriété intéressante est le lien qui existe avec les moments,

$$m_n = (-i)^n \frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi(t)|_{t=0}. \tag{1.1.8}$$

Sous quelques conditions et l'existence de tous les moments, nous pouvons écrire,

$$\phi(k) = \sum_n \frac{m_n}{n!}(ik)^n. \tag{1.1.9}$$

$\phi(x)$ est appelée première fonction caractéristique, la deuxième fonction caractéristique est $\ln\phi(x)$ et vérifie

$$c_n = (-i)^n \frac{\mathrm{d}^n}{\mathrm{d}t^n}\ln\phi(t)|_{t=0}, \tag{1.1.10}$$

où c_n est le cumulant d'ordre n. Nous pouvons donc écrire la fonction caractéristique par,

$$\phi(x) = \exp\left[\sum_n \frac{c_n}{n!}(ik)^n\right]. \tag{1.1.11}$$

Les quatres premiers cumulants sont donnés par,

$$\begin{aligned} c_1 &= m_1 \\ c_2 &= m_2 - m_1^2 \\ c_3 &= m_3 - 3m_2m_1 + 2m_1^3 \\ c_4 &= m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4. \end{aligned} \tag{1.1.12}$$

Le cumulant d'ordre 2 n'est autre que la mesure de déviation des observations, la variance,

$$\sigma^2 \equiv \mathbb{E}[(X - \mathbb{E}(X))^2] = m_2 - m_1^2. \tag{1.1.13}$$

Les cumulants normalisés sont définis par,

$$\lambda_n = \frac{c_n}{\sigma_n}. \tag{1.1.14}$$

La skewness, ou coefficient de dissymétrie d'une variable aléatoire, est le cumulant normalisé d'ordre 3,

$$\gamma_1 \equiv \frac{c_3}{\sigma^3} = \frac{\mathbb{E}[(X - \mathbb{E}(X))^3]}{\sigma^3}. \tag{1.1.15}$$

Si $\gamma_1 > 0$, alors la distribution sera étalée vers la droite, ramener à notre problématique de marché financier, cela revient à dire que si les rendements d'un actif financier ont une skweness positive, alors les rendements positifs sont plus fréquents. Inversement, si $\gamma_1 < 0$, alors les rendements négatifs sont plus fréquents. Notons pour que la loi qui nous servira de 'référence', le skewness est égal à 0.

Le degré de kurtosis d'une variable aléatoire X est le cumulant standardisé d'ordre 4,

$$\gamma_2 \equiv \frac{c_4}{\sigma^4} = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3. \tag{1.1.16}$$

Cette quantité mesure l'épaisseur des queues de distributions. Ce résultat s'interprète facilement en comparaison de loi normale pour qui $\gamma_2 = 0$. Si la variable aléatoire à des queues de distribution plus épaisses comparativement à la loi normale, $\gamma_2 > 0$ et on parlera de distribution leptokurtique, inversement, si $\gamma_2 < 0$, la distribution aura des queues plus fines que la loi normale, et elle sera dite platikurtique. Notons que la loi normale est dite mesokurtique. Donc si les rendements d'un actif sont tels que $\gamma_2 > 0$ cela veut dire que l'on observe 'beaucoup' d'évènements extrêmes, à la différence d'une toute douce loi normale, et inversement, si $\gamma_2 < 0$, les rendements sont très concentrés.

Empiriquement, les marchés financiers ont un degré de kurtosis supérieur à celui de la loi normale, les queues de distribution des rendements sont plus prononcées que pour la loi normale. La skewness est quant à elle négative, ce qui correspond à la dissymétrie dans les marchés.

Concluons, la modélisation des rendements financiers par une loi gaussienne n'est pas appropriée. Nous venons de le montrer par plusieurs calculs et graphes, il y avait beaucoup plus simple pour s'en rendre compte. Comme on a pu le voir précédemment, le skewness et le degré de kurtosis d'un loi gaussienne sont tous deux égaux à 0, ni assymétrie, ni queues épaisses. Les déviations à plus de deux fois l'écart type, c'est-à-dire $m \pm 2\sigma$ sont de l'ordre de 2.2% de chaque coté, la probabilité d'avoir une réalisation à plus ou moins 5 fois l'écart type est de 5.7×10^{-7}% et 10 fois de l'ordre de 1.5×10^{-23}%, soit 'quasiment jamais' observable si l'on parle en terme de jour d'une vie humaine. En fait, voir un crash boursier est possible avec une loi normale, à condition d'avoir bien sûr, eut la chance de vivre à cet instant qui ne se produit pourtant qu'une fois par million d'années et pourtant, les crash boursiers sont monnaie courante.

Définition 1.1.1 (Autocorrélation)**.** *La fonction d'autocorrélation est la corrélation croisée d'une série temporelle Y avec elle-même, c'est à dire de $Y(t)$ avec $Y(t+h)$,*

$$C(h) = \mathbb{C}or(Y(t), Y(t+h)) \,. \tag{1.1.17}$$

Pourquoi s'intéresser à cette quantité ? Si $C(h) \approx 0$ pour tout $h > p+1$, alors nous devrions pouvoir écrire,

$$\delta_\tau X(t) \sim a_0 + a_1 \delta_\tau X(t-1) + \cdots + a_p \delta_\tau X(t-p), \tag{1.1.18}$$

il s'agit d'un processus Auto-Régressif d'ordre p, AR(p). En revanche, si $C(h) \approx 0$ pour tout h, alors la modélisation par un processus AR n'est pas tout à fait adaptée puisque nous regardons uniquement comme variable explicative le retard en $t-1$ du cours alors que l'autocorrélation nous indique que les cours à t et $t-1$ sont décorrélés. Idem si nous souhaitons modéliser les rendements par un réseau de neurones, un arbre de décision ou autre (voir section (3.1) et (3.3)), il n'est pas nécessaire de prendre les retards des rendements comme variable explicative.

Définition 1.1.2 (Dépendance de long terme)**.** *La série temporelle Y possède une dépendance de long terme si sa fonction d'autocorrélation diminue en suivant une loi puissance de lag h,*

$$C(h) \sim \frac{L(h)}{h^{1-2d}} \qquad 0 < d < \frac{1}{2}, \tag{1.1.19}$$

où L est une fonction variant lentement à l'infinie, $\lim_{t\to\infty} \frac{L(at)}{L(t)} = 1$ *pour tout* $a > 0$.

Inversement, on parle de dépendance de court terme si la fonction d'autocorrélation décroît de manière géométrique,

$$\exists K > 0, c \in (0,1), |C(h)| \leq Kc^h. \tag{1.1.20}$$

Les études empiriques montrent largement que les rendements ne présentent pas d'autocorrélation significative quel-que soit le lag, cela est visible sur la figure (1.3) des rendements du S&P 500. Les amats de points, i.e. les clusters, nous indiquent qu'une forte de variation est généralement suivie d'une forte variation, mais pas nécessairement dans le même sens, d'où l'observation de clusters. Toujours pour les mêmes raisons, comme les fortes variations sont suivies de fortes variations, nous devrions observer un certaine persistance dans les rendements au carré, i.e. la volatilité, et bien qu'elle ne soit pas tracée ici, l'autocorrélation entre les rendements en valeurs absolues est encore plus plus marquée. Nous présentons l'autocorrélation des rendements et de la volatilité de notre référence sur la figure (1.13)

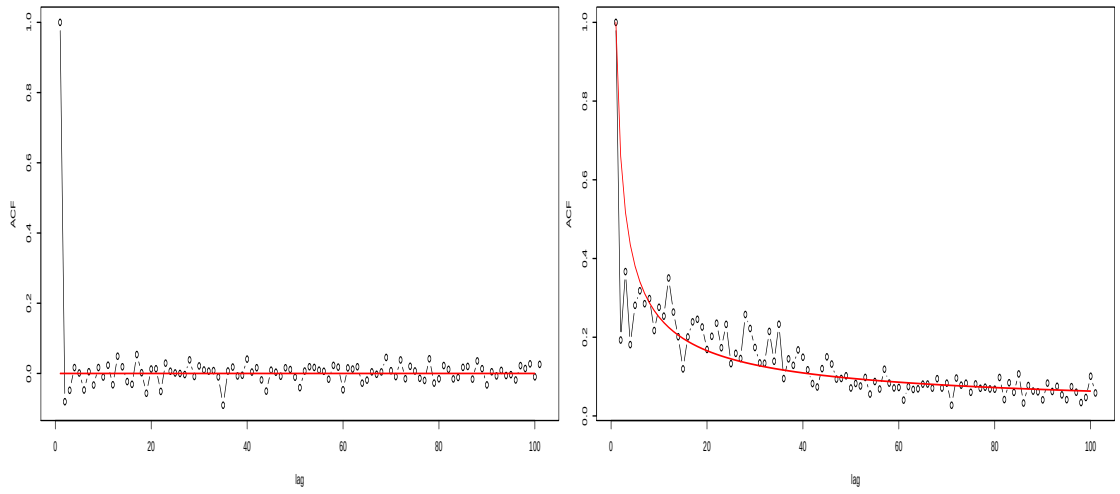


FIGURE 1.7 – autocorrélation des log-rendements du S&P 500 du 13/07/1998 au 16/03/2012, à gauche et de la volatilité à droite. Le trait plein rouge est le fit obtenu, pour les rendements, $y = 0$, pour la volatilité, $\propto h^{-b}$, $b > 0$. Dans ce cas, les rendements ne sont donc pas autocorrélés, en revanche la volatilité si, on parle de *dépendance de long terme*.

En plus de l'autocorrélation entre les mêmes observations (rendements avec retard des rendements, volatilité avec retard de la volatilité), nous pouvons nous

intéresser à la corrélation entre les rendements à la date t et la volatilité variant de $-T$ à T.

Définition 1.1.3 (Effet de levier)**.** *L'effet de levier (dans cette section) est la corrélation entre les rendements à date fixé et la volatilité à différents instants,*

$$\mathcal{L}(h) = \mathbb{C}or(\delta_\tau X(t), \delta_\tau^2 X(t+h)). \tag{1.1.21}$$

Sur certains actifs, dont les indices financiers comme le S&P 500, nous observons que la corrélation entre les rendements et la volatilité passée est en moyenne nulle. Si la corrélation était positive ou négative, un simple arbitrage en fonction de la volatilité connue serait alors possible, il est donc naturel de ne rien observer. En revanche, la corrélation avec la volatilité future est clairement négative et remonte en suivant une loi exponentielle. Ce phénomène peut s'interpréter comme un effet de panique, plus les rendements s'enfoncent dans le rouge, plus la volatilité à l'instant suivant va être importante (1.8).

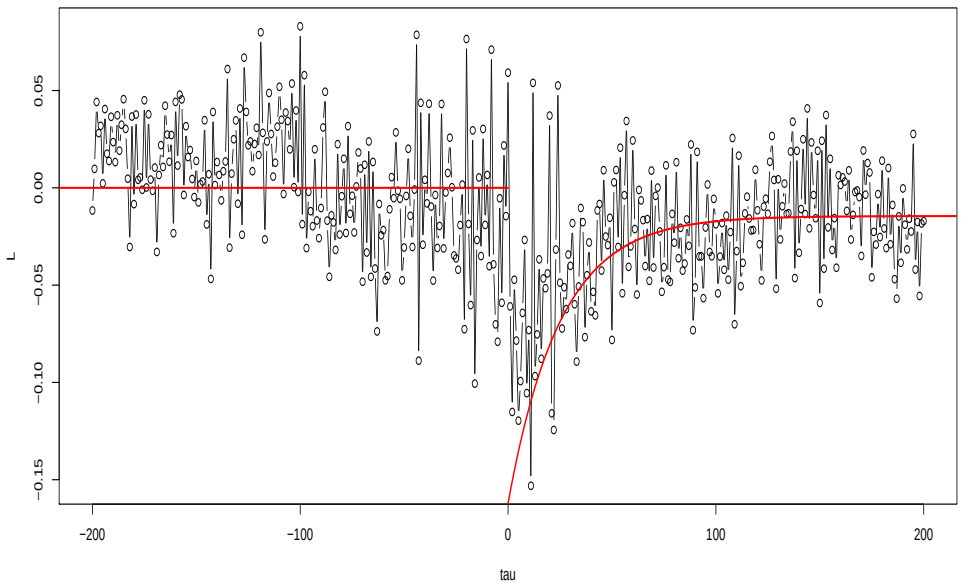


FIGURE 1.8 – Effet de levier, S&P 500 du 13/07/1998 au 16/03/2012.

Pour finir cette section, nous pouvons nous poser la question de comment se comportent les moments empiriques des rendements à différents lags,

$$m(q,\tau) = \mathbb{E}[|\delta_\tau X(t)|^q] = \mathbb{E}[|X(t+\tau) - X(t)|^q]. \tag{1.1.22}$$

Comme nous le montrons sur la figure (1.9), les moments semblent se comporter de la manière,

$$m(q,\tau) \sim C_q \tau^{\zeta(q)}, \tag{1.1.23}$$

où C_q est simplement le moment d'ordre q au lag 1 et $\zeta(q)$ est le spectre (multi) fractal. Nous reviendrons sur ce point dans la section suivante, sur la modélisation des données financières avec le modèle de marche aléatoire multifractale.

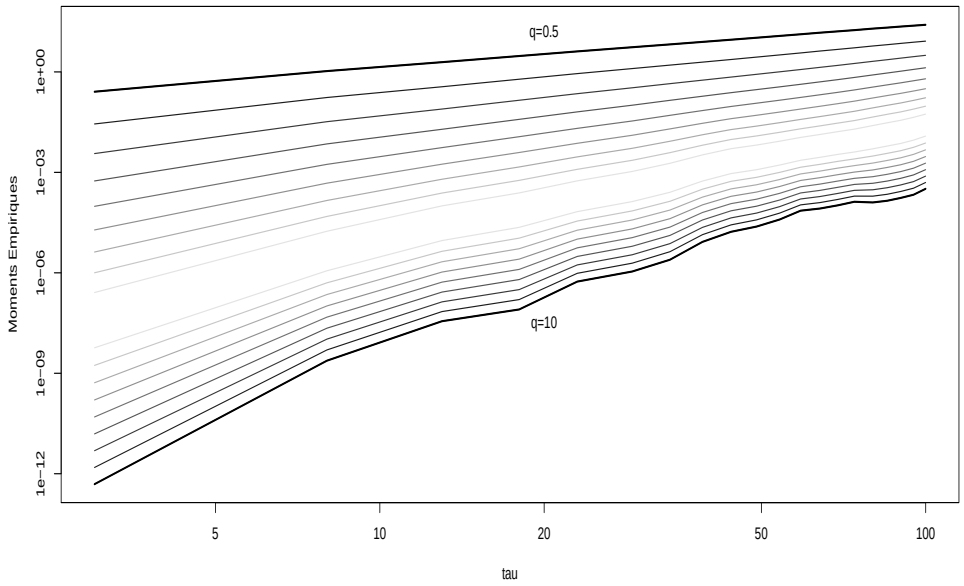


FIGURE 1.9 – Spectre fractal, S&P 500 du 13/07/1998 au 16/03/2012.

Les moments, sont une fonction du lag τ utilisé, on parle d'*invariance d'échelle*. Quelque soit le zoom que nous prenons sur l'observation des données, nous avons la même distribution, par rapport à τ, ce type de particularité est appelé une structure fractale ou multifractale, cela dépend de l'expression de l'exposant $\zeta(q)$.

Comment pouvons-nous interpréter ce comportement ? Les marchés financiers sont *hétérogènes*, c'est-à-dire que les agents ne sont pas tous informés de la même manière au même moment, en contradiction avec l'hypothèse d'E. Fama des marchés financiers efficients. De plus, tous les agents n'ont pas le même comportement, certains ont des profils d'investissement de long terme, d'autres de cour terme, les décisions prises sont donc basées sur des critères, sur des horizons différents. La distribution des rendements change donc avec l'échelle de temps, la fréquence, et se rapprochera d'une loi normale pour de plus grands intervalles.

Pour bien illustrer cette propriété d'invariance d'échelle, il suffit de zoomer au fur et à mesure sur le cours d'un actif financier et voir si les comportements semblent les mêmes, figure (1.10).

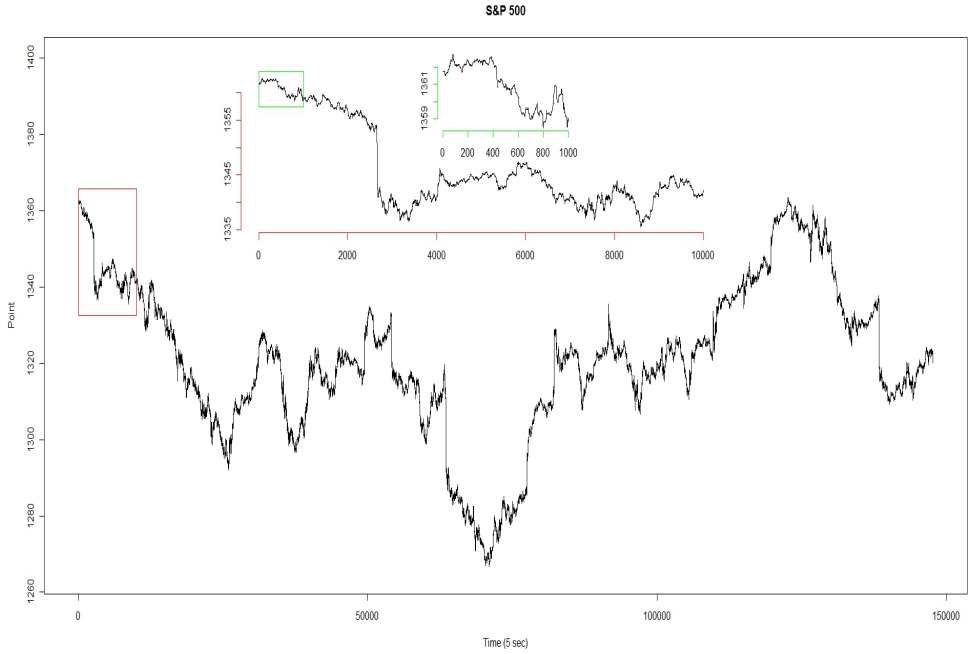


FIGURE 1.10 – Cours du S&P 500 du 11/05/2012, 18 :18 :40, au 26/02/2012, sampling 5 secondes, à l'intérieur, zoom du 11/05/2012 au 15/05/2012, 19 :11 :55, en rouge et en vert de 18 :18 :40 à 19 :41 :55 le 11/05/2012.

Pour un traitement plus complet des faits stylisés, on pourra regarder par exemple [MaSt07], [GoPlAmMeSt99], [BoPo04], [Vo05] pour une revue d'ensemble,

et [GeDaMuOlPi01] ou [ChMuPaAd11] pour des données à hautes fréquences.

1.2 Carnet d'Ordres

Sur les marchés financiers, un carnet d'ordres répertorie l'ensemble de l'offre et de la demande de tous les agents. Ce processus est dit de double enchère continue.

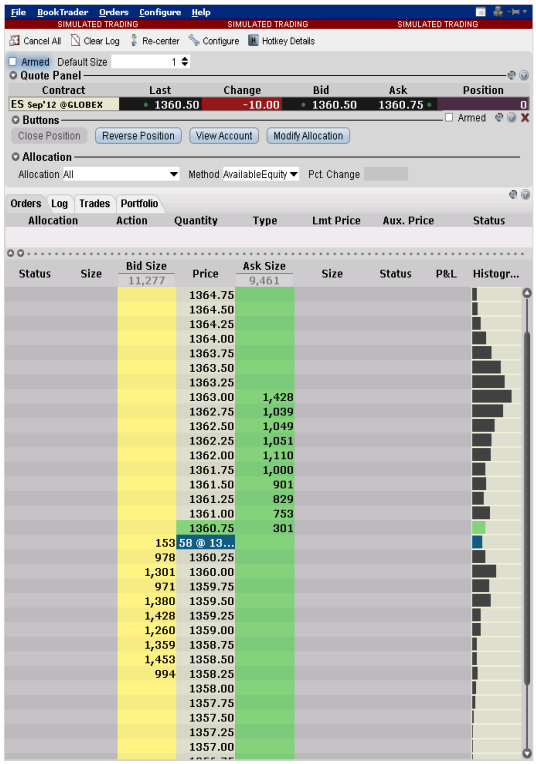


FIGURE 1.11 – Carnet d'ordres du contrat future E-mini le 02/08/2011 vers 17h15.

Un *ordre au marché* est un ordre d'achat/vente pour une certaine quantité d'un actif financier au meilleur prix possible dans le carnet d'ordre. Quand un ordre au marché apparaît, il est matché au meilleur prix disponible dans le carnet d'ordre et la transaction s'effectue. La priorité des ordres vient de l'ancienneté de ceux ci. La quantité disponible dans le carnet s'ajuste en fonction. Il est possible que l'ordre soit effectué en plusieurs fois si le volume donné par la contrepartie est de

taille inférieure.

Un *ordre limite* est un ordre d'achat/vente pour une certaine quantité et pour un prix bien fixé. Il est soit exécuté si une contrepartie est présente dans le carnet, soit, il est annulé. C'est le prix maximum (minimum) auquel l'agent est prêt à acheter (vendre) un volume précis. Il existe plusieurs types d'ordre limite, les options possibles sont immediate or cancel (IOC) et fill or kill (FOK). Un ordre IOC est exécuté instantanément, soit complétement, soit partiellement, soit pas du tout. Le volume demandé qui n'a pas pu être exécuté au prix limite voulu ou à un prix plus avantageux est immédiatement annulé. Un ordre FOK est un ordre IOC mais est exécuté si et seulement si tout le volume demandé est disponible.

D'un autre coté nous avons l'ordre good for a day (GFD), qui comme son nom l'indique, est retiré du carnet à la clôture du marché s'il n'a pas trouvé de contrepartie. Les ordres good till canceled (GTC) et good till date (GTD) sont également proposés. Un ordre GTC n'a pas de durée de validité, néanmoins, le broker peut décider de le retirer du carnet, ou pour la cas d'une option, si elle arrive à maturité, l'ordre sera annulé. L'ordre GTD permet de rentrer une date prédéfinie de cancellation.

Le carnet d'ordres est une capture de l'état du cours d'un actif à un instant donné. Il indique les ordres limites offerts et demandés. Le *ask price* est le plus petit montant dans le carnet auquel les agents sont prêts à vendre, il s'agit du prix auquel un ordre marché d'achat est exécuté. A contrario, le *bid price* est le plus grand montant dans le carnet auquel les agents sont prêts à acheter, il s'agit du prix auquel un ordre marché de vente est exécuté. Chaque nouvel ordre traité va donc se situer à l'extérieur de cet intervalle, sinon, si par exemple un ordre d'achat supérieur au bid price est passé, il devient le bid price, et donc le bid price augmente. Le *last price* est le dernier ordre traité, il se situe donc à l'intérieur de l'intervalle (fermé) du ask et bid price. Les prix ne sont pas cotés de manière continue dans le sens où seulement un multiple du tick est possible,

$$p(t) = k \cdot p_{tick}, \tag{1.2.1}$$

où p est le prix de l'actif, $k \in \mathbb{N}$ et p_{tick} la valeur du tick pour l'actif correspondant, par exemple, le future E-mini S&P 500 a un tick de 0.25, les prix peuvent donc varier uniquement de 0.25 en 0.25. Le taux de change EUR/USD, comme une bonne partie des instruments issus du forex, a un tick de 0.0001, le 24 juillet 2012 vers 18h il est à 1.2089, dans les prochaines secondes il sera peut être à 1.2088 ou

1.2090, mais en aucun cas à 1.20895 par exemple. La différence entre le bid price et le ask price est le spread,

$$s = p_{ask} - p_{bid}, \tag{1.2.2}$$

il est nécessairement positif puisque le ask price, donc le prix d'achat est toujours plus élevé que le prix de vente, le bid price, si l'inverse se produisait, les arbitrageur et market makeur interviendraient immédiatement sur le marché pour faire la correction (notons que le cas $s = 0$ est également non viable puisque dès que $s = 0$, les ordres au marché vont se matcher et le spread va redevenir positif). C'est une quantité usuellement regardée pour déterminer si l'actif en question est liquide ou non. Plus le spread se resserre et tend vers 0, plus il y a de liquidité, un actif très liquide a en moyenne un spread de 1 tick. Enfin, notons que le spread est à inclure dans les coûts de transactions, les fees. Disons que nous avons un spread constant égal à 1 tick et que le ask price vaut x\$, nous achetons donc à x\$ (nous avons acheté au marché), si nous revendons tout de suite, sans attendre de variations, nous revendons au bid price, i.e. $x\$ - 1\text{tick}$, multiplié par le capital investi, K, nous avons perdu $K \times 1\text{tick}$ \$, auquel on rajoute bien sur les frais de passage d'ordres. Avec le broker Interactive Borker, les frais sont de 2.5\$ par ordre si le montant échangé sur le mois représente une valeure inférieure à 10^9\$, $1.5 \cdot 10^{-5} \times K \times x$ si les transactions effectuées en 1 mois ont une valeur comprise entre 10^9\$ et 2×10^9\$ et $1 \cdot 10^{-5} \times K \times x$ si le montant excède 2×10^9\$. Chaque ordre ayant un coût minimum de 2.5\$.

Un autre type d'ordre est l'ordre iceberg, il est caractérisé par un certain degré de discrétion et est particulièrement utile pour placer des volumes importants. Nous avons trois paramètres à définir, le prix limite, le volume total et le peak volume. Le peak volume est la partie visible de l'iceberg dans le carnet d'ordre, chaque fois que le peak volume sera complètement exécuté, un nouveau peak volume va apparaître dans le carnet si le volume caché n'a pas encore été totalement absorbé.

Revenons un peu sur la priorité d'exécution des ordres, il s'agit d'une prioriété temps-prix, il est difficile de dire lequel des deux passe devant l'autre en fait. Supposons qu'au prix limite p^ℓ nous ayons dans le carnet un volume v^ℓ en attente, alors si l'on pose un nouvel ordre de volume k à la même limite, le volume devient $v^\ell + k$, pour être exécuté, nous devons attendre que la quantité v^ℓ soit absorbée, puis ce sera notre tour, il s'agit de la priorité temporelle, la règle de FIFO (first in, first out).

En plus de cette priorité, nous avons la priorité des prix, si nous avons deux prix bid en carnet, l'un à $p^\ell\$$ et l'autre à $p^\ell\$ + 1\text{tick}$, n'importe quel agent va préférer vendre au prix $p^\ell\$ + 1\text{tick}$ plutôt qu'à p^ℓ. Si nous plaçons un ordre d'achat à p^ℓ (un ordre placé donc coté bid en carnet), il faut d'abord que le bid price soit entièrement exécuté, puis le volume correspondant à $p_{bid} - 1\text{tick}$, puis, etc. jusqu'à ce que p^ℓ soit le nouvel ordre marché, le nouveau ask price. Dans tous les cas, si un nouvel ordre limite est placé à l'intérieur du spread, il sera le meilleur prix (que ce soit à l'achat ou à la vente) il a donc la priorité de prix et de temps.

C'est donc ce flux d'ordres qui forme les prix. Le carnet d'ordres est une des informations les plus importantes fournies par le marché puisqu'il nous indique à quel prix les agents sont prêts à exécuter leurs deal. C'est une vision 'microscopique' de l'actif. Un trader discrétionnaire devra faire attention à ne pas avoir un ordre limite trop élévé (pour la vente) au risque de ne pas trouver d'acheteur. Inversement, s'il propose un montant trop bas, il peut passer à coté d'un gain potentiel, mais minimise le risque de ne pas être exécuté.

L'évolution du carnet d'ordres nous indique le degré d'excitation de l'actif à chaque instant : la fréquence d'arrivée des ordres, les volumes et les prix proposés sont autant d'indications. C'est cette information que nous allons essayer d'extraire, de modéliser et, d'en tirer un profit dans la suite de ce cours.

On représente de manière schématique les différents ordres, marché, limite, cancel, et leurs évolutions sur la figure (1.12)

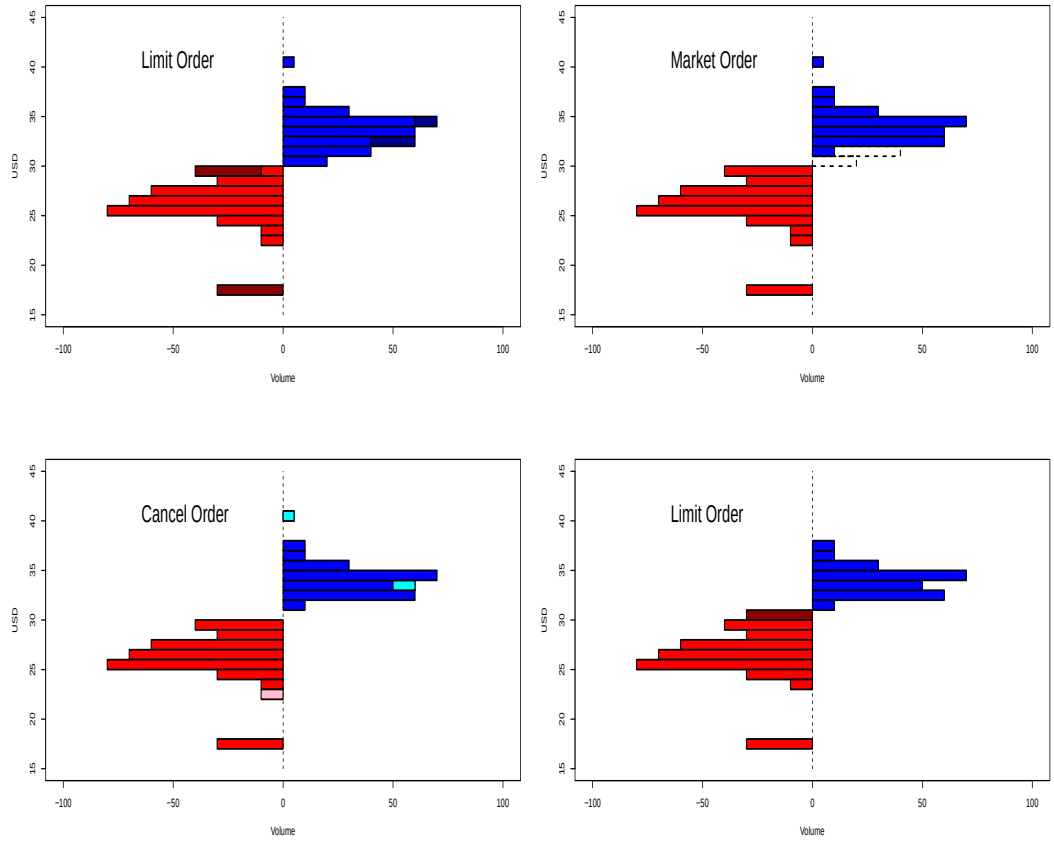


FIGURE 1.12 – Représentation des différents type d'ordre dans un carnet d'ordres, ordre marché, limite et cancel.

1.3 Microstructure

A très haute fréquence, ce qui dans cette partie correspond à la formation des prix, nous pouvons observer des caractéristiques très intéressantes, qui comme nous le verrons, sont très simples et intuitives.

Nous avons déjà (un peu) vu le comportement de la volatilité, à très haute fréquence, ce n'est pas tout à fait identique. A très haute fréquence, la volatilité 'explose', les variations n'étant pas continues mais arrivant de manière ponctuelle, nous avons nécessairement une volatilité qui augmente avec la fréquence. Ce phé-

nomène est usuellement appelé le signature plot. Formellement, le signature plot est défini par la quantité,

$$\mathbb{V}(\tau) = \frac{1}{T} \sum_{s=0}^{\lfloor T/\tau \rfloor} \left(\delta_\tau X(s)\right)^2 . \tag{1.3.1}$$

Plus le lag τ va être élevé, plus cette quantité va décroitre en suivant une loi puissance à cause de la discontuinité des fluctuations boursières à très haute fréquence. Inversement, plus les fluctuations sont observées à des échelles plus lointaines, ce qui conduit à voir des variations plus lisses, plus la volatilité va décroître jusqu'à une 'valeur d'équilibre'. Nous avons présenté cette caractéristique sur la figure **??**.

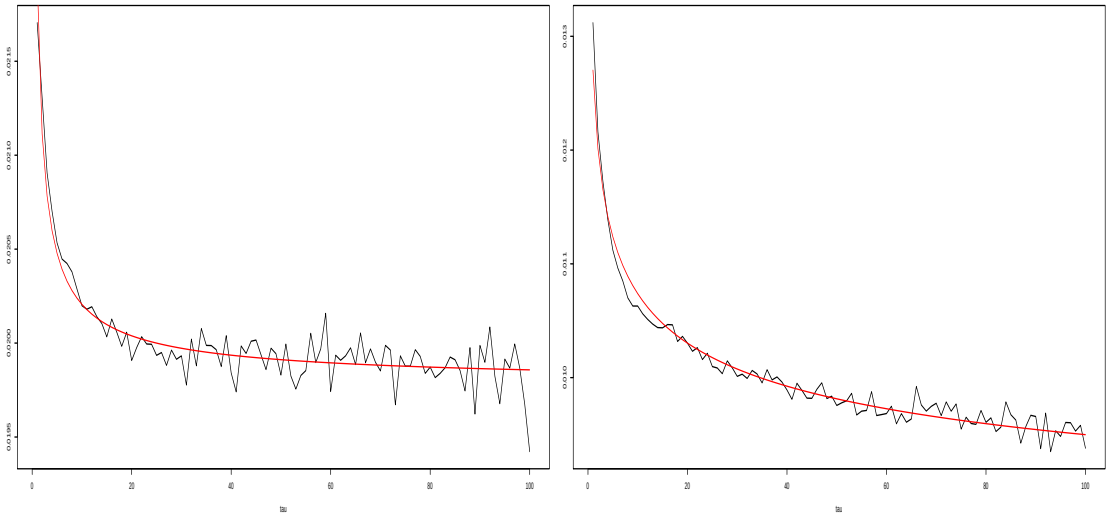


FIGURE 1.13 – Signature plot sur l'EUR/USD (haut), l'EUR/GBP (bas), les données partent de l'échelle 'tick' jusqu'a 100 secondes, du 30/01/12 00 :00 :00 jusqu'à 22h00 le même jour, soit 79200 secondes. En rouge le meilleur fit en loi puissance obtenu.

De la même manière que pour le signature plot, la corrélation entre deux actifs va évoluer selon la fréquence à laquelle nous allons nous placer. Les ordres d'achats et de ventes n'arrivent pas aux mêmes moments, il est très rare de pouvoir observer deux ordres sur les actifs Apple et IBM arriver exactement aux mêmes moments, à la même milliseconde. Ainsi, même si nous observons deux actifs extrêmement liquides, il sera compliqué de voir une relation se former à très haute fréquence.

Plus la fréquence va diminuer, plus la corrélation va se former et va arriver à son 'point d'équilibre'. Il s'agit de l'effet Epps et est quantifié par,

$$\rho_{1,2}(\tau) = \frac{\mathbb{C}\text{o}(X_1, X_2)}{\sqrt{\mathbb{V}_1(\tau)\mathbb{V}_2(\tau)}}, \tag{1.3.2}$$

où $\mathbb{V}_i(\tau)$ est la variation de l'actif i donnée (1.3.1) et la covariation empirique est donnée par,

$$\mathbb{C}\text{o}(X_1, X_2) = \frac{1}{T}\sum_{s=0}^{\lfloor T/\tau \rfloor} \delta_\tau X_1(s)\delta_\tau X_2(s). \tag{1.3.3}$$

Nous présentons sur la figure (1.14) ce phénomène.

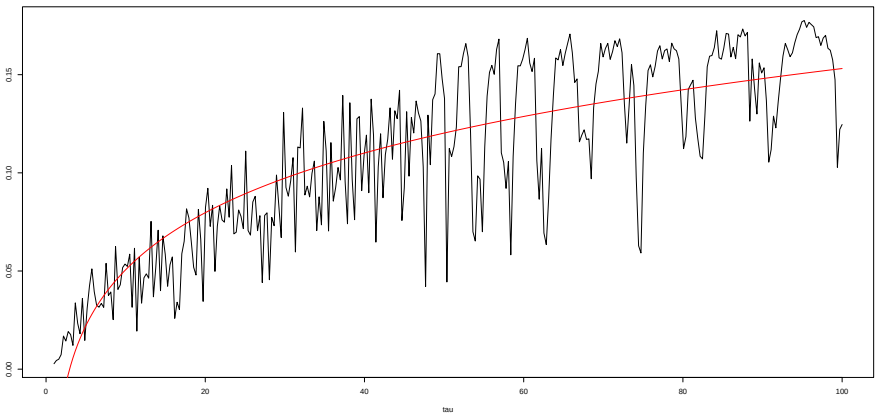


FIGURE 1.14 – Effet Epps sur l'EUR/USD et l'EUR/GBP, les données partent de l'échelle 'tick' jusqu'a 100 secondes, du 30/01/12 00 :00 :00 jusqu'à 22h00 le même jour, soit 79200 secondes.

La dernière propriété de ce type que nous pouvons citer est le lead-lag effect. Quel que soit le sampling sur lequel on se place, mais plus particulièrement à très haute fréquence, nous pouvons exhiber une paire d'actif (X_1, X_2) dont l'un des deux actifs, le lagger, ou suiveur, semble reproduire en partie les variations de l'autre (le leader ou meneur) après un certain lag.

Une propriété sur laquelle nous avons naturellement envie de nous pencher pour avoir une idée des variations est la duration, c'est-à-dire le temps passé entre

deux changements de prix. Schématiquement, si nous arrivons à bien modéliser le temps qu'il faut attendre pour observer un nouvel ordre, nous avons une idée de comment placer des ordres limites pour maximiser ses chances d'être executées. Le temps d'attente d entre deux ordres consécutifs est simplement donné par,

$$d_k = t_{k+1} - t_k, \tag{1.3.4}$$

où les t_i sont les instants d'arrivée des ordres. C'est précisément cette quantité qui nous produit le signature plot et l'effet de Epps. Nous avons présenté la distribution cumulée des durations pour les ordres bid et ask de l'EUR/USD et de l'EUR/GBP que la figure (1.3) ainsi que les fit obtenus avec la loi gaussienne et la loi de Pareto (loi puissance), $f \propto x^{-\alpha}, \alpha > 0$. Pour déterminer l'exposant α de la loi puissance, nous pouvons par exemple utilisé l'estimateur de Hill, $\hat{\alpha} = \frac{1}{k}\sum_{i=1}^{k} \frac{d_{(i)}}{d_{(k)}}$ où $d_{(1)} \geq \cdots \geq d_{(T)}$ est la statistique d'ordre des $d_1, \cdots, d_T$ et k et k est un paramètre à choisir pour obtenir la meilleure approximation des queues épaisses.

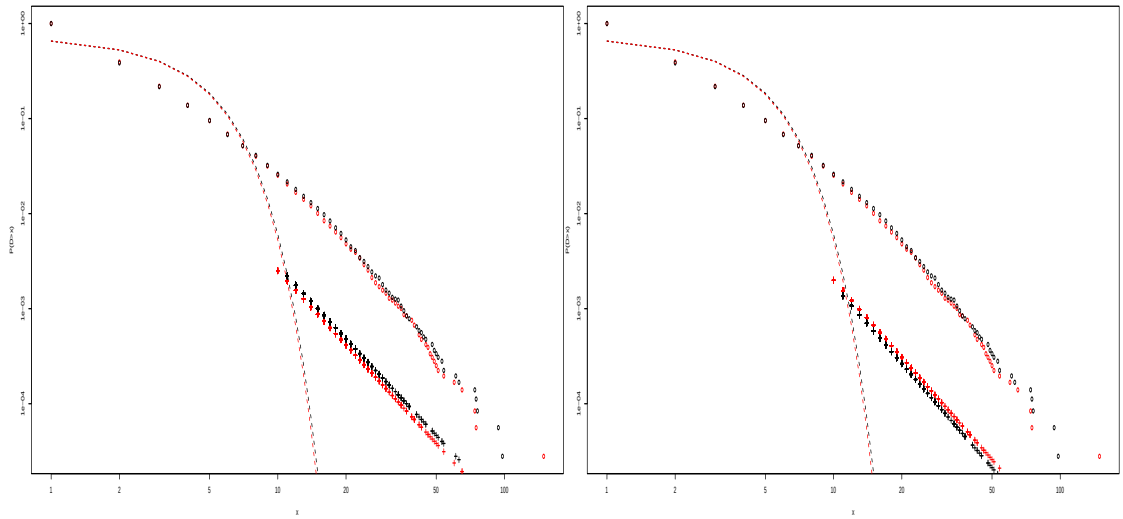


FIGURE 1.15 – Probabilité cumulée des durations pour l'EUR/USD (gauche) et l'EUR/GBP (droite), en rouge les ask prices, et noir les bid prices. Les pointillés sont la prévision par la loi gaussienne, les croix par la loi puissance.

L'autocorrelation est présenté sur (1.16)

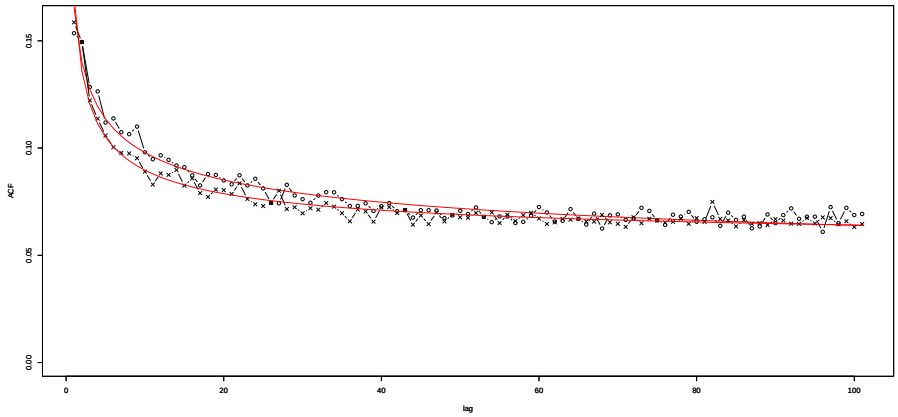


FIGURE 1.16 – Autocorrelation

Enfin, sur la figure (1.17), nous présentons le plot des durations en fonction des volumes échangés.

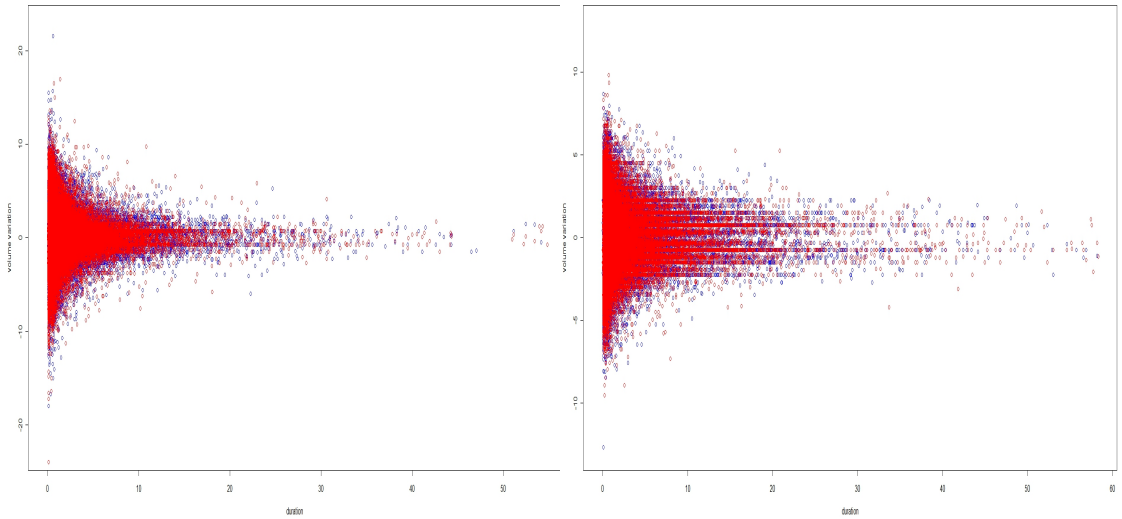


FIGURE 1.17 – Variation du volume en fonction des durations. Les variations sont données en Million d'USD. A gauche l'EUR/USD, à droite, l'EUR/GBP .

Nous renvoyons à [Ha12] pour une étude appronfondie des durations et [Le12] pour une étude sur la microstructure du marché et ses implications, et [BoMePo02],

[MoViMoGeFaVaLiMa09], [ChMuPaAd11] pour le carnet d'ordre et les impacts de marché.

Chapitre 2

Modélisation

La première question à se poser est comment décomposer le cours d'un actif financier en plusieurs éléments susceptibles d'être chacun modélisé.

Un cours boursier suit des tendances (des trends) en fonction des différents fondamentaux du sous-jacent, des différentes croyances des agents, des investisseurs, ne faisons pas l'amalgame dans l'interprétation entre tendance et 'vraie valeur'. La seconde composante fondamentale est la volatilité. La volatilité, comme nous l'avons vu dans la partie précédente, représente le degré de dispersion autour du trend, en d'autre terme, la température du marché. Plus la volatilité est importante, plus les fluctuations vont se faire de manière abrupte. Il s'agit dans un certain sens d'une mesure de risque, plus la volatilité est importante, plus les cours varient rapidement, plus les agents sont susceptibles de perdre leur capital.

Tendance et volatilité sont les deux principales composantes d'un cours boursier, il en existe d'autre bien entendu, on peut par exemple penser à la saisonnalité, celle intraday est particulièrement visible, les variations et les volumes échangés sont beaucoup plus importants au 'bord' de la journée, à l'ouverture et à la fermeture, qu'en plein milieu de la journée. On peut encore penser au 'saut', que l'on peut rajouter aux composantes de saisonnalité, de tendance et de volatilité pour prendre en compte, comme son nom l'indique, les sauts dans les cours boursiers, néanmoins, comme nous le verrons, il existe des modèles permettant d'intégrer ces sauts différemment, sans rajout, mais en changeant de classe de processus.

Notons $(X(t); t \geq 0)$ le cours d'un actif financier, Louis Bachelier dans sa thèse *Théorie de la Spéculation*, [Ba00] propose de modéliser le cours des actifs financiers par un mouvement brownien avec tendance,

$$X(t) = X(0) + \mu t + \sigma B(t), \tag{2.0.1}$$

μt est la tendance du cours (le drift), σ la volatilité et B est un mouvement brownien standard. Nous retrouvons bien nos deux principales composantes. Samuelson raffine cette modélisation en ne l'appliquant non plus aux cours eux-mêmes, mais aux log-rendements financiers,

$$\mathrm{d}X(t) = X(t)\left(\mu \mathrm{d}t + \sigma \mathrm{d}B(t)\right). \tag{2.0.2}$$

Dans ce chapitre, nous allons principalement nous occuper de la partie $\sigma \mathrm{d}B(t)$ plus que des autres. Une possibilité de modélisation et prévision de la tendance ne sera pas présente dans cette partie, mais dans la partie sur l'arbitrage statistique car certains outils mathématiques vont se croiser. Si l'on écrit le modèle de manière simplifiée nous avons,

$$\delta_\tau X(t) = \mu(t) + \varepsilon(t). \tag{2.0.3}$$

Concrètement, ce qui va nous intéresser principalement ici est le second terme, c'est-à-dire $\varepsilon(t)$.

2.1 Modélisation A Partir de l'Historique

Avant de présenter des modèles sophistiqués de prévision de volatilité, nous allons nous intéresser à des modèles beaucoup plus simples, moins 'sexy', mais qui finalement pourraient donner de très bons résultats en terme de prévision de variance.

Sans information particulière, l'idée la plus simple est de supposer que la variance est une marche aléatoire, et la prévision serait,

$$\hat{\sigma}(t+1) = \sigma(t). \tag{2.1.1}$$

Bien sûr, nous pourrions supposer qu'elle ne dépend pas seulement de sa dernière valeur, mais d'un historique plus lointain,

$$\hat{\sigma}(t+1) = \frac{\sigma(t) + \sigma(t-1) + \cdots + \sigma(t-N)}{N}. \tag{2.1.2}$$

On peut aisément se dire que les dernières valeurs ont plus d'importance que les valeurs plus éloignées, la moyenne mobile précédente se dérive alors en moyenne mobile exponentielle (EWMA),

$$\hat{\sigma}(t+1) = (1-\lambda)\sum_{i=0}^{N} \lambda^i \sigma(t-i), \;\; \lambda \in (0,1). \tag{2.1.3}$$

Le terme $(1-\lambda)$ provient du fait que $1+\lambda+\lambda^2+\cdots = (1-\lambda)^{-1}$. Le terme λ est un terme de persistence. Ce modèle est très connu, il fait parti de la suite RiskMetrics de JP Morgan. Un moyen de mieux s'adapter aux réactions de la volatilité avec son historique est de faire une régression linéaire,

$$\hat{\sigma}(t+1) = \alpha + \beta_0\sigma(t) + \beta_1\sigma(t-1) + \beta_2\sigma(t-2) + \cdots . \tag{2.1.4}$$

Intuitivement, cette dernière manière d'estimer la volatilité future en ne regardant que son passé devrait donner de meilleurs résultats car elle est plus adaptative. Néanmoins, le risque principal est de faire sur-apprendre le modèle lors de la détermination des poids (nous reviendrons sur le problème du sur-apprentissage quand nous aborderons l'apprentissage statistique).

2.2 Processus ARCH et Dérivées

Comme nous avons pu le voir dans les graphes précédents, une particularité courante et importante des séries financières est l'observation de *clusters*, i.e. un amas de points concentrés. Intuitivement, nous pouvons interpréter cela par l'arrivée de nouvelles sur le marché, la réaction de ces informations est d'acheter ou vendre 'frénétiquement', et quand la nouvelle a été digérée, le marché revient à la 'normale', à son niveau de volatilité avant la nouvelle. Le cours de l'actif va donc présenter une forte baisse ou hausse, empiriquement principalement des baisses (la finance comportementale peut apporter des réponses à ce phénomène), il s'agit de l'assymétrie des séries, la kurtosis. Cette particularité est appellée *heteroskacadicté*, chaque variation est corrélée à la suivante, on peut résumer cela grossièrement par un effet de mimétisme.

2.2.1 ARCH

La première solution est apportée par Engle [En82] avec le modèle ARCH, *AutoRegressive Conditional Heteroscedasticity*, partant de l'observation de clusters dans les rendements et des variations typiques de la volatilité, l'idée principale consiste à regarder la variance conditionnelle au lieu de la variance non conditionnelle.

La modélisation ARCH prend la forme,

$$\varepsilon(t) = z(t)h^{1/2}(t), \tag{2.2.1}$$

où $z(t)$ est un bruit blanc faible centré (nous allons le supposer gaussien, mais ce n'est pas un impératif, dans la pratique, il faut mieux prendre une distribution de Student ou GED par exemple). $h(t)$, la variance (conditionnelle) du processus qui elle-même est un processus autorégressif,

$$h(t) = \alpha_0 + \alpha_1\varepsilon^2(t-1) + \cdots + \alpha_q\varepsilon^2(t-q), \tag{2.2.2}$$

la variance est ainsi décrite comme fonction du carré du cours de l'actif, soit déterministe conditionnellement à l'historique $\mathcal{H}(t-1) = \{\varepsilon(t-1), \cdots, \varepsilon(t-q)\}$. La modélisation ARCH et les nombreux dérivés qui vont suivrent, ne permettent donc en rien de prévoir le cours d'un actif à l'instant suivant, uniquement sa volatilité. Par construction, et sous l'hypothèse de normalité de z, nous pouvons réécrire le modèle,

$$\begin{aligned} &\varepsilon(t)|\mathcal{H}(t-1) \sim \mathcal{N}(0, h(t)) \\ &h(t) = \alpha_0 + \alpha_1\varepsilon^2(t-1) + \cdots + \alpha_q\varepsilon^2(t-q). \end{aligned} \tag{2.2.3}$$

La première propriété intéressante à remarquer est que

$$\mathbb{E}[\varepsilon(t)|\mathcal{H}(t-1)] = h^{1/2}(t)\mathbb{E}[z(t)|\mathcal{H}(t-1)] = 0. \tag{2.2.4}$$

Par la règle des espérances itérées, on trouve également que $\mathbb{E}[\varepsilon(t)] = 0$. La variance conditionnelle de $\varepsilon(t)$ est quant à elle donnée par,

$$\begin{aligned} \mathbb{V}\text{ar}(\varepsilon(t)|\mathcal{H}(t-1)) &= \mathbb{E}[(\varepsilon(t) - \mathbb{E}[\varepsilon(t)|\mathcal{H}(t-1)])^2|\mathcal{H}(t-1)] \\ &= \mathbb{E}[\varepsilon^2(t)|\mathcal{H}(t-1)] \\ &= h(t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i\varepsilon^2(t-i). \end{aligned} \tag{2.2.5}$$

La variance conditionnelle évolue donc selon un processus auto-régressif. Après quelques calculs simples on peut également montrer que $\gamma_2 > 0$ (leptokurtik) et que $\mathbb{C}\text{ov}(\varepsilon(t), \varepsilon(t+h)|\mathcal{H}(t-1)) = 0$, donc cette représentation a le mérite de vérifier les caractéristiques de queues épaisses et d'absence d'autocorrélation des rendements financiers.

Pour pouvoir appliquer cette modélisation par exemple, pour prévoir, la volatilité d'un actif financier, nous avons besoin d'estimer les paramètres du modèle, ce que l'on appelle la *calibration*. Une des méthodes les plus courantes est celle du maximum de vraisemblance, en notant α l'ensemble des paramètres du modèle, $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_p)$ nous avons sous l'hypothèse de normalité de z,

$$\begin{aligned} L(\varepsilon(1), \cdots, \varepsilon(T); \alpha) &= \prod_{t=1}^{T} \ell(\varepsilon(t)) \\ &= \prod_{t=1}^{T} \left(\frac{1}{\sqrt{2\pi h(t)}} \exp\left\{ -\frac{1}{2} \frac{\varepsilon^2(t)}{h(t)} \right\} \right). \end{aligned} \tag{2.2.6}$$

La log vraisemblance est alors,

$$\begin{aligned} \log L(\varepsilon(1), \cdots, \varepsilon(T); \alpha) &= \sum_{t=1}^{T} \left(\log(\sqrt{2\pi h(t)}) - \frac{1}{2} \frac{\varepsilon^2(t)}{h(t)} \right) \\ &= -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} (\log(h(t))) - \frac{1}{2} \sum_{t=1}^{T} \left(\frac{\varepsilon^2(t)}{h(t)} \right). \end{aligned} \tag{2.2.7}$$

Les estimations $\hat{\alpha}_i$, $i = 1, \cdots, p+1$ sont données par maximisation de (2.2.7),

$$\hat{\alpha} = \arg\min_{\hat{\alpha}} \{ \log L(\varepsilon(1), \cdots, \varepsilon(T); \alpha) \}, \tag{2.2.8}$$

La maximisation se fait classiquement par un algorithme d'optimisation type BFGS (voir appendix). En dehors de la faiblesse du modèle à refléter correctement les faits stylilisés, le choix de l'ordre p n'est pas trivial. La méthode la plus courante est de minimiser l'AIC (Akaike Information Criterion),

$$AIC = 2(p+1) - 2 \log L. \tag{2.2.9}$$

Dans la pratique, ce n'est pas aussi évident et le critère AIC amène bien souvent à prendre un nombre de lag trop élevé. Le modèle GARCH(p, q) *Generalized Auto-Regressive Conditional Heteroskedasticity* proposé par Bollerslev en 1986 [Bo86] a, entre autre, le mérite de pallier à ce problème, il rend également le modèle plus réaliste. Les modèles GARCH sont très utilisés en finance et ont inspiré énormément de dérivés plus sophistiqués, nous en verrons quelques uns.

2.2.2 GARCH

Comme son nom l'indique, il s'agit d'une extension du modèle ARCH(p), au lieu de ne regarder que les retards des carrés du processus, nous regardons également les retards de la volatilité elle-même,

$$\begin{aligned} \varepsilon(t) &= z(t)h^{1/2}(t),\ z(t) \sim BB(0, \sigma_z^2) \\ h(t) &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon^2(t-i) + \sum_{j=1}^{q} \beta_j h(t-j), \end{aligned} \tag{2.2.10}$$

où $\alpha_0 > 0$, $\alpha_i \geq 0$ et $\beta_i \geq 0$. La représentation GARCH admet une unique solution stationnaire si

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1. \tag{2.2.11}$$

L'idée principale est que $h(t)$, la variance conditionnelle du processus $\varepsilon(t)$ sachant $\mathcal{H}(t-1)$ a une structure autorégressive et est corrélée positivement aux rendements aux carrés $\varepsilon^2(t-k)$. Cela permet d'avoir une volatilité (variance conditionnelle) persistante, de grandes valeurs de $\varepsilon^2(t)$ sont bien souvent suivies de grandes valeurs de $\varepsilon^2(t+k)$.

Les deux premiers moments conditionnels sont donnés par,

$$\mathbb{E}(\varepsilon(t)|\mathcal{H}(t-1)) = 0, \qquad \mathbb{E}(\varepsilon^2(t)|\mathcal{H}(t-1)) = h(t). \tag{2.2.12}$$

On en déduit directement la variance non-conditionnelle du processus,

$$\begin{aligned} \mathbb{V}\text{ar}(\varepsilon(t)) &= \mathbb{E}(\varepsilon^2(t)) - \mathbb{E}(\varepsilon(t))^2 \\ &= \mathbb{E}(\mathbb{E}(\varepsilon^2(t)|\mathcal{H}(t-1))) - \mathbb{E}(\mathbb{E}(\varepsilon(t)|\mathcal{H}(t-1)))^2 \\ &= \mathbb{E}(h(t)) \\ &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \mathbb{E}(\varepsilon^2(t-i)) + \sum_{j=1}^{q} \beta_j \mathbb{E}(h(t-j)) \\ &= \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}. \end{aligned} \tag{2.2.13}$$

On comprend maintenant la condition de stationnarité (2.2.11), si elle n'est pas respectée, la variance du processus explose. Si nous regardons la moyenne simple du processus,

$$\begin{aligned}\mathbb{E}(\varepsilon(t)) &= \mathbb{E}(z(t)h^{1/2}(t)) \\ &= \mathbb{E}[\mathbb{E}(z(t)h^{1/2}(t)|\mathcal{H}(t-1))] \\ &= \mathbb{E}[\mathbb{E}(z(t)|\mathcal{H}(t-1))h^{1/2}(t)] \\ &= \mathbb{E}[\mathbb{E}(z(t))h^{1/2}(t)] = 0.\end{aligned} \tag{2.2.14}$$

Montrons maintenant que les rendements ne sont pas corrélés entre eux. Comme $\mathbb{C}\text{ov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$, d'après (2.2.14) il suffit de montrer que $\mathbb{E}\varepsilon(t)\varepsilon(t-k) = 0$, $0 < k < t$,

$$\begin{aligned}\mathbb{E}(\varepsilon(t)\varepsilon(t-k)) &= \mathbb{E}(\varepsilon(t)h^{1/2}(t-k)z(t-k)) \\ &= \mathbb{E}[\mathbb{E}(\varepsilon(t)h^{1/2}(t-k)z(t-k)|\mathcal{H}(t-1))] \\ &= \mathbb{E}[\varepsilon(t)h^{1/2}(t-k)\mathbb{E}(z(t-k)|\mathcal{H}(t-1))] = 0.\end{aligned} \tag{2.2.15}$$

Pour le cas particulier d'un GARCH(1,1), paramétrisation classiquement utilisée en finance, qui nous évite donc d'avoir à déterminer les ordres p et q, par exemple à l'aide du critère d'Akaike, la kurtosis de $\varepsilon(t)$ est de la forme,

$$\gamma_2 = \frac{1-(\alpha_1+\beta_1)^2}{1-(\alpha_1+\beta_1)^2-\alpha_1^2(\mathbb{E}(z(t)^4)-1)}\gamma_2(z(t)). \tag{2.2.16}$$

Nous n'avons pas réécrit la vraisemblance du modèle, il s'agit exactement de la même que celle d'un processus ARCH (2.2.7), bien entendu la forme de la volatilité conditionnelle $h^{1/2}(t)$ est celle d'un GARCH et non plus d'un ARCH, et le jeu de paramètre à estimer n'est plus uniquement $\{\alpha_0, \alpha_1, \cdots, \alpha_p\}$ mais $\{\alpha_0, \alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q\}$.

Avant de raffiner les modèles, résumons cette première partie sur les modèles GARCH pour bien fixer les choses. En gardant la notation $\delta_\tau X(t)$ pour les log-rendements financiers, la modélisation GARCH(p,q) est,

$$\begin{aligned}\delta_\tau X(t) &= \mu(t) + \varepsilon(t) \\ \varepsilon(t) &= h^{1/2}(t)z(t), \qquad z(t) \sim BB(0,\sigma_z^2) \\ h(t) &= \alpha_0 + \sum_{i=1}^{p}\alpha_i\varepsilon^2(t-i) + \sum_{j=1}^{q}\beta_j h(t-j).\end{aligned} \tag{2.2.17}$$

$\varepsilon(t)$ est appelé le terme d'innovation, ou d'erreur. Une attention toute particulière doit être apportée au terme $\mu(t)$, il faut bien être conscient que la modèlisation GARCH est purement probabilistique et ne prend pas en compte de variables exogènes, on peut essayer de prédire en partie les rendements en fonction du rendement de marché $\delta_\tau X^m$, modèle CAPM (nous reviendrons sur ce point par la suite), d'un ensemble de variables explicatives $\boldsymbol{X}$, régression linéaire multiple, réseau de neurones, etc. (nous reviendrons également sur ce point par la suite), ou enfin simplement des retards des rendements, par exemple en les modélisant par un processus ARMA, modèle ARMA-GARCH (nous ne reviendrons pas sur ce point par la suite), etc. Soit, respectivement,

$$\begin{aligned}
&\delta_\tau X(t) = \delta_\tau X^m(t) + \varepsilon(t) \\
&\delta_\tau X(t) = \varphi(\boldsymbol{\beta}\delta_\tau \boldsymbol{X}) + \varepsilon(t), \;\; \varphi : \mathbb{R}^d \to \mathbb{R} \\
&\delta_\tau X(t) = \nu(t) + \sum_{i=1}^{p} \varphi_i \delta_\tau X(t-i) + \sum_{j=1}^{q} \theta_j \nu(t-j) + \varepsilon(t), \qquad \nu(t) \sim BB(0, \sigma_\nu^2).
\end{aligned} \tag{2.2.18}$$

Le choix de la distribution du bruit blanc $z(t)$ pour la représentation ARCH est également crucial, on pourra le choisir gaussien, student, GED, etc. La calibration des paramètres par maximum de vraisemblance sera naturellement différente pour chacune des lois choisies/testées. Pour le choix GED, rappelons que la densité est,

$$f(x) = \frac{\frac{\nu}{\sigma} \exp\left\{-\frac{1}{2}\left|\frac{x}{\lambda x}\right|^\nu\right\}}{\lambda 2^{(1+1/\nu)}\Gamma(1/\nu)}, \; \nu > 0. \tag{2.2.19}$$

avec $\lambda = \sqrt{2^{-(2/\nu)}\Gamma(1/\nu)/\Gamma(3/\nu)}$ et $\Gamma(\cdot)$ la fonction gamma définie par,

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} \mathrm{d}y. \tag{2.2.20}$$

Le paramètre ν est une mesure de l'épaisseur des queues de distributions. Le cas particulier $\nu = 2$ correspond au cas gaussien et si $\nu < 2$, la distribution admet des queues plus épaisses que la loi gaussienne. La fonction de vraisemblance du processus GARCH devient alors,

$$\begin{aligned}
\log L(\varepsilon(1), \cdots, \varepsilon(T); \alpha) =& T(\log(\nu) - \log(\lambda) - (1 + 1/\nu)\log(2) \\
&- \log(\Gamma(1/\nu))) - \frac{1}{2}\sum_{i=1} T \log(h(t)) - \frac{1}{2}\sum_{t=1}^{T}\left(\frac{\varepsilon^2(t)}{\lambda^2 h(t)}\right)^{\nu/2}.
\end{aligned} \tag{2.2.21}$$

L'un des inconvénients de ce type de processus et qu'en pratique, les paramètres α_i, β_i sont peu stables dans le temps, il s'agit d'un modèle discret. Un défaut du modèle GARCH est que la réponse de la variance conditionnelle aux innovations est linéaire. Seulement l'amplitude du shock est prise en compte pour déterminer la volatilité, pas qu'il soit positif ou négatif, en effet, comme $h(t)$ est déterminé en fonction du rendement au carré, $\varepsilon^2(t)$, il est clair que les effets de signe ne sont pas pris en compte.

Notons enfin que les queues de distribution produites sont trop épaisses par rapport à la réalité observée des marchés financiers. Enfin, les corrélations volatilité / volatilité décroissent exponentiellement et non en loi puissance, donc beaucoup trop rapidement.

Le modèle IGARCH, I pour *integrated* correspond au modèle GARCH(p, q) (2.2.10) non stationnaire, violant ainsi la condition (2.2.11) avec,

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1. \tag{2.2.22}$$

Prenons le cas d'un IGARCH(1,1), on a donc $\alpha_1 + \beta_1 = 1$, la représentation est alors,

$$h(t) = \alpha_0 + \alpha_1 \varepsilon(t-1)^2 + (1-\alpha_1)h(t-1) = \alpha_0 + h(t-1) + \alpha_1(\varepsilon(t-1)^2 - h(t-1)), \tag{2.2.23}$$

L'équation (2.2.23) nous donne donc un comportement persistant de la variance conditionnelle, elle ne décroît pas de manière abrupte.

2.2.3 Asymétrie

Passons maintenant aux modèles asymétriques. Le modèle EGARCH, E pour *Exponential* développé par Nelson en 91 [Ne91] autorise une forme d'asymétrie qui dépend non seulement du signe positif ou négatif de l'innovation, mais aussi de l'amplitude du chock,

$$\log h(t) = \alpha_0 + \sum_{i=1}^{q} \alpha_i g(z(t-i)) + \sum_{j=1}^{p} \beta_j \log h(t-j), \tag{2.2.24}$$

avec $z(t)$ un bruit blanc et la fonction g est telle que,

$$g(z(t-i)) = \theta z(t-i) + \gamma(|z(t-i)| - \mathbb{E}|z(t-i)|). \tag{2.2.25}$$

L'effet de signe est donné par le premier terme de g, $\theta z(t-i)$, le second terme, $\gamma(|z(t-i)| - \mathbb{E}|z(t-i)|)$ nous donne l'amplitude du chock.

Si nous supposons que la modélisation appropriée est une EGARCH(1,1) et que $\varepsilon(t)$ sont i.i.d. de loi normale centrée réduite, en notant θ l'ensemble des paramètres à estimer, la log-vraisemblance est alors,

$$\log L(\varepsilon(1), \cdots, \varepsilon(T); \theta) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(h(t)) - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon^2(t)}{h(t)}. \tag{2.2.26}$$

Bien entendu, elle ne change pas de précédemment, il peut être juste bon de la rappeler encore une fois. Une représentation alternative proposée par Bollerlev et al. 94 [BoEnNe94], pour l'impact des nouvelles est,

$$g(z(t-i)) = \sigma^{-2\theta_0}\frac{\theta_1}{1+\theta_2|z(t)|} + h - \gamma_0(t)\left[\frac{\gamma_1|z(t)|^\rho}{1+\gamma_2|z(t)|^\rho} - \mathbb{E}\left(\frac{\gamma_1|z(t)|^\rho}{1+\gamma_2|z(t)|^\rho}\right)\right]. \tag{2.2.27}$$

Les paramètres γ_0 et θ_0 nous permettent d'avoir une variance conditionnelle de $\log\sigma^2(t)$ et ca corrélation avec z_t variant d'ordre $\sigma^2(t)$.

Afin de prendre en compte la variation induite dans la variance conditionnelle d'un nouvel événement nous pouvons simplement introduire une indicatrice, il s'agit du modèle GJR-GARCH, GJR pour le nom des auteurs, Glosten, Jagannathan et Runkle en 93, [GlJaRu93]. La variance conditionnelle s'écrit alors,

$$h(t) = \alpha_0 + \sum_{i=1}^{q}(\alpha_i\varepsilon^2(t-i) + \gamma_i\mathbb{1}_{\varepsilon(t-i)<0}\varepsilon(t-i)^2) + \sum_{i=1}^{p}\beta_i h(t-i). \tag{2.2.28}$$

Ecrivons le cas (0,1),

$$h(t) = \alpha_0 + (\alpha_1 + \gamma_1\mathbb{1}_{\varepsilon(t-1)<0})\varepsilon^2(t-1), \tag{2.2.29}$$

nous voyons ainsi que nous n'avons que deux régimes possibles, soit uniquement α_1, soit $\alpha_1 + \gamma_1$, et cela de manière 'brutale', le changement de régime se faisant

à un instant bien précis. Nous pouvons très bien nous dire que les changements se font de manière plus 'lisses' et écrire,

$$h(t) = \alpha_0 + \sum_{i=1}^{q} (\gamma_1 g(\theta\varepsilon(t-i)) + \gamma_2(1 - g(\theta\varepsilon(t-i))))\varepsilon^2(t-i) + \sum_{i=1}^{p} \beta_i h(t-i), \quad (2.2.30)$$

avec $\theta > 0$ et la fonction g une sigmoïde donnée par,

$$g(x) = \frac{1}{1 + e^{-x}}, \quad (2.2.31)$$

il s'agit du modèle Logistic Smooth Transition GARCH de Gonzales-Rivera (1998) [Go98].

2.2.4 Mémoire Longue

Le modèle FIGARCH, FI pour fractionally integrated proposé par Baillie, Bollerslev et Mikkelsen (1996) [BaBoMi96], est une extension du modèle GARCH permettant d'obtenir une mémoire longue dans le processus de volatilité. La décroissance exponentielle de l'autocorrélation des modèles précédents se révèle trop rapide pour se conformer aux données de marchés, ils ne sont pas véritablement adaptés. Le processus FIGARCH admet quant à lui une décroissance plus lente, hyperbolique. Les équations du processus FIGARCH(p, d, q) sont données par,

$$\begin{aligned} \varepsilon(t) &= h^{1/2}(t) z(t) \\ \Phi(L)(1-L)^d \varepsilon^2(t) &= \omega + [1 - \beta(L)](\varepsilon^2(t) - h(t)), \end{aligned} \quad (2.2.32)$$

avec L l'opérateur de retard et,

$$\begin{aligned} \Phi(L) &= [1 - \alpha(L) - \beta(L)](1-L)^{-1} \\ \alpha(L) &= \sum_{k=0}^{q} L^k \\ \beta(L) &= \sum_{k=0}^{p} \beta_k L^k. \end{aligned} \quad (2.2.33)$$

Par souci de clarté, nous réécrivons les équations pour le cas $p = 1$, $q = 0$ et $v(t) = \varepsilon^2(t) - h(t)$,

$$(1-L)^d \varepsilon_t^2 = \omega + [1 - \beta_1 L] v(t), \quad (2.2.34)$$

ainsi, grace à la décomposition de Wold nous avons,

$$\begin{aligned}\sigma^2(t) &= \omega - \beta_1 v_{t-1} + [1-(1-L)^d]\varepsilon^2(t) \\ &= \omega - \beta_1 v_{t-1} - \sum_{k=1}^{\infty} \psi_k \varepsilon^2(t-k),\end{aligned} \tag{2.2.35}$$

avec,

$$\psi_k = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}, \tag{2.2.36}$$

où Γ est la fonction gamma, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$. Pour bien mettre en avant comment la longue mémoire est incorporée, nous traçons sur la figure (2.2.4) les fonctions ψ_k pour k variant de 0,5 à 1.

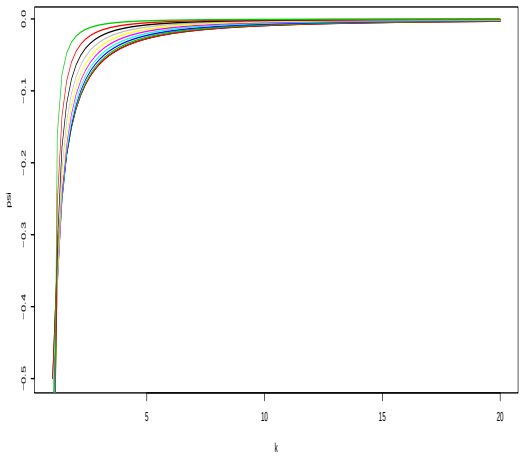


FIGURE 2.1 – Fonctions ψ_k pour $k = 0.5, 0.55 \cdots, 1$. Plus l'on s'approche de 1, plus les courbes croissent 'plus rapidement'.

Concluons cette partie sur la modélisation en utilisant des processus ARCH, ou plutôt, ne concluons pas ! Nous avons donné quelques résultats théoriques, certains ont été soit omis volontairement soit parce que les résultats ne sont pas (encore) connus, soit simplement parce que votre enseignant a oublié de les noter. Aucune simulation n'a été faite non plus, il est vivement conseillé aux étudiants de vérifier par eux-mêmes quel modèle vérifie quel fait stylisé, quel modèle donne les meilleurs

résultats en terme de prévision de volatilité.

Dans tous les cas, le caractère multifractal des marchés financiers n'est pas décrit par les processus ARCH.

2.3 Modèle Multifractal

Le terme *fractal* a été introduit par Mandelbrot pour pouvoir décrire la géométrie d'objets à dimension non entière. Avant de parler des processus multifractals, commençons par les processus fractals, caractérisés par une propriété d'autosimilarité.

Définition 2.3.1 (Autosimilaire)**.** *Un processus* $(X(t), t \geq 0)$ *est autosimilaire si pour tout* $c > 0$*, il existe* b *tel que,*

$$\mathcal{L}(\{X(ct), t \geq 0\}) = \mathcal{L}(\{bX(t), t \geq 0\}). \tag{2.3.1}$$

Pour le cas particulier où il existe $H > 0$ *tel que* $b = c^H$*, on parle de processus* $H-$*auto-similaire. L'exposant* H *est appelé l'exposant de Hurst, l'exposant d'echelle.*

Le cas particulier de $H = 1/2$ correspond au mouvement brownien standard. Introduisons maintenant le cas plus général du mouvement brownien fractionnaire introduit par Kolmogorov [Ko40] et étudié par Mandelbrot et Van Ness dans [MaNe68].

Définition 2.3.2 (Brownien fractionnaire)**.** *Un processus gaussien* $(B^H(t), t \geq 0)$ *est un mouvement Brownien fractionnaire (mBf) d'exposant de Hurst* $H \in (0, 1)$ *si* $\mathbb{E}(B^H(t)) = 0$ *et,*

$$\mathbb{E}(B^H(t)B^H(s)) = \frac{1}{2}(s^{2H} + t^{2H} - |t - s|^{2H}). \tag{2.3.2}$$

Pour l'historique, l'exposant H, a été introduit par le climatologue Hurst suite à une analyse statistique des crues du Nil [Hu10].

Le mBf est un processus H-autosimilaire,

$$\mathcal{L}(\{B^H(ct), t \geq 0\}) = \mathcal{L}(\{c^H B^H(t), t \geq 0\}), \tag{2.3.3}$$

et à accroissements stationnaires,

$$\mathcal{L}(\{B^H(t+s) - B^H(t), t \geq 0\}) = \mathcal{L}(\{B^H(s) - B^H(0), t \geq 0\}), \forall s \geq 0. \quad (2.3.4)$$

Enfin, dans le cas ou $1 > H > 1/2$, les corrélations sont positives et à longue portée, il possède la propriété de longue dépendance. Si $0 > H > 1/2$, les corrélations sont négatives, et si $H = 1/2$, les accroissements sont indépendants.

La notion de fractal s'étend naturellement au cas multifractal, cette fois nous n'avons plus simplement un H tel que $X(ct) = c^H X(t)$ mais quelque chose de plus général, $X(ct) = M(c)X(t)$, on parle d'invariance d'échelle. La notion d'invariance d'échelle est donc le coeur de l'analyse multifractale. On dit qu'une fonction f (un observable) est *invariante d'échelle* si,

$$f(\lambda x) = \mu f(x), \quad (2.3.5)$$

où μ est une certaine fonction. Classiquement x est une distance, le temps, en d'autres termes, à chaque échelle d'observation à laquelle on se place, on observe les mêmes propriétés. Propriété que les rendements financiers vérifient comme nous avons pu le voir sur les figures (1.9) et (1.10). Il n'y a pas que les rendements financiers qui ont cette particularité, Mandelbrot a listé un tas de cas ou cela se vérifie, comme par exemple les côtes maritimes de l'Angleterre, les fluctuations du vent ou le choux romanesco. Une illustration est donnée sur la figure (2.2)

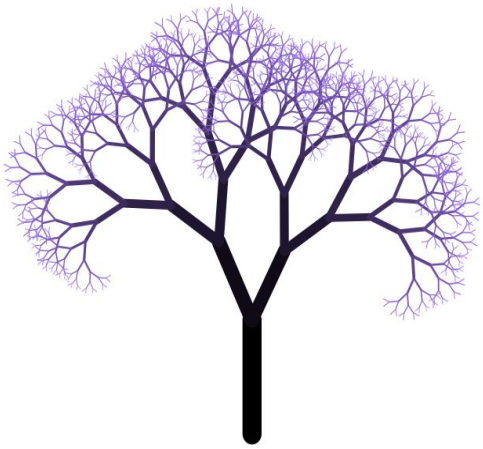


FIGURE 2.2 – Arbre fractal.

Nous l'avions déjà évoqué précédemment, les moments empiriques en valeurs absolues d'ordre q à différents lag τ des rendements sont donnés par,

$$m(q,\tau) = \mathbb{E}[|\delta_\tau X(t)|^q]. \tag{2.3.6}$$

la série de rendements (stationnaire) est alors invariante d'échelle si,

$$m(q,\tau) \sim C(q)\tau^{\zeta(q)}, \tag{2.3.7}$$

avec $C(q)$ le moment au lag 1 d'ordre q et $\zeta(q)$ une fonction dépendant de l'exposant de Hurst, H. Nous pouvons alors distinguer deux cas, si $\zeta(q)$, l'exposant de fractalité, est une fonction linéaire de q, on dira qu'il s'agit d'un processus fractal, et si $\zeta(q)$ est une fonction non linéaire, on dira qu'il s'agit d'un processus multifractal.

Il convient donc de s'intéresser de plus près à ce fameux exposant $\zeta(q)$.

Proposition 2.3.1. *Dans le cas où l'invariance d'échelle (2.3.7) est vérifiée pour tout $\tau \to 0$, alors l'exposant $\zeta(q)$ est une fonction concave de q.*

Preuve. Notons q_1 et q_2 deux moments et $\alpha \in [0,1]$, alors par l'inégalité d'Hölder nous avons,

$$\mathbb{E}[|\delta_\tau X(t)|^{\alpha q_1 + (1-\alpha)q_2}] \leq \mathbb{E}[|\delta_\tau X(t)|^{q_1}]^\alpha \mathbb{E}[|\delta_\tau X(t)|^{q_2}]^{1-\alpha}. \tag{2.3.8}$$

Par hypothèse, nous avons un processus stationnaire et invariant d'échelle pour $\tau \to 0$, donc,

$$\mathbb{E}[|\delta_\tau X(t)|^q] \sim_{\tau\to 0} C(q)\tau^{\zeta(q)} \tag{2.3.9}$$

ainsi, avec (2.3.8) et (2.3.9), en prenant le logarithme, nous avons pour $q = \alpha q_1 + (1-\alpha)q_2$,

$$\ln C(q) + \zeta(q)\ln(\tau) \leq \alpha(\ln C(q_1) + \zeta(q_1)\ln(\tau)) + (1-\alpha)(\ln C(q_2) + \zeta(q_2)\ln(\tau)). \tag{2.3.10}$$

Pour finir la preuve, il suffit de diviser l'égalité par $\ln\tau$ et faire tendre τ vers 0,

$$\zeta(\alpha q_1 + (1-\alpha)q_2) \geq \alpha\zeta(q_1) + (1-\alpha)\zeta(q_2). \tag{2.3.11}$$

□

Proposition 2.3.2. *Si l'exposant $\zeta(q)$ est non linéaire, alors la loi d'échelle (2.3.7) n'est pas valide à la fois pour le cas $\tau \to 0$ et $\tau \to \infty$*

Preuve. Idem que pour (2.3.1) □

Pourquoi est-ce important de voir ces deux propositions ? Simplement, cela nous dit que le lag τ doit appartenir à un intervalle borné $[0, T]$, T est appelé le *temps intégral* et dans ce cas, on dit que le processus a une invariance d'échelle exacte caractérisée par son exposant multifractal $\zeta(q)$. En dehors de cet intervalle, le processus n'est plus multifractal.

Nous introduisons maintenant la mesure multifractale, c'est elle qui va nous servir de base de construction des différents modèles de cette section.

2.3.1 Mesure Multifractale

Définition 2.3.3 (Mesure aléatoire)**.** *Une mesure aléatoire μ sur l'intervalle X est analague à une variable aléatoire. Il s'agit d'une application définie sur un espace de probabilité, et à valeurs dans la classe de toutes les mesures sur X. Pour un intervalle $I \subseteq X$, la masse $\mu(I)$ est une variable aléatoire.*

La mesure multifractale la plus simple est la mesure binomiale (aussi dit de Bernoulli ou Besicovitch) sur l'intervalle compact $[0, 1]$, c'est pourquoi nous commençons naturellement par cet exemple. Il s'agit de la limite d'une procédure itérative appelée *cascade multiplicative*.

Notons m_0 et m_1 deux nombres positifs tel que $m_0 + m_1 = 1$. A l'étape $k = 0$, nous commençons la mesure de probabilité uniforme sur $[0, 1]$. A l'étape $k = 1$, on affecte à la mesure μ_1 la masse m_0 sur le sous intervalle $[0, \frac{1}{2}]$ et la masse m_1 sur l'intervalle $[\frac{1}{2}, 1]$.

A l'étape $k = 2$, chaque intervalle est subdivisé en deux parties égales, par exemple pour $[0, \frac{1}{2}]$ on obtient $[0, \frac{1}{4}]$ et $[\frac{1}{4}, \frac{1}{2}]$, chacun des intervalles recevant une fraction m_0 et m_1 de la masse totale $\mu_1([0, \frac{1}{2}])$. En appliquant la même procédure sur l'ensemble dyadique $[\frac{1}{2}, 1]$ on obtient,

$$\begin{aligned}
&\mu_2\left(\left[0, \frac{1}{4}\right]\right) = \mu_1\left(\left[0, \frac{1}{2}\right]\right) m_0 = m_0 m_0, \quad \mu_2\left(\left[\frac{1}{4}, \frac{1}{2}\right]\right) = \mu_1\left(\left[0, \frac{1}{2}\right]\right) m_1 = m_0 m_1 \\
&\mu_2\left(\left[\frac{1}{2}, \frac{3}{4}\right]\right) = \mu_1\left(\left[\frac{1}{2}, 1\right]\right) m_0 = m_1 m_0, \quad \mu_2\left(\left[\frac{1}{4}, \frac{1}{2}\right]\right) = \mu_1\left(\left[\frac{1}{2}, 1\right]\right) m_1 = m_1 m_1.
\end{aligned} \tag{2.3.12}$$

On présente la construction jusqu'à l'ordre $k = 3$ sous forme d'une cascade (2.3).

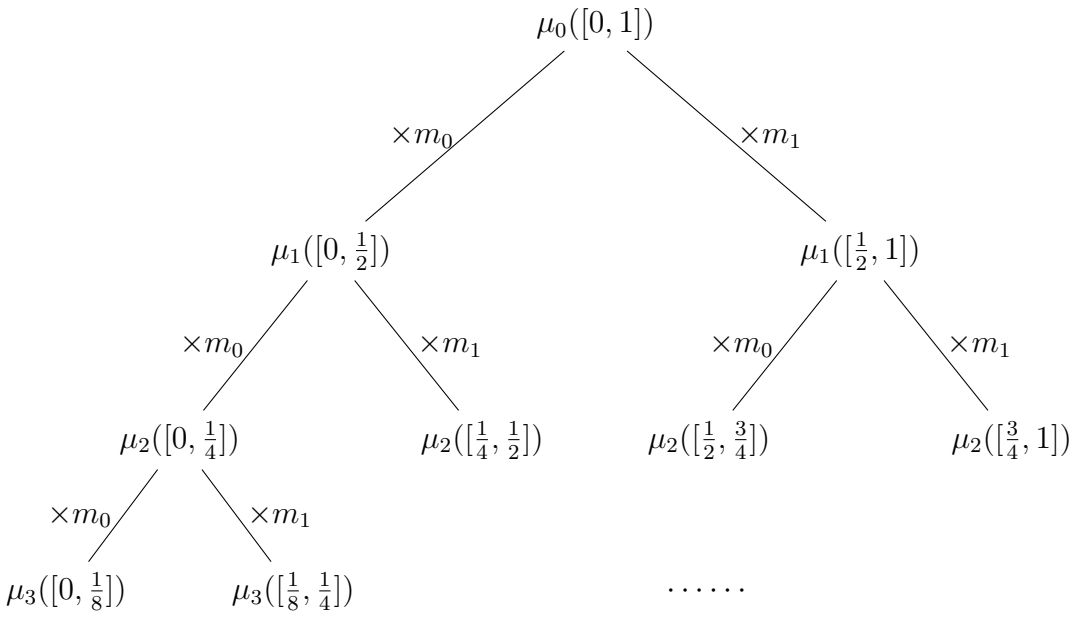


FIGURE 2.3 – Construction de la mesure binomiale.

A l'itération $k+1$, étant donnée la mesure μ_k et l'intervalle $[t, t+\frac{1}{2^k}]$ avec t le nombre dyadique de la forme,

$$t = 0.\eta_1\eta_2\cdots\eta_k = \sum_{i=1}^{k} \eta_i \frac{1}{2^i}, \tag{2.3.13}$$

dans la base $b=2$. La masse $\mu_k\left(\left[t, t+\frac{1}{2^k}\right]\right)$ est répartie uniformément sur les deux sous-intervalles $[t, t+\frac{1}{2^{k+1}}]$ et $[t+\frac{1}{2^{k+1}}, t+\frac{1}{2^k}]$ avec les proportions m_0 et m_1. En répétant la même procédure à tous les intervalles, nous arrivons ainsi à construire la mesure μ_{k+1}. La mesure binomiale μ est définie comme la limite de la séquence μ_k,

$$\mu = \lim_{k\to\infty} \mu_k. \tag{2.3.14}$$

Notons que comme $m_0+m_1 = 1$, chaque étape de la construction préserve la masse, pour cette raison, on appelle cette procédure *conservative* ou, *microcanonique*.

L'extension à la mesure multinomiale se fait assez naturellement. Au lieu de diviser chaque intervalle en 2, nous les divisons en sous-intervalles de tailles égales $b > 2$. Chacun de ces sous-intervalles indexé de gauche à droite par β, $(0 \leq \beta \leq$

$b-1$), reçoit une fraction de la masse totale égale à $m_0, \cdots, m_{b-1}$. Par conservation de la masse, ces fractions sont appelés *multiplieurs* tel que $\sum m_\beta = 1$.

Toujours de manière naturelle, l'extension suivante serait d'allouer les masses sur les sous-intervalles construits à chaque itération de manière aléatoire. Le multiplieur de chaque sous intervalle est ainsi une variable aléatoire discrète M_β positive prenant ses valeurs dans $m_0, m_1, \cdots, m_{b-1}$ avec les probabilités $p_0, p_1, \cdots, p_{b-1}$. La préservation de la masse imposant $\sum M_\beta = 1$. Cette procédure amenant à la construction de *mesures multiplicatives* est dite *cascade multiplicative*.

Détaillons un peu plus cette construction. On impose toujours la contrainte de normalisation, $\sum M_\beta = 1$, la masse intiale sur l'intervalle [0,1] est 1, et nous subdivisons l'intervalle en sous-intervalles de longueur $1/b$, chaque intervalle recevant un poids aléatoire M_β. Par répétition, l'intervalle $b-$adique de longueur $\Delta t = \frac{1}{b^k}$, partant de $t = 0.\eta_1 \cdots \eta_k = \sum_i \eta_i \frac{1}{b^i}$ a pour mesure,

$$\mu(\Delta t) = M(\eta_1)M(\eta_1\eta_2)\cdots M(\eta_1\cdots\eta_k). \tag{2.3.15}$$

Donc, $[\mu(\Delta t)]^q = M(\eta_1)^q M(\eta_1\eta_2)^q \cdots M(\eta_1\cdots\eta_k)^q$ pour tout $q \geq 0$. En passant à l'espérance, les multiplieurs $M(\eta_1), M(\eta_1\eta_2), \cdots, M(\eta_1\cdots\eta_k)$ étant indépendants et de même loi, nous obtenons,

$$\begin{aligned} \mathbb{E}[\mu(\Delta t)^q] &= \mathbb{E}[M(\eta_1)^q M(\eta_1\eta_2)^q \cdots M(\eta_1\cdots\eta_k)^q] \\ &= (\mathbb{E}[M^q])^k, \end{aligned} \tag{2.3.16}$$

soit,

$$\mathbb{E}[\mu(\Delta t)^q] = (\Delta t)^{\zeta(q)}, \quad \zeta(q) = -\log_b \mathbb{E}[M^q]. \tag{2.3.17}$$

Assumons maintenant que le poids soit conservé à chaque itération en espérance, $\mathbb{E}[\sum M_\beta] = 1$, autrement dit, qu'il s'agit d'une variable aléatoire, que nous noterons Ω, et dans ce cas,

$$\mu(\Delta t) = \Omega(\eta_1\cdots\eta_k)M(\eta_1)M(\eta_1\eta_2)\cdots M(\eta_1\cdots\eta_k), \tag{2.3.18}$$

et nous avons la relation d'échelle,

$$\mathbb{E}[\mu(\Delta t)^q] = \mathbb{E}[\Omega^q](\mathbb{E}[M^q])^k. \tag{2.3.19}$$

Soit,

$$\mathbb{E}[\mu(\Delta t)^q] = \mathbb{E}[\Omega^q](\mathbb{E}[M^q])^k, \tag{2.3.20}$$

qui caractérise le caractère multifractal.

2.3.2 Modèle Multifractal des Rendements Financiers

Rentrons dans le vif du sujet en présentant un premier modèle multifractal, il s'agit du modèle *Multifractal Model of Asset Return* (MMAR) introduit par Mandelbrot, Calvet et Fisher (1997) [MaFiCa97] et [FiCaMa97]. Il s'agit d'un processus subordonné, le premier modèle de ce type est attribué à Clark, [Cl73].

Définition 2.3.4. *Soit* $\{B(t)\}$ *un processus stochastique et* $\theta(t)$ *une fonction croissante de* t*. Alors le processus,*

$$X(t) = B^H(\theta(t)), \tag{2.3.21}$$

est un processus de composé. B^H *est un brownien fractionnaire d'exposant* H*, l'indice* t *dénote le temps physique et,* $\theta(t)$ *est le temps de trading, ou le processus de déformation du temps.*

Le modèle MMAR correspond au cas particulier où $\{\theta(t)\}$ et $\{B^H(t)\}$ sont indépendants et $\theta(t)$ la fonction de répartition d'une mesure multifractale définie sur $[0, T]$. Ainsi, $\theta(t)$ est une fonction à incréments continus, croissants et stationnaires. Le temps de trading étant issu d'une mesure multifractale, le processus $X(t)$ est également multifractal. En notant $\{\mathcal{F}_t^\theta; t \in [0, T]\}$ la filtration engendrée par le processus $\theta(t)$, on a,

$$\begin{aligned}
\mathbb{E}[|X(t)|^q] &= \mathbb{E}[|B^H(\theta(t))|^q] \\
&= \mathbb{E}[\mathbb{E}[|B^H(\theta(t))|^q|\mathcal{F}_t^\theta]] \\
&= \mathbb{E}[\mathbb{E}[|\theta(t)^H B^H(1)|^q|\mathcal{F}_t^\theta]] \\
&= \mathbb{E}[\theta(t)^{qH}]\mathbb{E}[|B^H(1)|^q].
\end{aligned} \tag{2.3.22}$$

En utilisant la définition de l'exposant multifractal on voit donc que,

$$\zeta_X(q) = \zeta_\theta(qH), \tag{2.3.23}$$

puisque,

$$C_X(q)t^{\zeta_X(q)} \sim \mathbb{E}[|X(t)|^q] = \mathbb{E}[\theta(t)^{qH}]\mathbb{E}[|B^H(1)|^q] \sim \mathbb{E}[|B^H(1)|^q]t^{\zeta_\theta(qH)}. \tag{2.3.24}$$

Par un simple calcul on peut montrer que dans le cas où θ une la fonction de répartition d'une mesure multiplicative log-normale l'exposant s'écrit,

$$\zeta_X^{ln}(q) = \zeta_\theta^{\ln}(qH) = qH(1 + 2\lambda^2) - 2\lambda^2 q^2 H^2. \tag{2.3.25}$$

Le processus vérifie certains des faits stylisés, l'autocovariance des rendements est nulle pour le cas où $H = 1/2$ (la preuve se base sur le fait que $X(t)$ soit une martingale et que donc, l'hypothèse de non-arbitrage est vérifiée avec cette modélisation), si $H > 1/2$ elle est positive et négative pour $H < 1/2$. En revanche, le modèle de Mandelbrot, Calvet et Fisher ne permet pas de vérifier l'effet de levier.

En dehors de cela (de toute facon, l'effet de levier n'est pas non plus vérifié par tous les actifs financiers), les avantages/inconvénients de ce modèle est qu'il n'y a pas de jeu de paramètre à estimer comme pour un modèle ARCH par exemple, en même temps, cela tombe bien, nous n'avons pas de formule fermée pour la vraisemblance ou d'autre méthode d'estimation comme la méthode généralisée des moments. Nous n'avons donc qu'à choisir la valeur de H, qui sera prise naturellement égale à 0,5 pour 'coller' aux données financières les plus classiques, exceptions faites pour des données à fréquence très élevée ainsi que quelques commodities et des produits très peu liquides susceptibles de présenter une certaine persistance dans leurs rendements. L'autre, et dernier paramètre à choisir est la taille des sous-intervalles, b ansi que la loi régissant la distribution des masses m_i. La construction est donc non causale.

2.3.3 Modèle Multifractal Markov Switching

Le modèle que nous allons présenter maintenant à été introduit par Calvet et Fisher [CaFi01], [CaFi04], il s'agit du modèle multifractal Markov switching ou encore multifrequency Markov switching. Les modèles à changement de régime sont principalement dus à Hamilton (1989, 1900). La plupart des modèles initiaux ne proposent que 2 (voir 4) fréquences différentes, simplement parce qu'ils sont destinés à des séries financières à basse fréquence où un petit nombre d'états apparaît comme suffisant pour décrire les cours boursiers. A la différence du modèle présenté précédemment, il ne s'agit pas du temps qui est issu d'une mesure multifractale, mais de la volatilité elle-même.

Le modèle multifractal Markov switching est un processus à volatilité stochastique construit à partir d'un processus de Markov du premier ordre de $\bar{k}$ composants (on parle d'un processus de Markov du premier ordre lorsque l'état à l'instant $t + \delta t$ dépend uniquement de l'état à l'instant t),

$$M_t = (M_t^{(1)}; M_t^{(2)}; \cdots; M_t^{(\bar{k})}) \in \mathbb{R}_+^{\bar{k}}. \tag{2.3.26}$$

Chacun des composants du vecteur $M_t^{(i)}$ a la même distribution marginale M mais évolue à différentes fréquences. La distribution M est supposée à support positif et tel que $\mathbb{E}M = 1$. Les multiplieurs sont également mutuellement indépendants, $M_t^{(k)}$ et $M_{t'}^{(k')}$ sont indépendants si k est différent de k'. Le vecteur d'état est construit à chaque nouvel instant à partir de l'instant précédent. Pour tout $k \in \{1, \cdots, \bar{k}\}$,

$$\begin{aligned} M_t^{(k)} &\sim M \text{ avec probabilité } \gamma_k \\ M_t^{(k)} &= M_{t-\delta t}^{(k)} \text{ avec probabilité } 1 - \gamma_k. \end{aligned} \tag{2.3.27}$$

Les probabilités de transitions $(\gamma_1, \cdots, \gamma_{\bar{k}})$ sont construites par itérations,

$$\gamma_k = 1 - (1 - \gamma_1)^{b^{k-1}}, \tag{2.3.28}$$

avec $\gamma_1 \in (0, 1)$ et $b \in (1, \infty)$. Pour de petites valeurs de k, la quantité $\gamma_1 b^{(k-1)}$ reste petite et la transition de probabilité vérifie,

$$\gamma_k \sim \gamma_1 b^{(k-1)}. \tag{2.3.29}$$

Pour de plus grande fréquence, c'est à dire k élevé, $\gamma_k \sim 1$.

Finalement, le modèle prend la forme,

$$\begin{aligned} \delta_\tau X(t) &= \sigma(M_t) z_t \\ \sigma(M_t) &= \bar{\sigma} \left(\prod_{i=1}^{\bar{k}} M_t^{(i)} \right)^{1/2}, \end{aligned} \tag{2.3.30}$$

avec $\bar{\sigma}$ une constante positive.

Quand les multiplieurs à de basses fréquence (petites valeurs de k) changent, la volatilité varie de manière discontinue avec une persistance forte, Les multiplieurs à plus haute fréquence (grandes valeurs de k) produisent quant à eux des outliers.

La construction multifractale du processus ne nous impose que quelques restrictions sur la distribution des multiplieurs, support positif et de moyenne unitaire, et comme nous avons $\gamma_1 < \cdots < \gamma_{\bar{k}} < 1 < b$, nous n'avons qu'à choisir $(b, \gamma_{\bar{k}})$ pour déterminer l'ensemble des probabilités de transitions. Ainsi, l'ensemble des paramètres du modèle est,

$$\Theta = \{m_0, \bar{\sigma}, b, \gamma_{\bar{k}}\} \in \mathbb{R}_+^4, \tag{2.3.31}$$

avec m_0 carctérisant la distribution des multiplieurs et $\bar{\sigma}$ la volatilité des rendements. Plusieurs méthodes sont possibles pour estimer le jeu de paramètre optimal, nous pouvons citer par exemple la méthode généralisée des moments ou l'estimateur du maximum de vraisemblance. Nous ne présentons ici que la méthode du maximum de vraisemblance et renvoyons l'étudiant curieux de voir comment faire par méthode des moments à [Lu08].

Supposons que M est discret et que la chaîne de Markov M_t prend un nombre fini de valeurs $m^1, \cdots, m^d \in \mathbb{R}_+^{\bar{k}}$. Sa dynamique est donnée par la matrice de transition $A = (a_{ij})_{1 \leq i,j \leq d}$ avec comme composants,

$$a_{ij} = \mathbb{P}(M_{t+1} = m^j | M_t = m^i). \tag{2.3.32}$$

Conditionnellement à l'état de la volatilité, le rendement à l'instant t, $\delta_\tau X(t)$ à pour densité,

$$f(\delta_\tau X | M_t = m^i) = \frac{1}{\sigma(m^i)} \phi\left(\frac{x}{\sigma(m^i)}\right), \tag{2.3.33}$$

où ϕ est la densité d'une loi normale centrée réduite. Dis autrement, la densité est gaussienne centrée de variance $\sigma^2(M_t)$. Les différents états ne sont pas observables, en revanche, nous pouvons calculer leurs différentes probabilités conditionnelles,

$$\Pi_t^j = \mathbb{P}(M_t = m^j | \mathcal{I}_t), \;\; \mathcal{I}_t = \{\delta_\tau X(1), \cdots, \delta_\tau X(t)\}. \tag{2.3.34}$$

En utilisant la règle de Bayes nous avons,

$$\Pi_t^j = \frac{\omega(\delta_\tau X(t)) * (\Pi_{t-1} A)}{[\omega(\delta_\tau X(t)) * (\Pi_{t-1} A)]\mathbf{1}^t}, \tag{2.3.35}$$

avec $\mathbf{1} = (1, \cdots, 1) \in \mathbb{R}^d$, $x * y$ dénote le produit de Hadamard pour tout $x, y \in \mathbb{R}^d$, $(x^1, x^2, \cdots, x^d) * (y^1, y^2, \cdots, y^d) = (x^1 y^1, \cdots, x^d y^d)$ et,

$$\omega(\delta_\tau X(t)) = \left(\frac{1}{\sigma(m^1)}\phi(\sigma(m^1)), \cdots, \frac{1}{\sigma(m^d)}\phi(\sigma(m^d))\right). \tag{2.3.36}$$

Nous avons donc besoin de choisir un vecteur d'initialisation du processus du Markov. Nous le choisissons ergodique et comme les multiplieurs sont mutuellement indépendants, la distribution ergodique initiale est donnée par $\Pi_0^j = \prod_{s=1}^{\bar{k}} \mathbb{P}(M = m_s^j)$. Ainsi, et toujours grâce à la règle de Bayes, nous trouvons la vraisemblance du modèle,

$$\begin{aligned} f(\delta_\tau X(t)|\mathcal{I}_t) &= \sum_{i=1}^{d} \mathbb{P}(M_t = m^i|\mathcal{I}_{t-1}) f(\delta_\tau X(t)|M_t = m^i) \\ &= \sum_{i=1}^{d} \mathbb{P}(M_t = m^i|\mathcal{I}_{t-1}) \frac{1}{\sigma(m^i)} \phi\left(\frac{x}{\sigma(m^i)}\right). \end{aligned} \tag{2.3.37}$$

La log-vraisemblance est donc,

$$\ln L(\delta_\tau X(1), \cdots, \delta_\tau X(T); \theta) = \sum_{t=1}^{T} \ln(\omega(\delta_\tau X(t))(\Pi_{t-1} A)). \tag{2.3.38}$$

Etudions maintenant les propriétés du modèle multifractal Markov switching. La démonstration de la forme assymptotique de la persistance dans les rendements d'ordre q en valeur absolue nécessitant quelques calculs fastidieux, mais intéressants [CaFi01], nous ne présentons que le résultat et l'interprétation.

En notant $C_q(h)$ la fonction d'autocorrélation du processus d'ordre q en valeur absolue,

$$C_q(h) = \mathbb{C}\text{or}(|\delta_\tau X(t)|^q, |\delta_\tau X(t+h)|^q), \tag{2.3.39}$$

alors si $\delta_\tau X(t)$ suit un processus multifractal Markov switching nous avons,

$$\sup_{n \in I_{\bar{k}}} \left| \frac{\ln C_q(h)}{\ln n^{-\beta(q)}} - 1 \right| \to 0 \text{ avec } \bar{k} \to \infty, \tag{2.3.40}$$

où $\beta(q) = \log_b(\mathbb{E}(M^q)) - \log_b(\mathbb{E}(M^{q/2})^2)$ et $I_{\bar{k}}$ est un ensemble d'entier tel que pour $\alpha_1 < \alpha_2 \in (0,1)$ choisi arbitrairement,

$$I_{\bar{k}} = \{n : \alpha_1 \log_b(b^{\bar{k}}) \leq \log_b n \leq \alpha_2 \log_b(b^{\bar{k}})\}. \tag{2.3.41}$$

L'autocorrélation du processus décroît donc de manière hyperbolique avec les pas de temps. On peut aussi montrer que pour de grande valeur de h, la décroissance ce fait de manière exponentielle.

Ce modèle peut également se prolonger en version continue,

$$\mathrm{d}X(t) = \mu \mathrm{d}t + \sigma(M_s)\mathrm{d}B(s), \tag{2.3.42}$$

avec B un mouvement brownien standard et $\sigma(M_s)$ défini comme précédemment par $\sigma(M_s) = \bar{\sigma}\left(\prod_{k=1}^{\bar{k}} M_t^{(k)}\right)^{1/2}$. L'intégrale stochastique étant bien définie puisque $\mathbb{E}\int_0^t \sigma^2(M_s)\mathrm{d}s = \bar{\sigma}^2 t < \infty$. A la différence de modèle MMAR, nous avons donc ici des incréments stationnaires, facilitant l'estimation des paramètres et la prévision de la volatilité. De plus les composants de la volatilité varient à des instants aléatoires et non prédéterminés.

2.3.4 Marche Aléatoire Multifractale

Avant d'introduire cette nouvelle modélisation, donnons quelques définitions pour pouvoir regarder la mesure multifractale sous jacente à la construction.

Définition 2.3.5. *Une variable aléatoire réelle X de loi μ a une distribution infiniment divisible si et seulement si pour tout n il existe $X_1, \cdots, X_n$ indépendante et de même loi μ_n tel que,*

$$\mathcal{L}(X) = \mathcal{L}(X_1 + \cdots + X_n). \tag{2.3.43}$$

Théorème 2.3.1. *Si X a une loi infiniment divisible, alors sa fonction caractéristique est,*

$$\phi(q) = \mathbb{E}[e^{iqX}] = \exp\left(imt + \int_{-\infty}^{\infty} \frac{e^{iqx} - 1 - iq\sin x}{x^2}\nu(\mathrm{d}x)\right), \tag{2.3.44}$$

avec $m \in \mathbb{R}$ et ν une mesure telle que les intégrales,

$$\int_x^{\infty} \frac{\nu(\mathrm{d}y)}{y^2} \quad et \quad \int_{-\infty}^{-x} \frac{\nu(\mathrm{d}y)}{y^2}, \tag{2.3.45}$$

sont toutes deux convergentes pour $x > 0$, elle est dite mesure de Lévy.

Proposition 2.3.3. *Soit $(X(t), t \geq 0)$ un processus de Lévy, i.e. tel que $X(t) - X(s)$ soit stationnaire et indépendant pour tout $t > s$, alors ses accroissements sont infiniment divisibles, réciproquement, une loi infiniment divisible est un processus de Lévy.*

Proposition 2.3.4. *Si X a une loi infiniment divisible, alors sa fonction caractéristique s'écrit comme une puissance n-ième d'une autre fonction caractéristique.*

Par exemple, si $X \sim \mathcal{N}(m, \sigma^2)$, alors sa fonction caractéristique est,

$$\phi(q) = \exp\left(imt - \frac{q^2\sigma^2}{2}\right) = \left[\exp\left(\frac{imq}{n} - \frac{t^2\sigma^2}{2n}\right)\right]^n \tag{2.3.46}$$

en d'autre terme, on a $X = X_1 + \cdots, X_n$ avec $X_i \sim \mathcal{N}\left(\frac{m}{n}, \frac{\sigma^2}{n}\right)$.

Définition 2.3.6. *La distribution de la variable aléatoire réelle X est dite stable si et seulement si pour tout k et $X_1, \cdots, X_k$, il existe $a_k > 0$ et $b_k \in \mathbb{R}$ tels que,*

$$\mathcal{L}(X_1 + \cdots + X_k) = \mathcal{L}(a_k X + b_k). \tag{2.3.47}$$

Proposition 2.3.5. *Si X est stable, X est infiniment divisible.*

Théorème 2.3.2. *Si X à une distribution stable, alors sa fonction caractéristique s'écrit,*

$$\phi(q) = \begin{cases} \exp\{imt - \gamma|q|(1 + i\beta sign(q)\frac{2}{\pi}\ln|q|)\}, & \alpha = 1 \\ \exp\{imt - \gamma|q|^\alpha(1 - i\beta sign(q)\tan\frac{\alpha\pi}{2})\}, & \alpha \neq 1, \end{cases} \tag{2.3.48}$$

avec $0 < \alpha \leq 2$, il caractérise les queues de distribution, plus α diminue, plus les queues sont épaisses. Si $\alpha = 2$, alors X suit une loi normale.

Le modèle MRW, [BaMuDe01], [BaMu02] repose sur une mesure aléatoire multifractale infiniment divisible, nous allons la constuire. Plaçons nous sur l'espace mesuré $(\mathcal{S}^+, \mu)$ avec $\mathcal{S}^+$ le demi-plan,

$$\mathcal{S}^+ = \{(t, \ell), t \in \mathbb{R}, \ell \in \mathbb{R}_*^+\}, \tag{2.3.49}$$

et μ, une mesure de Haar à gauche (invariante par le groupe de translations-dilatations agissant sur $\mathcal{S}^+$) tel que,

$$\mu(\mathrm{d}t, \mathrm{d}\ell) = \frac{\mathrm{d}t\mathrm{d}\ell}{\ell^2}. \tag{2.3.50}$$

$\mathcal{P}$ est un champ aléatoire infinement divisible $\mathcal{P}$ défini sur le demi-plan $(\mathcal{S}^+, \mu)$ si pour toute famille d'ensemble disjoint $\mathcal{A}_n$ de $\mathcal{S}^+$, $\{\mathcal{P}(\mathcal{A}_n)\}_n$ sont des variables aléatoires indépendantes qui vérifient,

$$\mathcal{P}(\cup_{n=1}^{\infty}\mathcal{A}_n) = \sum_{n=1}^{\infty}\mathcal{P}(\mathcal{A}_n), \;\; p.s.. \tag{2.3.51}$$

De plus, pour tout ensemble $\mathcal{A}$ μ-mesurable, $\mathcal{P}(\mathcal{A})$ est une variable infiniment divisible de fonction caractéristique,

$$\mathbb{E}\left(e^{iq\mathcal{P}(\mathcal{A})}\right) = e^{\varphi(q)\mu(\mathcal{A})}, \tag{2.3.52}$$

pour tout $q \geq 0$ et $\varphi(-iq) < \infty$. $\varphi(q)$ est donné par la formule de Lévy-Khintchine,

$$\varphi(q) = imq + \int \frac{e^{iqx} - 1 - iq\sin x}{x^2}\nu(\mathrm{d}x). \tag{2.3.53}$$

$\nu(\mathrm{d}x)$ est une mesure de Lévy (2.3.45). L'exposant de Laplace est donné par $\psi(q) = \varphi(-iq)$ et nous supposons que,

$$\begin{aligned} \psi(1) &= 0 \\ \psi(q(1+\varepsilon)) &< q(1+\varepsilon) - 1, \forall\varepsilon > 0. \end{aligned} \tag{2.3.54}$$

Enfin, le champs aléatoire $\mathcal{P}$ est défini sur la filtration $\mathcal{F}_\ell$ de l'espace de probabilité Ω,

$$\mathcal{F}_\ell = \sigma\{\mathcal{P}(\mathrm{d}t, \mathrm{d}\ell'); \ell \geq \ell'\}. \tag{2.3.55}$$

Nous pouvons maintenant donner la mesure multifractale aléatoire infiniment divisible.

Définition 2.3.7 (Mesure multifractal aléatoire infiniement divisible)**.** *Le processus* $w = (w_\ell(t); (t, \ell) \in \mathbb{R} \times (0, \infty))$ *est défini pour tout* $(t, \ell) \in \mathbb{R} \times (0, \infty)$ *par,*

$$w_{\ell,T}(t) = \mathcal{P}(\mathcal{A}_{\ell,T}(t)), \tag{2.3.56}$$

où $\mathcal{P}$ est le champ aléatoire infiniment divisible défini précédemment sur le demi-plan $\mathcal{S}^+$ et $\mathcal{A}_\ell(t)$ est la famille de cones donnée par,

$$\mathcal{A}_\ell(t) = \left\{(t', \ell'); \ell \leq \ell', |t' - t| \leq \frac{1}{2}\min(\ell', T)\right\}. \tag{2.3.57}$$

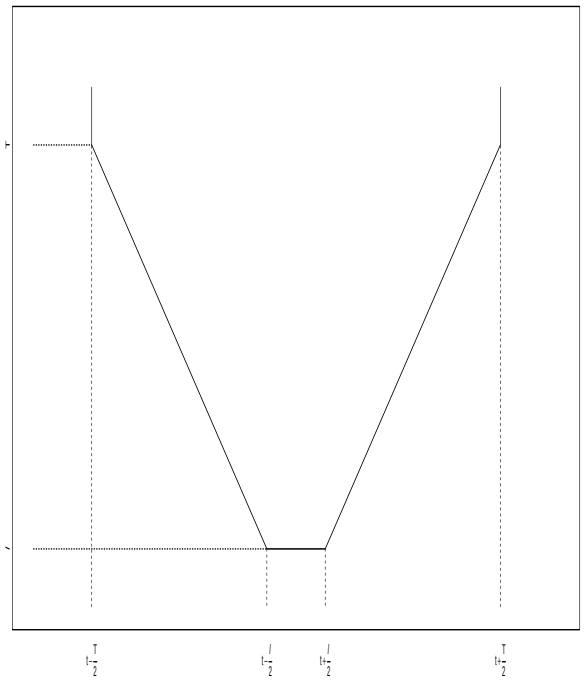


FIGURE 2.4 – Cone $\mathcal{A}_\ell(t)$.

La mesure multifractale aléatoire infiniment divisible M est alors,

$$M(t) = \sigma^2 \lim_{\ell \to 0} \int e^{w_\ell(t)} \mathrm{d}t,\ \sigma > 0. \tag{2.3.58}$$

La marche aléatoire multifractale est, de la même manière que le modèle multifractal pour rendements financiers, un processus subordonné,

$$X(t) = B(M(t)) = \sigma^2 \int_0^t e^{w(t)} \mathrm{d}B(t),\ \ t \geq 0, \tag{2.3.59}$$

avec B un mouvement Brownien standard indépendant de M. Notons que si la mesure de Lévy (2.3.44) ($\nu(\mathrm{d}x)$) est donnée par,

$$\nu(\mathrm{d}x) = 4\lambda^2\delta(x)\mathrm{d}x,\ \lambda^2 > 0, \tag{2.3.60}$$

alors nous pouvons montrer que $\psi(q)$ est la fonction génératrice des cumulants de la loi log-normale (c'est ce qui à priori, est un bon choix pour les distributions financières et, c'est dans ce cadre d'analyse que nous allons rester), $\psi(q) = -2\lambda^2 q + 2\lambda^2 q^2$ et l'exposant multifractal est une parabole,

$$\zeta_M^{\ln}(q) = q(1 + 2\lambda^2) - 2\lambda^2 q^2. \tag{2.3.61}$$

Le processus gaussien $\{w_\ell(t); t \geq 0\}$ est entièrement déterminé par sa moyenne et sa covariance,

$$\begin{aligned} \mathbb{E}(w_\ell(t)) &= -\lambda^2 \left(\ln \frac{T}{\ell} + 1\right) \\ \mathbb{C}\mathrm{ov}(w_\ell(t), w_\ell(t+\tau)) &= \begin{cases} \lambda^2(\ln\left(\frac{T}{l}\right) + 1 - \frac{\tau}{\ell}), & \text{si } 0 \leq \tau \leq \ell \\ \lambda^2 \ln\left(\frac{T}{l}\right), & \text{si } \ell \leq \tau \leq T \\ 0, & \text{si } 0 \leq \tau \leq \infty. \end{cases} \end{aligned} \tag{2.3.62}$$

Par construction, notons que la MRW est bien un processus multifractal,

$$\mathbb{E}(X(t)^q) = \mathbb{E}(B(M(t))^q) = \mathbb{E}(B(1)^q)\mathbb{E}(M(t)^q/2). \tag{2.3.63}$$

On en déduit que l'exposant multifractal de la marche aléatoire multifractale dans le cas log-normale est,

$$\zeta_X^{\ln}(q) = \zeta_M^{\ln}(q/2) = \frac{q}{2}(1 + 2\lambda^2) - \frac{\lambda^2 q^2}{2}. \tag{2.3.64}$$

L'autocovariance du processus semble donner une approximation correcte de celle estimée empiriquement sur les données de rendements financiers puisque nous pouvons montrer que pour le cas où $\tau < k < T$,

$$\mathbb{E}[\delta_\tau X^2(t)\delta_\tau X^2(t+k)] \sum \sigma^2\tau^2 \left(\frac{T}{k}\right)^{4\lambda^2}, \tag{2.3.65}$$

i.e. une loi puissance en k. Il est également possible, après quelques calculs, de voir que la fonction d'autocovariance des incréments d'une MRW en valeurs absolues

suit une loi puissance.

Pour pouvoir simuler le processus, il faut pouvoir simuler des variables aléatoires $\{w(k)\}$, gaussiennes, corrélées. Pour simuler des variables gaussiennes indépendantes, nous supposerons cela connue (cf. algorithme de Box-Muller ou Marsaglia-Bray), expliquons tout de même comment en faire des variables corrélées.

La méthode la plus simple, mais pas forcémenent la plus efficace, est de commencer par construire la matrice d'autocovariance souhaitée, Σ, donc à partir de (2.3.62), bien sûr, nous ne pouvons pas déterminer le paramètre ℓ, on le prendra arbitrairement petit. La matrice Σ étant supposée définie positive, il existe une matrice triangulaire C tel que $C^tC = \Sigma$, alors les composants du vecteur,

$$w = \mu + C^t\mathcal{N}(0, \mathbf{I}), \tag{2.3.66}$$

sont corrélés suivant une loi normale de matrice de covaraiance Σ. La méthode naturelle pour trouver la matrice triangulaire est la factorisation de Cholesky. Il n'est pas très difficile de trouver une librairie informatique capable de faire cette factorisation, néanmoins, même si elles sont en général codées par des experts, la méthode à ses limites (complexité en $\mathcal{O}(N^3)$) et il n'est plus possible de faire cette factorisation pour des matrices trop grandes. La deuxième méthode, qui semble plus compliquée mais qui reste très simple consiste à passer par la transformée de Fourrier rapide [BaLaOpPhTa03].

Venons en à l'estimation des paramètres du modèle, il y en a trois, la volatilité, σ, le coefficient d'intermittence λ^2 et le temps intégral T. L'estimation proposée de $\theta = \{\ln\sigma, \lambda, \ln T\}$ dans [BaKoMu08] repose sur la méthode des moments généralisés (je ne le détaille pas ici, car elle fera l'objet d'un projet).

Pour conclure, notons que la marche aléatoire que nous avons introduite semble être aujourd'hui l'un des modèles les plus performant pour reproduire les faits stylisés de rendements financiers et pour faire de la prévision de volatilité. Néanmoins, comme les processus B et M sont indépendants, l'effet de levier ne peut pas être vérifié. Une contruction alternative, basée sur le calcul de Malliavin, du processus dans le cas dépendant et, plus générale avec un mouvement Brownien fractionnaire au lieu du mouvement brownien standard, ce qui nous permet d'obtenir une persistance plus ou moins prononcée dans les rendements bruts (ce qui est le cas pour certaines commodities, des actifs peu liquides ou pour des données à haute fréquence) est proposée dans le très intéressant article [FaTu12].

2.4 Processus ponctuels

2.4.1 Processus de Hawkes

Pour finir ce chapitre, et cette partie, nous allons voir les processus ponctuels et leurs applications en finance, qui sont très vastes, nous ne verrons que quelques applications possibles. Commençons par quelques définitions théoriques pour introduire le processus.

Définition 2.4.1 (Processus Ponctuel)**.** *Soit t le temps physique et $\{t_i\}_{i\in\{1,2,\cdots\}}$ une suite croissante aléatoire de temps d'arrivées, $0 \leq t_i \leq t_{i+1}$, $\forall i$. Alors la séquence est un processus ponctuel sur le segment $[0,\infty)$.*

Définition 2.4.2 (Processus de Comptage)**.** *Soit $(N(t))_{t\geq 0}$ un processus stochastique à valeurs réelles tel que $N(t) = \sum_{i\geq 1} \mathbb{1}_{\{t_i\leq t\}}$. On dit que N est un processus de comptage si p.s., $N(0) = 0$, N est continu à droite et N est croissant.*

L'exemple certainement le plus connnu est le processus de Poisson, qui en plus des propriétés enoncées vérifie,

- pour tout $t, s \geq 0$, $N(t+s) - N(t)$ suit une loi de Poisson de paramètre λs (stationnarité)
- pour tout $t, s \geq 0$, les variables aléatoires $N(t+s) - N(s)$ indépendant de la sigma-algèbre $\sigma(N_u, u \leq s)$, (accroissement indépendant)

Définition 2.4.3 (Processus de Duration)**.** *Soit d_i le temps d'attente entre deux évements consécutifs,*

$$d_i = \begin{cases} t_i - t_{i-1} & si\ i = 1, 2, \cdots, n \\ t_i & si\ i = 1, \end{cases} \tag{2.4.1}$$

avec $t_0 = 0$. Alors, $\{d_i\}_{i\in\{1,2,\cdots\}}$ est le processus de duration associé au processus ponctuel $\{t_i\}_{i\in\{1,2,\cdots\}}$

Définition 2.4.4 (Intensité)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et N un processus de comptage adapté à la filtration $\mathcal{F}_t$. Le processus d'intensité continue à gauche et limite à droite est défini par,*

$$\begin{aligned} \lambda(t|\mathcal{F}_t) &= \lim_{h\searrow 0} \mathbb{E}\left[\frac{N(t+h) - N(t)}{h} \,|\mathcal{F}_t\right] \\ &= \lim_{h\searrow 0} \mathbb{P}\left[\left\{\frac{N(t+h) - N(t)}{h}\right\} > 0 \,|\mathcal{F}_t\right]. \end{aligned} \tag{2.4.2}$$

La $\mathcal{F}_t$-intensité caractérise l'évolution du processus $N(t)$ conditionellement à sa filtration naturelle. On peut l'interpréter comme la probabilité conditionnelle d'observer un évènement au prochain instant.

Notons qu'un processus ponctuel adapté à la fitration $\mathcal{F}_t$ est une $\mathcal{F}_t$-martingale. Pour tout $0 \leq s \leq t \leq \infty$ nous avons,

$$\mathbb{E}(N(t)|\mathcal{F}_s) = N(s) \quad p.s. \tag{2.4.3}$$

Et en particulier, c'est une surmartingale,

$$\mathbb{E}(N(t)|\mathcal{F}_s) \geq N(s) \quad \text{p.s.} \tag{2.4.4}$$

Alors, d'après la décomposition de Doob-Meyer, toute $\mathcal{F}_t$-surmartingale peut être décomposé (de manière unique) comme la somme d'une $\mathcal{F}_t$-martingale $M(t)$ de moyenne zéro et d'un $\mathcal{F}_t$-processus prédictible croissant $\Lambda(t)$,

$$N(t) = M(t) + \Lambda(t). \tag{2.4.5}$$

La partie prédictible intervenant dans la décomposition de Doob-Meyer est appelée le compensateur et est définie par,

$$\Lambda(t) = \int_0^t \lambda(s|\mathcal{F}_t)\mathrm{d}s. \tag{2.4.6}$$

En réarrangeant l'expression (2.4.5) nous pouvons donner une définition alternative de l'intensité,

$$N(t) - \int_0^t \lambda(s|\mathcal{F}_t)\mathrm{d}s = M(t). \tag{2.4.7}$$

$M(t)$ étant une martingale, nous obtenons,

$$\mathbb{E}[N(t)|\mathcal{F}_s] = \mathbb{E}[\int_0^t \lambda(u|\mathcal{F}_u)\mathrm{d}u|\mathcal{F}_s], \quad \text{p.s.} \tag{2.4.8}$$

Ou encore,

$$\mathbb{E}[N(t) - N(s)|\mathcal{F}_s] = \mathbb{E}[\int_s^t \lambda(u|\mathcal{F}_u)\mathrm{d}u|\mathcal{F}_s] = \mathbb{E}[\Lambda(s,t)|\mathcal{F}_t], \quad \text{p.s.} \tag{2.4.9}$$

avec, $\Lambda(s,t) = \Lambda(t) - \Lambda(s)$.

Théorème 2.4.1 (Changement de temps aléatoire)**.** *Soit $N(t)$ un processus ponctuel adapté à la filtration $\mathcal{F}_t$ avec pour intensité $\lambda(t|\mathcal{F}_t)$ et pour compensateur $\Lambda(t)$ avec $\Lambda(\infty) = \infty$ p.s. Pour tout t, nous définissons le temps d'arrêt $\tau(t)$ comme solution de,*

$$\int_0^{\tau(t)} \lambda(s|\mathcal{F}_t)\mathrm{d}s = t. \tag{2.4.10}$$

Alors, le processus ponctuel,

$$\tilde{N}(t) = N(\tau(t)) = N(\Lambda^{-1}(t)), \tag{2.4.11}$$

est un processus de Poisson homogène d'intensité $\lambda = 1$.

Lemme 2.4.1. *Soit $\tilde{t}_i$ les temps d'arrivés associés au processus ponctuel (transformé) $\tilde{N}(t)$ donnée par (2.4.11), alors,*

$$\tilde{t}_i = \int_0^{t_i} \lambda(s|\mathcal{F}_s)\mathrm{d}s, \quad \textit{pour tout } i. \tag{2.4.12}$$

Théorème 2.4.2. *Les durations $\tilde{t}_i$-$\tilde{t}_{i-1}$ associées au processus ponctuel transformé (2.4.11), $\tilde{N}(t)$ sont données par,*

$$\tilde{t}_i - \tilde{t}_{i-1} = \Lambda(t_i, t_{i-1}) = \int_{t_{i-1}}^{t_i} \lambda(s|\mathcal{F}_s)\mathrm{d}s, \tag{2.4.13}$$

et correspondent à une suite de variable aléatoire exponentielle d'intensité 1 i.i.d.,

$$\Lambda(t_i, t_{i-1}) \sim \textit{i.i.d. } Exp(1). \tag{2.4.14}$$

Ce dernier thèorème se déduit simplement du lemme (2.4.1). Comme nous le verrons par la suite, il va nous permettre de poser un algorithme de simulation d'un processus ponctuel.

Un des premiers papiers à avoir proposé une application des processus de Hawkes est celui de Y. Ogata sur la modélisation des séismes. En finance, ils peuvent, entre autre, servir à modéliser le carnet d'ordres (et exhiber des opportunités d'arbitrage), et par suite, modéliser le cours d'un actif à très haute fréquence, puisque ce sont les variations du carnet d'ordres qui forment le prix.

Les processus de Hawkes font donc partie de la classe des processus ponctuels et sont entièrement caractérisés par leur intensité.

Définition 2.4.5 (Processus de Hawkes)**.** *Soit N un processus de comptage adapté à la filtration $\mathcal{F}_t$ et associé au processus ponctuel $\{t_i\}_{i=1,2,\cdots}$, alors si son intensité conditionnelle est du type,*

$$\lambda(t|\mathcal{F}_t) = \eta + \nu \int_0^t w(t-s)N(\mathrm{d}s), \tag{2.4.15}$$

avec $\eta \geq 0$, $\nu \geq 0$ et $w\mathbb{R}^+ \to \mathbb{R}^+$, le processus N est un processus de Hawkes.

Notons déjà que l'intensité est une intégrale par rapport à une mesure de comptage, nous pouvons donc écrire,

$$\int_0^t w(t-s)N(\mathrm{d}s) := \sum_{t_i<t} w(t-t_i). \tag{2.4.16}$$

Ensuite, comment interpréter les différents paramètres ? η est l'intensité 'fixe' (un peu comme un coût fixe de production), quelle que soit la fréquence instantanée des arrivées, en utilisant le vocabulaire approprié nous devrions dire la fréquence des immigrants. w est le noyau de pénalité, c'est lui qui va donner les variations de l'intensité en fonction des temps d'arrivée. L'expression proposée par Hawkes est de prendre,

$$w(t-t_i) = \alpha e^{-\alpha(t-t_i)}, \tag{2.4.17}$$

avec un $\alpha > 0$. Ansi, plus t_i est loin de l'instant t, plus la fonction tend vers 0,

$$\lim_{(t-t_i)\to\infty} \alpha e^{-\alpha(t-t_i)} = 0, \tag{2.4.18}$$

et à contrario, plus t_i est proche de t, disons par exemple que les deux instants coïncident, $t_i = t$, alors $w(t-t_i) = w(0) = \alpha e^0 = \alpha$, nous avons donc une augmentation temporaire de l'intensité à chaque nouvelle arrivée d'un immigrant. Un autre noyau de pénalité envisageable est de prendre une fonction puissance,

$$w(t) = \frac{(\alpha-1)\beta}{(1+\beta(t-t_i))^\alpha}, \alpha > 2,\ \beta > 0. \tag{2.4.19}$$

Cette fonction correspond au modèle ETAS (Epidemic-Type Aftershock Sequence) pour la modélisation des séismes.

Pour pouvoir modéliser le carnet d'ordre ou les variations d'un actif financier à très haute fréquence, nous avons besoin d'étendre la définition d'un processus au

cas multivarié. En effet, les ordres faisant bouger le cours ne sont pas uniquement des ordres au marché, les ordres limites et d'annulations sont également une partie importante des variations observées.

Définition 2.4.6 (Processus de Hawkes Multivarié)**.** *Soit* $\boldsymbol{N} = (N_1(t), \cdots, N_d(t))$ *un processus de comptage adapté à la filtration* $\mathcal{F}_t$ *associé au processus ponctuel* $(t_{i,1}, \cdots, t_{i,d})$*,* $d \in \mathbb{N}$*, alors si l'intensité conditionnelle est de la forme,*

$$\lambda_j(t|\mathcal{F}_t) = \eta_j + \sum_{k=1}^{d} \nu_{jk} \int_0^t w_j(t-s) N_k(\mathrm{d}s), \tag{2.4.20}$$

alors $(N_1, \cdots, N_d)$ *est un processus de Hawkes multivarié.* $\boldsymbol{\nu} = \{\nu_{ij}\}_{i,j=1,\cdots,d}$ *est la matrice de branchement, c'est elle qui relit les différents types d'intensités entre elles et qui permet les interactions entre les différents processus que l'on cherche à modéliser.*

Sans rentrer dans les détails du pourquoi du comment (voir e.g. [DaVe05]), nous avons besoin d'imposer quelques restrictions pour que le processus soit bien défini,

Théorème 2.4.3 (Condition de Régularité)**.** *Pour qu'un processus de Hawkes multivarié soit bien défini et de manière unique, les conditions suivantes doivent être satisfaites :*

(i)

$$max\{|\lambda| : \lambda \in \Lambda(\boldsymbol{\nu})\} < 1 \tag{2.4.21}$$

(ii)

$$\int_0^\infty t w_j(t) \mathrm{d}t < \infty, \ \textit{pour tout } j = 1, \cdots, d. \tag{2.4.22}$$

En d'autre terme, on peut trouver un espace $(\Omega, \mathcal{F}, \mathbb{P})$ *ayant les conditions suffisantes pour supporter ce type de processus si les conditions précédentes sont vérifiées.*

Enfin, pour calibrer les paramètres du modèle, nous introduisons la vraisemblance du processus de Hawkes

$$\ln L = \sum_{j=1}^{d} \int_0^T \ln \lambda_j(t|\mathcal{F}_t) N(\mathrm{d}t) - \sum_{j=1}^{d} \Lambda_j(T). \tag{2.4.23}$$

Pour contrôler la qualité de la calibration, nous pouvons bien sûr nous servir des théorèmes (2.4.1) et (2.4.2) en traçant les qq-plot correspondants.

Sur la figure (2.5) nous avons fitté un processus de Hawkes sur les rendements de l'ETF SPY en 5 secondes du 4 janvier 2012 vers 11h. Pour bien mettre en évidence le comportement d'un processus de Hawkes, nous n'avons fitté que les rendements supérieurs à 98% de son historique.

Une fois que la calibration souhaitée est faite, nous pourrions avoir envie de simuler un processus de Hawkes correspondant, i.e. simuler le processus ponctuel $\{t_1, \cdots, t_n\}$. Le premier algorithme que nous allons présenter est celui de Shedler-Lewis, la seule difficulté étant la connaissance d'une borne M tel que,

$$\lambda(t|\mathcal{F}_t) \leq M. \tag{2.4.24}$$

Algorithme de Simulation d'un processus de Hawkes :

---

(1) Simulation d'une suite de variable aléatoire $U_1, \cdots, U_i$ i.i.d. uniforme sur (0,1).

(2) En utilisant la méthode d'inversion, on génère des variables aléatoires exponentielles avec $V_i = -\ln(1 - U_i), i = 1, \cdots, n$.

(3) On pose $t_0 = 0$. Si $\Lambda(0, t_1) = V_1$ peut être exprimé comme $t_1 = x_1 = \Lambda(0, V_1)^{-1}$, $t_1 = x_1$. Sinon, on résoud implicitement $\Lambda(t_0, t_1) = V_1$ en t_1.

(4) Pour $i = 2, \cdots, n$, si $\Lambda(t_{i-1}, t_i) = V_i$ peut être exprimé par $x_i = \Lambda(t_{i-1}, t_i)^{-1}$ alors $t_i = t_{i-1} + x_i$. Sinon, on résoud implicitement $\Lambda(t_{i-1}, t_i) = V_i$ en t_i.

(5) Output : $\{t_1, \cdots, t_n\}$.

2.4.2 Carnet d'Ordres

Venons en à la première application, comment modéliser un carnet d'ordres ? Comme nous l'avons vu sur l'animation (**??**), les ordres, quelque soit leur type, arrivent par intermittence, on va donc naturellement les modéliser par des processus ponctuels et plus particulièrement des processus de Hawkes. J. Large ([La07]) propose de modéliser les variations d'un carnet par un processus ponctuel de dimension 10, le but de son étude est de mesurer la résilience du marché, c'est-à-dire sa capacité à revenir à un état d'équilibre après un shock important. En effet,

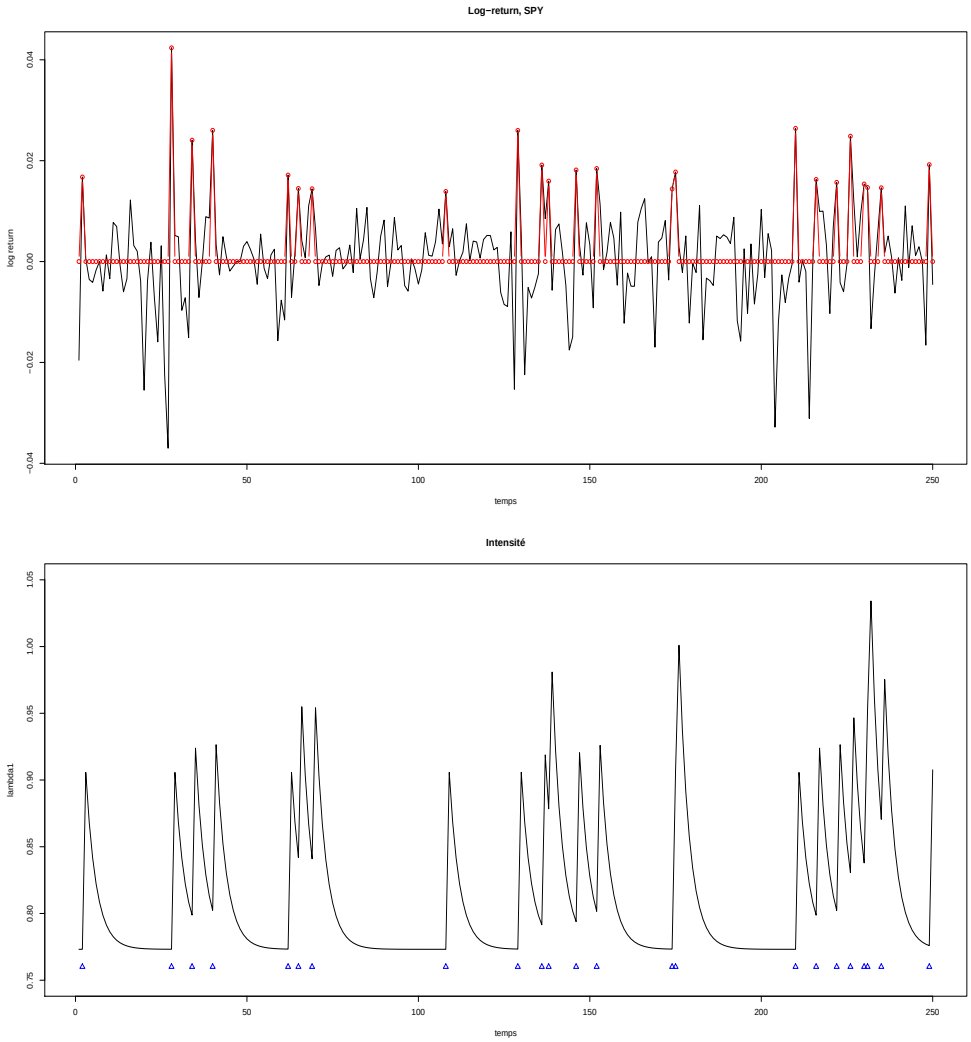


FIGURE 2.5 – Log-rendement de l'ETF SPY en 5 secondes du 4 janvier 2012 vers 11h, en rouge les rendements supérieurs à 98%. Sur la figure de droite, modélisation par un processus de hawkes, les triangles représentants les temps d'arrivés.

bien que chaque ordre arrive de manière non-continue, leurs proportions sont différentes. Sur le tableau (2.1) nous donnons le pourcentage de chaque type d'ordre par rapport à la totalité des ordres sur les actifs composant le CAC 40 entre le 29 octobre et le 26 novembre 1991 ([BiHiSp95]).

	+ ou -	Type d'ordre	%	Vol Moyen
1	Achat	marché ask qui déplace le ask price	3.13	4 830
2	Vente	marché bid qui déplace le bid price	3.70	4 790
3	Achat	limite bid à l'intérieur du spread	4.50	3 390
4	Vente	limite ask à l'intérieur du spread	4.49	3 940
5	Achat	marché ask qui ne déplace pas le ask price	5.01	2 660
6	Vente	marché bid qui ne déplace pas le bid price	4.35	2 550
7	Achat	limite bid inférieur au bid price	21.38	2 840
8	Vente	limite ask supérieur au ask price	19.09	3 110
9	Achat	Annulation d'un ordre bid	17.81	2 740
10	Vente	Annulation d'un ordre ask	16.54	2 820

TABLE 2.1 – Proportion des différents types d'ordres sur les actifs du CAC 40 entre le 29/10 et le 26/11 1991.

Notons bien sur que ces données ne sont qu'une photo et ne représentent pas le marché dans son ensemble, pour une modélisation nous intéressant, nous devrions collecter les données sur le marché qui nous intéresse, sur une période de temps suffisamment large pour être significative, correspondant à différent régime de marché et pas uniquement à un bear market par exemple, et bien sûr, récente.

Comme un processus de comptage est entièrement déterminé par son intensité, le modèle est donné par,

$$\lambda_j(t|\mathcal{F}_t) = \eta_j + \sum_{k=1}^{10} \nu_{jk} \int_0^t w_j(t-s) N_k(\mathrm{d}s). \tag{2.4.25}$$

La calibration du modèle se faisant par maximisation de la vraisemblance (2.4.23) par rapport aux données récoltées. La modélisation d'un carnet d'ordre se fait donc en diffusant chacun des processus N_i, $i = 1, \cdots, 10$, si les données utilisées pour fitter le modèle correspondent à celles présentées dans le tableau (2.1), nous devrions retrouver les mêmes proportions d'ordres.

Un tel raffinement dans le type d'ordre n'est pas forcément nécessaire, ([Mu10]) propose de modéliser le carnet d'ordre selon un modèle à deux agents, l'un impatient et envoyant uniquement des ordres marché, on parle de preneur de liquidité, l'autre patient et envoyant uniquement des ordres limites, on parle d'apporteur de liquidité, il lui arrive donc d'en annuler certains. De plus ce modèle propose de ne pas uniquement s'intéresser aux instants d'arrivées, mais également aux prix et volumes correspondants. Pour être tout à fait exact, ce n'est pas le prix en tant

que tel qui est modélisé mais la distance à laquelle l'ordre est placé par rapport au bid/ask price. La distribution du volume sera prise selon une loi de Student pour coller aux faits stylisés et pour la position de limite nous prendrons une distribution exponentielle.

- Agent Patient
  1. Instant de passage d'ordre limite selon l'intensité λ_L.
  2. Emplacement de la limite selon une loi de Student de paramêtre θ_L.
  3. Volume correspondant à la nouvelle limite selon une loi exponentielle de paramètre m_L.
  4. Instant d'annulation d'ordre selon l'intensité λ_C.
- Agent Impatient
  1. Instant de passage d'ordre marché selon l'intensité λ_M.
  2. Volume correspondant au nouvel ordre selon une loi exponentielle de paramètre m_M.
- Pour les deux agents : Equiprobabilité entre ordre marché bid ou ask.

Les intensités sont données par,

$$\begin{aligned}\lambda_M(t|\mathcal{F}_t) &= \eta_M + \int_0^t \alpha_{MM} e^{-\beta_{MM}(t-s)} N_M(\mathrm{d}s) \\ \lambda_L(t|\mathcal{F}_t) &= \eta_L + \int_0^t \alpha_{LM} e^{-\beta_{LM}(t-s)} N_M(\mathrm{d}s) + \int_0^t \alpha_{LL} e^{-\beta_{LL}(t-s)} N_L(\mathrm{d}s).\end{aligned} \tag{2.4.26}$$

Dans cette modélisation, nous avons donc supposé que les ordres marché arrivaient de manière indépendante des ordres limites, qu'ils s'auto-excitaient uniquement. En revanche, une interaction à été apportée pour les ordres limites qui eux dépendent des ordres marchés. Quand un ordre marché apparaît, l'intensité des ordres limites augmente au même instant, forçant la probabilité qu'un ordre limite soit le prochain évènement ce qui correspond aux comportements des market makeur. Notons que la formulation des intensités diffère légèrement de celle proposée précédemment. En fait, nous avons simplement absorbé le terme ν, c'est-à-dire la matrice de branchement, dans α, pour cela que nous n'avons pas commme noyau de pénalité $\alpha e^{-\alpha t}$ mais $\alpha e^{-\beta t}$ puisque le α n'est pas tout à fait le même.

Une méthodologie possible pour classifier les ordres selon leurs types est la suivante :

1. Si Le volume offert au bid price ou au ask price a diminué, on classifie les ordres comme ordres marché au prix et volume correspondant.
2. Si la quantité offerte à un prix donné a augmenté, l'ordre est considéré comme un ordre limite à un prix et volume correspondant.
3. Si la quantité offerte à un prix donné a diminué, l'ordre est considéré comme étant une annulation à un prix et volume correspondant.

2.4.3 Fluctuation à Très Haute Fréquence

Les modèles de carnet d'ordre devraient nous permettre de déduire des différents types d'ordre les variations à très hautes fréquences de l'actif sous-jacent. Néanmoins, nous pouvons faire beaucoup plus simple et nous verrons une limitation importante de ce type de modèlisation. La variation totale est la somme des variations positives et négatives au cours du temps. Si nous notons $X(t)$ le prix de l'actif à l'instant t, nous avons,

$$X(t) = N_+(t) - N_-(t), \tag{2.4.27}$$

avec N_+ et N_- deux processus de comptage associés respectivement aux sauts positifs et négatifs. Nous suivons les articles [BaDeHoMu12] et [BaDeHoMu11], les intensités de deux processus de Hawkes sont données,

$$\begin{aligned} \lambda_+(t|\mathcal{F}_t) &= \eta_+ + \int_0^t w(t-s)N_+(\mathrm{d}s) \\ \lambda_-(t|\mathcal{F}_t) &= \eta_- + \int_0^t w(t-s)N_-(\mathrm{d}s). \end{aligned} \tag{2.4.28}$$

Nous n'avons ici aucune interaction entre les intensités qui poussent le cours à la hausse et les intensités qui poussent à la baisse. De plus, nous n'avons notifié d'indice sur le noyau de pénalisation, en effet, nous avons $w_+ = w_- = w$, les intensités sont supposées symétriques à la hausse comme à la baisse.

Néanmoins, malgrè sa simplicité, ce modèle permet de reproduire le signature plot caractéristique des données à très haute fréquence. De plus, si nous augmentons la dimension du processus de Hawkes à 4, i.e.,

$$
\begin{aligned}
A : &\begin{cases} \lambda_{A,+}(t|\mathcal{F}_t) = \eta_A + \int_0^t w_+(t-s) N_{A,+}(\mathrm{d}s) \\ \lambda_{A,-}(t|\mathcal{F}_t) = \eta_A + \int_0^t w_-(t-s) N_{A,-}(\mathrm{d}s) \end{cases} \\
B : &\begin{cases} \lambda_{B,+}(t|\mathcal{F}_t) = \eta_B + \int_0^t w_+(t-s) N_{B,+}(\mathrm{d}s) \\ \lambda_{B,-}(t|\mathcal{F}_t) = \eta_B + \int_0^t w_-(t-s) N_{B,-}(\mathrm{d}s). \end{cases}
\end{aligned}
\tag{2.4.29}
$$

Nous avons supposé que les noyaux de pénalisation étaient les mêmes entre actif si à la hausse et si à la baisse. En d'autres termes, les sauts à la baisse (hausse) de l'actif A excitent les sauts à la baisse (hausse) de l'actif B, les intéractions n'existant pas. La diffusion des prix de deux actifs A et B est alors,

$$
X_A = N_{A,+} - N_{A,-}, \quad \text{et} \quad X_B = N_{B,+} - N_{B,-}. \tag{2.4.30}
$$

Les auteurs de [BaDeHoMu11] arrivent alors à montrer que dans ce cas, la covariation du modèle correspond à la covariation des faits stylisés, et que l'effet lead-lag est bien reproduit. De plus, à mesure que l'échelle de temps augmente, la limite du processus tend vers un mouvement Brownien.

Bibliographie

[Ba00] L. Bachelier, *Théorie de la Spéculation*, Annales Scientifiques de l'École Normale Supérieure 3 (17) : 21-86, 1900.

[BaDeHoMu12] E. Bacry, S. Delattre, M. Hoffmann et J-F. Muzy, *Modelling Microstructure Noise with Mutually Exciting Point Processes*, Submitted to Quantitative Finance, 2012, arXiv.org.

[BaDeHoMu11] E. Bacry, S. Delattre, M. Hoffmann et J-F. Muzy, *Scaling Limits for Hawkes Processes and Application to Financial Statistics*, Submitted to Annals of Applied Probability, 2011,arXiv.org.

[BaKoMu08] E. Bacry, A. Kozhemyak et J-F. Muzy, *Continuous Cascade Models for Asset Returns,* Journal of Economic Dynamics and Control, 32(1) :156-199, 2008.

[BaMu02] E. Bacry et J-F. Muzy, *Multifractal Stationary Random Measures and Multifractal Walks with Log-Infinitely Divisible Scaling Laws*, Phys. Rev. E, 66, 056121, 2002,arXiv.org.

[BaMuDe01] E. Bacry, J.F. Muzy et J. Delour, *Multifractal Random Walks*, Phys. Rev. E 64, 026103-026106, 2001, 2011,arXiv.org.

[BaBoMi96] R.T. Baillie, T. Bollerslev et H.O. Mikkelsen, *Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity*, Journal of Econometrics, 1996.

[BaLaOpPhTa03] J.M. Bardet, G. Lang, G. Oppenheim, A. Philippe et M. Taqqu, *Generators of Long-Range Dependent Processes : A Survey*, Long-range Dependence : Theory and Applications, Birkhauser, 2003.

[BiHiSp95] B. Biais, P. Hillion, et C. Spatt, *An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse*, Journal of Finance 50, 1655-1689, 1995.

[Bo86] T. Bollerslev, *Generalized Autoregressive Conditional Heteroskedasticity*, Journal of Econometrics, 31 :307-327, 1986.

[BoEnNe94] T. Bollerslev, R. Engle et D. Nelson, *ARCH Models*, Handbook of Econometrics, vol. IV, 1994.

[BoMePo02] J. P. Bouchaud, M. Mezard, et M. Potters, *Statistical properties of stock order books : empirical results and models*, Quantitative Finance, 2(4), 2002, arXiv.org.

[BoPo04] J.P. Bouchaud et M. Potters, *Theory of Financial Risk and Derivative Pricing*, Cambridge University Press, 400, 2003.

[CaFi01] L. Calvet et A. Fisher, *Forecasting Multifractal Volatility*. Journal of Econometrics 105 : 27-58, 2001.

[CaFi04] L. Calvet et A. Fisher, *How to Forecast Long-Run Volatility : Regime-Switching and the Estimation of Multifractal Processes*. Journal of Financial Econometrics 2 : 49-83, 2004.

[ChMuPaAd11] A. Chakraborti, I. Muni Toke, M. Patriarca, F. Abergel, *Econophysics Review : I. Empirical facts*, Quantitative Finance, vol.11, no.7, 991-1012, 2011, arXiv.org.

[Cl73] P.K. Clark, *A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices*, Econometrica 41, 135-156, 1973.

[GeDaMuOlPi01] R. Gencay, M. Dacorogna, U. Muller, R. Olsen, O. Pictet, *An Introduction to High-Frequency Finance*, Academic Press, New-York, 383 pp, 2001.

[DaVe05] D.J. Daley et D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer series in statistics, 2005.

[En82] R. Engle, *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation*, Econometrica, Vol. 50, No. 4, pp. 987-1007, 1982.

[FaTu12] A. Fauth et C. Tudor, *Multifractal Random Walk with Fractionnal Brownian Motion via Malliavin Calculus*, soumis, 2012.

[FiCaMa97] A. Fisher, L. Calvet, et B. Mandelbrot, *Multifractality of Deutsche-mark / US Dollar Exchange Rates*, Discussion Paper 1166, Cowles Foundation, 1997, SSRN.

[Go98] G. González-Rivera, *Smooth Transition GARCH models*, Studies in Nonlinear Dynamics and Econometrics, 3, 61-78, 1998.

[GoPlAmMeSt99] P. Gopikrishnan, V. Plerou, L.A. Amaral, M. Meyer et H.E Stanley, *Scaling of the Distribution of Fluctuations of Financial Market Indices*, Phys. Rev. E, 60, 5305-5316, 1999, arXiv.org.

[GlJaRu93] L. Glosten, R. Jagannathan et D. Runkle, *On the Relationship Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks*, Journal of Finance, 48, 1779-1801, 1993.

[Ha12] N. Hautsch, *Econometrics of Financial High-Frequency Data*, Springer, Berlin, 2012.

[Hu10] H. Hurst, *Long-Term Storage Capacity in Reservoirs*, Trans. Amer. Soc. Civil Eng. 166, 400-410, 1951.

[Ko40] A. N. Kolmogorov, *The Wiener Spiral and Some Other Interesting Curves in Hilbert Space*, Dokl. Akad. Nauk SSSR. 26 :2, 115-118. (Russian), 1940.

[La07] J. Large, *Measuring the Resiliency of an Electronic Limit Order Book*, Journal of Financial Markets 10, no. 1, 1-25, 2007.

[Le12] C.-A.Lehalle, *Market Microstructure Knowledge Needed to Control an Intra-Day Trading Process*, Forthcoming in Handbook on Systemic Risk, 2012.

[Lu08] T. Lux, *The Markov-Switching Multifractal Model of Asset Returns : GMM Estimation and Linear Forecasting of Volatility*. Journal of Business and Economic Statistics 26 (2) : 194-210, 2008.

[MaFiCa97] B. Mandelbrot, A. Fisher, et L. Calvet, *A Multifractal Model of Asset Returns*, Discussion Paper 1164, Cowles Foundation, 1997, SSRN.

[MaNe68] B. Mandelbrot et J. Van Ness, *Fractional Brownian Motions, Fractional Noises and Applications*, SIAM Review 10, 422-437, 1968.

[MaSt07] R.N. Mantegna et H.E. Stanley, *An Introduction to Econophysics : Correlations and Complexity in Finance*, Cambridge University Press, Cambridge, UK, 2000.

[MoViMoGeFaVaLiMa09] E. Moro, J. Vicente, L.G. Moyano, A. Gerig, J.D. Farmer, G. Vaglica, F. Lillo, et R.N. Mantegna, *Market Impact and Trading Profile of Large Orders in Stock Markets*. Phys. Rev. E, 80, 066102, 2009, arXiv.org.

[Mu10] I. Muni Toke, *"Market Making" in an Order Book Model and its Impact on the Bid-Ask Spread*, in Econophysics of Order-Driven Markets, New Economic Windows, Springer-Verlag Milan, 2010, arXiv.org.

[Ne91] D. Nelson, *Conditional Heteroskedasticity in Asset Returns : A New Approach*, Econometrica 59 : 347-370, 1991.

[Vo05] J. Voit, *The Statistical Mechanics of Financial Markets*, Springer, 3rd ed., 378, 2005.

Deuxième partie

Arbitrage Statistique

Cette partie sera faite en cours, je ne donne ici que quelques idées basiques. Les 'stratégies' ne sont pas écrites ici.

2.5 Performance

Définition 2.5.1. *Soit $(X(t), t \geq 0)$ le cours de l'actif financier et $\pi(t)$ la position à l'instant t. Si $\pi(t) = 1$ (-1) on dira qu'on se met "Long" ("Short") sur le marché si on achète (vend à découvert) l'actif financier, il s'agit d'un pari à la hausse (baisse). Si $\pi(t) = 0$ on est "Flat", aucune décision n'est prise à l'instant t. On peut interpréter de deux manières cette dernière position, soit on reste sur la position déterminée en $t-1$, soit on dénoue entièrement son portefeuille. De la même manière, si $\pi(t-1) = \pi(t) \neq 0$, on peut soit conserver la pose initiée en $t-1$, soit cumuler la position.*

Rappelons que le log-rendement de l'actif à l'instant t est donné par, $\delta_\tau X(t) = \log\left(\frac{X(t)}{X(t-\tau)}\right)$ et $\delta_\tau \mathbf{X} = (\delta_\tau X(1), \cdots, \delta_\tau X(t))$ le vecteur des rendements sur l'ensemble de l'historique, $\pi(t)$ la position donnée par la stratégie à l'instant t et $\boldsymbol{\pi} = (\pi(1), \pi(2), \cdots, \pi(t))$ l'ensemble des positions sur l'historique. Pour estimer la qualité de la stratégie nous pouvons déterminer (entre autres),

$$\begin{aligned}
\text{moyenne des rendements} &:= \frac{1}{t}\sum_{s=1}^{t} \pi(s)\delta_\tau X(s) \\
\text{écart type} &:= \frac{1}{t}\sqrt{\sum_{s=1}^{t}(\pi(s)\delta_\tau X(s) - \overline{\boldsymbol{\pi}\delta_\tau \mathbf{X}})^2} \\
\%\text{ de positions corrects} &:= \frac{100}{t}\sum_{s=1}^{t} \mathbb{1}_{\pi(s)=sign(\delta_\tau X(s))} \\
\text{max drawdown} &:= \max_{0\leq s\leq t}\left\{\max_{0\leq s\leq t}\left\{\sum_{s=1}^{t}\pi(s)\delta_\tau X(s)\right\} - \sum_{s=1}^{t}\pi(s)\delta_\tau X(s)\right\}.
\end{aligned} \tag{2.5.1}$$

L'écart type donne le risque de la stratégie, au sens de Markowitz, s'il est appliqué directement sur l'actif il nous donne sa volatilité (du moins c'est une manière de calculer la volatilité d'un actif). Le max-drawdown représente la plus grosse perte encaissée au cours de la période de trading. Un moyen visuel et rapide de se faire une idée est de tracer les rendements cumulés, l'*equity curve*, de la stratégie,

Définition 2.5.2 (Ratio de Sharpe)**.** *La solution la plus simple et la plus courante pour déterminer le risque est de calculer l'écart type des rendements. Le ratio de Sharpe est construit pour maximiser les rendements et pénaliser le risque,*

$$\max_{\Lambda_t} \frac{\overline{\boldsymbol{\pi}\delta_\tau \mathbf{X}} - \overline{\boldsymbol{\pi}^b \delta_\tau \mathbf{X}}}{\sqrt{Var(\boldsymbol{\pi}\delta_\tau \mathbf{X})}}. \tag{2.5.2}$$

$\boldsymbol{\theta}^b$ *est la stratégie benchmark, la stratégie non risquée, plus généralement, il s'agit de la stratégie définie par le gestionnaire comme celle à "battre".*

L'algorithme sélectionnera uniquement une stratégie qui a des rendements supérieurs au benchmark (numérateur), et qui n'aura pas un risque trop élevé (dénominateur).

Le problème du ratio de Sharpe est qu'il pénalise aussi bien les fortes variations de rendements négatifs que positifs.

Définition 2.5.3 (Ratio de Sortino)**.** *Le ratio de Sortino (dans sa version simplifiée) est équivalent au ratio de Sharpe dans le sens que l'on cherche à maximiser son espérence de rendement tout en pénalisant le risque de la stratégie. A la différence que le risque n'est pas défini comme l'écart type des rendements mais comme l'écart type des rendements uniquement négatifs.*

$$\max_{\Lambda} \frac{\overline{\boldsymbol{\pi}\delta_\tau \mathbf{X}} - \overline{\boldsymbol{\pi}^b \delta_\tau \mathbf{X}}}{\sqrt{Var(\boldsymbol{\pi}\delta_\tau \mathbf{X}\mathbb{1}_{\pi(s)\delta_\tau X<0})}}. \tag{2.5.3}$$

De manière plus formelle on cherche à résoudre le problème,

$$\mathcal{P} : \begin{cases} \max_{\Lambda} [\overline{\boldsymbol{\pi}\delta_\tau \mathbf{X}} - \overline{\boldsymbol{\pi}^b \delta_\tau \mathbf{X}}] \\ s.c. \ \sqrt{\mathrm{Var}(\boldsymbol{\pi}\delta_\tau \mathbf{X}\mathbb{1}_{\pi(s)\delta_\tau X<0})} \leq k \end{cases} \tag{2.5.4}$$

Le lagrangien associé est alors,

$$\max_{\Lambda} \left\{ [\overline{\boldsymbol{\pi}\delta_\tau \mathbf{X}} - \overline{\boldsymbol{\pi}^b \delta_\tau \mathbf{X}}] - \sqrt{\mathrm{Var}(\boldsymbol{\pi}\delta_\tau \mathbf{X}\mathbb{1}_{\pi(s)\delta_\tau X<0})} \right\}. \tag{2.5.5}$$

Le multiplicateur de lagrange, λ, peut être interprété comme le paramètre d'aversion au risque, libre au gestionnaire de l'ajuster selon son niveau d'aversion au risque.

Notons que cette phase d'optimisation ne peut être faite sur l'ensemble de l'historique disponible au risque du sur-apprentissage, *over-fitting*, i.e. d'avoir un jeu de paramètres très, trop, précis, correspondant uniquement à l'historique passé et ne se reproduisant plus (on donnera une définition plus rigoureuse du sur-apprentissage par la suite). Une première solution est de faire une optimisation séquentielle.

Bien sûr, il faut rajouter les fees, i.e. les coûts de transaction du broker ainsi que le spread bid-ask. Le spread n'est pas à rajouter si l'on dispose de l'historique bid et ask. Pour un ordre marché, on supposera simplement qu'il est tout le temps exécuté au ask price pour un ordre d'achat et au bid price pour un ordre de vente.

Il faut rajouter le temps de latence (voir tableau 2.2),

Départ	Arrivée	Temps (ms)
Paris	Londres	20
Paris	New York	110
Paris	Tokyo	300
Paris	Sydney	320

TABLE 2.2 – Temps de latence entre les principales places boursières et Paris.

Le plus simple étant bien sûr de louer un box dans le coeur de la place boursière qui nous intéresse pour diminuer drastiquement le temps de latence. Si nous recevons du flux en temps réel, le temps de le recevoir, de faire tourner les algorithmes qui vont prendre les décisions et, de renvoyer un ordre, nous perdons nécessairement du temps, le cours n'est donc plus le même. Si l'on envoie un ordre marché au broker, nous sommes sûr d'être exécuté, néanmoins, supposons l'ordre comme un ordre d'achat, pas au ask price sur lequel nous avons fait les calculs, mais sur le ask price au moment ou l'ordre aura été reçu. Il faut bien être conscient de cela en faisant les backtests, en estimant correctement le temps de latence pour pouvoir estimer le ask price effectivement traité.

Pour les ordres limites, la tâche est un peu plus compliquée. Notons $p_{ask}(t)$ le ask price et $p_{bid}(t)$ le bid price à l'instant où nous avons reçu l'information. Nous envoyons un ordre limite de vente à $p_{ask}(t)+k$tick, il y a deux principales approches

pour déterminer si nous aurions été exécuté ou non, le type 'defaut' ou 'libéral'.

Le type libéral est le plus avantageux, il suppose une exécution à partir du moment ou le prix à été touché au moins un fois, seulement il peut être rencontré mais ce n'est pas pour autant que notre ordre serait passé pour la simple et bonne raison que le volume correspondant à cette limite n'a pas totalement été absorbé. Si l'on se situe en bas de la pile, et que seulement une partie du volume a été absorbée, la priorité temps ne nous permet pas de passé.

Le deuxième type, 'defaut', est plus contrariant, il suppose que l'ordre aurait été exécuté ssi la totalité du volume a été prise. En d'autre terme, nous sommes exécutés à $p_{ask}(t) + k\text{tick}$ ssi le cours de l'actif est passé au moins à $p_{ask}(t) + (k+1)\text{tick}$, on parle de pénétration de la limite.

Pour plus de précision dans la réalisation du backtest, nous devrions avoir donc en plus de l'historique bid-ask, l'historique complet des volumes échangés pour chacune des limites.

2.6 Théorie du Signal

On présente ici les méthodes de filtrations d'un signal (un cours d'actif) erratiques. Il s'agit de 'lisser' les choses. Le méthode la plus baisque consiste à faire un moyenne mobile. On peut aller plus loin en utilisant les méthodes d'ondelettes, Feynman-Kack, etc. Cette partie sera fait en cours !

2.7 Agrégation des Stratégies

Finalement, il est 'facile' de produire des stratégies de trading, chacune donnant une performance plus ou moins bonne selon la période, selon l'actif, la phase de backtest permet de sélectionner son panier d'indicateurs. La question que l'on peut facilement se poser est comment prendre une décision à l'instant t face à $\pi_1(t), \pi_2(t), \cdots, \pi_K(t)$, i.e. K stratégies de trading.

L'idée la plus simple est de prendre la position revenant le plus souvent à l'instant t,

$$\hat{\pi}(t) = \text{sign}\left(\sum_{k=1}^{K} \pi_k(t)\right). \tag{2.7.1}$$

Cette méthode est dite *vote majoritaire*, sur la figure (2.6) nous avons testé 14 stratégies de trading et leur agrégation par vote majoritaire (2.7.1).

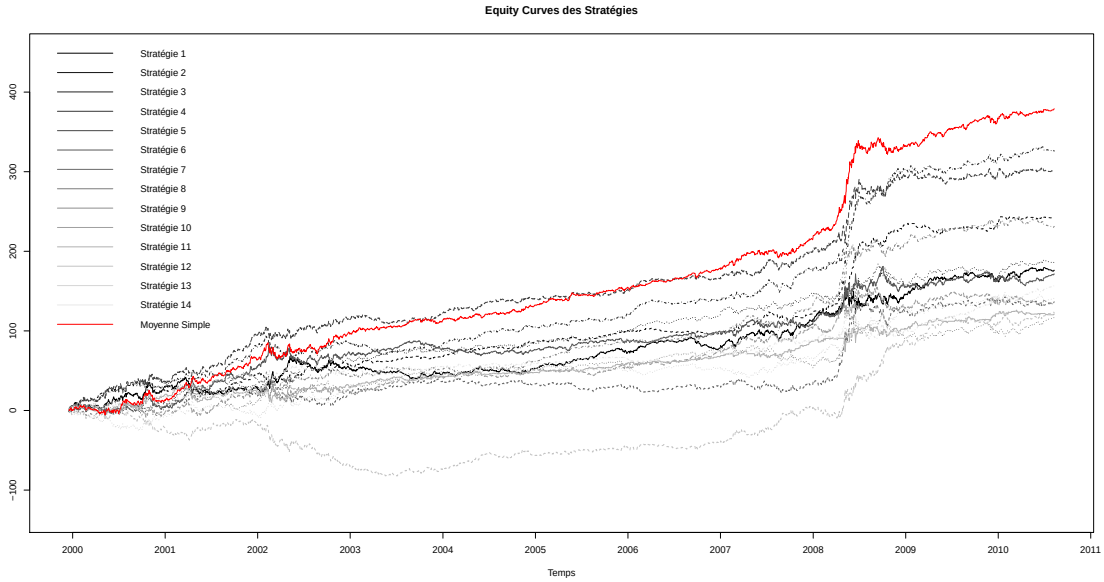


FIGURE 2.6 – 14 stratégies de trading appliquées au SPY du 31/08/2006 au 21/04/2011, journalier, 1169 points et leur agrégation par vote majoritaire.

On peut étendre cette idée en pondérant chacune des stratégies par leur performance passée,

$$w_k(t) = \begin{cases} (1+\lambda)w_k(t-1) & \text{si } \pi_k(t-1) = \text{sign}(\delta_\tau X(t-1)) \\ (1-\lambda)w_k(t-1) & \text{si } \pi_k(t-1) \neq \text{sign}(\delta_\tau X(t-1)) \end{cases} \quad \lambda \in [0,1]. \tag{2.7.2}$$

La règle de décision est alors,

$$\hat{\pi}(t) = sign\left(\frac{1}{K}\sum_{k=1}^{K} p_k(t)\pi_k(t)\right) \qquad \text{où} \qquad p_k(t) = \frac{w_k(t)}{\sum_{j=1}^{K} w_j(t)}. \tag{2.7.3}$$

Ce type d'algorithme peut se référer à la théorie des jeux.

Nous appellerons maintenant chacune des stratégies des *experts* ou des *bras*. Nous notons $\mathcal{D}$ l'espace de décision, $\pi_k(t) \in \mathcal{D}$ et $\mathcal{Y}$ l'espace d'arrivée, $\delta_\tau X(t) \in \mathcal{Y}$. A chaque instant, l'agent choisit un bras $\pi_k(t)$ et observe une *perte*, $\ell_k(t) = \ell(\pi_k(t), \delta_\tau X(t)) : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ et $\hat{\ell}_k(t) = \ell(\hat{\pi}_k(t), \delta_\tau X(t))$ la perte de la stratégie agrégée, ℓ est la fonction de perte. L'objectif est de minimiser le regret cumulé de la stratégie,

$$\mathcal{R}_k(t) = \hat{L}(t) - \min_{k=1,\cdots,K} L_k(t), \tag{2.7.4}$$

où $L_k(t) = \sum_{s=1}^{t} \ell_k(t)$ est la perte cumulée de l'expert k à l'instant t et $\hat{L}(t)$ la perte de la stratégie agrégée. On parle de regret puisque à l'instant t nous appliquons $\hat{\pi}$ alors que dans l'ensemble des stratégies, peut être que l'une d'entre elles avait la bonne décision, la perte minimum, $\min \ell$. Nous noterons également $r_k(t)$ le regret non-cumulé de la stratégie k à l'instant t, $r_k(t) = \hat{\ell}(t) - \min_{k=1,\cdots,K} \ell_k(t)$.

Le type d'algorithme présenté ici est dit d'exploration/exploitation, l'agent ne connaît pas les lois des différents experts, il doit donc acquérir de l'information en étudiant l'historique (exploration) pour pouvoir agir de manière optimale (exploitation).

Définition 2.7.1 (Prédiction avec avis d'experts)**.** *Le protocole de prédiction peut être naturellement interprété comme le jeu repété entre le prévisionniste (le gestionnaire) qui joue $\hat{\pi}(t)$ et "l'environnement" (le marché) qui révèle $\delta_\tau X(t)$.*

Le jeu le plus simple est de prendre le meilleur expert à l'instant t par rapport aux pertes cumulées,

$$\hat{\pi}(t) = \pi_k(t) \qquad \text{si } k = \arg\min_{k' \in \mathcal{K}} \sum_{s=1}^{t-1} \ell_{k'}(t). \tag{2.7.5}$$

Dans le cas de multiples minimiseurs, on choisit k de manière aléatoire.

Une famille d'algorithme susceptible de répondre à notre problématique est la classe des algorithmes de *bandit multi-armes contre adversaire. Bandit multi-armes* car on se réfère au jeu de bandit-manchot, le joueur place sa mise dans plusieurs machines possibles et son but est de gagner au moins autant que s'il avait connu

Algorithme : Prédiction avec avis d'experts
Paramètres : $\mathcal{D}$, $\mathcal{Y}$, ℓ, ensemble $\mathcal{K}$ des experts.

Pour tout tour $t = 1, 2, \cdots$

(1) L'environnement choisit le prochain rendement $\delta_\tau X(t)$ et les avis d'experts $\{\pi_k(t) \in \mathcal{D} : k \in \mathcal{K}\}$; les avis sont révélés au prévisionniste.
(2) Le prévisionniste choisit la prédiction $\hat{\pi}(t) \in \mathcal{D}$
(3) L'environnement révèle le rendement $\delta_\tau X(t) \in \mathcal{Y}$
(4) Le prévisionniste prend les pertes $\hat{\ell}_k(t)$ et chaque expert $\ell_k(t)$

en avance quelle machine aurait apporté le meilleur gain. *Contre adversaire*, i.e. quand l'environnement distribue des récompenses non i.i.d. Nous avons deux cas distincts, le premier est dans le cadre d'information parfaite, i.e. que l'on connaît les récompenses, ou pertes, de chacun des experts, le second dans le cas d'information partielle, i.e. que l'on ne connaît pas toutes les récompenses, ou pertes, de tous les experts sur un horizon 'infini'. Nous allons présenter seulement un cas correspondant à l'information complète.

Information Complète. Comme pour (2.7.2), une stratégie naturelle d'agrégation pour notre problématique est de construire une agrégation par moyenne pondérée des experts, nous donnons cette fois un cadre plus formel. La prédiction prend la forme,

$$\hat{\pi}(t) = \frac{\sum_{k=1}^{K} w_k(t-1)\pi_k(t)}{\sum_{j=1}^{K} w_j(t-1)}, \tag{2.7.6}$$

où $w_1(t-1), w_2(t-1), \cdots, w_K(t-1) \geq 0$ sont les poids associés aux experts à l'instant t avec $w_k(t) = \nabla\Phi(R_k(t))$. $\Phi(\cdot) : \mathbb{R} \to \mathbb{R}$ est le potentiel, fonction positive, convexe et croissante de la forme,

$$\Phi(\mathbf{u}) = \varphi\left(\sum_{k=1}^{K} \phi(u_i)\right), \tag{2.7.7}$$

où η est un paramètre positif, $\phi : \mathbb{R} \to \mathbb{R}$ une fonction positive, croissante, deux fois différentiable et $\varphi : \mathbb{R} \to \mathbb{R}$ positive, strictement croissante, concave et deux fois différentiable.

Lemme 2.7.1. *Si la fonction de perte est convexe en son premier argument,*

$$\sup_{y(t)\in\mathcal{Y}} \sum_{k=1}^{K} r_k(t)\phi'(R_i(t-1)) \leq 0. \tag{2.7.8}$$

Preuve. Avec le l'inégalité de Jensen on a directement,

$$\ell(\hat{\pi}(t), y(t)) = \ell\left(\frac{\sum_{k=1}^{K}\phi'(R_k(t-1))\pi_k(t)}{\sum_{k=1}^{K}\phi'(R_k(t-1))}, y(t)\right) \leq \frac{\phi'(R_k(t-1))\ell(\pi_k(t), y(t))}{\phi'(R_k(t-1))}. \tag{2.7.9}$$

□

Théorème 2.7.1. *Supposons que l'agent satisfait la condition de Blackwell, i.e.*

$$\sup_{y(t)\in\mathcal{Y}} r(t)\nabla\Phi(R(t-1)) \leq 0, \tag{2.7.10}$$

(la preuve est similaire à celle du lemme 2.7.1) alors pour tout $t = 1, 2, \cdots,$

$$\Phi(R(t)) \leq \Phi(0) + \frac{1}{2}\sum_{s=1}^{t} C(r(s)), \tag{2.7.11}$$

où

$$C(R(t)) = \sup_{u\in\mathbb{R}^K} \varphi'\left(\sum_{k=1}^{K}\phi(u_k)\right)\sum_{k=1}^{K}\phi''(u_k)r_k^2(t). \tag{2.7.12}$$

Preuve. D'après le théorème de Taylor on a,

$$\begin{aligned}
\Phi(\delta_\tau X(t)) &= \Phi(R(t-1) + \delta_\tau X(t)) \\
&= \Phi(R(t-1)) + \nabla\Phi(R(t-1)r(t)) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\partial^2\Phi}{\partial u_i\partial u_j}|_\xi r_i(t)r_j(t) \\
&\leq \Phi(R(t-1)) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\partial^2\Phi}{\partial u_i\partial u_j}|_\xi r_i(t)r_j(t).
\end{aligned} \tag{2.7.13}$$

□

Prenons l'exemple du potentiel exponentiel défini par,

$$\Phi_\eta(\mathbf{u}) = \frac{1}{\eta}\ln(\sum_{k=1}^{K} e^{\eta u_i}). \tag{2.7.14}$$

Il s'agit de la stratégie de *moyenne pondérée exponentiellement*. Dans ce cas, les poids prennent la forme,

$$w_k(t-1) = \nabla\Phi(R_k(t-1)) = e^{\eta R_k(t-1)}. \tag{2.7.15}$$

La prévision s'écrit alors,

$$\begin{aligned}\hat{\pi}_t =& \frac{\sum_{k=1}^K e^{\eta(\hat{L}(t-1)-L_k(t-1))}\pi_k(t)}{\sum_{j=1}^K e^{\eta(\hat{L}(t-1)-L_j(t-1))}} \\ =& \frac{\sum_{k=1}^K e^{-\eta L_k(t-1)}\pi_k(t)}{\sum_{j=1}^K e^{-\eta L_j(t-1)}},\end{aligned} \tag{2.7.16}$$

ce que l'on peut réécrire sous la forme $\hat{\pi}_t = \sum_{k=1}^K w_k(t-1)\pi_k(t)$ avec,

$$w_k(t) = \frac{w_k(t-1)e^{-\eta l(\pi_k(t),r(t))}}{\sum_{j=1}^K w_j(t-1)e^{-\eta l(\pi_j(t),r(t))}}. \tag{2.7.17}$$

Proposition 2.7.1. *Supposons que la fonction de perte, ℓ, est convexe en son premier argument et prend des valeurs dans [0,B]. Pour tout K et $\eta > 0$, le regret de la stratégie vérifie,*

$$\hat{L}(T) - \min_{k=1,\cdots,K} L_k(T) \leq \frac{\ln K}{\eta} + \frac{T\eta}{8}B^2. \tag{2.7.18}$$

Preuve. Pour tout $t > 0$, $k = 1, \cdots, K$, on note $w_{k,t} = e^{-\eta L_k(t-1)}$ et $W(t) = w_1(t) + \cdots + w_K(t)$, donc $p_k(t) = \frac{w_k(t)}{W(t)}$. Alors,

$$\begin{aligned}\ln\frac{W(T+1)}{W(1)} &= \ln\left(\sum_{k=1}^K e^{-\eta L_k(T)}\right) - \ln K \\ &\geq \ln\left(\max_{j=1,\cdots,K} e^{-\eta L_k(T)}\right) \\ &= -\eta \min_{k=1,\cdots,K} L_k(T) - \ln K.\end{aligned} \tag{2.7.19}$$

Pour continuer on a besoin d'introduire le lemme de Hoeffding-Azuma.

Lemme 2.7.2. *Soit X une variable aléatoire tel que $a \leq X \leq b$. Alors pour tout $s \in \mathbb{R}$,*

$$\ln \mathbb{E}(e^{sX}) \leq s\mathbb{E}X + \frac{s^2(b-a)}{8}. \tag{2.7.20}$$

Ainsi,

$$\begin{aligned}\ln \frac{W_{t+1}}{W(t)} &= \ln \frac{\sum_{k=1}^{K} e^{-\eta \ell_k(T)} e^{-\eta L_k(T-1)}}{\sum_{k=1}^{K} e^{-\eta L_k(t-1)}} \\ &= \ln \left(\sum_{k=1}^{K} p_k(t) e^{-\eta \ell_k(t)} \right) \\ &\leq -\eta \sum_{k=1}^{K} p_k(t) \ell_k(t) + \frac{\eta^2}{8} B^2.\end{aligned} \tag{2.7.21}$$

en sommant pour $t = 1, \cdots, T$ on trouve,

$$\ln \frac{W(T+1)}{W(1)} \leq \frac{T\eta^2}{8} B^2 - \eta \sum_{t=1}^{T} \tilde{\ell}(t). \tag{2.7.22}$$

On conclut la preuve en combinant la borne supérieure (2.7.22) avec la borne inférieure (2.7.19). □

Nous renvoyons à [CeLu06] et [AuCeFrSc01] pour plus d'algorithmes d'agrégation du même type.

Chapitre 3

Apprentissage Statistique

Dans ce chapitre nous allons présenter des algorithmes susceptibles d'apprendre une propriété pour pouvoir la reproduire. Par exemple, un taux de chômage 'faible', ou au moins plus bas qu'attendu, pousse l'indice du pays en question à la hausse, c'est une règle plus ou moins vérifiée empiriquement. Donc, dès que les chiffres du taux de chômage vont sortir, nous devrions voir les indices du pays soit s'envoler, soit dévisser, en fonction du taux. Il y a énormément de chiffre macroéconomique qui ont un impact sur les marchés financiers, citons entre autre l'inflation, CPI, GBP, etc., il est naturellement très délicat pour un gestionnaire discrétionnaire de prendre en compte toutes les informations, de réagir instantanément, et correctement. Les algorithmes d'apprentissage statistique se proposent de le faire.

Notons $\mathcal{X}$ un ensemble d'input, $\mathcal{X} = \{X_1, X_2, \cdots, X_d\} \in \mathbb{R}^{t\times d}$, X_i=$(x_i(1), \cdots, x_i(t))$, c'est l'ensemble des variables explicatives, et $\mathcal{Y}$ l'output, $\mathcal{Y} = (Y_1, \cdots, Y_u) \in \mathbb{R}^{t\times u}$, $Y_i = (y_i(1), \cdots, y_i(t))$, il s'agit de la cible, ce que l'on cherche à apprendre. Selon la problématique, $\mathcal{Y}$ pourra être les rendements de l'actif que l'on cherche à reproduire, le cours brut, ou uniquement le signe des rendements. Comme nous venons de le dire, $\mathcal{X}$ pourra être rempli de variables macroéconomiques, mais aussi des retards de $\mathcal{Y}$, d'autres d'actif financier, d'avis d'experts, diverse stratégies de trading. A la différence d'une simple agrégation par vote majoritaire ou par moyenne pondérée, nous pourrons essayer de trouver des règles de décisions en ne regardant pas uniquement directement les directions données mais les valeurs des indicateurs, l'algorithme se chargera de trouver des relations, du moins, s'il en trouve.

Le premier outil mathématique présenté est le réseau de neurones multicouche

feed forward, (multilayer perceptron feed-forward) et nous verrons qu'il peut s'écrire comme un problème de classification ou de prévision.

3.1 Réseau de Neurones

3.2 Perceptron Multicouche

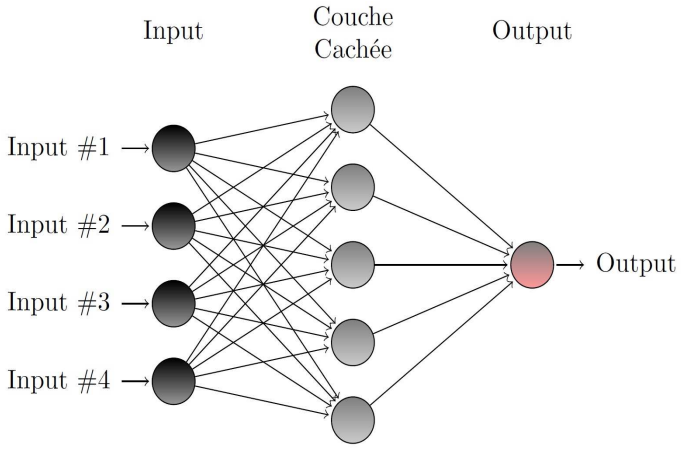


FIGURE 3.1 – Structure typique d'un réseau de neurones à 2 couches, 4 inputs et 5 neurones sur la couche cachée.

Un MLP (Multilayer Perceptron) est organisé en couches, chaque neurone prend en entrée tous les neurones de la couche inférieure. On définit alors une "couche d'entrée", une "couche de sortie", et n "couches cachées". Les inputs forment la première couche, ils passent aux travers des couches cachées (sur la Figure (3.1), une seule couche cachée) avant de produire la sortie. Ce type de réseau est très répandu du fait de son apprentissage aisé. Il ne possède pas de connexions intra-classes ni de cycles de retour sur les couches, une couche m ne peut utiliser que les sorties de la couche cachée $m-1$ (feed-forward). Enfin, ils sont totalement connectés, chaque neurone de la couche est relié à tous les neurones de la couche précédente.

Définition 3.2.1 (neurone)**.** *Un neurone est une fonction définie de* $\mathbb{R}^d$ *dans*

$\mathbb{R}$ *par,*

$$g(x) = \phi\left(\sum_{j=1}^{d} w_j x_j + w_0\right), \tag{3.2.1}$$

où w_j, $j = 1, 2, \cdots, d$ *est le vecteur des poids et* ϕ *est la fonction d'activation. On a donc un opérateur de sommation qui est activé s'il dépasse un certain seuil dépendant de la fonction d'activation. Les fonctions sigmoïdes sont très utilisées pour cette problématique.*

Définition 3.2.2 (sigmoïde)**.** *Une sigmoïde est une fonction croissante vérifiant,*

$$\begin{cases} \phi(x) & \underset{x\to-\infty}{\longrightarrow} & 0 \\ \phi(x) & \underset{x\to+\infty}{\longrightarrow} & 1 \end{cases}, \tag{3.2.2}$$

par exemple,

$$\phi(x) = \frac{1}{1+e^{-\lambda x}}, \quad \lambda \in \mathbb{R}. \tag{3.2.3}$$

D'autres fonctions d'activation sont possibles,

$$\phi(x) = \mathbb{1}_{x\geq 0}, \qquad \phi(x) = \tanh x. \tag{3.2.4}$$

Définition 3.2.3 (perceptron multicouche feed-forward)**.** *Un réseau de neurones à une couche cachée et* $c < \infty$ *neurones sur cette couche est défini par,*

$$\tilde{y}(t) := f(\mathbf{w}, X(t)) = \sum_{j=1}^{c} w_j^{(2)} \phi\left(\sum_{j=1}^{d} w_j^{(1)} x_j(t) + w_0^{(1)}\right) + w_0^{(2)}, \qquad t = 0, 1, \cdots. \tag{3.2.5}$$

$w^{(1)}$ *et* $w^{(2)}$ *sont respectivement les poids des liaisons entre les inputs et la couche cachée et, de la couche cachée vers l'output.* $w_0^{(i)}$, $i = 1, 2$ *sont les biais rajoutés au réseau, ils jouent le même rôle que l'*intercept *dans une régression linéaire.* $\tilde{y}_t$ *est la sortie du réseau.*

Pour le cas d'un problème de régression, on peut rajouter une liaison directe entre les inputs et la sortie,

$$\tilde{y}(t) := f(\mathbf{w}, X(t)) = \sum_{i=1}^{d} w_i^{(0)} x_i(t) + \sum_{j=1}^{c} w_j^{(2)} \phi^{(1)}\left(\sum_{k=1}^{d} w_{jk}^{(1)} x_k(t) + w_{j0}^{(1)}\right) + w_0^{(2)}, \tag{3.2.6}$$

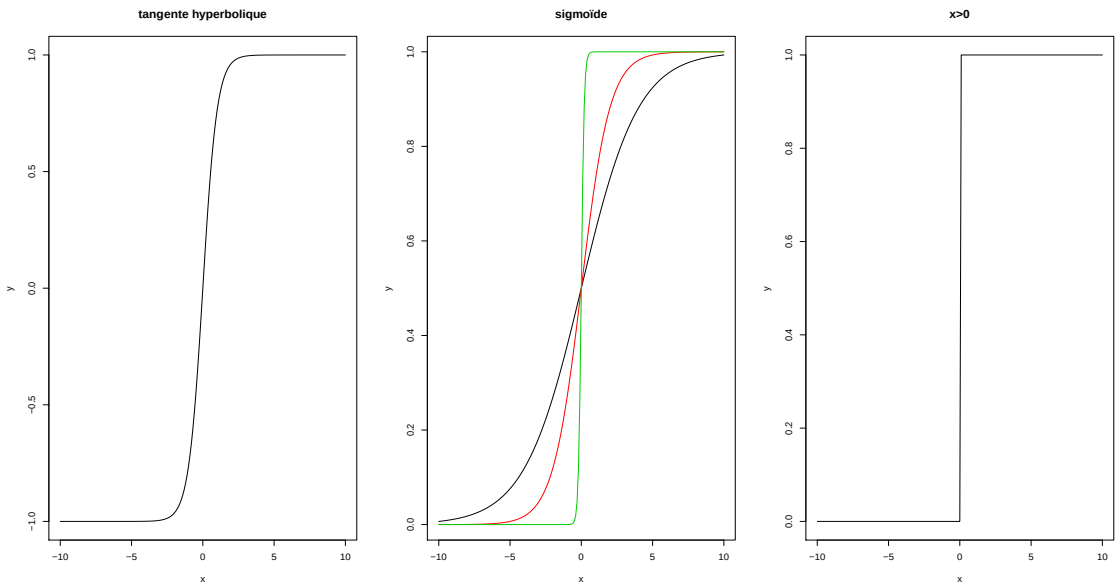


FIGURE 3.2 – De gauche à droite, tangente hyperbolique, sigmoïde avec $\lambda = 1/2$ (noir), $\lambda = 2$ (rouge), $\lambda = 10$, verte et, indicatrice.

$w^{(0)}$ *est le vecteur des poids des inputs vers l'output (liaisons directes non représentées sur la figure (3.1)),*

Le but d'un algorithme d'apprentissage, et donc d'un réseau de neurones est de minimiser le risque de la fonction de prédiction $g : \mathcal{X} \to \mathcal{Y}$, ℓ étant comme dans la section précédente une fonction de perte,

$$\hat{\mathcal{R}}(g, \mathbf{w}) = \mathbb{E}(\ell(Y, g(\mathbf{w}, X))), \tag{3.2.7}$$

approximée par,

$$\mathcal{R}(g, \mathbf{w}) = \frac{1}{t} \sum_{s=1}^{t} \ell(y(s), g(\mathbf{w}, x(s))). \tag{3.2.8}$$

Une fonction éligible devra donc vérifier,

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}(g, \mathbf{w}). \tag{3.2.9}$$

où $\mathcal{G}$ est un sous-ensemble de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, l'ensemble des fonctions de prédiction.

Dans le cadre d'un problème de prévision, en notant $Y = (y(1), y(2), \cdots, y(t))$ le cours d'un actif et $\tilde{Y} = (\tilde{y}(1), \tilde{y}(2), \cdots, \tilde{y}(t))$ le résultat de la modélisation sur la phase d'apprentissage, on peut chercher à minimiser l'erreur quadratique, $\ell(u, v) = (u - v)^2$,

$$\begin{aligned} \mathcal{R}(f, \mathbf{w}) &= \sum_{s=1}^{t} (y(s) - f(\mathbf{w}, x(s)))^2 \\ &= \sum_{s=1}^{t} (y(s) - \tilde{y}(s))^2. \end{aligned} \tag{3.2.10}$$

Pour un problème de classification, on peut choisir de calculer la cross-entropy,

$$\mathcal{R}(f, \mathbf{w}) = -\sum_{s=1}^{t} y(s) \log f(\mathbf{w}, x(s)). \tag{3.2.11}$$

Bien sûr, pour pouvoir produire la sortie $\tilde{Y}$ il faut pouvoir déterminer les poids $\mathbf{w}$. La minimisation de l'erreur n'est pas un problème d'optimisation convexe à cause de la forme particulière de la fonction d'activation (cf (3.3)), une manière commune est d'appliquer une descente de gradient (voir appendix pour plus de détails).

Pour notre problématique on parlera de rétro-propagation , c'est-à-dire que l'on calcule d'abord les erreurs pour tous les échantillons sans mettre à jour les poids (on additionne les erreurs) et lorsque l'ensemble des données est passé une fois dans le réseau, on rétro-propage l'erreur totale. Cette façon de faire est préférée pour des raisons de rapidité et de convergence. Si nous reprenons l'algorithme de descente de gradient,

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \rho \Delta \mathbf{w}, \tag{3.2.12}$$

avec $\Delta \mathbf{w} = \nabla_k \mathcal{R}_{\mathbf{w}}$, ρ est le paramètre d'apprentissage.

Les équations, dans le cadre où $u = 1$ s'écrivent,

$$\begin{aligned} \frac{\partial \mathcal{R}(s)}{\partial w_i^{(0)}} &= -2(y(s) - f(\mathbf{w}, x(s))) \frac{\partial f(\mathbf{w}, x(s))}{\partial w_i^{(0)}} \\ \frac{\partial \mathcal{R}(s)}{\partial w_{jk}^{(1)}} &= -2(y(s) - f(\mathbf{w}, x(s))) \frac{\partial f(\mathbf{w}, x(s))}{\partial w_{jk}^{(1)}} \\ \frac{\partial \mathcal{R}(s)}{\partial w_j^{(2)}} &= -2(y(s) - f(\mathbf{w}, x(s))) \frac{\partial f(\mathbf{w}, x(s))}{\partial w_j^{(2)}}. \end{aligned} \tag{3.2.13}$$

La mise à jour des poids par la descente de gradient s'écrit alors,

$$w_{k+1}^{(i)} = w_k^{(i)} - \rho \sum_{s=1}^{t} \frac{\partial \mathcal{R}(s)}{\partial w_k^{(i)}}, \quad i = 0, 1, 2. \tag{3.2.14}$$

Définition 3.2.4 (sur-apprentissage)**.** *Un problème très fréquent et ennuyeux est le phénomène de sur-apprentissage, i.e. quand le risque sur la partie empirique est bien inférieure au risque réel. Il existe presque une infinité de fonctions g pouvant répondre au critère (3.2.9). Sur la figure (3.3), la courbe rouge apprend par coeur les données mais a de fortes variations entre les points, elle sur-apprend $\mathcal{Y}$, la verte quant à elle semble mieux coller aux données.*

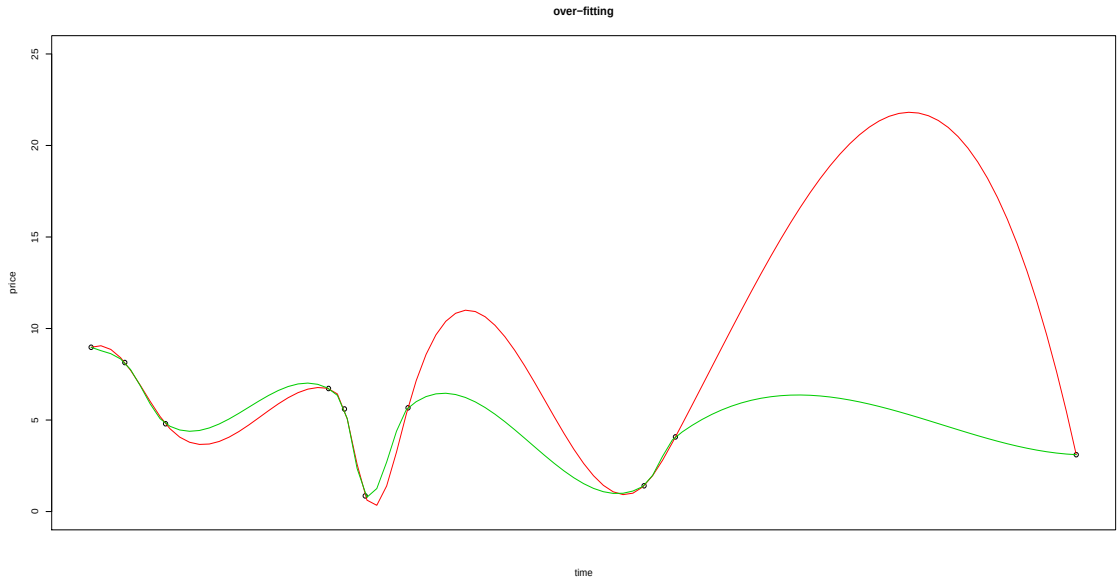


FIGURE 3.3 – Deux fonctions minimisant le risque empirique. Dans les deux cas $\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n}(y(i) - \tilde{y}(i))^2 = 0$

Des poids $\mathbf{w}$ avec des valeurs trop importantes ont tendance à over-fitter le modèle, on pense au cas d'une régression ridge (régularisation de Tikhonov) et on rajoute un terme de pénalité à l'erreur pour régulariser la solution afin d'éviter des coefficients instables,

$$\mathcal{R}^{\text{Pen}}(f, \mathbf{w}) = \mathcal{R}(f, \mathbf{w}) + \lambda ||\mathbf{w}||_2^2, \qquad \lambda \in \mathbb{R}, \tag{3.2.15}$$

λ est le paramètre d'ajustement. Une autre manière de voir (3.2.15) est que l'on ne prend jamais $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$, on se restreint à une classe de fonction ayant des paramètres pas trop élevés, on retrécie l'ensemble, on *shrink*. Un moyen complémentaire et nécessaire pour éviter le sur - apprentissage est d'effectuer une cross-validation, à utiliser avec parcimonie car cette méthode est très coûteuse en temps de calculs.

Pour pouvoir appliquer l'algorithme du gradient on a besoin d'initialiser les poids. S'ils sont trop près de zéro, la fonction d'activation ϕ disparaît et on se rapproche d'une régression linéaire. Néanmoins, on commence avec des poids aléatoires proches de zéro, et donc dans un cadre linéaire, la mise à jour des poids permet de passer progressivement dans un cadre non linéaire. Des poids initiaux strictement égaux à zéro ne permettent pas à l'algorithme de s'ajuster, on reste toujours avec des poids à zéro, à contrario, commencer avec des poids trop élevés donnent des résultats faibles.

Un point (un seul ?) n'a toujours pas été étudié, celui du nombre de couche cachée et du nombre de neurones sur chacune d'entre elles.

Théorème 3.2.1 (Approximation universelle)**.** *Soit g une fonction $\mathcal{C}^\infty$ bornée sur un compact de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Alors pour tout $\varepsilon > 0$, il existe un MLP $f(\mathbf{w}, \mathbf{x})$ à une couche cachée et $c < \infty$ neuronnes sur cette couche tel que,*

$$\int_{\mathbb{R}^d} ||g(\mathbf{x}) - f(\mathbf{w}, \mathbf{x})|| dx < \varepsilon. \tag{3.2.16}$$

Ce théorème explique le succès des perceptrons pour l'approximation de fonction. D'autant plus que le nombre de paramètres à optimiser est bien inférieur à celui d'un polynôme.

Il n'existe malheureusement pas de théorème donnant clairement ce montant. Certainement la meilleure manière pour déterminer ce nombre est d'effectuer une cross-validation, donner trop de choix à l'algorithme augmenterait le coup de calcul. Trois sont retenus, prendre 75%, la racine carrée, ou le nombre égal d'input.

Comme il s'agit d'un apprentissage supervisé, on a besoin de déterminer les positions $\tilde{\pi}(T-N), \cdots, \tilde{\pi}(T-1)$ pour construire le réseau. La manière la plus simple est de prendre $\tilde{\pi}(s) = \text{sign}(\delta_\tau X(s))$, $t - N \leq s < t$, néanmoins, il est possible que le réseau ait du mal à apprendre parfaitement chacune des positions, on risque surtout de sur-apprendre. Une idée simple est de décomposer le réseau en deux "sous-réseaux", l'un déterminant uniquement les positions longues et flat, l'autre short et flat.

$$\hat{\pi}(t) = \begin{cases} f^-(\mathbf{w}, X) \in \{-1, 0\} \\ f^+(\mathbf{w}, X) \in \{0, 1\} \end{cases}. \tag{3.2.17}$$

Il faut bien faire la distinction entre le pourcentage de "bonne position" et la moyenne des rendements. On ne cherche pas forcément à avoir un pourcentage le plus élevé possible, mais surtout à avoir un rendement moyen le plus élevé possible. On peut alors réduire la problématique,

$$\tilde{\pi}_s = \begin{cases} -1 \text{ si } \delta_\tau X(s) < q \\ 0 \text{ si } q \leq \delta_\tau X(s) \leq q' \\ 1 \text{ si } \delta_\tau X(s) > q', \end{cases} \tag{3.2.18}$$

q et q' sont les seuils à partir desquels on estime que le rendement de l'actif est suffisant pour le prendre en considération, on peut par exemple les déterminer en prenant les quantiles historiques à 20% et 80%.

Pour conclure cette section, nous présentons un autre moyen de déterminer les poids d'un réseau de neurones, cette fois nous n'avons plus besoin de déterminer la dérivée (ou une approximation) de la fonction à maximiser (ou minimiser).

Définition 3.2.5 (algorithme génétique)**.** *Un algorithme génétique est une fonction heuristique d'optimisation pour les cas extrêmes où le maximum (minimum) ne peut pas être déterminé analytiquement. L'idée est de reprendre la théorie de l'évolution, sélection naturelle, croisement et mutation. Il existe beaucoup de versions différentes des algorithmes génétiques (il s'agit en fait d'une classe d'algorithme), on présente ici un algorithme appartenant à cette classe pour optimiser les poids d'un réseau de neurones.*

3.3 Arbre de Décision

Dans la section précédente on a présenté une manière de construire son automate de trading en utilisant un réseau de neurones. Cette fois on va uniquement présenter comment constuire un arbre de décision, la mise en place d'un automate de trading avec cet algorithme est sensiblement la même.

Algorithme Génétique

Paramètres : q, p, p'.
Initialisation : On génère une population de possibilités $\mathbf{w}^i = (w_1^i, \cdots, w_m^i)$, $i = 1, \cdots, P$ avec m le nombre de paramètres à optimiser et P la taille de la population. Chacun des $\mathbf{w}^i$ est appelé un chromosome.

(1) Opérateur de sélection. On sélectionne les $[P/2]$ chromosomes donnant les meilleurs résultats par rapport à la fonction d'évaluation, pour nous, la fonction d'utilité. Les autres chromosomes sont enlevés de la population.

(2) Opérateur de croisement. Soit $\mathbf{w}^1 = (w_1^1, \cdots, w_m^1)$ et $\mathbf{w}^2 = (w_1^2, \cdots, w_m^2)$ deux chromosomes (en tout $q\%$ de chromosomes sont croisés), on choisit un entier aléatoirement tel que $0 \leq r \leq m$ et,

$$\mathbf{w}^{1'} = \{w_j | \text{ si } j \leq r, \text{ alors } w_j \in \mathbf{w}^1, \text{ sinon } w_j \in \mathbf{w}^2\}$$
$$\mathbf{w}^{2'} = \{w_j | \text{ si } j \leq r, \text{ alors } w_j \in \mathbf{w}^2, \text{ sinon } w_j \in \mathbf{w}^1\}.$$

(3) Opérateur de mutation. Soit $\mathbf{w}^1 = (w_1^1, \cdots, w_m^1)$ un chromosome (en tout $p\%$ de chromosomes sont mutés), on choisit de muter p' % du chromosome, on tire alors $mp'/100 = r$ entiers aléatoirement compris entre 0 et N et,

$$\mathbf{w}^{i''} = \{w_j | \text{ si } j \neq r, \text{ alors } w_j = w_j^i, \text{ sinon } w_j = \text{random}\}.$$

(4) on régénère $[P/2]$ nouveaux chromosomes et on répète les étapes 2 à 3 jusqu'à convergence.

3.3.1 CART

Il existe plusieurs types d'arbre, celui présenté ici est l'agorithme de Breiman, CART, *Classification And Regression Tree*. L'idée est très simple, l'espace d'arrivée est $\mathcal{Y} = \{-1, 0, 1\}$, on cherche en fonction des inputs $\mathcal{X} \in \mathbb{R}^{t \times d}$ à quelle classe on appartient en posant uniquement des questions binaires sur chacun des $X_i = (x_i(1), x_i(2), \cdots, x_i(t))$, $i = 1, \cdots, d$.

On prend chacun des X_i et pour toutes les valeurs du vecteur, on crée deux feuilles avec la question, $x_i < x_i(s)$ et $x_i \geq x_i(s)$. Pour déterminer l'impureté des

deux feuilles produites on calcule l'indice de Gini,

$$G(m) = \sum_{k \neq k'} p_{mk} p_{mk'}, \tag{3.3.1}$$

avec,

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathbb{1}_{y_i = k}, \tag{3.3.2}$$

où k est la classe considérée, m est le noeud, R_m est la région correspondante et N_m est le nombre total d'éléments dans la région.

La valeur de x qui sera choisie pour créer le premier noeud sera celle qui minimisera l'indice de Gini sur les deux régions,

$$x^* = \arg\min_x \left\{ G(m_d) + G(m_g) \right\}, \tag{3.3.3}$$

où m_d et m_g sont naturellement les feuilles droite et gauche issues de la région m. Le noeud créé est donc celui qui donne les feuilles les plus homogènes possibles. On continue à poser les mêmes questions pour déployer l'arbre jusqu'à ce qu'il ne soit plus possible d'homogénéiser les noeuds.

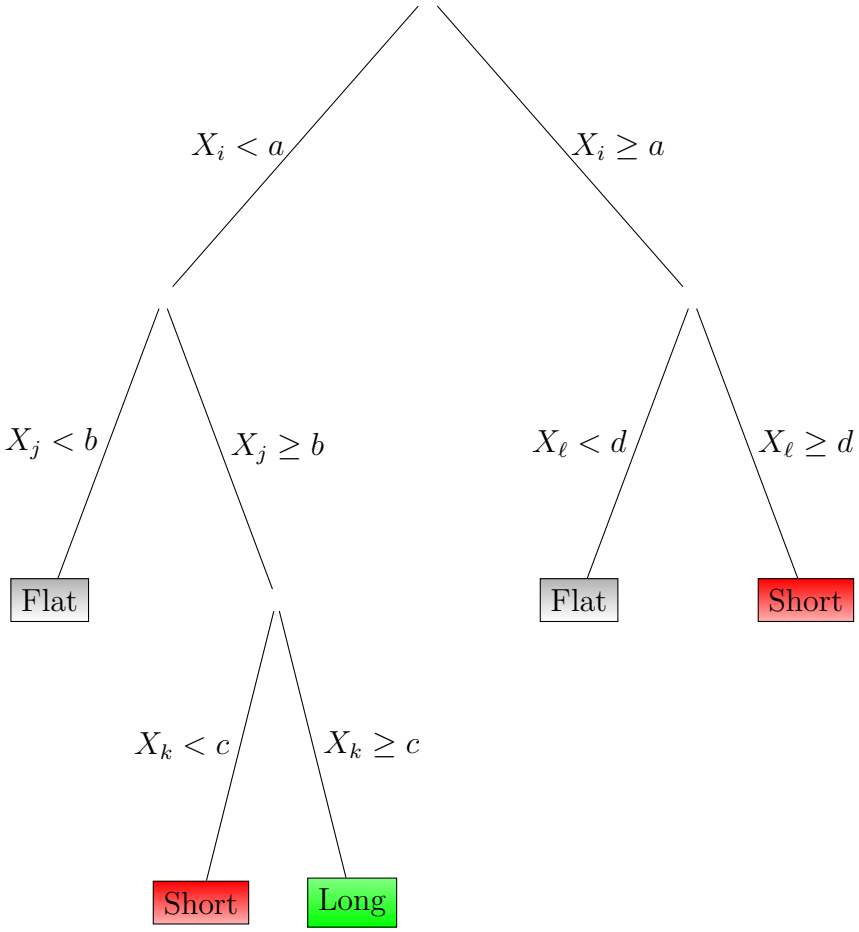


FIGURE 3.4 – Arbre de décision. $X \in \mathcal{X}$, $i, j, k, \ell \in 1, 2, \cdots, d$, et $a, b, c, d \in \mathbb{R}$ sont des valeurs appartenant à $\mathcal{X}$.

Sur la figure (3.4) on a construit un arbre avec l'algorithme présenté précédemment. On notera cet arbre $\mathcal{A}$, en fait il s'agit de $\mathcal{A}_{\max}$ puisque que toutes les feuilles possibles ont été déployées, il a la ramification maximale. Le premier noeud, la racine de l'arbre, est représenté par la question "est ce que $x_i(t)$ est supérieur à a ou inférieur ?", auquel cas on descend un niveau au dessous. La régle de décision est,

$$\hat{\theta}_t = \begin{cases} -1 \text{ si} & \Big\{\{x_i(t) < a\} \cap \{x_j(t) \geq b\} \cap \{x_k(t) < c\}\Big\} \cup \Big\{\{x_i(t) \geq a\} \cap \{x_\ell(t) \geq d\}\Big\} \\ 0 \text{ si} & \Big\{\{x_i(t) < a\} \cap \{x_j(t)\}\Big\} \cup \Big\{\{x_i(t) \geq a\} \cap \{x_\ell(t) < d\}\Big\} \\ 1 \text{ si} & \{x_i(t) < a\} \cap \{x_j(t) \geq b\} \cap \{x_k(t) \geq c\}. \end{cases} \tag{3.3.4}$$

Prendre $\mathcal{A}_{\max}$ pour notre automate n'est pas une bonne idée, il sur-apprend énormément, les relations trouvées sont beaucoup trop précises et ne se reproduiront probablement plus, de plus les feuilles n'ont pas assez d'éléments pour être significatives. On doit élaguer l'arbre.

On note $\mathcal{R}(\mathcal{A})$ l'erreur de l'arbre $\mathcal{A}$, on peut naturellement l'écrire comme la somme des erreurs de chacune des feuilles $\mathcal{R}(\mathcal{A}) = \sum_{m_i \in \mathcal{A}} \mathcal{R}(m_i)$, $i = d, g$. Un premier moyen est de minimiser le nombre de feuilles "à la main" en considérant l'erreur,

$$\mathcal{R}_\lambda(\mathcal{A}) = \sum_{m_i \in \tilde{\mathcal{A}}} \mathcal{R}(m_i) + \lambda|\mathcal{A}|, \qquad \lambda \geq 0. \tag{3.3.5}$$

avec $|\mathcal{A}|$ le nombre de feuilles sur l'arbre. On a bien évidement $\mathcal{R}(\mathcal{A}_{\max}) = 0$, mais $\mathcal{R}_\lambda(\mathcal{A}_{\max}) \neq 0$. Selon la valeur de λ choisie on a donc un sous arbre, Fig (3.3.1).

Une bonne manière de déterminer la valeur de λ optimale est de faire une cross-validation par rapport aux déviances.

3.3.2 Boosting

L'idée de déployer un arbre de décision pour avoir la position à prendre à l'instant t en fonction d'un ensemble d'inputs est assez séduisante. Néanmoins, en tant que tel, CART ne donnera pas forcément de bonnes prévisions et peut être apparenté à un classifieur *faible*. L'algorithme AdaBoost.M1 propose d'optimiser la performance de l'algorithme en créant un ensemble d'arbre, $\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_M$, chacun ayant des poids sur ses inputs issus de l'erreur de classification de l'arbre précédent. En faisant cela, on contraint le classifieur à donner plus d'importance aux données les plus difficiles à ajuster. L'importance d'une observation restera donc inchangée si elle est bien classée, sinon, elle croît proportionnellement à l'erreur du modèle. Au final, la décision sera,

$$\hat{\theta}_t = sign\left(\sum_{i=1}^{M} \alpha_m \mathcal{A}_m\right), \tag{3.3.6}$$

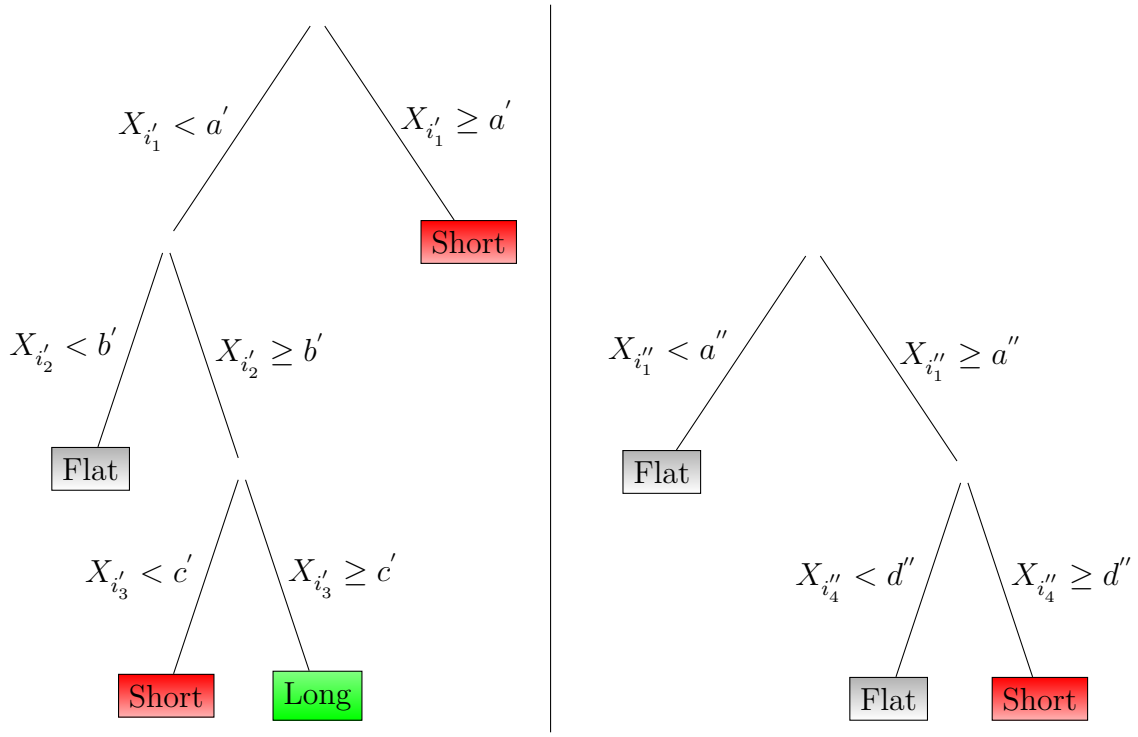


TABLE 3.1 – Exemple de deux sous arbres extraits de (3.4) pour des valeurs de λ différentes.

où α_m est déterminé par l'algorithme de boosting. L'agrégation nous permet d'éviter le sur-apprentissage. On présente dans l'encadré suivant la construction des classifieurs et des poids associés. Par abus de notation on notera $\mathcal{A}_m(X(s))$ la position donnée par le classifieur m à l'instant $0 < s \leq T$.

Remarquons que pour les inputs mal classés qui doivent être boostés, on doit avoir une erreur du classifieur inférieur à $1/2$, sinon α_m est négatif et les poids de l'algorithme sont mis à jour dans la mauvaise direction. Pour un problème à 2 classes, cette condition est similaire à prendre un aléa et est facilement atteignable. Pour un problème à K classe, cette précision peut être plus compliquée à atteindre. L'algorithme AdaBoost.M1 ne pourra donc pas répondre à notre problèmatique puisque $\mathcal{Y} \in \{-1, 0, 1\}$, $K = 3$.

Une solution simple pour passer dans un problème multi-classe est de décomposer le problème à K classes en sous problème binaire. On pourra donc séparer l'espace de décision en $\mathcal{Y}_1 = \{-1, 0\}$ et $\mathcal{Y}_2 = \{0, 1\}$ et appliquer l'algorithme Ada-

AdaBoost.M1

Initialisation : $w(s) = 1/t$, $s = 1, \cdots, t$

pour $m = 1, 2$,
(1) On construit le classifieur $\mathcal{A}_m(X)$ en utilisant les poids w_i.

(2) On calcule l'erreur,

$$\text{err}_m = \frac{\sum_{s=1}^{t} w(s)\mathbb{1}_{y(s)\neq\mathcal{A}_m(X(s))}}{\sum_{s=1}^{t} w(s)}.$$

(3) On calcule $\alpha_m = \log\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$.

(4) On réinitialise les poids,

$$w(s) = w(s)\exp\left(\alpha_m \mathbb{1}_{y(s)\neq\mathcal{A}_m(X(s))}\right), \qquad s = 1, \cdots, t.$$

Output :

$$\mathcal{A}(\mathbf{X}) = \text{sign}\sum_{m=1}^{M}\alpha_m\mathcal{A}(\mathbf{X}).$$

Boost.M1.

Une modification de l'algorithme AdaBoost.M1 pour avoir des résultats corrects dans le cas où $K > 2$ sans créer de sous problèmes $\mathcal{Y}_i$ est l'algorithme *Stagewise Additive Modeling using a Multi-class Exponential loss* (SAMME), Zhu et al., 06. En reprenant exactement le même algorithme que précédemment, la seule modification est,

$$\alpha_m = \log\left(\frac{1-\text{err}_m}{\text{err}_m}\right) + \log(K-1). \tag{3.3.7}$$

3.3.3 Bootstrap

Avant d'introduire les forêts aléatoires, nous avons besoin d'expliquer le principe de *bagging*. Il s'agit simplement de faire une estimation *bootstrap* pour réduire la variance du modèle. A la différence du boosting où chaque nouvel arbre construit dépendait de l'erreur du précédent, cette fois, chaque arbre est construit à partir d'un échantillon bootstrap de l'ensemble d'apprentissage $\mathbf{Z} = \{(\mathbf{X}(1), y(1)), \cdots,$

$(\mathbf{X}(t), y(t))\}$. On effectue un tirage avec remise de $\mathbf{Z}$ que l'on note $\mathbf{Z}_b$. La règle de décision est alors, après B bootstrap,

$$\hat{\theta}_t = sign\left(\sum_{b=1}^{B} \mathcal{A}(\mathbf{Z}_b)\right). \tag{3.3.8}$$

3.3.4 Forêt Aléatoire

Une forêt aléatoire prolonge cette idée en apportant une nouvelle partie randomizée à chaque arbre. Pour chacun des échantillons bootstrap on tire de manière aléatoire q inputs dans $\mathcal{X}$. L'intérêt est une nouvelle fois de diminuer la variance du classifieur ainsi que son biais. A la différence du bagging, les arbres construits vont pouvoir prendre en compte tous les inputs. Un modèle CART construit classiquement ne prendra pas en compte tous les inputs mais uniquement les plus importants, les moins 'dépendants' ne peuvent pas donner leurs avis. En randomizant $\mathcal{X}$, on laisse l'opportunité à tous les inputs d'être pris en compte. En fixant un nombre q d'input relativement bas, on évite l'arbre trop profond et, le surapprentissage. Néanmoins, il ne faut pas alimenter la forêt par des inputs n'ayant aucun rapport au risque du *data-snooping* (vrai quelque soit la méthodologie utilisée).

Chapitre 4

Pair Trading

Le *pair trading*, ancêtre de l'*arbitrage statistique*, est une stratégie permettant de détecter une opportunité d'arbitrage entre deux actifs financiers. En étudiant le *spread*, i.e. la différence entre deux actifs, on peut voir si l'un est sur-évalué par rapport à l'autre et inversement, si l'un est sous-évalué par rapport à l'autre. L'idée du pair trading est donc d'arbitrer certaines des fortes variations du spread dues à un shock temporaire en supposant qu'il revienne dans son intervalle historique bien défini. Cette stratégie est intuitivement assez simple mais néanmoins très risquée (cf. LTCM). En dehors de la stratégie en elle-même, i.e. savoir quand acheter ou shorter tel ou tel actif et à quel montant, il faut trouver deux instruments éligibles au pair trading. Le spread des deux instruments doit être *mean reverting*, c'est vraiment ce point qui mérite le plus d'attention quand on s'essaye au pair trading. Si cette dernière propriété n'est pas respectée, on risque de se retrouver avec en portefeuille deux actifs et attendre indéfiniment d'avoir un ordre de dénouement ou d'inversement de la pose si le spread ne revient pas à sa moyenne. En dehors de cette caractéristique qui doit être respectée, il faut intégrer un stop-loss à la stratégie. Enfin, une fois qu'une paire est sélectionnée, donc que le spread est mean-reverting, on a naturellement envie de le modéliser par un processus du type Ornstein Uhlenbeck. Pour commencer nous donnons une définition stricte de la stationnarité et du retour à la moyenne pour effectuer la sélection.

4.1 Cointégration

Définition 4.1.1 (stationnarité stricte)**.** *Un processus aléatoire* $Y = (Y_t, t \leq T)$ *est strictement stationnaire si* $\forall n \in \mathbb{N}^*$, $\forall (t_1, t_2, \cdots, t_n) \in [0,T]^n$, $\forall k \in [0,T]$,

$$\mathcal{L}(Y_{t_1}, Y_{t_2}, \cdots, Y_{t_n}) = \mathcal{L}(Y_{t_1+k}, Y_{t_2+k}, \cdots, Y_{t_n+k}). \tag{4.1.1}$$

En d'autre termes, la translation $\mathcal{T}(k) : t \to t+k$ *laisse la loi de* Y *invariante.*

Définition 4.1.2 (Stationnarité faible)**.** *Un processus aléatoire* $Y = (Y_t, t \in T)$ *est faiblement stationnaire (ou stationnaire au second ordre) si ses premiers et seconds moments sont indépendants du temps,*

- $\mathbb{E}Y_t^2 < \infty$, $\forall t \in T$.
- $\mathbb{E}Y_t = m$, $\forall t \in T$.
- $cov(Y_t, Y_{t+k})$ *ne dépend pas de* t *mais uniquement de* k.

Proposition 4.1.1. *Un processus strictement stationnaire et du second ordre est faiblement stationnaire, l'inverse n'est pas forcément vrai, excepté pour les processus gaussiens.*

Définition 4.1.3 (Intégration)**.** *Soit* (Y_t) *une série temporelle, si après* d *différenciations la série est stationnaire, on dira que la série est intégrée d'ordre* d*, noté* $X_t \sim I(d)$,

$$\begin{aligned} \Delta^{d-n} Y_t \text{ n'est pas stationnaire } \forall 1 \leq n < d \\ \Delta^d Y_t \text{ est stationnaire .} \end{aligned} \tag{4.1.2}$$

Définition 4.1.4 (cointégration)**.** *Soit* (X_t) *et* (Y_t) *deux séries temporelles respectivement intégrées de même ordre* d*. Alors s'il existe* $\alpha, \beta \in \mathbb{R}$ *tel que* $\alpha X_t + \beta Y_t \sim I(d-b)$ *pour* $0 < b \leq d$*, les séries* (X_t) *et* (Y_t) *sont cointégrées d'ordre* $d-b < d$*. Le vecteur* (α, β) *est appelé vecteur de cointégration. En d'autres termes, les séries temporelles sont cointégrées si leur spread, formé avec les coefficients* α *et* β *est stationnaire. Au lieu de parler de vecteur de cointégration, on peut également construire le spread avec juste un coefficient,* c*, le coefficient de cointégration, dans ce cas, le spread s'écrit* $Y_t + cX_t$*, ce qui est complètement équivalent à la forme précédente.*

La notion de cointégration a été introduite par deux prix Nobel, R. Engle et C. Granger, [EnGr87].

Une paire d'actif éligible au pair trading devra être cointégrée (attention, cela n'a rien à voir avec la corrélation !). Une manière de construire notre algorithme de sélection serait de déterminer l'ordre d'intégration de chacune des séries (s'il existe bien sûr), de les classer par rapport à ce dernier, puis de tester la cointégration avec un test de Dickey-Fuller. Cette manière de faire évite de tester des paires avec des ordres d'intégration différents et qui ne pourraient pas être cointégrées.

Proposition 4.1.2 (Test de Dickey-Fuller)**.** *Soit $(y_1, y_2, \cdots, y_t)$ une série temporelle. Le test de DF propose de tester l'hypothèse nulle, l'hypothèse de non stationnarité, $H_0 : \rho = 1$, où ρ est le paramètre du processus AR(1),*

$$y_t = \mu + \rho y_{t-1} + \epsilon_t, \ \epsilon \sim BB, \tag{4.1.3}$$

si $\rho = 1$, il s'agit alors d'une pure marche aléatoire. Une forme équivalente faisant apparaître une force de rappel est,

$$y_t - y_{t-1} = \Delta y_t = \mu + \gamma y_{t-1} + \epsilon_t, \ \epsilon \sim BB, \ \gamma = (\rho - 1), \ , H_0 : \gamma = 0. \tag{4.1.4}$$

De manière plus générale, on peut rajouter un terme de dérive au modèle,

$$\Delta y_t = \mu + \beta t + \gamma y_{t-1} + \epsilon_t, \ \epsilon \sim BB, \ \gamma = (\rho - 1), \ , H_0 : \gamma = 0. \tag{4.1.5}$$

Il est en fait recommandé de tester en premier lieu cette dernière forme en appliquant un test de Student sur chacun des coefficients, quitte à revenir à la forme plus restreinte d'un AR(1).

Soit S_1 et S_2 deux actifs financiers et $y_t = \log S_1(t) - c \log S_2(t)$ leur spread à l'instant t, $c > 0$ est le coefficient d'intégration. Pour pouvoir faire du pair trading entre deux actifs il faut que leur spread présente la propriété de retour à la moyenne pour pouvoir jouer sur les écarts de variations en étant en droit de supposer qu'ils reviendront dans un intervalle bien défini. En reprenant les définitions précédentes, une paire d'actif éligible devra vérifier l'hypothèse de stationnarité, par suite de retour à la moyenne, dans le cas du test de DF,

$$\Delta y_t = \mu + \beta t + \gamma y_{t-1} + \epsilon_t, \ \epsilon \sim BB, \quad H_0 : \gamma = 0. \tag{4.1.6}$$

ici, on a testé que y_t était directement stationnaire, on pourrait tester le cas $I(d-b)$.

Une fois qu'une paire éligible a été détectée, une stratégie simple est de se mettre long de c sur S_2 et short sur S_1 (cela permet d'être *market neutral* pour

ne pas supporter les risques du marché mais uniquement le risque de retour 'trop long' à la moyenne) quand le spread s'écarte de la moyenne de 2 fois son écart type et annuler sa position au retour à la moyenne.

FIGURE 4.1 – Exemple d'ETF susceptible d'être "pair-traidé".

Le cas présenté ci-dessus est un modèle discret, plutôt adapté pour du trading basse fréquence, pour un trading haute fréquence, en remarquant que pour $\Delta t = t + 1 - t$ et $\Delta y_t = y_t - y_{t-1}$,

$$\Delta y_t = \kappa(\theta - y_t)\Delta t + \sigma \Delta z_t, \tag{4.1.7}$$

est bien un AR(1) avec $\mu = \kappa\theta$, $\gamma = 1 - \kappa$ et z_t suit une loi normale centrée réduite, on peut alors étendre cette méthodologie au cas continu en modélisant Y par un processus de Vasicek,

$$dY(t) = \kappa(\theta - Y(t))dt + \sigma dW(t), \qquad Y_0 = x, \tag{4.1.8}$$

où W est un processus de Wiener.

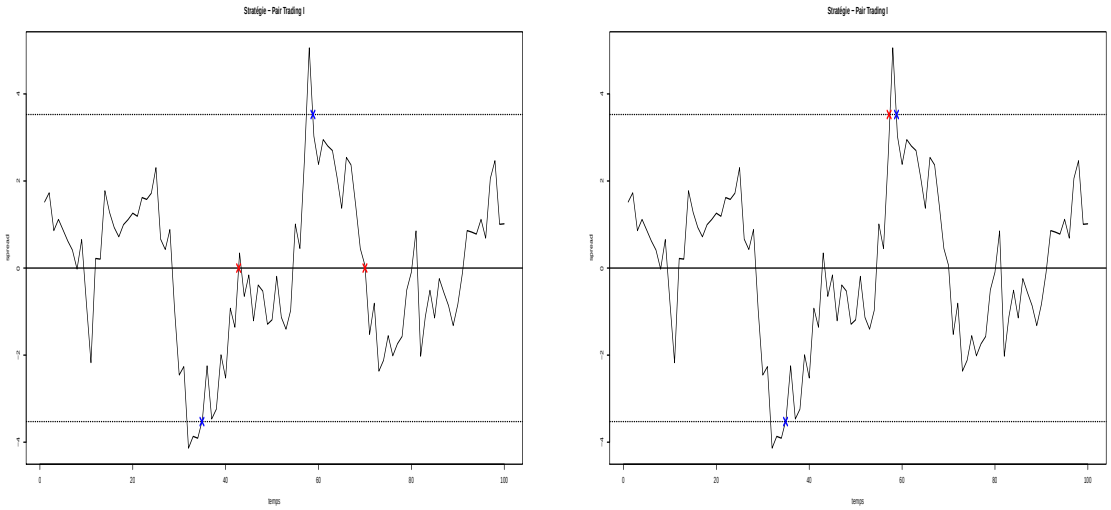


FIGURE 4.2 – Deux exemples de stratégies possibles de pair trading. Dans le cas I on dénoue sa position dès le retour à la moyenne, dans le cas II on attend de sortir de l'autre borne du spread pour se dénouer et reprendre une position dans la descente. Le cas I est évidemment moins risqué.

4.2 Contrôle Optimal

4.2.1 Fonction d'Utilité

Avant de continuer sur le pair trading, faisons quelques rappels sur la notion de fonction d'utilité. En notant V la richesse et u la fonction d'utilité, nous pouvons écrire la problématique par,

$$\max \mathbb{E}[u(V_{t+1})|\mathcal{F}_t], \tag{4.2.1}$$

où V_t est la richesse de l'agent à l'instant t. Pour être éligible, la fonction $u : \mathbb{R} \to \mathbb{R}$ doit vérifier quelques propriétés, la première étant qu'elle doit être au moins de classe C^2, ensuite,

$$\frac{\partial u}{\partial V} > 0. \tag{4.2.2}$$

La fonction u doit donc être croissante, ce qui signifie qu'une augmentation de richesse augmentera toujours l'utilité du gestionnaire, même aussi petite soit-elle

(hypothèse de non-saturation). La troisième propriété que doit vérifier la fonction d'utilité est,

$$\frac{\partial^2 u}{\partial V^2} < 0. \tag{4.2.3}$$

La fonction u est donc concave, ce qui signifie que les premiers gains donnent une plus grande satisfaction au gestionnaire que les gains suivants. L'inégalité stricte implique que le gestionnaire ne perdra pas son temps avec un gain qui ne l'intéresse plus, c'est la notion d'*aversion au risque*.

Au sens de Arrow-Pratt, l'aversion absolue au risque (ARA) est,

$$\mathrm{ARA}_u(V) = -\frac{u''(V)}{u'(V)}. \tag{4.2.4}$$

L'aversion relative au risque est donnée par,

$$\mathrm{RRA}_u(V) = -\frac{u''(V)}{u'(V)}V. \tag{4.2.5}$$

Enfin, la tolérance absolue au risque est,

$$\mathrm{ART}_u(V) = -\frac{u'(V)}{u''(V)}. \tag{4.2.6}$$

Notons que pour deux agents A et B ayant respectivement pour fonction d'utilité u et v, alors si A est plus averse au risque que B,

$$\mathrm{ARA}_u(V) > \mathrm{ARA}_v(V), \ \forall V. \tag{4.2.7}$$

Un exemple simple et déjà vu précédemment est de maximiser l'espérance de gain et diminuer la variance,

$$\mathbb{E}[u(V_{t+1})|\mathcal{F}_t] = \mathbb{E}[V_{t+1} - \frac{\gamma}{2}V_{t+1}^2|\mathcal{F}_t]. \tag{4.2.8}$$

La fonction d'utilité exponentielle est donnée par,

$$\mathbb{E}[u(V_{t+1})|\mathcal{F}_t] = \mathbb{E}[1 - e^{-\lambda V_{t+1}}|\mathcal{F}_t]. \tag{4.2.9}$$

Un autre exemple connu est la fonction d'utilité CRRA (constant relative risk aversion),

$$\mathbb{E}[u(V_{t+1})|\mathcal{F}_t] = \begin{cases} \mathbb{E}\left[\dfrac{V_{t+1}^{1-\gamma}}{1-\gamma}|\mathcal{F}_t\right] & \text{si } \gamma > 1 \\ \mathbb{E}[\ln V_{t+1}] & \text{si } \gamma = 1 \end{cases}. \tag{4.2.10}$$

Dans les trois cas, γ est le paramètre d'aversion au risque.

4.2.2 Problématique

Nous n'allons pas rentrer dans les détails et mener à terme les modèles que nous allons exposer mais juste écrire le problème sous forme d'un problème de contrôle stochastique, cela pourra donner lieu à des projets. Supposons déjà que les rendements des deux actifs S_1 et S_2 sont cointégrés et qu'ils peuvent s'écrire sous la forme,

$$\begin{aligned} \mathrm{d}S_1(t) &= \mu_1(S_1, S_2, t)\mathrm{d}t + \sigma_1(S_1, S_2, t)\mathrm{d}W_1(t) \\ \mathrm{d}S_2(t) &= \mu_2(S_1, S_2, t)\mathrm{d}t + \sigma_2(S_1, S_2, t)\mathrm{d}W_2(t), \end{aligned} \tag{4.2.11}$$

où μ_1, μ_2 sont les termes de dérive usuels et σ_1, σ_2 les volatilités respectives. On écrit le spread sous la forme,

$$Y(t) = S_1(t) - cS_2(t), \tag{4.2.12}$$

où c est le coefficient de cointégration. Puisque les deux actifs sont cointégrés le spread devrait pouvoir s'écrire sous la forme d'un processus mean-reverting,

$$\mathrm{d}Y(t) = \kappa(\theta - Y(y))\mathrm{d}t + \sigma_y \mathrm{d}W_y(t), \tag{4.2.13}$$

bien sur il faudrait prendre des formes particulières de μ_1, μ_2, σ_1, σ_2, et W_1, W_2 pour arriver à une équation de ce type et que κ soit positif pour assurer la stationnarité du processus.

Maintenant, notons π_i, $i = 1, 2$ le montant investi dans l'actif i, l'un des intérêts du pair-trading est d'être neutre face au marché, alors,

$$\pi_2 = c\pi_1. \tag{4.2.14}$$

Rajoutons un actif sans risque à notre portefeuille,

$$\mathrm{d}S_0(t) = rS_0(t)\mathrm{d}t, \tag{4.2.15}$$

S_0 est par exemple un bon du trésor américain à long terme, *Treasury Bond*, le 22/08/11 le taux était à 3.40, le 16/08/2012 à 2.941 (vers 19h CET (=GMT+2h)). La dynamique de la richesse du portefeuille auto-financé est ainsi donnée par,

$$\frac{\mathrm{d}V(t)}{V(t)} = \pi_1(t)\frac{\mathrm{d}S_1(t)}{S_1(t)} + \pi_2(t)\frac{\mathrm{d}S_2(t)}{S_2(t)} + \frac{\mathrm{d}S_0(t)}{S_0(t)}. \tag{4.2.16}$$

Maintenant que les équations de dynamique sont écrites, nous allons essayer de maximiser une fonction d'utilité, en espérance, pour déterminer la stratégie optimale (le contrôle optimal) π.

$$\mathcal{P} : \begin{cases} \sup_{\pi(t)} & \mathbb{E}(u(V^\pi(t))|V^\pi(t) = w, Y(t) = y) \\ sc & V^\pi(0) = w, Y(0) = y \\ & \mathrm{d}Y(t) = \kappa(\theta - Y(t))dt + \sigma_y \mathrm{d}W_y(t) \\ & \dfrac{\mathrm{d}V(t)}{V(t)} = c\pi_1(t)\dfrac{\mathrm{d}S_1(t)}{S_1(t)} + \pi_2(t)\dfrac{\mathrm{d}S_2(t)}{S_2(t)} + \dfrac{\mathrm{d}S_0(t)}{S_0(t)} \end{cases} . \tag{4.2.17}$$

notons $G(t, w, y) = \max_\pi \mathbb{E}\{G(t + \mathrm{d}t, W(t + \mathrm{d}t), Y(t + \mathrm{d}t))\}$ la fonction valeur à optimiser que l'on peut réécrire

$$G(t, w, y) = \max_\pi \mathbb{E}\{G(t, W(t), Y(t)) + \mathrm{d}G\}. \tag{4.2.18}$$

En appliquant le lemme d'Itô, la fonction G ne dépendant que de V et Y, on a,

$$\begin{aligned} \mathrm{d}G(t) =& \dot{G} + G_w(\mathrm{d}W(t)) + G_y(\mathrm{d}Y(t)) + \frac{1}{2}G_{ww}(\mathrm{d}(W(t)))^2 \\ &+ \frac{1}{2}G_{yy}(\mathrm{d}X(t))^2 + G_{yw}(\mathrm{d}Y(t)\mathrm{d}W(t)). \end{aligned} \tag{4.2.19}$$

en égalisant (4.2.18) et (4.2.19), avec $\mathrm{d}t \to 0$, le problème revient à résoudre,

$$\max_\pi \mathbb{E}\mathrm{d}G = 0. \tag{4.2.20}$$

Après quelques calculs 'très simples', on trouve des formules explicites pour les montants à investir, il s'agit d'un problème de contrôle stochastique, et les équations à résoudre sont dites équations de Hamilton-Jacobi-Bellman. Pour laisser l'opportunité de faire des projets sur le pair trading et le contrôle stochastique, nous nous arrétons ici sur les équations, pour plus de détail voir par exemple [MuPrWo08], [An11] et les références à l'intérieur.

Chapitre 5

Portefeuille Multi-Actifs

Dans les sections précédentes on a présenté des stratégies pour traiter un seul actif à la fois. On présente maintenant des stratégies multi-actifs. L'objectif est de tirer un gain de la diversification. Dans la première section la problèmatique est de minimiser le *regret*, on reprend les idées développées dans la partie sur l'agrégation des stratégies, *exploration/exploitation*. La construction se focalise sur la maximisation du rendement, la minimisation de la variance est implicite. Dans la seconde partie, on cherche à maximiser les rendements et à minimiser la variance. Il s'agit de construire le portefeuille de Markowitz, point central de la *Théorie Moderne du Portefeuille*.

5.1 Exploration/Exploitation

L'environnement est composé de N actifs, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N)$, $\mathbf{Y} \in \mathbb{R}_+^{T\times N}$. Le comportement du marché à l'instant t, c'est à dire les variations journalières, le close à l'instant t sur l'open à l'instant t est donné par la série $\mathbf{x}(t) = (x_1(t), \cdots, x_N(t)) \in \mathbb{R}_+^N$ et $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$. On note $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_N)$, $\pi_i \geq 0$, $\sum \pi_i = 1$ la proportion investie dans chaque actif. On présente ici une procédure pour allouer séquentiellement un montant π_i sur les N actifs dans le but de faire au moins mieux que la meilleure stratégie *Buy-and-Hold* (B&H) où la stratégie B&H est définie par,

$$\pi_i(u) = \frac{\frac{1}{N}\prod_{s=1}^{u} x_i(s)}{\sum_{k=1}^{N}\frac{1}{N}\prod_{s=1}^{u} x_k(s)}. \tag{5.1.1}$$

Pour tout $t \geq u$ on garde le même portefeuille, la richesse à l'instant t est alors

donnée par,

$$W_T(\boldsymbol{\pi}, \mathbf{X}) = \sum_{k=1}^{N} \pi_k(1) \prod_{s=1}^{T} x_k(s). \tag{5.1.2}$$

Ici, on a choisi d'initialiser sa pose initiale en fonction des rendements passés ,on aurait aussi pu intégrer les notions de volatilité (risque de marché) et de corrélation (risque de diversification). L'idée sous-jacente à cette stratégie est l'hypothèse d'efficience des marchés financiers, i.e. que les cours des actifs sont toujours à leur juste prix et qu'ils suivent donc une marche aléatoire. Il n'est donc pas possible d'espérer battre le marché mais plutôt de détenir un ensemble d'actif pour leur valeur intrinsèque (c'est d'ailleurs pour cela qu'il n'est pas fondamental de prendre en compte les risques du marché dans l'initialisation) et par suite la perception de dividendes.

FIGURE 5.1 – Cours de l'action Citigroup de ∼1987 à ∼2007, le capital initial investi avec la stratégie buy and hold aurait été multiplié par 30 !

Définition 5.1.1. *Soit* $\boldsymbol{\pi}$ *une stratégie, pour tout ce chapitre la richesse et la richesse logarithmique normalisée seront données par,*

FIGURE 5.2 – mais en suivant la stratégie, il aurait fallu avoir besoin de liquidité avant la fin 2007 pour avoir un gain...

$$W_T(\boldsymbol{\pi}, \mathbf{X}) = \prod_{s=1}^{T} \boldsymbol{\pi}(s) \cdot \mathbf{x}(s), \qquad LW_T(\boldsymbol{\pi}, \mathbf{X}) = \frac{1}{T} \sum_{s=1}^{T} \log(\boldsymbol{\pi}(s) \cdot \mathbf{x}(s)). \tag{5.1.3}$$

5.1.1 Portefeuille Universel

La première classe de portefeuille étudiée est le *portefeuille rebalancé constamment* (CRP). Cette stratégie a des poids rebalancés à chaque période de trading, aucun rajout de capital n'est effectué, on investit constamment le même montant,

$$\boldsymbol{\pi}(s) = \boldsymbol{\pi}, \;\; s = 1, \cdots, T. \tag{5.1.4}$$

Prenons l'exemple d'un environnement fictif à deux actifs selon la distribution (1,1/2), (1,2), (1,1/2), (1,2), etc., avec $\boldsymbol{\pi} = (1/2, 1/2)$,

$$\begin{array}{c} 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4} \\ 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2} \\ 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4} \\ \vdots \qquad\qquad \vdots \end{array} \tag{5.1.5}$$

Donc $S_1(\boldsymbol{\pi}) = 3/4$, $S_2(\boldsymbol{\pi}) = 9/8$, $S_3(\boldsymbol{\pi}) = 27/32$, on peut montrer par récurrence qu'après $2n$ période de trading la richesse atteint $(9/8)^n$. Plus généralement, la richesse à l'instant t est donnée par,

$$W_T(\boldsymbol{\pi}, \mathbf{X}) = \prod_{s=1}^{T} \left(\sum_{i=1}^{N} \pi_i x_i(s) \right). \tag{5.1.6}$$

La valeur de π maximisant la richesse pour l'environnement $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ est donnée par,

$$\boldsymbol{\pi}^* = \arg \max_{\boldsymbol{\pi}} W_T(\boldsymbol{\pi}, \mathbf{X}). \tag{5.1.7}$$

on notera cette stratégie $\boldsymbol{\pi}^{bcpr}$ et on appellera ce portefeuille le *meilleur portefeuille rebalancé constamment* (BCRP), [BlKa97], [BeCo88], [BuGoWa06], [Co91], [CoOr96], [He98].

Définition 5.1.2 (Portefeuille universel)**.** *Un portefeuille universel au sens de Cover (91) est un portefeuille qui vérifie,*

$$\lim_{t \to \infty} \max_{\mathbf{x}_t} (LW_T(\boldsymbol{\pi}^{bcpr}, \mathbf{X}) - LW_T(\boldsymbol{\pi}, \mathbf{X})) \leq 0. \tag{5.1.8}$$

Il s'agit donc d'un portefeuille qui a asymptotiquement au moins le même taux de croissance que le BCRP.

La définition suivante nous permet de généraliser ce cadre à un problème de borne min/max.

Définition 5.1.3 (Regret externe)**.** *Soit $\mathcal{Q}$ une classe de stratégie de portefeuille. On définit le regret externe de la stratégie, le worst case logarithmic ratio, P par,*

$$\mathcal{R}_T(\mathbf{v}, \mathcal{Q}) = \sup_{\mathbf{x}_T} \sup_{\mathbf{u} \in \mathcal{Q}} \log \frac{W_T(\mathbf{u}, \mathbf{X})}{W_T(\mathbf{v}, \mathbf{X})}. \tag{5.1.9}$$

Le but est de trouver une stratégie donnant un ratio aussi petit que possible. $\mathcal{R}_t(\mathbf{v}, \mathcal{Q}) = O(t)$ signifie que la stratégie d'investissement a au moins le même taux de croissance que la meilleure des stratégies de la classe $\mathcal{Q}$. En se restreignant au cas où $\mathbf{u} = \boldsymbol{v}^{bpcr}$, si le ratio tend vers 0 en t, on retrouve la définition d'un portefeuille universel au sens de Cover. La difficulté est bien entendu que nous devons investir séquentiellement, tandis que le BCRP ne peut être déterminé que rétrospectivement. Nous avons affaire à un problème d'apprentissage séquentiel.

On se propose de construire une stratégie de portefeuille $\boldsymbol{\pi}_t$, où $\boldsymbol{\pi}_t$ est déterminé au regard du passé $\mathbf{x}_1, \cdots, \mathbf{x}_{t-1}$, qui performe asymptotiquement aussi bien que le BCRP. Vouloir prendre en compte le passé est contradictoire avec l'hypothèse de E. Fama puisque la séquence des prix est arbitraire, le futur n'a aucune relation avec le passé. La stratégie proposée repose sur les performances passées,

$$\boldsymbol{\pi}(t+1) = \frac{\int \boldsymbol{\pi} W_t(\boldsymbol{\pi}, \mathbf{x}(t-1)) \mathrm{d}\mu(\boldsymbol{\pi})}{\int W_t(\boldsymbol{\pi}, \mathbf{x}(t-1)) \mathrm{d}\mu(\boldsymbol{\pi})}, \tag{5.1.10}$$

le vecteur des poids est initialisé uniformément, $\boldsymbol{\pi}_1 = (\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N})$ et $\mu(\boldsymbol{\pi})$ est une fonction de distribution dans le simplexe $\mathcal{D}$, l'ensemble des stratégies admissibles qui sont comprises dans le simplexe défini par,

$$\mathcal{D} = \{\boldsymbol{\pi} \in \mathbb{R}^N : \pi_i \geq 0, \sum_{i=1}^{N} \pi_i = 1\}, \tag{5.1.11}$$

on interdit donc les ventes à découverts. La richesse finale est,

$$W_T(\boldsymbol{\pi}, \mathbf{X}) = \prod_{s=1}^{T} \boldsymbol{\pi}(s) \cdot \mathbf{x}(s). \tag{5.1.12}$$

On peut prouver que dans le cas où μ est une densité uniforme sur le simplexe $\mathcal{D}$ de $\mathbb{R}^N$,

$$\mathcal{R}_t(\boldsymbol{\pi}^{univ}, \mathcal{Q}) \leq (N-1) \ln(1+T). \tag{5.1.13}$$

Si le portefeuille est construit selon la densité de Dirichlet $(1/2,\cdots,1/2)$ on a la borne,

$$\mathcal{R}_t(\boldsymbol{\pi}^{univ}, \mathcal{Q}) \leq \frac{N-1}{2} \ln T + \ln \frac{\Gamma(1/2)^N}{\Gamma(N/2)} + \frac{N-1}{2} \ln 2 + O(1). \tag{5.1.14}$$

5.1.2 Exponentiated Gradient

On présente maintenant l'algorithme Exponentiated Gradient (EG) [He98], on détermine les poids en connaissant uniquement le vecteur de valeurs relatives et les poids à l'instant t, $\mathbf{x}(t)$, $\boldsymbol{\pi}(t)$.

L'idée est de reprendre une méthode pour déterminer les poids dans une régression linéaire, Kivinen et Warmuth [KiWa97], montrent qu'un bon résultat est obtenu en gardant des poids "proches" à chaque nouvelle période. Pour notre problématique on cherche à maximiser la richesse logarithmique, on cherche donc à maximiser la fonction F définie par,

$$F(\boldsymbol{\pi}(t+1)) = \eta \log(\boldsymbol{\pi}_{t+1}\mathbf{x}(t)) - d(\boldsymbol{\pi}(t+1), \boldsymbol{\pi}(t)), \tag{5.1.15}$$

où $\eta > 0$ est le taux d'apprentissage et d est une mesure de distance servant de terme de pénalité servant à garder $\boldsymbol{\pi}_{t+1}$ dans un voisinage de $\boldsymbol{\pi}_t$. La distance utilisée est l'entropie relative,

$$D_{RE}(\mathbf{u}||\mathbf{v}) = \sum_{i=1}^{N} u_i \log \frac{u_i}{v_i} \tag{5.1.16}$$

Cette mesure est issue de la théorie de l'information, elle représente la quantité d'information délivrée par les deux vecteurs, il s'agit de la distance de Kullback-Leibler. Rappelons que nous sommes sous la contrainte $\sum_{i=1}^{N} \pi_i = 1$. Enfin, pour résoudre (approximativement) le problème, on fait un développement de Taylor au voisinage de $\boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(t)$,

$$\begin{aligned} \hat{F}(\boldsymbol{\pi}(t+1)) =& \eta \left(\log(\boldsymbol{\pi}(t) \cdot \mathbf{x}(t)) + \frac{\mathbf{x}(t)(\boldsymbol{\pi}(t+1) - \boldsymbol{\pi}(t))}{\boldsymbol{\pi}(t) \cdot \mathbf{x}(t)} \right) \\ & - d(\boldsymbol{\pi}(t+1), \boldsymbol{\pi}(t)) + \lambda(\sum_{i=1}^{N} \pi_i(t+1) - 1), \end{aligned} \tag{5.1.17}$$

λ est le multiplicateur de lagrange pour la contrainte de sommation. La condition du premier ordre s'écrit,

$$\begin{aligned}\frac{\partial \hat{F}(\boldsymbol{\pi}(t+1))}{\partial \pi_i(t+1)} = \eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot \mathbf{x}(t)} - \frac{\partial d(\boldsymbol{\pi}(t+1), \boldsymbol{\pi}(t))}{\partial \pi_i(t+1)} + \lambda = 0 \\ \eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot \mathbf{x}(t)} - \left(\log \frac{\pi_i(t+1)}{\pi_i(t)}\right) + \lambda = 0 \\ \pi_i(t+1) = \pi_i(t) \exp\left(\eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot \mathbf{x}(t)} + \lambda - 1\right),\end{aligned} \tag{5.1.18}$$

en normalisant, on obtient le portefeuille *exponentiated gradient* $\mathrm{EG}(\eta)$,

$$\pi_i(t+1) = \frac{\pi_i(t) \exp\left(\eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot \mathbf{x}(t)}\right)}{\sum_{j=1}^{m} \pi_j(t) \exp\left(\eta \frac{x_j(t)}{\boldsymbol{\pi}(t)\cdot \mathbf{x}(t)}\right)}. \tag{5.1.19}$$

Un choix raisonnable pour le portefeuille initial est de prendre $\boldsymbol{\pi}(1) = (1/N, 1/N, \cdots, 1/N)$.

Théorème 5.1.1. *Soit* $\mathbf{u} \in \mathcal{Q}$ *un portefeuille et* $\mathbf{x}(1), \cdots, \mathbf{x}(t)$ *une séquence de prix relatifs avec* $x_i(t) \geq r > 0$, $\forall i, t$ *et* $\max_i x_i(t) = 1$ *pour tout* t. *Alors pour tout* $\eta > 0$, *le pire ratio de richesse logarithmique pour la stratégie* $\boldsymbol{\pi}^{\mathrm{EG}}$ *est borné par,*

$$\mathcal{R}_T(\boldsymbol{\pi}^{\mathrm{EG}}, \mathcal{Q}) = \log \frac{W_T(\mathbf{u}, \mathbf{X})}{W_T(\boldsymbol{\pi}^{\mathrm{EG}}, \mathbf{X})} \leq \frac{D_{RE}(\mathbf{u}||\boldsymbol{\pi_1})}{\eta} + \frac{\eta T}{8r^2}, \tag{5.1.20}$$

et pour le cas où $\boldsymbol{\pi}(1) = (1/N, 1/N, \cdots, 1/N)$,

$$\mathcal{R}_T(\boldsymbol{\pi}^{\mathrm{EG}}, \mathcal{Q}) \leq \frac{\ln N}{\eta} + \frac{\eta T}{8r^2}, \tag{5.1.21}$$

pour η *optimal, la borne minimale est alors* $\frac{1}{r}\sqrt{\frac{T}{2}\ln N}$.

Preuve de (5.1.20). Posons $\Delta_t = DR_{RE}(\mathbf{u}||\boldsymbol{\pi}(t+1)) - DR_{RE}(\mathbf{u}||\boldsymbol{\pi}(t))$ et

$Z_t = \sum_{j=1}^{N} \pi_{j,t} \exp\left(\eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}\right)$, alors (avec $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathrm{EG}}$),

$$\begin{aligned}\Delta(t) &= -\sum_i u_i \log \frac{\pi_i(t+1)}{\pi_i(t)} \\ &= -\sum_i u_i \left(\frac{\eta x_i(t+1)}{\boldsymbol{\pi}_t \cdot \mathbf{x}(t)} - \log Z(t)\right) \\ &= -\eta \frac{\mathbf{u}\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)} + \log Z(t).\end{aligned} \tag{5.1.22}$$

Pour tout $\alpha \in [0,1]$, $\beta > 0$, $x \in \mathbb{R}$ et $y \in [0,1]$,

$$\log(1-\alpha(1-e^x)) \le (\alpha x + x^2/8) \quad (a), \qquad \beta^y \le (1-(1-\beta)y) \quad (b).$$

Alors, comme $x_i(t) \in [0,1]$, d'après (b) on a,

$$\begin{aligned}Z(t) &= \sum_i \pi_i(t) \exp\left(\eta \frac{x_i(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}\right) \\ &\le \sum_i \pi_i(t)(1-(1-\exp\left(\frac{\eta}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}\right))x_i(t)) \\ &= 1-(1-\exp\left(\frac{\eta}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}\right))\boldsymbol{\pi}(t)\cdot\mathbf{x}(t),\end{aligned} \tag{5.1.23}$$

et d'après (a),

$$\log Z(t) \le \eta + \frac{\eta^2}{8(\boldsymbol{\pi}(t)\cdot\mathbf{x}(t))}. \tag{5.1.24}$$

Ainsi,

$$\begin{aligned}\Delta(t) \le \eta\left(1 - \frac{\mathbf{u}\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\mathbf{x}(t)}\right) + \frac{\eta^2}{8(\boldsymbol{\pi}(t)\cdot\mathbf{x}(t))^2} \\ - \eta \log \frac{\mathbf{u}\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)} + \frac{\eta^2}{8(\boldsymbol{\pi}(t)\cdot\mathbf{x}(t))^2}.\end{aligned} \tag{5.1.25}$$

Enfin, comme $x_i(t) \ge r$, en sommant sur t on obtient,

$$\begin{aligned}-D_{RE}(\mathbf{u}||\boldsymbol{\pi}_1) &\le D_{RE}(\mathbf{u}||\boldsymbol{\pi}(t+1)) - D_{RE}(\mathbf{u}||\boldsymbol{\pi}(1)) \\ &\le \eta \sum_{s=1}^{T} (\log(\boldsymbol{\pi}(s)\cdot\mathbf{x}(s)) - \log(\mathbf{u}(s)\cdot\mathbf{x}(s))) + \frac{\eta^2 T}{8r^2}.\end{aligned} \tag{5.1.26}$$

□

Preuve de (5.1.21). D'après (5.1.20) et remarquant que $D_{RE}(\mathbf{u}||\boldsymbol{\pi}(1)) \leq \log m$ quand $\boldsymbol{\pi}(1)$ est le vecteur uniforme, la preuve est évidente.

□

EG(η) requiert de connaître la valeur de r (le minimum des prix relatifs normalisés) à l'avance pour calculer η, pour pallier à ce problème Helmbold et al. [He98] propose une version modifiée de EG(η) que l'on notera $\tilde{\text{EG}}(\eta, \alpha)$. La première modification consiste à shrinker par un réel $\alpha \in [0, 1]$ les prix relatifs,

$$\tilde{\mathbf{x}}(t) = (1 - \alpha/N)\mathbf{x}(t) + (\alpha/N)\mathbf{1}. \tag{5.1.27}$$

Ce qui nous permet d'avoir une borne minimale $\tilde{x}_i(t) \geq \alpha/N$. On construit un sous portefeuille identique à EG(η) en remplaçant $\mathbf{x}(t)$ par $\tilde{\mathbf{x}}(t)$,

$$\pi_{i,t+1} = \frac{\pi_i(t) \exp\left(\eta \frac{\tilde{x}_i(t)}{\boldsymbol{\pi}_t \cdot \mathbf{x}}(t)\right)}{\sum_{j=1}^{N} \pi_j(t) \exp\left(\eta \frac{\tilde{x}_j(t)}{\boldsymbol{\pi}(t) \cdot \mathbf{x}(t)}\right)}. \tag{5.1.28}$$

Enfin, on reparamétrise le portefeuille,

$$\tilde{\boldsymbol{\pi}}(t) = (1 - \alpha)\boldsymbol{\pi}(t) + (\alpha/N)\mathbf{1}. \tag{5.1.29}$$

Théorème 5.1.2. *Soit $\mathbf{u} \in \mathcal{Q}$ un portefeuille et $\mathbf{x}(1), \cdots, \mathbf{x}(t)$ une séquence de prix relatifs avec $x_i(t) \geq r > 0$, $\forall i, t$ et $\max_i x_i(t) = 1$ pour tout t. Pour tout $\alpha \in (0, 1/2]$ et $\eta > 0$, le pire ratio de richesse logarithmique du portefeuille $\tilde{\boldsymbol{\pi}}^{\text{EG}}$ est borné par,*

$$\mathcal{R}_T(\tilde{\boldsymbol{\pi}}^{\text{EG}}, \mathcal{Q}) \leq 2\alpha T - \frac{D_{RE}(\mathbf{u}||\tilde{\boldsymbol{\pi}}^{\text{EG}})}{\eta} - \frac{\eta T}{8(\alpha/N)^2}, \tag{5.1.30}$$

et pour le cas où $\boldsymbol{\pi}(1) = (1/N, 1/N, \cdots, 1/N)$, $T \geq 2N^2 \log m$, avec $\alpha = (N \log N/(8T))^{1/4}$ et $\eta = \sqrt{8\alpha^2 \log N/(N^2 T)}$ on a,

$$\mathcal{R}_T(\tilde{\boldsymbol{\pi}}^{\text{EG}}, \mathcal{Q}) \leq 2(2N^2 \log N)^{1/4} T^{3/4}. \tag{5.1.31}$$

Preuve de (5.1.30). Comme $\max_i x_{i,t} = 1$,

$$\frac{\tilde{\boldsymbol{\pi}}(t) \cdot \mathbf{x}(t)}{\boldsymbol{\pi}(t) \cdot \mathbf{x}(t)} \geq \frac{(1 - \alpha)\boldsymbol{\pi}(t) \cdot \mathbf{x}(t) + \alpha/N}{(1 - \alpha/N)\boldsymbol{\pi}(t) \cdot \mathbf{x}(t) + \alpha/N}, \tag{5.1.32}$$

le membre de droite est une fonction décroissante en $\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)$ et est donc minimum pour $\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)=1$, donc,

$$\frac{\tilde{\boldsymbol{\pi}}(t)\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}\geq(1-\alpha)+\alpha/N, \tag{5.1.33}$$

et de manière équivalente, sachant que $\ln(1-\alpha+\alpha/N)\geq\ln(1-\alpha)\geq-2\alpha$ pour tout $\alpha\in(0,1/2]$,

$$\begin{aligned}\log\frac{\tilde{\boldsymbol{\pi}}(t)\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}&\geq\ln((1-\alpha)+\alpha/N)\\ \log\frac{\tilde{\boldsymbol{\pi}}(t)\cdot\mathbf{x}(t)}{\boldsymbol{\pi}(t)\cdot\mathbf{x}(t)}&\geq 2\alpha.\end{aligned} \tag{5.1.34}$$

En appliquant le théorème 5.1.1 au prix relatifs $\tilde{\mathbf{x}}(t)$ on obtient,

$$RLS_T(\tilde{\boldsymbol{\pi}}^{\mathrm{EG}},\mathcal{Q})\geq\frac{D_{RE}(\mathbf{u}||\boldsymbol{\pi}(1))}{\eta}-\frac{\eta T}{8(\alpha/N)^2}, \tag{5.1.35}$$

car $\tilde{x}_i(t)\geq\alpha/N$. Pour conclure la preuve, on combine les deux dernières inégalités en sommant sur t. □

Preuve de (5.1.31). Comme pour la preuve du théorème 5.1.1, on utilise le fait que $D_{RE}(\mathbf{u},\boldsymbol{\pi}(1))\leq N$ quand $\boldsymbol{\pi}(1)$ est le vecteur uniforme.

□

Corollaire 5.1.1. *Le portefeuille* $\tilde{\mathrm{EG}}(\eta,\alpha)$ *est un portefeuille universel.*

Preuve. Soit $b=\lceil\log_2(T/(2N^2\log N))\rceil$, par le théorème 5.1.2 on à,

$$\begin{aligned}RLS_T(\tilde{\boldsymbol{\pi}}^{\mathrm{EG}},\mathcal{Q})\leq&4N^2\log N+\sum_{i=1}^{b}2^{1/4}(2^i)^{3/4}2N^2\log N\\ \leq&2^{5/4}N^2\log N\left(1+\sum_{i=0}^{b}(2^{3/4})^i\right)\\ =&2^{5/4}N^2\log N\left(1+\frac{(2^{3/4})^{b+1}-1}{2^{3/4}-1}\right)\\ \leq&6N^2\log N(1+(2^{3/4})^b)\\ \leq&6N^2\log N\left(1+\left(\frac{T}{2N^2\log N}\right)^{3/4}\right).\end{aligned} \tag{5.1.36}$$

Comme l'inégalité est vraie pour toute stratégie $\mathbf{u} \in \mathcal{Q}$, elle est également vraie pour le cas où $\mathbf{u} = \boldsymbol{\pi}^{bcrp}$, on a alors,

$$\log \frac{S_T(\boldsymbol{\pi}^{bpcr}, \mathbf{X})}{S_T(\tilde{\boldsymbol{\pi}}^{EG}, \mathbf{X})} \leq \frac{6N^2 \log N \left(1 + \left(\frac{T}{2N^2 \log N^2}\right)^{3/4}\right)}{T} \longrightarrow 0 \text{ quand } T \to \infty. \tag{5.1.37}$$

□

Pour conclure ce chapitre nous présentons un nouvel algorithme du type EG que nous appellerons EG-UNIV cf [StLu05], cette fois ne connaissant pas la valeur finale de T, c'est-à-dire la période de trading à l'avance. En initialisant toujours avec $\boldsymbol{\pi} = (1/N, \cdots, 1/N)$, et avec η et α adaptatif au cours du temps,

$$\tilde{\mathbf{x}}(t) = (1 - \alpha(t)/N)\mathbf{x}(t) + (\alpha(t)/N)\mathbf{1}, \qquad \alpha(t) = t^{-1/3}/2, \tag{5.1.38}$$

$$\tilde{\boldsymbol{\pi}}_{t+1} = -\exp\left\{\eta_t \sum_{s=0}^{t} \frac{\tilde{\mathbf{x}}_s}{\tilde{\boldsymbol{\pi}}_s \cdot \tilde{\mathbf{x}}_s}\right\}, \qquad \eta_t = t^{-2/3}/4. \tag{5.1.39}$$

Le portefeuille est alors,

$$\boldsymbol{\pi}(t+1) = (1 - \alpha(t))\frac{\tilde{\boldsymbol{\pi}}(t+1)}{\sum_{i=1}^{N} \boldsymbol{\pi}_i(t+1)} + (\alpha(t)/N)\mathbf{1}. \tag{5.1.40}$$

Théorème 5.1.3. *Soit $\mathbf{u} \in \mathcal{Q}$ un portefeuille et $\mathbf{x}(1), \cdots, \mathbf{x}(t)$ une séquence de prix relatifs avec $x_i(t) \geq r > 0$, $\forall i, t$ et $\max_i x_i(t) = 1$ pour tout t. Le regret externe de $\boldsymbol{\pi}^{\text{EG-UNIV}}$ est borné par,*

$$\mathcal{R}_t \leq 10NT^{2/3}. \tag{5.1.41}$$

Preuve.

$$\begin{aligned}\sum_{t=1}^{T}(\log(\mathbf{u}\cdot\mathbf{x}(t)) - \log(\boldsymbol{\pi}^{\text{EG-UNIV}}\cdot\mathbf{x}(t))) \leq &\sum_{t=1}^{T}(\log(\mathbf{u}\cdot\mathbf{x}(t)) - \log(\mathbf{u}\cdot\tilde{\mathbf{x}}(t))) \\ &+ \sum_{t=1}^{T}(\log(\mathbf{u}\cdot\tilde{\mathbf{x}}(t)) - \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}}\cdot\tilde{\mathbf{x}}(t))) \\ &\sum_{t=1}^{T}(\log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}}\cdot\tilde{\mathbf{x}}(t)) \\ &\quad - \log(\boldsymbol{\pi}^{\text{EG-UNIV}}\cdot\tilde{\mathbf{x}}(t))).\end{aligned} \tag{5.1.42}$$

Pour le terme de droite, comme $\mathbf{x}(t)$ est renormalisé tel que ses composants soient inférieurs ou égaux à 1, la première somme est négative. La troisième somme est inférieure à $2(\alpha(1) + \cdots + \alpha(T))$ car,

$$\log(\boldsymbol{\pi}^{\text{EG-UNIV}}(t) \cdot \mathbf{x}(t)) \geq \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}}(t) \cdot \tilde{\mathbf{x}}(t)) - 2\alpha(t). \tag{5.1.43}$$

Il nous reste donc à travailler sur la seconde somme. Notons $\ell_i(t) = -\dfrac{\tilde{x}_i(t)}{\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}}(t)$, alors,

$$\sum_{t=1}^{T} \left(\log(\mathbf{u} \cdot \tilde{\mathbf{x}}(t)) - \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t))\right) \leq \sum_{j=1}^{N} u_j \left(\sum_{t=1}^{T}\sum_{i=1}^{N} \tilde{\pi}_i^{\text{EG-UNIV}}(t)\ell_i(t) - \ell_j(t)\right), \tag{5.1.44}$$

$\eta(t)$ est croissant, et comme (on assume l'inégalité suivante),

$$\begin{aligned}\sum_{t=1}^{T} \boldsymbol{\pi}(t) \cdot \mathbf{x}(t) - \max_i \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t) \geq &- \left(\frac{2}{\eta(T+1)} - \frac{1}{\eta(1)}\right) \log N \\ &- \sum_{t=1}^{T} \frac{1}{\eta(t)} \log \sum_{i=1}^{N} \pi_i(t) \exp\left(\eta_t(x_i(t) - \pi_i(t)x_i(t))\right).\end{aligned} \tag{5.1.45}$$

On a,

$$\begin{aligned}\sum_{t=1}^{T} \left(\log(\mathbf{u} \cdot \tilde{\mathbf{x}}(t)) - \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t))\right) \leq &\left(\frac{2}{\eta(T+1)} - \frac{1}{\eta(1)}\right) \log N \\ &+ \sum_{t=1}^{T} \frac{1}{\eta(t)} \log \sum_{i=1}^{N} \pi_i(t) \exp\left(\eta_t(x_i(t) - \pi_i(t)x_i(t))\right).\end{aligned} \tag{5.1.46}$$

Pour continuer l'algorithme on a besoin d'introduire la proposition suivante (on omet la preuve),

Proposition 5.1.1. *Soit $(x_1(t), \cdots, x_N(t))$ le vecteur des prix relatifs, alors pour tout $\eta(t) \geq 0$ on a,*

$$\frac{1}{\eta(t)} \log \sum_{i=1}^{N} \pi_{i=1}^{N} \exp(\eta_t(x_i(t) - \pi_i(t)x_i(t))) \leq 8\eta(t) \left(\sum_{i=1}^{N} \pi_i(t)x_i^2(t) - \left(\sum_{i=1}^{N} \pi_i(t)x_i(t)\right)^2\right). \tag{5.1.47}$$

Alors d'après l'algorithme de EG-UNIV on a $\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t) \geq \alpha(t)/N$ et comme $\ell_i(t) \in [-N/\alpha_t, 0]$, le choix de $\alpha(t)$ avec la proposition 5.1.1 et $N\eta(t)/\alpha(t) \leq 1$,

$$\sum_{t=1}^{T} \Big(\log(\mathbf{u} \cdot \tilde{\mathbf{x}}(t)) - \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t))\Big) \leq \left(\frac{2}{\eta(T+1)} - \frac{1}{\eta(1)}\right) \log N + 8 \sum_{t=1}^{T} \eta_t \sum_{i=1}^{N} \tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \frac{\tilde{x}_i^2(t)}{(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t))^2}, \tag{5.1.48}$$

réutilisant le fait que $\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t) \geq \alpha(t)/N$ et $N\eta(t)/\alpha(t) \leq 1$,

$$\sum_{t=1}^{T} \Big(\log(\mathbf{u} \cdot \tilde{\mathbf{x}}(t)) - \log(\tilde{\boldsymbol{\pi}}^{\text{EG-UNIV}} \cdot \tilde{\mathbf{x}}(t))\Big) \leq \left(\frac{2}{\eta(T+1)} - \frac{1}{\eta(1)}\right) \log N + 8 \sum_{t=1}^{T} \frac{\eta(t)}{\alpha(t)}, \tag{5.1.49}$$

et enfin,

$$\sum_{t=1}^{T} \Big(\log(\mathbf{u} \cdot \mathbf{x}(t)) - \log(\boldsymbol{\pi}^{\text{EG-UNIV}} \cdot \mathbf{x}(t))\Big) \leq \left(\frac{2}{\eta(T+1)} - \frac{1}{\eta(1)}\right) \log N + 8N \sum_{t=1}^{T} \frac{\eta(t)}{\alpha(t)} + 2 \sum_{t=1}^{T} \alpha(t), \tag{5.1.50}$$

le remplacement de $\eta(t)$ et $\alpha(t)$ par les valeurs proposées conclut la preuve. □

Pour conclure, notons que l'algorithme initial de Cover a un regret externe de l'ordre de T, $\tilde{\text{EG}}$ de l'ordre de $T^{3/4}$ et EG-UNIV de l'ordre de $T^{2/3}$.

5.2 Portefeuille de Markowitz

5.2.1 Moyenne - Variance

Dans cette partie, on va essayer de maximiser le rendement du portefeuille et de minimiser son risque. On va parler de *portefeuille de variance minimum*, au lieu de chercher à maximiser son rendement ou de minimiser son regret, on cherche à minimiser la variance. Le problème s'écrit simplement,

$$\mathcal{P} : \begin{cases} \min\limits_{\boldsymbol{\pi}} \dfrac{1}{2} \boldsymbol{\pi}^\top \Sigma \boldsymbol{\pi} \\ \boldsymbol{\pi}^\top \mu = \mu^* \\ \displaystyle\sum_{i=1}^{N} \pi_i = 1. \end{cases} \tag{5.2.1}$$

où $\boldsymbol{\pi}$ est le vecteur des poids, μ la moyenne des rendements, Σ la matrice de covariance et μ^* le rendement espéré. Il s'agit du portefeuille de Markowitz. Le lagrangien s'écrit, avec λ et γ les multiplicateurs,

$$\mathcal{L} = \frac{1}{2}\boldsymbol{\pi}^\top \boldsymbol{\Sigma}\boldsymbol{\pi} - \lambda(\boldsymbol{\pi}^\top \mu - \mu^*) - \gamma(\boldsymbol{\pi}^\top \mathbf{1} - 1). \tag{5.2.2}$$

La condition du premier ordre s'écrit,

$$\begin{cases} \dfrac{d\mathcal{L}}{d\boldsymbol{\pi}} = 0 \\ \dfrac{d\mathcal{L}}{d\lambda} = 0 \\ \dfrac{d\mathcal{L}}{d\boldsymbol{\pi}} = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \boldsymbol{\Sigma}\boldsymbol{\pi} - \lambda\mu - \gamma\mathbf{1} = 0 \\ \boldsymbol{\pi}^\top \mu - \mu^* = 0 \\ \boldsymbol{\pi}^\top \mathbf{1} - 1 = 0, \end{cases} \tag{5.2.3}$$

ainsi,

$$\boldsymbol{\pi} = \boldsymbol{\Sigma}^{-1}(\lambda\mu + \gamma\mathbf{1}). \tag{5.2.4}$$

En arrangeant les équations on obtient,

$$\lambda\mu^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} + \gamma\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} = 1, \tag{5.2.5}$$

et

$$\mu^* = \lambda\mu^\top \boldsymbol{\Sigma}^{-1}\mu + \gamma\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mu. \tag{5.2.6}$$

En écrivant sous forme matricielle les deux dernières équations on a,

$$\begin{pmatrix} \mu^* \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda\mu^\top \boldsymbol{\Sigma}^{-1}\mu + \gamma\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mu \\ \lambda\mu^\top \boldsymbol{\Sigma}^{-1}\mu + \gamma\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} \end{pmatrix} = \begin{pmatrix} \mu^\top \boldsymbol{\Sigma}^{-1}\mu & \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mu \\ \mu^\top \boldsymbol{\Sigma}^{-1}\mu & \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} \end{pmatrix} \begin{pmatrix} \lambda \\ \gamma \end{pmatrix}. \tag{5.2.7}$$

En posant

$$A = \mu^\top \boldsymbol{\Sigma}^{-1}\mu, \qquad B = \mu^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}, \qquad C = \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}, \tag{5.2.8}$$

on a

$$\begin{pmatrix} \lambda \\ \gamma \end{pmatrix} = \frac{1}{AC - BB} \begin{pmatrix} C & -B \\ -B & A \end{pmatrix} \begin{pmatrix} \mu^* \\ 1 \end{pmatrix}. \tag{5.2.9}$$

Les multiplicateurs sont donc,

$$\lambda = \frac{C\mu^* - B}{AC - BB}, \qquad \gamma = \frac{-B\mu^* + A}{AC - BB} \tag{5.2.10}$$

La stratégie optimale est alors,

$$\begin{aligned}\boldsymbol{\pi}^* =& \boldsymbol{\Sigma}^{-1}(\lambda\mu + \gamma\mathbf{1}) \\ =& \boldsymbol{\Sigma}^{-1}\frac{(A\mathbf{1} - B\mu) + \mu^*(C\mu - B\mathbf{1})}{AC - BB},\end{aligned} \tag{5.2.11}$$

avec pour variance minimale,

$$\begin{aligned}\sigma^* =& \boldsymbol{\pi}^\top \boldsymbol{\Sigma}\boldsymbol{\pi} \\ =& \lambda\mu^* + \gamma \\ =& \frac{C\mu^* - 2B\mu^* + A}{AC - BB}.\end{aligned} \tag{5.2.12}$$

Le point qui mérite le plus d'attention est l'estimation de la matrice de covariance. On rappelle déjà que pour X et Y deux variables aléatoires réelles on a,

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y), \tag{5.2.13}$$

et pour N variables aléatoires, $X_1, \cdots, X_N$, $X_i \in \mathbb{R}^T$, $i = 1, \cdots, N$, la matrice de covariance est,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}X_1 & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_N) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}X_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_N, X_1) & \cdots & \cdots & \mathrm{Var}X_N \end{pmatrix}. \tag{5.2.14}$$

Notons $\boldsymbol{\Sigma}$ la vraie matrice de covariance et $\mathbf{S}$ son estimation. Le risk 'in-sample', donc construit avec $\mathbf{S}$, est donné par,

$$\mathcal{R}_{\text{in}} = \mathbf{w}_{\mathbf{S}}^\top \mathbf{S}\mathbf{w}_{\mathbf{S}}. \tag{5.2.15}$$

Le 'vrai' risque minimal, quand $\boldsymbol{\Sigma}$ est connu parfaitement est,

$$\mathcal{R}_{\text{true}} = \mathbf{w}_{\boldsymbol{\Sigma}}^\top \boldsymbol{\Sigma}\mathbf{w}_{\boldsymbol{\Sigma}}. \tag{5.2.16}$$

Le risque 'out-of-sample', finalement celui qui nous intéresse le plus, est naturellement,

$$\mathcal{R}_{\text{out}} = \mathbf{w}_{\mathbf{S}}^{\top} \mathbf{\Sigma} \mathbf{w}_{\mathbf{S}}. \tag{5.2.17}$$

Intuitivement, on a,

$$\mathcal{R}_{\text{in}} \leq \mathcal{R}_{\text{true}} \leq \mathcal{R}_{\text{out}}. \tag{5.2.18}$$

On peut montrer que [PoBoLa05],

$$\mathcal{R}_{\text{in}} = \mathcal{R}_{\text{true}} \sqrt{1 - N/T} = \mathcal{R}_{\text{out}}(1 - N/T). \tag{5.2.19}$$

Supposons que nous essayons de construire un portefeuille consitué de tous les éléments de S&P 500 et que nous ayons $t = 2500$, ce qui correspond à 10 ans d'historique, alors on a $q := N/t = 0.2$, or, pour que l'estimation ait quelques degrés d'universalité, il faudrait que $q \to 0$, ainsi, $\mathcal{R}_{\text{in}} = \mathcal{R}_{\text{true}} = \mathcal{R}_{\text{out}}$. Intuitivement, les données ne contiennent pas assez d'information pour estimer la 'vraie' matrice de covariance. On propose deux manières distinctes pour donner une 'meilleure' estimation de la matrice de covariance.

5.2.2 Shrinkage

Une fois de plus dans ce cours, on va shrinker l'estimateur pour avoir un résultat plus stable. Une idée simple est de prendre comme estimateur,

$$\mathbf{\Sigma}^{SH} = \rho \nu_F \mathbf{I} + \rho \mathbf{S}, \tag{5.2.20}$$

où ρ et ν sont les paramètres de shrinkage (on verra par la suite que seul ρ est le paramètre de shrinkage), $\mathbf{I} \in \mathbb{R}^{N \times N}$ la matrice identitée et S est l'estimateur classique de la matrice de covariance,

$$\mathbf{S} = \frac{1}{T} \mathbf{X}_c \mathbf{X}_c^{\top}, \tag{5.2.21}$$

où $\mathbf{X}_c$ est la matrice des séries centrées.

Il nous faut maintenant déterminer les valeurs optimales de ν et ρ. On cherche toujours à minimiser une fonction de perte.

Définition 5.2.1 (norme de Frobenius)**.** *Pour* $\mathbf{Z} = \{z_{ij}\}_{i,j=1,\cdots,N}$ *une matrice* $N \times N$ *symétrique avec pour valeurs propres* $(\lambda_i)_{i=1\cdots,N}$*, la norme de Frobenius est définie par,*

$$||\mathbf{Z}||^2 = tr\mathbf{Z}^2 = \sum_{i=1}^{N}\sum_{i=1}^{N} z_{ij}^2/N = \sum_{i=1}^{N} \lambda_i^2/N. \tag{5.2.22}$$

Le scalaire associé est, pour $\mathbf{Z}_1$ *et* $\mathbf{Z}_2$ *deux matrices symétriques,*

$$< \mathbf{Z}_1, \mathbf{Z_2} >= tr(\mathbf{Z}_1\mathbf{Z}_2^\top)/N. \tag{5.2.23}$$

Proposition 5.2.1. *Utilisant (5.2.21) comme estimateur de la matrice de covariance et la norme de Frobenius comme fonction de perte,* $\ell(\rho, \nu) = ||\mathbf{\Sigma}^{SH} - \mathbf{\Sigma}||^2$*, le problème s'écrit,*

$$\mathcal{P} : \begin{cases} \min\limits_{\rho,\nu} & \mathbb{E}||\mathbf{\Sigma}^{SH} - \mathbf{\Sigma}||^2 \\ s.c & \mathbf{\Sigma}^{SH} = \rho\nu\mathbf{I} + (1-\rho)\mathbf{S}. \end{cases} \tag{5.2.24}$$

La solution est donnée par

$$\frac{\mathbb{E}||\mathbf{S} - \mathbf{\Sigma}||^2}{\mathbb{E}||\mathbf{S} - < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2} < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I} + \frac{||\mathbf{\Sigma} - < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2}{\mathbb{E}||\mathbf{S} - < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2}\mathbf{S}, \tag{5.2.25}$$

et le risque est,

$$\mathbb{E}||\mathbf{\Sigma}^{SH} - \mathbf{\Sigma}||^2 = \frac{||\mathbf{\Sigma} - < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2 \mathbb{E}||\mathbf{S} - \mathbf{\Sigma}||^2}{\mathbb{E}||\mathbf{S} - < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2}. \tag{5.2.26}$$

Preuve.

$$\begin{aligned} \mathbb{E}||\mathbf{\Sigma}^{SH} - \mathbf{\Sigma}||^2 =& \mathbb{E}||(\rho\nu\mathbf{I} + (1-\rho)\mathbf{S}) - \mathbf{\Sigma}||^2 \\ =& \mathbb{E}||\rho\nu\mathbf{I} + \rho\mathbf{\Sigma} - \rho\mathbf{\Sigma} + (1-\rho)\mathbf{S} - \mathbf{\Sigma}||^2 \\ =& \mathbb{E}||\rho\nu\mathbf{I} - \rho\mathbf{\Sigma} + (1-\rho)(\mathbf{S} - \mathbf{\Sigma})||^2 \\ =& \mathbb{E}||\rho\nu\mathbf{I} - \rho\mathbf{\Sigma}||^2 + \mathbb{E}||(1-\rho)(\mathbf{S} - \mathbf{\Sigma})||^2 \\ &+ 2\rho(1-\rho)\mathbb{E} < \nu\mathbf{I} - \mathbf{\Sigma}, \mathbf{S} - \mathbf{\Sigma} > \\ =& \rho^2\mathbb{E}||\nu\mathbf{I} - \mathbf{\Sigma}||^2 + (1-\rho)^2\mathbb{E}||\mathbf{S} - \mathbf{\Sigma}||^2, \end{aligned} \tag{5.2.27}$$

comme $||\nu\mathbf{I} - \mathbf{\Sigma}||^2 = \nu^2 - 2\nu < \mathbf{\Sigma}, \mathbf{I} > + ||\mathbf{\Sigma}||^2$, la condition du premier ordre pour ν s'écrit,

$$\frac{\partial\mathbb{E}||\mathbf{\Sigma}^{SH} - \mathbf{\Sigma}||^2}{\partial\nu} = 2\nu - 2 < \mathbf{\Sigma}, \mathbf{I} >= 0. \tag{5.2.28}$$

Ainsi,

$$\nu^* =< \mathbf{\Sigma}, \mathbf{I} > . \tag{5.2.29}$$

En dérivant maintenant par rapport à ρ on trouve,

$$\frac{\partial \mathbb{E}\ell(\rho,\nu)}{\partial \rho} = 2\rho\mathbb{E}||\nu\mathbf{I}-\mathbf{\Sigma}||^2 - 2(1-\rho)\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2. \tag{5.2.30}$$

La condition du premier ordre est,

$$\begin{aligned} 2\rho\mathbb{E}||\nu\mathbf{I}-\mathbf{\Sigma}||^2 - 2(1-\rho)\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2 &= 0 \\ \rho &= \frac{\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2}{\mathbb{E}||\nu\mathbf{I}-\mathbf{\Sigma}||^2 + \mathbb{E}||\mathbf{S}-\Sigma||^2}. \end{aligned} \tag{5.2.31}$$

Remarquons maintenant que,

$$\begin{aligned} \mathbb{E}||\mathbf{S}-\nu\mathbf{I}||^2 =& \mathbb{E}||\mathbf{S}-\mathbf{\Sigma}+\mathbf{\Sigma}-\nu\mathbf{I}||^2 \\ =& \mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2 + \mathbb{E}||\mathbf{\Sigma}-\nu\mathbf{I}||^2 + \mathbb{E} < \mathbf{S}-\mathbf{\Sigma}, \mathbf{\Sigma}-\nu\mathbf{I} > \\ =& \mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2 + \mathbb{E}||-(\mathbf{\Sigma}+\nu\mathbf{I})||^2 \\ =& \mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2 + \mathbb{E}||-\mathbf{\Sigma}+\nu\mathbf{I}||^2. \end{aligned} \tag{5.2.32}$$

Ainsi,

$$\rho^* = \frac{\mathbb{E}||\mathbf{S}-\Sigma||^2}{\mathbb{E}||\mathbf{S}-\nu\mathbf{I}||^2}. \tag{5.2.33}$$

□

Le pourcentage de gain relatif du shrinkage est,

$$\frac{\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2 - \mathbb{E}||\mathbf{\Sigma}^{SH}-\mathbf{\Sigma}||^2}{\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2} = \frac{\mathbb{E}||\mathbf{S}-\mathbf{\Sigma}||^2}{\mathbb{E}||\mathbf{S}- < \mathbf{\Sigma}, \mathbf{I} > \mathbf{I}||^2}. \tag{5.2.34}$$

Les paramètres ρ^* et ν^* dépendent de $\mathbf{\Sigma}$, la vraie matrice de covariance, qui est inobservable, on ne peut donc pas effectuer le shrinkage tel quel. On a les estimateurs suivants,

$$\begin{aligned} a =& < \mathbf{S}_t, \mathbf{I}_t >_t \longrightarrow < \mathbf{\Sigma}_t, \mathbf{I}_t >_t \\ b =& ||\mathbf{S}_t - a\mathbf{I}_t||^2 \longrightarrow \mathbb{E}||\mathbf{S}_t - a\mathbf{I}_t||_t^2 \\ c =& \min\left(\frac{1}{t^2}\sum_{k=1}^{t} ||\mathbf{X}_k\mathbf{X}_k^\top - \mathbf{S}_t||_t^2, b\right) \longrightarrow \mathbb{E}||\mathbf{S}_n - \mathbf{\Sigma}_t||^2 \\ d =& b - c \longrightarrow ||\mathbf{\Sigma}_t - a\mathbf{I}_t||_t^2. \end{aligned} \tag{5.2.35}$$

Dans cette partie on a supposé aucune distribution particulière des rendements, ce qui est agréable. Néanmoins, pour que les convergences (5.2.35) se produisent, il faut que pour chacun des rendements le moment d'ordre 4 existe.

On peut maintenant effectuer un autre shrinkage, un peu plus intuitif, on estime deux manières différentes la matrice de covariance, une première fois classiquement avec l'estimateur non biaisé (5.2.21) et en prenant la covariance de la modélisation des rendements,

$$x_i(t) = \beta_{0,i} + \beta_{m,i} x_m(t) + \varepsilon_i(t), \qquad i = 1, \cdots, N, \;\; t = 1, \cdots, T, \tag{5.2.36}$$

où x_m est le rendement du marché et ε_i est un bruit blanc de variance δ_{ii}. Il s'agit (à une constante près, le taux sans risque devrait être retranché des rendements) du modèle CAPM. β est le fameux "beta" du marché, il s'agit de la sensibilité de l'actif par rapport au marché. On l'interprète souvent comme une mesure de risque, s'il est proche de 1, c'est qu'il va dans le même sens que le marché, si proche de 0, c'est qu'il va dans "n'importe quel" sens. Bien sûr, si on pense que c'est une mesure de risque, c'est que l'on pense que le marché est non risqué et qu'il croît de manière constante et bien gentiment... Habituellement, on ne note pas β_0 mais α, il s'agit du α de Jensen, c'est-à-dire le rendement espéré de l'actif ceteris paribus. En d'autre termes, c'est l'excès de rendement en dehors du marché, nous pourrons nous reporter aux travaux de Fama, Tobin, etc. pour plus de détails.

Enfin, pour continuer dans la parenthèse du modèle CAPM, notons que le portefeuille au sens de Markowitz nous permet de construire ce que l'on appelle la frontière efficiente, c'est-à-dire l'ensemble des portefeuilles optimaux en fonction du risque et du rendement. Intuitivement, plus le risque augmente, plus le rendement du portefeuille augmente. Il faut donc faire un choix, soit on augmente le rendement et donc mécaniquement le risque, mais nous allons un peu plus 'transpirer le soir' avec une formule 'offensive' (comprendra les initiés) soit on diminue les deux. La frontière n'est pas linéaire mais forme un ensemble convexe, un 'U' tourné, la partie ouverte à droite. En bas, on parle de portefeuilles dominés, en haut, de portefeuilles dominants. Pour trouver le 'meilleur' choix, on peut tracer la droite issue du rendement sans risque et pour coefficient directeur β. Le portefeuille optimal étant le point de rencontre du CAPM avec la frontière efficiente.

En réécrivant le modèle (5.2.36) sous forme vectoriel en i on a,

$$\mathbf{x}(t) = \boldsymbol{\beta}_0 + \boldsymbol{\beta} x_m(t) + \boldsymbol{\varepsilon}(t), \qquad t = 1, \cdots, T, \tag{5.2.37}$$

en utilisant l'indépendance des ε on obtient la matrice de covariance,

$$\begin{aligned}\boldsymbol{\Phi} =& \mathrm{Cov}(\mathbf{x}(t), \mathbf{x}(t)) \\ =& \mathrm{Cov}(\boldsymbol{\beta}_0 + \boldsymbol{\beta} x_m(t) + \boldsymbol{\varepsilon}(t), \boldsymbol{\beta}_0 + \boldsymbol{\beta} x_m(t) + \boldsymbol{\varepsilon}(t)) \\ =& \mathrm{Cov}(\boldsymbol{\beta} x_m(t), \boldsymbol{\beta} x_m(t)) + 2\mathrm{Cov}(\boldsymbol{\beta} x_m(t), \varepsilon(t)) + \mathrm{Cov}(\boldsymbol{\varepsilon}(t), \boldsymbol{\varepsilon}(t)) \\ =& \sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}^\top + \Delta,\end{aligned} \tag{5.2.38}$$

où σ_m^2 est la variance du marché et $\boldsymbol{\Delta}$ est la matrice de covariance des bruits blancs, $\boldsymbol{\Delta} = \{\delta_{ij}\}_{i=j=1,\cdots,N}$ si $i = j$.

Bien évidemment $\mathbf{S} \neq \Phi$ puisque Φ est construite uniquement à partir du marché.

Le shrinkage étudié va être,

$$\Sigma^{CAPM} = \alpha \boldsymbol{\Phi} + (1 - \alpha)\mathbf{S}. \tag{5.2.39}$$

Pour déterminer le coefficient de shrinkage α nous réutilisons la norme de Frobenuis et la fonction de perte est,

$$\ell(\alpha) = ||\alpha \boldsymbol{\Phi} + (1 - \alpha)\mathbf{S} - \boldsymbol{\Sigma}||^2. \tag{5.2.40}$$

En notant $\tilde{\phi}_{ij}$ les $i - j$ entrées estimées de $\boldsymbol{\Phi}$, s_{ij} celle de $\mathbf{S}$, et σ_{ij} celle de $\boldsymbol{\Sigma}$, le risque à minimiser est,

$$
\begin{aligned}
\mathbb{E}\ell(\alpha) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}(\alpha\tilde{\phi}_{ij} + (1-\alpha)s_{ij} - \sigma_{ij})^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Var}(\alpha\tilde{\phi}_{ij} + (1-\alpha)s_{ij}) + (\mathbb{E}(\alpha\tilde{\phi}_{ij} + (1-\alpha)s_{ij} - \sigma_{ij}))^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^2\mathrm{Var}(\tilde{\phi}_{ij}) + (1-\alpha)^2\mathrm{Var}(s_{ij}) + 2\alpha(1-\alpha)\mathrm{Cov}(\tilde{\phi}_{ij}, s_{ij}) \\
&\quad + (\mathbb{E}(\alpha\tilde{\phi}_{ij} + s_{ij} - \alpha s_{ij} + \alpha\sigma_{ij} - \alpha\sigma_{ij} - \sigma_{ij}))^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^2\mathrm{Var}(\tilde{\phi}_{ij}) + (1-\alpha)^2\mathrm{Var}(s_{ij}) + 2\alpha(1-\alpha)\mathrm{Cov}(\tilde{\phi}_{ij}, s_{ij}) \\
&\quad + (\alpha^2\mathbb{E}(\tilde{\phi}_{ij} - \sigma_{ij}))^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^2\mathrm{Var}(\tilde{\phi}_{ij}) + (1-\alpha)^2\mathrm{Var}(s_{ij}) + 2\alpha(1-\alpha)\mathrm{Cov}(\tilde{\phi}_{ij}, s_{ij}) \\
&\quad + \alpha^2(\phi_{ij} - \sigma_{ij})^2.
\end{aligned}
\tag{5.2.41}
$$

La dérivée du risque par rapport à α est,

$$
\frac{\partial\mathbb{E}\ell(\alpha)}{\partial\alpha} = 2\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha\mathrm{Var}(\tilde{\phi}_{ij}) - (1-\alpha)\mathrm{Var}(s_{ij}) + (1-2\alpha)\mathrm{Cov}(\tilde{\phi}_{ij}, s_{ij}) + \alpha(\tilde{\phi}_{ij} - \sigma_{ij})^2. \tag{5.2.42}
$$

La condition du premier ordre nous donne,

$$
\alpha^* = \frac{\sum_{i,j}^{N}\mathrm{Var}(s_{ij}) - \mathrm{Cov}(\tilde{\phi}_{ij}, s_{ij})}{\sum_{i,j=1}^{N}\mathrm{Var}(\tilde{\phi}_{ij} - s_{ij}) + (\phi_{ij} - \sigma_{ij})^2}. \tag{5.2.43}
$$

Comme précédemment, l'intensité de shrinkage optimale dépend de paramètre non observable. Après quelques calculs, voir [LeWo00] pour plus de détails, on peut montrer $\tilde{\alpha}$ est un estimateur consistant de α, où,

$$
\tilde{\alpha} = \frac{1}{T}\frac{p-r}{c}, \tag{5.2.44}
$$

avec,

$$\begin{aligned} p_{ij} =& \sum_{t=1}^{T} \left\{ ((x_{it} - m_i)(x_{it} - m_i) - s_{ij})^2 \right\} \\ r_{ijt} =& \frac{s_{jm}s_{mm}(s_{it} - m_i) + s_{im}s_{mm}(s_{jt} - m_j) - s_{im}s_{jm}(s_{mt} - m_m)}{s_{mm}^2} \times \\ & (x_{mt} - m_m)(x_{it} - m_i)(x_{jt} - m_j) - f_{ij}s_{ij} \\ c_{ij} =& (\tilde{\phi}_{ij} - s_{ij})^2. \end{aligned} \tag{5.2.45}$$

Pour conclure, notons que la dérivée seconde du risque est,

$$\frac{\partial^2 \mathbb{E}\ell(\alpha)}{\partial \alpha^2} = 2 \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Var}(\tilde{\phi}_{ij} - s_{ij}) + (\phi_{ij} - \sigma_{ij})^2, \tag{5.2.46}$$

$\frac{\partial^2 \mathbb{E}\ell(\alpha)}{\partial \alpha^2}$ est donc positif, donc α^* est bien un minimum.

5.2.3 Matrice Aléatoire

La théorie des matrices aléatoires a été originellement développée pour la physique nucléaire et la représentation des interactions dans un noyau. Une matrice aléatoire est constituée d'éléments aléatoires réels de variance unitaire et de moyenne zéro. Dans une étude portant sur une centaine d'actifs du NYSE [LaCiBoPo99], il est montré que la distribution spectrale des valeurs propres de la matrice de covariance empirique, sauf les plus importantes, est très proche de la distribution spectrale d'une matrice aléatoire symétrique semi-définie positive. Si l'ensemble d'actifs étudiés peut être prédit par la théorie des matrices aléatoires, alors aucune information ne peut être extraite et (5.2.21) ne peut être appliquée. En revanche, si quelques valeurs propres dévient de cette théorie, on peut essayer de se focaliser sur celle-ci puisque les autres ne donnent pas d'information. Le problème adressé est d'extraire l'information "importante" et d'enlever le bruit, on parlera de *cleaning matrix*. L'étude de la statistique des valeurs propres est le point central de cette théorie.

Définition 5.2.2 (matrice hermitienne)**.** *Une matrice hermitienne, ou auto-adjointe, est une matrice carrée avec des éléments complexes dont la transposée de la matrice conjugée est égale à la matrice. En particulier, une matrice à éléments réels est hermitienne ssi elle est symétrique. Notons de plus qu'une telle matrice est orthogonalement diagonalisable et toutes ses valeurs propres sont réelles.*

F. Dyson a introduit une classification pour les ensembles de matrices gaussiennes. Nous présentons les trois classes. Pour toute la suite on se place sur l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ où Ω est l'ensemble des matrices, $\mathcal{F}$ est une σ-algèbre et $\mathbb{P}$ est une mesure définie sur cet espace.

Définition 5.2.3 (gaussien orthogonal)**.** *Soit* $\mathbf{X}$ *une matrice* $N \times N$ *avec pour entrée des variables réelles iid* $N(0,1)$*. L'ensemble des matrices gaussiennes orthogonales (GOE) est l'ensemble des matrices Hermitiennes tel que* $\mathbf{H}_N = (\mathbf{X} + \mathbf{X}^\top)/2$*. Les éléments sur la diagonale sont iid* $N(0,1)$ *et en dehors iid* $N(0,1/2)$*. On peut écrire les entrées de la matrice comme suit,*

$$h_{lm} \sim \quad N\left(0, \frac{1+\delta_{lm}}{2}\right), \quad 1 \leq l \leq m \leq N, \tag{5.2.47}$$

avec δ_{lm} *le symbole de Kronecker.*

Définition 5.2.4 (gaussien unitaire)**.** *Soit* $\mathbf{X}$ *une matrice* $N \times N$ *constituée d'éléments complexes iid* $N(0,1)$*. L'ensemble des matrices gaussiennes unitaires (GUE) est l'ensemble des matrices Hermitienne tel que* $\mathbf{H}_N = (\mathbf{X} + \mathbf{X}^*)/2$ *où* $\mathbf{X}^*$ *est la transposée Hermitienne de* $\mathbf{X}$*,*

$$(\mathbf{X}^*)_{ij} = (\bar{\mathbf{X}})_{ji}. \tag{5.2.48}$$

Les entrées de $\mathbf{H}_N$ *sont donc* $N(0,1)$ *iid sur la diagonale et iid* $N(0,1/2)$ *en dehors de la diagonale. On peut écrire les entrées de la matrice comme suit,*

$$\begin{gathered} \Re h_{lm},\, \Im h_{lm} \sim N\left(0, \frac{1}{2}\right), \quad 1 \leq l < m \leq N \\ h_{kk} \sim N(0,1), 1 \leq k \leq N. \end{gathered} \tag{5.2.49}$$

Ces deux premiers ensembles appartiennent à l'ensemble des matrices de Wigner, [Wi55], réelle ou complexe. Dans cet ensemble, les entrées ne sont plus nécessairement gaussiennes, mais nous nous contenterons pour ce cours du cas gaussien.

Avant d'introduire la dernière classe rappellons qu'un quaternion réel q est construit par,

$$q = a + ib + jc + kd, \quad a, b, c, d \in \mathbb{R}, \tag{5.2.50}$$

et i, j, k vérifient,

$$i^2 = j^2 = k^2 = ijk = -1. \tag{5.2.51}$$

Définition 5.2.5 (gaussien symplectique)**.** *Soit* $\mathbf{X}$ *une matrice* $N \times N$ *constituée de quaternions réels et iid* $N(0,1)$*. L'ensemble des matrices gaussiennes symplectiques (GSE) est l'ensemble des matrices Hermitiennes tel que* $\mathbf{H}_N = (\mathbf{X} + \mathbf{X}^D)/2$ *où* D *est l'opérateur de transposé dual. Les éléments diagonaux de* $\mathbf{H}_N$ *sont iid* $N(0,1)$ *et en dehors iid* $N(0,1/2)$*. On peut écrire les entrées de la matrice comme suit,*

$$\begin{aligned} &h_{lm} = a_{lm} + ib_{lm} + jc_{lm} + kd_{lm}, \;\; 1 \leq l \leq m < N \\ &h_{kk} \sim N(0,1), 1 \leq k \leq N, \end{aligned} \tag{5.2.52}$$

où chacun des éléments des matrices $\mathbf{A}$*,* $\mathbf{B}$*,* $\mathbf{C}$*,* $\mathbf{D}$*, sont donnés par,* $a_{lm}, b_{lm}, c_{lm}, d_{lm} \sim$ *i.i.d.*$N\left(0, \frac{1}{2}\right), \;\; 1 \leq l \leq m < N$

Les matrices présentées ont une densité de probabilité du type,

$$p(\mathbf{H}_N) \propto \exp\{-\beta \mathrm{tr} V(\mathbf{H}_N)\}, \tag{5.2.53}$$

où V est une fonction de $\mathbf{H}_N$. Le coefficient β dépend de la nature de la matrice, $\beta = 1$ pour le cas réel, $\beta = 2$ pour le cas complexe et $\beta = 4$ pour le cas quaternion réel. Si $V(\mathbf{H}_N) \propto \mathbf{H}_N^2$, alors, $\mathrm{tr}(\mathbf{H}_N) = \sum_{i,j} h_{ij}^2$ ce qui correspond à l'ensemble gaussien.

Proposition 5.2.2 (distribution jointe)**.** *Pour le cas d'une matrice Hermitienne GOE* ($\beta = 1$)*, GUE* ($\beta = 2$) *ou GSE* ($\beta = 4$)*, la densité jointe des valeurs propres* $\lambda_1 > \lambda_2 \cdots > \lambda_N$ *de* $\mathbf{H}_N$ *est donnée par,*

$$p_\beta(\lambda_1, \cdots, \lambda_N) = c'_{(N,\beta)} \prod_{j=1}^{N} [w_\beta(\lambda_j)]^{1/2} \prod_{1 \leq j < k \leq N} |\lambda_j - \lambda_k|^\beta, \tag{5.2.54}$$

avec

$$w_\beta(\lambda) = e^{-\beta\lambda^2}, \tag{5.2.55}$$

et $c'_{(N,\beta)}$ *est la constante de normalisation,*

$$c'_{(N,\beta)} = (2\pi)^{N/2} \beta^{N/2 + \beta N(N-1)/4} \prod_{i=1}^{N} \frac{\Gamma\left(1 + \frac{\beta}{2}\right)}{\Gamma\left(1 + \frac{\beta i}{2}\right)}. \tag{5.2.56}$$

Définition 5.2.6. *Soit* $\mathbf{X}$ *une matrice* $N \times N$ *avec pour valeurs propres* $\lambda_1 > \cdots > \lambda_N$*. La fonction de répartition empirique des valeurs propres de* $\mathbf{X}$ *est donnée par,*

$$G_N(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\lambda_i \leq x}. \tag{5.2.57}$$

Au sens des distributions, la fonction de Dirac, δ, est la dérivée de la fonction indicatrice. La densité des valeurs propres est donnée par,

$$g(x) = \frac{1}{d}\sum_{i=1}^{d}\delta(x-\lambda_i). \tag{5.2.58}$$

Théorème 5.2.1 (loi du semi-cercle de Wigner)**.** *Soit $\mathbf{H}_N$ une matrice de Wigner, complexe ou réelle, gaussienne ou non. Si les moments d'ordres 2 des éléments en dehors de la diagonale existent, la fonction de répartition empirique converge presque surement quand N tend vers l'infinie vers $G(\lambda)$ avec pour densité de probabilité,*

$$g(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & |x| \leq 1 \\ 0 & |x| > 1. \end{cases} \tag{5.2.59}$$

Cette densité forme un semi-cercle de rayon 2.

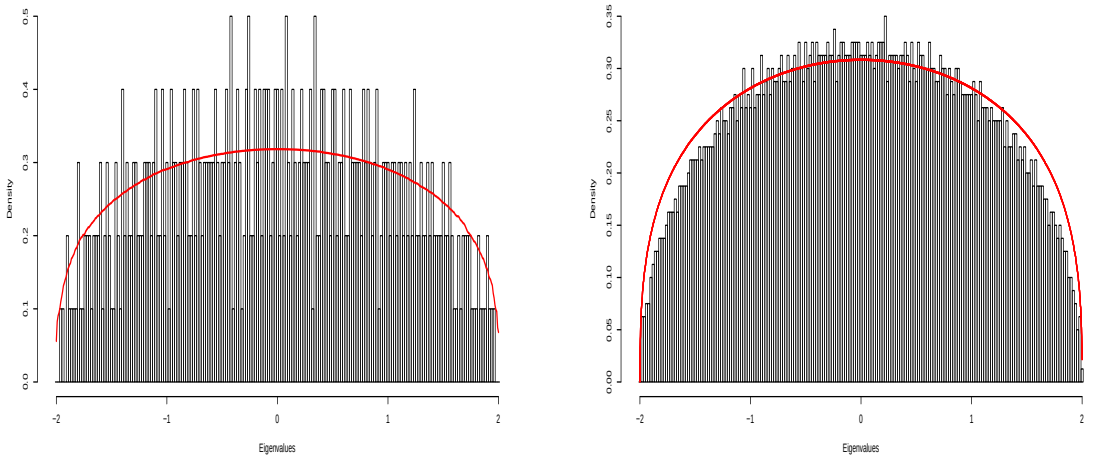


FIGURE 5.3 – Matrices aléatoires normales et le semi-cercle Wigner par dessus. A gauche 500 réalisations, à droite, 4000 réalisations.

Ce qui nous intéresse c'est d'étudier la distribution de la plus grande des valeurs propres pour la limite $N \to \infty$. En posant,

$$F_{N,\beta}(t) := \mathbb{P}_{\beta,N}(\lambda_{\max} < t), \quad \beta = 1, 2, 4, \tag{5.2.60}$$

la fonction de distribution de la plus grande valeur propre, alors, les lois limites basiques nous donnent,

$$F_\beta(x) := \lim_{N\to\infty} F_{N,\beta}\left(2\sigma\sqrt{N} + \frac{\sigma x}{N^{1/6}}\right), \ \beta = 1, 2, 4, \tag{5.2.61}$$

et sont données explicitement par,

$$F_2(x) = \exp\left(-\int_x^\infty (y-x)q^2(y)\mathrm{d}y\right), \tag{5.2.62}$$

où q est l'unique solution de l'équation de Painlevé II

$$\frac{d^2q}{qx^2} = xq + 2q^3, \tag{5.2.63}$$

satisfaisant la condition au borne,

$$q(x) \sim \mathrm{Ai}(x), \ x \to \infty. \tag{5.2.64}$$

$\mathrm{Ai}(s)$ étant la fonction spéciale de Airy,

$$\mathrm{Ai}(x) = \frac{1}{\pi}\int_0^\infty \cos(\frac{1}{3}t^3 + xt)\mathrm{d}t. \tag{5.2.65}$$

La solution de l'équation de Painlevé II est,

$$q(x) = \sqrt{-\frac{x}{2}}\left(1 + \frac{1}{8x^3} + O\left(\frac{1}{x^6}\right)\right), \ \text{si } x \to -\infty. \tag{5.2.66}$$

Pour les ensemble GOE et GSE on à,

$$\begin{aligned} F_1(x) &= \exp\left(-\frac{1}{2}\int_x^\infty q(x)d(y)\right)(F_2(x))^{1/2} \\ F_4(x/\sqrt{2}) &= \cosh\left(-\frac{1}{2}\int_x^\infty q(x)d(y)\right)(F_2(x))^{1/2}. \end{aligned} \tag{5.2.67}$$

Introduisons maintenant les deux fonctions E et F,

$$\begin{aligned} F(x) &= \exp\left(-\frac{1}{2}\int_x^\infty (y-x)q^2(y)dy\right) \\ E(x) &= \exp\left(-\frac{1}{2}\int_x^\infty q(y)dy\right). \end{aligned} \tag{5.2.68}$$

Ainsi,

$$\begin{aligned} F_1(x) &= E(x)F(x) \\ F_2(x) &= F^2(x) \\ F_4(x/\sqrt{2}) &= \frac{1}{2}\left(E(x) + \frac{1}{E(x)}\right)F(x), \end{aligned} \tag{5.2.69}$$

Alors, pour $x \to \infty$,

$$\begin{aligned} F(x) &= 1 - \frac{e^{-\frac{4}{3}x^{3/2}}}{32\pi x^{3/2}}\left(1 + O\left(\frac{1}{x^{3/2}}\right)\right) \\ E(x) &= 1 - \frac{e^{-\frac{2}{3}x^{3/2}}}{4\sqrt{\pi}x^{3/2}}\left(1 + O\left(\frac{1}{x^{3/2}}\right)\right), \end{aligned} \tag{5.2.70}$$

On peut maintenant écrire les lois asymptotiques F_β pour $x \to -\infty$,

$$\begin{aligned} F_1(x) &= \tau_1 \frac{e^{-\frac{1}{24}|x|^3 - \frac{1}{3\sqrt{2}}|x|^{3/2}}}{|x|^{1/16}}\left(1 - \frac{1}{24\sqrt{2}|x|^{3/2}} + O(|x|^{-3})\right) \\ F_2(x) &= \tau_2 \frac{e^{-\frac{1}{12}|x|^3}}{|x|^{1/8}}\left(1 + \frac{1}{2^6|x|^3} + O(|x|^{-6})\right) \\ F_3(x) &= \tau_4 \frac{e^{-\frac{1}{24}|x|^3 + \frac{1}{3\sqrt{2}}|x|^{3/2}}}{|x|^{1/16}}\left(1 - \frac{1}{24\sqrt{2}|x|^{3/2}} + O(|x|^{-3})\right), \end{aligned} \tag{5.2.71}$$

où,

$$\tau_1 = 2^{-11/48}e^{\frac{1}{2}\xi'(-1)},\ \ \tau_2 = 2^{\frac{1}{24}}e^{\xi'(-1)},\ \ \tau_4 = 2^{-35/48}e^{-\frac{1}{2}\xi'(-1)}, \tag{5.2.72}$$

et $\xi'(-1) = -0,1654211437\cdots$ est la dérivée de la fonction zeta de Riemann au point -1.

On vient de voir qu'il existe des lois limites pour caractériser les matrices de Wigner, des matrices construite par la somme d'une matrice aléatoire avec sa transposée. Nous allons maintenant nous intéresser non à la somme, mais à la multiplication. Rappelons tout d'abord que pour notre problème d'optimisation du portefeuille de Markowitz nous devons estimer la matrice de covariance, avec

maintenant $\mathbf{X}$ une matrice aléatoire de taille $N \times T$

$$\begin{aligned}\mathbf{\Sigma} &= \mathbb{E}(\mathbf{X}-\mu)(\mathbf{X}-\mu), \qquad \mu = \mathbb{E}\mathbf{X},\\ \mathbf{S} &= \frac{1}{T}(\mathbf{X}-\bar{\mathbf{X}})^\top(\mathbf{X}-\bar{\mathbf{X}})\\ &= \frac{1}{T}(\mathbf{X}(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T))^\top(\mathbf{X}(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T))\\ &= \frac{1}{T}\mathbf{X}_c^\top\mathbf{X}_c,\end{aligned} \tag{5.2.73}$$

où $\mathbf{J}_T = \mathbf{1}_T^\top\mathbf{1}_T$.

Définition 5.2.7 (matrice de Wishart)**.** *Soit $\mathbf{X}_c$ une matrice aléatoire $\mathbf{X_c} \sim N_{T,N}(0,\mathbf{\Sigma})$, alors $T\mathbf{S} = \mathbf{X}_c^\top\mathbf{X}_c$ suit la distribution de Wishart de paramètre d'échelle $\mathbf{\Sigma}$ et T degrés de libertés. Ce que l'on note,*

$$T\mathbf{S} \sim W_N(T,\mathbf{\Sigma}). \tag{5.2.74}$$

Si $\mathbf{\Sigma} = \mathbf{I}_N$, alors on dira que la matrice suit une distribution de Wishart blanche.

La densité d'une telle matrice est donnée par,

$$p(\mathbf{S}) = c_{N,T}|\mathbf{\Sigma}|^{-1/2}|\mathbf{S}|^{\frac{T-N-1}{2}}e^{-\frac{1}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{S})}, \tag{5.2.75}$$

avec,

$$c_{N,T}^{-1} = 2^{NT/2}\pi^{N(N-1)/4}\prod_{i=1}^{N}\Gamma\left(\frac{T-i+1}{2}\right). \tag{5.2.76}$$

Avant de continuer nous faisons un petit rappel sur la fonction hypergéométrique multivariée.

Définition 5.2.8 (Fonction hypergéométrique multivariée)**.** *Soit $\mathbf{X} \in \mathbb{C}^{N\times N}$, $\kappa = (k_1,\cdots,N)$ est une partition de l'entier k tel que $k_1 \geq \cdots \geq k_N \geq 0$ et $k = k_1 + \cdots + k_N$. On définit $[\cdot]_j$ par,*

$$[a]_\kappa = \prod_{i=1}^{N}(a-i+1)_{k_i}, \tag{5.2.77}$$

avec $(a)_k = a(a+1)\cdots(a+k-1)$. La fonction hypergéométrique multivariée d'une matrice complexe est donnée par,

$${}_pF_q(a_1,\cdots,a_p;b_1,\cdots,b_q;\mathbf{X}) = \sum_{k=0}^{\infty}\sum_{\kappa}\frac{[a_1]_\kappa\cdots[a_p]_\kappa}{[q_1]_\kappa\cdots[b_q]_\kappa}\frac{C_\kappa(\mathbf{X})}{k!}, \tag{5.2.78}$$

où $\sum_\kappa$ *est la somme sur toutes les partitions* κ *de* k. C_κ *est le polynome zonal complexe défini par,*

$$C_\kappa(\mathbf{X}) = \chi_{[\kappa]}(1)\chi_{[\kappa]}(\mathbf{X}), \tag{5.2.79}$$

avec $\chi_{[\kappa]}(1)$ *et* $\chi_{[\kappa]}(\mathbf{X})$ *donnée par,*

$$\chi_{[\kappa]}(1) = k! \frac{\prod_{i<j}^{N}(k_i - k_j - i + j)}{\prod_{i=1}^{N}(k_i + m - i)!}, \tag{5.2.80}$$

et,

$$\chi_{[\kappa]}(\mathbf{X}) = \frac{det(\lambda_i^{k+m-j})}{det(\lambda_i^{m-j})}. \tag{5.2.81}$$

On peut retenir en plus deux points particuliers,

$$\begin{aligned} {}_0F_0(\mathbf{X}) &= etr(\mathbf{X}) \\ {}_1F_0(a;\mathbf{X}) &= det(\mathbf{I} - \mathbf{X})^{-a}. \end{aligned} \tag{5.2.82}$$

Proposition 5.2.3 (densité jointe)**.** *Comme pour les matrices de Wigner, on peut écrire la densité jointe des valeurs propres* $\lambda_1 > \cdots, > \lambda_N$ *d'une matrice de covariance* $\mathbf{S} \sim W_N(T, \mathbf{\Sigma})$*, en posant* $t = T - 1$*,*

$$p(\lambda_1, \ldots, \lambda_N) = c_{N,t} \prod_{j=1}^{N} \lambda_j^{\frac{T-N-1}{2}} \prod_{1\le j<k\le N} |\lambda_j - \lambda_k| {}_0F_0\left(-\frac{1}{2}t\mathbf{L}, \mathbf{\Sigma}^{-1}\right), \tag{5.2.83}$$

où

$$c_{N,t} = \frac{\pi^{N^2/2}(det\mathbf{\Sigma})^{-t/2}}{\Gamma_N\left(\frac{t}{2}\right)\Gamma_N\left(\frac{N}{2}\right)} \frac{t^{-tN/2}}{2}, \tag{5.2.84}$$

Γ_p *est la fonction gamma multivariée,*

$$\Gamma_N(z) = \pi^{N(N-1)/4} \prod_{k=1}^{N} \Gamma(z - \frac{1}{2}(k-1)), \ \ \Re z > \frac{1}{2}(N-1), \tag{5.2.85}$$

enfin, ${}_0F_0$ *est la fonction hypergéométrique multivariée.*

Proposition 5.2.4. *Soit* $\mathbf{S}$ *une matrice de covariance* $N \times N$ *constuite avec des variables* $N_{T\times N}(\mu, \mathbf{\Sigma})$*. Alors, les valeurs propres* $\lambda_1 > \lambda_2 > \cdots > \lambda_N$ *de* $\mathbf{S}$ *convergent en distribution vers,*

$$\sqrt{T}(\lambda_i - \lambda_i') \ \to \ N(0, 2\lambda_i'), \ \ i = 1, \cdots, N, \tag{5.2.86}$$

où $\{\lambda_i'\}$ *sont les valeurs propres distinctes de la matrice* $\mathbf{\Sigma}$*.*

Il y a une sorte d'analogie à la loi du semi cercle de Wigner pour les matrices de covariance dans le cas d'une distribution gaussienne, la distribution de Marčenko et Pastur, la loi du quart de cercle.

Théorème 5.2.2 (Marčenko et Pastur)**.** *Le théorème de Marčenko et Pastur établit que pour une matrice de covariance* $\mathbf{S} \sim W_N(T, \mathbf{I}_N)$*, avec* $q = N/T$*,*

$$G_N \to G_{mp}(\cdot, q), \;\; p.s., \tag{5.2.87}$$

où G_N *est la fonction de répartition des valeurs propres* λ_i *de* $\mathbf{S}/T$*,*

$$G_N : x \to \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\lambda_i \leq x}, \tag{5.2.88}$$

et G_{mp} *est la fonction de répartition de Marčenko et Pastur,*

$$G_{mp}(x, q) = G_{mp}^{Dir}(x, q) + G_{mp}^{Leb}(x, q), \qquad \forall \lambda \in \mathbb{R}, \tag{5.2.89}$$

la partie de Dirac étant donnée par,

$$G_{mp}^{Dir}(x, q) : x \to \begin{cases} 1 - q & si \; x \geq 0, \;\; 0 < q < 1 \\ 0 & sinon, \end{cases} \tag{5.2.90}$$

et la partie de Lebesgue par $G_{mp}^{Leb}(\cdot, q) : x \to \int_0^{\lambda_+} g_{mp}^{Leb}(x, q)dx$ *, avec,*

$$g_{mp}^{Leb}(\cdot, q) : x \to \frac{1}{2q\pi} \frac{\sqrt{(\lambda_+ - x)^+ (x - \lambda_-)^+}}{x}, \tag{5.2.91}$$

avec, enfin,

$$\lambda_\pm = \left(1 \pm \sqrt{q}\right)^2. \tag{5.2.92}$$

Comme on le voit sur la figure (5.2.3), on parle bien cette fois-ci de quart de cercle.

Notons que λ_1 et $\lambda_{min\{T,N\}}$ convergent vers le support $[\lambda_-, \lambda_+]$,

$$\begin{aligned} \lambda_1 &\to (1 + q^{1/2})^2 \;\; \text{p.s.} \\ \lambda_{\min\{t,d\}} &\to (1 + q^{1/2})^2. \;\; \text{p.s.} \end{aligned} \tag{5.2.93}$$

et si $T < N$, alors les valeurs propres $\lambda_{t+1}, \cdots, \lambda_d$ sont égales à zéro.

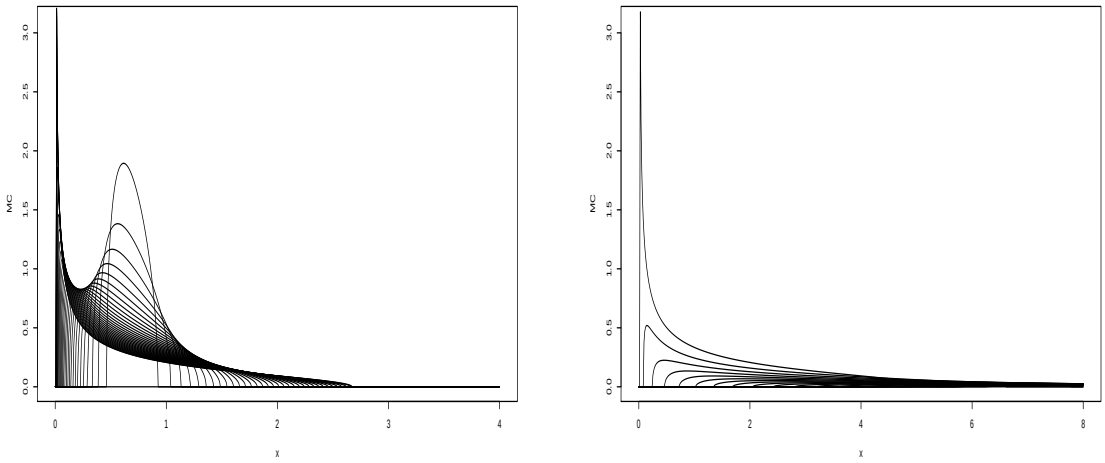


FIGURE 5.4 – Exemple de différentes densités de Marčenko et Pastur pour différentes valeurs de q. A gauche $q \leq 1$, à droite $q \geq 1$.

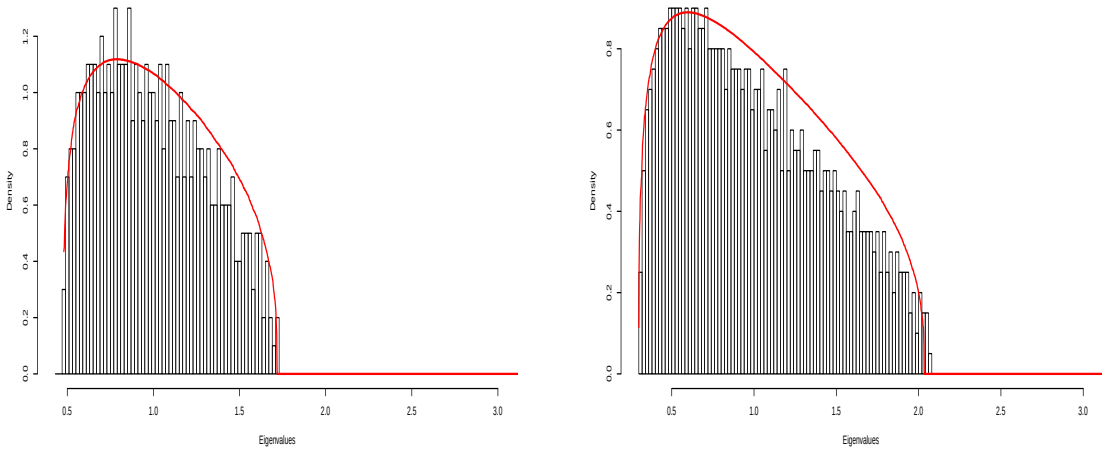


FIGURE 5.5 – Exemple de deux histogrammes issus correspondant à des matrices de covariances construitent à partir de loi normale et la densité de Marčenko et Pastur prédite.

Théorème 5.2.3 (comportement asymptotique de la plus grande valeur propre)**.** *Soit* $\mathbf{W}$ *une matrice de Wishart blanche réelle et* λ_1 *sa plus grande valeur*

propre. Alors,

$$\mathbb{P}(T\lambda_1 \leq \mu_{NT} + \sigma_{NT}x | H_0 = \mathbf{W} \sim W_N(T, \mathbf{I}_N)), \tag{5.2.94}$$

où les constantes de normalisation sont,

$$\begin{aligned} \mu_{NT} &= \left(\sqrt{T - \frac{1}{2}} + \sqrt{N - \frac{1}{2}}\right)^2 \\ \sigma_{NT} &= (\sqrt{N} + \sqrt{T}) \left(\frac{1}{\sqrt{N - \frac{1}{2}}} + \frac{1}{\sqrt{T} - \frac{1}{2}}\right)^{1/3}, \end{aligned} \tag{5.2.95}$$

et F_1 est la fonction de répartition de la loi de Tracy-Widom d'ordre 1 (5.2.67).

Nous avons montré quelques théorèmes de convergence des plus grandes valeurs propres d'une matrice de covariance. Le cas où la matrice $\mathbf{S}$ est de Wishart blanche correspond à une matrice de corrélation. En effet, rappelons que la corrélation est donnée par,

$$r_{X,Y} = \frac{\mathbb{C}\text{ov}(X, Y)}{\sigma_X \sigma_Y}. \tag{5.2.96}$$

donc pour pouvoir utiliser le théorème de Marčenko et Pastur, il suffit de réduire les séries d'actifs et construire $\mathbf{S}$ comme (5.2.73), et,

$$\begin{aligned} \pi_i^* &= \frac{\sum_{j=1}^N c_{i,j}^{-1}}{\sum_{i=1}^N \sum_{j=1}^N c_{ij}^{-1}} \\ &= \frac{\sum_{j=1}^N r_{i,j}^{-1}/\sigma_{x_j}}{\sum_{i=1}^N \sum_{j=1}^N r_{ij}^{-1}/(\sigma_{x_i}\sigma_{x_j})}. \end{aligned} \tag{5.2.97}$$

Pour résumer, si $q \to 0$ et, si les séries utilisées pour la construction de la matrice de covariance suivent des lois normales iid, alors il ne devrait pas y avoir de valeurs propres à l'extérieur de l'intervalle $[\lambda_-, \lambda_+]$. Remarquons que nous avons la décomposition suivante,

$$\text{Tr}\mathbf{C} = \sum_{k=1}^N \lambda_k = \sum_{i \in [\lambda_-, \lambda_+]} \lambda_i + \sum_{j \notin (\lambda_-, \lambda_+)} \lambda_j, \tag{5.2.98}$$

$\lambda_j \in [\lambda_1, \lambda_-) \cup (\lambda_+, \lambda_N]$ est l'ensemble des valeurs propres qui n'obéissent pas aux conditions de la théorie des matrices aléatoires. On vient de décomposer la trace en la partie bruitée et la partie contenant l'information.

Seules les valeurs propres supérieures à $\approx \lambda_+$ comportent de l'information. ' $\approx'$ car on peut chercher 'à la main' la valeur de q de (5.2.91) permettant de fitter au mieux la densité trouvée en regardant par exemple le coefficient de détermination, le R^2,

$$R^2 = \frac{\sum_i (\tilde{y}_i - \bar{y})}{\sum_i (y_i - \bar{y})}. \tag{5.2.99}$$

Néanmoins, il serait bien que la trace initiale soit conservée, on peut par exemple ajouter une valeur propre permettant de satisfaire à cette condition,

$$\lambda = \mathrm{tr}\mathbf{\Sigma} - \sum_j \lambda_j \mathbb{1}_{\lambda_j \in [\lambda_-, \lambda_+]}, \tag{5.2.100}$$

ainsi, la trace de la matrice de covariance est conservée. La dernière étape consiste donc à calculer les vecteurs propres à partir des nouvelles valeurs propres et reconstruire la matrice de covariance nettoyée du bruit.

On présente sur la figure (5.2.3) la densité des valeurs propres de la matrice de covariance construite avec 88 des composants de l'indice NYSE 100 du 02/04/2004 au 15/08/2011, 1856 points, $q = 0.047$. On voit clairement que le marché se détache des autres valeurs propres et que la plus forte concentration décrit la densité d'une matrice de correlation aléatoire.

En lisant ce cours, il peut paraître étonnant que nous ayons présenté cette théorie avec des entrées gaussiennes alors que toute la première partie nous indique le contraire que les fluctuations financières ne sont pas normales.

Il est un peu plus compliqué de se placer dans un cadre 'plus réaliste', non-gaussien. Rappelons que la distribution de Student est donnée par,

$$f(x) = \frac{1}{\sqrt{(\pi)}} \frac{\Gamma\left(\frac{1+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)} \frac{a^\mu}{(x^2 + a^2)^{\frac{1+\mu}{2}}}, \tag{5.2.101}$$

avec μ et a des paramètres positifs, pour notre problématique on suppose généralement μ entre 3 et 5 et Γ la fonction d'Euler,

$$\Gamma(t) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t. \tag{5.2.102}$$

L'estimation de la matrice de corrélation peut se faire par maximum de vraisemblance. La distribution jointe de $(x_1, \cdots, x_N)$ est

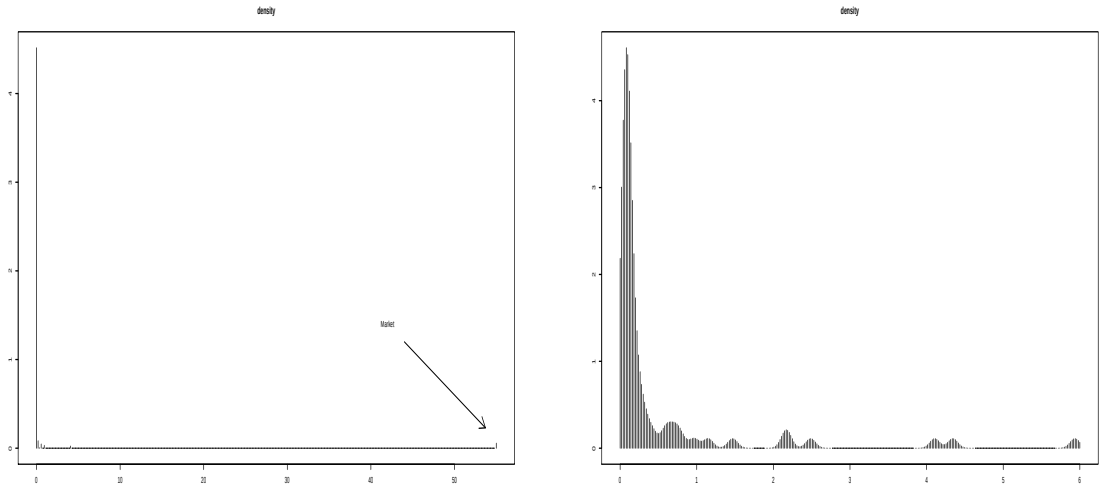


FIGURE 5.6 – Densité des valeurs propres de la matrice de covariance construite avec 88 des composants de l'indice NYSE 100 du 02/04/2004 au 15/08/2011, 1856 points. A gauche la densité totale, à droite un zoom de la partie gauche.

$$f(x_1, \cdots, x_N) = \frac{\Gamma\left(\frac{N+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)\sqrt{(\mu\pi)^N|\mathbf{S}|}} \frac{1}{\left(1+\frac{1}{\mu}\sum_{i,j} x_i(S^{-1})_{ij}x_j\right)^{\frac{N+\mu}{2}}}. \tag{5.2.103}$$

où $|A|$ dénote le déterminant de A. La log vraisemblance est donc,

$$\begin{aligned}\log L &= \log \prod_{t=1}^{T} \frac{\Gamma\left(\frac{N+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)\sqrt{(\mu\pi)^N|\mathbf{S}|}} \frac{1}{\left(1+\frac{1}{\mu}\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)\right)^{\frac{N+\mu}{2}}} \\ &\propto T\log\left(|\mathbf{S}|^{-1/2}\right) + \sum_{t=1}^{T}\log\left(\left(1+\frac{1}{\mu}\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)\right)^{-\frac{N+\mu}{2}}\right).\end{aligned} \tag{5.2.104}$$

Pour plus de clarté nous dérivons les deux termes séparement. Pour le premier terme, $\log\left(|\mathbf{S}|^{-1/2}\right)$, nous avons besoin de noter que,

$$|A^{-1}| = |A|^{-1}, \quad \frac{\partial \log |A|}{\partial A} = A^{-1}. \tag{5.2.105}$$

Alors,

$$\frac{\partial \log \left(|\mathbf{S}|^{-1/2}\right)}{\partial (S^{-1})_{ij}} = \frac{1}{2}\frac{\partial \log \left(|\mathbf{S}^{-1}|\right)}{\partial (S^{-1})_{ij}} = \frac{1}{2}S_{ij}. \tag{5.2.106}$$

Pour le second terme, $\sum_{t=1}^{T} \log \left(\left(1 + \frac{1}{\mu}\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)\right)^{-\frac{N+\mu}{2}} \right)$, sachant que,

$$\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t) = \mathrm{tr}(\mathbf{S}^{-1}\mathbf{x}\mathbf{x}^\top), \quad \frac{\partial \mathrm{tr}(AB)}{\partial A} = B^\top, \tag{5.2.107}$$

alors,

$$\frac{\partial \log \left(\left(1 + \frac{1}{\mu}\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)\right)^{-\frac{N+\mu}{2}} \right)}{\partial (S^{-1})_{ij}} = -\frac{N+\mu}{2}\frac{x_i(t)x_j(t)/\mu}{1 + \frac{1}{\mu}\sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)}. \tag{5.2.108}$$

Ainsi la condition du premier ordre est,

$$\frac{T}{2}S_{ij} - \frac{N+\mu}{2}\sum_{i,j}\frac{x_i(t)x_j(t)}{\mu + \sum_{i,j} x_i(t)(S^{-1})_{ij}x_j(t)} = 0. \tag{5.2.109}$$

L'estimateur du maximum de vraisemblance est,

$$\hat{S}_{ij} = \frac{N+\mu}{T}\sum_{t=1}^{T}\frac{x_i(t)x_j(t)}{\mu + \sum_{i,j} x_i(t)(\hat{S}^{-1})_{ij}x_j(t)}. \tag{5.2.110}$$

Comme nous le voyons, nous ne disposons pas de formule fermée pour la distribution de Student. Le calul se faisant récursivement, demandant ainsi un certain coût de calcul. Néanmoins, nous avons un résultat très intéressant, dans [BuGoWa06], nous avons la densité des valeurs propres pour le cas d'une matrice de corrélation avec des entrées suivant une loi de Student multivariée,

$$g(\lambda) = \frac{\left(\frac{\mu}{2}\right)^{\mu/2}}{2\pi q \Gamma\left(\frac{\mu}{2}\right)}\lambda^{-\alpha/2-1}\int_{\lambda_-}^{\lambda^+}\sqrt{(\lambda_+ - x)(x-\lambda_-)}e^{-\frac{\alpha x}{2\lambda}}x^{\alpha/2-1}\mathrm{d}x, \tag{5.2.111}$$

avec q, $\lambda_\pm$ défini comme précédement.

Pour conclure cette section, notons simplement que dans la partie sur la modélisation des données financières, le modèle de marche aléatoire multifractale apparaît comme un sérieux candidat aujourd'hui. Il existe un théorème du type Marčento Pastur pour cette 'loi' [AlRhVa], il donnera lieu à un projet.

5.2.4 Données Asynchrones

Dans cette partie nous n'avons pas parlé de la fréquence. Est-ce que la théorie des matrices aléatoires s'applique à toutes les fréquences ? Nous avons présenté une théorie pour des variables gaussiennes, et, dans la première partie, nous avons vu que les actifs financiers tendent vers une loi gaussienne pour des lags 'importants', un mois par exemple. Tout de même, pour une distribution de Student, nous avons un résultat très intéressant (si on suppose que la RMT est intéressante et/ou utile pour notre problématique bien sûr) de densité théorique des valeurs propres. La distribution de Student est mieux adaptée pour les données à plus hautes fréquences que mensuelles. Idem pour le cas des marches aléatoires multifractales.

Si l'on se place à une fréquence beaucoup plus élevée, lors de la formation des prix, au tick, les choses sont bien différentes. Quelque soit la fréquence, nous avons fait la supposition que nous regardions des barres et non des données tick by tick. Construire une matrice de covariance avec des données pré-samplé n'est pas cohérent, nous n'avons pas forcément de transactions à chaque minute précisément, donc quelles données prendre ? A très haute fréquence, les données arrivent ponctuellement, sans prendre nécessairement des actifs non liquides, les données étant samplé à la milli-seconde, nous ne pouvons fondamentalement pas voir les transactions arriver aux mêmes instants. Nous retrouvons ici les concepts d'effet Epps et lead-lag.

Notons $X^i(t) = \ln S^i(t)$ le log-prix de l'actif S^i à l'intant t, i=1,2. La covariance des actifs S^1 et S^2 est donnée par,

$$\mathbb{C}\text{ov} = \sum_{s'=s=1}^{T} (X^1(s) - X^1(s-1))(X^2(s') - X^2(s'-1)). \tag{5.2.112}$$

A très haute fréquence, nous n'avons donc pas nécessairement $s = s'$ et l'équa-

tion (5.2.112) n'a plus de sens. La manière la plus simple de traiter les données asynchrones est de faire une interpolation, [GeDaMuOlPi01]. Il s'agit de resynchroniser les observations.

La première idée est de prendre le dernier prix,

$$t_i = \arg\max_{t_i}\{t_i \leq t\}, i = 1, 2, \cdots, n. \tag{5.2.113}$$

Cette méthode est l'interpolation previous-tick. Nous pouvons bien sûr penser à interpoler entre les deux dates les plus proches. En supposant que t_i, est construit par (5.2.113) pour tout i,

$$S(t) = S(t_i) + \frac{t - t_i}{t_{i+1} - t_i}(S(t_{i+1}) - S(t_i)). \tag{5.2.114}$$

Cette méthode est dites d'interpolation linéaire.

Les deux méthodes sont intéressantes. L'interpolation previous-tick n'utilise que le passé, l'information disponible avant et égale à la date t, elle est donc causale. L'interpolation linéaire utilise l'information future, nous devons connaître la valeur de la transaction arrivant juste après la date t. Dans le cas d'une longue période de creux, l'interpolation previous-tick peut nous donner une valeur extrême et fausser les analyses statistiques de l'actif en question, particulièrement pour les valeurs extrêmes puisqu'il ne s'agirait que d'un artefact. L'interpolation linéaire semble donc mieux appropriée pour mener à bien des analyses de marché. Néanmoins, si l'on utilise les données dans un contexte de temps réel, comme pour une stratégie passant des ordres sur le marché et non en backtest, il est impossible d'utiliser cette méthode, uniquement le previous-tick. Nous avons présenté ces deux méthodes sur la figure (5.7)

Nous présentons maintenant deux des estimateurs les plus courament utilisés pour les données asynchrones, l'estimateur de Fourier et de Hayashi-Yoshida. Pour ces deux estimateurs nous supposons que les fluctuations suivent des processus d'Itô. Commencons par l'estimateur de Fourier [MaMa02].

Soit $(X(t) = (X_1(t), \cdots, X_d(t)))$ d les log-prix de d actifs financiers observés sur $[0, T]$. On suppose qu'ils ont pour diffusion,

$$\mathrm{d}X_i(t) = \sigma_i(t)\mathrm{d}W_i(t), i = 1, \cdots, d, \tag{5.2.115}$$

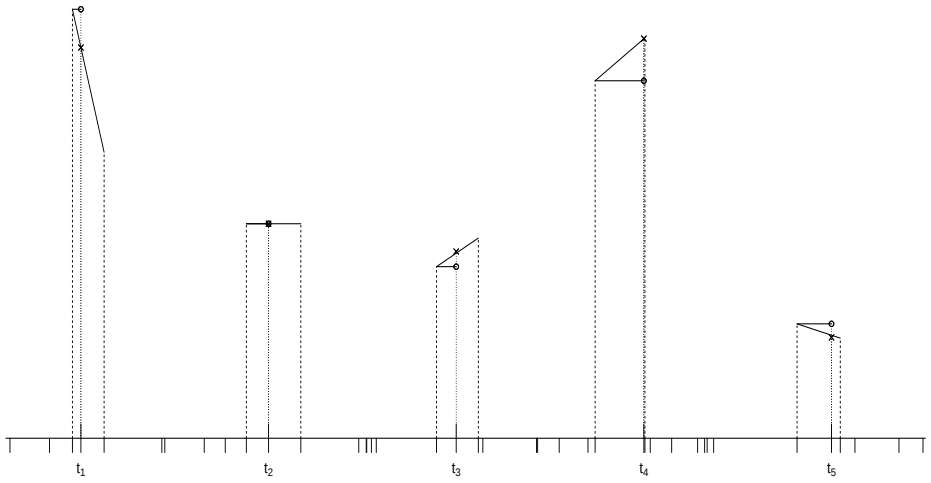


FIGURE 5.7 – Représentation des schémas d'interpolation. Les t_i sont equidistants, les autres tirets représentent les instants d'arrivées des transactions. Les points correspondent a l'interpolation previous-tick, les croix à l'interpolation linéaire.

avec σ_i la volatilité instantanée de l'actif i et W_i deux mouvements browniens de corrélation constante, $\mathrm{d} < W_i(t), W_j(t) >_t = \rho_{ij}\mathrm{d}t$, pour tout $i, j = 1, \cdots, d$. La matrice de covariance $\Sigma(t)$ a donc pour éléments, $\Sigma_{ij}(t) = \rho_{ij}\sigma_i(t)\sigma_j(t)$. L'estimation de la matrice de covariance se fait alors en utilisant l'analyse harmonique. Commençons par écrire les coefficients de Fourier associés aux rendements $\mathrm{d}X_i$,

$$\begin{aligned} a_0(\mathrm{d}X_i) &= \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}X_i \\ a_q(\mathrm{d}X_i) &= \frac{1}{\pi}\int_0^{2\pi} \cos(qt)\mathrm{d}X_i \\ a_q(\mathrm{d}X_i) &= \frac{1}{\pi}\int_0^{2\pi} \sin(qt)\mathrm{d}X_i. \end{aligned} \tag{5.2.116}$$

Il ne s'agit pas des coefficients usuels mais leurs équivalents sous forme d'intégrale stochastique. Introduisons les deux conditions,

1. Le processus de prix $X_i(t)$ est observé n fois aux instants $0 = t_0 < t_1 < \cdots < t_n = T$, non nécessairement espacé uniformément.

2. Le processus de prix $X_i(t)$ est constant et continue à gauche sur l'intervalle $[t_{k-1}, t_k]$, i.e. $X_i(t) = X_i(t_{k-1})$ pour $t \in [t_{k-1}, t_k]$ et $k = 1, \cdots, n$.

Alors en intégrant par partie,

$$a_q(\mathrm{d}X_i(t)) = \frac{1}{\pi}\int_0^{2\pi} \cos(qt)\mathrm{d}X_i = \frac{X_i(2\pi) - X_i(0)}{\pi} - \frac{q}{\pi}\int_0^{2\pi} \sin(qt)X_i(t)\mathrm{d}t. \tag{5.2.117}$$

Et,

$$\begin{aligned}\frac{q}{\pi}\int_{t_{k-1}}^{t_k} \sin(qt)X_i(t)\mathrm{d}t =& X_i(t_{k-1})\frac{q}{\pi}\int_{t_k}^{t_{k-1}} \sin(qt)\mathrm{d}t \\ =& X_i(t_{k-1})\frac{1}{\pi}(\cos(qt_k) - \cos(qt_{k-1})).\end{aligned} \tag{5.2.118}$$

En considérant que $\frac{X_i(2\pi)-X_i(0)}{\pi}$ est le drift du log-prix,

$$\tilde{X}_i(t) = X_i(t) - \frac{X_i(2\pi) - X_i(0)}{\pi}t \tag{5.2.119}$$

il est ainsi égal à zéro et,

$$\begin{aligned}a_q(\mathrm{d}X_i) &\approx \frac{1}{\pi}\sum_{k=1}^{n}[\cos(qt_k) - \cos(qt_{k-1})]X(t_{k-1}) \\ b_q(\mathrm{d}X_i) &\approx \frac{1}{\pi}\sum_{k=1}^{n}[\sin(qt_k) - \sin(qt_{k-1})]X(t_{k-1}).\end{aligned} \tag{5.2.120}$$

La transformée de Fourier inverse de la matrice de covariance est donnée par,

$$\hat{\Sigma}^{i,j}_{N,M} = a_0(\Sigma_{ij}) + \lim_{M\to\infty}\sum_{q=1}^{N}\frac{1}{2}[a_q(\mathrm{d}X_i)a_q(\mathrm{d}X_j) + b_q(\mathrm{d}X_i)b_q(\mathrm{d}X_j)], \tag{5.2.121}$$

les coefficients étant donné par,

$$\begin{aligned}a_0(\Sigma_{ij}) &= \lim_{N\to\infty}\frac{\pi}{N}\sum_{q=1}^{N}\frac{1}{2}[a_q(\mathrm{d}X_i)a_q(\mathrm{d}X_i) + b_q(\mathrm{d}X_i)b_q(\mathrm{d}X_j)] \\ a_s(\Sigma_{ij}) &= \lim_{N\to\infty}\frac{\pi}{N}\sum_{q=1}^{N-s}\frac{1}{2}[a_q(\mathrm{d}X_i)a_{q+s}(\mathrm{d}X_i) + b_q(\mathrm{d}X_i)b_{q+s}(\mathrm{d}X_j)] \\ b_s(\Sigma_{ij}) &= \lim_{N\to\infty}\frac{\pi}{N}\sum_{q=1}^{N-s}\frac{1}{2}[a_q(\mathrm{d}X_i)b_{q+s}(\mathrm{d}X_i) - b_q(\mathrm{d}X_i)a_{q+s}(\mathrm{d}X_j)].\end{aligned} \tag{5.2.122}$$

avec N la fréquence choisie, on choisira inférieur à $T/2$ (voir e.g. [GeRe04] et [MaTuIoMa11] pour un choix de la valeur). 5.2.121 étant la covariance instantanée, en integrant nous obtenons,

$$(\Sigma^{Fourier})_{ij} = \int_0^{2\pi} \hat{\Sigma}_{ij}(t)\mathrm{d}t = 2\pi a_0(\Sigma_{ij}). \tag{5.2.123}$$

Il s'en suit que la matrice de corrélation est quant à elle estimée par,

$$(\rho^{Fourier})_{ij} = \frac{a_0(\Sigma_{ij})}{\sqrt{a_0(\Sigma_{ii})a_0(\Sigma_{jj})}}, \quad i,j = 1,\cdots,p. \tag{5.2.124}$$

L'estimateur de la matrice de correlation de Hayashi-Yoshida, [HaYo05] peut être appliqué directement sur les transactions boursières, sans qu'elles aient été redéfinies sur une même grille. Les instants de transactions sont supposés être distribués selon un processus ponctuel, $\{t_i^1 : i = 1,\cdots,n_1\}$, $\{t_j^2 : j = 1,\cdots,n_2\}$, $\{t_k^d : k = 1,\cdots,n_d\}$, $\cdots$ n_1, n_2, n_d étant donc le nombre de transactions des actifs $S^1, S^2, \cdots, S^d$ de $t_0 = 0$ jusqu'à $t_{n_1} = t_{n_2} = t_{n_d} = T$. En notant I^k l'intervalle $I^k =]t_{k-1}, t_k]$, la covariance est alors donnée par,

$$(\Sigma^{HY})_{ij} = \sum_i^{n_i}\sum_j^{n_j} \delta X_{I^i}^i \delta X_{I^j}^j \mathbb{1}_{I^i \cap I^j \neq \emptyset}, i,j = 1,\cdots,d. \tag{5.2.125}$$

Nous avons donc chaque rendement de l'actif i, disons à l'instant t_k, multiplié par tous les rendements de l'actif j compris entre $]t_{k-1}, t_k]$. L'estimateur est ainsi consistant.

L'estimateur de la matrice de corrélation est alors,

$$(\rho^{HY})_{ij} = \frac{\sum_i^{n_i}\sum_j^{n_j} \delta X_{I^i}^i \delta X_{I^j}^j \mathbb{1}_{I^i \cap I^j \neq \emptyset}}{\sqrt{\sum_i^{n_i}\left(\delta X_{I^i}^i\right)^2 \sum_j^{n_j}\left(\delta X_{I^j}^j\right)^2}}, i,j = 1,\cdots,d. \tag{5.2.126}$$

Notons enfin que dans le cas où les actifs sont des processus d'Itô,

$$\mathrm{d}S^k(t) = S^k(t)(\mu^k(t)\mathrm{d}t + \sigma^k(t)\mathrm{d}W^k(t)), \tag{5.2.127}$$

Si les processus μ^k et σ^k sont adaptés, les mouvements browniens corrélés entre eux, $\mathrm{d} < W^k, W^l >_t = \rho_{k,l}(l)\mathrm{d}t$ et les instants d'arrivées $I^k, k = 1,\cdots,d$ pouvant

être dépendant entre eux mais indépendant de S^k, $k = 1, \cdots, d$, alors il est possible que l'estimateur soit consitant et asymptotiquement normale.

Bibliographie

[AlRhVa] R. Allez, R. Rhodes et V. Vargas, *Marchenko Pastur type Theorem for Independent MRW Processes : Convergence of the Empirical Spectral Measure.* Arxiv : 1106.5891, 2011.

[An11] B. Angoshtari, *Portfolio Choice With Cointegrated Assets*, The Oxford-Man Institute, University of Oxford Working paper, OMI11.01, 34, 2011, arXiv.org.

[AuCeFrSc01] P. Auer, N. Cesa-Bianchi, Y. Freund et R. E. Schapire, *The Non-Stochastic Multi-Armed Bandit Problem.* SIAM Journal on Computing, 32 :48-77, 2001.

[BeCo88] R. Bell et T.M. Cover, *Game-Theoretic Optimal Portfolio*, Mgmt. Sci. 34, 724-733, 1988.

[BlKa97] A. Blum, A. and A. Kalai, *Universal Portfolios With and Without Transaction Costs*, in Proceedings of the Tenth Annual Conference on Computational Learning Theory, pages 309-313. ACM Press, 1997.

[BuGoWa06] Z. Burda, A. Görlich et B. Waclaw, *Spectral Properties of Empirical Covariance Matrices for Data with Power-Law Tails*, Phys. Rev. E 74, 041129, 2006, arXiv.org.

[CeLu06] N. Cesa-Bianchi et G. Lugosi, *Prediction, Learning, and Games,* Cambridge University Press, 406, 2006.

[Co91] T.M. Cover, *Universal Portfolio*, Mathematical Finance, 1 :1-29, 1991.

[CoOr96] T.M. Cover and E. Ordentlich, *Universal Portfolios With Side Information*, IEEE Transactions on Information Theory, 42 :348-363, 1996.

[EnGr87] R. Engle, et C.W. Granger, *Co-Integration and Error Correction : Representation, Estimation, and Testing*, Econometrica, Vol. 55(2), 251-276

[GeDaMuOlPi01] R. Gencay, M. Dacorogna, U. Muller, R. Olsen, O. Pictet, *An Introduction to High-Frequency Finance*, Academic Press, New-York, 383 pp, 2001.

[GeRe04] C. Genovese et R. Renò, *Modeling International Market Correlations with High Frequency Data*, Correlated Data Modelling 2004, Franco Angeli Editore, Milano, Italy (2008) 99-113.

[HaTiFr09] T. Hastie, R. Tibshirini and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference and Prediction* (Second Edition). Springer-Verlag, New York, 2009.

[HaYo05] T. Hayashi, et N. Yoshida, *On Covariance Estimation of Non-Synchronous Observed Diffusion Processes*, Bernoulli, 11, 359-379, 2005.

[He98] D. P. Helmbold et al., *On-line Portfolio Selection Using Multiplicative Updates*, Mathematical Finance, 8 :325-344, 1998.

[LaCiBoPo99] L. Laloux, P. Cizeau, J.P. Bouchaud et M. Potters, *Noise Dressing of Financial Correlation Matrices*. Phys. Rev. Lett. 83, 1467, 1999, arXiv.org.

[KiWa97] J. Kivinen et M.K. Warmuth, *Additive versus Exponentiated Gradient Updates for Linear Prediction*, Info Computation 132(1), 1-64, 1997.

[LeWo00] O. Ledoit et M. Wolf, *A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices*. Working paper, Departamento de Estadistica y Econometria, Universidad Carlos III de Madrid, 47pp, 2000.

[LeWo00] O. Ledoit et M. Wolf, *Improved Estimation of the Covariance Matrix of Stock Returns With an Application to Portfolio Selection*, Working paper, Departamento de Estadistica y Econometria, Universidad Carlos III de Madrid, 21pp, 2000.

[MaMa02] P. Malliavin et M. Mancino, *Fourier Series Method for Measurement of Multivariate Volatilities*. Fin. Stoch., 6, 2002. 49Ű61.

[MaTuIoMa11] V. Mattiussi, M. Tumminello, G. Iori et R.N. Mantegna, *Comparing Correlation Matrix Estimators Via Kullback-Leibler Divergence*, 20 pp, 2011, SSRN.

[MuPrWo08] S. Mudchanatongsuk, J.A. Primbs et W. Wong, *Optimal Pairs Trading : A Stochastic Control Approach*, 2008 American Control Conference Westin Seattle Hotel, 5, 2008.

[PoBoLa05] M. Potters, J.P. Bouchaud, L. Laloux, *Financial Application of Random Matrix Theory : Old Laces and New Pieces*, Acta Phys. Pol. B 36, 2767, 2005, arXiv.org.

[stockcharts.com] http ://stockcharts.com/school/doku.php ?id=chart_school :technical_indicators.

[StLu05] G. Stoltz et G. Lugosi, *Internal Regret in On-Line Portfolio Selection*, Machine Learning, 59 :125-159, 2005.

[Wi55] E. P. Wigner, *Characteristic Vectors of Bordered Matrices with Infinite Dimensions*, Annals of Mathematics, 62, 548-564, 1955.

Troisième partie

Appendix

Chapitre 6

Optimisation non linéaire

Définition 6.0.9 (descente de gradient)**.** *Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction C^1 à minimiser. La dérivée partielle de f s'écrit,*

$$\frac{\partial f}{\partial x} = \lim_{\delta x \to 0} \frac{f(x + \delta x) - f(x)}{\delta x}, \qquad x \in \mathbb{R}^d, \tag{6.0.1}$$

cela nous indique comment un changement de x affecte f. Si $\dfrac{\partial f}{\partial x_i} < 0$ cela signifie qu'une augmentation infinitésimale du paramètre x_i ferait baisser la fonction f. On cherche donc à changer x dans la direction du vecteur de gradient ∇f,

$$\begin{cases} x_0 \in \mathbb{R}^d \\ x_{k+1} = x_k - \alpha \nabla_k f \end{cases}, \tag{6.0.2}$$

$\alpha > 0$ est le pas de déplacement, s'il est assez petit on sera assuré d'avoir une réduction de f à chaque itération, mais si trop petit, la convergence sera très lente vers un minimum local, si trop grand, divergence possible.

On peut proposer une autre interprétation de cet algorithme, si on écrit le développement de Taylor de f à l'ordre 1 par rapport à x_k on obtient,

$$f(x) = f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2\alpha}||x - x_k||^2. \tag{6.0.3}$$

Comme on aboutit à une forme quadratique (le fait qu'il s'agisse d'un min ou d'un max sera fait en cours), on peut chercher à minimiser simplement (6.0.3),

$$\nabla_x(f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2\alpha}||x - x_k||^2) = \nabla f(x_k) + \frac{1}{\alpha}(x - x_k). \quad (6.0.4)$$

On cherche donc à résoudre,

$$\begin{aligned} \nabla f(x_k) + \frac{1}{\alpha}(x - x_k) = 0 \\ x_{k+1} = x_k - \alpha \nabla f(x_k). \end{aligned} \quad (6.0.5)$$

On retrouve bien l'algorithme de descente du gradient.

Il s'agit d'un exemple simple d'un algorithme de descente. Les idées générales d'un *algorithme de descente* sont les suivantes :

- Recherche d'une direction de descente d_k, direction dans laquelle il est possible de faire décroître f
- Recherche d'un certain α qui permet une décroissance significative

La forme usuelle est,

$$x_{k+1} = x_k + \alpha d_k. \quad (6.0.6)$$

Une fois la direction d_k déterminée, il faut s'occuper de déterminer le pas de déplacement optimal, α. Comme on cherche à minimiser f, on peut vouloir minimiser le critère le long de d_k et donc de déterminer le α_k comme solution du problème,

$$\alpha_k = \arg\min_{\alpha} f(x_k + \alpha d_k). \quad (6.0.7)$$

Il s'agit de la règle de Cauchy. La règle suivante s'appelle la règle de Curry,

$$\alpha_k = \inf\{\alpha \geq 0 : f'(x_k + \alpha d_k) = 0, f(x_k + \alpha d_k) < f(x_k)\}. \quad (6.0.8)$$

Dans ce cas, α_k est donc choisi parmi les points stationnaires.

Ces deux manières de déterminer le pas optimal α_k sont dites de *recherche exacte*. Prenons le cas d'une fonction quadratique, i.e. $\nabla f(x) = Ax + b$, nous avons simplement,

$$d_k^\top (Ax_k + b + \alpha_k A d_k) = d_k^\top \nabla f(x_k) + \alpha_k d_k^\top A(d_k) = 0, \quad (6.0.9)$$

si A est définie positive, et donc que la fonction est convexe, la valeur du pas de déplacement est simplement donnée par,

$$\alpha_k = -\frac{d_k^\top \nabla f(x_k)}{||d_k||_A^2}. \tag{6.0.10}$$

Si la fonction à minimiser n'a pas toutes les 'bonnes propriétés' requises, alors il se peut que les règles (et donc les pas fournis) de Cauchy ou Curry n'existent pas. De plus, la détermination exacte de α_k (au sens de Cauchy ou Curry) peut s'avérer très coûteuse en temps de calcul, ce qui n'est pas forcément compensé par le gain de temps dans la minimisation de f. On va naturellement essayer de construire des algorithmes peu gourmands approchant au moins l'une des deux équations (6.0.7), (6.0.8).

Pour essayer d'approcher la condition de Cauchy on peut rechercher α_k tel que,

$$f(x_k + \alpha_k d_k) \leq f(x_k) + w_1 \alpha_k \nabla f_(x_k) d_k^\top, \tag{6.0.11}$$

où $w_1 \in]0,1[$. Il s'agit de la condition d'Armijo. Pour arriver à trouver un pas de déplacement qui vérifie l'équation précédente on applique l'algorithme suivant,

Algorithme : Recherche Linéaire approchée - Armijo

Initialisation : $\alpha_k^1 > 0$, $\tau \in]0,1[$

Tant que (6.0.11) n'est pas vérifiée i.e.,

$$f(x_k + \alpha_k^i d_k) > f(x_k) + w_1 \alpha_k^i \nabla f(x_x) d_k^\top.$$

(1) On prend

$$\alpha_k^{i+1} \in [\tau \alpha_k^i, (1-\tau)\alpha_k^i].$$

(2) i=i+1

Le pas déterminé par cet algorithme a la fâcheuse tendance à être trop petit et donné une convergence trop lente. La règle de Goldstein est,

$$f(x_k + \alpha_k d_k) \geq f(x_k) + w_1' \alpha_k \nabla f(x_k) d_k^\top. \tag{6.0.12}$$

où $w_1' \in [w_1, 1]$. Cette règle permet de ne pas choisir de α_k trop petit. Les conditions de Wolfe permettent également d'avoir un pas de déplacement pas trop "petit" et vérifient la condition de décroissance linéaire (6.0.11),

$$\begin{aligned} f(x_k + \alpha_k d_k) &\leq f(x_k) + w_1 \alpha_k \nabla f(x_k) d_k^\top. \\ \nabla f(x_k + \alpha_k d_k) &\geq w_2 g_k d_k^\top \end{aligned} \tag{6.0.13}$$

Pour déterminer α vérifiant les conditions de Wolfe, on peut appliquer l'algorithme de Fletcher-Lemaréchal

Algorithme : Recherche Linéaire approchée - Fletcher-Lemaréchal

Initialisation : $\alpha_k^1 = 0$, Soient $\underline{\alpha}^1$=0, $\overline{\alpha}^1 = \infty$, $\tau \in]0, \frac{1}{2}[$ et $\tau_e > 1$

(1) Si

$$f(x_k + \alpha_k^i d_k) \leq f(x_k) + w_1 \alpha_k^i \nabla f(x_k) d_k^\top.$$

n'est pas vérifiée, alors $\overline{\alpha}^i = \alpha^i$ et,

$$\alpha^i \in [(1-\tau)\underline{\alpha}^i + \tau\overline{\alpha}^i, \tau\underline{\alpha}^i + (1-\tau)\overline{\alpha}^i].$$

(2) Sinon,

(2.1) Si

$$\nabla f(x_k + \alpha_k^i d_k) \geq w_2 g_k d_k^\top.$$

est vérifiée, alors stop.

(2.2) Sinon, $\underline{\alpha}^i = \alpha^i$.

(2.3) Si $\overline{\alpha}^i = +\infty$, alors prendre $\alpha^i \in [\tau_e \underline{\alpha}^i, \infty]$, sinon,

$$\alpha^i \in [(1-\tau)\underline{\alpha}^i + \tau\overline{\alpha}^i, \tau\underline{\alpha}^i + (1-\tau)\overline{\alpha}^i].$$

Précédemment nous avions montré que d_k pouvait être égale à $-\nabla f(x_k)$, si nous faisons maintenant le développement de Taylor de f au second ordre (on doit donc avoir une fonction f de classe C^2) nous obtenons,

$$f(x) = f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^\top Hf(x_k)(x - x_k), \tag{6.0.14}$$

où $Hf(x)$ est le hessien de f au point x. Les conditions d'optimalité sont,

$$\nabla f(x_k) + Hf(x_k)(x_{k+1} - x_k) = 0. \tag{6.0.15}$$

Et donc, la mise à jour est donnée par,

$$x_{k+1} = x_k - Hf(x_k)^{-1} \nabla f(x_k). \tag{6.0.16}$$

Il s'agit de la méthode de Newton. Le problème ici est le temps de calcul de la matrice hessienne Hf, on essaye d'approximer Hf, ou directement $H^{-1}f$,

$$x_{k+1} = x_k - \alpha B_k \nabla f(x_k), \tag{6.0.17}$$

où B_k est l'approximation de $H^{-1}f$ (on parle alors de quasi Newton).

Pour toute la suite on pose $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ et $s_k = x_{k+1} - x_k$. L'approximation de Broyden est donnée par,

$$B_{k+1} = B_k + \frac{(s_k - B_k y_k)(s_k - B_k y_k)^\top}{(s_k - B_k y_k)^\top y_k}. \tag{6.0.18}$$

Algorithme : Quasi Newton - Broyden

Initialisation : $\alpha_0 = 1$, $H_0 = I$
pour $k = 1, 2, \cdots$

(1) Calcul de la direction de descente, $d_k = -B_k \nabla f(x_k)$
(2) On détermine α_k en appliquant une recherche linéaire de Wolfe.
(3) On calcule
$$B_{k+1} = B_k + \frac{(s_k - B_k y_k)(s_k - B_k y_k)^\top}{(s_k - B_k y_k)^\top y_k}.$$
(4) Stop si $||\nabla f(x_k)|| \leq \epsilon$, alors $x^* = x_k$, sinon, retour à l'étape 1.

Le problème ici est que les B_k ne sont pas forcément définis positifs. La mise à jour de Davidson, Fletcher et Powell permet de pallier à ce problème, elle est donnée par,

$$B_{k+1} = B_k + \frac{s_k s_k^\top}{s_k y_k} - \frac{B_k y_k y_k^\top B_k}{y_k^\top B_k y_k}. \tag{6.0.19}$$

Nous présentons enfin la méthode de Broyden, Fletcher, Goldfard et Shanno (BFGS),

On peut montrer dans ce cas que la convergence est superlinéaire, i.e. que,

$$\lim_{\tau \to \infty} \frac{||\mathbf{w}_{k+1} - \mathbf{w}^*||}{||\mathbf{w}_k - \mathbf{w}^*||} = 0. \tag{6.0.20}$$

Algorithme : Quasi Newton - BFGS

Initialisation : $\alpha_0 = 1$, $H_0 = I$
pour $k = 1, 2, \cdots$

(1) Calcule de la direction de descente, $d_k = -H_k^{-1} \nabla f(x_k)$
(2) On détermine α_k en appliquant une recherche linéaire de Wolfe.
(3) On calcul

$$B_{k+1} = B_k - \frac{s_k y_k^\top B_k + B_k y_k s_k^\top}{y_k \top s_k} + \left(1 + \frac{y_k^\top B_k y_k}{y_k^\top}\right) \frac{s_k s_k^\top}{y_k^\top s_k}.$$

(4) Stop si $||\nabla f(x_k)|| \leq \epsilon$, alors $x^* = x_k$, sinon, retour à l'étape 1.

Notons qu'il existe une version modifiée de l'algorithme BFGS ne nécessitant pas de calculer explicitement Hf et donc d'avoir à stocker un trop grand nombre de données, il s'agit de l'algorithme Low Memory BFGS.

Chapitre 7

Introduction à R

7.1 Premiers Pas (ou pas !)

R est un logiciel libre sous la licence GNU GPL disponible sur la plupart des OS. Il s'agit d'un langage de haut niveau principalement utilisé pour les statistiques, mais on peut facilement, comme on le verra, coder tout type de fonction mathématique.

La première étape consiste bien sûr au téléchargement du logiciel, il suffit de se rendre sur : http ://cran.r-project.org/, choisir son OS et, se laisser guider pour l'installation. Une fois la procédure terminée et R lancé, nous pouvons avoir besoin d'installer des 'packages', rien de plus simple ! Il suffit de cliquer sur l'onglet 'Packages', sélectionner 'installer le(s) package(s)', une liste de site miroir est proposée, on pourra par exemple choisir Lyon 1 puis, sélectionner le package désiré. Prenons par exemple RExcelInstaller, on voit que l'installation a nécessité l'installation des packages rcom et rscproxy. Une autre procédure possible pour installer un package aurait été de se rendre sur la page web de la liste des packages disponibles, de sélectionner RExcelInstaller, de le charger au format zip, et de l'installer à partir de R en sélectionnant 'installer à partir du fichier zip'. Mais cela aurait été plus fastidieux puisque nous aurions dû répéter les mêmes étapes pour les dépendances rcom et rcsproxy. Pour charger le package dans la session ouverte, il suffit de taper dans la console, library(RExcelInstaller).

On pourra manipuler des objets : vecteurs, matrices, tableaux de données, listes, etc. Commençons par définir un vecteur :

```
x=c(1,4)
```

Deux possibilités, soit on tape la ligne de code précédente directement dans la console et la ligne est compilée, soit on la tape dans un script, et pour la compiler il suffit de sélectionner la ligne et ctrl + R. Si on tape 'x' dans la console, on obtient naturellement le vecteur '1 2'. Pour vérifier à quelle classe appartient l'objet 'x', on tape CLASS(x), le résultat sera NUMERIC. Pour modifier un élément du vecteur x, disons le premier :

```
> x[1]=2
> x
[1]  2  4
```

Pour multiplier deux vecteurs entre eux, on pourra utiliser les opérateurs '*' ou '%*%' :

```
x=c(1,4)
y=c(2,3)
> x*y
[1]  2 12
> x%*%y
[1]    14
```

On a donc bien compris que l'opérateur '*' multiplie élément par élément et que '%*%' effectue un produit scalaire, de même, l'addition, '+', la soustraction '-' et la division '/', se font élément par élément :

```
> x/2
[1] 0.5 2.0
> x/y
[1] 0.500000 1.333333
```

Les mêmes règles s'appliquent aux matrices,

```
x=matrix(c(1,-3,4,2), ncol=2, nrow=2)
y=matrix(c(-2,3,1,5), ncol=2, nrow=2)
```

```
> x                 |    > y                 |    > x+y
     [,1] [,2]      |         [,1] [,2]      |         [,1] [,2]
[1,]    1    4      |    [1,]   -2    1      |    [1,]   -1    5
[2,]   -3    2      |    [2,]    3    5      |    [2,]    0    7

> x*y               |    > x%*%y             |    > x/y
     [,1] [,2]      |         [,1] [,2]      |         [,1] [,2]
[1,]   -2    4      |    [1,]   10   21      |    [1,] -0.5  4.0
[2,]   -9   10      |    [2,]   12    7      |    [2,] -1.0  0.4
```

Au passage, on vient d'écraser l'objet 'x' initialement défini comme le vecteur (1;4), pour afficher la liste d'objets en mémoire on tape LS() et pour effacer l'objet 'x', RM(x). Il existe d'autres opérateurs matriciels, par exemple :

```
> t(x) # transposée | > solve(x) #inversion      |   > det(x)
     [,1] [,2]      |      [,1] [,2]             |   [1] 14
[1,]    1   -3      | [1,] 0.1428571 -0.28571429 |
[2,]    4    2      | [2,] 0.2142857  0.07142857 |

v=eigen(x) # valeurs et vecteurs propres
> v$values
[1] 1.5+3.427827i 1.5-3.427827i
> v$vectors
                        [,1]                    [,2]
[1,] 0.7559289+0.0000000i 0.7559289+0.0000000i
[2,] 0.0944911+0.6477985i 0.0944911-0.6477985i
```

Si on souhaite importer un fichier de données, deux solutions, soit à l'aide du package RCOM installé et chargé précédemment, il suffit alors sur un fichier excel d'aller dans -> compléments -> RExcel -> connect R (ou start R si R n'est pas encore ouvert ou que le package n'a pas été chargé. Pour obtenir de l'aide sur le package rcom, ou n'importe quel package 'machin', on tape '?machin'.) puis de sélectionner la plage de données souhaitées et Put R var ou R DataFrame selon. L'autre moyen d'importer un jeu de données est de taper,

```
data=read.csv("C:/chemin/fichier.csv",header=TRUE,sep=";",dec=".")
```

Détaillons un peu la fonction READ.CSV, l'argument 'header', défini à TRUE ou FALSE signifie que la première ligne du fichier comporte les noms des colonnes, 'sep' notifie quel est le symbole qui sépare les colonnes (usuellement un ';' pour un csv) et 'dec' pour le symbole des décimales, si français une virgule, si anglosaxon, un point. Pour afficher les autres arguments disponibles de la fonction READ.CSV, ou tout autre, soit encore une fois on affiche l'aide ?READ.CSV, soit on tape simplement le nom de la fonction dans la console, mais il faut déjà avoir une idée des arguments et de leurs rôles pour que cela soit utile. Si le fichier .csv (R n'arrivera pas à importer un .xls, .xlsx, etc., inutile de s'acharner) ne contient pas de nom, on aura donc renseigné header comme FALSE et pour pouvoir en rajouter,

```
colnames(data)[1]=''Date''
colnames(data)[2]=''Open''
etc.
```

pour définir un élément comme un character nous l'avons donc entouré de " ". Sans grande surprise, 'rownames' permet de mettre des noms aux lignes. Il est possible que l'affectation d'un nom à une colonne soit laborieux, une solution rapide est de changer le type de 'data' en data.frame,

```
data=as.data.frame(data)
```

Passons maintenant à des fonctions un peu plus sophistiquées proposées par R. Supposons que le fichier de données chargé précédemment est l'historique d'une action contenant 6 colonnes, date, open, high, low, close et volume et que l'on souhaite calculer la moyenne du close, on pourra avoir recourt à une boucle :

```
close=data[,''Close'']
somme=0
for ( t in 1:length(close) ) {
  somme=somme+close[t]
}
moyenne=somme/length(close)
```

Plusieurs remarques, 'data' est une matrice, ou une data frame mais peu importe, l'objet comporte donc des lignes et des colonnes -> 'data[ligne, colonne]', comme on souhaite faire la moyenne sur le close et que la colonne comportant le close s'appelle 'Close', on cherche 'data[,"Close"]' (étonnant non ?) comme il s'agit de la 5ème colonne, 'data[,5]' revenait au même. La fonction LENGTH(x) nous donne le nombre d'éléments de 'x' (DIM(X) nous renvoit les dimensions de la matrice X), la boucle 'for' va donc faire varier t sur toute la série de données, l'opérateur ' :' permet de construire la suite de nombre entier partant du terme à sa gauche jusqu'au terme à sa droite, on aurait pu mettre 'c(a :b)' mais cela revient au même que 'a :b'. Nous pouvons aussi aller bien plus rapidement,

```
close=data[,''Close'']                | close=data[,''Close'']
moyenne=sum(close)/length(close)      | moyenne=mean(close)
```

Nous voulons maintenant calculer la moyenne de l'open, du high, du low et du close, comment faire rapidement ? On fait du calcul parallèle !! Pas à proprement parler car cela serait un peu compliqué à expliquer en quelques lignes (qui voudra pourra regarder le package SNOWFALL) mais on peut déjà se servir de la fonction APPLY de R. Avant de faire ce petit exemple, nous allons introduire la notion de fonction pour faire quelque chose de plus clair. Pour définir une fonction en R il suffit de taper :

```
mafonction=function(arguments){
liste commandes
return(résultat)
}
```

Revenons à notre problème de moyenne, nous écrivons tout sous forme de fonction, nous verrons pourquoi après, la solution naïve serait de faire simplement,

```
moyenne_une_boucle=function(table, sur_quoi){
  moyenne=c() #on définit le type, ici un vecteur
  j=0
  for ( i in sur_quoi ){
    j=j+1
    moyenne[j]=mean(table[,i])
  }
  return(moyenne)
```

```
}
```

ou pire encore :

```
moyenne_deux_boucles=function(table, sur_quoi){
  moyenne=c()
  j=0
  for ( i in sur_quoi ){
    j=j+1
    somme=0
    for ( t in 1:length(table[,i]) ) {
      somme=somme+table[t,i]
    }
    moyenne[j]=somme/length(table[,i])
  }
  return(moyenne)
}
```

On a ici deux boucles imbriquées, c'est exactement ce qu'il faut éviter à tout prix pour que les temps de calculs n'explosent pas ! Une solution propre est :

```
moyenne_pas_de_boucle=function(table, sur_quoi){
  moyenne=apply(table[,sur_quoi], 2 ,mean)
  return(moyenne)
}
```

Nous allons comparer les temps d'éxecution, le fichier utilisé comporte 30446 lignes,

```
> colonnes=c(''Open'', ''High, ''Low'', Close'')

> system.time(moyenne_une_boucle(table=data, sur_quoi=colonnes))
utilisateur      système       écoulé
       0.02         0.00         0.02
> system.time(moyenne_deux_boucles(table=data, sur_quoi=colonnes))
utilisateur      système       écoulé
       6.99         0.00         6.99
> system.time(moyenne_pas_de_boucle(table=data, sur_quoi=colonnes))
utilisateur      système       écoulé
       0.01         0.00         0.01
```

On vient seulement de montrer le temps de calcul nécessaire pour les trois fonctions, comme on peut le voir, avoir imbriqué deux fonctions fait complétement exploser les temps de calculs. Pour avoir le résultat, c'est quand même notre premier objectif,

```
> colonnes=c(''Open'', ''High, ''Low'', Close'')
> moyenne_une_boucle(table=data, sur_quoi=colonnes)
```

```
[1] 1186.041 1186.806 1185.253 1186.038
```

on obtient le même résultat quelle que soit la fonction utilisée, seuls les temps de calculs changent. Notons que pour définir des arguments par défauts, par exemple si dans la majorité des cas on souhaite effectuer la moyenne uniquement sur le close,

```
moyenne_pas_de_boucle=function(table, sur_quoi=''Close'')){
  moyenne=apply(table[,sur_quoi], 2 ,mean)
  return(moyenne)
}
```

dans ce cas, inutile de renseigner sur quelle colonne vous souhaitez effectuer le calcul. Pour (presque) terminer cette présentation du logiciel R nous allons regarder la fonction LM car un bon nombre de fonctions R ont la même structure. Rappelons qu'une régression linéaire multiple est,

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots, \beta_N x_{N,t}, \qquad t = 1, \cdots, T \tag{7.1.1}$$

ou sous forme matricielle,

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} \tag{7.1.2}$$

où $\mathbf{X}$ est la matrice $T \times (N+1)$, la première colonne étant constituée uniquement de 1 pour faire apparaître le coefficient β_0, l'intercept. La solution par minimisation de l'erreur quadratique, après quelques calculs,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{7.1.3}$$

Si on code tout à la main on a,

```
> Y=data[,''Close'']
> var_exp=c(''Open'', ''High'', ''Low'')
> X=cbind(rep(1, dim(data)[1]),data[,var_exp])
> beta=solve(t(X)%*%X)%*%t(X)%*%Y
```

La fonction LM de R nous permet de faire la même chose, les détails statistiques en plus,

```
> fit_simple=lm(data[,''Close'']~data[,''Open''])
> fit_multiple=lm(data[,''Close'']~data[,''Open'',''High'', ''Low''])
```

Attention à l'exemple de régression multiple proposé, dans la 'vraie vie', si vous faites une régression linéaire pour essayer de prévoir le close, vous pouvez connaître l'open selon à quel moment vous souhaitez prendre un pari, mais vous ne connaissez ni le plus bas ni le plus haut de la journée avant la fermeture, il s'agit juste d'un exemple ! la fonction LM nous retourne un ensemble de valeurs, premièrement, les

betas, ensuite, en tapant SUMMARY(fit) on obtient un ensemble de statistiques propres à la régression, à vous de tester ! Pour tracer les valeurs reconstruites par le modèle de régression linéaire sur l'apprentissage on utilise la fonction PLOT,

```
> valeurs=fit_multiple$fitted.values
> plot(data[,''Close''], type='l', ylab=''machin'', xlab=''truc'',
> main=''Cours à la fermeture'')
> lines(y=valeurs, type='p', col='red')
```

l'objet résultant de la régression linéaire, fit_multiple, comporte plusieurs éléments comme les coefficients, le R2, etc. nous, nous voulons les valeurs reconstruites, pour extraire uniquement cette information on utilise l'opérateur '$' suivi de ce qui nous intéresse,

```
> coef=fit$coefficients
```

Pour la fonction PLOT nous avons écrit directement le résultat sans se soucier de l'axe des abscisses, pour écrire la date on aurait écrit,

```
> plot(data[,''Date''], data[,''Close''], type='l', ylab=''machin'', ...)
```

par défaut R comprend que le premier argument de la fonction PLOT est x et le deuxième y. Enfin, pour faire une prévision du close à partir du modèle construit,

```
prev=predict(fit_multiple, newdata)
```

bien sûr, newdata doit avoir trois colonnes comme lors de la construction.

Par exemple, le perceptron à deux couches NNET du package éponyme et l'arbre CART, RPART, toujours issu du package éponyme fonctionnent de la même manière.

Il existe énormément de packages proposant un tas de fonctions, néanmoins, il est quand même parfois utile de vraiment savoir coder proprement et de ne pas se reposer uniquement sur l'existant, par exemple le perceptron NNET ne permet d'avoir qu'une seule couche cachée, mais comme on l'a déjà dit précédemment, le théorème de Cybenko nous assure que nous n'avons pas besoin de plus de couches, en revanche, on pourrait peut être avoir envie de définir une autre fonction de minimisation que l'erreur quadratique pénalisée, et le package ne nous permet pas cela, on pourra donc avoir besoin de développer nous-même. En revanche, il est vivement déconseillé de vouloir coder nous-même un produit matriciel par exemple, pour les plus septiques, je vais essayer d'expliquer pourquoi. Bien sûr que si l'on utilise l'opérateur '%*%', il y a des boucles derrière, mais les 'personnes' qui l'ont codé sont bien meilleures que nous ! Faisons des tests ! On va se contenter de regarder le produit scalaire de deux matrices carrées A et B de taille $N \times N$, le résultat est $C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj}$. Le code le plus naïf est,

```
prod1=function(A,B){
N=dim(A)[1]
C=matrix(nrow=N, ncol=N)
for (i in 1:N ) {
      for (j in 1:N ) {
           C[i,j]=0
           for (k in 1:N ) {
                C[i,j]=C[i,j]+A[i,k]*B[k,j]
           }
     }
}
return(C)
}
```

Le problème ici est la gestion de la mémoire, on préfèrera,

```
prod2=function(A,B){
N=dim(A)[1]
C=matrix(nrow=N, ncol=N)
for (i in 1:N ) {
      for (j in 1:N ) {
           tmp=0
           for (k in 1:N ) {
                tmp=tmp+A[i,k]*B[k,j]
           }
     C[i,j]=tmp
     }
}
return(C)
}
```

Enfin, la dernière idée que nous allons tester est qu'au lieu de lire de la première ligne à la dernière, idem pour les colonnes, nous allons faire le chemin inverse, de la dernière à la première,

```
prod3=function(A,B){
N=dim(A)[1]
C=matrix(nrow=N, ncol=N)
for (i in N:1 ) {
      for (j in N:1 ) {
           tmp=0
           for (k in N:1 ) {
                tmp=tmp+A[i,k]*B[k,j]
           }
     C[i,j]=tmp
     }
}
```

```
return(C)
}
```

Pour faire le test nous construisons deux matrices aléatoires de taille 100 par 100

```
P=100
A=matrix(rnorm(P*P, 0, 1), ncol=P, nrow=P)
B=matrix(rnorm(P*P, 0, 1), ncol=P, nrow=P)
```

Bien sûr, nous allons comparer par rapport au produit matriciel de R :

```
> system.time(prod1(A,B))
utilisateur      système       écoulé
       4.11         0.00         4.11
> system.time(prod2(A,B))
utilisateur      système       écoulé
       2.03         0.00         2.03
> system.time(prod3(A,B))
utilisateur      système       écoulé
       1.99         0.00         2.00
> system.time(A%*%B)
utilisateur      système       écoulé
          0            0            0
```

Les résultats sont donc sans appel ! Comment expliquer cela, un peu compliqué à vrai dire. Ce que vous devez retenir de ces exemples, utiliser pleinement le calcul matriciel de R ! Eviter à tout prix d'imbriquer des boucles ! Ne pas aller de 1 à N mais plutôt de N à 1 !

Sur les packages, être bien critique et vérifier que la fonction proposée correspond bien à votre problématique, ne pas hésiter à faire votre code si nécessaire. Et pour information, le produit matriciel de R n'est en fait pas codé en R mais dans un langage de bien plus bas niveau. L'algorithme utilisé n'est pas celui que nous avons appris durant nos premières années d'études et utilisé plus haut, il est de complexité bien plus basse (on pourra par exemple travailler non pas avec A et B mais plutôt avec la transposée pour diminuer les chemins, travailler par blocs et non par éléments ($O(n^{2.807})$) au lieu de n^3 pour le produit naïf), etc.

Nous concluons (pour de vraie !) cette introduction par un tableau récapitulatif de quelques fonctions R qui, espérons le, vous sera utile.

Général	
ls()	affiche la liste des objets en mémoire
rm(x)	supprime l'élément x
save.image()	sauvegarde une image de la session courante
library(machin)	charge le package machin
?machin	affiche l'aide du package machin
<; >; <=; >=; ==; !=	opérateur de comparaison
x<1 & x>=0	$0 \leq x < 1$
x&y (x|y)	ET (OU) logique opérant sur tous les éléments de x et y
x&&y (x||y)	ET (OU) logique opérant sur le premier élément de x et y

On présente un code R pour l'algorithme de descente du gradient à pas constant pour optimiser la fonction $f(x_1, x_2) = x_1^2 + x_2$,

```
##############################
#                            #
#  On choisit une fonction   #
#                            #
##############################

f=function(x){
f=x[1]^2+x[2]
return(f)
}

##############################
#                            #
#  On calcule le gradient    #
#                            #
##############################

grad=function(g, x, delta){
y=x
der=c()
for ( i in 1:length(x)){
      y[i]=x[i]+delta
      der[i]=(g(y)-g(x))/delta
      y=x
}
return(der)
}
```

Vecteur	
x=c(1,2,3)	vecteur (1,2,3)
x=1 :10	suite d'entier de 1 à 10
x=rep(1,10)	répète 10 fois le chiffre 1
x=rep(1 :2, 10)	répète 10 fois la séquence 1, 2
x=seq(1, 10, by=0.1)	séquence allant de 1 à 10 avec un pas de 0.1
x=seq(1, 10, length=100)	séquence allant de 1 à 10 de taille 100
%/%	division entière
%%	congrue
x=c("a", "b")	vecteur contenant les caractères a et b
x=c(y,z)	vecteur qui concatène les vecteurs y et z
x[i]	ième élément du vecteur x
x[-i]	supprime le ième élément de x
x[x>1]	uniquement les éléments de x supérieur à 1
unique(x)	donne les éléments uniques du vecteur x
round(x, p)	arrondi les éléments de x à p chiffres après la virgule
floor(x) (ceiling(x))	entier le plus bas (haut)
rev(x)	inverse l'ordre des élements du vecteur x
sort(x)	tri croissant des éléments de x
length(x)	taille du vecteur x
sum(x)	somme des éléments de x
prod(x)	produit des éléments de x
cumsum(x)	somme cumulée des éléments de x
cumprod(x)	produit cumulé des éléments de x
max(x) (min(x))	max (min) des éléments de x
which.max(x) (which.min(x))	position du maximum (minimum) du vecteur x
which(x==p)	donne la position des éléments de x égaux à p
sin(x) (cos(x))	sinus (cosinus) de tous les éléments de x
exp(x)	exponentielle de tous les éléments de x
log(x, a)	logarithme de tous les éléments de x en base a

```
##############################
#                            #
#  Algorithme de descente    #
#                            #
##############################
```

Matrice	
x=matrix(c(1,...), ncol=, nrow=)	matrice
x[i,j]	élément de la matrice à la ième ligne et jème colonne
x[,-j]	matrice x sans la jème colonne
cbind(x,y)	concatène les matrices x et y par colonnes
rbind(x,y)	concatène les matrices x et y par lignes
apply(x, 1, machin)	applique la fonction machin à toutes les colonnes de x
apply(x, 2, machin)	applique la fonction machin à toutes les lignes de x
x*y	produit de x par y élément par élément
x%*%y	produit scalaire de x y
as.matrix(x)	change le format de x en matrice, pareil pour .vector, .numeric, etc.
is.matrix(x)	renseigne si x est une matrice ou non, pareil pour .vector, etc.
diag(n)	matrice identité $n \times n$
diag(1 :n)	matrice diagonale avec le vecteur 1 :n sur la diagonale
t(x)	transposée de x
det(x)	déterminant de x
eigen(x)	valeurs et vecteurs propres de x
solve(x)	inverse de x
dim(x)	dimension de la matrice x
qr(x)	décomposition QR de x
chol(x)	décomposition de Cholesky de x
svd(x)	décomposition en valeur singulière de x

```
xk=rep(rnorm(1),2)
nabla=grad(f, xk, 0.0001)
alpha=0.001

i=1
while (sqrt(sum(nabla^2))>0.0001){
nabla=grad(f, xk, 0.0001)
xk=xk-alpha*nabla
i=i+1
if ( i==10000) break
}
```

Statistique	
mean(x)	moyenne de x
median(x)	médianne de x
quantile(x, c(0.1,0.2))	quantile à 10 et 20% de x
sd(x)	écart type de x
corr(x,y)	corrélation de x et y
cov(x,y)	covariance de x et y
sample(x, n, replace=TRUE)	tirage aléatoire de x nombre du vecteur n avec remise, sans si replace=FALSE.
rLOI(n, paramètres)	génére n variables aléatoires de la LOI (norm pour normal, unif pour uniforme, etc.)
qLOI(q, paramètres)	quantile à q% de la LOI
ecf(x)	fonction de répartition de x
density(x)	densité de x
hist(x)	histogramme de x
qqplot(x, y)	qqplot du vecteur x par rapport au vecteur y
lm(x∼y)	régression de x par y

Boucle	
for (i in vecteur) { commandes }	boucle sur éléments du vecteur
if (conditions) { commandes }	exécute les commandes si conditions est à TRUE
else { commandes }	exécute les commandes si les conditions précédentes sont à FALSE
if else (conditions) { commandes }	exécute les commandes si la condition est à TRUE
while (conditions) { commandes }	exécute les commandes tant que les conditions sont vérifiées
repeat { commandes }	boucle infinie répétant les commandes
if (conditions) { break }	stop les commandes si conditions vérifiées