CentraleSupélec

Campus de Gif-Sur-Yvette

Rapport du projet innovation

Etude de prix de microstructure

Module: Projet Innovation - Laboratoire MICS

Abouseir Amine
Layad Naoufal
Oubaik Ilyass

Encadrant du Projet Innovation:
Ioane Muni Toke

Promotion 2019

Avril 2018

Sommaire

List of Figures

1 Introduction

Le prix d'un actif financier peut être défini de plusieurs manières : en se basant sur le prix que l'animateur de marché est prêt à investir pour acheter un titre (**Bid**), le prix auquel l'animateur de marché décide de le vendre (**Ask**), la demi somme de ces prix (**Mid**). La théorie de microstructure, a été mise en place, pour étudier les différents mécanismes mis en oeuvre dans les salles de marchés, et comment ils peuvent influencer la formation du prix du marché. De ce fait, des prix dits de microstructure ont été introduits, ces prix prennent en compte les différentes observations sur les carnets d'ordre, comme le rapport **offre/demande** et le **déséquilibre**, afin de mieux prédire l'évolution à court-terme des prix de transaction.

2 Première définition de micro-prix

Nous nous sommes intéressés dans un premier temps à l'article **"The Micro-Price: A high frequency estimator of future prices"** écrit par Sasha Stoikov, où l'auteur définit le micro-prix comme étant la limite d'une séquence de prix moyens attendus, et fournit les conditions de l'existence de cette limite. Le micro-prix est, par construction, une martingale s'appuyant sur les informations contenues dans le carnet d'ordre, qu'on peut considérer comme le prix **juste** d'un actif. Ce micro-prix dépendra de l'écart (Ask-Bid), ainsi que du déséquilibre.

On cherchera dans la suite à estimer ce micro-prix, en utilisant des données en haute fréquence, et on étudiera les propriétés de ce prédicteur.

2.1 Le prix moyen, le prix moyen pondéré et le micro-prix

La quantité naturelle que nous pouvons définir, à partir des flux des carnets d'ordre, est le **prix moyen**, qui est défini de la manière suivante :

$$M = \frac{1}{2}(P^a + P^b)$$

où P^b est le meilleur **bid price** et P^a est le meilleur **ask price**. Cette quantité est utilisée dans plusieurs cas comme le prix "juste" d'un actif, sauf qu'elle a plusieurs inconvénients. Parmi ces inconvénients, on peut noter que les changements du **prix moyen** sont fortement auto-corrélés, en outre, ils ont lieu rarement, si on compare leurs fréquences aux taux des mises à jours, et finalement cette quantité ne prend pas en compte les volumes aux meilleurs prix.

Afin de prendre en compte les volumes aux meilleurs prix, une autre quantité a été introduite, qui est le **prix moyen pondéré**, défini par:

$$W = IP^a + (1 - I)P^b$$

où le poids I est défini à partir du déséquilibre

$$I = \frac{Q^b}{Q^a + Q^b}$$

où Q^b est le volume total au meilleur **bid price** et Q^a est le volume total au meilleur **ask price**.

De même, le prix moyen pondéré a plusieurs inconvénients, parmi lesquels, on peut noter les changements du prix qui ont lieu après chaque mise à jour du déséquilibre. En plus, il n'y a aucune justification théorique pour le considérer comme le prix "juste" d'un actif, vu qu'il n'est pas nécessairement une martingale. On peut même constater intuitivement que l'arrivée d'un nouvel ordre d'achat, avec un nouveau **ask price**, va diminuer le prix "juste" d'un actif, sauf que l'arrivée de cet ordre d'achat tend à augmenter le prix moyen pondéré, dans certains cas, ce qui est contre l'intuition, cette observation n'est pas générale. Pour cela, il est peut-être légitime d'utiliser les données du level 2.

Pour toutes ces raisons, le document introduit la définition du prix de microstructure qui est équitable et efficace. Ce prix est une martingale, dont la définition est basée sur le **bid price**, le **ask price** et les volumes. L'auteur a défini **micro-prix** comme ce qui suit :

$$P^{micro} = M + g(I, S)$$

où M est le prix moyen, g est une fonction qu'on va chercher à estimer, I est le déséquilibre et S est le spread $S = P^a - P^b$.

Le **micro-prix** est, par construction, une martingale conditionnée par l'état du carnet d'ordre, et peut être considéré comme le prix juste d'un actif. Dans la suite, nous nous intéresserons à l'estimation de la fonction g, ce qui nous permettra d'obtenir le **micro-prix**.

2.2 Cadre théorique

L'auteur introduit les prédictions du i-ème prix moyen :

$$P_t^i = E[M_{\tau_i} | \mathcal{F}_t]$$

où $\mathcal{F}_t$ est l'information contenue dans le carnet d'ordre à l'instant t les τ_i sont des temps d'arrêt qui représentent les changements des prix moyens M_t. On pose

$$\tau_1 = inf\{u > t | M_u - M_{u-} \neq 0\}$$
$$\tau_{i+1} = inf\{u > \tau_i | M_u - M_{u-} \neq 0\}$$

Le micro-prix est défini comme la limite :

$$P_t^{micro} = \lim_{i \to +\infty} P_t^i$$

et si cette limite existe, on obtient une martingale pout tout $t \geq 0$.

Afin de d'étudier le micro-prix, l'auteur a adopté deux hypothèses :

Hypothèse 1:
L'information contenue dans le carnet d'ordre est donnée par la filtration générée par un processus de Markov tri-dimensionnel $\mathcal{F}_i = \sigma(M_t, I_t, S_t)$ où

$$M_t = \frac{1}{2}(P_t^b + P_t^a)$$

est le prix moyen,

$$S_t = \frac{1}{2}(P_t^b - P_t^a)$$

est le spread(divisé par 2),

$$I_t = \frac{Q_t^b}{Q_t^b + Q_t^a}$$

est le déséquilibre dans le carnet d'ordre.

Cette hypothèse peut être généralisée, en incluant d'autres variables générées par les données du Level 2.

Hypothèse 2:
Les variations du prix moyen sont indépendantes du prix moyen

$$E\big[M_{\tau_i} - M_{\tau_{i-1}} \big| M_t = M, I_t = I, S_t = S\big] = E\big[M_{\tau_i} - M_{\tau_{i-1}} \big| S_t = S, I_t = I\big]$$
$$t \leq \tau_{i-1}$$

Cette hypothèse assure que la dynamique du prix reste la même pour chaque tick.

En adoptant les hypothèses ci-dessus, on peut démontrer le théorème suivant qui exprime les prédictions du prix moyen en fonction des trois variables d'état introduites dans l'hypothèse 1.

Théorème 1 Etant données les hypothèses 1 et 2, la prédiction du i ème prix moyen peut être écrite comme ce qui suit

$$P_t^i = P_t + \sum_{k=1}^{i} g^k(I_t, S_t)$$

où

$$g^1(S, I) = E[M_{\tau_1} - M_t | S_t = S, I_t = I]$$

et

$$g^{i+1}(S, I) = E\big[g^i(I_{\tau_1}, S_{\tau_1}) \big| S_t = S, I_t = I\big], \forall i \geq 0$$

Ces quantités sont calculées récursivement.

2.3 Espace fini d'états

Dans cette section, l'auteur introduit une implémentation du modèle du micro-prix. L'auteur discrétise le temps et fait l'hypothèse que le pas de temps est un entier naturel. De même, le déséquilibre I_t prend des valeurs discrètes $1 \leq i \leq n$ et le **Spread** prend des valeurs $1 \leq s \leq m$. Ainsi, le déséquilibre est discrétisé de la manière suivante : $I_t = i$ si $\frac{i-1}{n} \leq I_t \leq \frac{i}{n}$ et le **Spread** est exprimé en fonction du nombre des ticks. Les variations du prix moyen entre deux instants consécutifs prennent des valeurs dans les multiples du demi-tick (la variation minimale du prix).

Dans la suite, on utilise $x = (i, s)$ pour représenter le vecteur d'états $X_t = (I_t, S_t)$.

Nous pouvons calculer l'ajustement du premier ordre du micro-prix, en utilisant les techniques habituelles sur les processus de Markov discrets avec des états absorbants. On trouve dans ce cas:

$$G^1(x) = E[M_{\tau_1} - M_t | X_t = x] = \sum_{k \in K} kP(M_{\tau_1} - M_t = k | X_t = x)$$

$$= \sum_{k \in K} \sum_{u} kP(M_{\tau_1} - M_t = k \wedge \tau_1 - t = u | X_t = x)$$

L'auteur définit les probabilités des états absorbants et transients, respectivement, de la manière suivante :

$$R^1_{xk} = P(M_{t+1} - M_t = k | X_t = x)$$

$$Q_{xy} = P(M_{t+1} - M_t = 0 \wedge X_{t+1} = y | X_t = x)$$

On peut remarquer que la matrice R^1 est de dimension $nm \times$Taille du vecteur K, et que la matrice Q est de dimension $nm \times nm$. L'ajustement du premier ordre du micro-prix peut être reécrit en fonction de ces matrices de transition :

$$G^1(x) = (\sum_{s} Q^{s-1} R^1) K = (1 - Q)^{-1} R^1 K$$

Afin de calculer cette formule récursive $G^{i+1}(x) = E[G^i(X_{\tau_1}) | X_t = x]$ sous forme matricielle, l'auteur définit une nouvelle matrice des états absorbants:

$$R^2_{xy} = P(M_{t+1} - M_t \neq 0 \wedge X_{t+1} = y | X_t = x)$$

En utilisant cette nouvelle matrice, on trouve

$$G^{i+1}(x) = (1 - Q)^{-1} R^2 G^i(x)$$

On remarque que la matrice $B = (I - Q)^{-1} R^2$ est de dimension $nm \times nm$. En utilisant les matrices introduites précédemment, on réecrit la i-ème prédiction du prix moyen en fonction des puissances de la matrice B, et on retrouve

$$P^i_t = M_t + \sum_{k=0}^{i} B^k G^1$$

L'expression ci-dessus présente une méthode pour calculer le micro-prix, mais on ne peut pas garantir la convergence de la somme. Le théorème suivant établit une condition de la convergence du micro-prix.

Théorème 2 Si $B^* = \lim_{k\to\infty} B^k$ et $B^*G^1 = 0$, alors

$$\lim_{i\to\infty} P_t^i = P_t^{micro}$$

existe et elle est finie.

2.4 Analyse des données utilisées et estimation du micro-prix

Afin d'illustrer ce qu'on a décrit dans les parties précédentes et d'étudier l'évolution du micro-prix en exploitant des données réelles, nous avons utilisé un fichier de données, fourni par notre encadrant Monsieur **Ioane Muni Toke**. Ce fichier comprend toutes les modifications du carnet d'ordre pour **le stock BNPP.PA le 22 janvier 2015**. Chaque ligne de ce fichier Excel décrit un changement ; en précisant l'instant du changement t_s heures (en secondes) , **type** du prix changé (Bid ou Ask),**level** qui représente le niveau du changement, **price** qui donne le nouveau prix et **qty** détermine la nouvelle quantité. L'approche de l'auteur consiste à n'utiliser que les données du level 1. Pour cette raison, nous devons filtrer, dans un premier temps, les données qu'on va utiliser, en ne conservant que les données du **level 1**. Les données de ce level représentent les changements des meilleures offres. En outre, on a décidé de ne traiter que les données du carnet d'ordre générées entre 10h et 17h, dans le but d'éliminer les lignes qui ne reflètent pas de changement, et qui peuvent gêner le fonctionnement de nos algorithmes.

Pour chaque instant t, représentant l'arrivée d'un nouvel ordre, on calcule le déséquilibre I_t et le Spread S_t. Ces deux quantités sont discrétisées en un nombre fini d'états $1 \leq x \leq nm$. On calcule, en plus, la variation du prix moyen $dM_t = M_{t+1} - M_t$ qui sera souvent nulle.

Après l'étude du cadre théorique dans un premier temps, nous nous sommes intéressés à la mise en place de la procédure de l'estimation du micro-prix suivante, proposée par l'auteur. Cette procédure consiste en :

*Symétriser les données, à chaque observation $(I_t, S_t, I_{t+1}, S_{t+1}, dM)$, on introduit une observation symétrique $(1 - I_t, S_t, 1 - I_{t+1}, S_{t+1}, -dM)$. Cette symétrisation assure que $B^*G^1 = 0$, cette conséquence nous permet, grâce au **Théorème 2**, de garantir la convergence du micro-prix.

*Estimer les matrices de probabilités de transition Q, R_2 et R_1.

*Calculer $G^1 = (1 - Q)^{-1}R^1K$

*Calculer $B = (1 - Q)^{-1}R^2$

*Calculer le micro-prix:

$$G^* = P^{micro} - M = G^1 + \sum_{i=1}^{\infty} B^i G^1$$

En pratique cette somme converge rapidement, c'est ce qu'on illustrera dans la suite.

2.4.1 Etude des données

Après le filtrage des données, nous nous sommes intéressés, ci-dessous, à la distribution des Spreads dans le stock qu'on étudie.

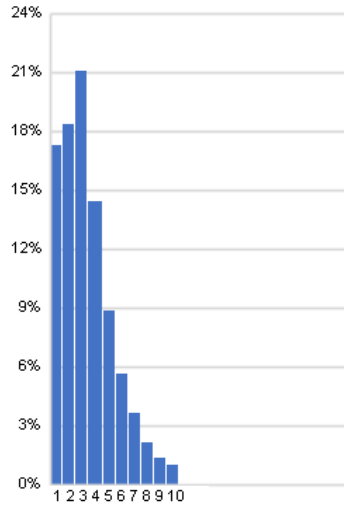


Figure 1: L'histogramme du Spread dans le stock BNPP.PA du 22 janvier 2015, pour les données du level 1

Tout d'abord, on constate que le tick caractérisant notre stock est de 0.005,

ce tick correspond à la plus petite variation possible du prix. En plus, on remarque clairement que le stock étudié a une distribution de Spread large et dispersée, la hauteur des pics augmente jusqu'au troisième pic, qui correspond à 3 ticks, donc le spread le plus fréquent est de **0.015** avec une hauteur de 70 000, puis commence à décroître. L'histogramme montre que les pics correspondant aux valeurs de Spread élevées, sont négligeables devant ceux des trois premiers Spreads.

De même, et vu que la procédure de l'estimation du micro-prix fait intervenir la variation du prix moyen dM_t, nous avons décidé de tracer l'histogramme de la variation des prix moyens. Et on trouve l'histogramme suivant :

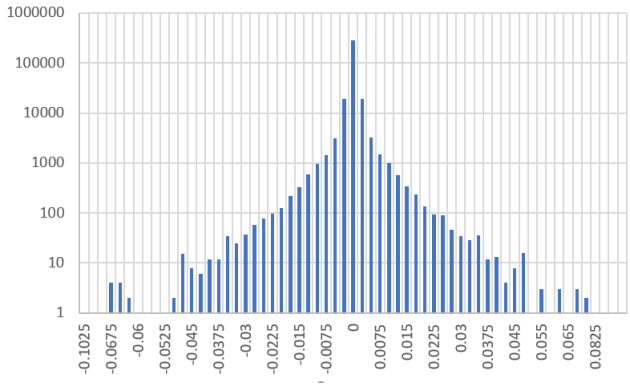


Figure 2: L'histogramme de la variation du prix moyen dM_t, avec une échelle semi-logarithmique dans le stock BNPP.PA du 22 janvier 2015, pour les données du level 1

L'histogramme ci-dessus affirme la remarque qu'on a faite dans la partie précédente, que la variation du prix moyen entre deux instants successifs est souvent nulle. Dans le cas de notre stock, on constate que la hauteur du pic, correspondant à une variation nulle, atteint 280 000, ce pic est entouré par d'autres pics ayant des hauteurs très inférieures à la hauteur du pic central. On peut même noter que l'histogramme est quasi-symétrique par rapport à 0, ce qui est normal, vu que l'augmentation du prix et sa diminution sont équiprobables.

2.5 Implémentation du programme de l'estimation du micro-prix

On appliquera la méthode du calcul du micro-prix à la base de données contenant toutes les modifications du carnet d'ordre pour le stock BNPP.PA le 22 janvier 2015. Comme nous l'avons déjà précisé précédemment, chaque ligne contient les cinq variables suivantes et qui permettent de représenter un changement :

- **ts** : date du changement (en secondes).
- **type** : côté du changement (A pour ask et B pour bid).
- **level** : niveau du changement.
- **price** : nouveau prix.
- **qty** : nouveau volume.

On s'intéresse aux meilleures offres de l'ask et du bid, donc on ne conserve que les changements dont le niveau (level) est égal à 1. On construit alors la série temporelle (a, b, qA, qB) des prix ask, bid et quantités ask et bid qui nous permettra de calculer les différentes matrices qui interviennent dans la détermination du micro-prix. On choisit de coder en Python le code nous permettant de lire la base de données, en extraire la série temporelle (a, b, qA, qB) puis en déduire le micro-prix.

Ce code est constitué des 12 fonctions suivantes :

• **csv2txt(directory)** : Cette fonction prend en entrée l'emplacement du fichier DB.csv qui continent les changements du carnet d'ordre. Elle lit son contenu en le considérant comme un fichier texte puis l'explore ligne par ligne et ne traite que celle dont le niveau est égal à 1. On construit alors un vecteur T, qui contient les dates des changements de niveau 1, et une matrice X à quatre colonnes (a, b, qA, qB) où la ième ligne représente le ième changement de niveau 1. Pour ne pas devoir lire le fichier **DB.csv** à chaque exécution, on stocke ces derniers sous forme de fichiers texte en utilisant la fonction numpy.savetxt, on obtient alors deux fichiers **DB.txt** et **TS.txt** qu'on lira par la suite à l'aide de la fonction **numpy.loadtxt**.

• **format-csv(directory)** : Cette fonction effectue le même traitement que la fonction précédente mais permet quant à elle de générer un fichier **new-DB.csv** qui contient les changements du carnet d'ordre de niveau 1 représentés par (ts, a, b, qA, qB). Ce fichier sera utilisé pour visualiser les données plus facilement et effectuer quelques graphiques statistiques (histogramme des spreads ...).

• **calc(directory)** : Cette fonction effectue le calcul du prix moyen M_t, le spread S_t et du déséquilibre I_t pour chaque changement. Elle charge le fichier DB.txt en mémoire en utilisant numpy.loadtxt et calcule une matrice Y qui contient trois colonnes M,S et I.

• **sym(directory)** : Comme l'avait précisé l'article, il faut symétriser les données de telle façon que pour chaque observation (I_t, S_t, I_{t+1}, S_{t+1}, dM_t) l'observation ($1 - I_t$, S_t, $1 - I_{t+1}$, S_{t+1}, $-dM_t$) existe aussi. En utilisant la matrice que retourne la fonction précédente, on construit une matrice qui contient les ob-

servations (I_t, S_t, I_{t+1}, S_{t+1}, dM_t) et on symétrise cette dernière pour pouvoir garantir la convergence du micro prix.

• **discrete(directory,n)** : Cette fonction discrétise les observations (I_t, S_t, I_{t+1}, S_{t+1},dM_t), c'est-à-dire qu'on exprime les variables S_t, S_{t+1} et dM_t en nombre de demi-ticks. Il faut noter qu'on utilise des demi-ticks car $S_t = \frac{A_t - B_t}{2}$ et $M_t = \frac{A_t + B_t}{2}$ donc quand l'ask ou le bid varie d'un tick St et M_t (et donc dM_t) varie d'un demi-tick. En ce qui concerne le déséquilibre I_t, on choisit n et on discrétise I_t de telle façon que $I_t = i$ si $\frac{i-1}{n} \leq I_t \leq \frac{i}{n}$. Cette fonction retourne la matrice des observations discrétisées, le vecteur K des valeurs que prend dM_t et le nombre m qui représente la valeur maximale de S_t. Il est aussi important de noter qu'on a choisi dans cette fonction de supprimer les observations pour lesquelles le spread prend une valeur supérieure à 20. Ce choix nous permet d'améliorer d'une façon considérable le temps d'exécution et n'influence pas sur le calcul du micro prix car ces observations sont peu nombreuses comme le montre l'histogramme du spread suivant :

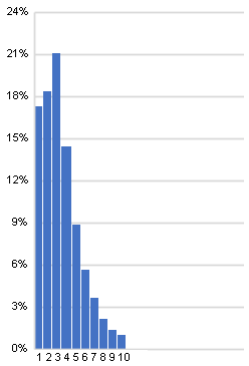


Figure 3: L'histogramme du Spread dans le stock BNPP.PA du 22 janvier 2015, pour les données du level 1

• **calc-Q(directory,n)** : Cette fonction permet de calculer la matrice Q. Cette matrice est définie par : $Q_{xy} = P(M_{t+1} - M_t = 0 \wedge X_{t+1} = y | X_t = x)$ avec $X_t = (I_t, S_t)$ Les variables It, St, Mt+1 et Mt sont obtenues en utilisant la fonction discrete(directory,n). La matrice Q est alors de taille $nm \times nm$, on a donc intérêt à ne pas choisir une valeur très grande pour n et on voit aussi l'utilité du choix qu'on a fait dans la fonction précédente c'est à dire supprimer les observations pour lesquelles le spread prend une valeur supérieure à 20 cela permet d'avoir un m inférieur ou égal à 20.

• **calc-R1(directory,n)** : Cette fonction permet de calculer la matrice R^1. Cette matrice est définie par : $R^1_{xk} = P(M_{t+1} - M_t = k | X_t = x)$ Cette matrice est de taille $nm \times l_K$ où l_K est la taille du vecteur K des valeurs que prend dMt que nous avons introduit dans la fonction discrete(directory,n).

• **calc-R2(directory,n)** : Cette fonction permet de calculer la matrice R^2 de taille nm x nm. Cette matrice est définie par :$R^2_{xy} = P(M_{t+1} - M_t \neq 0 \wedge X_{t+1} = y | X_t = x)$

• **calc-G1(directory,n)** : Cette fonction permet de calculer la matrice G^1. Cette matrice est définie par : $G^1 = (1-Q)^{-1}R^1K$

• **calc-B(directory,n)** : Cette fonction permet de calculer la matrice B. Cette matrice est définie par : $B = (1-Q)^{-1}R^2$

• **calc-Gs(directory,n,l,override)**: Cette fonction permet de calculer la matrice G^*. Cette matrice est définie par :

$$G^* = G^1 + \sum_{i=1}^{\infty} B^i G^1$$

Afin de calculer la somme infinie des B^iG^1 qui intervient dans l'expression de G^*, on va tronquer la somme à un nombre fini de termes. Pour cela, on va étudier tout d'abord la convergence de la série. On aimerait dans l'idéal sommer le plus possible de ces termes. Mais ceci a des conséquences négatives sur le temps d'exécution de la fonction. On décide alors d'étudier la convergence de la série pour déterminer une valeur sure du paramètre l qui garantit une erreur moindre.

On calcule alors la valeur de $\frac{\|B^{i+1}G^1\|_\infty}{\|B^iG^1\|_\infty}$ (On choisit la norme infinie car les normes sont équivalentes en dimension finie), car si on arrive à prouver que ce quotient est inférieur, à partir d'un certain rang, à un ϵ qui est strictement inférieur à 1, on montrera que le reste de la série de terme général $\|B^iG^1\|_\infty$ converge vers 0, car à partir d'un certain rang, chaque terme du reste $\|B^iG^1\|_\infty$ est majoré par ϵ^{i-m_0}. On obtient, donc, la courbe suivante, qui nous justifie le constat déjà fait, que la série du micro-prix converge rapidement, pour cela et afin de diminuer la complexité de nos fonctions, on peut se contenter d'un n qui assurera la convergence.

En s'intéressant à la convergence absolue de la série en utilisant la norme 1, on obtient la courbe ci-dessous. Cette courbe justifie la convergence absolue de la série donc la convergence simple de la série.

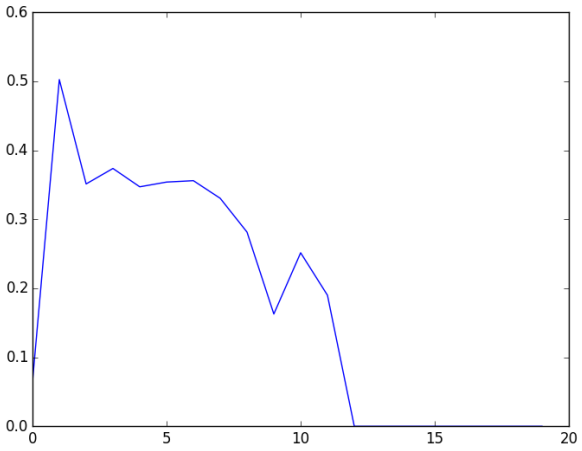


Figure 4: La courbe de convergence du quotient $\frac{\|B^{i+1}G^1\|_\infty}{\|B^iG^1\|_\infty}$

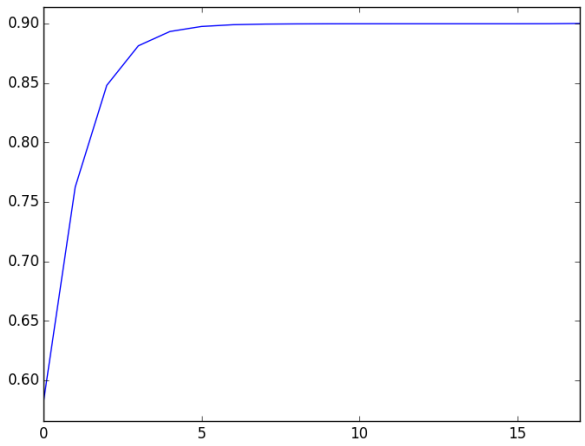


Figure 5: La somme partielle de la norme 1 des termes de la série en fonction de n.

On remarque que la convergence des B^iG^1 est rapide comme l'avait précisé l'article, il suffit de prendre l=12 pour garantir qu'on est proche de G^*.

Remarque :

Le calcul de G^* nécessite le calcul de B et de G^1, ces derniers emploient les matrices Q, R^1 et R^2. Donc à chaque exécution, on est contraint d'effectuer ces calculs qui sont très couteux en termes de temps. Nous avons alors décider de stocker les matrices B, G^1 et G^* dans des fichiers texte pour les initialiser rapidement. Par ailleurs, si on souhaite refaire le calcul, on donne au booléen **override** la valeur True pour ignorer le fichier texte Gs.txt et refaire le calcul de la série. Si on souhaite refaire tous les calculs, il suffit de supprimer les fichiers **B.txt**, **G1.txt** et **Gs.txt** qui se trouve dans le dossier Files.

• **IvsGs(directory,n,l,spreads,override)**: Cette fonction permet de tracer G^* en fonction du déséquilibre pour des valeurs de spread données. On obtient par exemple lorsqu'on exécute **IvsGs(directory,20,15,[1,3,5],False)** la figure ci-dessous :

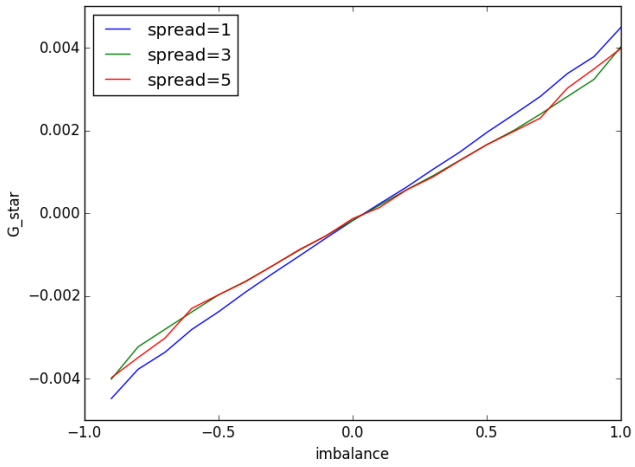


Figure 6: L'évolution du micro-prix en fonction du déséquilibre, pour les cas Spread=1,3 et 5

• **empiricalE(directory,n,spread,delta)**: Cette fonction de calculer l'espérance empirique $E[M_{t+\delta} - M_t|I_t]$ pour les différentes valeurs de I_t et pour un spread donné puis la trace. On obtient par exemple qu'on exécute empiricalE(directory,50,3,5) la figure 7.

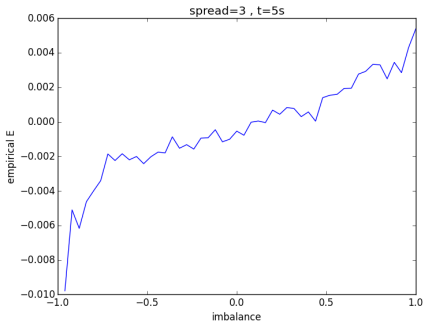


Figure 7: L'espérence empirique en fonction du déséquilibre pour un spread égale à 3 et t=5s

- **empiricalEw(directory,n,spread,delta)**: Cette fonction de calculer l'espérance empirique $E[W_{t+\delta} - M_t|I_t]$ pour les différentes valeurs de I_t et pour un spread donné puis la trace. On obtient par exemple qu'on exécute empiricalEw(directory,50,3,5) la figure 8.

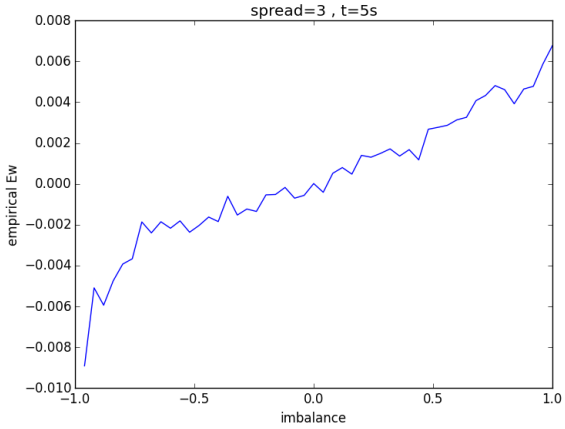


Figure 8: L'espérence empirique $E[W_{t+\delta} - M_t|I_t]$ en fonction du déséquilibre pour un spread égale à 3 et $\delta = 5s$

2.6 Résultats

Les premiers résultats qu'on a obtenus, décrivant l'évolution du micro-prix en fonction du déséquilibre, pour de différentes valeurs du Spread, sont les suivants :

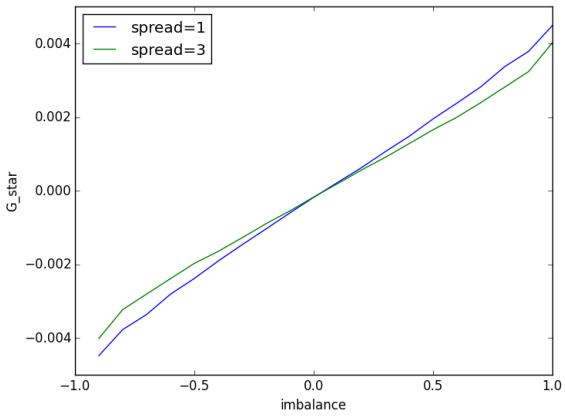


Figure 9: L'évolution de G^* en fonction du déséquilibre pour S=1,3

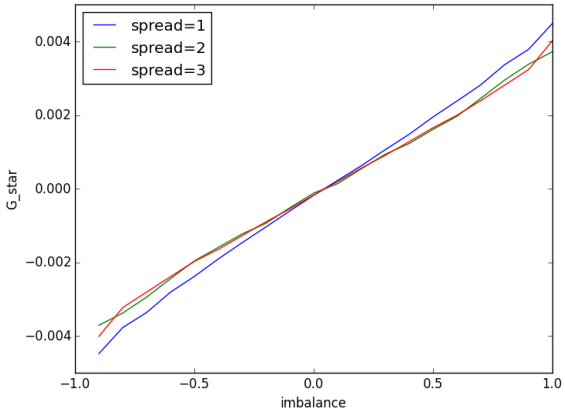


Figure 10: L'évolution de G^* en fonction du déséquilibre pour S=1,2,3

Comme on le constate dans les figures 9 et 10, lorsque le spread augmente, la courbe de G^* en fonction du déséquilibre devient plus aplatie, ce qui est normal, vu que si la différence Bid-Ask est grande, les possibilités de transactions sont nombreuses. De ce fait le micro-prix sera proche du prix moyen pour les différentes valeurs du déséquilibre.

Dans les figures 11 et 12, ce constat n'est pas très clair. On peut justifier ceci, par la taille des données traitées, dans le document, l'auteur utilise les données générées pendant un mois, cependant nous utilisons les données d'une seule journée.

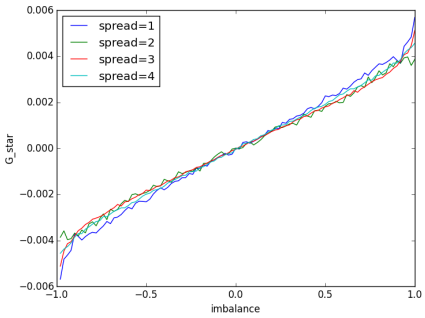


Figure 11: L'évolution de G^* en fonction du déséquilibre pour S=1,2,3 et 4

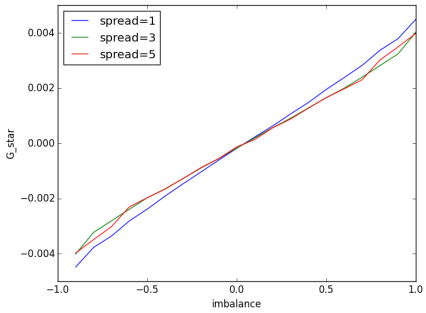
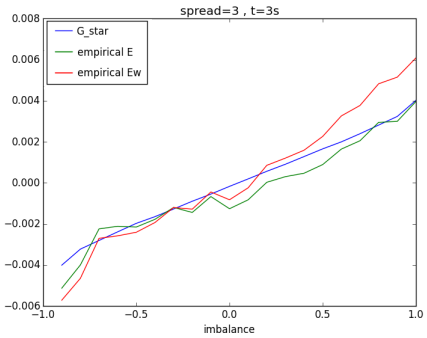
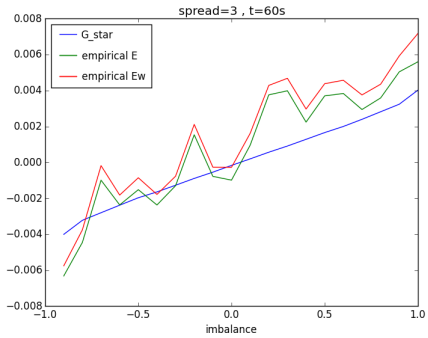


Figure 12: L'évolution de G^* en fonction du déséquilibre pour S=1,3,5

Lorsqu'on superpose les figures des du G_*, $E[M_{t+\delta} - M_t|I_t]$ et $E[W_{t+\delta} - M_t|I_t]$ pour un spread $= 3$ et $\delta = 3$ et $\delta = 60s$, on obtient les figures ci-dessous :





Il est difficile de prouver que le micro-prix est un bon prédicteur des prix futurs. Mais, on peut déterminer empiriquement son efficacité à prédire le prix futur, il faut choisir une échelle de temps donnée. Si l'échelle de temps est trop courte, les changements du prix empirique vont souffrir du bruit de la microstructure, selon que l'on essaie de prédire le prix moyen ou le prix moyen pondéré. Si l'échelle de temps est trop grande, les changements du prix empirique seront bruyants et leurs barres d'erreur seront grandes. On peut remarquer qu'à cette échelle de temps, le micro-prix est un meilleur prédicteur du prix moyen que du prix moyen pondéré. Cependant, l'auteur trouve que le micro-prix est un meilleur prédicteur du prix moyen pondéré que du prix moyen, et justifie ceci, par le bruit de microstructure dans le prix moyen, qui souffre souvent de **rebond bid-ask** à cette échelle de temps. Dans le cas d'une échelle de temps

assez grande, l'auteur annonce que les effets de bruit microstructure se dissipent et les prédictions à prix moyen et à prix moyen pondérées coïncident. Les barres d'erreur sont, dans ce cas, beaucoup plus grande, et la validation du micro-prix à plus long terme n'est pas très réaliste. Néanmoins, nous n'observons pas ceci sur la deuxième courbe, dessinée pour un $\delta = 60s$. On peut justifier cette différence entre nos résultats et les résultats de l'auteur, par la taille des données, qu'on utilise pour estimer le micro-prix. En effet, nous traitons des données générées en une journée, cependant l'auteur utilise des données générées pendant un mois.

Nous allons travailler, dans la suite de notre projet, sur des données balayant plusieurs journées, afin d'obtenir des résultats plus interprétables.

2.7 Vérification des résultats sur des nouvelles données

Afin de vérifier la performance de nos algorithmes et la cohérence entre nos résultats et ceux du document, nous avons utilisé comme base de données, les données **BNPP.PA** de la totalité du mois de Janvier 2015, pour voir si la taille des données va améliorer la compatibilité de nos résultats avec ceux trouvés dans le document.

En traçant la courbe de G^* en fonction du déséquilibre, pour de différentes valeurs du **spread** (1,2 et 3), on trouve la courbe ci-dessous, sur laquelle nous constatons qu'effectivement, la figure des spreads 1,2 et 3 présente effectivement la propriété d'aplatissement des courbes de G* au fur et à mesure que le spread augmente ce qui n'était pas le cas lorsqu'on n'utilisait que les données d'une seule journée.

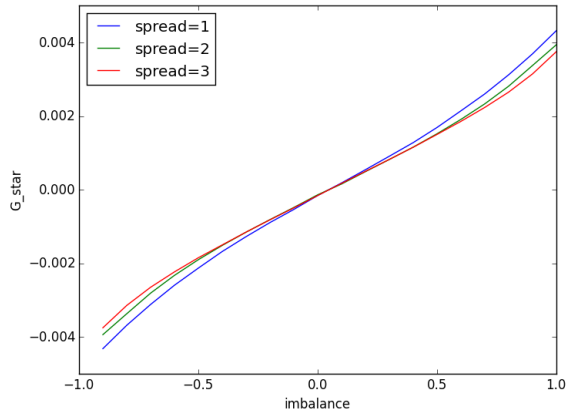


Figure 15: L'évolution de G_* en fonction du déséquilibre pour S=1,2,3

En ce qui concerne les courbes de comparaison de G* et des espérances empiriques, nous avons effectué les modifications nécessaires sur la fonction **empiricalE** pour symétriser les données lors du calcul des espérances empiriques. Nous

avons aussi optimisé cette fonction pour avoir un temps de d'exécution beaucoup plus rapide et indépendant du pas de temps choisi (complexité : O(n) où n est la taille des données). A titre d'exemple, les nouveaux temps d'exécution (en minutes) sur les données d'un mois sont les suivants : 2:21, 2:23, 2:26 pour dt=1s, dt=10s et dt=60s respectivement pour un spread valant 3, alors que pour l'ancien code on obtenait : 2:17, 3:27 et 9:25. En choisissant un spread valant 1 et pour un pas de temps $\delta = 1s$, on obtient la courbe ci-dessous,

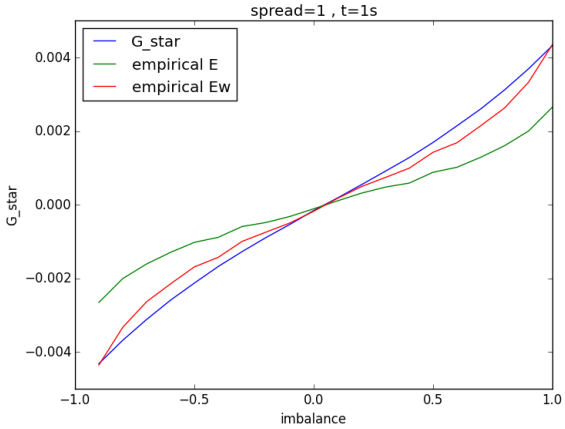
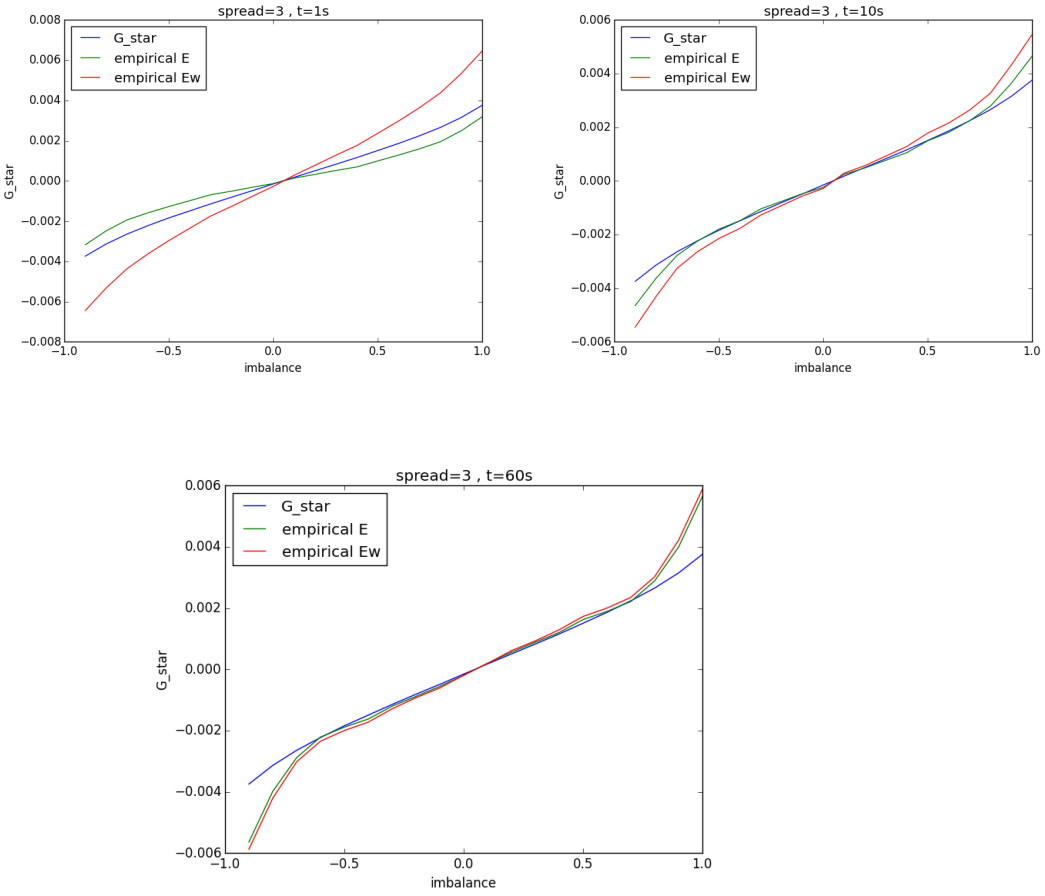


Figure 16: Les trois courbes superposées pour un Spread=1 et $\delta = 10$ s

Avec un petit pas de temps, on constate que le micro-prix estime mieux le prix moyen pondéré que le prix moyen, d'après l'article, ceci est dû à l'effet du bruit de microstructure sur le prix moyen.

En prenant un Spread S=3, qui est le plus fréquent, et les pas de temps $\delta = 1, 10$ et 60, on obtient les courbes ci-dessous, à partir desquelles on constate que les deux courbes du prix moyen et du prix moyen pondéré se confondent lorsque le pas du temps augmente. En outre, on remarque que le micro prix estime parfaitement les deux prix sur des intervalles du déséquilibre centrés en 0, par exemple, dans le cas de $\delta = 60s$, les trois courbes sont confondues sue un intervalle du déséquilibre $[-0.75, 0.75]$:



3 Deuxième définition du micro-prix

Nous nous sommes intéressés dans la seconde partie du projet, à un deuxième article intitulé **A continuous and efficient fundamental price on the discrete order book grid**, écrit par **Julius Bonart** et **Fabrizio Lillo**. Les auteurs de l'article traitent et essaient d'adapter le modèle de MRR (Madhavan, Richardson, et Roomans (1997)) de formation de prix aux carnets d'ordres, en introduisant un prix dit fondamental, qui est un prix continu et efficace et peut prendre des valeurs en dehors de l'intervalle défini par les meilleurs Bid et Ask. Les auteurs de l'article étudient les prédictions de leur modèle en se basant sur des tests empiriques. Finalement, ils utilisent leur modèle afin de proposer un estimateur du prix fondamental en étudiant sa performance par rapport aux autres estimateurs.

Comme précédemment, un prix juste ou efficace est un prix qui contient,

sans ambiguïté, toutes les informations disponibles. Cependant, les différentes frictions microstructurales empêchent le prix observé de refléter librement le prix efficace. Parmi celles-ci, la mise en œuvre de la discrétisation des prix dans la plupart des marchés modernes et qui, joue un rôle primordial dans la formation de prix. Par conséquent, il n'y aucune raison pour que les meilleurs prix de l'Ask et du Bid, ou leur prix moyen soient confondus avec un prix efficace fondamental.

3.1 Données utilisées

Afin de vérifier les relations et les figures présentes dans cet article, nous utilisons des fichiers de données différents de ceux employés précédemment. Les définitions qui ont été introduites nécessitent la connaissance des transactions effectuées sur un carnet d'ordre et non l'état du carnet d'ordre. Ainsi, on utilisera un autre format de fichier différents des précédents mais toujours pour le mois de janvier 2015 pour le stock BNPP.PA. Ces fichiers regroupent l'ensemble des ordres, leur type (Market, Limit ou Cancel), leur côté (ask ou bid), leur niveau, leur prix et leur quantité. On précise aussi le spread et les quantités des meilleurs ask et bid à l'instant de la soumission de l'ordre.

Dans notre cas, on ne s'intéresse qu'aux ordres de marché. On utilisera alors des fonctions implémentées en Python, similaires à ceux utilisées pour le calcul du micro-prix, afin de lire ces fichiers et extraire les données et construire la suite $u(t) = (a(t), b(t), V_a(t), V_b(t), e(t))$ qui représente les prix et les quantités des meilleurs ask et bid à l'instant de la soumission du t-ième ordre de marché, et l'indicateur e_t qui est égal à 1 s'il s'agit d'ordre d'achat ou -1 s'il s'agit d'ordre de vente.

Cette suite nous permettra de vérifier les résultats et les figures présentés dans l'article pour le stock BNPP.PA. L'article classifie les actions selon la taille du tick en calculant l'espérance du spread. Une action est considérée à large tick si $E(s_t) \leq 1.3 \times tick$ et elle est considérée à petit tick si $E(s_t) \geq 4 \times tick$. Les autres actions sont considérées comme étant à tick moyen.

Le calcul de l'espérance su spread pour l'action BNPP.PA donne une valeur $E(s_t) = 0.008 = 1.6 \times tick$ où le tick vaut 0.005. D'après la définition précédente, cette action est à tick moyen mais l'espérance de son spread est plus proche du domaine des actions à large tick.

3.2 La formation des prix

La formation des prix décrit le processus par lequel l'information, la liquidité et le flux des ordres influencent le prix du marché. Le mécanisme de l'impact sur le prix n'est pas du tout évident. Il est connu qu'empiriquement les flux d'ordres sont fortement corrélés, en revanche les prix moyens résultants sont presque décorrélés.

Un simple modèle structurel de formation de prix a été développé, en supposant l'existence d'un prix fondamental du stock, qu'on va noter p_t dans toute la suite, où t représente le temps de la transaction et p_t est le prix fondamental

immédiatement avant la transaction ayant lieu avant l'instant t. Si une transaction à l'instant t est initiée par l'acheteur, on lui affecte l'indicateur $e_t = 1$, et si elle est initiée par le vendeur, on lui affecte l'indicateur $e_t = -1$. En outre, le modèle MRR suppose que les prix sont positivement corrélés à l'innovation dans les flux d'ordres.

On peut formaliser ce qu'on vient de dire, de la manière suivante :

$$p_{t+1} - p_t = G[e_t - \hat{e}_t] + W_t$$

où

$$\hat{e}_t = E_{t-1}[\hat{e}_t]$$

est le signe de transaction attendu à un instant t, en se basant sur l'information publique jusqu'à l'instant t. Les auteurs de l'article définissent cette information publique à travers l'historique des transactions, qui ont eu lieu dans le passé, $e_{t-1}, e_{t-2}...$ et les chocs $W_{t-1}, W_{t-2}...$ qui décrivent les informations externes. Les auteurs supposent que W_t est un bruit blanc de moyenne nulle, qui est décorrélé avec l'historique des transactions. D'une autre part, l'information privée du prix se manifeste à travers le terme $G[e_t - \hat{e}_t]$, vu que W_t fait partie de l'information publique, le contenu de la transaction ne dépend donc que de sa partie inattendue $e_t - \hat{e}_t$.

Par construction, ce mécanisme de formation assure l'efficacité du prix, vu que ce prix vérifie ce qui suit:

$$E_{t-1}[p_{t+1} - p_t] = E_{t-1}[G[e_t - \hat{e}_t] + W_t]$$

or W_t est indépendant avec l'historique des transactions et il est de moyenne nulle donc

$$E_{t-1}[W_t] = E[W_t] = 0$$

$$E_{t-1}[p_{t+1} - p_t] = E_{t-1}[G[e_t - \hat{e}_t]]$$

$$E_{t-1}[p_{t+1} - p_t] = GE_{t-1}[e_t - \hat{e}_t]$$

or

$$E_{t-1}[e_t] = \hat{e}_t$$

donc

$$E_{t-1}[p_{t+1}] = p_t$$

indépendamment des corrélations des (e_t).

Dans un marché concurrentiel, les teneurs de marchés (Market makers) mettent en place les Ask Prix et le Bid prix, de la manière suivante :

$$a_t = p_t + G[1 - \hat{e}_t]$$

$$b_t = p_t + G[-1 - \hat{e}_t]$$

ce qui assure une moyenne de gain nulle. Par conséquent, on obtient

$$x_t = \frac{a_t + b_t}{2} = p_t - G\hat{e}_t$$

$$s_t = a_t - b_t = 2G$$

Et vu que les teneurs de marchés cherchent à anticiper l'effet de la prochaine transaction sur les prix, le prix moyen à l'instant t ne sera pas égal à p_t. Et d'après de ce qui précède, l'équation de son évolution sera la suivante :

$$x_{t+1} - x_t = G[e_t - \hat{e}_t] + W_t$$

Les trois chercheurs qui ont mis en place le modèle MRR ont étudié l'évolution de cette équation et ont comparé le spread implicite aux spreads historiques. Et ils ont concluent que la discrétisation du prix constitue la limite à l'applicabilité de leur modèle.

Afin d'étudier l'évolution de cette équation, les auteurs de l'article traduisent l'équation ci-dessus, en introduisant des quantités observables, qui sont faciles à calculer empiriquement. Ils définissent donc la fonction de réponse de la transaction.

$$R(l) = E[e_t(x_{t+l} - x_t)]$$

ce qui devient

$$R(l) = GE[1 - e_t\hat{e}_{t+l}] = G - GE[e_tE_t(\hat{e}_{t+l} - e_{t+l})] - GE[e_te_{t+l}] = G[1 - C(l)]$$

En utilisant le fait que

$$E_t[\hat{e}_{t+l} - e_{t+l}] = 0$$

et en définissant la fonction C

$$C(l) = E[e_{t+l}e_t]$$

Ces deux quantités $R(l)$ et $C(l)$ sont facilement calculables en utilisant des données empiriques, et on peut vérifier, sur des données empiriques facilement que

$$R(1) = R(l)\frac{1 - C(1)}{1 - C(l)}$$

pour les différentes valeurs de l, on peut retrouver la valeur de G. Cette équation n'apparaît pas dans le papier original du MRR, mais on la retrouve dans d'autres papiers.

	Données
$R(1)$	0.00259603
$R(2)\frac{1-C(1)}{1-C(2)}$	0.00272346
$R(3)\frac{1-C(1)}{1-C(3)}$	0.00276392
$R(4)\frac{1-C(1)}{1-C(4)}$	0.00276137
$R(5)\frac{1-C(1)}{1-C(5)}$	0.00275858
$R(6)\frac{1-C(1)}{1-C(6)}$	0.00275785
$R(7)\frac{1-C(1)}{1-C(7)}$	0.00272646
$R(8)\frac{1-C(1)}{1-C(8)}$	0.00269459
$R(9)\frac{1-C(1)}{1-C(9)}$	0.00266932
$R(10)\frac{1-C(1)}{1-C(10)}$	0.00264911
$R(11)\frac{1-C(1)}{1-C(11)}$	0.00264212
$R(12)\frac{1-C(1)}{1-C(12)}$	0.00262917
$R(13)\frac{1-C(1)}{1-C(13)}$	0.002622
$R(14)\frac{1-C(1)}{1-C(14)}$	0.00263522
$R(15)\frac{1-C(1)}{1-C(15)}$	0.00265663
$R(16)\frac{1-C(1)}{1-C(16)}$	0.00267745

Le tableau ci-dessus montre, qu'effectivement, l'équation d'égalité $R(1) = R(l)\frac{1-C(1)}{1-C(l)}$ est bien vérifiée, en utilisant nos données, avec un tick valant 0.005. Dans l'article, les auteurs ont dressé un tableau pour des données de ticks larges et petits, et la relation $R(1) = R(l)\frac{1-C(1)}{1-C(l)}$ a été approximativement toujours vérifiée.

Dans la suite de l'article, les auteurs essayent d'expliquer, en utilisant le minimum des hypothèses, pourquoi la relation reste vérifiée même dans le cas d'un tick large.

Dans la suite, nous utiliserons le même tick de 0.005 en le considérant comme un tick large, comme précisé avant.

3.3 La formation des prix en cas de tick large

On suppose l'existence d'un prix fondamental qui prend des valeurs continues et satisfait l'équation MRR. Quand le tick est assez grand, le meilleur bid et le meilleur ask sont séparés par un seul tick :$a_t - b_t = \tau$, les équations $a_t = p_t + G[1 - \hat{e}_t]$ et $b_t = p_t + G[-1 - \hat{e}_t]$ sont, dans ce cas, incorrectes car elles ne prennent pas en considération la discrétisation du prix.

Avec un tick large, le prix moyen change quand l'ordre limite est mis en file d'attente. Nous supposons que les fournisseurs de liquidité possèdent le volume aux meilleures cotations tant qu'il est marginalement rentable, c'est-à-dire tant que son prix d'exécution attendu est égal au prix fondamental immédiatement après la transaction.

Un ordre limite reste rentable tant que :

$$p_t \in (b_t - r - G[-1 - \hat{e}_t], a_t + r - G[1 - \hat{e}_t])$$

où $r > 0$ est la remise sur le marché par action offerte par la bourse, $p_t + G[-1 - \hat{e}_t]$ est le prix fondamental après l'exécution de l'ordre limite d'achat en b_t, et $p_t + G[1 - \hat{e}_t]$ est le prix fondamental après l'exécution de l'ordre limite de vente à a_t , on peut donc dire qu'un ordre limite peut être profitable à son propriétaire même lorsque le prix fondamental est en dehors de l'intervalle (b_t,a_t).

La discrétisation des prix empêche la vérification de l'équation dynamique de MRR pour x_t, $(x_{t+1} - x_t = G[e_t - \hat{e}_t] + W_t)$, mais nous pouvons montrer qu'elle est encore vraie en moyenne. Nous supposons que la distribution de p_t dans l'intervalle caractéristique est symétrique par rapport à son centre. Dans ce cas, le prix fondamental attendu, compte tenu du prix moyen, , et d'après l'équation: $p_t \in (b_t - r - G[-1 - \hat{e}_t], a_t + r - G[1 - \hat{e}_t])$, est égal à la moyenne des bords de l'intervalle caractéristique autour de x_t :

$$E[p_t|x_t] = x_t + G\hat{e}_t$$

Nous pouvons appliquer l'espérance conditionnelle à la relation ci-dessus, en considérant les instants passés jusqu'à $t' \leq t$, et à une moyenne de x_t donnée. Ceci est utile vu que l'équation MRR implique , $GE_{t'}[p_{t+1} - p_t] = GE_{t'}[e_t - \hat{e}_t]$ pour $t' \leq t$.

Tout cela nous permet d'écrire :

$$GE_{t'}[e_t - \hat{e}t] = E_{t'}[x_{t+1} - x_t] + GE_{t'}[\hat{e}_{t+1} - \hat{e}_t]$$

ou simplement

$$E_{t'}[x_{t+1} - x_t] = E_{t'}[e_t - \hat{e}_{t+1}]$$

notons que $t' \leq t$ est arbitraire. On peut sommer les deux équations précédentes pour obtenir

$$E_t[x_{t+l} - x_t] = GE_t[e_t - \hat{e}_{t+1+l}]$$

Nous pouvons maintenant procéder en calculant la fonction de réponse d'une transaction :

$$R(l) = E[e_t(x_{t+l} - x_t)] = G[1 - C(l)]$$

C'est exactement la relation MRR $R(1) = R(l)\frac{1-C(1)}{1-C(l)}$, que nous avons bien testée sur des données empiriques. Vu que MRR a supposé une échelle de prix continue, il n'était pas clair pourquoi ses prédictions se sont révélées être vraies aussi bien pour les grands ticks. Alors que la discrétisation des prix implique que l'équation MRR $x_{t+1} - x_t = G[e_t - \hat{e}_t] + W_t$ est incorrecte en l'état, notre équation $E_t[x_{t+l} - x_t] = GE_t[e_t - \hat{e}_{t+1+l}]$ démontre qu'il est néanmoins vrai en moyenne: Les relations entre les quantités qui dépendent linéairement du prix moyen ne sont donc pas affectées par la taille du tick.

Dans ce qui suit, nous introduisons la fonction d'autocovariance des rendements au prix moyen. Parce que l'autocovariance dépend quadratiquement du prix moyen, l'espérance mathématique en conduit à des résultats qui sont très différents de ce que l'original MRR prédit.

3.3.1 L'effet du déséquilibre

Si nous acceptons qu'un prix fondamental doit refléter l'information publiquement disponible, il est alors facile de montrer empiriquement que le prix moyen ne peut être le prix fondamental efficace. On définit le volume du déséquilibre par :

$$I(t) = \frac{V_b(t) - V_a(t)}{V_b(t) + V_a(t)}$$

où $V_a(t)$ et $V_b(t)$ indiquent les volumes disponibles au prix de l'Ask et du Bid,

La quantité $I(t)$ peut être interprétée comme la pression exercée par les teneurs de marché sur le prix, en ce sens que lorsque $I(t) \geq 0$, plus de commandes à cours limité ont été soumises et non annulées du côté des achats, et les prix devraient augmenter. Quand $I(t) \leq 0$, plus de commandes à cours limité ont été soumises du côté des ventes et les prix devraient baisser. Nous pouvons confirmer quantitativement cette intuition en considérant l'impact prix d'un déséquilibre, à savoir :

$$R(l|I) = E[x_{x+l} - x_t | I(t) = I]$$

qui est la variation de prix attendue après l transactions étant donné qu'un déséquilibre a été observé à t, Si x_t était efficace, tous les prédicteurs de prix construits avec des données publiques n'auraient aucun impact: en particulier, la fonction de réponse conditionnée par le déséquilibre de file d'attente I disparaîtrait, $R(l|I) = 0$. Ceci contraste toutefois fortement avec les observations empiriques: La figure ci-dessous montre $R(l|I)$ en fonction de l pour différents intervalles de valeurs de I. Dès que $I \neq 0$, le changement de prix anticipé suivant est non nul. Fait intéressant, pour les grands $|I|$ la variation de prix absolue attendue dépasse un demi-tick. Nous trouvons également empiriquement que $R(\infty|I = 1) = -R(\infty|I = -1)$ est approximativement égal à l'impact absolu moyen d'une déplétion sur le prix moyen, noté $r_\infty(t)$.

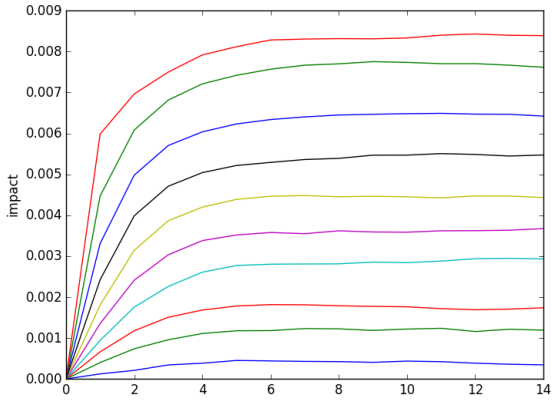


Figure 20: Impact symétrisé du déséquilibre pour les différentes valeurs de $|I(t)| \in [k \times 0.1, (k+1) \times 0.1)$ pour k=1,2,3,....,9

3.3.2 Covariance des rendements

Nous avons montré que le prix moyen n'est pas efficace dans le cas d'un tick large car il n'intègre pas l'information véhiculée par le déséquilibre de volume. Lorsque la taille du tick est faible, le déséquilibre du volume perd sa puissance prédictive. Un moyen direct d'évaluer l'efficacité du prix moyen dans les petits segments de marché est de considérer la fonction de covariance empirique des rendements au prix moyen.

Le prix moyen n'est pas efficace, que ce soit en petit tick ou en large tick dans un carnet d'ordre, car nous observons que la fonction de covariance de lag 1 est dans la plupart des cas négative, c'est-à-dire que la moyenne des prix revient après un rendement non nul.

Dans cette section, nous montrons que la relation MRR $x_{t+1} - x_t = G[e_t - \hat{e}_t] + W_t(*)$ nous permet de prévoir la covariance des rendements au prix moyen pour les petits ticks. Mais l'espérance mathématique dans l'équation correspondante $E_t[x_{t+l} - x_t] = GE_t[e_t - \hat{e}_{t+1+l}]$ de notre modèle pour les grands ticks nous empêche de prévoir la covariance des rendements au prix moyen dans les stocks de ticks larges et nous allons l'observer explicitement de manière empirique. Lorsque les teneurs de marché obtiennent des remises, établissent des devis selon $a_t = p_t + G[1 - \hat{e}_t] - r$ et $b_t = p_t + G[-1 - \hat{e}_t] + r$ encore le prix moyen $x_t = \frac{1}{2}(b_t + a_t)$ n'est pas affecté par r et on peut montrer d'après l'équation (*) que :

$$cov_x(l) = G^2(C(l) - C(l-1) - E[e_{t+l}\hat{e}_{t+1}] + E[e_{t+l+1}\hat{e}_1]) + E[e_{t+l+1}\hat{e}_{t+1}] + GE[e_{t+l}W_t] - GE[e_{t+l+1}W_t]$$

Il est important de noter que nous ne supposons pas que le flux de commandes futures n'est pas corrélé avec W_t. En fait, il est intuitivement clair que les

participants au marché réagissent aux rendements passés.

Malgré cette difficulté, il est possible d'obtenir une prédiction non paramétrique de la covariance des rendements à prix moyen. Nous définissons la fonction de réponse décalée du signe :

$$R_k(l) = E[e_{t+k}(x_{t+l} - x_t]$$

$R_k(l)$ est facilement mesurable sur des données empiriques. Étonnamment, en utilisant l'équation MRR (*), il est possible d'exprimer la covariance des rendements des prix moyens, de la manière suivante :

$$cov_x(l) = G[R_l - R_{l+1}(1)] = R(\infty)[R_l(1) - Rl + 1(1)]$$

Cette relation est un test subtil de l'équation (*), ce qui est valable quelles que soient les corrélations de la dynamique des prix (c'est-à-dire le bruit W_t) avec le flux de commandes futures.

3.4 Estimation du prix fondamental pour un tick valant 0.005

L'objectif des auteurs était de mettre en place un estimateur du prix fondamental pour un tick large, en n'utilisant que les données publiques et disponibles concernant les carnets d'ordre. Dans cette section, les auteurs introduisent un estimateur de prix fondamental pour un tick large, en utilisant les carrés des volumes aux meilleurs prix :

$$\hat{p}_t = \frac{V_a^2(b_t - r) + V_b^2(a_t + r)}{V_a^2(t) + V_b^2(t)}$$

Cet estimateur a plusieurs propriétés, en particulier, il est facile à calculer, et il peut prendre des valeurs en dehors de l'intervalle défini par les prix de l'Ask et du Bid. En outre, dans le cas d'un carnet d'ordre équilibré, $V_a = V_b$, l'estimateur coïncide avec le prix moyen. On a comparé les performances de cet estimateur avec celles d'autres estimateurs :

$$\hat{p}'_t = \frac{V_a(t)(b_t - r) + V_b(t)(a_t - r)}{V_a(t) + V_b(t)}$$

$$\hat{p}''_t = \frac{V_a(t)b_t + V_b(t)a_t}{V_a(t) + V_b(t)}$$

En montrant que les performances de l'etimateur $\hat{p}_t$ sont mieux que les performances des autres estimateurs, on justifie l'adoption de $\hat{p}_t$ pour notre modèle.

3.4.1 Comparaison des performances des différents estimateurs

On commence tout d'abord par étudier l'impact du déséquilibre sur le prix, afin de déterminer la proportion de l'information concernant le déséquilibre. Pour cela, on a calculé l'impact du déséquilibre sur $\hat{p}_t$, en utilisant une remise $r = 0.1 \times tick = 0.0005$. On obtient donc la courbe suivante :

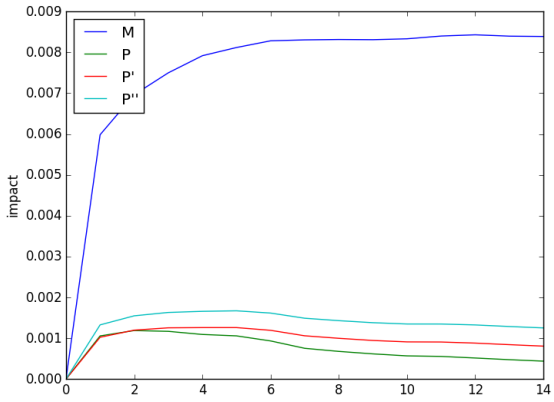


Figure 21: L'impact symétrisé du déséquilibre sur le prix moyen, $\hat{p}_t$, $\hat{p}_t$' et $\hat{p}_t$"

On constate que le quotient de l'impact du déséquilibre entre $\hat{p}_t$ et le prix moyen vaut 6.25%, on interprète la partie manquante comme l'information incluse dans $\hat{p}_t$. Donc on a plus que 93% de l'information qui est transmise à $\hat{p}_t$ par le déséquilibre. De même, on constate que l'information transmise à $\hat{p}_t$' par le déséquilibre, est aux environs de 90%. En ce qui concerne $\hat{p}_t$" l'information incluse est inférieure par rapport aux estimateurs précédents. Donc, on peut conclure que les estimateurs tenant compte de la remise sont plus performants que les autres.

En outre, $\hat{p}_t$ se comporte approximativement comme un prix efficace, dans le cadre du modèle de MRR, vu qu'il satisfait approximativement l'équation qui définit le modèle

$$p_{t+1} - p_t = G[e_t - \hat{e}_t] + W_t$$

On définit la fonction retardée de la transaction de $\hat{p}_t$:

$$R^{\hat{p}}(l) = E[e_t(\hat{p}_{t+l} - \hat{p}_t)]$$

On trace dans la courbe ci-dessous, la courbe de $R^{\hat{p}}(l), R^{\hat{p}'}(l), R^{\hat{p}''}(l), R(l)$. On constate que l'impact permanent de $\hat{p}_t$ est deux fois plus petit que l'impact permanent du prix moyen, on peut interpréter cette différence par le fait que $\hat{p}_t$ inclut une grande partie des corrélations des flux d'ordre passés et futurs. En plus, on peut s'intéresser à la différence des fonctions de réponses entre la

valeur 1 et l. On remarque que la différence la plus petite entre $R(\infty)$ et $R(1)$ est dans le cas de l'estimateur $\hat{p}_t$, ce qui montre bien qu'il est plus performant par rapport $\hat{p}_t$' et $\hat{p}_t$". En effet, plus cette différence est faible plus l'estimateur est proche du prix fondamental qui lui vérifie $R(\infty) = R(l)$ pour tout $l \geq 1$

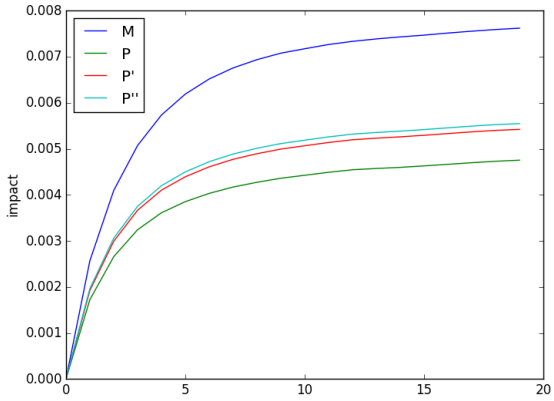


Figure 22: Les fonctions de réponses des transactions $R^{\hat{p}}(l), R^{\hat{p}'}(l), R^{\hat{p}''}(l), R(l)$

En se basant sur ce qui précède, il est plus légitime de choisir l'estimateur $\hat{p}_t$.

3.5 Comparaison de l'estimateur $\hat{p}_t$ avec le micro prix du premier article

Nous souhaitons, dès le début de notre projet, comparer les différentes définitions de micro-prix. Nous avons donc décider de comparer $\hat{p}_t$ avec la fonction G^* définie dans le premier article, en utilisant plusieurs espérances empiriques faisant intervenir $\hat{p}_t$. La comparaison que nous avons adoptée est entre G^* et $E[\hat{p}_{t+l} - M_t]$, pour de différentes valeurs de l. Ces deux quantités sont calculées en utilisant les temps de transactions. On obtient les courbes suivantes :

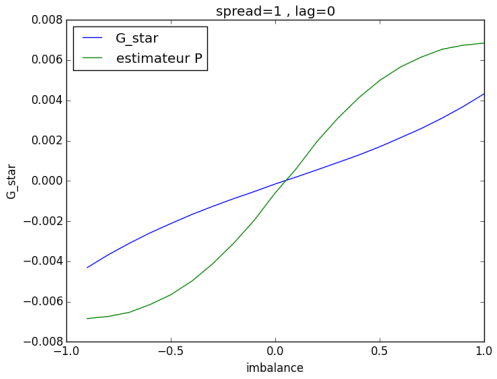


Figure 23: G^* et $E[\hat{p}_{t+l} - M_t]$ avec l=0

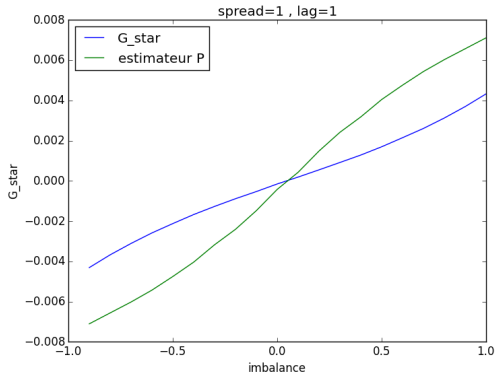


Figure 24: G^* et $E[\hat{p}_{t+l} - M_t]$ avec l=1

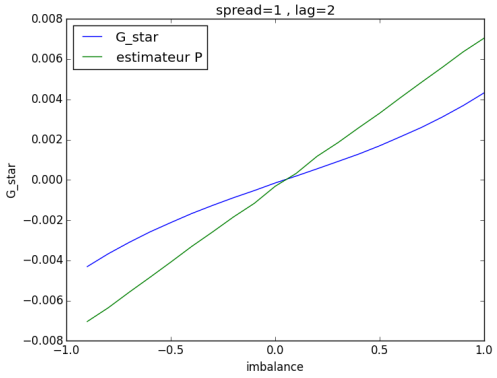


Figure 25: G^* et $E[\hat{p}_{t+l} - M_t]$ avec l=2

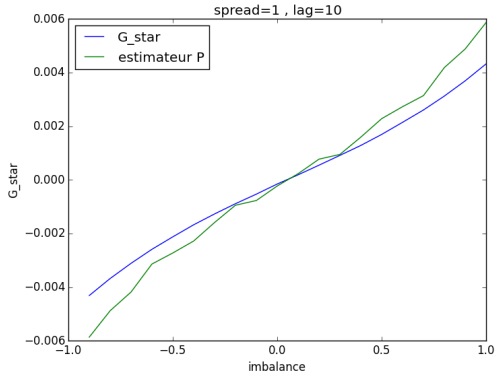


Figure 26: G^* et $E[\hat{p}_{t+l} - M_t]$ avec l=10

Nous remarquons que lorsque le pas l augmente, les deux courbes de G^* et $E[\hat{p}_{t+l} - M_t]$ se rapprochent. A titre d'exemple, on obtient deux courbes assez proches pour l=10. Ce constat est compatible avec ce que nous avons remarqué dans le premier article, quand le pas augmente, les courbes des moyennes empiriques du prix moyen et du prix moyen pondéré rejoingnent la courbe de G^*.

Nous pouvons expliquer cette similarité, par le fait que l'estimateur $\hat{p}$ ainsi que les autres estimateurs empiriques introduits dans le premier article, retirent la totalité de l'information sur l'actif, à travers les volumes de l'Ask et du Bid.

Donc le micro prix défini dans le premier article, à un grand pas de temps, constitue un bon prédicteur du $\hat{p}$.

4 Conclusion

Dans la première moitié de notre projet innovation, nous avons étudié l'article de Monsieur **Sasha Stoikov**, qui nous a été proposé par notre encadrant Monsieur **Ioane Muni Toke**. Cette étude nous a permis de découvrir la notion du micro-prix, sous l'angle de l'auteur. Nous avions pu montrer en calculant du prix de microstructure pour un carnet d'ordre pour le stock BNPP.PA au long du mois de janvier 2015, que celle-ci converge rapidement et a la même allure que celle obtenue dans l'article. Pour les valeurs de spread 1,2,3 et 4 (qui figurent le plus dans le carnet d'ordre), on retrouve des courbes lisses et qui s'aplatissent plus le spread est plus grand, tant dit que pour des valeurs de spread plus rares, ces deux propriétés ne sont pas nécessairement vérifiées. Les courbes des espérances empiriques $E[M_{t+\delta} - M_t|I_t]$ et $E[W_{t+\delta} - M_t|I_t]$ sont assez proche de la courbe de G_*, et on obtient, pour un spread=1, que le micro-prix est un meilleur estimateur du prix moyen pondéré que le prix moyen. Pour des spreads plus grands, on trouve le contraire, le micro prix estime mieux le prix moyen. Dans un second temps, nous avons étudié l'article de Monsieur **Julius Bonart** et Monsieur **Fabrizio Lillo** qui présente une généralisation du modèle MRR pour le cas de large tick. Cet article compare notamment le prix moyen à trois estimateurs du prix fondamental pour conclure que $\hat{p}_t$ est le meilleur des quatre car il inclut 80% de l'information sur le déséquilibre. Nous avions pu vérifier et tracer la majorité des résultats et des figures présentées par l'article en utilisant les données sur les ordres de marché du stock BNPP.PA au long du mois de janvier 2015. Cette action est considérée comme ayant un tick moyen d'après la définition introduite dans l'article, mais elle est tout de même proche du domaine du large tick, donc son utilisation pour vérifier les résultats est pertinente. On retrouve notamment le fait que l'estimateur $\hat{p}_t$ a le plus faible impact du déséquilibre ce qui signifie qu'il incorpore le plus d'informations sur celle-ci et qu'il est ainsi un meilleur estimateur du prix fondamental.

Dans la dernière partie, nous avons comparé le micro prix défini à travers le premier article et l'estimateur $\hat{p}_t$ introduit dans le deuxième article. Nous avons pu constater que les deux courbes de G^* et $E[\hat{p}_{t+l} - M_t]$ sont plus en plus confondues, lorsque le pas l augmente.

Les deux définitions de prix que nous avons abordées ne sont pas uniques. L'objectif de notre projet était d'étudier plusieurs définitions de micro prix, nous avons étudié deux articles, qui mettent en place deux approches de calcul du micro prix, ainsi que les tests et les calculs permettant de valider les modèles. Nous estimons que nous avons réussi à atteindre plusieurs de nos objectifs dans le cadre de ce projet.

Nous tenons à remercier notre encadrant et le responsable de notre projet Monsieur **Ioane Muni Toke**, pour son encadrement, ses aides et ses explications, tout au long du semestre.

5 Références

. [1] **Sasha Stoikov**, ***The Micro-Price:*** *A High Frequency Estimator of Future Prices*, *November 2017*

[2] **Julius Bonart** et **Fabrizio Lillo**, ***A continuous and efficient fundamental price on the discrete order book grid***, *November 2016*