École Centrale Lyon

MIR
Projet d'option
Rapport

Dynamiques hautes fréquences dans le carnet d'ordre

Élèves :
Ryan Houcine
Marvin-Evans Kameni de Djani

Enseignants :
Christian De Peretti
Matthieu Bonnivard

13 novembre 2024

Table des matières

1 Introduction

Les marchés financiers ont connu une transformation significative ces dernières décennies, marquée par l'émergence des opérations à haute fréquence (HFT). Ces transactions ultra-rapides, caractérisées par l'exécution d'ordres en fractions de seconde, ont profondément modifié la structure et le fonctionnement des marchés, introduisant de nouveaux défis pour les acteurs du marché et les chercheurs en finance quantitative.

Au cœur de cette révolution se trouve le carnet d'ordres à cours limité, un outil central pour comprendre les dynamiques des marchés financiers. Le carnet d'ordres enregistre et affiche tous les ordres d'achat et de vente en attente pour un actif financier donné, fournissant une image en temps réel de la demande et de l'offre sur le marché. L'analyse des dynamiques du carnet d'ordres à haute fréquence est cruciale pour anticiper les variations de prix, comprendre les comportements des acteurs du marché et gérer les risques de manière efficace.

Pourtant, malgré l'importance croissante des dynamiques haute fréquence dans le carnet d'ordres, la modélisation précise et robuste de ces phénomènes reste un défi de taille.

1.1 Présentation du carnet d'ordres

Le carnet d'ordre est un système dynamique discret qui caractérise l'état de l'offre et de la demande pour un produit financier quelconque s'échangeant sur les marchés organisés. Du côté de la demande (Bid) on retrouve en ordonnée la quantité de produit demandée à un certain prix affiché en abscisses. Symétriquement on retrouve la quantité de produit proposée à l'offre (Ask). Le mid price est la moyenne entre le bid et l'ask, et le spread la différence entre ces deux valeurs.

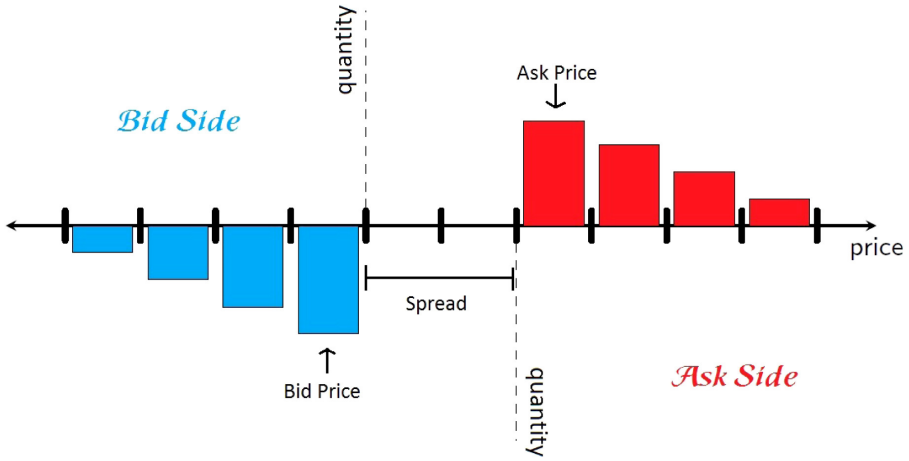


FIGURE 1 – Représentation graphique du Limit Order Book [2]

Il existe trois types d'interaction avec le marché pour ce carnet d'ordre :

— **Ordre marché (market order) :** Les acteurs du marché vont soustraire une certaine quantité disponible au bid ou à l'ask. Par exemple, une entreprise va acheter une obligation au prix le plus bas disponible du côté de l'ask, et la quantité

d'obligation disponible à ce prix là va diminuer d'une unité. Ce genre d'ordre est exécuté immédiatement.

— **Ordre limite (limit order) :** Les acteurs vont placer des offres d'achat ou de vente à un prix différent du meilleur prix disponible. En effet, un acheteur désirant se procurer l'obligation précédente à un prix plus attractif pourra placer une offre d'achat au meilleur bid (le bid le plus haut) sans avoir à sacrifier le spread dans son opération. Il prend le risque que son ordre ne soit pas exécuté car il se positionne alors en dernier dans la file d'attente à ce prix là. Cet ordre ajoute une quantité positive dans l'order book.

— **Annulation (cancel order) :** Tout simplement une annulation de limit order. Un acteur peut, tant que son ordre n'est pas exécuté (c'est à dire qu'il ne trouve pas de contrepartie désirant acheter ou vendre au prix qu'il aura fixé) annuler son offre de vente ou d'achat.

Ainsi, après avoir examiné les interactions fondamentales avec le carnet d'ordres, nous allons maintenant explorer différentes approches de modélisation pour mieux comprendre et analyser les dynamiques du marché financier à travers les modèles suivants.

2 État de l'art

Dans cette étude, nous nous proposons d'explorer l'état de l'art des modèles de dynamiques haute fréquence dans les carnets d'ordres, en mettant en lumière les défis et les lacunes actuelles.

Pour notre étude, nous avons fait le choix délibéré de nous focaliser sur les modèles basés sur les processus stochastiques. Ainsi, nous mettons volontairement de côté les modèles de réseaux de neurones, les machines à support de vecteur et autres. Les modèles stochastiques sont bien adaptés pour capturer les événements à haute fréquence dans le carnet d'ordres, tels que les changements rapides de prix et les exécutions d'ordres instantanées.

Nous nous concentrerons sur les questions suivantes :

1. **Capture des phénomènes de microstructure :** Comment les modèles existants parviennent-ils à modéliser la formation des prix, la propagation des ordres et les effets de latence dans les carnets d'ordres à haute fréquence ? Dans quelle mesure ces modèles reflètent-ils fidèlement la réalité des transactions ultra-rapides sur les marchés financiers ?

2. **Volatilité, liquidité et comportement des acteurs du marché :** Comment les modèles existants capturent-ils la volatilité et la liquidité des marchés à haute fréquence ? Dans quelle mesure ces modèles intègrent-ils les comportements des acteurs du marché, tels que les stratégies de trading algorithmique et les réactions

aux événements externes ?

En adressant ces questions, cet état de l'art vise à identifier les défis clés de la modélisation des dynamiques haute fréquence dans les carnets d'ordres et à proposer des pistes d'amélioration pour une meilleure compréhension et gestion des marchés financiers à l'ère de la haute fréquence.

2.1 Le modèle zéro-intelligence

2.1.1 L'intérêt de la modélisation zero-intelligence

Les marchés financiers contemporains sont des environnements complexes et dynamiques, où une multitude d'acteurs interagissent dans le but de prendre des décisions d'investissement et de trading. Dans ce contexte, la compréhension des mécanismes sous-jacents qui régissent les mouvements des prix et l'allocation des ressources est d'une importance capitale pour les investisseurs, les chercheurs et les régulateurs.

L'une des approches les plus fascinantes pour étudier les dynamiques des marchés est l'utilisation de modèles de Carnet d'Ordres à Zero-Intelligence. Ces modèles simplifiés permettent de simuler le comportement des traders sans recourir à des hypothèses complexes sur leur rationalité ou leurs stratégies de trading. Au lieu de cela, ils se fondent sur des règles simples et aléatoires pour générer des ordres d'achat et de vente, reproduisant ainsi les interactions fondamentales qui se produisent sur les marchés réels.

L'étude des modèles Zero-Intelligence de Carnet d'Ordres revêt un intérêt particulier pour plusieurs raisons. Tout d'abord, ces modèles fournissent un cadre théorique élégant pour analyser les propriétés émergentes des marchés financiers, telles que la formation des prix, la liquidité et la volatilité. En simulant le comportement des traders de manière simplifiée, ils permettent d'explorer comment les interactions microscopiques entre les différents ordres influencent les tendances macroscopiques du marché.

De plus, l'étude des modèles Zero-Intelligence de Carnet d'Ordres offre des perspectives précieuses sur la capacité des marchés à atteindre une efficience allocative, même en l'absence de rationalité individuelle chez les traders. En observant comment ces modèles parviennent à converger vers des équilibres de marché malgré le comportement aléatoire des acteurs, nous pouvons mieux comprendre les mécanismes qui sous-tendent l'efficacité des marchés financiers dans la réalité.

Enfin, l'analyse des modèles Zero-Intelligence de Carnet d'Ordres peut également fournir des insights pratiques pour les traders et les gestionnaires de portefeuille. En comprenant comment les variations dans la composition et le comportement des ordres peuvent influencer les prix et les rendements, les professionnels de la finance peuvent mieux anticiper les mouvements du marché et prendre des décisions plus éclairées.

2.1.2 Les principes du modèle

Le modèle de Zero Intelligence (ZI) est une approche simplifiée utilisée pour simuler le comportement des échanges et la formation des prix sur les marchés financiers, en particulier dans le contexte de la dynamique des carnets d'ordres. Ses principes reposent sur l'idée de simplicité et de hasard, visant à capturer les interactions de base entre les traders sans supposer de stratégies de trading spécifiques ou de processus de prise de décision rationnelle. Le modèle de Zero Intelligence (ZI) repose sur plusieurs principes fondamentaux qui définissent son fonctionnement et ses hypothèses [1]. Tout d'abord, il suppose que les traders passent des ordres de manière aléatoire, sans stratégie prédéfinie ni préférences spécifiques. Cette approche vise à reproduire l'absence de comportement stratégique chez les traders. De plus, le modèle garantit une probabilité égale pour tous les traders de passer des ordres d'achat ou de vente à tout moment, assurant ainsi une représentation équilibrée de la pression d'achat et de vente sur le marché. Les ordres sont générés indépendamment les uns des autres et sont placés à des prix aléatoires, tandis que la correspondance des ordres se fait selon une règle de priorité prix-temps. En fin de compte, les prix sont déterminés par l'intersection de l'offre et de la demande, sans considération pour les facteurs économiques, ce qui permet de simuler le trading continu et d'explorer la dynamique des carnets d'ordres et la formation des prix.

Dans l'ensemble, le modèle de ZI offre un cadre simple mais puissant pour étudier les marchés financiers, en permettant de comprendre l'impact des interactions aléatoires entre les traders sur la microstructure du marché et l'efficience des prix. Bien que ses hypothèses puissent être simplistes, le modèle reste précieux pour explorer les concepts fondamentaux des marchés financiers et analyser le rôle du hasard dans les échanges.

2.1.3 Formalisme mathématique

Le modèle de Daniels et al. [4] suppose une enchère double continue, qui représente la méthode la plus couramment utilisée pour la formation des prix sur les marchés financiers modernes. Dans ce cadre, il existe deux types fondamentaux d'ordres de trading : les traders impatients passent des ordres au marché pour acheter ou vendre immédiatement un nombre défini d'actions au meilleur prix disponible, tandis que les traders plus patients soumettent des ordres limites qui incluent le prix le moins favorable pour la transaction [5]. Ces ordres limites peuvent ne pas être exécutés immédiatement et sont stockés dans une file d'attente appelée carnet d'ordres limites. Lorsque des ordres d'achat arrivent, ils sont exécutés contre des ordres de vente limites accumulés ayant un prix de vente inférieur, en fonction de la priorité de prix et de l'heure d'arrivée. Le processus est similaire pour les ordres de vente. Les prix bid et ask sont déterminés par les ordres limites disponibles dans le carnet d'ordres, et ils évoluent en fonction des nouveaux ordres arrivant sur le marché.

Le modèle tel que décrit par Eric Smith et al. [7] suppose que deux types d'agents de zéro intelligence placent et annulent des ordres de manière aléatoire. Les agents impatients passent des ordres au marché de taille σ, qui arrivent à un taux de μ actions par unité de temps. Les agents patients placent des ordres limites de même taille σ, qui arrivent avec une densité de taux constante α actions par prix par unité de temps. Ces agents peuvent être considérés comme des fournisseurs et des demandeurs de liquidité. Les ordres limites en attente sont annulés à un taux constant δ, avec des dimensions

de 1/temps. Les prix changent par incréments discrets appelés ticks, de taille dp. Pour maintenir le modèle aussi simple que possible, il existe des taux égaux pour l'achat et la vente, et le placement et l'annulation d'ordres sont des processus de Poisson. Tous ces processus sont indépendants, sauf pour le couplage à travers leurs conditions aux limites : les ordres limites d'achat arrivent avec une densité constante α sur l'intervalle semi-infini $-\infty < p < a(t)$, où p est le logarithme du prix, et les ordres limites de vente arrivent avec une densité constante α sur l'intervalle semi-infini $b(t) < p < \infty$. En conséquence des processus aléatoires d'arrivée des ordres, $a(t)$ et $b(t)$ effectuent chacun une marche aléatoire, mais en raison du couplage des processus d'achat et de vente, l'écart entre les prix bid et ask $s(t) \equiv a(t) - b(t)$ est une variable aléatoire stationnaire. À mesure que de nouveaux ordres arrivent, ils peuvent modifier les meilleurs prix $a(t)$ et $b(t)$, ce qui modifie à son tour les conditions aux limites pour le placement ultérieur d'ordres limites. Par exemple, l'arrivée d'un ordre limite d'achat à l'intérieur de l'écart modifiera le meilleur bid $b(t)$, ce qui modifie immédiatement la condition aux limites pour le placement d'ordres limites de vente. C'est cette rétroaction entre le placement d'ordres et la diffusion des prix qui rend ce modèle intéressant, et malgré sa simplicité apparente, assez difficile à comprendre analytiquement.

L'article de J. Doyne Farmer [5] nous donne une relation entre le *spread* moyen $\hat{s}$ et :
- la taille σ des ordres placés (qui est la même pour les *market orders* et pour les *limit orders*),
- la densité de taux constante des *limit orders* α en actions par prix par unité de temps,
- le taux de *market orders* μ en actions par unité de temps,
- le taux constant d'annulation des ordres δ
- et le ratio de suppression des ordres par annulation sur celui de suppression par *market order* : $\epsilon = \sigma\delta/\mu$.

La valeur moyenne du *spread* prédit sur la base d'une analyse de la théorie du champ moyen du modèle [7] dans la limite où la taille des *ticks* dp (en supposant que les prix varient discrètement selon ce pas d'incrémentation) tend vers 0 :

$$\hat{s} = (\mu/\alpha) f(\sigma\delta/\mu) \qquad \hat{D} = k\mu^{5/2}\delta^{1/2}\sigma^{-1/2}\alpha^{-2}$$

où k est une constante et f est estimée numériquement à $f(\epsilon) = 0.28 + 1.86\epsilon^{3/4}$.

2.1.4 Limites du modèle

Le modèle de zéro intelligence de l'order book présente plusieurs limites importantes à prendre en compte lors de son utilisation dans l'analyse des marchés financiers.

Tout d'abord, sa simplification excessive peut ne pas refléter de manière précise le comportement réel des participants sur le marché, en particulier dans des situations où des stratégies sophistiquées ou des comportements irrationnels sont présents comme dans les périodes de crise financière.

De plus, le modèle ne prend pas en compte les aspects de la dynamique des ordres tels que la corrélation entre les ordres d'achat et de vente, les réactions des traders aux informations du marché, ou les effets de réseau entre les différents acteurs du marché. En outre, le modèle suppose souvent une symétrie parfaite entre les acheteurs et les vendeurs,

ce qui peut ne pas correspondre à la réalité des marchés financiers. Enfin, l'absence de paramètres de calibration spécifiques peut rendre difficile la validation empirique du modèle et son application dans des contextes réels.

2.2 Les processus de Hawkes

2.2.1 Présentation du modèle

La principale différence entre le modèle de zéro intelligence (ZI) et les processus de Hawkes réside dans leur approche de modélisation des dynamiques temporelles. Alors que le modèle ZI considère les événements comme étant indépendants les uns des autres, les processus de Hawkes capturent les dépendances temporelles entre les événements, ce qui permet une modélisation plus réaliste des phénomènes d'auto-excitation et d'auto-inhibition.

Dans cette perspective, les processus de Hawkes représentent une classe importante de modèles stochastiques utilisés pour décrire et analyser ces phénomènes dans les événements ponctuels. Originellement développés dans le domaine de la séismologie pour modéliser les répliques sismiques, ces processus ont trouvé des applications étendues dans divers domaines, y compris la finance. Ils offrent un cadre mathématique puissant pour modéliser les dépendances temporelles entre les événements, ce qui les rend particulièrement pertinents pour l'analyse des données financières à haute fréquence.

Dans cette section, nous introduirons les principes fondamentaux des processus de Hawkes, en expliquant leur structure, leurs propriétés et leurs applications dans le domaine financier. Nous examinerons également les défis associés à l'utilisation de ces modèles et les limites associées.

2.2.2 Les fondements théoriques

Les processus de Hawkes sont des modèles stochastiques utilisés pour modéliser les événements ponctuels, caractérisés par leur auto-excitation et leur auto-inhibition. Les caractéristiques fondamentales de ces processus comprennent l'intensité ponctuelle et la fonction de noyau.

L'intensité ponctuelle d'un processus de Hawkes, notée $\lambda(t)$, représente la probabilité qu'un événement se produise à l'instant t. Elle est définie comme la somme des contributions de tous les événements passés, modulée par une fonction de noyau :

$$\lambda(t) = \mu + \sum_{t_i<t} \phi(t - t_i),$$

où μ est l'intensité de base du processus et $\phi(\cdot)$ est la fonction de noyau.

La fonction de noyau $\phi(t)$ mesure l'influence des événements passés sur l'intensité future du processus. Elle décroît généralement avec le temps écoulé depuis l'événement précédent. Une fonction de noyau typique est la fonction exponentielle décroissante :

$$\phi(t) = \alpha e^{-\beta t},$$

où α et β sont des paramètres de contrôle qui déterminent la force et la rapidité de décroissance de l'influence.

Les processus de Hawkes présentent une propriété d'auto-excitation et d'auto-inhibition, ce qui signifie que la probabilité d'occurrence d'un événement est influencée par les événements passés. L'auto-excitation se produit lorsque la survenue d'un événement accroît la probabilité d'occurrence d'événements futurs similaires, tandis que l'auto-inhibition se produit lorsque la survenue d'un événement diminue cette probabilité.

2.2.3 Limites du modèle

Le modèle de Hawkes, malgré ses avantages, comporte certaines limites. Tout d'abord, il repose souvent sur des structures paramétriques fixes, ce qui peut ne pas refléter pleinement la complexité des dynamiques observées sur les marchés financiers. De plus, l'estimation des paramètres peut être difficile, en particulier avec des données bruitées ou des processus multivariés. Le modèle peut également avoir du mal à capturer les événements rares ou extrêmes et à modéliser les effets de mémoire longue. En résumé, bien que le modèle de Hawkes soit utile, il convient de reconnaître ses limites et de compléter son utilisation avec d'autres approches pour une meilleure compréhension de la microstructure du marché.

2.3 Conclusion de l'état de l'art

Les modèles examinés dans l'état de l'art abordent efficacement la capture des phénomènes de microstructure dans les carnets d'ordres à haute fréquence. Le modèle ZI, en supposant un comportement aléatoire des traders et une correspondance simple des ordres, parvient à modéliser la formation des prix et la propagation des ordres dans un environnement de marché dynamique. Cependant, sa capacité à refléter fidèlement la réalité des transactions ultra-rapides peut être limitée par son approche simplifiée, qui ne prend pas en compte les stratégies de trading sophistiquées utilisées par les acteurs du marché réel.

D'autre part, les Processus de Hawkes offrent une approche plus sophistiquée en capturant les interactions endogènes entre les événements dans le temps, ce qui permet de modéliser les effets de latence et les réactions en cascade dans les carnets d'ordres. Cette approche permet une meilleure représentation de la dynamique temporelle des transactions à haute fréquence. Cependant, malgré cette sophistication, les Processus de Hawkes peuvent encore être confrontés à des défis pour modéliser pleinement la volatilité et la liquidité des marchés à haute fréquence, ainsi que pour intégrer de manière exhaustive les comportements complexes des acteurs du marché, tels que les stratégies de trading algorithmique et les réactions aux événements externes.

3 Introduction aux séries temporelles

En étudiant l'impact sur les prix des événements du carnet d'ordres, Rama Cont montre en 2011 que l'*order flow imbalance* est le principal facteur qui gouverne les chan-

gements de prix [3].

Dans cette section, nous nous concentrons sur l'analyse de trois aspects clés du marché financier : le déséquilibre du carnet d'ordres (*order book imbalance*, le déséquilibre du flux d'ordres *order flow imbalance* et la réversion à la moyenne (*mean reversion*). Ces facteurs sont essentiels pour comprendre la dynamique des marchés financiers et peuvent jouer un rôle crucial dans la prise de décision des traders et des investisseurs. En examinant ces aspects, nous visons à approfondir notre compréhension des mécanismes sous-jacents qui influent sur les fluctuations des prix et à explorer les implications potentielles pour les stratégies de trading et d'investissement.

3.1 Les données

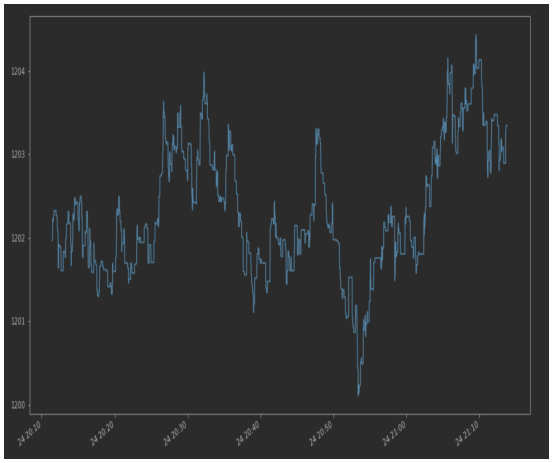
Dans le cadre de notre étude, nous utilisons des données d'évolution du carnet d'ordres (*limit order book*) du Bitcoin (BTC) provenant directement du site Binance, l'une des principales plateformes d'échange de cryptomonnaies. Ces données fournissent des informations telles que le meilleur prix d'achat (best bid), le meilleur prix de vente (best ask), la taille des ordres d'achat (bid size) et la taille des ordres de vente (ask size) à des intervalles de temps réguliers, généralement toutes les 200-300 millisecondes. Ces données représentent une source précieuse pour comprendre la dynamique du marché et explorer les interactions entre les différents acteurs sur la plateforme d'échange. Nous avons uniquement récupéré la première couche du bid/ask (le *best ask* et le *best bid*).

3.1.1 Visualisation des données

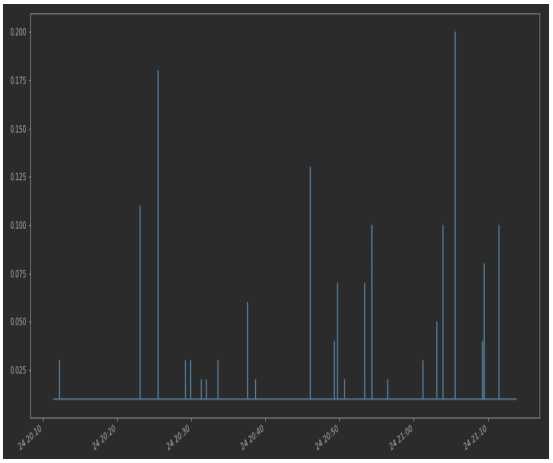
Les données présentes dans l'order book sont modifiées afin d'obtenir une matrice qui nous permettra d'effectuer tous les calculs nécessaires. Nous introduisons :

— Le mid qui est le prix moyen entre le ask et le bid.
— eA et eB qui sont les variations de volume au bid et au ask entre deux timestamps.
— Le spread, qui est la différence entre le prix à l'ask et le prix au bid.
— L'average depth, qui est la moyenne glissante des volume à l'ask et au bid.

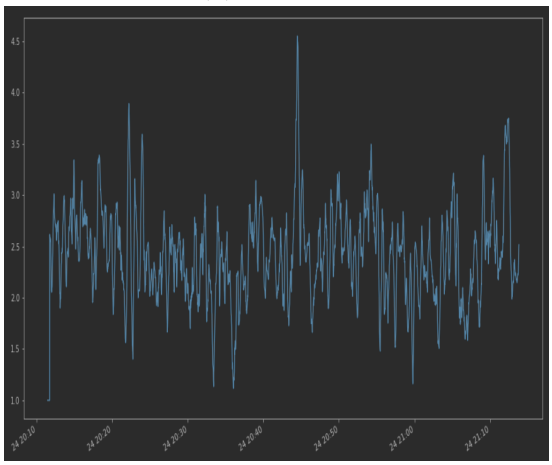
Ces données sont tracées sur les figures suivantes :



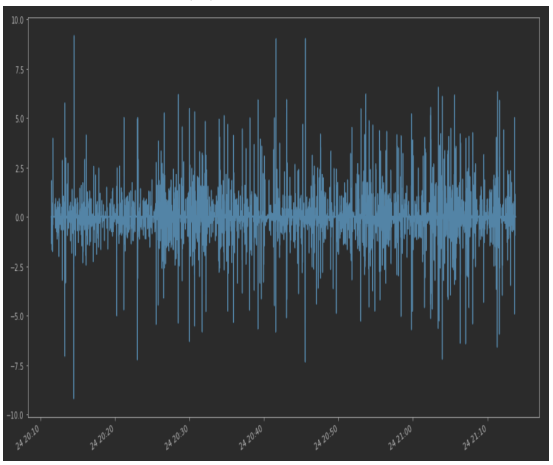
(a) Prix du BTC



(b) Spread du BTC



(c) Profondeur des couches limites



(d) Variation des volumes à l'ask

Comme on peut le voir, ces séries temporelles ne semblent pas être stationnaire. Pour les séries financières non transformées, il est impossible de calculer une espérance et une variance de manière consistante pour les modéliser. En effet, dans la figure (a) et la figure (c) on voit très bien que l'espérance et la variance ne sont pas constantes dans le temps. Il en est de même pour les séries du spread et des variations de volume à l'ask. On peut néanmoins trouver des relations de co-intégration et d'autres modèles plus compliqués pour décrire les comportement économétriques de ces séries. Le but du prochain paragraphe est de décrire des features permettant de modéliser, à l'aide de séries stationnaires, les variations futures du prix du Bitcoin. Ces features sont issues de la description physique d'un carnet d'ordre ont un sens logique et ne sont pas des artefacts dues aux données de marché.

Les stationnarités des séries précédentes ont été testées à l'aide de la fonction adfuller de la bibliothèque statmodel qui réalise un test de l'hypothèse de non stationnarité des séries précédentes. Ces tests sont fournis en annexe (voir notebook section test de stationnarité) et rejettent tous l'hypothèse de stationnarité des séries.

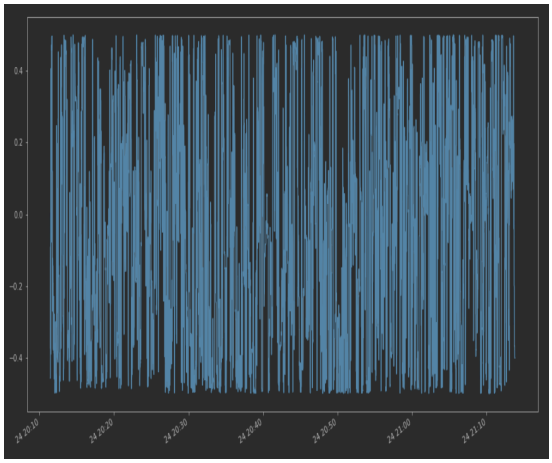
3.2 Order Book Imbalance

Le déséquilibre du carnet d'ordres (OBI) est une mesure cruciale dans l'analyse du marché financier, offrant des indices sur les pressions d'achat et de vente. Cette mesure évalue la disparité entre les quantités d'ordres d'achat (*bids*) et de vente (*asks*) présentes dans le carnet d'ordres à un instant donné. Une imbalance positive indique une prédominance des ordres d'achat, tandis qu'une imbalance négative signale une prédominance des ordres de vente. Elle s'exprime comme suit :

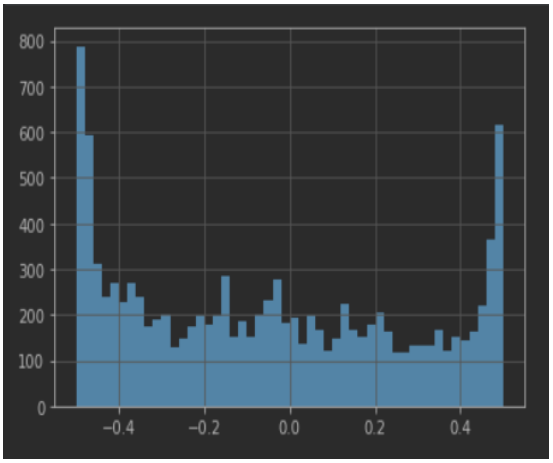
$$OBI_t = \frac{V_t^b - V_t^a}{(V_t^b + V_t^a) \times spread_t} \tag{1}$$

Les traders et les analystes utilisent souvent l'OBI comme un indicateur supplémentaire pour confirmer ou infirmer les signaux provenant d'autres analyses techniques. Une imbalance croissante peut signaler un afflux d'ordres dans une direction particulière, indiquant un possible mouvement de prix à venir. En revanche, une diminution de l'imbalance peut indiquer un équilibre retrouvé entre l'offre et la demande, suggérant une pause dans le mouvement des prix.

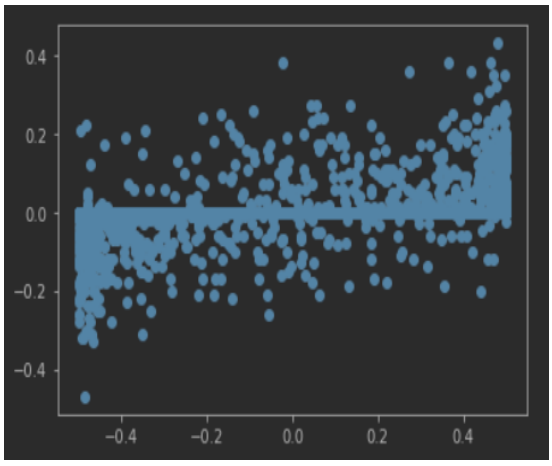
3.2.1 Analyse des données d'order book imbalance



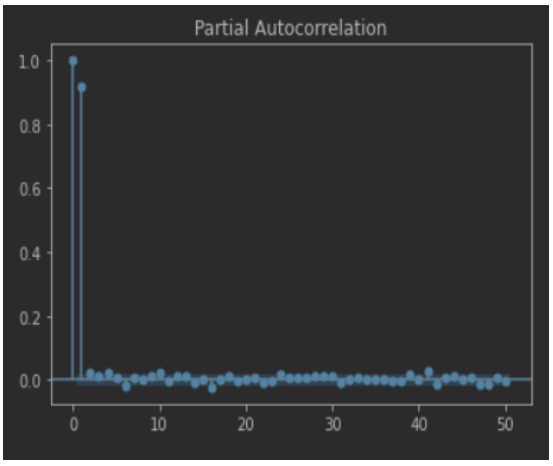
(a) Order Book Imbalance



(b) Order Book Imbalance Hist



(c) Order Book Imbalance vs δP



(d) Order Book Imbalance Auto Corr

On remarque plusieurs choses :
— La séries est bornée, elle est de moyenne nulle de variance à priori constante.
— La répartition est uniforme sauf aux bords où elle atteint un pic.
— A vue d'oeil, il existe une relation soit logistique soit linéaire avec la variation de prix.
— Le signal OBI n'est pas auto corrélé.

En effectuant un test de stationnarité adfuller, on accepte l'hypothèse de stationnarité. En fait, c'est le fait de diviser par le spread qui borne le signal d'order book imbalance.

L'hypothèse de non regression est rejetée mais le R-squared est très faible pour plusieurs raisons :
— A très haute fréquence d'échantillonage, le prix varie très peu localement et énormément de valeurs de δP sont nulles.
— Les données financières sont très difficiles à modéliser

Néanmoins en conditionnant par le fait qu'il n'y avait pas eu de variation de prix depuis quelques timestamps, le signal devient significatif et la régression donne un R-squared de 0.421, une p-value de 0 et un coefficient de regression de 0.2.

3.3 Order Flow Imbalance

Le déséquilibre du flux d'ordres (OFI), également connue sous le nom d'imbalance de l'activité de trading, est une mesure qui évalue la différence entre le volume total d'ordres d'achat et de vente exécutés sur le marché à un moment donné. Contrairement à l'OBI, qui se concentre sur les ordres en attente dans le carnet d'ordres, l'OFI se réfère aux transactions réelles qui ont eu lieu.

Pour calculer l'OFI, on compare le volume total des ordres d'achat exécutés (par exemple, le volume cumulé des achats réalisés à chaque tick) au volume total des ordres de vente exécutés sur une période donnée. Une imbalance positive indique une prédominance des achats, tandis qu'une imbalance négative signale une prédominance des ventes.

L'OFI sur des intervalles de temps $[t_{k-1}; t_k]$ st défini comme une somme des contributions individuelles des événements sur ces intervalles :

$$OFI_k = \sum_{n=N(t_{k-1})+1}^{N(t_k)} e_n;$$

où $N(t_{k-1})+1$ et $N(t_k)$ sont l'indice du premier et du dernier événement dans l'intervalle $[t_{k-1}; t_k]$. L'OFI est une mesure du déséquilibre entre l'offre et la demande, qui englobe les transactions, les ordres limites et les annulations.

Comme pour l'OBI, l'OFI peut fournir des indications sur les pressions acheteuses ou vendeuses sur le marché. Une forte imbalance positive peut indiquer un fort intérêt des acheteurs et une possible hausse des prix à venir, tandis qu'une forte imbalance négative peut signaler une pression vendeuse accrue et une baisse potentielle des prix.

Les traders utilisent souvent l'OFI comme un indicateur de momentum ou de confirmation pour valider les signaux provenant d'autres analyses techniques. Une augmentation soudaine de l'imbalance du flux d'ordres dans un sens particulier peut être interprétée comme un signal d'accélération du mouvement des prix dans cette direction.

3.3.1 Analyse des données d'Order Flow Imbalance

Premièrement, il faut noter que l'order flow imbalance a été modifié en divisant par la taille moyenne average depth afin de considérer une relation linéaire entre l'OFI et la variation de prix passée qui a été décrite dans l'article de Cont et al sur l'order book imbalance. Autrement dit, l'OFI est avant tout une variable descriptive de la variation de prix passée.
Néanmoins, il existe également un pouvoir prédictif lorsque celui-ci est couplé à l'order

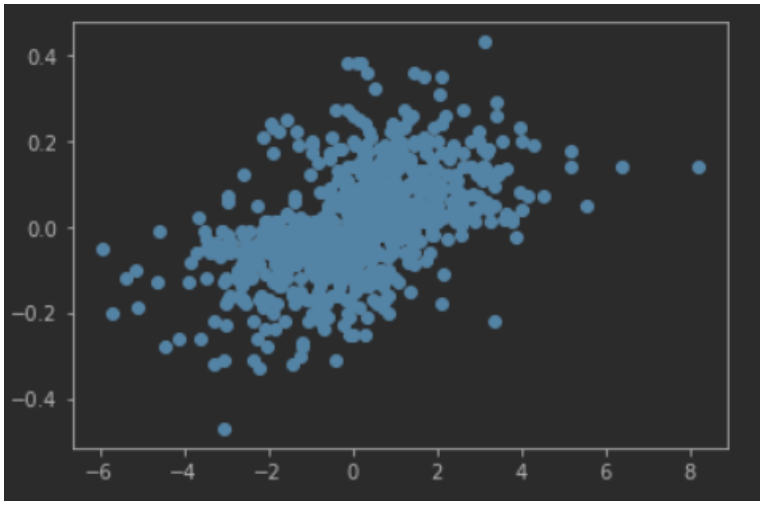


FIGURE 4 – Past variation of price vs OFI over a few seconds

book imbalance. Le pouvoir prédictif est visible à l'oeil nu sur cette figure :

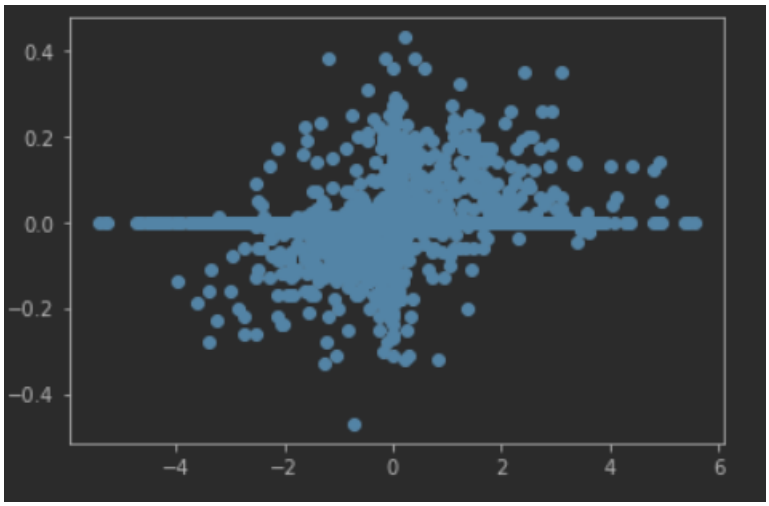
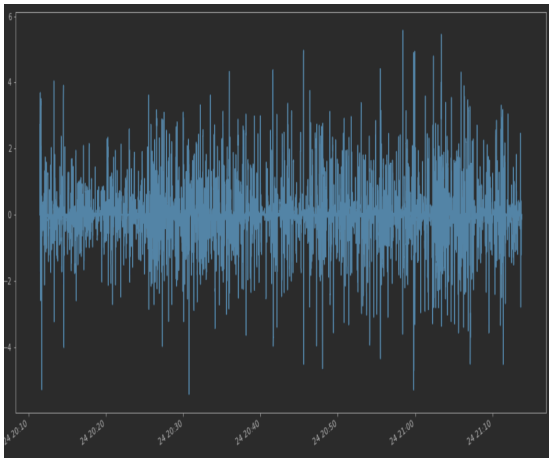
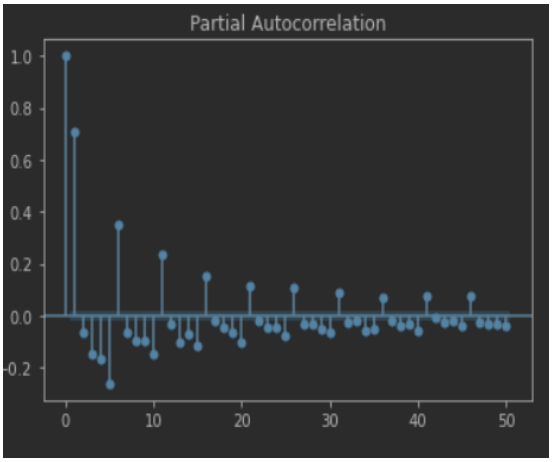


FIGURE 5 – Future Price variation over OFI

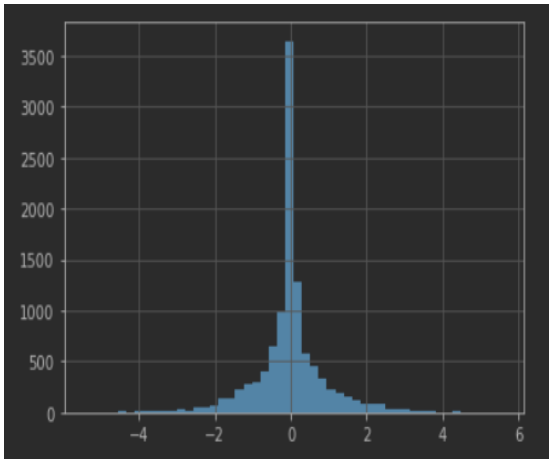
Une regression linéaire confirme ce pouvoir prédictif (Voir le notebook).



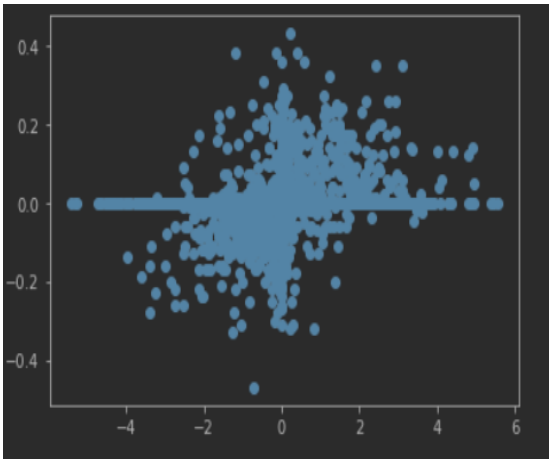
(a) Order Flow Imbalance



(b) OFI Autocorr



(c) OFI hist



(d) OFI vs δP

3.4 Mean Reversion

La réversion à la moyenne est un phénomène observé dans de nombreux marchés financiers où les prix tendent à revenir vers une moyenne à long terme après avoir dévié de celle-ci [6]. Ce concept est souvent utilisé pour développer des stratégies de trading. Mathématiquement, la réversion à la moyenne peut être modélisée par une équation de la forme :

$$dx(t) = -\theta(x(t) - \mu)dt + \sigma dW(t)$$

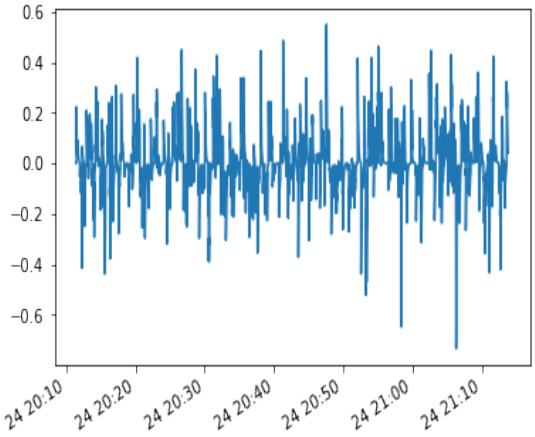
où $x(t)$ représente le prix à un instant t, μ est la moyenne à long terme, θ est le coefficient de réversion, σ est la volatilité du processus, et $dW(t)$ est un processus de Wiener différentiel représentant le bruit aléatoire. Cette équation décrit comment le prix $x(t)$ évolue au fil du temps, avec une force de réversion à la moyenne qui ramène le prix vers μ et un terme stochastique représentant les fluctuations aléatoires du marché.

Dans le contexte de l'analyse de la mean reversion, le calcul en Python s'appuie sur la méthode de la Moyenne Mobile Exponentielle Pondérée (EWMA), qui est utilisée pour

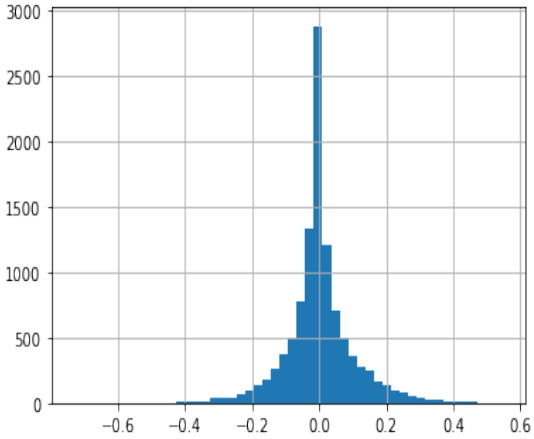
estimer la tendance ou le comportement moyen d'une série temporelle. Plus précisément, pour calculer la mean reversion (MR) d'une série de prix (P), on soustrait la valeur actuelle des prix à l'EWMA des prix sur la même période. Cette approche permet de mettre en évidence les écarts par rapport à la tendance moyenne et de quantifier le degré de mean reversion dans la série de prix. En d'autres termes, la mean reversion est déterminée par la différence entre les prix observés et la moyenne mobile exponentielle, fournissant ainsi des indications sur la dynamique de retour à la moyenne des prix sur une période donnée. Ce calcul constitue une méthode courante pour évaluer et analyser la mean reversion dans les données financières et est largement utilisé dans les études empiriques et les stratégies de trading algorithmique.

3.5 Analyse des données de Mean Reversion

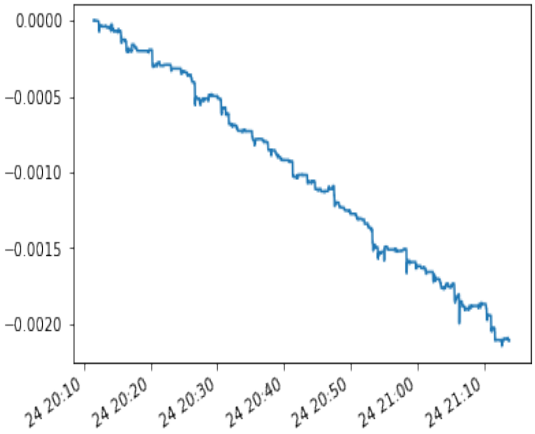
Nous examinons en figure 14 en détail quatre aspects essentiels : la courbe de mean reversion elle-même, l'histogramme de mean reversion pour étudier sa distribution, la somme cumulée de la covariance de la mean reversion pour comprendre ses tendances, et enfin l'auto-corrélation de la mean reversion pour évaluer ses corrélations temporelles.



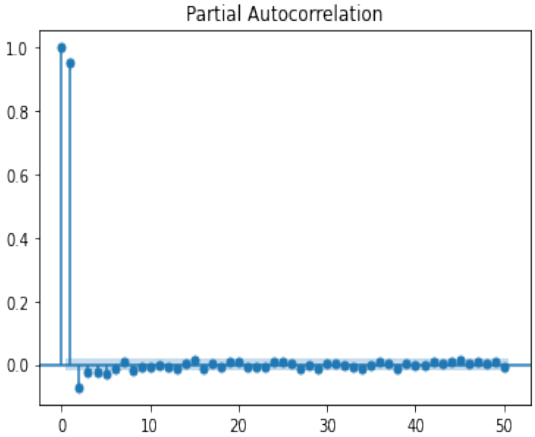
(a) Mean Reversion



(b) Mean Reversion Hist



(c) Mean Reversion cumulative covariance



(d) Mean Reversion Auto Corr

Le signal obtenu est bien stationnaire, avec une distribution centrée en zéro et symétrique. Cela signifie que les périodes de réversion positive sont équilibrées par les périodes

de réversion négative. De plus, le signal n'est pas auto-corrélé.

4 Prise en compte des interaction et analyse de composante principale

Maintenant que l'on a montré le pouvoir potentiel de prédiction des trois features, il faut maintenant construire un modèle de prédiction prenant en compte également les interactions. Le modèle s'écrit de la manière suivante :

$$\delta P_i = \sum_j \alpha_j s_j^i + \beta_j int_j^i + \epsilon_i \tag{2}$$

où les s_j sont les signaux de base, les int_j sont les interactions et ϵ est un bruit blanc de moyenne nulle et de variance constante.

4.1 Analyse des interactions

En effectuant des régressions linéaires multiples sur les données normalisées et en éliminant pas à pas les variables de p-value les moins significatives on arrive à conserver 4 variables :

- OBI
- OFI
- MR
- $OFI \times MR$

Les p-values sont toutes inférieures au threshold de 0.05, les coefficients ne s'annulent pas et sont donc validés.

D'autre part, on peut directement observer les correlation des signaux sur les scatter plot ci-dessous :

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0009	0.000	2.383	0.017	0.000	0.002
x1	0.0155	0.001	13.917	0.000	0.013	0.018
x2	0.0037	0.000	9.747	0.000	0.003	0.004
x3	0.0114	0.003	3.484	0.000	0.005	0.018
x4	-0.0009	0.001	-0.978	0.328	-0.003	0.001
x5	0.0044	0.010	0.428	0.668	-0.016	0.025
x6	-0.0057	0.003	-2.184	0.029	-0.011	-0.001

Omnibus:	4373.674	Durbin-Watson:	1.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783931.895
Skew:	0.832	Prob(JB):	0.00
Kurtosis:	44.705	Cond. No.	32.9

(a) Regression multiple

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0008	0.000	2.477	0.013	0.000	0.001
x1	0.0155	0.001	13.925	0.000	0.013	0.018
x2	0.0037	0.000	9.731	0.000	0.003	0.004
x3	0.0112	0.003	3.423	0.001	0.005	0.018
x4	-0.0058	0.002	-2.533	0.011	-0.010	-0.001

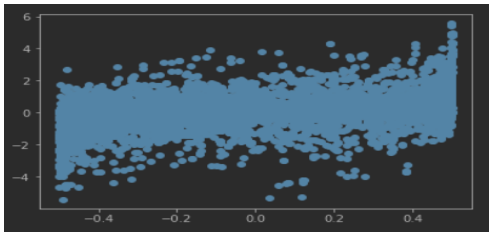
Omnibus:	4354.228	Durbin-Watson:	1.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783581.850
Skew:	0.821	Prob(JB):	0.00
Kurtosis:	44.696	Cond. No.	10.4

(b) Regression Multiple Filtrée

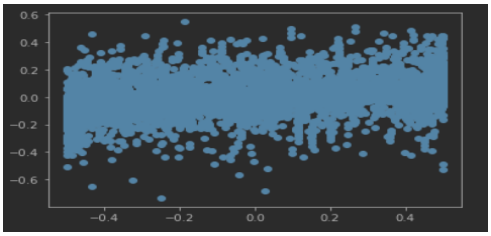
4.2 Analyse en composantes principales

Maintenant que l'on a effectué l'analyse multiple, on va essayer d'orthogonaliser les signaux, de sorte qu'ils soient le plus dé-corrélés possible. La procédure que l'on va suivre s'appelle Principal Component Analysis (PCA) ou Analyse en Composante Principale. Elle consiste à visualiser l'espace des variables et de trouver, par un processus d'orthogonalisation itératif, les directions qui expliquent le plus la variance des observations des variables.

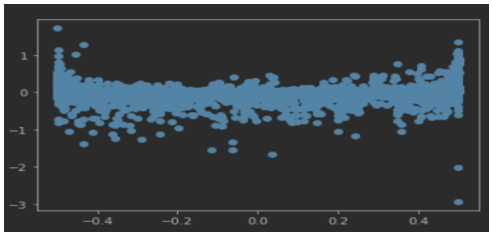
Le processus d'orthogonalisation donne 4 signaux orthogonaux s_1,s_2,s_3,s_4 qui quand ils sont exprimés dans la base originale donnent : On remarque que le dernier vecteur est clairement non significatif dans l'explication de l'espace des variables (c'est essentiellement le vecteur nul). On peut donc conserver s_1,s_2,s_3. Si on normalise les signaux et que l'on trace les séries temporelles ainsi que les répartitions statistiques, on remarque que s1 et s2 sont assez similaires, que s3 a une variance très volatile et que s4 est non significatif. Les vecteurs obtenus sont bien de covariance nulle ou alors non significatif :



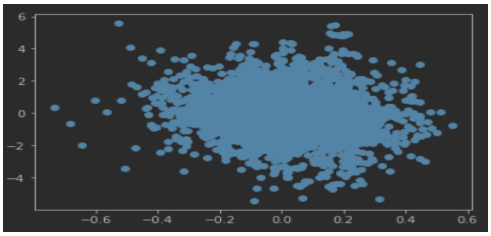
(a) Scatter OBI OFI



(b) Scatter OBI MR



(c) Scatter OBI interaction

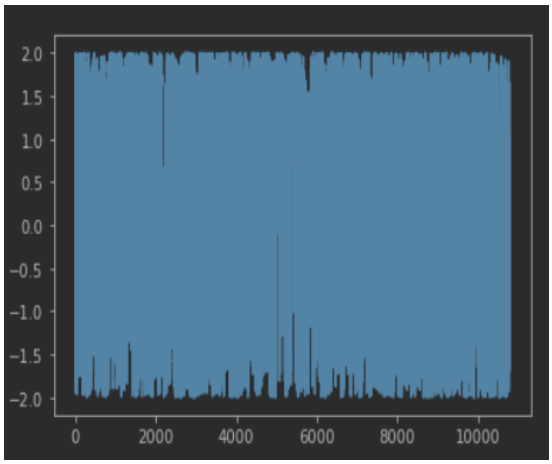


(d) Scatter MR OFI

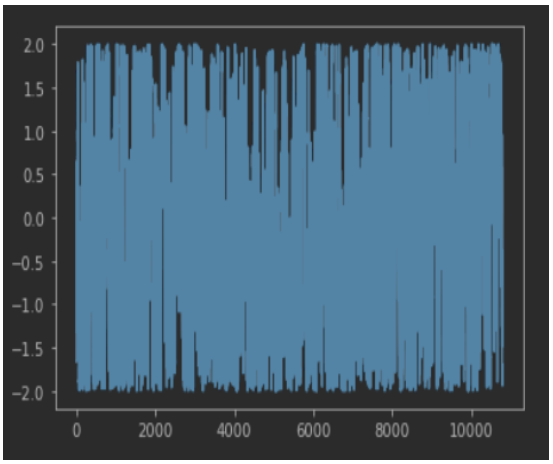
	0	1	2	3
0	-72.975138	97.303724	1.378517	1.230583e-14
1	126.709134	29.271497	-1.844214	1.230583e-14
2	-32.388153	-62.480757	-31.172540	1.230583e-14
3	-21.345843	-64.094463	31.638237	1.230583e-14

FIGURE 10 – Coefficients des vecteurs écrits dans la base e1,e2,e3,e4

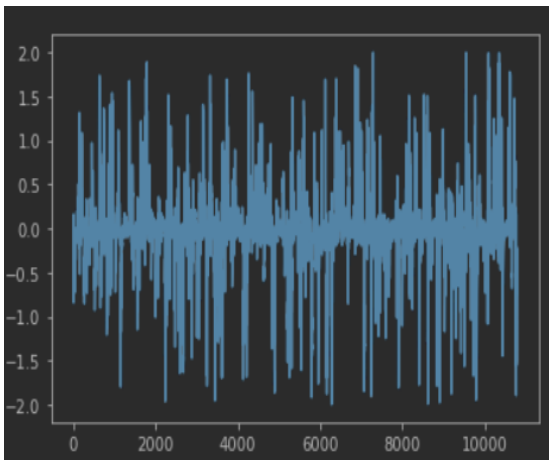
FIGURE 11 – Tracé des séries ortho



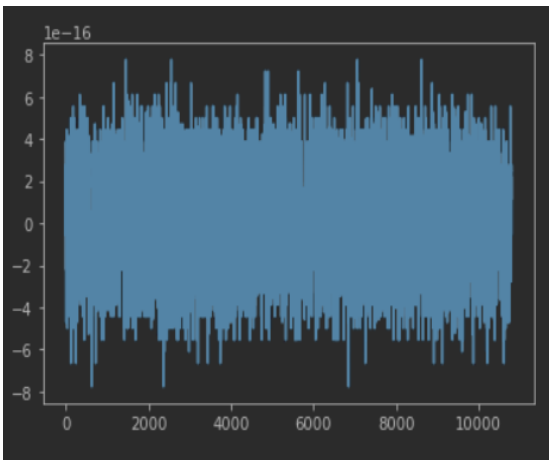
(a) signal 1



(b) signal 2

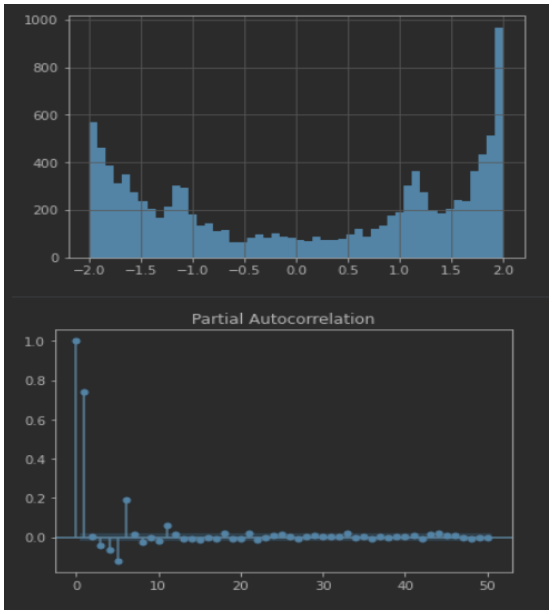


(c) signal 3

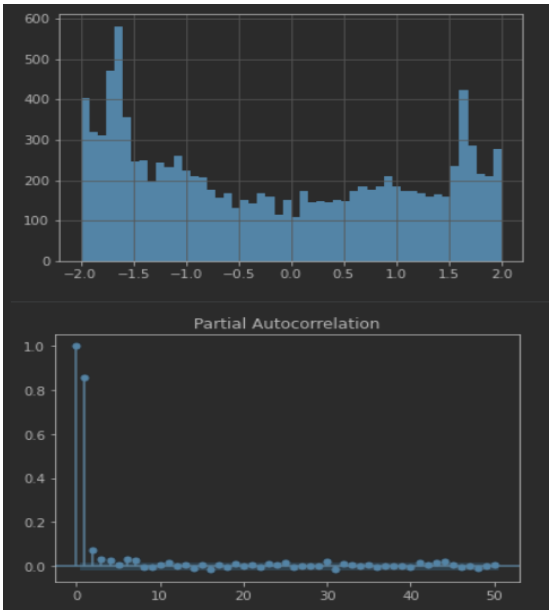


(d) signal 4 (non significatif)

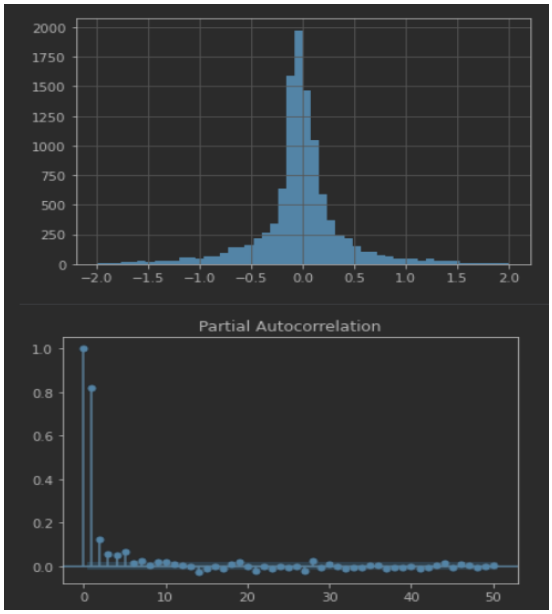
FIGURE 12 – Répartitions statistiques des signaux



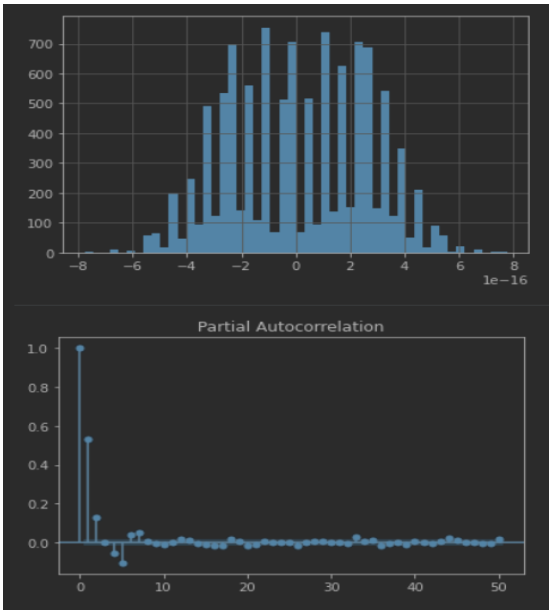
(a) signal 1



(b) signal 2

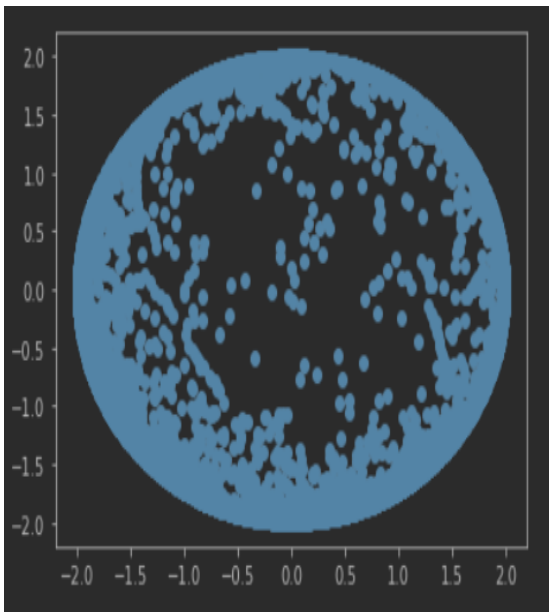


(c) signal 3

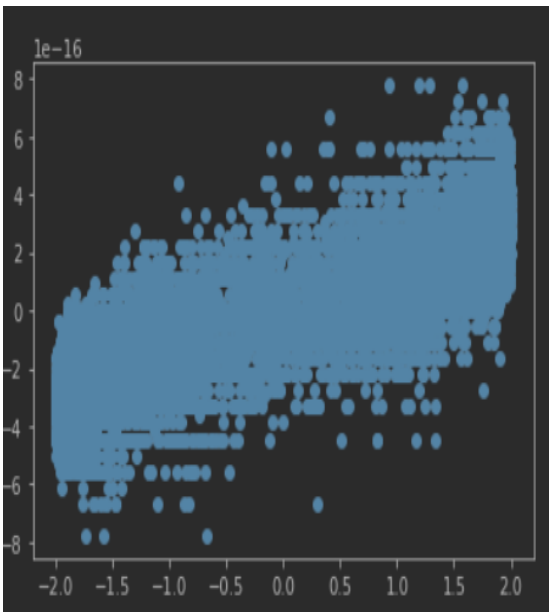


(d) signal 4 (non significatif)

FIGURE 13 – Représentation des interactions



(a) signal 1 et signal 2



(b) signal 1 et signal 4

5 Cross Validaton

Dans cette dernière partie, il s'agit de tester la robustesse de nos signaux de prédiction. Un bon moyen est de tester la stabilité des coefficients de régression du modèle sur des dataset indépendants. La méthode utilisée est la suivante :

- On génère les signaux de PCA sur l'ensemble complet des données d'entraînement.
- On crée des random seeds de 4 sous ensembles d'entraînement et 1 de test de manière uniforme sur le jeu de données complet.
- On entrainement le modèle de regression linéaire sur les données d'entraînement et on compare les performances de chaque jeu d'alpha sur les données de test
- Le résultat est satisfaisant si le signe et la norme des coefficients de régression reste constant à quelques exceptions près que l'on doit savoir expliquer. Et que les significativité des coefficients restent les mêmes.

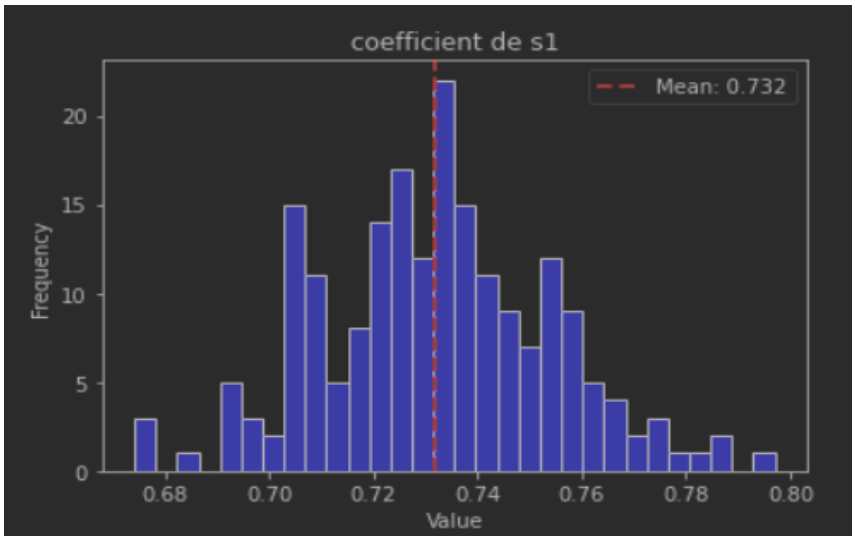
Dans la figure 14 ci-dessous, on observe premièrement que les coefficients de régression sont très robustes, ils ne changent pas de signe et restent dans une bande très proche en terme de valeur.

On va utiliser une dernière métrique utilisée par les chercheurs quantitatifs pour mesurer les performances d'un signal de trading, la cumulative covariance :

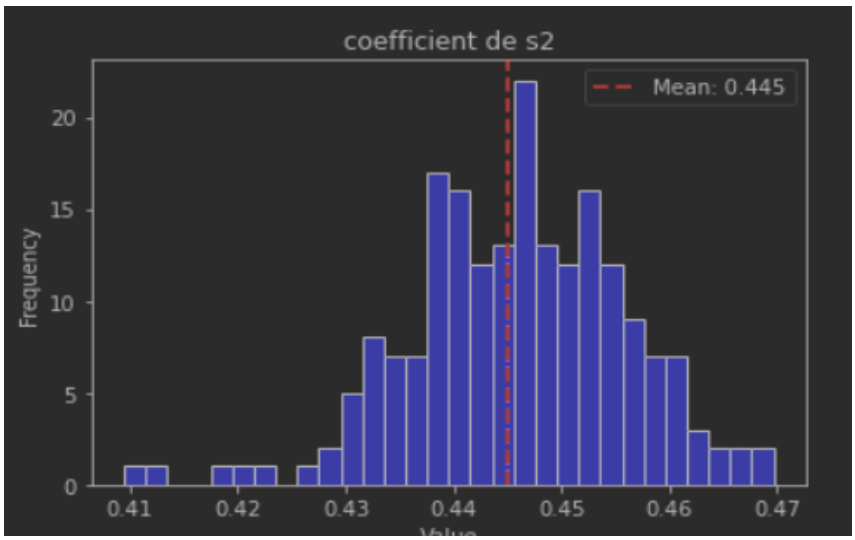
$$CumCov^n(Signal, Prix) = \sum_j^n Signal_j \times Prix_j \tag{3}$$

De cette manière on peut observer la stabilité des coefficients de la prédiction totale.

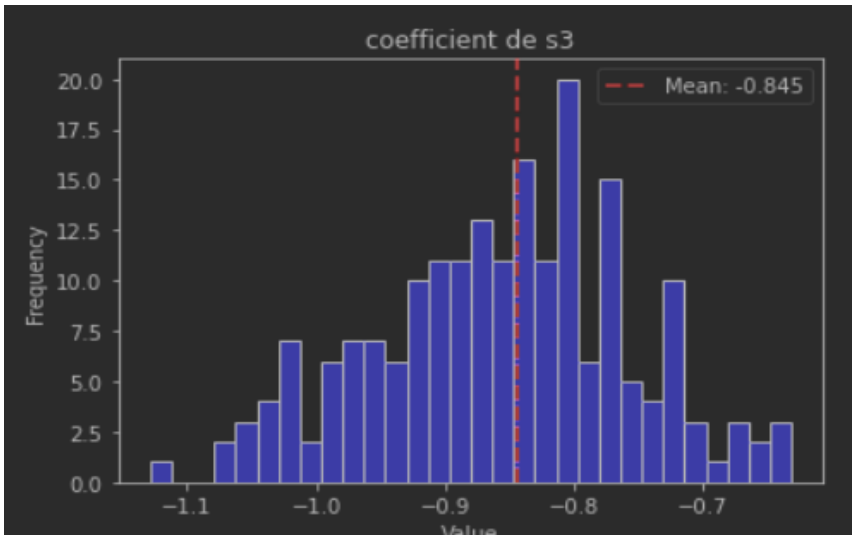
FIGURE 14 – Répartition des coefficients de régression



(a) signal 1



(b) signal 2



(c) signal 3

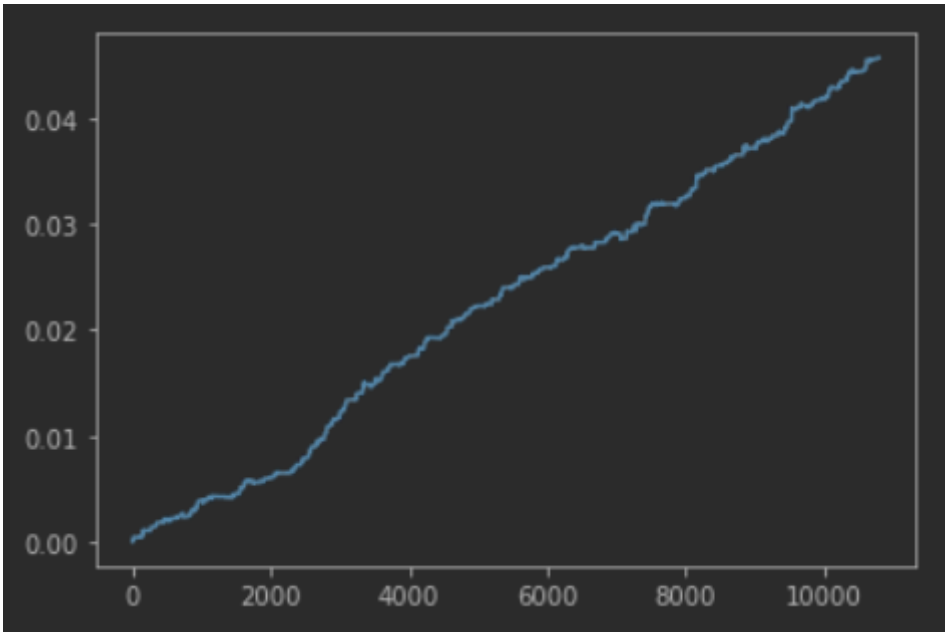


FIGURE 15 – Cumulative covariance of predicted price move vs actual price move

6 Conclusion

Nous avons lors de ce projet d'option, pu appliquer une multitude de connaissances et de techniques enseignées en option mathématiques et ingénierie du risque. Les séries financières étant très complexes à prédire, il peut paraître étonnant d'obtenir de si bon résultats. En réalité, nous n'avons pas pris en compte dans l'étude des performances, les coûts de transaction qui vont, souvent, venir contrebalancer les performances d'algorithmes simples comme celui-ci. Nous avons pu néanmoins obtenir des résultats satisfaisants. Une possible ouverture serait de s'intéresser à l'étude de couches supplémentaires de l'order book.

Références

[1] Fredéric Abergel, Marouane Anane, Anirban Chakraborti, Aymen Jedidi, and Ioane Muni Toke. *Limit Order Books*. Princeton University Press, Princeton, NJ, 2012.

[2] Jonathan A. Chavez-Casillas and Jose E. Figueroa-López. A one-level limit order book model with memory and variable spread. 2016. Received 9 June 2015 ; received in revised form 8 September 2016 ; accepted 27 November 2016. Available online 7 December 2016.

[3] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. March 2011.

[4] Marcus Daniels, J. Doyne Farmer, Léon Gillemot, Giulia Iori, and Douglas E. Smith. Quantitative model of price diffusion and market friction based on trading as a mechanistic random process. *Physical Review Letters*, 90(19), October 2003. DOI : 10.1103/PhysRevLett.90.108102.

[5] J. Doyne Farmer, Paolo Patelli, and Ilija I. Zovko. The predictive power of zero intelligence in financial markets. *arXiv preprint physics/0404006*, 2004. Dated : Original version Sept. 9, 2003 ; this version Feb. 9, 2004.

[6] Boming Ning and Kiseop Lee. Advanced statistical arbitrage with reinforcement learning. *arXiv preprint arXiv :2403.12180*, Mar 2024.

[7] Eric Smith, J. Doyne Farmer, László Gillemot, and Supriya Krishnamurthy. Statistical theory of the continuous double auction. *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501*, February 2008. Dated : February 1, 2008.