Macé Lewis
Servière Adrien

ENSAE 3^{eme} année
2022-2023

Résumé et critique

Market making by an FX dealer : tiers, pricing ladders and hedging rates for optimal risk control

Alexander Barzykin, Philippe Bergault, Olivier Guéant

Avril 2023

Table des matières

1 Présentation de l'article

1.1 Introduction

Un market maker dans le monde du FOREX fait face à deux types de risques, l'un statique et l'autre dynamique. Le premier concerne le volume ciblé par ordre et la marge que le trader peut en tirer. Une fois l'ordre confirmé, le book du trader est directement impacté et soumis aux fluctuations du marché, ce qui induit un risque dynamique, dit « d'inventaire ». Face à ce risque, le dealer dans le contexte spécifique du marché FX peut se tourner vers deux entités pour le couvrir : soit il patiente jusqu'à l'arrivée d'autres clients pour rééquilibrer son inventaire tout en augmentant son skew (internalisation), ce qui lui laisse davantage de liberté de prix (reform) et un impact sur le marché moindre contre d'un temps d'attente incertain, ou bien il peut se tourner vers d'autres dealers capables de lui fournir une couverture rapide au prix d'un coût d'exécution plus élevé et d'un impact sur le marché plus important (externalisation).

L'article étudié s'insère dans un contexte où les modèles de market making n'intègrent pas ce dilemme de contrôle du risque d'inventaire. En reprenant une solution proposée par ses auteurs [3] il allie une approche de market making optimal et d'exécution optimale qui – fait nouveau –dépend de la catégorie du client auquel le dealer fait face. Cette catégorie est estimée à l'aide d'un clustering des paramètres de fonctions d'intensité pour chaque client, de sorte que pour chaque client le modèle propose une grille de tarification pour diverses tailles au bid et à l'ask optimaux (on parle de pricing ladder) et un taux de couverture dédié à l'externalisation.

1.2 Modèle

L'approche générale du modèle trouve ses fondements dans les travaux d'Avellanda et Stoikov [2] complété par les résultats de Guéant, Lehalle et Fernandez-Tapia [10]. Ce modèle, par rapport à celui précedemment cité, permet de modeliser différentes tailles de trade, classer les clients en tiers, et d'externaliser le risque dynamique pour se hedger, se rapprochant encore plus des conditions réélles en pratique.

1.2.1 Flux et intensités

Il est nécessaire de comprendre les sensibilités des clients aux prix pour trouver une solution d'internalisation optimale. On commence par modéliser la pricing ladder bid ask pour K tailles d'ordres différentes. En partant d'un prix de référence S_t, on a, pour le bid à la date t et pour une taille z un prix $S_t^b(t,z) = S_t(1-\delta^b(t,z))$. On modélise pour le bid et l'ask des fonctions d'intensité qui feront le lien avec la probabilité d'arrivée d'un flux client sur un intervalle $[t, t+dt]$ pour une taille $[z, z+dz]$: $\Lambda^b(z, \delta^b(t,z))dzdt$. Du fait de l'ensemble discret de tailles au choix, les intensités sont approximées par des mesures discrètes.

Si l'on s'en tient à une approche rejoignant un contexte de régression logistique, on considère un flux de clients décidés à traiter $\lambda_k^{b/a}$ avec une probabilité $\frac{1}{1+e^{\alpha_k^{b/a}+\delta\beta_k^{b/a}}}$. Dans ce cas, la fonction d'intensité le concernant est obtenue en pondérant la probabilité par le flux associé. En accord avec les données dont disposent les auteurs, chaque trade peut être catégorisé selon six volumes de référence. Les fonctions d'intensité sont ensuite estimées par maximum de vraisemblance, en tenant compte des fonctions de distribution des quotes et des trades réalisés. Dans ce but, on associe aux prix proposés les trades exécutés pour chaque client. Au cours de l'estimation il est apparu que les intensités au bid et à l'ask n'étaient pas significativement différentes. De fait, les intensités proposés à la suite de l'estimation par log-vraisemblance sont une sorte de moyenne propres aux estimations sur les côtes bid et ask.

1.2.2 Modèle de market making

Dans le contexte étudié, le market maker fait face à deux intermédiaires d'échange :

— Les clients au travers de leurs flux de requêtes. En accord avec les résultats du clustering des intensités, les auteurs considèrent que les clients se divisent en deux catégories : chaque groupe fera face à sa propre pricing ladder optimale pour le market maker .
— Les autres market makers, contreparties dédiées à l'externalisation du risque, comme on l'a évoqué auparavant. L'interaction avec cette partie est incarnée par un taux d'exécution dénoté $(v_t)_t$.

La dynamique du prix dans cet univers de marché est découpée en deux parties : la première montre des fluctuations exogènes (le terme classique de la dynamique de Black-Scholes σdW_t), la seconde est endogène et traduit l'impact de marché de l'exécution d'un trade externalisé (liant le taux d'exécution v_t à une magnitude d'impact k). Dans l'approche log-normale qui est choisie, cela donne :

$$dS_t = S_t(\sigma dW_t + kv_t dt)$$

Comme décrit initialement, l'activité commerciale du dealer implique l'existence d'un inventaire (noté $(q_t)_t$) soumis au risque de marché. Cet inventaire est affecté par les demandes des clients et les interactions d'externalisation. En notant $J^{(b,n)}(dt, dz)$ et $J^{(a,n)}(dt, dz)$ les mesures des variables aléatoires qui modélisent la taille et la date des trades acheteurs et vendeurs du tier n, on a la dynamique de l'inventaire :

$$dq_t = \sum_{n=1}^{N} \int_{z=0}^{\infty} zJ^{b,n}(dt, dz) - \sum_{n=1}^{N} \int_{z=0}^{\infty} zJ^{a,n}(dt, dz) + v_t dt$$

Notons que les noyaux d'intensité $J^{(a,\cdot)}$ et $J^{(b,\cdot)}$ font le lien entre inventaire et fonctions d'intensité. Le processus sur lequel repose la fonction objectif du dealer est celui du cash

$(X_t)_t$ du dealer, en tenant compte d'un coût d'exécution à l'externalisation $L(v_t)S_t$(L est une fonction typiquement convexe, décroissante sur R^-et croissante sur R^+) :

$$dX_t = \sum_{n=1}^{N} \int_{z=0}^{\infty} S^{a,n}(t,z) z J^{a,n}(dt,dz) - \sum_{n=1}^{N} \int_{z=0}^{\infty} S^{b,n}(t,z) z J^{b,n}(dt,dz) - v_t S_t dt - L\left(v_t\right) S_t dt$$

Le dealer calque ses choix selon une fonction d'objectif visant à maximiser la valeur de marché de son portefeuille à un horizon T tout en pénalisant son risque. Son portefeuille étant composé de son cash et de l'inventaire qu'il détient, le contrôle optimale est donné par :

$$E\left[X_T + q_T S_T - \frac{C}{2}\int_0^T q_t^2 d[S]_t\right]$$

Les étapes suivantes nécessitent l'application de la formule d'Itô au processus $X_t + q_t S_t$. Une formulation équivalente du problème en se plaçant à des horizons courts permet de prendre $S_T = S_0$. Finalement, en introduisant une fonction de valeur adaptée au problème de contrôle stochastique et faisant quelques hypothèses sur les fonctions d'intensité (différentiabilité, décroissantes selon l'écart du prix moyen, etc.) on peut formuler une équation d'Hamilton-Jacobi dont les solutions sont les pricing ladders adaptées à chaque tier, quantitativement évaluables à l'aide d'un schéma d'Euler implicite.

1.3 Résultats généraux de la modélisation

1.3.1 Sur l'estimation des fonctions d'intensité et la catégorisation des clients

L'estimation des paramètres de la fonction d'intensité nécessite pour chaque client l'historique des quotes associé aux trades exécutés. C'est en répétant cette opération pour plusieurs clients d'HSBC que les auteurs sont en mesure d'isoler deux catégories de clients après clustering des projections dans l'espace (α, β) des paramètres estimés. La figure 2 nous renseigne sur le comportement de chacune d'entre elles : la courbe bleue étant associée aux clients qui semblent les moins sensibles à des variations de prix. Les résultats sont cohérents et montrent la persistance de cette dualité de catégories pour la variété des tailles observées.

1.3.2 Sur le modèle de market making

Sur un échantillon de clients d'HSBC, les amplitudes d'intensité sont égales pour les deux groupes de clients, de sorte à rester en ligne avec une estimation antérieure de Butz et Oomen [6]. La fonction de coût et le coefficient d'impact de marché sont estimés en suivant la procédure proposée par Almgren et al pour le marché action [1]. Le temps terminal d'optimisation T est choisi faible pour assurer une convergence vers des prix stationnaires.

Les auteurs constatent pour chaque catégorie de client la présence d'une zone d'internalisation pure pour un inventaire proche de zéro. Au-delà, le taux d'externalisation

v_t est linéairement lié au niveau de l'inventaire. Cette zone est dépendante du paramètre ϕ de la fonction de coût d'externalisation, tout comme l'aversion au risque, la franchise des clients, la volatilité, les coûts d'exécution et l'impact sur le marché dans le cas d'une externalisation (figure 4). En outre, l'écart bid-ask est fonction du signe du flux client et de sa catégorie : l'estimation pour un inventaire neutre de cet écart pour un volume de 1M€ double quasiment selon la catégorie de sensitivité du client, passant de 0.26 bps pour les plus sensibles à 0.55 bps pour les moins sensibles, quantités en ligne avec les observations faites sur le marché. Un cas particulier de prix à l'achat pour le groupe 1 et à la vente pour le groupe 2, complété par le tracé du taux d'externalisation selon l'inventaire est présenté dans la figure 3.

La praticité de ce modèle permet son transfert simple à un contexte de simulation une fois que les intensités et les pricing ladders ont été estimés. Cette adaptabilité permet d'évaluer le modèle dans différents contextes paramétriques pour observer leur effet sur le ratio d'internalisation/externalisation du risque. La figure 5 montre la proportion des trades clients (catégories 1 et 2) et externalisés chez les autres dealers ainsi qu'un temps caractéristique d'externalisation du risque (l'intégrale de la fonction d'autocorrélation de l'inventaire), selon différents niveaux d'aversion au risque. Les résultats permettent de constater que la part des clients les plus sensibles aux prix est progressivement sacrifiée dans un contexte d'aversion au risque croissante appelant à plus d'externalisation.

Enfin, une dernière étude est menée sur la frontière efficiente pouvant être tirée du modèle (figure 6). Cette frontière – par analogie avec l'approche à la Markovitz d'un portefeuille efficient – est obtenue après plusieurs simulations du PnL pour différentes valeurs d'aversion au risque (ligne bleue). On observe aussi sur cette figure le résultat d'autres stratégies qui dévient aléatoirement de la solution optimale par des écarts de prix aléatoires (les points rouges). Finalement, les auteurs tracent aussi le PnL maximal attendu avec une aversion au risque minimale (et donc sans aucune externalisation du risque).

Il se trouve que les simulations déviant de la solution du modèle se situent majoritairement en dessous de la frontière représentée par la stratégie optimale. Ce résultat est un appui supplémentaire en faveur du modèle proposé ; malgré une fonction d'objectif n'étant pas directement de moyenne-variance , il s'avère bien efficace dans cette configuration risque/récompense. A côté de cette observation, le PnL augmente progressivement avec l'aversion au risque, mais contrairement au cadre général de Markovitz, il est borné par le cas extrême sans externalisation du risque.

2 Critique et commentaires

2.1 Sur les hypothèses et choix du modèle

De ce que nous savons sur le market-making compte-tenu du cours et de nos connaissances personnelles, ce modèle spécifié pour le marché OTC du FX nous a paru pertinent

du point de vue de la modélisation des quantités clés. Tout d'abord l'arrivée des clients, modélisée par un processus de comptage dont l'intensité dépend de la quantité d'actif traitée et des prix bid-ask nous parait être très générale et peu restrictive. On pourrait par ailleurs imaginer utiliser des processus auto-excitants tels que les processus de Hawkes. L'intensité devient donc dépendante des réalisations passées :

$$\lambda_t = \lambda_0(t) + \int_0^t \nu(t-s)dN_s = \lambda_0(t) + \sum_{0<t_i<t} \nu\left(t - t_i\right)$$

$\lambda_0(t)$ étant l'intensité de base, ν un kernel (décroissant le plus souvent), t_i les temps d'évènements. Ainsi, s'il y a récemment eu beaucoup d'ordres, cela augmente l'intensité et donc la probabilité de saut dans un cours intervalle de temps à venir.

Ces processus sont surtout utilisés dans le cadre de la modélisation de Limit Order Books (LOB) [11, 12], car ils permettent d'incorporer des faits stylisés très importants à prendre en compte lorsque les acteurs du marché voient les ordres des autres, ce qui n'est par contre pas le cas dans un marché OTC. En effet sur un LOB assez liquide on observe des faits de renforcement de consensus (par exemple un ordre limite à l'achat implique un autre ordre limite à l'achat), d'effets dit de 'momentum' (par exemple un ordre de marché à l'achat implique un autre ordre de marché à l'achat), et bien d'autres, qui n'existe pas dans un marché OTC car les clients n'ont pas accès aux ordres des autres. En revanche on pourrait considérer des faits stylisés tels que les market impacts, des effets de foules à la suite d'une nouvelle exogène, qui feraient en sorte que les clients feraient plus de transactions aux dealers FX et pourraient être modélisés par les processus susmentionnés. On peut également introduire des corrélations entres clients via ces processus, ce qui pourrait s'inscrire dans le cadre du clustering des clients, qui est déjà justifiée par l'approche de l'article.

Cependant, nous ignorons dans quelle mesure cette remarque est utile au contexte de l'OTC, ni si cela améliorera les résultats de l'article qui est déjà très général, mais dans le cas assez liquide du marché FX peut-être que beaucoup d'ordres peuvent arriver en même temps à une certaine période (par 'clusters') si bien que l'intensité augmente. L'estimation de cette nouvelle intensité de Hakwes peut se faire assez facilement par maximum de vraisemblance via la librairie `Hawkes` de python, pourvu que les types de kernels, leurs nombres et le type de baseline soient spécifiés. Une boucle `for` suffit pour trouver ces types.

Ensuite l'estimation des paramètres α et β justifie bien la création de tiers, comme on le voit dans leur figure 2, où les deux clusters de paramètres symbolisent des clients différents. Le problème de contrôle optimal stochastique est ensuite posé puis résolu (sous des conditions de régularité citées en référence) pour minimiser le critère mean-variance, et utilise un résultat de Cartea et Jaimungal cité en cours qui justifie de ne tenir compte que de la variance du 'price risk' et d'omettre celle du 'spread capture' dans la décomposition du PnL.

Les résultats numériques découlent ensuite de l'application du modèle et présentent

de très bon résultats avec un PnL positif et contrôlé par le niveau de risque choisi par le Market maker, comme on peut le voir dans leur figure 6 qui présente une frontière efficace. C'est sur ce niveau de risque choisit que nous voulions cependant nous arrêter. Le choix des paramètres d'aversion au risque, reste cependant à déterminer selon les objectifs du dealer.

2.2 Sur la part de marché des dealers

Les auteurs font mention dans leur article de l'importance de la part de marché des dealers dans leur modèle commercial sans pour autant l'inclure dans leur fonction objectif. La littérature liée au market making qui cible spécifiquement cette multi-objectivité est d'après nos recherches assez maigre.

On peut néanmoins envisager que la prise en compte de la part de marché serait à moduler avec le risque général pris par le dealer : dans un cas limite il ne serait pas soumis à cette considération, ce qui donnerait la solution uniquement conditionnée sur le PnL ; dans le cas opposé la part de marché serait centrale, le dealer exécuterait un maximum d'ordres et internaliserait au maximum sa gestion de risque. De fait, nous supposons que la conséquence principale de la prise en compte de la part de marché dans la fonction objectif élargirait la zone d'internalisation du risque d'inventaire.

2.3 Sur le changement de régime de volatilité du marché FX et la calibration

Le marché s'est repositionné au niveau de la volatilité par rapport aux observations de 2019, note la revue semestrielle de BIS [7], notamment du fait des évènements macroéconomiques rencontrés début 2022. Le turnover moyen journalier FX (spot et dérivés confondus) a augmenté de 14% par rapport au dernier sondage. Les observations récentes confirment la poursuite de la fragmentation du marché FX, essentiellement orienté vers les plateformes propriétaires des dealers majeurs de ce marché. Côté risque, les données montrent une augmentation de la part d'externalisation, avec un progrès des échanges inter-dealer spot en hausse de 43%.

En faisant le lien avec le modèle étudié, on peut noter comment seraient amenées à changer les solutions optimales dans ce contexte nouveau et corroborer les observations décrites dans le contexte actuel : l'augmentation de la volatilité serait pour partie contrebalancée par un niveau d'aversion au risque plus élevé côté dealer, mais donnant globalement une plus forte externalisation.

Concernant la calibration de la fonction de coût et de l'impact permanent, les auteurs précisent que leur méthodologie est tirée d'un article d'Almgren et al [1]. Cette méthodologie fait aujourd'hui office de référence dans de nombreux articles, indépendamment du fait qu'elle ait été initialement pensée pour le marché Equity.

2.4 Vers le cas du marché FX complet

La suite logique de ce travail est l'adaptation du modèle à un marché FX non réduit à une paire. Cet article étant déjà une extension d'un article précèdent des mêmes auteurs [3] qui intégrant la gestion du risque dynamique d'inventaire via l'externalisation, le classement de différentes tailles de trades, et la séparation de différents types de clients. Dans cette continuité de développement, les auteurs étendent ces résultats à un marché FX composé de plusieurs monnaies et non plus d'une seule paire dans un papier plus récent [4]. Ils utilisent notamment des paramètres de l'article étudié.

Ils citent également un article de [9, 8] qui propose un cadre mathématique de market-making avec plusieurs actifs. Cependant la résolution de ce dernier pour obtenir les quotes optimales nécessite la résolution d'une équation différentielle qui souffre de la malédiction de la dimension. La nouvelle approche de [4] se focalise sur le marché FX, et généralise en fait l'étude présentée plus haut avec $d > 1$ monnaies, ce qui transforme les équations vues plus haut en des équations matricielles, qui intègrent également la corrélation entre actifs par une matrice de corrélation. Des méthodes de résolutions (deep reinforcement learning, réduction des facteurs de risque) ont été étudiées, mais la nouvelle approche de [4] utilise des résultats de [5], permettant d'obtenir des approximations de la stratégie optimale (de la fonction de contrôle optimale) en résolvant une équation de Riccati matricielle en dimension raisonnablement faible, linéaire en le nombre de paires échangées. Selon les auteurs, cette nouvelle approche est utilisable en pratique, ils la testent d'ailleurs sur 5 monnaies : USD, EUR, JPY, GBP et CHF.

Dans un tel marché, il n'y a plus vraiment de bid-ask car un trade revient simplement à échanger une monnaie pour une autre. Ils prennent donc un numéraire de référence (l'USD) pour exprimer les autres. L'inventaire peut être exprimé en différentes monnaies. Les figures et résultats sont semblables à ceux de leur article étudié plus haut : nous avons bien une zone d'internalisation pure si l'inventaire ne dépasse pas une certaine limite (dépendante de l'aversion au risque, mais également des corrélations entre actifs dans cette nouvelle configuration) puis une externalisation approximativement linéaire.

Conclusion

Pour conclure, ce modèle nous semble pertinent pour une banque voulant développer une activité de market-making FX en OTC en se rémunérant sur le spread (sans spéculation donc), et la littérature qui s'appuie sur cet article le généralise déjà à un marché plus gros à plusieurs paires. Nous avons aussi mentionné des pistes, ou des nuances dues à la conjoncture pour développer davantage ce modèle qui prend déjà en compte la plupart des aspects essentiels à cette activité.

Références

[1] Robert Almgren et al. "Direct estimation of equity market impact". In : *Risk* (2005).

[2] Marco Avellaneda et Sasha Stoikov. "High-frequency trading in a limit order book." In : *Quantitative Finance* (2008).

[3] Alexander Barzykin, Philippe Bergault et Olivier Guéant. "Algorithmic market making in foreign exchange cash markets with hedging and market impact". In : *arXiv preprint arXiv:2106.06974* (2021).

[4] Alexander Barzykin, Philippe Bergault et Olivier Guéant. "Dealing with multi-currency inventory risk in FX cash markets". In : *arXiv preprint arXiv:2207.04100* (2022).

[5] Philippe Bergault et al. "Closed-form approximations in multi-asset market making". In : *Applied Mathematical Finance* 28.2 (2021), p. 101-142.

[6] Maximilian Butz et Roel Oomen. "Internalisation by electronic fx spot dealers". In : *Quantitative Finance* (2019).

[7] Julián Caballero et al. "BIS Quarterly Review, December 2022 49 The internationalisation of EME currency trading". In : *BIS Quarterly Review* (Decembre, 2022), p. 49-64.

[8] Olivier Guéant. "Optimal market making". In : *Applied Mathematical Finance* 24.2 (2017), p. 112-154.

[9] Olivier Guéant. *The Financial Mathematics of Market Liquidity: From optimal execution to market making*. T. 33. CRC Press, 2016.

[10] Olivier Guéant, Charles-Albert Lehalle et Joaquin Fernandez-Tapia. "Dealing with the inventory risk: a solution to the market making problem". In : *Mathematics and financial economics* (2013).

[11] Xiaofei Lu et Frédéric Abergel. "High-dimensional Hawkes processes for limit order books: modelling, empirical analysis and numerical calibration". In : *Quantitative Finance* 18.2 (2018), p. 249-264.

[12] Ioane Muni Toke et Nakahiro Yoshida. "Modelling intensities of order flows in a limit order book". In : *Quantitative Finance* 17.5 (2017), p. 683-701.