Université de Montréal

Mesure et prévision de la volatilité pour les actifs liquides

par
Selma Chaker

Département de sciences économiques
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en sciences économiques

Avril, 2012

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

Mesure et prévision de la volatilité pour les actifs liquides

présentée par:

Selma Chaker

a été évaluée par un jury composé des personnes suivantes:

Marine Carrasco	président-rapporteur
Silvia Gonçalves	directeur de recherche
Ilze Kalnina	membre du jury
Tim Bollerslev	examinateur externe
Michael Huberman	représentant du doyen de la FES

RÉSUMÉ

Le prix efficient est latent, il est contaminé par les frictions microstructurelles ou bruit. On explore la mesure et la prévision de la volatilité fondamentale en utilisant les données à haute fréquence.

Dans le premier papier, en maintenant le cadre standard du modèle additif du bruit et le prix efficient, on montre qu'en utilisant le volume de transaction, les volumes d'achat et de vente, l'indicateur de la direction de transaction et la différence entre prix d'achat et prix de vente pour absorber le bruit, on améliore la précision des estimateurs de volatilité. Si le bruit n'est que partiellement absorbé, le bruit résiduel est plus proche d'un bruit blanc que le bruit original, ce qui diminue la misspécification des caractéristiques du bruit.

Dans le deuxième papier, on part d'un fait empirique qu'on modélise par une forme linéaire de la variance du bruit microstructure en la volatilité fondamentale. Grâce à la représentation de la classe générale des modèles de volatilité stochastique, on explore la performance de prévision de différentes mesures de volatilité sous les hypothèses de notre modèle.

Dans le troisième papier, on dérive de nouvelles mesures réalizées en utilisant les prix et les volumes d'achat et de vente. Comme alternative au modèle additif standard pour les prix contaminés avec le bruit microstructure, on fait des hypothèses sur la distribution du prix sans frictions qui est supposé borné par les prix de vente et d'achat.

Mots clés : Volatilité réalisée, bruit microstructure du marché, volume.

ABSTRACT

The high frequency observed price series is contaminated with market microstructure frictions or noise. We explore the measurement and forecasting of the fundamental volatility through novel approaches to the frictions' problem.

In the first paper, while maintaining the standard framework of a noise-frictionless price additive model, we use the trading volume, quoted depths, trade direction indicator and bid-ask spread to get rid of the noise. The econometric model is a price impact linear regression. We show that incorporating the cited liquidity costs variables delivers more precise volatility estimators. If the noise is only partially absorbed, the remaining noise is closer to a white noise than the original one, which lessens misspecification of the noise characteristics. Our approach is also robust to a specific form of endogeneity under which the common robust to noise measures are inconsistent.

In the second paper, we model the variance of the market microstructure noise that contaminates the frictionless price as an affine function of the fundamental volatility. Under our model, the noise is time-varying intradaily. Using the eigenfunction representation of the general stochastic volatility class of models, we quantify the forecasting performance of several volatility measures under our model assumptions.

In the third paper, instead of assuming the standard additive model for the observed price series, we specify the conditional distribution of the frictionless price given the available information which includes quotes and volumes. We come up with new volatility measures by characterizing the conditional mean of the integrated variance.

Keywords: Realized volatility, market microstructure noise, volume, depths.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

À mes parents.

REMERCIEMENTS

Je remercie chaleureusement tous ceux et celles qui m'ont aidée. Plus particulièrement, j'adresse mes vifs remerciements ma directrice de recherche, Sílvia Gonçalves, pour son judicieux encadrement et tout ce que j'ai pu apprendre sous sa direction.

J'exprime ma reconnaissance à Nour Meddahi qui était mon directeur de recherche quand j'ai commencé ma thèse. Grâce à lui, j'ai développé la précieuse qualité de pouvoir mener mes recherches de façon indépendante. Je le remercie aussi de m'avoir poussée à explorer de nouvelles littératures.

Mes remerciements s'adressent aux institutions qui m'ont accueillie durant ma thèse notamment Imperial College Business School à Londres en Angleterre et Toulouse Business School en France. Je témoigne toute ma gratitude à Nour Meddahi pour ces opportunités.

Merci pour la confiance et l'octroi de bourses de recherche des organismes tels que le Conseil de Recherche en Sciences Humaines du Canada (CRSH), l'Institut de Finance Mathématique de Montréal (IFM2), le Centre Interuniversitaire de Recherche en Analyse des Organisations (CIRANO), le Centre Interuniversitaire de Recherche en Économie Quantitative (CIREQ) et la Banque Laurentienne.

Une pensée sincère pour Bruno Feunou Kamkui, Ilze Kalnina, Michel Poitevin, Jean-Sébastien Fontaine, Andrew Patton, Bruno Biais, Tim Bollerslev et tous mes professeurs de l'Université de Montréal.

Un grand merci à mes adorables frères, mes chers amis et toute ma famille pour m'avoir soutenue.

INTRODUCTION

La volatilité mesure la fluctuation ou encore la variabilité d'une série temporelle donnée. Il est important de souligner le caractère latent et stochastique de la volatilité. Utilisée aussi bien en évaluation d'‘actifs, de couverture de risque et de gestion de portefeuille, la volatilité doit être estimée et prédite avec précision. La littérature classique GARCH (Generalized Autoregressive Conditional Heteroskedasticity) présente des modèles paramétriques de la volatilité. Une récente approche de l'économétrie financière estime la volatilité de façon nonparamétrique et ma thèse s'inscrit dans le cadre de cette ligne de recherche. L'approche nonparamétrique se base sur la théorie de la variation quadratique. En outre, ce qui a particulièrement alimenté l'avancement de la recherche est la disponibilité de plus en plus grandissante des données à haute fréquence. Il s'agit des prix et volumes des actifs liquides qui sont transigés à titre de milliers de fois par jour dans les grandes places boursières. Ainsi, approximer le processus du prix par un processus en temps continu devient très légitime. Plus précisément, le processus du prix p_t est modélisé par la semimartingale,

$$dp_t = \mu_t dt + \sigma_t dW_t,\ t \in [0,1]$$

où W_t est un processus Brownian standard et σ_t est la volatilité spot. Dans ma thèse la variable d'intérêt est la volatilité intégrée IV, définie comme

$$IV = \int_0^1 \sigma_s^2 ds.$$

L'intervalle $[0,1]$ réfère à une journée par exemple. Ainsi, on observe presque des rendements infinitésimaux dp_t ou à haute fréquence tout le long de la journée $[0,1]$. La fréquence d'échantillonnage dt atteint une seconde. Cependant, la base de données Trades and Quotes (TAQ) qui contient généralement des données à une fréquence d'une seconde a récemmment disposé de données à la fréquence d'une milli-seconde. Un tel progrès ouvre la porte pour la recherche sur la volatilité spot par exemple.

Un estimateur consistent de la volatilité intégrée est la volatilité réalisée RV définie par,

$$RV = \sum_{i=1}^{N} r_i^2,$$

où N est la taille de l'échantillon et r_i est le rendement à haute fréquence, à une seconde par exemple. RV est la variation quadratique du processus de prix. Le type de théorie asymptotique utilisé pour démontrer la convergence de RV vers IV n'est pas l'asymptotique Standard puisque l'intervalle couvert $[0,1]$ reste fixe. On utilise plutôt l'asymptotique Infill, qui remplit de plus en plus d'observations l'intervalle fixe. Ainsi passer de la seconde à la milli-seconde comme fréquence d'échantillonnage permet à la théorie asymptotique d'être une approximation plus réaliste.

Un problème inhérent aux prix à haute fréquence est qu'ils sont contaminées par des frictions dues au marché qu'on appelle bruit microstructurel du marché. En effet, les actifs sur le marché ne sont pas transigés à leur valeur fondamentale. Les frictions incluent les coûts de transaction ainsi que l'asymétrie d'information sur le marché. Concrètement, le bruit rend la volatilité réalisée inconsistente pour la volatilité intégrée. Le premier estimateur de volatilité robuste à la présence du bruit microstructure a été proposé en 2005. L'idée est basée sur un échantillonnage de la série des prix et combine deux échelles de temps : haute fréquence et basse fréquence. Une seconde approche basée sur les autocovariances a été développée plutard et délivre des estimateurs consistents avec une vitesse de convergence optimale de $N^{-1/4}$. La méthode de pre-averaging introduite en 2009 est principalement la même que les deux approches précédentes, mais avec un meilleur traitement des effets de bord et des hypothèses sur le bruit moins contraingnantes.

L'apport fondamental de ma thèse est de spécifier le bruit microstructurel du marché par des variables économiques ; de façon directe dans les deux premiers papiers et indirecte pour le troisième papier. Ainsi, j'améliore la mesure et la prévision de la volatilité intégrée. En effet, concevoir les frictions comme un bruit ou une erreur de mesure est une approche statistique qui ignore la nature même de ces frictions. En plus, il a été démontré que pour un bruit de magnitude stochastique d'ordre un, la vitesse de convergence optimale est $N^{-1/4}$. Or, l'estimateur pre-averaging a déjà atteint cet optimum. Outre la vitesse de convergence, techniquement les hypothèses sur le bruit sont bien limitées et ne peuvent pas être relaxées indéfiniment. D'ailleurs, la majorité des travaux de recherche suppose un bruit identiquement et indépendemment distribué pour simplifier la théorie. Introduire de nouvelles variables dans l'estimation de la volatilité s'inspire naturellement de la théorie microstructure des marchés. En effet, cette théorie étudie le processus de formation des prix ainsi que les mécanismes de marché. La qualité d'un marché est bonne si le niveau de frictions est faible. Une mesure de la qualité de marché est la variance du bruit microstructure qui contamine les prix observés. Aussi, la différence entre prix d'achat et prix de vente d'un actif ou spread, constitue une mesure de friction. Un spread large peut être engendré par une forte volatilité fondamentale. D'où un lien entre la variance du bruit microstructure et la variance fondamentale. J'étudie les implications de ce lien dans le deuxième papier. D'autres observables tel que le volume de transaction et les volumes d'achat et de vente (correspondant au maximum qu'on peut acheter ou vendre aux prix d'achat et de vente respectivement) peuvent entrer dans les mesures de friction.

Le premier papier de ma thèse est motivé par quelques modèles de frictions de la littérature théorie microstructurelle des marchés. On considère le modèle standard additif où le prix observé est la somme du prix sans frictions et du bruit microstructure. On utilise cinq variables explicatives pour capturer sous une forme linéaire le bruit. Quatre des ces variables sont observables, le volume de transaction, le spread, les volumes d'achat et de vente. L'unique variable explicative qui est inférée est une variable binaire prenant la valeur +1 si la transaction est initiée par un acheteur ou -1 si la transaction est initiée par un vendeur. On utilise un algorithme classique pour inférer la série de cette variable binaire. Le modèle se présente comme une régression d'impacte du prix, où la variable dépendente est la série des rendements et les variables indépendantes sont les variations des cinq variables explicatives du bruit. Économétriquement, il s'agit d'une régression Infill. Le résidu de cette régression est le rendement sans frictions en plus d'un bruit blanc non capturé par les variables de frictions. On distingue formellement le cas où le bruit blanc est nul, ce qui correspond à ce que les variables de friction ont capturé tout le bruit microstructure. Ainsi, la série des résidus de la regression de l'impact du prix correspond à des rendements moins contaminés par le bruit que les rendements observés. On montre, aussi bien théoriquement qu'empiriquement qu'on améliore la précision de l'estimation de la volatilité si on utilise la série des résidus de la régression de l'impact du prix à la place des rendements observés.

Dans le deuxième papier, on suppose que la variance du bruit microstructure varie en fonction du temps. Plus précisément, on modélise la variance du bruit microstructure comme fonction affine de la volatilité fondamentale. Dans le cadre théorique des fonctions propres de volatilité stochastique, on examine les conséquences de notre modèle sur la prévision de la volatilité. Le cas du bruit identiquement et indépendamment distribué a été étudié dans ce même cadre théorique. On s'interroge ainsi sur l'apport du caractère variable en fonction du temps du bruit pour la prévision de la volatilité. Le cadre théorique des fonctions propres de volatilité stochastique permet d'avoir des formules exactes pour les mesures de performance du forecasting en fonction des paramètres du modèle de volatilité stochastique supposé pour σ_t. On constate que la volatilité réalisée RV performe mieux que les estimateurs robustes au bruit microstructure en ce qui concerne la prévision de la volatilité. En effet, si la variance du bruit est informative à propos de la variance fondamentale alors tout estimateur robuste au bruit n'exploite pas cette information. On aborde aussi le problème de correction de biais de prévision dans ce papier. En effet, vu le caractère latent de IV, tout modèle de prévision nécessite un proxy de IV pour l'utiliser comme variable dépendante. Sous notre modèle, on montre qu'utiliser RV comme proxy de IV engendre un biais qui dépend du temps. Si la variance du bruit est constante, le biais de prévision serait constant aussi. Cette conclusion engendre des implications pratiques quand l'économètre calcule la volatilité prédite.

Dans le troisième papier, on se distingue de la littérature en se basant sur d'autres hypothèses que le modèle additif stipulant que le prix observé est la somme du prix sans frictions et du bruit microstructure. L'intuition de ce papier remonte à la théorie de l'identification partielle selon laquelle relaxer certaines hypothèses diminue le risque de mauvaise spécification du modèle mais vient au prix d'une identification partielle et non totale de l'objet d'intérêt. Un exemple simple d'identification partielle est de dériver des bornes pour la variable d'intérêt au lieu d'avoir une estimation ponctuelle. En supposant que le prix sans frictions est compris entre le prix de vente et le prix d'achat, on a dérivé des bornes pour RV. Empiriquement, ces bornes ne sont pas informatives car trop larges. On a donc exploré des distributions conditionnelles du prix sans frictions à support borné par le prix de vente et le prix d'achat. L'ensemble de conditionnement contient les prix et les volumes d'achat et de vente. En effet, en introduisant le volume dans les paramètres des distributions, on peut examiner l'impact sur la prévision de l'information apportée par le volume. On dérive de nouvelles mesures de volatilité correspondant à l'espérance conditionnelle de RV pour chacune des distributions supposée pour le prix sans frictions. Le spread apparait naturellement dans les expressions de certaines nouvelles measures. Pour la prévision de la volatilité, on montre empiriquement que l'utilisation du spread et du volume peut être bénéfique.

ARTICLE 1

VOLATILITY AND LIQUIDITY COSTS

Abstract [1]

This paper proposes a new estimator of the integrated volatility using some liquidity costs variables to absorb the market microstructure noise that contaminates high frequency prices. More specifically, I model the noise as a linear function of the inferred trade direction indicator, the signed trading volume, the bid-ask spread and the quoted depths. These liquidity cost measures can totally or partially absorb the market microstructure noise. In either case, I argue that the difference between the observed prices and the estimated liquidity cost measures yield an adjusted price series that is more likely to satisfy the usual semi-martingale assumption. I formally test this assumption. Empirically, I estimate daily integrated volatility for the stock Alcoa from the NYSE. For more than half of 01/2009-03/2011 business days, a linear liquidity cost function captures all the noise and the sum of squared adjusted returns is a consistent volatility estimator.

Key phrases : Realized volatility, bid-ask spread, trading volume, quoted depths.

[1] I am grateful to Nour Meddahi for helpful discussions. I also thank Sílvia Gonçalves for feedbacks. I express my gratitude to Ilze Kalnina for comments and discussions about the paper assumptions', and I thank Andrew Patton, George Tauchen, Tim Bollerslev, Bruno Feunou, and all the participants of the financial econometrics lunch group at Duke University, Torben Andersen, Dobrislav Dobrev, and all the participants of the NBER-NSF time series conference 2011, Jean-Sébastien Fontaine, Scott Hendry, Antonio Diez, Jun Yang, Nikil Chande, and all the participants of Bank of Canada seminar, Rim Lahmandi and all the participants of the MASE seminar at École Polytechnique de Tunisie. I gratefully acknowledge Bruno Feunou for providing me with Alcoa high frequency data. Many thanks to Walid Chaker, Federico Bandi, Imen Ghattassi, Prosper Dovonon, Sophie Moinas, and Bruno Biais.

1.1 Introduction

The advent of large intraday financial databases in the last quarter of the twentieth century has created new opportunities to improve risk management and asset pricing. While it is natural to think that with respect to data “more is always better”, this is not the case when dealing with high frequency asset returns. The reason is that as the sampling frequency increases, observed prices are more contaminated with trading frictions such as bid-ask bounce, rounding errors, discrete trading prices, etc. These so-called market microstructure effects create a discrepancy between the frictionless price process and the observed prices, resulting in the inconsistency of realized volatility as an estimator of integrated volatility (see Zhang et al. (2005) and Bandi and Russell (2008)). To overcome this problem, the existing literature on the econometrics of high frequency data has proposed a number of alternative estimators that are consistent for integrated volatility when observed prices are modeled as the sum of the frictionless price process and an error term that captures the market microstructure effects (see, among others, Zhou (1996), Zhang et al. (2005), Hansen and Lunde (2006), Bandi and Russell (2008), Barndorff-Nielsen et al. (2008), Podolskij and Vetter (2009), and Jacod et al. (2009)).

Although these alternative approaches to estimating volatility are robust to some form of market microstructure noise, they do not exploit any of the possible driving forces of the market microstructure effects. Nevertheless, the market microstructure theory literature suggests a number of potential variables that can explain the existence of this noise term. For instance, Roll (1984) models the trade price noise as a linear function of the trade direction indicator (which takes the value +1 if the trade is buyer initiated and -1 if the trade is seller-initiated) whereas Glosten and Harris (1988) models it as a linear function of the trade direction indicator and the signed volume. The bid-ask spread is also a common source of frictions as in Stoll (2000) and Huang and Stoll (1997). A larger spread is associated with more illiquid stocks and is thus a natural measure of frictions. In Kavajecz (1999), the ask (bid) depth that specifies the maximum quantity for which the ask (bid) applies is used to capture inventory control costs as well as asymmetric information costs. In this context, a larger depth implies an increase in liquidity.

In this paper, instead of leaving the noise unspecified [2], I explicitly model the market microstructure noise as a linear function of liquidity cost measures such as the trade direction indicator, the signed volume and the depth. This model is estimated by least squares and yields a adjusted return series that can then be used in conjunction with existing realized volatility-like estimators. This approach not only provides a more structural interpretation to the noise inspired in the microstructure theory literature, but can also result in more efficient estimators of volatility. Improved estimation is due to the fact that the adjusted returns are more likely to conform to the assumptions that justify the use of nonparametric estimators. In the extreme case in which the explanatory variables absorb all the noise, the resulting adjusted returns effectively measure the frictionless returns and a realized volatility estimator can be used. This estimator is squared root convergent, which is the maximum rate of convergence attainable in this context. If instead the liquidity costs only partially absorb the noise, I argue that the remaining noise is likely better behaved than the original noise series (i.e. it is closer to being exogeneous and it is less dependent). This may result in more efficient estimators of integrated volatility based on existing noise robust estimators applied to the adjusted returns series.

[2] In Carrasco and Kotchoni (2011), the market microstructure noise is modeled semiparametrically and depends on the frequency at which the prices are recorded.

Specifically, suppose that the Glosten and Harris (1988) model holds. In this case, the residual obtained by regressing intraday trade returns on the trade direction indicator and the signed volume variation corresponds exactly to the frictionless return. Therefore, the sum of squared residuals (i.e. the realized volatility estimator applied to the residuals) is the sum of squared frictionless returns and it is a consistent estimator of integrated volatility with the maximum possible convergence rate.

Assuming that a few explanatory variables such as the trade direction indicator and the signed volume can fully absorb all the noise is perhaps too strong an assumption. Therefore, I allow for the possibility that the explanatory variables related to liquidity costs only partially absorb the noise. In this more realistic scenario, the regression residuals (i.e. the adjusted returns) no longer represent the frictionless returns and a nonparametric noise robust estimator should be used. One important assumption underlying many of the existing estimators (such as the subsampling approach of Zhang et al. (2005) and the pre-averaging approach of Jacod et al. (2009)) is the independence between the noise and the frictionless price process. This exogeneity assumption has been questioned empirically by Hansen and Lunde (2006). Li and Mykland (2007) show that endogeneity causes the inconsistency of the subsampling two-times scale volatility estimator. Kalnina and Linton (2008) discuss an alternative estimator, but require strong assumptions on the endogeneity form. In this paper, I assume that the explanatory variables capture all the endogeneity component of the noise and therefore justify the use of the subsampling approach in conjunction with the residuals obtained from the regression of intraday returns on liquidity costs variables. The assumption that these variables capture the endogeneity in the noise is consistent with the asymmetric information models of Glosten and Harris

(1988) and Hasbrouck (1991), where informational frictions are due to adverse selection, resulting in endogeneity.

Another assumption that typically underlies the justification of many existing nonparametric estimators of volatility that are robust to market microstructure noise is the assumption that the noise is i.i.d. This implies first order negative autocorrelation for observed returns. Although mathematically convenient, this assumption does not hold empirically (see Hansen and Lunde (2006) and Diebold and Strasser (2008)). I argue that by filtering the observed returns with a regression that uses liquidity cost measures, we obtain a less correlated adjusted return series. Therefore, the i.i.d. assumption on the remaining noise is more likely to be satisfied. As a consequence, I show here that the pre-averaging estimator of Jacod et al. (2009) applied to the adjusted returns is more efficient asymptotically than the pre-averaging estimator applied to the original contaminated returns. This result is shown when there is no endogeneity problem in the noise and therefore the precision gain associated with the adjusted approach of this paper comes solely from reducing the degree of dependence in the adjusted returns.

By comparing the finite sample simulation results with those predicted by the asymptotic theory, I find that the method advocated in this paper outperforms the pre-averaging approach. Empirically, I use intraday data for Alcoa from the NYSE covering the business days of January 2009 to March 2011 in order to estimate daily volatility. I compare the realized variance estimator using the adjusted prices with the pre-averaging robust to dependent noise estimator. Using trade price and trade direction indicator, volume, spread and depths as liquidity costs explanatory variables, the noise is completely soaked up for more than the half of the business days. For the remaining business days, the noise is partially absorbed and I use the pre-averaging robust to noise volatility estimator. In an artificial option trading market, I find that the agent using the realized variance based on adjusted prices to estimate volatility achieves profits compared to the agent using the pre-averaging estimator who endure losses.

The rest of this paper is organized as follows. In Section 2, I present the model for market microstructure noise based on liquidity costs. I discuss the estimation of this model and present a test for the performance of the liquidy cost measure. In Section 3, I discuss volatility estimation based on adjusted prices using the liquidity measures introduced in Section 2. Section 4 offers a simulation exercise. Section 5 is an empirical application where I compare the estimation accuracy of the volatility estimator of this paper to the pre-averaging estimator. In Section 6, I study option trading in an artificial market, and Section 7 concludes.

1.2 Liquidity Costs Measurement

The standard additive model of the high frequency literature is given by,

$$p_t = p_t^* + \varepsilon_t, \;\; t \in [0,1], \tag{1.1}$$

where p_t is the observed price, p_t^* is the frictionless price, and ε_t is the noise. The fixed interval $[0,1]$ is a day, for example. In the literature, there are some attempts to link the noise to market frictions. For instance, Aït-Sahalia and Yu (2009) relate the market microstructure noise to financial measures of the stock liquidity, and Bandi and Russell (2006) distinguish between the adverse selection frictions and transaction costs in the noise. From an empirical stand point, Hansen and Lunde (2006) show that ε_t is not a white noise as commonly assumed, but time-dependent instead. In this paper, I explicitly specify the noise in the standard model (1.1) as,

$$\varepsilon_t = F_t^{'}\beta + \xi_t,$$

where F is an M- vector of liquidity costs observables, and ξ is a residual independent white noise. Then, the observed price is given by

$$p_t = p_t^* + F_t^{'}\beta + \xi_t. \tag{1.2}$$

The Glosten and Harris (1988) model is nested in the previous model and is given by,

$$p_t = p_t^* + \underbrace{\underbrace{\beta_1}_{fixed\ transaction\ costs} q_t + \underbrace{\beta_2}_{size\ varying\ transaction\ costs} q_t v_t}_{=\varepsilon_t}. \tag{1.3}$$

For $\beta_2 = 0$, the Glosten and Harris (1988) model is reduced to the Roll (1984) model. In this section, I extend the Glosten-Harris (1988) linear model with adding other explanatory variables in the noise. I first make assumptions about the validity of the noise explanatory variables. Then I provide consistent estimators of the noise parameters β. Finally, I test whether the noise is completely or partially absorbed using the observables F_t. If the high frequency adjusted returns are free from first order autocorrelation then the frictionless return is recovered. But if the adjusted returns are still autocorrelated, I quantify the improvement done so far by comparing the characteristics of the original noise and the remaining noise.

The one-dimensional price process, which is evolving in continuous time over the fixed interval [0,1], is defined on a complete probability space $(\mho, F, \mathbf{P})$. I consider an information filtration, the increasing family of σ-fields $(F_t)_{t\in[0,1]} \subseteq F$, which satisfies the usual conditions of **P**-completeness and right continuity. The prices and explanatory noise variables are included in the information set F_t.

The arbitrage-free log price p^* is assumed to follow the continuous semimartingale dynamics

$$dp_t^* = \mu_t dt + \sigma_t dW_t,$$

where W_t is standard Brownian motion and σ_t is a *càdlàg* volatility function, which is independent (no leverage). The object of interest of the next section is the integrated variance $IV = \int_0^1 \sigma_u^2 du$. I dispose of N equidistant observations at $i = 0, 1, .., N$ over [0,1]. For notation simplicity, an intraday variable Y_i stands for $Y_{i/N}$. The sample size N goes to infinity, because I use the highest data frequency available. Convergence in probability is denoted $\xrightarrow{P}$, whereas convergence in law is denoted $\xrightarrow{L}$. For mixed normal limit distributions, I denote the stable convergence as $\xrightarrow{st}$.

I denote r_i and r_i^* the intraday observed and latent returns $p_i - p_{i-1}$ and $p_i^* - p_{i-1}^*$, respectively. The noise variation $\Delta\varepsilon_i$ is given by $\varepsilon_i - \varepsilon_{i-1}$. The first differences or variations of the regressors and the residual noise are denoted by $X_i = F_i - F_{i-1}$ and $\Delta\xi_i = \xi_i - \xi_{i-1}$, respectively. The semimartingale assumption for the frictionless price is fully exploited if I regress returns instead of prices. The price impact regression is given by,

$$\underbrace{r_i}_{regressand} = \underbrace{X_i^{'}}_{regressors} \beta + \underbrace{r_i^* + \Delta\xi_i}_{residual} \;\; ; \;\; i = 1, ..., N. \tag{1.4}$$

In matrix notation, the regression is written as,

$$r = X\beta + r^* + \Delta\xi, \tag{1.5}$$

where

$$r = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix} \;,\; X = \begin{pmatrix} X_1^{(1)} & \dots & X_1^{(M)} \\ \vdots & \vdots & \vdots \\ X_N^{(1)} & \dots & X_N^{(M)} \end{pmatrix} \;,\; \Delta\xi = \begin{pmatrix} \Delta\xi_1 \\ \vdots \\ \Delta\xi_N \end{pmatrix}. \tag{1.6}$$

Observe that, in such a regression, the object of interest -the frictionless return- is part of the regression residual. In terms of relative magnitudes of the frictionless return compared to the liquidity costs increments, we make in the next subsection the assumptions to identify each component.

1.2.1 Assumptions

I make the following set of assumptions.

Assumption A

(i) ξ is i.i.d. and independent from p^* and F, $E[\xi] = 0$.

(ii) $F_t = f(p_t^*) + \tilde{F}_t$ such that $E[\tilde{F}_t | p_t^*] = 0$; $f(.)$ is a smooth M-dimensional function.

Assumption $A(ii)$ specifies the form of endogeneity between the frictionless price and the liquidity costs variables F. To my knowledge, this paper is the first to consistently measure the frictionless price volatility under an endogeneity assumption as described in Assumption $A(ii)$. Under this assumption, nonparametric robust to noise volatility estimators converge to the volatility of $p^* + f(p^*)^{'}\beta$ instead of the volatility of p^*. This result is derived by applying Theorem 1 in Li and Mykland (2007). However, if I relax the assumption that the unabsorbed noise ξ is exogenous described in Assumption $A(i)$, it would be impossible to estimate the volatility of p^*. Thus, I am assuming that the liquidity costs variables capture all sources of endogeneity. The next assumption specifies the magnitude of the noise.

Assumption B

(i) $E[(F_t - F_{t-h})(F_t^{'} - F_{t-h}^{'})] = \Omega^{(t,h)}$, a positive definite matrix ; $t \in [0,1]$, $h > 0$.

(ii) $\frac{1}{N}\sum_{i=1}^{N}\Omega_i \xrightarrow{P} \Omega$, a positive definite matrix where $\Omega_i = \Omega^{(i/N,1/N)}$.

(iii) $\frac{1}{N}\sum_{i=1}^{N}(X_i X_i^{'}) \xrightarrow{P} \Omega$.

(iv) $\frac{1}{N}E\left[(\sum_{i=1}^{N} X_i \Delta\xi_i)(\sum_{i=1}^{N}\Delta\xi_i X_i^{'})\right] \xrightarrow{P} S$.

Assumption B concerns the stochastic magnitude of the noise variation compared to the frictionless return, it also defines the validity of the noise explanatory variables. The part (iv) is useful to derive the asymptotic distribution of the OLS estimator of β.

Assumption B imposes bounds for the second moment of X, and the order of the variations X is assumed to be $\mathscr{O}(1)$. That is, the variance of X does not vanish when the sample size N grows. Being of $\mathscr{O}(1)$ is a fundamental identifying assumption, because the order of noise variations is $\mathscr{O}(1)$. To be part of the noise, any explanatory variable candidate must be $\mathscr{O}(1)$,

$$r = \underbrace{r^*}_{\mathscr{O}(1/\sqrt{N})} + \underbrace{X\beta}_{\mathscr{O}(1)} + \underbrace{\Delta\xi}_{\mathscr{O}(1)}. \tag{1.7}$$

The notation $\mathscr{O}(1/\sqrt{N})$ for the frictionless return means that its variance is of order $1/N$, which is the variance size of the Brownian motion increment. Noise variations dominate frictionless returns at ultra high frequencies, which translates into the explosion of the realized variance (or the sum of squared returns) signature plot at high frequencies. If the frictions ε are an exogenous white noise, then

$$E[\sum_{i=1}^{N} r_i^2] = E[IV] + 2NE[\varepsilon^2].$$

At ultra high frequencies, N goes to infinity and the signature plot explodes because of the noise term.

In the presence of informational frictions or endogeneity, the noise increments have two components : the real frictions component and the informational frictions component. In this case, the first component dominates the second at ultra high frequencies.

The next assumption will be useful when the liquidity costs explanatory variables capture all the noise and the remaining noise ξ is zero. In that case, the residual of the regression (1.5) is the frictionless return only. The heteroscedasticity of the frictionless return r^* under stochastic volatility will impact the asymptotic distribution of the price impact regression parameters. To handle this case, we make the Assumption C.

Assumption C $\sum_{i=1}^{N} r_i^{*2} X_i X_i^{'} \xrightarrow{P} \Omega^*$.

Assumption C is not needed when $\xi \neq 0$. Indeed, in that case, the dominating regression residual term is $\Delta\xi$ (r^* is negligible). Under Assumption $A(i)$, this dominating term is homoscedastic.

In the next subsection, I derive the asymptotic theory for the estimators of the liquidity costs parameters.

1.2.2 Inference

In this section, I show consistency and asymptotic normality of the OLS estimator of β. Let $\widehat{\beta}$ be the OLS estimator of β defined by $\widehat{\beta} = (X'X)^{-1}X'r$. Two cases are possible. First, the case where there is a residual exogenous white noise ξ non captured by the liquidity costs variables F. Second, if the price impact regression residual is the frictionless return. All proofs are in the Appendix.

Proposition 1. *Under Assumptions A and B, and if* $\xi \neq 0$ *a.s.,*

$$(i)\ \widehat{\beta} \overset{P}{\rightarrow} \beta.$$

$$(ii) \sqrt{N}(\widehat{\beta} - \beta) \overset{L}{\rightarrow} \mathcal{N}\left((0)_{M\times 1}, \Omega^{-1}S\Omega^{-1}\right).$$

Usually, endogeneity causes inconsistency of the OLS estimator. In this case, consistency holds even in the presence of endogeneity because of the relatively small magnitude of the endogeneity. Moreover, I obtain the usual $\sqrt{N}$ rate of convergence. The frictionless return moments do not appear in the asymptotic variance of the OLS estimator. Indeed, the stochastic magnitude of the frictionless return is negligible. Let the adjusted returns $\widehat{r}$ be defined as,

$$\widehat{r_i} = r_i - X_i'\widehat{\beta} \tag{1.8}$$

Then, if $\xi \neq 0$ a.s., Proposition 1 applies and,

$$\widehat{r_i} = r_i^* + X_i^{'} \underbrace{(\beta - \widehat{\beta})}_{\mathscr{O}(1/\sqrt{N})} + \Delta\xi_i. \tag{1.9}$$

Since $\widehat{\beta}$ is $\sqrt{N}$-consistent, the order of the estimating error is $\mathscr{O}(\widehat{\beta} - \beta) = \mathscr{O}(1/\sqrt{N}) = \mathscr{O}(r^*)$. Therefore, based on their order, the frictionless returns and the estimation error of $\widehat{\beta}$ are not distinguishable.

If the noise is completely captured by the liquidity costs variables, the following result holds.

Proposition 2. *Under Assumptions A, B, C, and if* $\xi = 0$ *a.s.,*

$$(i)\ \widehat{\beta} \xrightarrow{P} \beta.$$

$$(ii) N(\widehat{\beta} - \beta) \xrightarrow{L} \mathscr{N}\left((0)_{M\times 1}, \Omega^{-1}\Omega^*\Omega^{-1}\right).$$

Consistency is then obtained with a faster rate of convergence than that of the Proposition 1 because the residual of the price impact regression is of smaller order of magnitude than the case where $\xi \neq 0$,

$$\widehat{r_i} = r_i^* + X_i^{'} \underbrace{(\beta - \widehat{\beta})}_{\mathscr{O}(1/N)}, \tag{1.10}$$

in which case the frictionless returns are now the dominant term in the adjusted return expression.

1.2.3 Assessing the quality of the liquidity costs measure

As the previous section pointed out, the order of magnitude of the adjusted return $\widehat{r}$ is different depending on whether the noise is completely absorbed (i.e. $\xi = 0$) or partially absorbed ($\xi \neq 0$). In this section, I present a formal test for serial correlation in the adjusted returns. The null hypothesis of zero first order serial covariance in the adjusted returns corresponds to the case where the noise is completely absorbed. The alternative hypothesis of non zero serial covariance in adjusted returns corresponds to the case where the noise is partially absorbed.

Within the microstructure literature, this test may be interpreted as a test for the quality of measurement of trading costs. If the null hypothesis is not rejected, then the explanatory variables in the noise capture all the frictions related to trading costs. If the null hypothesis is rejected, the trading cost measures do not capture all the real frictions.

The null hypothesis H_0 and the alternative hypothesis H_1 are, respectively,

$$\begin{aligned} H_0 &: \xi = 0 \; a.s. \\ H_1 &: \xi \neq 0 \; a.s. \end{aligned} \tag{1.11}$$

The hypothesis H_0 and H_1 are respectively equivalent to $RC1 = 0 \; a.s.$ and $H_1 : RC1 < 0 \; a.s.$, where the realized autocovariance of order one $RC1$ for the adjusted returns is given by,

$$RC1 = \left(\sum_{i=1}^{N} \widehat{r}_i \widehat{r}_{i-1} + \sum_{i=1}^{N} \widehat{r}_i \widehat{r}_{i+1} \right) /2. \tag{1.12}$$

In Barndorff-Nielsen et al. (2008), the asymptotic distribution of the first order autocovariance of a continuous semimartingale increments is given. In particular, if the residual noise ξ is totally absorbed, that is, if H_0 is true, the asymptotic distribution derived by Barndorff-Nielsen et al. (2008) applies.

Proposition 3. *Under H_0, and given Assumptions A, B, and C,*

$$\sqrt{N}RC1 \xrightarrow{st} \mathcal{N}(0, IQ),$$

where the integrated quarticity is defined by $IQ = \int_0^1 \sigma_u^4 du$.

Suppose $\hat{IQ}$ is a consistent estimator for the integrated quarticity IQ. Then, the test statistic S_N is given by

$$S_N = \frac{\sqrt{N}RC1}{\sqrt{\hat{IQ}}}. \tag{1.13}$$

Observe that under $H_0 : RC1 = 0$ a.s.,

$$S_N \xrightarrow{d} \mathcal{N}(0,1). \tag{1.14}$$

I reject H_0 at confidence level α when

$$|S_N| > c_{1-\frac{\alpha}{2}}, \tag{1.15}$$

where $c_{1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$-quantile of the $\mathcal{N}(0,1)$ distribution. Notice that this test is consistent against the alternative $H_1 : RC1 \neq 0$ a.s.

1.3 Volatility Estimation

Using the adjusted high frequency returns, I estimate integrated volatility in this section. First, for the case where the frictionless price is recovered, the realized variance is a consistent estimator with optimal convergence rate. Second, if the liquidity costs are only partially absorbed, a robust to noise volatility estimator is still needed. The asymptotic theory is presented for the two cases.

1.3.1 The frictionless return identification case

I denote $[L,L] = \sum_{i=1}^{N}(\Delta L_i)^2$ the realized variation of a series L_i, and $\langle L^*, L^* \rangle = lim_{N\to\infty}[L^*, L^*]$ where L_t^* is a semimartingale, and $\langle L^*, L^* \rangle$ is the quadratic variation.

Theorem 1. *Under Assumptions A, B, C, and if* $\xi \equiv 0$ *a.s.,*

$$\textbf{\textit{(i)}}\ [\widehat{p},\widehat{p}] \xrightarrow{P} IV.$$

$$\textbf{\textit{(ii)}}\ \sqrt{N}([\widehat{p},\widehat{p}] - IV) \xrightarrow{st} \mathcal{N}(0, 2\, IQ).$$

According to Theorem 1, if the liquidity costs measures fully absorb the noise, the the realized volatility of the adjusted price process $\widehat{p}$ is a consistent estimator of IV, and its asymptotic distribution is the usual distribution of the realized volatility when no market microstructure noise exists. In particular, estimation error in $\hat{\beta}$ does neither impact the consistency nor the asymptotic distribution of the estimator based on the adjusted returns because this error is of smaller order of magnitude (it is $\mathcal{O}(1/N)$). To compute confidence intervals for the integrated volatility, a feasible estimator of the integrated quarticity is needed. I show in the appendix that the sum of adjusted returns to the fourth power is a consistent estimator of the integrated quarticity.

1.3.2 The partially absorbed liquidity costs case

In this section, I treat the case where the noise is partially absorbed. Among the existing nonparametric noise robust *IV* estimators I choose the pre-averaging method of Jacod et al. (2009) because it provides a consistent estimator of the integrated quarticity in the presence of market microstructure noise. The integrated quarticity is needed in the asymptotic distribution of robust to noise volatility estimators. Moreover, under heteroscedastic and autocorrelated noise, the pre-averaging estimator converges to the integrated variance at the optimal rate[3] of $N^{1/4}$. Let L_t be a given semimartingale contaminated with noise. The sum of the pre-averaged increments $[L,L]^{avg}$ is defined as,

$$[L,L]^{avg} = \sum_{i=0}^{N-k} \left\{ \sum_{j=1}^{k} \phi\left(\frac{j}{k}\right) \Delta L_{i+j} \right\}^2 ,$$

where $\Delta L_j = L_j - L_{j-1}$, $\frac{k}{\sqrt{N}} = \theta + \mathscr{O}(N^{-1/4})$ for some $\theta > 0$, and $\phi(x) = min(x, 1-x)$. To reduce the influence of the noise, the pre-averaging approach averages the increments of L.

Hautsch and Podolskij (2010) extend the original pre-averaging method of Jacod et al. (2009) to allow for autocorrelated market microstructure noise. I compare the estimator of Hautsch and Podolskij (2010) using original returns to the Jacod et al. (2009) estimator using adjusted returns. I find that using adjusted returns in the pre-averaging estimator of Jacod et al. (2009) achieves consistency of the integrated volatility estimator even if

[3] The kernel estimator of Barndorff-Nielsen et al. (2011) is also robust to heteroscedastic and autocorrelated noise but converges at the slower rate of $N^{1/5}$.

there is endogeneity. The pre-averaging estimator of Jacod et al. (2009) or Hautsch and Podolskij (2010) using the original returns are inconsistent in the presence of endogeneity. When there is no endogeneity, the pre-averaging estimator of Jacod et al. (2009) using adjusted returns is more precise than the pre-averaging estimator of Hautsch and Podolskij (2010) using original returns (which is consistent in the absence of endogeneity).

To describe my next result, some additional notation is required. In particular, let $(\tilde{F}_t)_{t\geq 0}$ be a stationary q-dependent sequence, $B(q) = E[\xi^2] + E[(\tilde{F}^{'}\beta)^2] + 2\sum_{m=1}^{q}\rho(m)$, where $\rho(m) = cov(\tilde{F}_t^{'}\beta, \tilde{F}_{t+m}^{'}\beta)$. Let $\hat{B}(q)$ be a consistent estimator of $B(q)$. The pre-averaging estimator of Hautsch and Podolskij (2010) using original prices is defined as $[p,p]^{pre} = \frac{12}{\theta\sqrt{N}}[p,p]^{avg} - \frac{12}{\theta^2}\hat{B}(q)$.

Proposition 4. *Suppose Assumptions A and B hold. In the case* $\xi \neq 0$ *a.s.,*

i) If $f(.) \neq 0$, $[p,p]^{pre}$ *is inconsistent.*

ii) If $f(.) \equiv 0$,

$$N^{1/4}\left([p,p]^{pre} - IV\right) \xrightarrow{st} \mathcal{N}(0, \Gamma_\varepsilon(q)),$$

where $\Gamma_\varepsilon(q) = \frac{151}{140}\theta\, IQ + 12\frac{B(q)}{\theta}IV + \frac{96}{\theta^3}B(q)^2$.

According to Proposition 4 ii), the pre-averaging estimator is consistent when there is no endogeneity at the usual $N^{1/4}$ rate of convergence. However, as showed in Part i), in the presence of endogeneity via $f(.)$, the pre-averaging estimator based on original prices is inconsistent. My next theorem characterizes the limiting distribution of the pre-averaging estimator based on adjusted prices $\widehat{p}$. Let the pre-averaging estimator of

Jacod et al. (2009) using the adjusted prices defined as $[\widehat{p},\widehat{p}]^{pre} = \frac{12}{\theta\sqrt{N}}[\widehat{p},\widehat{p}]^{avg} - (\frac{6}{\theta^2 N} + \frac{1}{N})[\widehat{p},\widehat{p}]$.

Theorem 2. *Suppose Assumptions A and B hold. In the case* $\xi \neq 0$ *a.s.,*

(i) $[\widehat{p},\widehat{p}]^{pre} \xrightarrow{P} IV$.

(ii) $N^{1/4}\left([\widehat{p},\widehat{p}]^{pre} - IV\right) \xrightarrow{st} \mathcal{N}(0,\Gamma_\xi)$,

where $\Gamma_\xi = \frac{151}{140}\theta\, IQ + 12\frac{E[\xi^2]}{\theta}IV + \frac{96}{\theta^3}E[\xi^2]^2$.

(iii) $\Gamma_\varepsilon(q) - \Gamma_\xi > 0$,

if $f(.) \equiv 0$ *and* $(\tilde{F}_t)_{t\geq 0}$ *is a stationary q-dependent sequence.*

Theorem 2 (i) shows that the pre-averaging estimator based on adjusted prices is consistent even in the presence of endogeneity. Part (ii) gives the asymptotic distribution of $[\widehat{p},\widehat{p}]^{pre}$. To compare the precision of the usual estimator $[p,p]^{pre}$ and the $[\widehat{p},\widehat{p}]^{pre}$ estimator, I assume that $f(.) \equiv 0$ in (iii) because $[p,p]^{pre}$ is inconsistent otherwise. Result (iii) shows a precision gain if $[\widehat{p},\widehat{p}]^{pre}$ is used to estimate volatility. To conclude, when the noise is only partially absorbed, the pre-averaging estimator based on adjusted prices is robust to endogeneity and is more precise than the usual estimator based on the original prices.

In the next section, I provide a simulation exercise to examine the finite sample properties of the noise parameters estimators and the volatility estimators.

1.4 Monte Carlo evidence

In this section I compare the finite sample simulation results with those predicted by the aforementioned asymptotic theory.

I use a two-factor affine stochastic volatility model as in Andersen et al. (2011). Recall the frictionless price dynamics,

$$dp_t^* = \mu_t dt + \sigma_t dW_t.$$

I take a constant drift $\mu_t = \mu = 0.0314$. The first volatility model $M1$ is a GARCH diffusion model. The instantaneous volatility is defined by the process,

$$d\sigma_t^2 = \kappa(\theta - \sigma_t^2)dt + \sigma\sigma_t^2 dW_t^{(1)},$$

where $\kappa = 0.035,\ \theta = 0.636$, and $\sigma = 0.1439$.

The second model $M2$ is a two-factor affine model. The instantaneous volatility follows a two-factor affine dynamics given by,

$$\sigma_t^2 = \sigma_{1,t}^2 + \sigma_{2,t}^2 \quad d\sigma_{j,t}^2 = \kappa_j(\theta_j - \sigma_{j,t}^2)dt + \eta_j\sigma_{j,t}^2 dW_t^{(j+1)}, \quad j = 1,2,$$

where $\kappa_1 = 0.5708,\ \theta_1 = 0.3257,\ \eta_1 = 0.2286,\ \kappa_2 = 0.0757,\ \theta_2 = 0.1786$, and $\eta_2 = 0.1096$, implying a very volatile first factor and a much more slowly mean reverting second factor.

Now, I turn to the market microstructure noise explanatory variables dynamics. The vector of the noise explanatory variables is $F_t = \begin{pmatrix} q_t & q_t v_t & q_t s_t & d_t^a & d_t^b \end{pmatrix}'$ which defines the trade direction indicator, the signed volume, the signed spread, the ask depth and the bid depth, respectively.

1.4.1 The trade direction indicator

The direction of the trade q_t is triggered by a Bernoulli process with clustering. Trades cluster as buys are likely followed by buys and sells are likely followed by sells. Moreover, some big volume trades are divided into small volume trades and executed consecutively as a series of sells or buys. The Bernoulli process is originally a sequence of random binary variables which are independent. A generalization of a Bernoulli process which incorporates a dependence structure was given by Klotz (1972), in which he considered $q_1, q_2, ..., q_N$, as a stationary two-state Markov chain with state space $\{-1, 1\}$. The parameters of the process are $\alpha = Prob(q_i = 1)$ and λ, which measures the degree of persistence in the chain. The transition matrix is given by,

$$T(\alpha, \lambda) = \begin{pmatrix} \frac{1-2\alpha+\lambda\alpha}{1-\alpha} & \frac{(1-\lambda)\alpha}{1-\alpha} \\ 1-\lambda & \lambda \end{pmatrix}. \tag{1.16}$$

I use the parameters $\alpha = 1/3$ and $\lambda = 0.8$ to simulate the trade direction sequence.

1.4.2 The trading volume

For the trading volume, the process - inspired from Hasbrouck (1999) - is given by,

$$v_i = \mu_i^v + \phi^v (v_{i-1} - \mu_{i-1}^v) + \varepsilon_i^v,$$

where ε^v follows a $\mathcal{N}(0, 0.8)$ and $\phi^v = 0.3$. To allow for an intraday U effect, the deterministic component μ^v of the volume process is specified as a combination of exponential decay functions,

$$\mu_i^v = k_1 + k_2^{open}\exp(-k_3^{open}\tau_i^{open}) + k_2^{close}\exp(-k_3^{close}\tau_i^{close}),$$

where τ_i^{open} is the elapsed time since the opening trade of the day (in hours) and τ_i^{close} is the time remaining before the scheduled market close (in hours). I calibrate the parameters as $k_1 = 15$, $k_2^{open} = 0.5$, $k_3^{open} = 2.5$, $k_2^{close} = 0.2$, and $k_3^{close} = 3.5$.

1.4.3 The bid-ask spread

To simulate the spread series, I follow Hasbrouck (1999) model defined as,

$$\begin{aligned} s_i &= log(A_i - B_i), \\ A_i &= Ceiling[(\exp(p_i^*) + c_i^a)/Tick]Tick, \\ B_i &= Floor[(\exp(p_i^*) - c_i^b)/Tick]Tick, \end{aligned}$$

where the quote exposure costs are assumed to evolve as,

$$\begin{aligned} c_i^a &= \mu_i^c + \phi^c(c_{i-1}^a - \mu_{i-1}^c) + \varepsilon_i^{c^a}, \\ c_i^b &= \mu_i^c + \phi^c(c_{i-1}^b - \mu_{i-1}^c) + \varepsilon_i^{c^b}, \\ \mu_i^c &= z_1 + z_2^{open}\exp(-z_3^{open}\tau_i^{open}) + z_2^{close}\exp(-z_3^{close}\tau_i^{close}), \end{aligned}$$

where τ_i^{open} is the elapsed time since the opening trade of the day (in hours) and τ_i^{close} is the time remaining before the scheduled market close (in hours). I calibrate the parameters as $z_1 = 10$, $z_2^{open} = 0.4$, $z_3^{open} = 1.5$, $z_2^{close} = 0.1$, and $z_3^{close} = 2.5$. The innovations ε^{c^a} and ε^{c^b} are independently distributed as $\mathcal{N}(0, 0.9)$. The Tick size or minimum price variation is 0.01\$. The New York Stock Exchange tick size changed from $1/16$\$ to 0.01\$ on January 29, 2001. Technological innovation is indeed propelling the move in financial markets away from fractional trading and towards decimal trading.

1.4.4 The quoted depths

I generate the quoted depths series using the following AR dynamics,

$$d_i^a = \mu^d + \phi^d (d_{i-1}^a - \mu^d) + \varepsilon_i^{d^a},$$
$$d_i^b = \mu^d + \phi^d (d_{i-1}^b - \mu^d) + \varepsilon_i^{d^b},$$

where ε^{d^a} and ε^{d^b} are independently distributed as $\mathcal{N}(0, 0.5)$, and $\mu^d = 10$.

1.4.5 Other parameters

For the endogeneity, I fix the parameters of the function $f(p^*)$ as follows,

$$f(p^*) = \begin{pmatrix} 0 & 10^{-7} & 2\ 10^{-7} & 5\ 10^{-8} & 5\ 10^{-8} \end{pmatrix}' p^*.$$

The first element of the function vector is null because it corresponds to the trade direction variable which is binary and could not have a semimartingale component. The true parameter β is fixed as,

$$\beta = \begin{pmatrix} 5\ 10^{-4} & -0.6\ 10^{-4} & -3\ 10^{-4} & 2\ 10^{-4} & -2\ 10^{-4} \end{pmatrix}'.$$

I add a white noise ξ for a randomly chosen half of the intraday prices. Precisely, I take $\xi \sim \mathcal{N}(0, 5\ 10^{-6})$.

1.4.6 The results

I run 1000 replications or days. For each day a trade occurs every 5 seconds. A business day has 6.5 working hours. For the simulation results, I report in Table $1.I$ and $1.II$, the bias, variance, and RMSE of the interest variables for model $M1$ and $M2$ respectively. The rows marked ”relative” report the corresponding results in percentage terms. I compare three volatility measures. The first measure is the pre-averaging estimator denoted $[p,p]^{pre}$, the second is the sum of squared adjusted returns denoted $[\widehat{p},\widehat{p}]$, and the third measure is $[p,p]^{this\ paper}$ which is the sum of squared adjusted returns if $\xi \equiv 0$ and the pre-averaging estimator using adjusted returns else. The results show that the price impact regression parameters -β- are estimated very precisely. The $[\widehat{p},\widehat{p}]$ estimator has the largest bias because of the residual noise ξ. The bias of the pre-averaging estimator is due to the inconsistency for integrated volatility of this estimator in the presence of endogeneity between the frictionless price and the liquidity costs. The $[p,p]^{this\ paper}$ has the best performance as advocated by this paper asymptotic theory. I measure performance using the root mean squared error criterion. Both volatility models M1 and M2 have similar results.

1.5 Empirical analysis

This section is organized as follows. First, I check that the noise explanatory variables are of order one. Second, I present results for the noise parameters estimation. I find that all the coefficients are significant at the 95% level. Third, I graphically compare the realized variance and the realized first order covariance for the original prices and the adjusted prices. I find that adjusted prices are closer to a semimartingale than original prices. I propose a formal test to check if the adjusted price is a semimartingale. Fourth, I estimate integrated volatility using original prices and adjusted prices.

I use Alcoa data, listed on the NYSE covering the 01/2006-03/2011 period. The noise explanatory variable q_t is not directly observed, but I infer it from observed series using the Lee-Ready (1991) trade classification algorithm. A trade is classified as a buy if the trade price is closer to the ask than the bid, $q_t = +1$. It is classified as a sale if the trade price is closer to the bid, $q_t = -1$. When matching trades and quotes, I assume a zero time lag because I use recent data. Appendix A details the data manipulation procedure.

As stated before, the volatility signature plot of Andersen et al. (2000) draws the mean of daily realized variances across the sampling frequency of the underlying returns. This plot illustrates the main problem of ultra high frequency data which is the noise contamination problem.

An explanatory variable is valid (i.e. $\mathscr{O}(1)$) if its quadratic variation explodes at high frequencies, as in Assumption B. Since q_t has a Bernoulli distribution, we know that the quadratic variation of q_t explodes at a high frequency. Figure 1.2 uses the signature plot visual tool to verify that the quadratic variation of volume, v_t, explodes at high frequencies. The same explosion of the sum of squared quoted depths and the spread is presented in Figures $1.3 - 1.5$. Therefore, they are valid noise explanatory variables.

I find that the noise explanatory variables coefficients are significant at the 95% confidence level for almost all the business days (cf. Figures 1.6-1.10). The confidence intervals are computed for the worst case $\xi \neq 0$. Indeed, in such case the confidence intervals are larger that for the best case $\xi = 0$. In the former case, they are of order $\mathscr{O}(1/\sqrt{N})$, and in the latter case of order $\mathscr{O}(1/N)$.

The trade indicator q coefficient is positive for all days except one. The signed volume qv coefficient is negative for all days. A transaction with higher number of shares generates a lower cost per share. For the signed spread qs, the coefficient is negative. A wider spread is associated with a smaller buy price and a bigger sell price. The quoted depths coefficients are positive for the ask volume and negative for the bid volume. This is consistent with inventory control matters. If increasing the ask volume, this makes the price higher in an attempt to elicit sales. The same is true for the bid volume.

Figure 1.11 presents the $RC1$ test results. For 344 among 565 business days, the test is not rejected.

Realized variances with the highest frequency underlying returns are plotted in Figure 1.12. Observe that the explosion is less problematic at high frequencies for the adjusted prices. Consequently, the unabsorbed noise has a smaller magnitude than the original noise. The first order autocorrelation for the noise are plotted in Figure 1.11. The original noise first order autocorrelation is negative whereas the unabsorbed noise first order autocorrelation is of smaller magnitude and tends to be rather positive than negative.

To estimate daily integrated volatility, I use the sum of squared adjusted returns as in Theorem 1 if the zero noise test of section 1.2.3 is not rejected, and the pre-averaging estimator robust to dependent noise as in Theorem 2 with adjusted prices if the zero noise test is rejected. I denote such an estimator $RV^{this\ paper}$. I also compute the pre-averaging estimator using original prices. Figure 1.13 plots the pre-averaging estimator and the $RV^{this\ paper}$. I find that, for 202 business days among 565, the confidence intervals of the pre-averaging estimator and the $RV^{this\ paper}$ are non overlapping. This result shows that since for many days, the pre-averaging estimator is statistically different from this paper's estimator, one suspects that the pre-averaging approach underlying assumptions' are unrealistic. Moreover, for Hautsch and Podolskij (2010) empirical results, the estimators are not necessarily positive in all cases and the authors bound them from below by zero. I do the same in this section.

I also divided each daily high frequency data sample into 3 sub-samples : morning period, lunch-time and afternoon period. I find that all the empirical conclusions also apply for the sub-samples (see Fig. 1.14-1.26 in Appendix C). This exercise is helpful if one is interested in the intradaily instead of daily volatility.

1.6 Volatility forecasting and option trading

In this section I evaluate the proposed integrated volatility forecasts in the context of the profits from option pricing and trading economic metric. Using alternative forecasts obtained from alternative volatility estimates, agents price short-term options on Alcoa stock before trading with each other at average prices. The average profit is used as the criterion to evaluate alternative volatility estimates and the corresponding forecasts.

I construct an artificial option market as in Bandi et al. (2008) in order to quantify the economic gain or loss for using alternative Integrated Volatility measures.

Our hypothetical market has 3 traders. Each trader uses a different measure. The first measure is the pre-averaging estimator denoted $[p,p]^{pre}$, the second is the sum of squared adjusted returns denoted $[\widehat{p},\widehat{p}]$, and the third measure is $[p,p]^{this\ paper}$ which is the sum of squared adjusted returns if $\xi \equiv 0$ and the pre-averaging estimator using adjusted returns otherwise.

First, each trader constructs an out-of sample one day ahead variance forecast using his daily variances series and computes his predicted Black-Scholes option price. I focus on an at-the-money price of a 1-day option on a 1 Dollar share of Alcoa or General Motors. The risk free rate is taken to be zero.

Second, the pair-wise trades take place. For two given traders, if the forecast of the first one is higher than the mid-point of the forecasts of the two traders, than the option is perceived as underpriced. And the first trader will buy a straddle (one call and one put) from his counterpart. Then the positions are hedged using the deltas of the options.

Finally, I compute the profits or losses. Each trader averages the two profits or losses from pair-wise trading. I report the average profits across all days in the sample.

The option trading and profit results are computed as in the following three steps,

1-Let σ_t denote the volatility forecast for a given measure. The Black-Scholes option price P_t is given by,

$P_t = 2\Phi(\frac{1}{2}\sigma_t) - 1$, where Φ is the cumulative normal distribution.

2-The daily profit for a trader who buys the straddle is :

$\mid R_t \mid -2P_t + R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$, where the last term corresponds to the hedging, and R_t is the daily return for day t.

The daily profit for a trader who sells the straddle is :

$2P_t - \mid R_t \mid -R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$.

3- I then average the profits and obtain the metric.

I report in Table 1.III the in-sample and the out-of-sample R^2 of the Mincer-Zarnowitz regressions of the realized variance using low frequency returns on a constant and the forecast of volatility using $[p,p]^{pre}$, $[\widehat{p},\widehat{p}]$, and $[p,p]^{this\ paper}$, respectively. The forecasting model is an AR(3) with a rolling window of 100 days. I use Alcoa data, listed in the

NYSE covering the 01/2006-03/2011 period. Both in-sample and out-of-sample R^2 of the $[\widehat{p}, \widehat{p}]$ forecast are the best among the three forecasts. However, the R^2 of $[\widehat{p}, \widehat{p}]$ and $[p, p]^{pre}$ are very close. The same ranking is obtained for the profits, losses in Cents for the option trading exercise that uses the out-of-sample forecasts of the previous forecasting model. The biggest loss is endured by the agent using the pre-averaging estimator whereas the agent using $[\widehat{p}, \widehat{p}]$ has the best profits. The agent using $[p, p]^{this\ paper}$ is ranked as the second and endures a loss.

1.7 Conclusion

I use measures of friction from microstructure theory to absorb the noise that contaminates high frequency prices. I find that explicitly modeling the noise improves the measurement of volatility. If the noise is completely absorbed, the volatility estimator has a convergence rate of $N^{1/2}$ instead of $N^{1/4}$. Instead, if the noise is partially absorbed, the unabsorbed noise is closer to an exogenous white noise than the original noise. In that case, the volatility estimator is more precise since the asymptotic variance is smaller. I focus on integrated volatility estimation, but the approach could improve measurement of intraday quantities such as spot volatility, powers of volatility, leverage effect, and integrated betas in a multivariate setting. Potentially a nonlinear liquidity costs function would capture more noise than a linear one. Indeed, nonlinearities are well documented in market microstructure theory. The resulting econometric model would be a nonparametric price impact regression. Finally, adding jumps in the frictionless price dynamics may not alter the approach of this paper.

	Bias	Variance	RMSE	Relative bias	Relative variance	Relative RMSE
$\hat{\beta}_1$	$3.8091\ 10^{-6}$	$8.0683\ 10^{-9}$	$8.9905\ 10^{-5}$	$7.6182\ 10^{-4}$	$1.6137\ 10^{-6}$	0.0180
$\hat{\beta}_2$	$-5.9750\ 10^{-8}$	$4.1245\ 10^{-12}$	$2.0318\ 10^{-6}$	$9.9583\ 10^{-6}$	$-6.8742\ 10^{-10}$	$-3.3863\ 10^{-4}$
$\hat{\beta}_3$	$-1.0100\ 10^{-6}$	$8.1296\ 10^{-10}$	$2.8530\ 10^{-5}$	$3.3667\ 10^{-4}$	$-2.7099\ 10^{-7}$	-0.0095
$\hat{\beta}_4$	$4.3873\ 10^{-8}$	$7.0284\ 10^{-12}$	$2.6515\ 10^{-6}$	$2.1936\ 10^{-5}$	$3.5142\ 10^{-9}$	0.0013
$\hat{\beta}_5$	$1.6256\ 10^{-8}$	$7.3964\ 10^{-12}$	$2.7197\ 10^{-6}$	$-8.1280\ 10^{-6}$	$-3.6982\ 10^{-9}$	-0.0014
$[p,p]^{pre}$	0.1178	0.0048	0.1366	0.1850	0.0075	0.2146
$[\widehat{p},\widehat{p}]$	5.8962	0.0029	5.8964	9.2617	0.0045	9.2621
$[p,p]^{this\ paper}$	-0.0051	0.0027	0.0519	-0.0080	0.0042	0.0816

Table 1.I – Simulation results, model M1

	Bias	Variance	RMSE	Relative bias	Relative variance	Relative RMSE
$\hat{\beta}_1$	$3.8009\ 10^{-6}$	$7.9438\ 10^{-9}$	$8.9209\ 10^{-5}$	$7.6019\ 10^{-4}$	$1.5888\ 10^{-6}$	0.0178
$\hat{\beta}_2$	$-6.4180\ 10^{-8}$	$4.0616\ 10^{-12}$	$2.0164\ 10^{-6}$	$1.0697\ 10^{-5}$	$-6.7694\ 10^{-10}$	$-3.3606\ 10^{-4}$
$\hat{\beta}_3$	$-9.8321\ 10^{-7}$	$8.0020\ 10^{-10}$	$2.8305\ 10^{-5}$	$3.2774\ 10^{-4}$	$-2.6673\ 10^{-7}$	-0.0094
$\hat{\beta}_4$	$4.4943\ 10^{-8}$	$6.9074\ 10^{-12}$	$2.6286\ 10^{-6}$	$2.2471\ 10^{-5}$	$3.4537\ 10^{-9}$	0.0013
$\hat{\beta}_5$	$1.7845\ 10^{-8}$	$7.2848\ 10^{-12}$	$2.6991\ 10^{-6}$	$-8.9226\ 10^{-6}$	$-3.6424\ 10^{-9}$	-0.0013
$[p,p]^{pre}$	0.0955	0.0026	0.1084	0.1894	0.0052	0.2148
$[\widehat{p},\widehat{p}]$	5.8962	0.0015	5.8964	11.6867	0.0030	11.6869
$[p,p]^{this\ paper}$	-0.0048	0.0013	0.0368	-0.0096	0.0026	0.0730

Table 1.II – Simulation results, model M2

Figure 1.1 – The trade price signature plot.

Figure 1.2 – The signed volume signature plot.

Figure 1.3 – The signed spread signature plot.

Figure 1.4 – The ask depth signature plot.

Figure 1.5 – The bid depth signature plot.

Figure 1.6 – The trade indicator coefficient with 95% confidence interval.

Figure 1.7 – The signed volume coefficient with 95% confidence interval.

Figure 1.8 – The signed spread coefficient with 95% confidence interval.

Figure 1.9 – The ask depth coefficient with 95% confidence interval.

Figure 1.10 – The bid depth coefficient with 95% confidence interval.

Figure 1.11 – Realized covariance test results with 95% confidence band.

Figure 1.12 – The original and adjusted realized variance.

Figure 1.13 – Time series of daily realized measures.

	R^2 in-sample	R^2 out-of-sample	profits/losses in Cents
$[p,p]^{pre}$	0.5113	0.5102	-0.1062
$[\widehat{p},\widehat{p}]$	0.5144	0.5199	0.0319
$[p,p]^{this\ paper}$	0.4935	0.4957	-0.0535

Table 1.III – Alcoa forecasting performance.

1.8 Appendices

Appendix A : data manipulations

As in Barndorff-Nielsen, Hansen, Lunde and Shephard (2008), I do the following :

1-All data :

P1. Delete entries with a time stamp outside the 9 :30 am to 4 pm window when the exchange is open.

P2. Delete entries with a bid, ask or transaction price equal to zero.

P3. Retain entries originating from a single exchange (NYSE in our application). Delete other entries.

2-Quote data only :

Q1. When multiple quotes have the same time stamp, I replace all these with a single entry with the median bid and median ask price.

Q2. Delete entries for which the spread is negative.

Q3. Delete entries for which the spread is more that 50 times the median spread on that day.

Q4. Delete entries for which the mid-quote deviated by more than 10 mean absolute deviations from a rolling centered median (excluding the observation under consideration) of 50 observations (25 observations before and 25 after).

3-Trade data only :

T1. Delete entries with corrected trades. (Trades with a Correction Indicator, CORR 6 = 0).

T2. Delete entries with abnormal Sale Condition. (Trades where COND has a letter code, except for "E" and "F"). See the TAQ 3 User's Guide for additional details about sale conditions.

T3. If multiple transactions have the same time stamp : use the median price.

T4. Delete entries with prices that are above the ask plus the bid-ask spread. Similar for entries with prices below the bid minus the bid-ask spread.

Appendix B : technical proofs

Proof of Proposition 1

(i) Consistency

We have,

$$\begin{aligned}\widehat{\beta}-\beta &= (X^{'}X)^{-1}X^{'}r-\beta\\ &= (X^{'}X)^{-1}X^{'}(r^{*}+X\beta+\Delta\xi)-\beta\\ &= (X^{'}X)^{-1}X^{'}(r^{*}+\Delta\xi)\\ &= \left[N^{-1}X^{'}X\right]^{-1}\left[N^{-1}X^{'}(r^{*}+\Delta\xi)\right]\end{aligned} \tag{A.1}$$

Assumption A (3) gives the limit of the first term

$$\left[N^{-1}X^{'}X\right]\longrightarrow\Omega \tag{A.2}$$

For the second term,

$$\left[N^{-1}X^{'}(r^{*}+\Delta\xi)\right]=\left[N^{-1}X^{'}\Delta\xi\right]+\mathscr{O}(1/\sqrt{N}) \tag{A.3}$$

$$\begin{aligned}\left[N^{-1}X^{'}\Delta\xi\right] &= \left[N^{-1}F^{'}\xi\right]-\left[N^{-1}F^{'}lag(\xi)\right]\\ &-\left[N^{-1}lag(F)^{'}\xi\right]+\left[N^{-1}lag(F)^{'}lag(\xi)\right]\\ &\to 0\end{aligned} \tag{A.4}$$

because F and ξ are independent and $E[F]=E[\xi]=0$. Therefore the second term converges to 0, it implies along with $(A.2)$ the consistency of β.

(ii) The central limit theorem

Recall from $(A.1)$,

$$\sqrt{N}(\widehat{\beta}-\beta)=\left[N^{-1}X^{'}X\right]^{-1}\left[\sqrt{N}^{-1}X^{'}(r^{*}+\Delta\xi)\right] \tag{A.5}$$

We have,

$$\begin{aligned}\sqrt{N}^{-1}X^{'}(r^{*}+\Delta\xi) &= \sqrt{N}^{-1}X^{'}r^{*}+\sqrt{N}^{-1}X^{'}\Delta\xi\\ &= \sqrt{N}^{-1}\sum_{i=1}^{N}(\int_{i-1}^{i}df(p_u^{*})+\tilde{F}_i-\tilde{F}_{i-1})r_i^{*}+\sqrt{N}^{-1}X^{'}\Delta\xi\\ &= \sqrt{N}^{-1}\sum_{i=1}^{N}(\int_{i-1}^{i}df(p_u^{*})\int_{i-1}^{i}dp_u^{*}+(\tilde{F}_i-\tilde{F}_{i-1})r_i^{*})+\underbrace{\sqrt{N}^{-1}X^{'}\Delta\xi}_{the\ dominant\ term}\end{aligned} \tag{A.6}$$

Using Assumption B (iv), we have

$$\sqrt{N}^{-1}X'\Delta\xi \longrightarrow \mathcal{N}((0)_{M\times 1},S) \tag{A.7}$$

Using Theorem 1 of Barndorff-Nielsen et Al. (2008) for example to find the convergence rates of the two first terms in $(A.6)$,

$$\begin{aligned} &\mathcal{O}\left(\sum_{i=1}^{N}(\int_{i-1}^{i} df(p_u^*)\int_{i-1}^{i} dp_u^*)\right) = \mathcal{O}(1/\sqrt{N}) \\ &\mathcal{O}\left(\sum_{i=1}^{N}(\tilde{F}_i-\tilde{F}_{i-1})r_i^*\right) = \mathcal{O}(1) \end{aligned} \tag{A.8}$$

Combining (A.6)-(A.8) gives

$$\sqrt{N}(\hat{\beta}-\beta) = \sqrt{N}^{-1}(\mathcal{O}(1/\sqrt{N})+\mathcal{O}(1)) + \underbrace{\left[N^{-1}X'X\right]^{-1}}_{\to\Omega^{-1}}\underbrace{\sqrt{N}^{-1}X'\Delta\xi}_{\mathcal{O}(1)}$$

which implies the CLT using (A.9).

□

Proof of Proposition 2

(i) Consistency

If $\xi \equiv 0$, we have

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^{'}X)^{-1}X^{'}r - \beta \\
&= (X^{'}X)^{-1}X^{'}(r^{*} + X\beta) - \beta \\
&= (X^{'}X)^{-1}X^{'}r^{*} \\
&= \left[N^{-1}X^{'}X\right]^{-1}\left[N^{-1}X^{'}r^{*}\right]
\end{aligned} \tag{A.9}$$

Using the Cauchy-Schwarz inequality for each M-vector element, we obtain that

$$N^{-1}X^{(m)}r^{*} \leq \sqrt{N}^{-1}\sqrt{\frac{\sum_{i=1}^{N} X_i^{(m)2}}{N}}\sqrt{\sum_{i=1}^{N} r_i^{*2}}\ ,\ m = 1..M \tag{A.10}$$

Since $\sum_{i=1}^{N} r_i^{*2} \xrightarrow{P} IV$ and $\frac{\sum_{i=1}^{N} X_i^{(m)2}}{N} = \mathscr{O}(1)$ than $N^{-1}X^{(m)}r^{*} \xrightarrow{P} 0$.

So $N^{-1}X^{'}r^{*} \xrightarrow{P} 0$. Using $(A.11)$, we obtain $\widehat{\beta}$ consistency using (3) in Assumption B stipulating that $N^{-1}X^{'}X = \mathscr{O}(1)$.

(ii) The central limit theorem

$$\begin{aligned}
X^{'}r^{*} &= \sum_{i=1}^{N}(\int_{i-1}^{i} df(p_u^{*}) + \tilde{F}_i - \tilde{F}_{i-1})r_i^{*} \\
&= \sum_{i=1}^{N}(\int_{i-1}^{i} df(p_u^{*})\int_{i-1}^{i} dp_u^{*} + (\tilde{F}_i - \tilde{F}_{i-1})r_i^{*})
\end{aligned} \tag{A.11}$$

Using Theorem 1 of Barndorff-Nielsen et Al. (2008) we have,

$$\begin{aligned} &\mathscr{O}\left(\sum_{i=1}^{N}(\int_{i-1}^{i} df(p_u^*)\int_{i-1}^{i} dp_u^*)\right) = \mathscr{O}(1/\sqrt{N}) \\ &\sum_{i=1}^{N}(\tilde{F}_i - \tilde{F}_{i-1})r_i^* \xrightarrow{st} \mathscr{N}((0)_{M\times 1}, \Omega^*) \end{aligned} \tag{A.12}$$

notice that $lim\left[\sum_{i=1}^{N} r_i^{*2} X_i X_i^{'}\right] = lim\left[\sum_{i=1}^{N} r_i^{*2}\tilde{X}_i\tilde{X}_i^{'}\right] = \Omega^*$. We have,

$$X^{'}r^* \xrightarrow{st} \mathscr{N}((0)_{M\times 1}, \Omega^*) \tag{A.13}$$

Recall,

$$N(\widehat{\beta} - \beta) = \left[N^{-1}X^{'}X\right]^{-1}\left[X^{'}r^*\right] \tag{A.14}$$

Then we obtain,

$$N(\widehat{\beta} - \beta) \xrightarrow{st} \mathscr{N}((0)_{M\times 1}, \Omega^{-1}\Omega^*\Omega^{-1}) \tag{A.15}$$

□

Proof of Proposition 3

Recall,

$$\begin{aligned} \widehat{r}_i &= r_i - X_i^{'}\widehat{\beta} \\ &= r_i^* + X_i^{'}(\beta - \widehat{\beta}) + \Delta\xi_i. \end{aligned} \tag{A.16}$$

under H_0,

$$\widehat{r}_i = \underbrace{r_i^*}_{\mathscr{O}(1/\sqrt{N})} + \underbrace{X_i^{'}(\beta - \widehat{\beta})}_{\mathscr{O}(1/N)}. \tag{A.17}$$

so the frictionless return dominates the adjusted return. Therefore, we can use the Theorem 1 of Barndorff-Nielsen et Al. (2008) to obtain that,

$$\sqrt{N}\left(\sum_{i=1}^{N}\widehat{r}_i\widehat{r}_{i-1} + \sum_{i=1}^{N}\widehat{r}_i\widehat{r}_{i+1}\right) \xrightarrow{st} \mathscr{N}(0,4IQ) \tag{A.18}$$

so $\sqrt{N}RC1 \xrightarrow{st} \mathscr{N}(0,IQ)$.

□

Proof of Theorem 1

We have in the zero residual noise case,

$$\widehat{r} = r^* + \underbrace{(X-\tilde{X})(\beta-\hat{\beta})}_{endogenous\ noise} + \underbrace{\tilde{X}(\beta-\hat{\beta})}_{exogenous\ noise} \tag{A.19}$$

Since $\mathscr{O}(\beta-\hat{\beta}) = \mathscr{O}(1/N)$. Therefore,

$$\widehat{r} = \underbrace{r^*}_{\mathscr{O}(1/\sqrt{N})} + \underbrace{(X-\tilde{X})(\beta-\hat{\beta})}_{\mathscr{O}(1/N\sqrt{N})} + \underbrace{\tilde{X}(\beta-\hat{\beta})}_{\mathscr{O}(1/N)} \tag{A.20}$$

So the frictionless return dominates the frictions increment and the adjusted return is almost the equal to the frictionless return. So consistency and limit distribution results are the same if the frictionless return were observed i.e.

(i) $[p^*, p^*] \xrightarrow{P} IV.$

(ii) $\sqrt{N}([p^*, p^*] - IV) \xrightarrow{st} \mathcal{N}(0, 2\, IQ).$

□

Proof of Theorem 2

For the usual pre-averaging estimator,

$$\begin{aligned}
&[p,p]^{pre} = [p^* + F'\beta + \xi, p^* + F'\beta + \xi]^{pre} \\
&= [p^* + (f(p^*)' + \tilde{F}')\beta + \xi, p^* + (f(p^*)' + \tilde{F}')\beta + \xi]^{pre} \\
&= [\underbrace{p^* + f(p^*)'\beta}_{semimartingale} + \underbrace{(\tilde{F}'\beta + \xi)}_{autocorrelated\ noise}, p^* + f(p^*)'\beta + (\tilde{F}'\beta + \xi)]^{pre}
\end{aligned} \tag{A.21}$$

Since the noise is autocorrelated, we need an autocorrelated-noise-robust estimator. In Hautsch and Podolskij (2010), the authors extend the pre-averaging estimator to the case of autocorrelated noise. But they restrict the noise to stationarity and q-dependence type. By applying their Lemma 3.1, we have

$$\begin{aligned}
&\frac{12}{\theta\sqrt{N}}[p,p]^{pre} \xrightarrow{P} \left\langle p^* + f(p^*)^{'}\beta, p^* + f(p^*)^{'}\beta \right\rangle \\
&+ \frac{12}{\theta^2}\left(E[\xi^2] + E[(\tilde{F}^{'}\beta)^2] + 2\sum_{m=1}^{q}\rho(m) \right)
\end{aligned} \tag{A.22}$$

which proves the proposition result along with a straightforward application of the theorem 3.3 of Hautsch and Podolskij (2010).

(i) the pre-averaging estimator using adjusted prices.

$$
\begin{aligned}
[\widehat{p},\widehat{p}]^{pre} &= [p^* + F^{'}(\beta-\hat{\beta})+\xi, p^* + F^{'}(\beta-\hat{\beta})+\xi]^{pre} \\
&= [p^* + (f(p^*)^{'} + \tilde{F}^{'})(\beta-\hat{\beta})+\xi, p^* + (f(p^*)^{'} + \tilde{F}^{'})(\beta-\hat{\beta})+\xi]^{pre} \\
&= [\underbrace{p^* + \underbrace{f(p^*)^{'}(\beta-\hat{\beta})}_{endogenous\ noise}}_{semimartingale} + \underbrace{\tilde{F}^{'}(\beta-\hat{\beta})+\xi}_{exogenous\ noise}, p^* + f(p^*)^{'}(\beta-\hat{\beta}) + \tilde{F}^{'}(\beta-\hat{\beta})+\xi]^{pre}
\end{aligned}
\tag{A.23}
$$

In terms of orders of magnitude, the intuition is

$$
\begin{aligned}
\widehat{r} &= r^* + \underbrace{X^*(\beta-\hat{\beta})}_{very\ small\ endogenous\ noise} + \underbrace{\underbrace{\tilde{X}(\beta-\hat{\beta})}_{small} + \underbrace{\Delta\xi}_{big}}_{exogenous\ noise} \\
&= \underbrace{r^* + X(\beta-\hat{\beta})}_{\mathscr{O}(1/\sqrt{N})} + \underbrace{\Delta\xi}_{\mathscr{O}(1)}
\end{aligned}
\tag{A.24}
$$

Since $\mathscr{O}(\beta-\hat{\beta}) = \mathscr{O}(1/\sqrt{N})$. Therefore,

$$
\frac{12}{\theta\sqrt{N}}[\widehat{p},\widehat{p}]^{pre} \longrightarrow plim\left(\sum_{i=1}^{N}(r_i^* + X_i^{'}(\beta-\hat{\beta}))^2\right) + \frac{12}{\theta^2}E[\xi^2] \tag{A.25}
$$

we have,

$$
\begin{aligned}
&\sum_{i=1}^{N}(r_i^* + X_i^{'}(\beta-\hat{\beta}))^2 = \sum_{i=1}^{N}(r_i^*)^2 + \sum_{i=1}^{N}(X_i^{'}(\beta-\hat{\beta}))^2 \\
&+ 2\sum_{i=1}^{N} r_i^* X_i^{'}(\beta-\hat{\beta})
\end{aligned}
\tag{A.26}
$$

The first term converges to IV. For the second term,

$$\begin{aligned}
&\sum_{i=1}^{N}(X_i^{'}(\beta-\hat{\beta}))^2=\sum_{i=1}^{N}(\beta-\hat{\beta})^{'}X_iX_i^{'}(\beta-\hat{\beta})\\
&=(\beta-\hat{\beta})^{'}(\sum_{i=1}^{N}X_iX_i^{'})(\beta-\hat{\beta})\\
&=trace\left((\beta-\hat{\beta})^{'}(\sum_{i=1}^{N}X_iX_i^{'})(\beta-\hat{\beta})\right)\\
&=trace\left(\underbrace{(\sum_{i=1}^{N}X_iX_i^{'})}_{\to N\Omega}\underbrace{(\beta-\hat{\beta})(\beta-\hat{\beta})^{'}}_{\to 2E[\xi^2]\Omega^{-1}/N}\right)
\end{aligned}\tag{A.27}$$

we deduce that $plim(\sum_{i=1}^{N}(X_i^{'}(\beta-\hat{\beta}))^2)=2E[\xi^2]$.

For the third term,

$$\begin{aligned}
&2\sum_{i=1}^{N}r_i^{*}X_i^{'}(\beta-\hat{\beta})=2\sum_{i=1}^{N}r_i^{*}(X_i^{*}+\tilde{X}_i)^{'}(\beta-\hat{\beta})\\
&=\underbrace{2\sum_{i=1}^{N}r_i^{*}(X_i^{*})^{'}}_{\mathscr{O}(1)}\underbrace{(\beta-\hat{\beta})}_{\mathscr{O}(1/\sqrt{N})}+2\sum_{i=1}^{N}r_i^{*}\tilde{X}_i^{'}(\beta-\hat{\beta})\\
&=\mathscr{O}(1/\sqrt{N})+2\underbrace{\sum_{i=1}^{N}r_i^{*}\tilde{X}_i^{'}}_{\to 0}(\beta-\hat{\beta})
\end{aligned}\tag{A.28}$$

so the third term converges to 0. Combining (A.28)-(A.30) gives the result (i).

(ii) the CLT of the pre-averaging estimator using the adjusted prices.

Recall (A.14),

$$\widehat{r}=\underbrace{r^{*}+X(\beta-\hat{\beta})}_{\mathscr{O}(1/\sqrt{N})}+\underbrace{\Delta\xi}_{\mathscr{O}(1)}\tag{A.29}$$

So the CLT is a direct application of the usual pre-averaging estimator CLT with a biais given by (ii).

To correct the biais, we use $\frac{1}{2N}[\widehat{p},\widehat{p}] \to E[\xi^2]$.

(iii) Efficiency gain.

Straightforward because $B(q) > E[\xi^2]$.

□

Appendix C : Intraday results

Figure 1.14 – Signature plot using trade price (intraday).

Figure 1.15 – Signature plot signed volume (intraday).

$$RV(h) = \sum_i (q_i v_i - q_{i-h} v_{i-h})^2$$

Figure 1.16 – Signature plot signed spread (intraday).

$$RV(h) = \sum_i (q_i s_i - q_{i-h} s_{i-h})^2$$

Figure 1.17 – Signature plot ask depth (intraday).

$$RV(h) = \sum_i (D_i - D_{i-h})^2$$

Figure 1.18 – Signature plot bid depth (intraday).

$$RV(h) = \sum_i (D_i - D_{i-h})^2$$

Figure 1.19 – The trade direction indicator coefficient (intraday).

Figure 1.20 – The signed volume coefficient (intraday).

Figure 1.21 – The signed spread coefficient (intraday).

Figure 1.22 – The ask depth coefficient (intraday).

Figure 1.23 – The bid depth coefficient (intraday).

Figure 1.24 – Realized covariance test results (intraday).

The test using adjusted returns is not rejected for 359, 471, 399 among 565 (respectively for mornings, lunch-times, and afternoons).

Figure 1.25 – The original and adjusted realized variance (intraday).

Figure 1.26 – Time series of daily realized variances (intraday).

For 148, 151, 185 among 565 (respectively for mornings, lunch-times, and afternoons), the confidence intervals of the pre-averaging estimator and the $RV^{this\ paper}$ are non overlapping.

ARTICLE 2

VOLATILITY FORECASTING WHEN THE NOISE VARIANCE IS TIME-VARYING

Abstract

This paper analyzes the forecasting of integrated variance when one observes high frequency noisy prices of assets. The paper departs from the literature by assuming that the variance of the noise is time-varying. We assume that the conditional variance of the noise is an affine function of the instantaneous variance of the frictionless price. In this setting, we revisit the results of Andersen et al. (2011) and quantify analytically the predictive ability of various measures of integrated variance. Importantly, the time-varying aspect of the noise variance implies that the forecast of the integrated variance is different from the forecast of a realized measure. We characterize this difference, which is time-varying, and we propose a feasible bias correction. We assess numerically the usefulness of our approach for realistic models. We then study the empirical implication of our method when one deals with forecasting integrated variance or trading option. The empirical results highlight the improvements achieved by assuming a time-varying noise variance.

Key phrases : Realized volatility, volatility forecasting, heteroscedastic noise, eigenfunction stochastic volatility models.

2.1 Introduction

Volatility forecasts are central to many financial issues such as empirical asset pricing finance and risk management ; see Andersen et al. (2006) for different forecast usages. The good performance of the return volatility forecasts using high frequency data was first shown in Andersen et al. (2003).

A problem volatility forecasters face is how to deal with the noise that contaminates the latent frictionless high frequency prices. One answer is to construct volatility forecasts based on low frequency returns in order to limit the impact of the noise accumulation. For instance, Andersen et al. (2003) use intraday returns sampled at a thirty minutes frequency. Another answer is to use robust to market microstructure noise volatility estimators such as the two time scales estimator of Zhang et al. (2005).

The innovation of this paper is to propose a framework under which the realized variance - defined as the sum of the squared intraday returns - based on the highest available frequency returns may improve volatility forecasting if the noise variance is an affine function of the frictionless return volatility. The intuition behind this result is that under this assumption the noise variance contains information about the fundamental volatility. Consequently, the realized volatility measure, although inconsistent, also carries information about the fundamental volatility. Moreover, by properly centering and scaling the realized volatility, we obtain a consistent volatility estimator.

The standard homoscedastic assumption on the noise is convenient to derive consistent robust to noise volatility estimators but it can be rather unrealistic ; see Hansen and Lunde (2006) for the empirical properties of the market microstructure noise. However, the pre-averaging estimator of Jacod et al. (2009) allows for heteroscedasticity in the noise of a general form. Whether the noise is homoscedastic or heteroscedastic, the realized volatility is inconsistent and dominated by robust to noise forecasts.

The i.i.d. assumption for the noise is assumed in most of the forecasting studies. Aït-Sahalia and Mancini (2008) analyze the out-of-sample forecast performance of the two time scales volatility estimator. This estimator is robust to i.i.d. noise but could be inconsistent under heteroscedastic noise as showed in Kalnina and Linton (2008). Apart from individual forecasts, Patton and Sheppard (2009) study optimal forecasts combinations where the forecasts are the commonly used estimators of integrated variance.

Heteroscedasticity for the noise variance is treated in the literature but this paper is the first to assume the presence of the fundamental volatility in the noise variance. Kalnina and Linton (2008) introduce a diurnal heteroscedasticity motivated by the stylized fact in market microstructure theory of the U-shape intradaily spreads. Indeed, the bid-ask spread as a friction measure is an important component of the market microstructure noise. In Bandi et al. (2010), the variance and the kurtosis of the noise are varying across days but not intradaily. Barndorff-Nielsen et al. (2011) allow for intradaily heteroscedastic noise that is independent from the fundamental volatility and derive a consistent kernel estimator.

Our model is empirically motivated by the high R^2 that we obtain by regressing the RV^{all} - the sum of squared returns computed at the highest frequency - on a constant and RV^{pre}, the pre-averaging estimator. This estimator, derived by Jacod et al. (2009), is a consistent estimator of the integrated volatility even under the assumption of heteroscedastic market microstructure noise. We find an R^2 of 0.94 for Alcoa data covering the 01/2009-03/2011 period. Theoretically, under independent and white noise assumptions for the noise, this regression has a small R^2. In this paper we assume that the noise variance is an affine function of the fundamental spot volatility. Our model also nests the common iid noise model in the literature.

This paper is also motivated by a fact observed in financial markets. We observe that during high volatility times - such as for 2008 financial crisis - transitory volatility (which is the noise volatility) is also high. For instance, one observes wide bid-ask spreads, and transaction costs - one of the sources of market microstructure noise - are highly volatile during crisis periods. Consequently, hedging strategies that work well under normal market conditions may deteriorate in performance during crisis periods. In Stoll (2000), the asset volatility is used as an explanatory variable for the bid-ask spread. We plot in Figure 2.1 the time series of the transitory or the noise variance measured by the RV^{all} estimator and the fundamental variance measured by the RV^{pre} -a proxy for fundamental volatility- estimator for Alcoa during 01/2009-03/2011. We observe a clustering of these measures during highly volatile periods.

To theoretically examine the performance of volatility estimators in terms of forecasting, Andersen et al. (2011) use the eigenfunction representation of the general stochastic volatility class of models developed by Meddahi (2001) for a standard i.i.d. market microstructure noise. In this paper, we extend the Andersen et al. (2011) work to analyze the impact - in terms of forecasting performance - of a specific market microstructure noise form. Using the theoretical framework of Andersen et al. (2011), we quantify the forecasting performance improvement if the noise variance is a function of the fundamental volatility. Andersen et al. (2004) and Sizova (2011) also use the eigenfunction stochastic volatility (ESV) framework.

We then present a numerical study with two stochastic volatility models : a GARCH diffusion model and a two-factor affine model. We find that, under our noise variance form assumption, the traditional realized variance based on the highest frequency returns outperforms the kernel, the two time scales, and the pre-averaging estimators under the Mincer-Zarnowitz R^2 metric. The pre-averaging estimator is robust to heteroscedastic noise and is supposed to perform better than the noisy realized variance in terms of forecasting.

To confront with real data our numerical results about the potential out-of-sample performance of the realized volatility for forecasting, we conduct an empirical application with Alcoa data covering the 01/2009-03/2011 period. The competing forecasts of the daily integrated volatility are the realized variance based on the highest frequency returns and some common robust to noise volatility estimators. To assess the performance of the forecasts, we use a Mincer-Zarnowitz type regression as in the theoretical section and an option trading economic gain measure derived by Bandi et al. (2008). Using alternative forecasts, agents price short-term options on the Alcoa stock before trading with each other at average prices. The average profits are used as the criteria to evaluate alternative volatility forecasts. We find that the traditional realized variance based on the highest frequency returns is the best forecast for short and long term horizons as it achieves the highest R^2 in the Mincer-Zarnowitz type regression, and it allows to reach a good option trading gain compared to the overall realized measures that we use.

The rest of the paper is structured as follows. In the next section, we present our model as well as the setting. Section 3 revisits the common realized measures under the heteroscedasticity model of this paper. In the Section 4, we compute analytically the R^2 of the Mincer-Zarnowitz regression that measures the forecasting efficiency of the alternative realized measures using the ESV framework. In Sections 5 and 6, we provide the forecasts in practice as well as estimators of the noise variance parameters, respectively. Sections 7, 8, and 9 present numerical results for two calibrated volatility models and empirical results for Alcoa data. The last section concludes.

2.2 The model

The main goal of this section is to describe our theoretical framework. In particular, we state our assumptions and introduce the model for the variance of the noise. We also define the realized volatility estimator.

We are interested in forecasting the volatility of the frictionless log price denoted p_s^*, and evolving as a semimartingale given by,

$$dp_s^* = \sigma_s dW_s,\ s \in [0,T], \tag{2.1}$$

where W_s is a Wiener process and σ_s is a *càdlàg* volatility function. By assumption, the drift term is zero and W_s and σ_s are independent to exclude leverage and drift effects. These simplifying assumptions could be relaxed using the ESV framework. Andersen et al. (2006) provide a starting point for a direct analytical exploration and quantification of such effects in the case of white noise. In this paper, we are interested in forecasting the latent integrated volatility over one-period horizon,

$$IV_{t+1} = \int_t^{t+1} \sigma_s^2 ds, \tag{2.2}$$

and m-periods horizon,

$$IV_{t+1:t+m} = \sum_{i=1}^{m} IV_{t+i}, \tag{2.3}$$

where m is a positive integer, and $0 < t$. We assume the usual additive form contamination for the observed log price denoted p_s,

$$p_s = p_s^* + u_s, \tag{2.4}$$

where u_s is the market microstructure noise. The standard assumption on the noise is that u_s is i.i.d. and independent from the frictionless price p_s^*. Heteroscedasticity in the noise is accounted for in Kalnina and Linton (2008), Barndorff-Nielsen et al. (2011), Jacod et al. (2009) etc. The two time scales estimator of Zhang et al. (2005), derived under the standard assumption for the noise, has been extended to the multi time scales estimator of Aït-Sahalia et al. (2011) to allow for serial correlation in the noise.

This paper is the first to model the noise heteroscedasticity as a function of the fundamental volatility σ_s. We assume that, given the volatility path, the noise variance is an affine function of the fundamental volatility. Formally, we make the following set of assumptions.

Assumption A

$\forall s, q \in [0, T]$, and conditioning on the volatility path $\{\sigma_\tau, 0 \leq \tau \leq T\}$,

i) $u_s \; and \; u_q$ are independent.

ii) $u_s \; and \; W_q$ are independent.

iii) $Var[u_s \mid \sigma_\tau, 0 \leq \tau \leq T] = a + b\sigma_s^2$, where $a, b \geq 0$.

If $b = 0,\ a \neq 0$, the noise is i.i.d. and it is independent from the frictionless return. This corresponds to the same framework as Andersen et al. (2011). We generalize their framework by allowing $b \neq 0$, in which case the parameter a can be zero. The case where $a = 0$ provides a noise variance that is proportional to the fundamental volatility. In either cases, the analytical ESV framework helps to quantify the impact of each parameter. In Assumption A, the noise parameters a and b are constant across days. An interesting extension to this model is to assume time-varying parameters across days.

A consistent integrated volatility estimator when there is no market microstructure noise is the standard realized volatility given by

$$RV_t^*(h) = \sum_{i=1}^{1/h} r_{t-1+ih}^{*2}, \tag{2.5}$$

where $h = 1/N$ and $r_s^* = p_s^* - p_{s-h}^*$. In practice, the frictionless returns are not observed. We rather dispose of the h-period returns $r_s = p_s - p_{s-h}$. The contaminated and frictionless returns are linked as $r_s = r_s^* + e_s$, where $e_s = u_s - u_{s-h}$. The feasible realized volatility measure based on observed high frequency returns is,

$$RV_t(h) = \sum_{i=1}^{1/h} r_{t-1+ih}^2. \tag{2.6}$$

The realized volatility is inconsistent for integrated volatility estimation because of the noise. We now turn to the analysis of the realized volatility forecasting performance under the noise model of Assumption A.

2.3 The common realized measures under the heteroscedasticity model

The realized variance $RV_t(h)$ is inconsistent under our Assumption A because of the noise. However, since the noise variance is affine in the fundamental volatility, we show in the Proposition 1 how we scale $RV_t(h)$ to obtain a consistent estimator.

Proposition 1. *Under Assumption A,*

$$\frac{hRV_t(h) - 2a}{2b + h} \to IV_t, \tag{2.7}$$

when h goes to zero.

All the technical proofs are in Appendix A. The pre-averaging estimator of Jacod et al. (2009) is robust to our heteroscedasticity noise form. Therefore the pre-averaging estimator is consistent under Assumption A. For the two time scales estimator of Zhang et al. (2005) and the kernel estimator of Barndorff-Nielsen et al. (2008), we do not know whether consistency is achieved under Assumption A.

A standard approach in the literature is to compute the optimal sampling frequency for returns underlying the realized variance RV_t ; see Bandi and Russell (2008) and Zhang et al. (2005). Indeed, while low sampling frequencies reduce the bias of RV_t, they increase its variance. Consequently, we can optimally trading-off bias and variance by choosing the frequency that minimizes the mean squared error. In this section, we aim to find the optimal h in the sense of minimizing the conditional mean squared error (on the volatility path) for RV_t denoted MSE and defined as,

$$MSE(h) = E_\sigma\left[(RV_t(h) - IV_t)^2\right].$$

Proposition 2 gives the optimal sampling frequency expression.

Proposition 2. *Under Assumption A,*

$$MSE(h) = 2hQ_t + \frac{4}{h^2}(a + bIV_t)^2 + o(h). \tag{2.8}$$

When the optimal sampling frequency is high, the following rule-of-thumb applies for the optimal frequency h^*,

$$h^* = \sqrt[3]{\frac{4(a+bIV_t)^2}{Q_t}}, \tag{2.9}$$

where the quarticity Q_t is defined as $\int_{t-1}^{t} \sigma_s^4 ds$.

The form of the optimal frequency given in Proposition 2 is basically the same as the one in Bandi et al. (2010) where the authors find that $h^* = \sqrt[3]{\frac{E[e_t^2]^2}{Q_t}}$. Their optimal frequency is derived under the assumption that the second moment of the noise is constant intradaily but varies across days.

Here we derive the optimal frequency to minimize an estimation error. For the sake of forecasting, one would minimize a forecasting error and find another optimal frequency.

2.4 Forecasting integrated volatility within the ESV framework

Our procedure builds directly on the eigenfunction representation of the general stochastic volatility (ESV) class of models developed by Meddahi (2001). We first describe the ESV framework. Then, we derive the analytical expressions of the Mincer-Zarnowitz regression R^2 which is our main forecast evaluation tool.

2.4.1 The ESV framework

We assume that the spot volatility process is in the ESV class introduced by Meddahi (2001). If we assume that volatility is driven by a single state variable f_t, the spot volatility takes the form,

$$\sigma_t^2 = \sum_{n=0}^{p} a_n P_n(f_t), \tag{2.10}$$

where the integer p may be infinite. We assume the normalization $P_0(f_t) = 1$. The latent state variable evolves as

$$df_t = m(f_t)dt + \sqrt{v(f_t)}dW_t^f, \tag{2.11}$$

where the W_t^f Brownian motion is independent of the W_t Brownian motion driving the frictionless price. Furthermore, the a_n coefficients are real numbers and the $P_n(f_t)$'s denote the eigenfunctions of the infinitesimal generator associated with f_t. In particular, $P_n(f_t)$ are orthogonal and centered at zero,

$$E[P_n(f_t)P_j(f_t)] = 0 \qquad E[P_n(f_t)] = 0, \tag{2.12}$$

and follow first-order autoregressive processes,

$$\forall l > 0, n > 0,\ E[P_n(f_{t+l}) \mid f_\tau, \tau \leq t] = \exp(-\lambda_n l)P_n(f_t), \tag{2.13}$$

where $(-\lambda_n)$ denote the corresponding eigenvalues.

The above class of models includes most diffusive stochastic volatility models in the literature. We now turn to the forecast evaluation within the ESV framework.

2.4.2 Analytical Mincer-Zarnowitz style regression

In this section, we examine the forecasting performance of several volatility estimators. The traditional volatility forecasts are the realized variance $RV_t(h)$ with various sampling frequencies h of intraday returns. The robust to noise forecasts compete with the realized variance. To facilitate the analysis of the realized measures RM_t, whether traditional or robust to noise, we use the quadratic form representation. For a sampling frequency h of intraday returns, the quadratic form representation is given by,

$$RM_t(h) = \sum_{1 \leq i,j \leq 1/h} q_{ij} r_{t-1+ih} r_{t-1+jh}, \tag{2.14}$$

where q_{ij} are weights to be chosen for each realized measure. For instance, the realized variance based on the highest frequency returns available RV_t^{all} is a realized measure with $q_{ij}^{all} = 1$ if $i = j$ and $q_{ij}^{all} = 0$ else. In Andersen et al. (2011), quadratic forms representation of the two time scales estimator, the Zhou's (1996) and the kernel estimators are provided. We recall these forms in Appendix B. Here, we derive the pre-averaging estimator quadratic form ; see the Appendix B for the proof. We show that,

$$q_{ij}^{pre} = \frac{12}{\theta\sqrt{N}} q_{ij}^{\phi} - \frac{6}{\theta^2 N} q_{ij}^{all}, \tag{2.15}$$

where

$$q_{ij}^{\phi} = \sum_{l=0}^{N-k} \delta_{l+1 \leq i \leq l+k} \delta_{l+1 \leq j \leq l+k} \phi\left(\frac{i-l}{k}\right) \phi\left(\frac{j-l}{k}\right), \tag{2.16}$$

and $\delta_{a\leq b\leq c}$ is the indicator function equal to 1 when $a \leq b \leq c$ and 0 otherwise. The tuning parameters of the pre-averaging estimator are θ, the function $\phi(.)$ and the integer k.

The R^2 from the Mincer-Zarnowitz style regression of IV_{t+1} onto a constant and the $RM_t(h)$, is expressed as,

$$R^2(IV_{t+1}, RM_t(h)) = \frac{Cov[IV_{t+1}, RM_t(h)]^2}{Var[IV_{t+1}]Var[RM_t(h)]}. \tag{2.17}$$

This R^2 is our forecasting performance measure. Proposition 3 is instrumental in deriving the following moments in order to compute this measure under our new heteroscedastic noise assumptions (Assumption A). We have,

$$\begin{aligned}
&Cov[IV_{t+1}, RM_t(h)] = \sum_{1\leq i,j\leq 1/h} q_{ij} Cov[IV_{t+1}, r_{t-1+ih} r_{t-1+jh}],\\
&Var[RM_t(h)] = E[RM_t^2(h)] - E[RM_t(h)]^2,\\
&E[RM_t^2(h)] = \sum_{1\leq i,j,k,l\leq 1/h} q_{ij} q_{kl} E[r_{t-1+ih} r_{t-1+jh} r_{t-1+kh} r_{t-1+lh}],\\
&E[RM_t(h)] = \sum_{1\leq i,j\leq 1/h} q_{ij}(h) E[r_{t-1+ih} r_{t-1+jh}].
\end{aligned}$$

More precisely, we derive in Proposition 3 the expressions of $Cov[IV_{t+1}, r_{t-1+ih} r_{t-1+jh}]$, $E[r_{t-1+ih} r_{t-1+jh} r_{t-1+kh} r_{t-1+lh}]$, $E[r_{t-1+ih} r_{t-1+jh}]$, and $Var[IV_{t+1}]$. We denote $E[u_t^2] = V_u$ and $E[u_t^4] = K_u V_u^2$.

Proposition 3. *Under Assumption A,*

$$(a)\ E[r_{t-1+ih}r_{t-1+jh}] = a_0 h + 2V_u\ \ if\ \ i=j$$

$$= -V_u\ \ for\ \ |i-j| = 1.$$

$$(b)\ Cov[IV_{t+1}, r_{t-1+ih}r_{t-1+jh}]$$
$$= \delta_{i,j}(\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}(1-\exp(-\lambda_n h))(1-\exp(-\lambda_n))\exp(-\lambda_n(1-ih))$$
$$+b(1-\delta_{i,j-1})\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-ih))-\exp(-\lambda_n(2-ih))}{\lambda_n}$$
$$+b(1-\delta_{i-1,j})\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-(i-1)h))-\exp(-\lambda_n(2-(i-1)h))}{\lambda_n}),$$

where $\delta_{i,j} = 1$ *if* $i = j$*, and* 0 *otherwise.*

$(c)\ E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}]$

$$= 3a_0^2h^2 + 2(K_u+3)V_u^2 + 12a_0hV_u + 6\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[-1+\lambda_n h+\exp(-\lambda_n h)]$$

$$+6b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h) + 12b\sum_{n=0}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n} \ \ if\ \ i=j=k=l,$$

$$= -(K_u+3)V_u^2 - 3a_0hV_u - 3b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n} - 3b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h)$$

$if\ \ i=j=k=l+1\ or\ i=j+1=k+1=l+1,$

$$= a_0^2h^2 + (K_u+3)V_u^2 + 4a_0hV_u + \sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[1-\exp(-\lambda_n h)]^2 + 2b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h)-\exp(-2\lambda_n h)}{\lambda_n}$$

$$+2b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n} + 2b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h) + b^2\sum_{n=1}^{p}a_n^2\exp(-2\lambda_n h) \ \ if\ \ i=j=k+1=l+1,$$

$$= a_0^2h^2 + 4a_0hV_u + 4V_u^2 + \sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[1-\exp(-\lambda_n h)]^2\exp(-\lambda_n(i-k-1)h)$$

$$+2b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h(i-k))-\exp(-\lambda_n h(i-k-1))}{-\lambda_n}$$

$$+b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h(i-k+1))-\exp(-\lambda_n h(i-k))}{-\lambda_n}$$

$$+b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h(i-k))-\exp(-\lambda_n h(i-k+1))}{\lambda_n}$$

$$+2b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k)) + b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k+1))$$

$$+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k-1)) \ \ if\ \ i=j>k+1, k=l,$$

$$= 2(V_u^2 + b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h)) \ \ if\ \ i=j+1, j=k=l+1,$$

$$= -a_0hV_u - 2V_u^2 - b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k+1)) - b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k))$$

$$-b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h(i-k+1))-\exp(-\lambda_n h(i-k))}{-\lambda_n} \ \ if\ \ i=j>k, k=l+1\ or\ i=j+1, j>k, k=l,$$

$$= V_u^2 + b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k)) \ \ if\ \ i=j+1, j>k, k=l+1,$$

$= 0\ \ else.$

$$(d)\ Var[IV_{t+1}] = 2\sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2}[\exp(-\lambda_n) + \lambda_n - 1].$$

By taking $b = 0$ in Proposition 1, we find the same results as Proposition 2.1 of Andersen et al. (2011). This is coherent with their i.i.d. noise assumption corresponding to $b = 0$ in our framework. In the numerical results subsection, we use Proposition 3 to quantify the forecasting gain for two specific stochastic volatility models.

For longer forecasting horizons $m > 1$, the R^2 from the Mincer-Zarnowitz regression of $IV_{t+1:t+m}$ onto a constant and the $RM_t(h)$ is expressed as,

$$R^2(IV_{t+1:t+m}, RM_t(h)) = \frac{Cov[IV_{t+1:t+m}, RM_t(h)]^2}{Var[IV_{t+1:t+m}]Var[RM_t(h)]}. \tag{2.18}$$

For the numerator we have,

$$Cov[IV_{t+1:t+m}, RM_t(h)] = \sum_{1\leq i,j\leq 1/h} q_{ij} Cov[IV_{t+1:t+m}, r_{t-1+ih} r_{t-1+jh}].$$

Proposition 4 gives the needed expressions to compute R^2 for $m > 1$.

Proposition 4. *Under Assumption A,*

$$
\begin{aligned}
&(a)\ Cov[IV_{t+1:t+m}, r_{t-1+ih}r_{t-1+jh}]\\
&= \delta_{i,j}(\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}(1-\exp(-\lambda_n h))(1-\exp(-\lambda_n m))\exp(-\lambda_n(1-ih))\\
&+b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-ih))-\exp(-\lambda_n(m+1-ih))}{\lambda_n}\\
&+b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-(i-1)h))-\exp(-\lambda_n(m+1-(i-1)h))}{\lambda_n})\\
&-\delta_{i,j-1}b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-ih))-\exp(-\lambda_n(m+1-ih))}{\lambda_n}\\
&-\delta_{i-1,j}b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n(1-(i-1)h))-\exp(-\lambda_n(m+1-(i-1)h))}{\lambda_n}.
\end{aligned}
$$

where $\delta_{i,j} = 1$ *if* $i = j$*, and* 0 *otherwise.*

$$(b)\ Var[IV_{t+1:t+m}] = 2\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[\exp(-\lambda_n m)+\lambda_n m-1].$$

As mentioned for the one-horizon forecasting, by setting $b = 0$ in the multi-period volatility forecasting we find the same expressions as Andersen et al. (2011) i.i.d. noise case.

2.5 The forecast in practice

In the previous section, we assess the forecasting performance for each realized measure. In this section, we explicitly give the forecast under Assumption A and a bias correction. Then, we provide a method to assess the forecasting performance of the realized measures when the latent dependent variable in Mincer-Zarnowitz regression is replaced by a feasible measure of integrated variance.

Let $E_t[.]$ denote the expectation operator conditional on all the past up to time t. Using the quadratic form representation of RM_{t+1}, we have

$$\begin{aligned} E_t[RM_{t+1}] &= \sum_{1\leq i,j\leq 1/h} q_{ij} E_t[r_{t+ih} r_{t+jh}] \\ &= \sum_{1\leq i,j\leq 1/h} q_{ij} E_t[(r^*_{t+ih} + e_{t+ih})(r^*_{t+jh} + e_{t+jh})] \\ &= \sum_{1\leq i,j\leq 1/h} q_{ij} (\underbrace{E_t[r^*_{t+ih} r^*_{t+jh}]}_{=\delta_{ij} E_t[\int_{t+(i-1)h}^{t+ih} \sigma_s^2 ds]} + E_t[e_{t+ih} e_{t+jh}]). \end{aligned} \tag{2.19}$$

If we suppose that $q_{ii} = 1, \forall i = 1..N$, then we have

$$E_t[RM_{t+1}] = E_t[IV_{t+1}] + \sum_{1\leq i,j\leq 1/h} q_{ij} E_t[e_{t+ih} e_{t+jh}]. \tag{2.20}$$

A bias correction is given by $E_t[RM_{t+1}] - \sum_{1\leq i,j\leq 1/h} q_{ij} E_t[e_{t+ih} e_{t+jh}]$ for the realized measures such that $q_{ii} = 1, \forall i = 1..N$. We conclude that, under Assumption A, the forecasting bias is time-varying. If $b = 0$, $E_t[e_{t+ih} e_{t+jh}]$ is constant, and so is the bias correction.

In the R^2 expression of equation (2.17), the integrated volatility regressand is latent. In this section, we replace IV_{t+1} by a feasible estimator denoted $\overline{RM}_{t+1}$ among the realized measures. The R^2 is then written as,

$$R^2(\overline{RM}_{t+1}(h), RM_t(h)) = \frac{Cov[\overline{RM}_{t+1}(h), RM_t(h)]^2}{Var[\overline{RM}_{t+1}(h)] Var[RM_t(h)]}. \tag{2.21}$$

Using the quadratic form representation of $\overline{RM}_{t+1}$ and $RM_t(h)$, we could compute the requisite moments. And, observe that we could maximize the R^2 to find the optimal sampling frequency h for forecasting.

2.6 Estimating the noise parameters

In this section, we examine the estimation of the noise parameters a and b. We also provide a centered and scaled version of the realized variance to obtain a consistent estimator of the integrated variance under Assumption A. Using Proposition 1 and since the pre-averaging estimator is consistent under Assumption A, we have

$$hRV_t(h) = 2a + (2b+h)RV_t^{pre} + \eta_t, \tag{2.22}$$

where η_t is a zero mean residual term, and h is fixed. Seen as a regression of $hRV_t(h)$ on a constant and RV_t^{pre}, the equation (2.22) delivers estimators of the noise parameters. More precisely, the regression constant is $2a$ and the slope is $2b+h$.

We denote $\widehat{a}$ and $\widehat{b}$ the OLS estimators (when T is big and h is fixed) for a and b respectively. Their expressions are,

$$\begin{aligned} \widehat{b} &= \frac{1}{2}\left(\frac{h\sum_{t=1}^{T} RV_t(h)RV_t^{pre}}{\sum_{t=1}^{T}(RV_t^{pre})^2} - h\right), \\ \widehat{a} &= \frac{1}{2}\left(\frac{h\sum_{t=1}^{T} RV_t(h)}{T} - (2\widehat{b}+h)\frac{\sum_{t=1}^{T} RV_t^{pre}}{T}\right). \end{aligned} \tag{2.23}$$

We propose a realized measure that results from our noise heteroscedasticity specific form. We denote $RV_t^{a,b}(h)$ the new realized measure if the noise parameters a and b are known,

$$RV_t^{a,b}(h) = \frac{hRV_t(h) - 2a}{2b + h}. \tag{2.24}$$

and $RV_t^{\widehat{a},\widehat{b}}(h)$ the new realized measure if the noise parameters a and b are estimated by $\widehat{a}$ and $\widehat{b}$ respectively.

$$RV_t^{\widehat{a},\widehat{b}}(h) = \frac{hRV_t(h) - 2\widehat{a}}{2\widehat{b} + h}. \tag{2.25}$$

We show in Proposition 1 the consistency of $RV_t^{a,b}(h)$ when h goes to zero. However, we do not derive the asymptotic distributions of $RV_t^{a,b}(h)$, $\widehat{a}$, $\widehat{b}$, and $RV_t^{\widehat{a},\widehat{b}}(h)$ if T goes to infinity and h goes to zero. This question is important for future work. A first step would be to fix h and let T goes to infinity, then allow h to go to zero while T goes to infinity.

2.7 Numerical results

We follow Andersen et al. (2011) for the volatility models choice. The first model $M1$ is a GARCH diffusion model. The instantaneous volatility is defined by the process,

$$d\sigma_t^2 = \kappa(\theta - \sigma_t^2)dt + \sigma\sigma_t^2 dW_t^{(2)},$$

where $\kappa = 0.035,\ \theta = 0.636$, and $\ \psi = 0.296$.

The second model $M2$ is a two-factor affine model. The instantaneous volatility is given by,

$$\sigma_t^2 = \sigma_{1,t}^2 + \sigma_{2,t}^2 \quad d\sigma_{j,t}^2 = \kappa_j(\theta_j - \sigma_{j,t}^2)dt + \eta_j\sigma_{j,t}dW_t^{(j+1)}, \quad j = 1,2,$$

where $\kappa_1 = 0.5708,\ \theta_1 = 0.3257,\ \eta_1 = 0.2286,\ \kappa_2 = 0.0757,\ \theta_2 = 0.1786$, and $\eta_2 = 0.1096$, implying a very volatile first factor and a much more slowly mean reverting second factor.

We chose two scenarios for the parameter b. The sample size is $N = 1/h = 1440$ which is equivalent to a trade each 15 seconds for a 6 hours daily market. The realized variance RV^{mse} is based on the frequency $\sqrt[3]{\frac{4(a+bE[IV_t])^2}{E[Q_t]}}$ instead of the hard-to-estimate frequency given in (2.9). Andersen et al. (2011) also replace the quarticity by its unconditional expectation. The expressions of the realized measures are given in Appendix B. The alternative realized measures are : RV^{all} (the realized variance based on the highest frequency returns), RV^{sparse} (the realized variance based on sub-sampled returns $1/h = 1440/5$), $RV^{average}$ (the average of the sparse estimators that differs in the first used observation to compute RV^{sparse}), RV^{TS} (the two time scales estimator), RV^{Zhou} (the Zhou estimator), RV^{Kernel} (the kernel estimator), RV^{pre} (the pre-averaging estimator), and RV^{mse} (the realized variance based on optimal frequency returns).

For each scenario and each model, we report in Table 2.I the mean, variance and mean-squared-error for the competing realized measures. As anticipated, the 'all' estimator is heavily biased whereas the new realized measure reduces the 'all' estimator bias. When varying b, the pre-averaging estimator characteristics are almost unchanged which is coherent with its robustness to heteroscedastic noise property. The two time scales estimator achieve a very good performance as measured by the MSE for both scenarios and models.

In tables 2.II and 2.III we report the correlations among the alternatives realized measures for models $M1$ and $M2$ respectively. We provide in Appendix C the analytical expressions for the true volatility and realized measures correlations. The realized variance RV^{mse} is the most correlated with the pre-averaging estimator.

In Table 2.IV, we compute R^2 for different values of b. The pre-averaging estimator is robust to heteroscedastic noise so varying b do not change R^2. However, the traditional noisy realized variance estimator computed at the highest frequency dominates when b is high. This corroborates Assumption A intuition of the noise containing information about the frictionless return volatility. As the forecasting horizon increases, each realized measure has a bigger R^2. Moreover, the realized variance based on the highest frequency returns performs well for multi-period volatility forecasting. Finally, we notice that $RV^{average}$ achieves a very good performance, as it was also the case in Andersen et al. (2011) where $b = 0$.

In the next section, we turn to the empirical forecasting gain of the realized volatility using real data.

2.8 Forecasting analysis with real data

The goal of this section is to investigate with real data the forecasting performance of the competing volatility estimators. To evaluate alternative predictors, we use the Mincer-Zarnowitz regression as in section 4. The proxy for the true IV or the dependent variable in the Mincer-Zarnowitz regression is the realized variance where returns are sampled every 300 ticks, RV^{low}. I focus on a one-day, 5-days and 20-days ahead forecast horizon. The Mincer-Zarnowitz regression is given by,

$$IV_j = b_0 + b_1\widehat{IV}_{1,j} + b_2\widehat{IV}_{2,j} + error_j, \tag{2.26}$$

where $\widehat{IV}_{1,j}$ and $\widehat{IV}_{2,j}$ are the predictors and RV^{low} is a proxy for IV. The subscript j refers to the days of the sample.

We use trade prices of Alcoa during $01/2009-03/2011$. In Table 2.V and 2.VI, we present descriptive statistics and correlations for the alternative realized measure. We report the Mincer-Zarnowitz R^2 in Table 2.VII showing that the pre-averaging predictor does not necessarily outperform the inconsistent realized volatility out of sample. As the forecasting horizon grows, the forecasting performance increases for the realized measures. When adding the 'all' estimator in the Mincer-Zarnowitz regression, we notice an important improvement in R^2 as advocated in the theoretical study of this paper. We can further improve the R^2 providing a practical adjustment as in Andersen et al. (2005)[1]. In the next section, we propose another forecasting performance measure.

2.9 Option trading analysis with real data

In this section we evaluate the proposed integrated volatility forecasts in the context of the profits from option pricing and trading economic metric. Using alternative forecasts obtained in the previous section, agents price short-term options on Alcoa stock before trading with each other at average price. The average profit is used as the criterion to evaluate alternative volatility estimates and the corresponding forecasts.

We construct an hypothetical option market as in Bandi et al. (2008) in order to quantify the economic gain or loss for using alternative Integrated Volatility measures.

Our hypothetical market has 8 traders. Each trader uses one from the following realized measures : RV^{all}, RV^{sparse}, $RV^{average}$, RV^{TS}, RV^{Zhou}, RV^{Kernel}, RV^{pre}, and RV^{mse}. The

[1] see the Appendix D for more details.

quadratic form representations of the realized measures are given in Appendix B.

First each trader constructs an out-of sample one day ahead variance forecast using his daily variances series and computes his predicted Black-Scholes option price. We focus on an at-the-money price of a 1-day option on a 1 Dollar share of Alcoa or General Motors. The risk free rate is taken to be zero.

Second, the pair-wise trades take place. For two given traders, if the forecast of the first one is higher than the mid-point of the forecasts of the two traders, than the option is perceived as underpriced. And the first trader will buy a straddle (one call and one put) from his counterpart. Then the positions are hedged using the deltas of the options.

Finally, we compute the profits or losses. Each trader averages the 8 profits or losses from pair-wise trading. We then average across days.

Option trading and profit results are computed as in the following three steps.

1-Let σ_t the volatility forecast for a given measure. The Black-Scholes option price P_t is given by :

$P_t = 2\Phi(\frac{1}{2}\sigma_t) - 1$, where Φ is the cumulative normal distribution.

2-The daily profit for a trader who buys the straddle is :

$\mid R_t \mid -2P_t + R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$, where the last term corresponds to the hedging, and R_t is the daily return for day t.

The daily profit for a trader who sells the straddle is :

$2P_t - \mid R_t \mid -R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$.

3- We then average the profits and obtain the metric.

We obtain the profits, losses in Cents in Table 2.VIII for different realized measures. The traditional realized variance RV^{all} achieves the best profit. All of the estimators

$RV^{pre}, RV^{kernel}, RV^{TS}, RV^{Zhou}, RV^{mse}$ endure losses for the agents using them as forecasts. Compared with the forecasting performance results using the Mincer-Zarnowitz regression, the option trading exercise provide similar rankings.

2.10 Conclusion

This paper quantifies the gain for volatility forecasting performance if the noise heteroscedasticity form is an affine function of the fundamental volatility. We use the eigenfunction stochastic volatility theoretical framework. If our model is true, using robust to noise volatility estimator for forecasting do not make profit of the fundamental volatility information in the noise volatility. The traditional realized variance computed using high frequency intraday returns exploit this information though.

Figure 2.1 – Time series for Alcoa trade price realized variance.

Model	M1			M2		
	Mean	Variance	MSE	Mean	Variance	MSE
IV_t	0.6360	0.1681	0.1681	0.5043	0.0262	0.0262
$b = 0.35\%$						
RV_t^{all}	9.7943	20.8353	104.7104	7.7662	3.3443	56.0798
RV_t^{sparse}	2.4676	1.5894	4.9444	1.9566	0.2742	2.3836
$RV_t^{average}$	2.4668	1.5419	4.8937	1.9512	0.2459	2.3396
RV_t^{TS}	0.5133	0.1205	0.1356	0.4023	0.0231	0.0335
RV_t^{Zhou}	0.6423	0.3346	0.3346	0.5093	0.1166	0.1167
RV_t^{Kernel}	0.6423	0.2016	0.2016	0.5093	0.0437	0.0438
RV_t^{pre}	0.5793	0.2065	0.2097	0.4723	0.0683	0.0693
$RV_t^{a,b}$	0.6990	0.2050	0.2090	0.5543	0.0329	0.0354
RV_t^{mse}	0.6423	0.3391	0.3391	0.5093	0.1188	0.1189
$b = 0.45\%$						
RV_t^{all}	9.7943	32.9596	116.8347	7.7662	5.2379	57.9734
RV_t^{sparse}	2.4676	2.2309	5.5858	1.9566	0.3747	2.4841
$RV_t^{average}$	2.4668	2.1819	5.5338	1.9512	0.3457	2.4393
RV_t^{TS}	0.5133	0.1220	0.1370	0.4023	0.0235	0.0339
RV_t^{Zhou}	0.6423	0.3526	0.3526	0.5093	0.1200	0.1201
RV_t^{Kernel}	0.6423	0.2053	0.2053	0.5093	0.0447	0.0447
RV_t^{pre}	0.5793	0.2121	0.2153	0.4723	0.0701	0.0712
$RV_t^{a,b}$	0.6850	0.1962	0.1986	0.5432	0.0311	0.0326
RV_t^{mse}	0.6423	0.3570	0.3571	0.5093	0.1222	0.1223

Table 2.I – Mean, Variance and MSE of the realized measures.

Note : The size of the intraday return h is 1/1440 for RV^{all}. RV_t^{sparse} is computed with $h = 5/1440$ as well as the $RV^{average}$ estimator. The noise-to-signal ratio is equal to 0.5%, which is defined as $V_u/E[IV_t]$. Recall that $V_u = a + bE[\sigma_t^2]$ under Assumption A.

	RV_t^{all}	RV_t^{sparse}	$RV_t^{average}$	RV_t^{TS}	RV_t^{Zhou}	RV_t^{Kernel}	RV_t^{pre}	$RV_t^{a,b}$	RV_t^{mse}
Model M1									
$b = 0.35\%$									
IV_t	0.9952	0.9808	0.9953	0.9503	0.7137	0.9194	0.8258	0.9952	0.7089
RV_t^{all}	1.00	0.9807	0.9952	0.9373	0.6842	0.9018	0.8201	1.0000	0.7060
RV_t^{sparse}	-	1.00	0.9854	0.9526	0.7093	0.9151	0.8394	0.9807	0.7263
$RV_t^{average}$	-	-	1.00	0.9667	0.7197	0.9286	0.8518	0.9952	0.7370
RV_t^{TS}	-	-	-	1.00	0.7801	0.9565	0.8961	0.9373	0.7844
RV_t^{Zhou}	-	-	-	-	1.00	0.7955	0.6139	0.6842	0.5146
RV_t^{Kernel}	-	-	-	-	-	1.00	0.7726	0.9018	0.6369
RV_t^{pre}	-	-	-	-	-	-	1.00	0.8201	0.9761
$RV_t^{a,b}$	-	-	-	-	-	-	-	1.00	0.7060
RV_t^{mse}	-	-	-	-	-	-	-	-	1.00
$b = 0.45\%$									
IV_t	0.9969	0.9860	0.9966	0.9465	0.6966	0.9129	0.8122	0.9969	0.6923
RV_t^{all}	1.00	0.9859	0.9965	0.9362	0.6720	0.8986	0.8082	1.0000	0.6905
RV_t^{sparse}	-	1.00	0.9894	0.9519	0.6951	0.9122	0.8275	0.9859	0.7104
$RV_t^{average}$	-	-	1.00	0.9621	0.7025	0.9219	0.8364	0.9965	0.7179
RV_t^{TS}	-	-	-	1.00	0.7679	0.9528	0.8878	0.9362	0.7724
RV_t^{Zhou}	-	-	-	-	1.00	0.7839	0.5916	0.6720	0.4913
RV_t^{Kernel}	-	-	-	-	-	1.00	0.7548	0.8986	0.6153
RV_t^{pre}	-	-	-	-	-	-	1.00	0.8082	0.9759
$RV_t^{a,b}$	-	-	-	-	-	-	-	1.00	0.6905
RV_t^{mse}	-	-	-	-	-	-	-	-	1.00

Table 2.II – Correlations of the realized measures under model M1.

	RV_t^{all}	RV_t^{sparse}	$RV_t^{average}$	RV_t^{TS}	RV_t^{Zhou}	RV_t^{Kernel}	RV_t^{pre}	$RV_t^{a,b}$	RV_t^{mse}
Model M2									
$b = 0.35\%$									
IV_t	0.5116	0.6048	0.6370	0.8497	0.4757	0.7767	0.5808	0.5116	0.3929
RV_t^{all}	1.00	0.9339	0.9835	0.8085	0.4081	0.7266	0.5661	1.0000	0.4655
RV_t^{sparse}	-	1.00	0.9495	0.8560	0.4669	0.7661	0.6125	0.9339	0.5108
$RV_t^{average}$	-	-	1.00	0.9015	0.4916	0.8065	0.6452	0.9835	0.5376
RV_t^{TS}	-	-	-	1.00	0.6238	0.8867	0.7460	0.8085	0.6363
RV_t^{Zhou}	-	-	-	-	1.00	0.6514	0.3106	0.4081	0.2415
RV_t^{Kernel}	-	-	-	-	-	1.00	0.4627	0.7266	0.3413
RV_t^{pre}	-	-	-	-	-	-	1.00	0.5661	0.9850
$RV_t^{a,b}$	-	-	-	-	-	-	-	1.00	0.4655
RV_t^{mse}	-	-	-	-	-	-	-	-	1.00
$b = 0.45\%$									
IV_t	0.5053	0.5896	0.6123	0.8426	0.4693	0.7691	0.5750	0.5053	0.3882
RV_t^{all}	1.00	0.9518	0.9883	0.8122	0.4168	0.7323	0.5643	1.0000	0.4630
RV_t^{sparse}	-	1.00	0.9630	0.8586	0.4682	0.7704	0.6089	0.9518	0.5059
$RV_t^{average}$	-	-	1.00	0.8915	0.4860	0.7997	0.6324	0.9883	0.5251
RV_t^{TS}	-	-	-	1.00	0.6223	0.8855	0.7444	0.8122	0.6346
RV_t^{Zhou}	-	-	-	-	1.00	0.6497	0.3076	0.4168	0.2388
RV_t^{Kernel}	-	-	-	-	-	1.00	0.4589	0.7323	0.3375
RV_t^{pre}	-	-	-	-	-	-	1.00	0.5643	0.9850
$RV_t^{a,b}$	-	-	-	-	-	-	-	1.00	0.4630
RV_t^{mse}	-	-	-	-	-	-	-	-	1.00

Table 2.III – Correlations of the realized measures under model M2.

b	Model	M1			M2		
	Horizon m	1	5	20	1	5	20
0.35%	$R^2(IV_{t+1:t+m}, RV_t^{all})$	0.9455	0.8626	0.6238	0.6644	0.4291	0.2063
	$R^2(IV_{t+1:t+m}, RV_t^{a,b})$	0.9455	0.8626	0.6238	0.6644	0.4291	0.2063
	$R^2(IV_{t+1:t+m}, RV_t^{sparse})$	0.9183	0.8379	0.6058	0.6004	0.3877	0.1864
	$R^2(IV_{t+1:t+m}, RV_t^{average})$	0.9457	0.8629	0.6239	0.6656	0.4299	0.2067
	$R^2(IV_{t+1:t+m}, RV_t^{TS})$	0.8622	0.7867	0.5689	0.4972	0.3211	0.1544
	$R^2(IV_{t+1:t+m}, RV_t^{Zhou})$	0.4862	0.4436	0.3208	0.1573	0.1015	0.0488
	$R^2(IV_{t+1:t+m}, RV_t^{kernel})$	0.8070	0.7363	0.5324	0.4193	0.2708	0.1302
	$R^2(IV_{t+1:t+m}, RV_t^{pre})$	0.6509	0.5939	0.4294	0.2306	0.1489	0.0716
	$R^2(IV_{t+1:t+m}, RV_t^{mse})$	0.4798	0.4378	0.3165	0.1544	0.0997	0.0479
0.45%	$R^2(IV_{t+1:t+m}, RV_t^{all})$	0.9488	0.8656	0.6259	0.6734	0.4349	0.2091
	$R^2(IV_{t+1:t+m}, RV_t^{a,b})$	0.9488	0.8656	0.6259	0.6734	0.4349	0.2091
	$R^2(IV_{t+1:t+m}, RV_t^{sparse})$	0.9280	0.8467	0.6123	0.6232	0.4025	0.1935
	$R^2(IV_{t+1:t+m}, RV_t^{average})$	0.9481	0.8650	0.6255	0.6717	0.4338	0.2086
	$R^2(IV_{t+1:t+m}, RV_t^{TS})$	0.8554	0.7805	0.5643	0.4888	0.3157	0.1518
	$R^2(IV_{t+1:t+m}, RV_t^{Zhou})$	0.4633	0.4227	0.3056	0.1534	0.0991	0.0476
	$R^2(IV_{t+1:t+m}, RV_t^{kernel})$	0.7957	0.7260	0.5250	0.4120	0.2661	0.1279
	$R^2(IV_{t+1:t+m}, RV_t^{pre})$	0.6295	0.5744	0.4153	0.2255	0.1457	0.0700
	$R^2(IV_{t+1:t+m}, RV_t^{mse})$	0.4575	0.4174	0.3018	0.1507	0.0973	0.0468

Table 2.IV – R^2 for the integrated variance forecasts.

	Mean	Variance	Skewness	Kurtosis	Minimum
RV_t^{low}	7.649	79.0	3.178	20.111	0.779
RV_t^{all}	11.540	180.1	2.724	12.200	1.334
RV_t^{sparse}	8.387	93.4	2.775	12.950	0.885
$RV_t^{average}$	8.323	91.3	2.741	12.588	0.864
RV_t^{TS}	7.520	75.4	2.782	12.910	0.723
RV_t^{Zhou}	7.288	67.0	2.867	14.222	0.654
RV_t^{Kernel}	7.489	74.4	2.854	13.931	0.668
RV_t^{pre}	7.506	81.9	2.959	15.080	0.681
RV_t^{mse}	8.178	87.5	2.818	13.438	0.979

Table 2.V – Descriptive statistics for the realized measures, Alcoa data.

	RV_t^{low}	RV_t^{all}	RV_t^{sparse}	$RV_t^{average}$	RV_t^{TS}	RV_t^{Zhou}	RV_t^{Kernel}	RV_t^{pre}	RV_t^{mse}
RV_t^{low}	1.00	0.922	0.954	0.953	0.955	0.950	0.957	0.973	0.962
RV_t^{all}	-	1.00	0.979	0.980	0.961	0.958	0.960	0.957	0.968
RV_t^{sparse}	-	-	1.00	0.999	0.996	0.993	0.994	0.988	0.993
$RV_t^{average}$	-	-	-	1.00	0.997	0.993	0.995	0.988	0.994
RV_t^{TS}	-	-	-	-	1.00	0.996	0.997	0.989	0.993
RV_t^{Zhou}	-	-	-	-	-	1.00	0.995	0.981	0.985
RV_t^{Kernel}	-	-	-	-	-	-	1.00	0.989	0.991
RV_t^{pre}	-	-	-	-	-	-	-	1.00	0.994
RV_t^{mse}	-	-	-	-	-	-	-	-	1.00

Table 2.VI – Correlations of the realized measures, Alcoa data.

Horizon	1	5	20
RV_t^{all}	0.701	0.802	0.773
RV_t^{sparse}	0.647	0.723	0.695
$RV_t^{average}$	0.646	0.726	0.698
RV_t^{TS}	0.611	0.683	0.655
RV_t^{Zhou}	0.597	0.668	0.634
RV_t^{Kernel}	0.611	0.682	0.651
RV_t^{pre}	0.612	0.678	0.658
RV_t^{mse}	0.640	0.706	0.684
RV_t^{sparse},RV_t^{all}	0.707	0.820	0.791
$RV_t^{average}$,RV_t^{all}	0.709	0.819	0.790
RV_t^{TS},RV_t^{all}	0.709	0.819	0.790
RV_t^{Zhou},RV_t^{all}	0.712	0.824	0.800
RV_t^{Kernel},RV_t^{all}	0.708	0.818	0.791
RV_t^{pre},RV_t^{all}	0.706	0.817	0.784
RV_t^{mse},RV_t^{all}	0.703	0.814	0.782

Table 2.VII – R^2 for volatility forecasts, Alcoa data.

	Profits/Losses	Ranking
RV_t^{all}	0.172	1
RV_t^{sparse}	0.034	3
$RV_t^{average}$	0.109	2
RV_t^{TS}	-0.387	5
RV_t^{Zhou}	-0.482	7
RV_t^{Kernel}	-0.477	6
RV_t^{pre}	-0.312	4
RV_t^{mse}	-0.487	8

Table 2.VIII – Rank by annualized daily profits, Alcoa data.

2.11 Appendices

Appendix A : Technical proofs

Proof of Proposition 1 :

We have,

$$RV_t(h) = RV_t^*(h) + \sum_{i=1}^{1/h} e_{t-1+ih}^2 + 2\sum_{i=1}^{1/h} r_{t-1+ih}^* e_{t-1+ih}.$$

When h goes to zero, the first term $RV_t^*(h)$ converges to IV_t and the last term goes to zero. Therefore, along with Assumption A iii) we obtain that

$$hRV_t(h) = 2a + (h+2b)IV_t + o(h), \tag{A.1}$$

which gives (2.7).

Proof of Proposition 2 :

$$\begin{aligned} MSE(h) &= E_\sigma\left[(RV_t(h) - IV_t)^2\right] \\ &= Var_\sigma[RV_t(h)] + (E_\sigma[RV_t(h)] - IV_t)^2 \end{aligned} \tag{A.2}$$

Recall the equality,

$$RV_t(h) = RV_t^*(h) + \sum_{i=1}^{1/h} e_{t-1+ih}^2 + 2\sum_{i=1}^{1/h} r_{t-1+ih}^* e_{t-1+ih}. \tag{A.3}$$

For the bias term,

$$\begin{aligned}
E_\sigma[RV_t(h)] &= E_\sigma[RV_t^*(h)] + \sum_{i=1}^{1/h} E_\sigma[e_{t-1+ih}^2] + 2\sum_{i=1}^{1/h} \underbrace{E_\sigma[r_{t-1+ih}^* e_{t-1+ih}]}_{=0} \\
&= \sum_{i=1}^{1/h} E_\sigma(r_{t-1+ih}^{*2}) + E_\sigma[(u_{t-1+ih} - u_{t-1+(i-1)h})^2] \\
&= \sum_{i=1}^{1/h} \int_{t-1+(i-1)h}^{t-1+ih} \sigma_s^2 ds + \sum_{i=1}^{1/h} (Var_\sigma[u_{t-1+ih}] + Var_\sigma[u_{t-1+(i-1)h}]) \\
&= IV_t + 2a/h + b\sum_{i=1}^{1/h} (\sigma_{t-1+ih}^2 + \sigma_{t-1+(i-1)h}^2)
\end{aligned} \tag{A.4}$$

For the variance term,

$$\begin{aligned}
Var_\sigma[RV_t(h)] &= Var_\sigma[RV_t^*(h)] + Var_\sigma\left[\sum_{i=1}^{1/h} e_{t-1+ih}^2\right] + Var_\sigma\left[2\sum_{i=1}^{1/h} r_{t-1+ih}^* e_{t-1+ih}\right] \\
&+ 2Cov_\sigma\left[RV_t^*(h), \sum_{i=1}^{1/h} e_{t-1+ih}^2\right] + 2Cov_\sigma\left[RV_t^*(h), 2\sum_{i=1}^{1/h} r_{t-1+ih}^* e_{t-1+ih}\right] \\
&+ 2Cov_\sigma\left[\sum_{i=1}^{1/h} e_{t-1+ih}^2, 2\sum_{i=1}^{1/h} r_{t-1+ih}^* e_{t-1+ih}\right]
\end{aligned} \tag{A.5}$$

$$Var_\sigma[RV_t^*(h)] = 2hQ_t + o(h) \tag{A.6}$$

where $Q_t = \int_{t-1}^{t} \sigma_s^4 ds$ is the integrated quarticity.

$$
\begin{aligned}
&Var_\sigma[\sum_{i=1}^{1/h} e_{t-1+ih}^2] = \sum_{i=1}^{1/h} Var_\sigma[e_{t-1+ih}^2] + \sum_{i,j=1:i\neq j}^{1/h} Cov_\sigma[e_{t-1+ih}^2, e_{t-1+jh}^2] \\
&= \sum_{i=1}^{1/h} \Big(Var_\sigma[u_{t-1+ih}^2] + Var_\sigma[u_{t-1+(i-1)h}^2] + 4Var_\sigma[u_{t-1+ih}]Var_\sigma[u_{t-1+(i-1)h}] \Big) \\
&+ \sum_{i,j=1:i\neq j}^{1/h} Cov_\sigma[(u_{t-1+ih} - u_{t-1+(i-1)h})^2, (u_{t-1+jh} - u_{t-1+(j-1)h})^2] \\
&= \sum_{i=1}^{1/h} \Big(Var_\sigma[u_{t-1+ih}^2] + Var_\sigma[u_{t-1+(i-1)h}^2] + 4Var_\sigma[u_{t-1+ih}]Var_\sigma[u_{t-1+(i-1)h}] \Big) \\
&+ 2 \sum_{i=1}^{1/h-1} Var_\sigma[u_{t-1+ih}^2] \\
&= 4 \sum_{i=1}^{1/h-1} Var_\sigma[u_{t-1+ih}^2] + 4 \sum_{i=1}^{1/h} [a + b\sigma_{t-1+ih}^2][a + b\sigma_{t-1+(i-1)h}^2] + Var_\sigma[u_{t-1}^2] + Var[u_t^2] \\
&= 4 \sum_{i=1}^{1/h-1} \big[K_u V_u^2 - (a + b\sigma_{t-1+ih}^2)^2 \big] + 4 \sum_{i=1}^{1/h} [a + b\sigma_{t-1+ih}^2][a + b\sigma_{t-1+(i-1)h}^2] \\
&+ \big[K_u V_u^2 - (a + b\sigma_{t-1}^2)^2 \big] + \big[K_u V_u^2 - (a + b\sigma_t^2)^2 \big]
\end{aligned}
\tag{A.7}
$$

$$\begin{aligned}
&Var_{\sigma}[2\sum_{i=1}^{1/h} r^*_{t-1+ih} e_{t-1+ih}] = 4\sum_{i,j=1}^{1/h} Cov_{\sigma}[r^*_{t-1+ih}e_{t-1+ih}, r^*_{t-1+jh}e_{t-1+jh}] \\
&= 4\sum_{i=1}^{1/h} Var_{\sigma}[r^*_{t-1+ih}e_{t-1+ih}] \\
&= 4\sum_{i=1}^{1/h} E_{\sigma}[r^{*2}_{t-1+ih}e^2_{t-1+ih}] \\
&= 4\sum_{i=1}^{1/h} E_{\sigma}[r^{*2}_{t-1+ih}]E_{\sigma}[e^2_{t-1+ih}] \\
&= 4\sum_{i=1}^{1/h}(\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds)[2a + b\sigma^2_{t-1+ih} + b\sigma^2_{t-1+(i-1)h}] \\
&= 8aIV_t + 4b\sum_{i=1}^{1/h}(\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds)[\sigma^2_{t-1+ih} + \sigma^2_{t-1+(i-1)h}]
\end{aligned} \tag{A.8}$$

$$2Cov_{\sigma}\left[RV_t^*(h), \sum_{i=1}^{1/h} e^2_{t-1+ih}\right] = 0 \tag{A.9}$$

$$2Cov_{\sigma}\left[RV_t^*(h), 2\sum_{i=1}^{1/h} r^*_{t-1+ih}e_{t-1+ih}\right] = 0 \tag{A.10}$$

$$2Cov_{\sigma}\left[\sum_{i=1}^{1/h} e^2_{t-1+ih}, 2\sum_{i=1}^{1/h} r^*_{t-1+ih}e_{t-1+ih}\right] = 0 \tag{A.11}$$

To summarize, we have

$$
\begin{aligned}
&MSE(h) = Var_{\sigma}[RV_t(h)] + (E_{\sigma}[RV_t(h)] - IV_t)^2 \\
&= 2hQ_t + o(h) + 4\sum_{i=1}^{1/h-1}\left[K_u V_u^2 - (a + b\sigma_{t-1+ih}^2)^2\right] + 4\sum_{i=1}^{1/h}[a + b\sigma_{t-1+ih}^2][a + b\sigma_{t-1+(i-1)h}^2] \\
&+ \left[K_u V_u^2 - (a + b\sigma_{t-1}^2)^2\right] + \left[K_u V_u^2 - (a + b\sigma_t^2)^2\right] \\
&+ 8aIV_t + 4b\sum_{i=1}^{1/h}(\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds)[\sigma_{t-1+ih}^2 + \sigma_{t-1+(i-1)h}^2] \\
&+ (2a/h + b\sum_{i=1}^{1/h}(\sigma_{t-1+ih}^2 + \sigma_{t-1+(i-1)h}^2))^2 \\
&= 2hQ_t + (2a/h + b\sum_{i=1}^{1/h}(\sigma_{t-1+ih}^2 + \sigma_{t-1+(i-1)h}^2))^2 + o(1/h) + f(t) \\
&\approx 2hQ_t + \frac{4}{h^2}(a + bIV_t)^2
\end{aligned}
\tag{A.12}
$$

Proof of Proposition 3 :

$$(a)E[r_{t-1+ih}r_{t-1+jh}] = E[(r_{t-1+ih}^* + e_{t-1+ih})(r_{t-1+jh}^* + e_{t-1+jh})] \tag{A.13}$$

If $i = j$,

$$E[r_{t-1+ih}r_{t-1+jh}] = E[r_{t-1+ih}^{*2} + e_{t-1+ih}^2] = a_0 h + 2V_u \tag{A.14}$$

If $|i - j| = 1$,

$$E[r_{t-1+ih}r_{t-1+jh}] = -E[u_{t-1+(i-1)h}^2] = -V_u \tag{A.15}$$

Else,

$$E[r_{t-1+ih}r_{t-1+jh}] = 0. \tag{A.16}$$

(b)

$$\begin{aligned}
&Cov[IV_{t+1}, r_{t-1+ih}r_{t-1+jh}] \\
&= Cov[IV_{t+1}, (r^*_{t-1+ih} + e_{t-1+ih})(r^*_{t-1+jh} + e_{t-1+jh})] \\
&= \delta_{i,j} Cov[IV_{t+1}, r^{*2}_{t-1+ih}] + Cov[IV_{t+1}, e_{t-1+ih}e_{t-1+jh}] \\
&= \delta_{i,j} Cov[IV_{t+1}, r^{*2}_{t-1+ih}] - \delta_{i,j-1} Cov[IV_{t+1}, u^2_{t-1+ih}] - \delta_{i-1,j} Cov[IV_{t+1}, u^2_{t-1+(i-1)h}] \\
&+ \delta_{i,j} Cov[IV_{t+1}, u^2_{t-1+ih}] + \delta_{i,j} Cov[IV_{t+1}, u^2_{t-1+(i-1)h}]
\end{aligned} \tag{A.17}$$

Using (2.21) in ABM(2011),

$$Cov[IV_{t+1}, r^{*2}_{t-1+ih}] = \sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2}(1 - \exp(-\lambda_n h))(1 - \exp(-\lambda_n))\exp(-\lambda_n(1 - ih)) \tag{A.18}$$

We have,

$$
\begin{aligned}
Cov[IV_{t+1}, u^2_{t-1+ih}] &= E[E_\sigma[IV_{t+1}u^2_{t-1+ih}]] - \underbrace{E[IV_{t+1}]}_{=a_0}\underbrace{E[u^2_{t-1+ih}]}_{=V_u=a+ba_0} \\
&= aa_0 + bE[\sigma^2_{t-1+ih}\int_t^{t+1}\sigma_s^2 ds] - a_0V_u \\
&= aa_0 + bE[(a_0 + \sum_{n=1}^{p} a_nP_n(f_{t-1+ih}))\int_t^{t+1}(a_0 + \sum_{m=1}^{p} a_mP_m(f_s))ds] \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+1} E[P_n(f_{t-1+ih})P_m(f_s)]ds \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+1} E[E[P_n(f_{t-1+ih})P_m(f_s)|f_\tau, \tau \leq t-1+ih]]ds \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+1} E[P_n(f_{t-1+ih})\ \underbrace{E[P_m(f_s)|f_\tau, \tau \leq t-1+ih]}_{=\exp(-\lambda_m(s-(t-1+ih)))P_m(f_{t-1+ih})}\ ]ds \\
&= b\sum_{n=1}^{p} a_n^2\int_t^{t+1}\exp(-\lambda_n(s-(t-1+ih)))ds \\
&= b\sum_{n=1}^{p} a_n^2\frac{\exp(-\lambda_n(1-ih)) - \exp(-\lambda_n(2-ih))}{\lambda_n}
\end{aligned}
\tag{A.19}
$$

The same for,

$$
Cov[IV_{t+1}, u^2_{t-1+(i-1)h}] = b\sum_{n=1}^{p} a_n^2\frac{\exp(-\lambda_n(1-(i-1)h)) - \exp(-\lambda_n(2-(i-1)h))}{\lambda_n}
\tag{A.20}
$$

To recapitulate,

$$\begin{aligned}
&Cov[IV_{t+1}, r_{t-1+ih} r_{t-1+jh}] \\
&= \delta_{i,j}(\sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2}(1-\exp(-\lambda_n h))(1-\exp(-\lambda_n))\exp(-\lambda_n(1-ih)) \\
&+ b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-ih)) - \exp(-\lambda_n(2-ih))}{\lambda_n} \\
&+ b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-(i-1)h)) - \exp(-\lambda_n(2-(i-1)h))}{\lambda_n}) \\
&- \delta_{i,j-1} b \sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-ih)) - \exp(-\lambda_n(2-ih))}{\lambda_n} \\
&- \delta_{i-1,j} b \sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-(i-1)h)) - \exp(-\lambda_n(2-(i-1)h))}{\lambda_n}
\end{aligned} \tag{A.21}$$

(c)

$$\begin{aligned}
&E[r_{t-1+ih} r_{t-1+jh} r_{t-1+kh} r_{t-1+lh}] \\
&= E[(r^*_{t-1+ih} + e_{t-1+ih})(r^*_{t-1+jh} + e_{t-1+jh})(r^*_{t-1+kh} + e_{t-1+kh})(r^*_{t-1+lh} + e_{t-1+lh})]
\end{aligned} \tag{A.22}$$

If $i = j = k = l$,

$$\begin{aligned}
&E[r_{t-1+ih} r_{t-1+jh} r_{t-1+kh} r_{t-1+lh}] = E[r^{*4}_{t-1+ih}] + E[e^4_{t-1+ih}] + 6E[r^{*2}_{t-1+ih} e^2_{t-1+ih}] \\
&= E[r^{*4}_{t-1+ih}] + E[u^4_{t-1+ih}] + E[u^4_{t-1+(i-1)h}] + 6E[u^2_{t-1+ih} u^2_{t-1+(i-1)h}] + 6E[r^{*2}_{t-1+ih} e^2_{t-1+ih}]
\end{aligned} \tag{A.23}$$

Equation (2.17) in ABM(2011) gives,

$$E[r^{*4}_{t-1+ih}] = 3a_0^2h^2 + 6\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[-1+\lambda_n h+\exp(-\lambda_n h)] \tag{A.24}$$

We have,

$$E[u^4_{t-1+ih}] = E[u^4_{t-1+(i-1)h}] = K_u V_u^2. \tag{A.25}$$

$$
\begin{aligned}
&E[u_{t-1+ih}^2 u_{t-1+(i-1)h}^2] = E[E_\sigma[u_{t-1+ih}^2 u_{t-1+(i-1)h}^2]] \\
&= E[E_\sigma[u_{t-1+ih}^2] E_\sigma[u_{t-1+(i-1)h}^2]] \\
&= E[(a + b\sigma_{t-1+ih}^2)(a + b\sigma_{t-1+(i-1)h}^2)] \\
&= a^2 + b^2 E[\sigma_{t-1+ih}^2 \sigma_{t-1+(i-1)h}^2] + abE[\sigma_{t-1+ih}^2] + abE[\sigma_{t-1+(i-1)h}^2] \\
&= a^2 + b^2 E[\sigma_{t-1+ih}^2 \sigma_{t-1+(i-1)h}^2] + 2aba_0 \\
&= a^2 + b^2 E[(a_0 + \sum_{n=1}^{p} a_n P_n(f_{t-1+ih}))(a_0 + \sum_{m=1}^{p} a_m P_m(f_{t-1+(i-1)h}))] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[P_n(f_{t-1+ih}) P_m(f_{t-1+(i-1)h})] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[E[P_n(f_{t-1+ih}) P_m(f_{t-1+(i-1)h}) | f_{\tau,\tau \le t-1+(i-1)h}]] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[P_m(f_{t-1+(i-1)h}) \underbrace{E[P_n(f_{t-1+ih}) | f_{\tau,\tau \le t-1+(i-1)h}]}_{= \exp(-\lambda_n h) P_n(f_{t-1+(i-1)h})}] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m \exp(-\lambda_n h) E[P_m(f_{t-1+(i-1)h}) P_n(f_{t-1+(i-1)h})] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h) + 2aba_0,
\end{aligned}
\tag{A.26}
$$

$$
\begin{aligned}
&E[r_{t-1+ih}^{*2} e_{t-1+ih}^2] = E[r_{t-1+ih}^{*2}(u_{t-1+ih}^2 + u_{t-1+(i-1)h}^2 - 2u_{t-1+ih} u_{t-1+(i-1)h})] \\
&= E[r_{t-1+ih}^{*2} u_{t-1+ih}^2] + E[r_{t-1+ih}^{*2} u_{t-1+(i-1)h}^2]
\end{aligned}
\tag{A.27}
$$

$$
\begin{aligned}
&E[r^{*2}_{t-1+ih}u^2_{t-1+ih}] = E[E_\sigma[r^{*2}_{t-1+ih}]E_\sigma[u^2_{t-1+ih}]] = E[(a+b\sigma^2_{t-1+ih})\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds] \\
&= aa_0h + bE[\sigma^2_{t-1+ih}\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds] \\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+ih}))\int_{t-1+(i-1)h}^{t-1+ih}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds] \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+ih})P_m(f_s)]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[E[P_n(f_{t-1+ih})P_m(f_s)|f_\tau,\tau\le s]]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}\underbrace{E[P_m(f_s)E[P_n(f_{t-1+ih})|f_\tau,\tau\le s]}_{\exp(-\lambda_n(t-1+ih-s))P_m(f_s)}]ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(i-1)h}^{t-1+ih}\exp(-\lambda_n(t-1+ih-s))ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_nh)}{\lambda_n}
\end{aligned}
\tag{A.28}
$$

$$
\begin{aligned}
&E[r^{*2}_{t-1+ih}u^2_{t-1+(i-1)h}] = E[E_\sigma[r^{*2}_{t-1+ih}]E_\sigma[u^2_{t-1+(i-1)h}]]\\
&= E[(a+b\sigma^2_{t-1+(i-1)h})\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds]\\
&= aa_0h + bE[\sigma^2_{t-1+(i-1)h}\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds]\\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p} a_nP_n(f_{t-1+(i-1)h}))\int_{t-1+(i-1)h}^{t-1+ih}(a_0+\sum_{m=1}^{p} a_mP_m(f_s))ds]\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p} a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+(i-1)h})P_m(f_s)]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p} a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[E[P_n(f_{t-1+(i-1)h})P_m(f_s)|f_\tau,\tau\le t-1+(i-1)h]]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p} a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+(i-1)h})\ \underbrace{E[P_m(f_s)|f_\tau,\tau\le t-1+(i-1)h]}_{=\exp(-\lambda_m(s-(t-1+(i-1)h)))P_m(f_{t-1+(i-1)h})}\ ]ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p} a_n^2\int_{t-1+(i-1)h}^{t-1+ih}\exp(-\lambda_n(s-(t-1+(i-1)h)))ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p} a_n^2\frac{1-\exp(-\lambda_nh)}{\lambda_n}
\end{aligned}
\tag{A.29}
$$

Consequently if $i = j = k = l$,

$$\begin{aligned}
&E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = 3a_0^2h^2 + 6\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[-1+\lambda_n h+\exp(-\lambda_n h)] + 2K_uV_u^2 \\
&+6(V_u^2+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h)) + 12(a_0hV_u + b\sum_{n=0}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n}) \\
&= 3a_0^2h^2 + 2(K_u+3)V_u^2 + 12a_0hV_u + 6\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[-1+\lambda_n h+\exp(-\lambda_n h)] \\
&+6b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h) + 12b\sum_{n=0}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n}
\end{aligned} \tag{A.30}$$

If $i = j = k = l+1$ or $i = j+1 = k+1 = l+1$,

$$\begin{aligned}
&E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r_{t-1+ih}^3 r_{t-1+(i-1)h}] \\
&= E[(r_{t-1+ih}^{*3} + 3r_{t-1+ih}^{*2}e_{t-1+ih} + 3r_{t-1+ih}^{*}e_{t-1+ih}^2 + e_{t-1+ih}^3)(r_{t-1+(i-1)h}^{*} + e_{t-1+(i-1)h})] \\
&= E[(3r_{t-1+ih}^{*2}e_{t-1+ih} + e_{t-1+ih}^3)e_{t-1+(i-1)h}] \\
&= -3E[r_{t-1+ih}^{*2}u_{t-1+(i-1)h}^2] - 3E[u_{t-1+ih}^2u_{t-1+(i-1)h}^2] - E[u_{t-1+(i-1)h}^4] \\
&= -(K_u+3)V_u^2 - 3a_0hV_u - 3b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n}) - 3b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h)
\end{aligned} \tag{A.31}$$

If $i = j = k+1 = l+1$,

$$E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r^2_{t-1+ih}r^2_{t-1+(i-1)h}]$$
$$= E[(r^*_{t-1+ih} + e_{t-1+ih})^2(r^*_{t-1+(i-1)h} + e_{t-1+(i-1)h})^2]$$
$$= E[r^{*2}_{t-1+ih}r^{*2}_{t-1+(i-1)h}] + E[r^{*2}_{t-1+ih}e^2_{t-1+(i-1)h}] + E[r^{*2}_{t-1+(i-1)h}e^2_{t-1+ih}] + E[e^2_{t-1+ih}e^2_{t-1+(i-1)h}]$$
$$= E[r^{*2}_{t-1+ih}r^{*2}_{t-1+(i-1)h}] + E[r^{*2}_{t-1+ih}(u^2_{t-1+(i-1)h} + u^2_{t-1+(i-2)h})]$$
$$+ E[r^{*2}_{t-1+(i-1)h}(u^2_{t-1+ih} + u^2_{t-1+(i-1)h})] + E[(u^2_{t-1+ih} + u^2_{t-1+(i-1)h})(u^2_{t-1+(i-1)h} + u^2_{t-1+(i-2)h})]$$
$$= E[r^{*2}_{t-1+ih}r^{*2}_{t-1+(i-1)h}] + E[r^{*2}_{t-1+ih}u^2_{t-1+(i-1)h}] + E[r^{*2}_{t-1+ih}u^2_{t-1+(i-2)h}]$$
$$+ E[r^{*2}_{t-1+(i-1)h}u^2_{t-1+ih}] + E[r^{*2}_{t-1+(i-1)h}u^2_{t-1+(i-1)h}]$$
$$+ E[u^2_{t-1+ih}u^2_{t-1+(i-1)h}] + E[u^2_{t-1+ih}u^2_{t-1+(i-2)h}] + E[u^4_{t-1+(i-1)h}] + E[u^2_{t-1+(i-1)h}u^2_{t-1+(i-2)h}] \tag{A.32}$$

Using (2.18) in ABM(2011),

$$E[r^{*2}_{t-1+ih}r^{*2}_{t-1+(i-1)h}] = a_0^2h^2 + \sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[1 - \exp(-\lambda_n h)]^2 \tag{A.33}$$

From (A.14), we have

$$E[r^{*2}_{t-1+ih}u^2_{t-1+(i-1)h}] = aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{1 - \exp(-\lambda_n h)}{\lambda_n} \tag{A.34}$$

The third term in (A.17) is written as

$$
\begin{aligned}
&E[r^{*2}_{t-1+ih}u^2_{t-1+(i-2)h}] = E[E_\sigma[r^{*2}_{t-1+ih}]E_\sigma[u^2_{t-1+(i-2)h}]] \\
&= E[(a+b\sigma^2_{t-1+(i-2)h})\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds] \\
&= aa_0h + bE[\sigma^2_{t-1+(i-2)h}\int_{t-1+(i-1)h}^{t-1+ih}\sigma_s^2 ds] \\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+(i-2)h}))\int_{t-1+(i-1)h}^{t-1+ih}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds] \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+(i-2)h})P_m(f_s)]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[E[P_n(f_{t-1+(i-2)h})P_m(f_s)|f_\tau,\tau\leq t-1+(i-2)h]]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+(i-2)h})\ \underbrace{E[P_m(f_s)|f_\tau,\tau\leq t-1+(i-2)h]}_{=\exp(-\lambda_m(s-(t-1+(i-2)h)))P_m(f_{t-1+(i-2)h})}\ ]ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(i-1)h}^{t-1+ih}\exp(-\lambda_n(s-(t-1+(i-2)h)))ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{\exp(-2\lambda_nh)-\exp(-\lambda_nh)}{-\lambda_n}
\end{aligned}
$$

(A.35)

$$
\begin{aligned}
&E[r^{*2}_{t-1+(i-1)h}u^2_{t-1+ih}] = E[E_\sigma[r^{*2}_{t-1+(i-1)h}]E_\sigma[u^2_{t-1+ih}]] \\
&= E[(a+b\sigma^2_{t-1+ih})\int_{t-1+(i-2)h}^{t-1+(i-1)h}\sigma^2_s ds] \\
&= aa_0h + bE[\sigma^2_{t-1+ih}\int_{t-1+(i-2)h}^{t-1+(i-1)h}\sigma^2_s ds] \\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+ih}))\int_{t-1+(i-2)h}^{t-1+(i-1)h}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds] \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}E[P_n(f_{t-1+ih})P_m(f_s)]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}E[E[P_n(f_{t-1+ih})P_m(f_s)|f_\tau, \tau \leq t-1+(i-1)h]]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}E[P_m(f_s)\underbrace{E[P_n(f_{t-1+ih})|f_\tau, \tau \leq s]}_{=\exp(-\lambda_m((t-1+ih)-s))P_n(f_s)}]ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(i-2)h}^{t-1+(i-1)h}\exp(-\lambda_n((t-1+ih)-s))ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_nh)-\exp(-2\lambda_nh)}{\lambda_n}
\end{aligned}
\tag{A.36}
$$

$$\begin{aligned}
&E[r^{*2}_{t-1+(i-1)h}u^{2}_{t-1+(i-1)h}] = E[E_{\sigma}[r^{*2}_{t-1+(i-1)h}]E_{\sigma}[u^{2}_{t-1+(i-1)h}]]\\
&= E[(a+b\sigma^{2}_{t-1+(i-1)h})\int_{t-1+(i-2)h}^{t-1+(i-1)h}\sigma_s^2 ds]\\
&= aa_0h + bE[\sigma^{2}_{t-1+(i-1)h}\int_{t-1+(i-2)h}^{t-1+(i-1)h}\sigma_s^2 ds]\\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+(i-1)h}))\int_{t-1+(i-2)h}^{t-1+(i-1)h}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds]\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}E[P_n(f_{t-1+(i-1)h})P_m(f_s)]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}E[E[P_n(f_{t-1+(i-1)h})P_m(f_s)|f_\tau,\tau\le s]]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-2)h}^{t-1+(i-1)h}\underbrace{E[P_m(f_s)E[P_n(f_{t-1+(i-1)h})|f_\tau,\tau\le s]}_{\exp(-\lambda_n(t-1+(i-1)h-s))P_m(f_s)}]ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(i-2)h}^{t-1+(i-1)h}\exp(-\lambda_n(t-1+(i-1)h-s))ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_nh)}{\lambda_n}
\end{aligned}\tag{A.37}$$

The equation (A.11) gives

$$E[u^{2}_{t-1+ih}u^{2}_{t-1+(i-1)h}] = a^2 + b^2a_0^2 + b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_nh) + 2aba_0 \tag{A.38}$$

$$
\begin{aligned}
&E[u_{t-1+ih}^2 u_{t-1+(i-2)h}^2] = E[E_\sigma[u_{t-1+ih}^2 u_{t-1+(i-2)h}^2]] \\
&= E[E_\sigma[u_{t-1+ih}^2]E_\sigma[u_{t-1+(i-2)h}^2]] \\
&= E[(a+b\sigma_{t-1+ih}^2)(a+b\sigma_{t-1+(i-2)h}^2)] \\
&= a^2 + b^2 E[\sigma_{t-1+ih}^2 \sigma_{t-1+(i-2)h}^2] + abE[\sigma_{t-1+ih}^2] + abE[\sigma_{t-1+(i-2)h}^2] \\
&= a^2 + b^2 E[\sigma_{t-1+ih}^2 \sigma_{t-1+(i-2)h}^2] + 2aba_0 \\
&= a^2 + b^2 E[(a_0 + \sum_{n=1}^{p} a_n P_n(f_{t-1+ih}))(a_0 + \sum_{m=1}^{p} a_m P_m(f_{t-1+(i-2)h}))] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[P_n(f_{t-1+ih}) P_m(f_{t-1+(i-2)h})] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[E[P_n(f_{t-1+ih}) P_m(f_{t-1+(i-2)h}) | f_{\tau,\tau \le t-1+(i-2)h}]] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m E[P_m(f_{t-1+(i-2)h}) \underbrace{E[P_n(f_{t-1+ih}) | f_{\tau,\tau \le t-1+(i-2)h}]}_{=\exp(-2\lambda_n h) P_n(f_{t-1+(i-2)h})}] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n,m=1}^{p} a_n a_m \exp(-2\lambda_n h) E[P_m(f_{t-1+(i-2)h}) P_n(f_{t-1+(i-2)h})] + 2aba_0 \\
&= a^2 + b^2 a_0^2 + b^2 \sum_{n=1}^{p} a_n^2 \exp(-2\lambda_n h) + 2aba_0,
\end{aligned}
\tag{A.39}
$$

$$
\begin{aligned}
&E[u^2_{t-1+(i-1)h}u^2_{t-1+(i-2)h}] = E[E_\sigma[u^2_{t-1+(i-1)h}u^2_{t-1+(i-2)h}]] \\
&= E[E_\sigma[u^2_{t-1+(i-1)h}]E_\sigma[u^2_{t-1+(i-2)h}]] \\
&= E[(a+b\sigma^2_{t-1+(i-1)h})(a+b\sigma^2_{t-1+(i-2)h})] \\
&= a^2+b^2E[\sigma^2_{t-1+(i-1)h}\sigma^2_{t-1+(i-2)h}]+abE[\sigma^2_{t-1+(i-1)h}]+abE[\sigma^2_{t-1+(i-2)h}] \\
&= a^2+b^2E[\sigma^2_{t-1+(i-1)h}\sigma^2_{t-1+(i-2)h}]+2aba_0 \\
&= a^2+b^2E[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+(i-1)h}))(a_0+\sum_{m=1}^{p}a_mP_m(f_{t-1+(i-2)h}))]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[P_n(f_{t-1+(i-1)h})P_m(f_{t-1+(i-2)h})]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[E[P_n(f_{t-1+(i-1)h})P_m(f_{t-1+(i-2)h})|f_{\tau,\tau\leq t-1+(i-2)h}]]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[P_m(f_{t-1+(i-2)h})\underbrace{E[P_n(f_{t-1+(i-1)h})|f_{\tau,\tau\leq t-1+(i-2)h}]}_{=\exp(-\lambda_nh)P_n(f_{t-1+(i-2)h})}]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_m\exp(-\lambda_nh)E[P_m(f_{t-1+(i-2)h})P_n(f_{t-1+(i-2)h})]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_nh)+2aba_0,
\end{aligned}
\tag{A.40}
$$

To summarize, if $i = j = k+1 = l+1$,

$$E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = a_0^2h^2 + (K_u+3)V_u^2 + 4a_0hV_u + \sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[1-\exp(-\lambda_n h)]^2$$
$$+2b\sum_{n=1}^{p}a_n^2\frac{1-\exp(-\lambda_n h)}{\lambda_n} + 2b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h)-\exp(-2\lambda_n h)}{\lambda_n}$$
$$+2b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h)) + b^2\sum_{n=1}^{p}a_n^2\exp(-2\lambda_n h) \tag{A.41}$$

If $i = j > k+1, k = l$,

$$E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r_{t-1+ih}^2 r_{t-1+kh}^2]$$
$$= E[(r_{t-1+ih}^* + e_{t-1+ih})^2(r_{t-1+kh}^* + e_{t-1+kh})^2]$$
$$= E[r_{t-1+ih}^{*2}r_{t-1+kh}^{*2}] + E[r_{t-1+ih}^{*2}e_{t-1+kh}^2] + E[r_{t-1+kh}^{*2}e_{t-1+ih}^2] + E[e_{t-1+ih}^2e_{t-1+kh}^2]$$
$$= E[r_{t-1+ih}^{*2}r_{t-1+kh}^{*2}] + E[r_{t-1+ih}^{*2}(u_{t-1+kh}^2 + u_{t-1+(k-1)h}^2)]$$
$$+E[r_{t-1+kh}^{*2}(u_{t-1+ih}^2 + u_{t-1+(i-1)h}^2)] + E[(u_{t-1+ih}^2 + u_{t-1+(i-1)h}^2)(u_{t-1+kh}^2 + u_{t-1+(k-1)h}^2)]$$
$$= E[r_{t-1+ih}^{*2}r_{t-1+kh}^{*2}] + E[r_{t-1+ih}^{*2}u_{t-1+kh}^2] + E[r_{t-1+ih}^{*2}u_{t-1+(k-1)h}^2]$$
$$+E[r_{t-1+kh}^{*2}u_{t-1+ih}^2] + E[r_{t-1+kh}^{*2}u_{t-1+(i-1)h}^2]$$
$$+E[u_{t-1+ih}^2u_{t-1+kh}^2] + E[u_{t-1+ih}^2u_{t-1+(k-1)h}^2] + E[u_{t-1+(i-1)h}^2u_{t-1+kh}^2] + E[u_{t-1+(i-1)h}^2u_{t-1+(k-1)h}^2] \tag{A.42}$$

From (2.18) in ABM (2011),

$$E[r_{t-1+ih}^{*2}r_{t-1+kh}^{*2}] = a_0^2h^2 + \sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}[1-\exp(-\lambda_n h)]^2\exp(-\lambda_n(i-k-1)h) \tag{A.43}$$

$$
\begin{aligned}
&E[r^{*2}_{t-1+ih}u^2_{t-1+kh}] = E[E_\sigma[r^{*2}_{t-1+ih}]E_\sigma[u^2_{t-1+kh}]]\\
&= E[(a+b\sigma^2_{t-1+kh})\int_{t-1+(i-1)h}^{t-1+ih}\sigma^2_s ds]\\
&= aa_0h + bE[\sigma^2_{t-1+kh}\int_{t-1+(i-1)h}^{t-1+ih}\sigma^2_s ds]\\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+kh}))\int_{t-1+(i-1)h}^{t-1+ih}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds]\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+kh})P_m(f_s)]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[E[P_n(f_{t-1+kh})P_m(f_s)|f_\tau,\tau\le t-1+kh]]ds\\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(i-1)h}^{t-1+ih}E[P_n(f_{t-1+kh})\underbrace{E[P_m(f_s)|f_\tau,\tau\le t-1+kh]}_{=\exp(-\lambda_m(s-(t-1+kh)))P_m(f_{t-1+kh})}]ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(i-1)h}^{t-1+ih}\exp(-\lambda_n(s-(t-1+kh)))ds\\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_nh(i-k))-\exp(-\lambda_nh(i-k-1))}{-\lambda_n}
\end{aligned}
\tag{A.44}
$$

Using the previous equation,

$$
E[r^{*2}_{t-1+ih}u^2_{t-1+(k-1)h}] = aa_0h + ba_0^2h + b\sum_{n=0}^{p}a_n^2\frac{\exp(-\lambda_nh(i-k+1))-\exp(-\lambda_nh(i-k))}{-\lambda_n}
\tag{A.45}
$$

$$
\begin{aligned}
&E[r^{*2}_{t-1+kh}u^2_{t-1+ih}] = E[E_\sigma[r^{*2}_{t-1+kh}]E_\sigma[u^2_{t-1+ih}]] \\
&= E[(a+b\sigma^2_{t-1+ih})\int_{t-1+(k-1)h}^{t-1+kh}\sigma_s^2 ds] \\
&= aa_0h + bE[\sigma^2_{t-1+ih}\int_{t-1+(k-1)h}^{t-1+kh}\sigma_s^2 ds] \\
&= aa_0h + bE[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+ih}))\int_{t-1+(k-1)h}^{t-1+kh}(a_0+\sum_{m=1}^{p}a_mP_m(f_s))ds] \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(k-1)h}^{t-1+kh}E[P_n(f_{t-1+ih})P_m(f_s)]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(k-1)h}^{t-1+kh}E[E[P_n(f_{t-1+ih})P_m(f_s)|f_\tau,\tau\le t-1+(i-1)h]]ds \\
&= aa_0h + ba_0^2h + b\sum_{n,m=1}^{p}a_na_m\int_{t-1+(k-1)h}^{t-1+kh}E[P_m(f_s)\underbrace{E[P_n(f_{t-1+ih})|f_\tau,\tau\le s]}_{=\exp(-\lambda_m((t-1+ih)-s))P_n(f_s)}]ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\int_{t-1+(k-1)h}^{t-1+kh}\exp(-\lambda_n((t-1+ih)-s))ds \\
&= aa_0h + ba_0^2h + b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_nh(i-k))-\exp(-\lambda_nh(i-k+1))}{\lambda_n}
\end{aligned}
\tag{A.46}
$$

From the previous equation we have,

$$
E[r^{*2}_{t-1+kh}u^2_{t-1+(i-1)h}] = aa_0h + ba_0^2h + b\sum_{n=0}^{p}a_n^2\frac{\exp(-\lambda_nh(i-1-k))-\exp(-\lambda_nh(i-k))}{\lambda_n}
\tag{A.47}
$$

$$\begin{aligned}
&E[u^2_{t-1+ih}u^2_{t-1+kh}] = E[E_\sigma[u^2_{t-1+ih}u^2_{t-1+kh}]] \\
&= E[E_\sigma[u^2_{t-1+ih}]E_\sigma[u^2_{t-1+kh}]] \\
&= E[(a+b\sigma^2_{t-1+ih})(a+b\sigma^2_{t-1+kh})] \\
&= a^2+b^2E[\sigma^2_{t-1+ih}\sigma^2_{t-1+kh}]+abE[\sigma^2_{t-1+ih}]+abE[\sigma^2_{t-1+kh}] \\
&= a^2+b^2E[\sigma^2_{t-1+ih}\sigma^2_{t-1+kh}]+2aba_0 \\
&= a^2+b^2E[(a_0+\sum_{n=1}^{p}a_nP_n(f_{t-1+ih}))(a_0+\sum_{m=1}^{p}a_mP_m(f_{t-1+kh}))]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[P_n(f_{t-1+ih})P_m(f_{t-1+kh})]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[E[P_n(f_{t-1+ih})P_m(f_{t-1+kh})|f_{\tau,\tau\le t-1+kh}]]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_mE[P_m(f_{t-1+kh})\underbrace{E[P_n(f_{t-1+ih})|f_{\tau,\tau\le t-1+kh}]}_{=\exp(-\lambda_nh(i-k))P_n(f_{t-1+kh})}]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n,m=1}^{p}a_na_m\exp(-\lambda_nh(i-k))E[P_m(f_{t-1+kh})P_n(f_{t-1+kh})]+2aba_0 \\
&= a^2+b^2a_0^2+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_nh(i-k))+2aba_0,
\end{aligned} \tag{A.48}$$

The same for,

$$E[u^2_{t-1+ih}u^2_{t-1+(k-1)h}] = a^2+b^2a_0^2+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_nh(i-k+1))+2aba_0 \tag{A.49}$$

$$E[u^2_{t-1+(i-1)h}u^2_{t-1+kh}] = a^2 + b^2 a_0^2 + b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k-1)) + 2aba_0 \quad \text{(A.50)}$$

$$E[u^2_{t-1+(i-1)h}u^2_{t-1+(k-1)h}] = a^2 + b^2 a_0^2 + b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k)) + 2aba_0 \quad \text{(A.51)}$$

To summarize, if $i = j > k+1, k = l$,

$$\begin{aligned}
&E[r_{t-1+ih} r_{t-1+jh} r_{t-1+kh} r_{t-1+lh}] \\
&= a_0^2 h^2 + 4a_0 h V_u + 4V_u^2 + \sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2} [1-\exp(-\lambda_n h)]^2 \exp(-\lambda_n (i-k-1)h) \\
&+ 2b \sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n h(i-k)) - \exp(-\lambda_n h(i-k-1))}{-\lambda_n} \\
&+ b \sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n h(i-k+1)) - \exp(-\lambda_n h(i-k))}{-\lambda_n} \\
&+ b \sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n h(i-k)) - \exp(-\lambda_n h(i-k+1))}{\lambda_n} \\
&+ 2b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k)) \\
&+ b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k+1)) \\
&+ b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k-1))
\end{aligned} \quad \text{(A.52)}$$

If $i = j+1, j = k = l+1$,

$$
\begin{aligned}
&E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r_{t-1+ih}r^2_{t-1+(i-1)h}r_{t-1+(i-2)h}] \\
&= E[(r^*_{t-1+ih}+e_{t-1+ih})(r^*_{t-1+(i-1)h}+e_{t-1+(i-1)h})^2(r^*_{t-1+(i-2)h}+e_{t-1+(i-2)h})] \\
&= E[(r^*_{t-1+ih}+e_{t-1+ih})(r^{*2}_{t-1+(i-1)h}+e^2_{t-1+(i-1)h})(r^*_{t-1+(i-2)h}+e_{t-1+(i-2)h})] \\
&= E[r^{*2}_{t-1+(i-1)h}e_{t-1+ih}e_{t-1+(i-2)h}]+E[e^2_{t-1+(i-1)h}e_{t-1+ih}e_{t-1+(i-2)h}] \\
&= 2E[u^2_{t-1+(i-1)h}u^2_{t-1+(i-2)h}] \\
&= 2(V_u^2+b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h))
\end{aligned}
\tag{A.53}
$$

If $i = j > k, k = l+1$ or $i = j+1, j > k, k = l$,

$$
\begin{aligned}
&E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r^2_{t-1+ih}r_{t-1+kh}r_{t-1+(k-1)h}] \\
&= E[(r^*_{t-1+ih}+e_{t-1+ih})^2(r^*_{t-1+kh}+e_{t-1+kh})(r^*_{t-1+(k-1)h}+e_{t-1+(k-1)h})] \\
&= E[e^2_{t-1+ih}e_{t-1+kh}e_{t-1+(k-1)h}]+E[r^{*2}_{t-1+ih}e_{t-1+kh}e_{t-1+(k-1)h}] \\
&= -E[u^2_{t-1+(k-1)h}(u^2_{t-1+ih}+u^2_{t-1+(i-1)h})]-E[r^{*2}_{t-1+ih}u^2_{t-1+(k-1)h}] \\
&= -a_0hV_u-2V_u^2-b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k+1)) \\
&-b^2\sum_{n=1}^{p}a_n^2\exp(-\lambda_n h(i-k)) \\
&-b\sum_{n=1}^{p}a_n^2\frac{\exp(-\lambda_n h(i-k+1))-\exp(-\lambda_n h(i-k))}{-\lambda_n}
\end{aligned}
\tag{A.54}
$$

If $i = j+1, j > k, k = l+1$,

$$\begin{aligned}
&E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = E[r_{t-1+ih}r_{t-1+(i-1)h}r_{t-1+kh}r_{t-1+(k-1)h}] \\
&= E[e_{t-1+ih}e_{t-1+(i-1)h}e_{t-1+kh}e_{t-1+(k-1)h}] \qquad \text{(A.55)} \\
&= E[u^2_{t-1+(i-1)h}u^2_{t-1+(k-1)h}] = V_u^2 + b^2 \sum_{n=1}^{p} a_n^2 \exp(-\lambda_n h(i-k))
\end{aligned}$$

Else, $E[r_{t-1+ih}r_{t-1+jh}r_{t-1+kh}r_{t-1+lh}] = 0$.

(d) The variance expression is the same as in Andersen et al. (2011).

Proof of Proposition 4 :

$$\begin{aligned}
&(a)Cov[IV_{t+1:t+m}, r_{t-1+ih}r_{t-1+jh}] \\
&= Cov[IV_{t+1:t+m}, (r^*_{t-1+ih} + e_{t-1+ih})(r^*_{t-1+jh} + e_{t-1+jh})] \\
&= \delta_{i,j}Cov[IV_{t+1:t+m}, r^{*2}_{t-1+ih}] + Cov[IV_{t+1:t+m}, e_{t-1+ih}e_{t-1+jh}] \\
&= \delta_{i,j}Cov[IV_{t+1:t+m}, r^{*2}_{t-1+ih}] - \delta_{i,j-1}Cov[IV_{t+1:t+m}, u^2_{t-1+ih}] - \delta_{i-1,j}Cov[IV_{t+1:t+m}, u^2_{t-1+(i-1)h}] \\
&+ \delta_{i,j}Cov[IV_{t+1:t+m}, u^2_{t-1+ih}] + \delta_{i,j}Cov[IV_{t+1:t+m}, u^2_{t-1+(i-1)h}]
\end{aligned} \qquad \text{(A.56)}$$

Using (2.21) in ABM(2011),

$$Cov[IV_{t+1:t+m}, r^{*2}_{t-1+ih}] = \sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2}(1-\exp(-\lambda_n h))(1-\exp(-\lambda_n m))\exp(-\lambda_n(1-ih)) \qquad \text{(A.57)}$$

We have,

$$
\begin{aligned}
Cov[IV_{t+1:t+m}, u^2_{t-1+ih}] &= E[E_\sigma[IV_{t+1:t+m}u^2_{t-1+ih}]] - \underbrace{E[IV_{t+1:t+m}]}_{=ma_0}\underbrace{E[u^2_{t-1+ih}]}_{=V_u=a+ba_0} \\
&= aa_0 + bE[\sigma^2_{t-1+ih}\int_t^{t+m}\sigma^2_s ds] - a_0V_u \\
&= aa_0 + bE[(a_0 + \sum_{n=1}^{p} a_nP_n(f_{t-1+ih}))\int_t^{t+m}(a_0 + \sum_{m=1}^{p} a_mP_m(f_s))ds] - a_0V_u \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+m}E[P_n(f_{t-1+ih})P_m(f_s)]ds \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+m}E[E[P_n(f_{t-1+ih})P_m(f_s)|f_\tau, \tau \leq t-1+ih]]ds \\
&= b\sum_{n,m=1}^{p} a_na_m\int_t^{t+m}E[P_n(f_{t-1+ih})\ \underbrace{E[P_m(f_s)|f_\tau, \tau \leq t-1+ih]}_{=\exp(-\lambda_m(s-(t-1+ih)))P_m(f_{t-1+ih})}\ ]ds \\
&= b\sum_{n=1}^{p} a_n^2\int_t^{t+m}\exp(-\lambda_n(s-(t-1+ih)))ds \\
&= b\sum_{n=1}^{p} a_n^2\frac{\exp(-\lambda_n(1-ih)) - \exp(-\lambda_n(m+1-ih))}{\lambda_n}
\end{aligned}
\tag{A.58}
$$

The same for,

$$
Cov[IV_{t+1:t+m}, u^2_{t-1+(i-1)h}] = b\sum_{n=1}^{p} a_n^2\frac{\exp(-\lambda_n(1-(i-1)h)) - \exp(-\lambda_n(m+1-(i-1)h))}{\lambda_n}
\tag{A.59}
$$

To recapitulate,

$$
\begin{aligned}
&Cov[IV_{t+1:t+m}, r_{t-1+ih}r_{t-1+jh}] \\
&= \delta_{i,j}(\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}(1-\exp(-\lambda_n h))(1-\exp(-\lambda_n m))\exp(-\lambda_n(1-ih)) \\
&+ b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-ih))-\exp(-\lambda_n(m+1-ih))}{\lambda_n} \\
&+ b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-(i-1)h))-\exp(-\lambda_n(m+1-(i-1)h))}{\lambda_n}) \\
&- \delta_{i,j-1} b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-ih))-\exp(-\lambda_n(m+1-ih))}{\lambda_n} \\
&- \delta_{i-1,j} b\sum_{n=1}^{p} a_n^2 \frac{\exp(-\lambda_n(1-(i-1)h))-\exp(-\lambda_n(m+1-(i-1)h))}{\lambda_n}
\end{aligned}
\tag{A.60}
$$

(b) The variance expression is the same as in Andersen et al. (2011).

Appendix B : Quadratic form representation for the realized measures

-The all RV Estimator,

$$
\begin{aligned}
q_{ij}^{all}(\underline{h}) &= 1 \;\text{ for }\; i = j \\
&= 0 \;\text{ otherwise.}
\end{aligned}
\tag{B.1}
$$

-The average RV Estimator,

$$
q_{ij}^{average}(h) = \frac{1}{n_h}\sum_{k=0}^{n_h-1} q_{ij}^{sparse}(h,k). \tag{B.2}
$$

where,

$$\begin{aligned} q_{ij}^{sparse}(h,k) = 1 \quad & \text{for } \; k+1 \leq i = j \leq N_k + k, \\ = 1 \quad & \text{for } i \neq j, \; (s-1)n_h + 1 + k \leq i, j \leq sn_h + k, \; s = 1, ..., N_k/n_h, \\ = 0 \quad & \text{otherwise.} \end{aligned} \tag{B.3}$$

-The two time scales estimator,

$$q_{ij}^{TS}(h) = q_{ij}^{average}(h) - \bar{n}\underline{h}q_{ij}^{all}(\underline{h}). \tag{B.4}$$

-The kernel-Based RV Estimator,

$$\begin{aligned} q_{ij}^{Kernel}(K(\cdot), L) &= 1 \text{ for } \; i = j \\ &= K\left(\frac{l-1}{L}\right) \text{ for } |i-j| = l, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{B.5}$$

In the specific calculations below, we use the modified Tukey-Hanning kernel advocated by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006),

$$K(x) \; = \; (1 \; - \; cos\pi(1-x)^2)/2. \tag{B.6}$$

-The pre-averaging estimator

The following is the proof of the quadratic form representation for the pre-averaging.

$$
\begin{aligned}
RV_t^{pre} &= \frac{12}{\theta\sqrt{N}} \sum_{i=0}^{N-k} \left(\sum_{j=1}^{k} \phi\left(\frac{j}{k}\right) r_{i+j} \right)^2 - \frac{6}{\theta^2 N} RV_t^{all} \\
&= \frac{12}{\theta\sqrt{N}} \sum_{i=0}^{N-k} \left(\sum_{l,m=1}^{k} \phi\left(\frac{l}{k}\right) \phi\left(\frac{m}{k}\right) r_{l+i} r_{m+i} \right) - \frac{6}{\theta^2 N} RV_t^{all} \\
&= \frac{12}{\theta\sqrt{N}} \sum_{i=0}^{N-k} \left(\sum_{I,J=1+i}^{k+i} \phi\left(\frac{I-i}{k}\right) \phi\left(\frac{J-i}{k}\right) r_I r_J \right) - \frac{6}{\theta^2 N} RV_t^{all} \\
&= \frac{12}{\theta\sqrt{N}} \sum_{i=0}^{N-k} \left(\sum_{I,J=1}^{N} \delta_{1\le I-i\le k} \delta_{1\le J-i\le k} \phi\left(\frac{I-i}{k}\right) \phi\left(\frac{J-i}{k}\right) r_I r_J \right) - \frac{6}{\theta^2 N} RV_t^{all} \\
&= \frac{12}{\theta\sqrt{N}} \sum_{I,J=1}^{N} \left(\underbrace{\sum_{i=0}^{N-k} \delta_{1\le I-i\le k} \delta_{1\le J-i\le k} \phi\left(\frac{I-i}{k}\right) \phi\left(\frac{J-i}{k}\right)}_{=q_{IJ}^{\phi}} \right) r_I r_J - \frac{6}{\theta^2 N} RV_t^{all} \\
&= \frac{12}{\theta\sqrt{N}} \sum_{I,J=1}^{N} q_{IJ}^{\phi} r_I r_J - \frac{6}{\theta^2 N} \sum_{I,J=1}^{N} q_{IJ}^{all} r_I r_J \\
&= \sum_{I,J=1}^{N} \underbrace{(\frac{12}{\theta\sqrt{N}} q_{IJ}^{\phi} - \frac{6}{\theta^2 N} q_{IJ}^{all})}_{=q_{IJ}^{pre}} r_I r_J \\
&= \sum_{I,J=1}^{N} q_{IJ}^{pre} r_I r_J
\end{aligned}
\tag{B.7}
$$

-The RV_t^{mse} estimator :

$$
RV_t^{mse} = \sum_{i=1}^{1/h} r_{t-1+ih^*}^2 = \sum_{i,j=1}^{N} q_{ij}^{mse} r_{t-1+ih} r_{t-1+jh}, \tag{B.8}
$$

where $q_{ij}^{mse} = q_{ij}^{sparse}(h^*)$.

Appendix C : The true volatility and realized measures correlations

We prove that the covariances between the integrated variance and the realized measures (needed to compute the correlations in Table 2.III) are given by,

$$Cov[IV_t, RM_t(h)] = \sum_{1 \leq i,j \leq 1/h} q_{ij} Cov[IV_t, r_{t-1+ih} r_{t-1+jh}], \tag{C.1}$$

where,

$$\begin{aligned}
Cov[IV_t, r_{t-1+ih} r_{t-1+jh}] &= \delta_{i,j}(2 \sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2}(\exp(-\lambda_n h) + \lambda_n h - 1) \\
&+ \sum_{n=1}^{p} \frac{a_n^2}{\lambda_n^2} (2 - \exp(-\lambda_n (i-1)h) - \exp(-\lambda_n (1-ih)))\,(1 - \exp(-\lambda_n h))) \\
&- \delta_{i,j-1} b \sum_{n=1}^{p} a_n^2 \frac{2 - \exp(-\lambda_n ih) - \exp(-\lambda_n (1-ih))}{\lambda_n} \\
&- \delta_{i-1,j} b \sum_{n=1}^{p} a_n^2 \frac{2 - \exp(-\lambda_n (i-1)h) - \exp(-\lambda_n (1-(i-1)h))}{\lambda_n} \\
&+ \delta_{i,j} b \sum_{n=1}^{p} a_n^2 \frac{2 - \exp(-\lambda_n ih) - \exp(-\lambda_n (1-ih))}{\lambda_n} \\
&+ \delta_{i,j} b \sum_{n=1}^{p} a_n^2 \frac{2 - \exp(-\lambda_n (i-1)h) - \exp(-\lambda_n (1-(i-1)h))}{\lambda_n}
\end{aligned} \tag{C.2}$$

Indeed,

$$\begin{aligned}
Cov[IV_t, r_{t-1+ih} r_{t-1+jh}] &= Cov[IV_t, (r^*_{t-1+ih} + e_{t-1+ih})(r^*_{t-1+jh} + e_{t-1+jh})] \\
&= \delta_{i,j} Cov[IV_t, r^{*2}_{t-1+ih}] + Cov[IV_t, e_{t-1+ih} e_{t-1+jh}] \\
&= \delta_{i,j} Cov[IV_t, r^{*2}_{t-1+ih}] - \delta_{i,j-1} Cov[IV_t, u^2_{t-1+ih}] - \delta_{i-1,j} Cov[IV_t, u^2_{t-1+(i-1)h}] \\
&+ \delta_{i,j} Cov[IV_t, u^2_{t-1+ih}] + \delta_{i,j} Cov[IV_t, u^2_{t-1+(i-1)h}]
\end{aligned} \tag{C.3}$$

For the first term, using (2.20) in ABM(2011),

$$
\begin{aligned}
Cov[IV_t, r^{*2}_{t-1+ih}] &= 2\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}(\exp(-\lambda_n h)+\lambda_n h-1) \\
&+\sum_{n=1}^{p}\frac{a_n^2}{\lambda_n^2}\left(2-\exp(-\lambda_n(i-1)h)-\exp(-\lambda_n(1-ih))\right)\left(1-\exp(-\lambda_n h)\right).
\end{aligned} \tag{C.4}
$$

For a given integer k, we have

$$
\begin{aligned}
&Cov[IV_t, u^2_{t-1+kh}] = E[E_\sigma[IV_t u^2_{t-1+kh}]] - \underbrace{E[IV_t]}_{=a_0}\underbrace{E[u^2_{t-1+kh}]}_{=V_u=a+ba_0} \\
&= aa_0 + bE[\sigma^2_{t-1+kh}\int_{t-1}^{t}\sigma^2_s ds] - a_0V_u \\
&= aa_0 + bE[(a_0 + \sum_{n=1}^{p} a_n P_n(f_{t-1+kh}))\int_{t-1}^{t}(a_0 + \sum_{m=1}^{p} a_m P_m(f_s))ds] - a_0V_u \\
&= b\sum_{n,m=1}^{p} a_n a_m \int_{t-1}^{t} E[P_n(f_{t-1+kh})P_m(f_s)]ds \\
&= b\sum_{n,m=1}^{p} a_n a_m \int_{t-1}^{t-1+kh} E[P_n(f_{t-1+kh})P_m(f_s)]ds + b\sum_{n,m=1}^{p} a_n a_m \int_{t-1+kh}^{t} E[P_n(f_{t-1+kh})P_m(f_s)]ds \\
&= b\sum_{n,m=1}^{p} a_n a_m \int_{t-1}^{t-1+kh} E[E[P_n(f_{t-1+kh})P_m(f_s)|f_\tau, \tau \le s]]ds \\
&+ b\sum_{n,m=1}^{p} a_n a_m \int_{t-1+kh}^{t} E[E[P_n(f_{t-1+kh})P_m(f_s)|f_\tau, \tau \le t-1+kh]]ds \\
&= b\sum_{n,m=1}^{p} a_n a_m \int_{t-1}^{t-1+kh} E[P_m(f_s)\ \underbrace{E[P_n(f_{t-1+kh})|f_\tau, \tau \le s]}_{=\exp(-\lambda_n((t-1+kh)-s))P_n(f_s)}]ds \\
&+ b\sum_{n,m=1}^{p} a_n a_m \int_{t-1+kh}^{t} E[P_n(f_{t-1+ih})\ \underbrace{E[P_m(f_s)|f_\tau, \tau \le t-1+kh]}_{=\exp(-\lambda_m(s-(t-1+kh)))P_m(f_{t-1+kh})}\ ]ds \\
&= b\sum_{n=1}^{p} a_n^2 \int_{t-1}^{t-1+kh} \exp(-\lambda_n((t-1+kh)-s))ds + b\sum_{n=1}^{p} a_n^2 \int_{t-1+kh}^{t} \exp(-\lambda_n(s-(t-1+kh)))ds \\
&= b\sum_{n=1}^{p} a_n^2 \frac{2-\exp(-\lambda_n kh)-\exp(-\lambda_n(1-kh))}{\lambda_n}
\end{aligned}
\tag{C.5}
$$

Appendix D : A practical adjustment

For the alternative realized measures, we can correct the the Mincer-Zarnowitz regression R^2 in practice. In Andersen et al. (2005), practical error corrections are provided using the fact that the integrated volatility is latent. Recall the R^2 expression,

$$R^2(IV_{t+1}, RM_t(h)) = \frac{Cov[IV_{t+1}, RM_t(h)]^2}{Var[IV_{t+1}]Var[RM_t(h)]}.$$

A practical adjustment is to replace $Var[IV_{t+1}]$ in the denominator by

$$Var[\widehat{IV}_{t+1}] - f(h)Avar(\widehat{IV}_{t+1} - IV_{t+1}) + o(f(h)),$$

where $\widehat{IV}_{t+1}$ is a consistent estimator of IV_{t+1}, $Avar(\widehat{IV}_{t+1} - IV_{t+1})$ is the asymptotic variance, and $f(h)$ is the convergence rate. This adjustment does not apply for non consistent estimators of integrated volatility. For our model, only robust to heteroscedastic noise volatility estimators are consistent. The two time scales estimator is not robust to heteroscedastic noise. The traditional estimator namely, the realized variance computed at the highest frequency and the average estimator, are not robust to any type of noise. However, the only estimators for which a practical adjustment could be applied are the pre-averaging, the kernel estimators because they are robust to heteroscedastic noise, and obviously our new realized measure.

ARTICLE 3

A DISTRIBUTIONAL APPROACH TO REALIZED VOLATILITY

Abstract [1]

This paper proposes new measures of the integrated variance that uses high frequency bid-ask spreads and volumes. The traditional approach assumes that the mid-quote is a good measure of frictionless price. The recent econometric literature explicitly assumes that the mid-quote is a noisy measure of the frictionless price and proposed new and robust measures of the integrated variance. This paper departs from the literature by specifying the conditional distribution of the frictionless price given the available information which includes quotes and volumes. The distributional assumption allows one to characterize the conditional mean of the integrated variance, which we take as new measures of the integrated variance. We then compare empirically the new measures with the robust ones when one deals with forecasting integrated variance or trading options. We show that the new measures dominate in some cases the traditional measures.

Key phrases : Realized variance, bid-ask spread, quoted depths, volatility forecasting, option trading.

[1] We thank Bruno Feunou for providing us with the high frequency data used in this paper.

3.1 Introduction

We are interested in measuring the integrated variance of asset returns using bid and ask prices. Measuring volatility using high frequency data has attracted a growing interest for many reasons. First, thanks to large data samples availability, we can observe almost continuous data processes, which in turn justifies the continuous time framework use. The Trades and Quotes (TAQ) database usually releases one-second frequency prices and quotes, but recently it released one-millisecond frequency data. Such an ultra high frequency dataset opens up research opportunities to explore intraday volatility features and spot volatility estimation. The second major reason for the growing interest in using high frequency data to measure volatility, is that the model free approach of the theory of quadratic variation is not vulnerable to model misspecification, as is the case with other approaches from the parametric literature.

In this paper we assume conditional distributional assumptions on the frictionless price. The common approach is to assume that the mid-quotes price - the bid and ask prices average - is the sum of the frictionless price and a noise term. By making assumptions on the noise, one could derive consistent estimators of the integrated variance ; see, e.g., Zhang et al. (2005), Zhang (2006), Barndorff-Nielsen et al. (2008), and Jacod et al. (2009). Early assumptions hypothesize an exogenous iid dynamic for the noise. Later on, it was relaxed to allow for some forms of endogeneity with the frictionless price and an autocorrelated noise. The problem is that, since noise is not observed it is difficult to be precise about its time-varying characteristics.

The present paper follows a novel approach. As a first attempt, we derive bounds on the integrated variance when assuming that the frictionless price lies between the bid and the ask prices. Such a non-point-identification (also known as partial identification) approach was initiated by Manski (2003) and later surveyed by Tamer (2010). Unfortunately, this approach leads to wide bounds, implying that one needs to make additional assumptions. Our main approach consists in making distributional assumptions on the frictionless price conditioned on quoted data (the bid, ask, and depths). We then derive new realized volatility measures. One important feature of the new measures is the explicit presence of the bid-ask spread variable. So far, in market microstructure theory, the spread has been only implicitly shown to impound information about volatility ; see Hasbrouck (1999).

We consider different distributions and evaluate them through three directions. First, the ability to capture the noise at high frequency using the signature plot ; see Andersen et al. (1999) [2]. Second, the forecast performance in-sample and out-of-sample. For instance Andersen et al. (2003) evaluates the forecasting abilities of the standard realized variance and Aït-Sahalia and Mancini (2008) study the forecasting of integrated volatility using the robust to noise estimator ; the two time scales estimator. Third, we quantify the pecuniary gain or loss for option pricing in a hypothetical market as in Bandi et al. (2008). We show that some new measures outperform the existing measures.

We carry out our analysis by adding the quoted depths (the ask volume and the bid volume) to the conditioning information set. The ask (resp. bid) volume is the maximum number of shares to buy (resp. sell) at the ask (resp. bid) price. The quoted depths reveal information about the stock liquidity and inventory control ; see Kavajecz (1999). Consequently, using the depths may lessen the microstructure frictions effect. The boun-

[2] The signature plot of the realized variance draws the realized variance against sampling frequencies.

ded distributions for the frictionless price that we use are the uniform and the triangular over the bid-ask interval. We also accommodate the normal distribution to a bounded support. We explicitly model the correlation between successive prices.

When it comes to the empirical section, we use data from the Alcoa stock traded in the New York Stock Exchange during the 01/2009-03/2011 period. We find that the best measures stemming from the forecasting exercise are different from those based on the option trading exercise. Moreover, the new realized measures could outperform the traditional robust to noise volatility estimators.

The rest of the paper is structured as follows. First we present the common realized measures, the Mincer-Zarnowitz regression for forecasting evaluation, and the option trading exercise. Then, we state the distributional assumptions and the new volatility measures that they imply. We also assess the forecasting performance of each new realized measure. Finally, we provide a conclusion.

3.2 The forecasting performance of the realized measures

In what follows, t stands for the day. One observes a sample of size N of intra-day bids and asks denoted b_{t-1+ih}, a_{t-1+ih} in log terms, where $i = 1..N$ and h is the sampling frequency. The logarithm of the frictionless price is latent and denoted p_{t-1+ih}. In all the paper, we let b_i, a_i, and p_i stand for $b_{t-1+i/N}$, $a_{t-1+i/N}$, and $p_{t-1+i/N}$, respectively. The intra-day return is given by

$$r_i = p_i - p_{i-1}, \tag{3.1}$$

We suppose that the frictionless price follows a semimartingale given by,

$$dp_s = \mu_s ds + \sigma_s dW_s, \tag{3.2}$$

where W_s is a Wiener process and σ_t is a *càdlàg* volatility function. The object of interest is the integrated variance for a given day t defined as,

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds \tag{3.3}$$

The realized variance is defined as

$$RV_t(h) = \sum_{i=1}^{1/h} r_{t-1+ih}^2, \tag{3.4}$$

where $r_{t-1+ih} = p_{t-1+ih} - p_{t-1+(i-1)h}$. The realized variance computed with the highest frequency returns would be a consistent estimator of the integrated variance if the observed price is equal to the frictionless price ; see Jacod (1994), and Barndorff-Nielsen and Shephard (2002).

Let m_{t-1+ih} and s_{t-1+ih} denote the mid-quote and the spread, respectively. We have,

$$m_{t-1+ih} = \frac{a_{t-1+ih} + b_{t-1+ih}}{2}, \tag{3.5}$$

and

$$s_{t-1+ih} = a_{t-1+ih} - b_{t-1+ih}. \tag{3.6}$$

In this paper, we make assumptions about the distribution of the frictionless price conditioning on the quotes data. These data include bid, ask prices and, quoted depths (bid and ask depth) that specify the maximum quantity for which the ask (bid) price applies. Such assumption is not conflicting with the previous semimartingale assumption for the price. In this paper, we work in a discrete time setting as we directly make distributional hypothesis about successive intraday prices.

3.2.1 The common realized measures

The realized variance defined in equation (3.4) is an inconsistent estimator of the integrated variance because of the market microstructure noise that contaminates frictionless prices. An empirical evidence for the presence of the noise is the signature plot introduced by Andersen et al. (1999). The signature plot draws a sample average of daily realized measure of volatility as a function of the underlying returns sampling frequency. A graph that explodes at high frequencies is an evidence for market microstructure noise severity. At low frequencies, the plot converges to the integrated variance measure and the noise effect disappears. Fig. 3.1 presents the signature plot for Alcoa. We use data covering the 01/2009-03/2011 period in all the paper.

If the highest frequency returns are used to compute the realized variance, we obtain the all estimator given by

$$RV_t^{all} = \sum_{i=1}^{N} r_i^2. \tag{3.7}$$

If a low frequency $\overline{h}$ is used to compute the returns, we obtain the following estimator

$$RV_t^{low} = RV_t(\overline{h}) = \sum_{i=1}^{1/\overline{h}} r^2_{t-1+i\overline{h}}. \tag{3.8}$$

The two time scales estimator of Zhang et al. (2005), which is consistent under i.i.d. market microstructure noise, is defined as,

$$RV_t^{TS} = RV_t^{average} - \frac{\overline{N}}{N} RV_t^{all}, \tag{3.9}$$

where the average estimator $RV^{average}$ is the mean of several sparse estimators, formally

$$RV_t^{average} = \frac{1}{K} \sum_{k=1}^{K} RV_t^{(k)}, \tag{3.10}$$

where $RV_t^{(k)} = \sum_{i=1}^{N-k+1} r^2_{t-1+(i+k-1)h}$, and h is the sampling frequency.

The kernel estimator of Barndorff-Nielsen et al. (2008) achieves a faster rate of convergence than the RV^{TS}, and is defined as

$$RV_t^{kernel} = \gamma_0 + \sum_{l=1}^{L} f\left(\frac{l-1}{L}\right) \{\gamma_l + \gamma_{-l}\}, \tag{3.11}$$

where $\gamma_l = \sum_{j=1}^{N} r_{t-1+jh} r_{t-1+(j-l)h}$, $f(x) = \; (1 \; - \; cos\pi(1-x)^2)/2$, and L is the bandwidth.

The pre-averaging estimator introduced by Jacod et al. (2009) is robust to heteroscedastic market microstructure noise, and achieves an optimal rate of convergence. We denote

RV_t^{pre} the pre-averaging estimator given by,

$$RV_t^{pre} = \sum_{i=0}^{N-k} \left\{ \sum_{j=1}^{k} \phi\left(\frac{j}{k}\right) r_{i+j} \right\}^2 - \frac{6}{\theta^2} RV_t^{all},$$

where $\frac{k}{\sqrt{N}} = \theta + \mathscr{O}(N^{-1/4})$ for some $\theta > 0$, and $\phi(x) = min(x, 1-x)$.

3.2.2 Mincer-Zarnowitz regression

In order to assess the forecasting performance of a realized measures RM_t, we use a Mincer-Zarnowitz regression given by,

$$IV_{t+1} = \alpha + \beta RM_{t+1|t} + \eta_{t+1}, \tag{3.12}$$

where $RM_{t+1|t}$ is the forecast at time t of IV_{t+1} using the measure RM, and η_{t+1} is an error term. A forecast $RM_{t+1|t}$ is good if $\alpha = 0$, $\beta = 1$, and a high R^2. Since the dependent variable is latent, we use RV_{t+1}^{low} as a proxy for IV_{t+1}. In second paper of this thesis, we discuss a bias correction that results from not observing the dependent variable. For longer forecasting horizon H, the Mincer-Zarnowitz regression is given by,

$$IV_{t:t+H} = \alpha + \beta RM_{t+H|t} + \eta_{t+H}. \tag{3.13}$$

where $IV_{t:t+H} = \int_t^{t+H} \sigma_s^2 ds$, and $RM_{t+H|t}$ is the forecast at time t of $IV_{t:t+H}$ using the measure RM, and η_{t+H} is an error term. The forecasting model that we use is an AR(3) model. We conduct an in-sample and an out-of-sample forecasting exercise.

In Table 3.I, we report the R^2 for the in-sample and out-of-sample forecasts for one day and 5 days horizon. We use an AR(3) forecasting model (the dependent variable is RV^{low}) with a 100 days rolling window. For the short horizon, the pre-averaging estimator achieves the highest R^2 whether in-sample or out-of-sample. The realized variance RV^{all} has the least R^2. However, the R^2 for the estimators RV^{all}, RV^{TS}, RV^{kernel}, and RV^{pre} are close. For the longer horizon of 5 days, we find that the overall forecasting performance of the 4 estimators has improved upon the short horizon. The RV^{TS} becomes the best forecast whether in-sample and out-of-sample when H equals 5 days.

3.2.3 Option trading

In this section we evaluate the proposed integrated volatility estimates in the context of the profits from option pricing and trading economic metric. Using alternative forecasts obtained in the previous section, agents price short-term options on Alcoa stock before trading with each other at average prices. The average profits is used as the criteria to evaluate alternative volatility estimates and the corresponding forecasts.

We construct an hypothetical option market as in Bandi, Russell and Yang (2008) in order to quantify the economic gain or loss for using alternative integrated volatility measures. Our artificial market has as many traders as alternative forecasts. Each trader uses a different measure from the set of realized measures.

First, each trader constructs an out-of sample one day ahead variance forecast using his daily variances series and computes his predicted Black-Scholes option price. We focus on an at-the-money price of a 1-day or 5-days options on a 1 Dollar share of Alcoa. The risk free rate is taken to be zero.

Second, the pair-wise trades take place. For two given traders, if the forecast of the first one is higher than the mid-point of the forecasts of the two traders, than the option is perceived as underpriced. And the first trader will buy a straddle (one call and one put) from his counterpart. Then the positions are hedged using the deltas of the options.

Finally, we compute the profits or losses. Each trader averages the profits or losses from pair-wise trading. We report the average profits across all days in the sample.

The option trading and profit results are computed as in the following three steps,

1-Let σ_t denote the volatility forecast for a given measure. The Black-Scholes option price P_t is given by,

$P_t = 2\Phi(\frac{1}{2}\sigma_t) - 1$, where Φ is the cumulative normal distribution.

2-The daily profit for a trader who buys the straddle is :

$\mid R_t \mid -2P_t + R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$, where the last term corresponds to the hedging, and R_t is the daily return for day t.

The daily profit for a trader who sells the straddle is :

$2P_t - \mid R_t \mid -R_t(1 - 2\Phi(\frac{1}{2}\sigma_t))$.

3- We then average the profits and obtain the metric.

We use the out-of-sample forecasts of the section 3.2.2. We report the profits/losses in Cents in Tables 3.II when all the realized measures are used in the trading game. We find that the agents using the 4 traditional measures RV^{all}, RV^{TS}, RV^{kernel}, and RV^{pre} endure losses. For the one day horizon, the RV^{pre} is the worst estimator, whereas it becomes the best at the 5 days horizon. The inverse is observed for RV^{all} where it is the best at short horizon but the worst at long horizon.

3.3 Simple bounds

We suppose that the frictionless price is bounded by the bid and the ask. This assumption is restrictive since the frictionless price could be less than the bid or higher than the ask. In the future, it would be interesting to consider the case where the frictionless price could lie outside the bid ask interval. Formally we have,

Assumption A $b_i \leq p_i \leq a_i$, $i = 1...N$.

In the next proposition, we compute the realized variance bounds under Assumption A.

Proposition 1. *Under Assumption A,*

$$RV_t^{inf} \leq RV_t \leq RV_t^{sup},$$

where

$$\begin{aligned} RV_t^{sup} &= \sum_{i=1}^{1/h} \overline{r}_i^2, \\ RV_t^{inf} &= \sum_{i=1}^{1/h} \underline{r}_i^2, \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} \overline{r}_i^2 &= Max\left\{(b_i - a_{i-1})^2; (a_i - b_{i-1})^2\right\}, \\ \underline{r}_i^2 &= \left\{ \begin{array}{cc} 0 & if\ (b_i - a_{i-1})(a_i - b_{i-1}) \leq 0 \\ Min\left\{(b_i - a_{i-1})^2; (a_i - b_{i-1})^2\right\} & else. \end{array} \right\} \end{aligned} \tag{3.15}$$

The bounds derived in Proposition 1 are not tight to be informative. Indeed, they are based on very weak assumptions. We draw the signature plot of RV^{inf} and RV^{sup} in Fig.

3.1. At high frequencies, the interval $[RV^{inf}, RV^{sup}]$ is very wide. At low frequencies, this interval becomes narrow. For the forecasting results, we find in Table 3.I that the bound RV^{sup} beats all the traditional measures RV^{all}, RV^{TS}, RV^{kernel}, and RV^{pre}, whether in-sample, out-of-sample, short or long horizon forecasting. The lower bound RV^{inf} has the worst results compared to the traditional measures at short horizon but beats them at long forecasting horizon. The profits/losses from the option trading exercise are reported in Table 3.II. At long horizon, the upper bound RV^{sup} achieves a big profit whereas the lower bound RV^{inf} has a big loss.

In the following section, our goal is to examine some distributional restrictions for the price. In fact, imposing much more restrictions may provide better volatility estimators.

3.4 A distributional approach

In this section, we impose more restrictions on the distribution of the price. We evaluate the new realized measures using signature plots, Mincer-Zarnowitz regression for forecasting performance, and option trading outcome.

3.4.1 A Dirac measure

If we assume that the frictionless price p_i follows a Dirac measure in the mid quotes, we obtain the usual expression for the the realized volatility,

$$RV_t^{all} = \sum_{i=1}^{N} (m_i - m_{i-1})^2. \tag{3.16}$$

If we assume that the frictionless price p_i follows a Dirac measure in the bid, we obtain the measure,

$$RV_t^{bid} = \sum_{i=1}^{N} (b_i - b_{i-1})^2, \tag{3.17}$$

The same applies for the ask, and the corresponding measure is given by,

$$RV_t^{ask} = \sum_{i=1}^{N} (a_i - a_{i-1})^2. \tag{3.18}$$

The signature plots in Fig. 3.1 of RV_t^{bid} and RV_t^{ask} are very close and more noisy than RV_t^{all} at high frequencies. In Table 3.I reporting the forecasting results, show that RV_t^{bid} and RV_t^{ask} have similar forecasting performance with the traditional realized measures. However, the option trading profits/losses in Table 3.II are positive for the 5 days horizon contrarily to the negative outcome of the traditional realized measures. At short horizon, the traders using RV_t^{bid} and RV_t^{ask} endure losses comparable to the traditional measures.

3.4.2 Univariate distributions

We take the set of the intraday quotes as the conditioning set,

$$I = \{b_j, a_j, j = 1, ..., N\}. \tag{3.19}$$

We derive the components of the squared return conditional expectation,

$$\begin{aligned} E[r_i^2 \mid I] &= (E[r_i \mid I])^2 + Var[r_i \mid I] \\ &= (E[p_i \mid I] - E[p_{i-1} \mid I])^2 \\ &+ Var[p_i \mid I] + Var[p_{i-1} \mid I] - 2Cov[p_i, p_{i-1} \mid I]. \end{aligned} \tag{3.20}$$

We make the following assumption.

Assumption B

Conditionally on I, r_{t-1+ih} and $p_{t-1+(i-1)h}$ are independent.

Assumption B specifies that any intraday return is conditionally independent from the previous price. In Proposition 2, we use Assumption B and the expression (3.20) to derive the conditional expectation of the realized variance.

Proposition 2. *Under Assumption A and B,*

$$\begin{aligned}
&E[RV_t(h) \mid I] \\
&= \sum_{i=1}^{1/h} (E[p_{t-1+ih} \mid I] - E[p_{t-1+(i-1)h} \mid I])^2 + Var[p_{t-1+ih} \mid I] + Var[p_{t-1+(i-1)h} \mid I] \\
&- 2Min\left(\sqrt{\frac{Var[p_{t-1+(i-1)h} \mid I]}{Var[p_{t-1+ih} \mid I]}}, 1\right)\sqrt{Var[p_{t-1+ih} \mid I]}\sqrt{Var[p_{t-1+(i-1)h} \mid I]}.
\end{aligned} \tag{3.21}$$

The equation (3.21) is only function of the expectation and the variance. Therefore, by varying the distributional hypothesis about the intraday price we obtain different estimators. We define the resulting realized measure as,

$$RM_t = E[RV_t \mid I], \tag{3.22}$$

We specifically examine the realized measures based on uniform and triangular distributional assumptions.

3.4.2.1 The uniform distribution

We suppose that the intraday price follows a uniform distribution. Formally,

$$p_i \mid I \sim \; Uniform[b_i, a_i]. \tag{3.23}$$

The uniform distribution is such that all intervals of the same length on the distribution's support $[b_i, a_i]$ are equally probable. The first two moments are given by,

$$\begin{aligned} E[p_i \mid I] &= m_i, \\ Var[p_i \mid I] &= \frac{s_i^2}{12}. \end{aligned} \tag{3.24}$$

We define $RV^{uniform}$ using equations (3.21) and (3.22),

$$RV_t^{uniform} = \sum_{i=1}^{1/h} \left\{ (m_i - m_{i-1})^2 + v_i + v_{i-1} - 2Min\{\sqrt{\frac{v_{i-1}}{v_i}} 1\} \sqrt{v_i}\sqrt{v_{i-1}} \right\}, \tag{3.25}$$

where $v_i = \frac{s_i^2}{12}$. The bid-ask spread appears in the new realized variance $RV^{uniform}$. The spread is a friction measure that is not yet explored -to our knowledge- in the high frequency literature to measure volatility.

The empirical results of $RV^{uniform}$ show similar forecasting performance to the traditional realized measures as reported in Table 3.I. However, short horizon the trader using $RV^{uniform}$ endures the smallest loss among the realized measures introduced so far as shown in Table 3.II. At 5 days horizon, the trader using $RV^{uniform}$ achieves a profit. The signature plots in Figure 3.1 show that at high frequencies $RV^{uniform}$ is more noisy than the realized variance RV^{all}.

3.4.2.2 The triangular distribution

Let the frictionless price follow a centered triangular distribution,

$$p_i \mid I \sim \; Centered\ Triangular[b_i, a_i]. \tag{3.26}$$

The centered triangular distribution has an affine probability density function. The mid-quotes is the most probable of the distribution support. The expectation and variance expressions are respectively,

$$\begin{aligned} E[p_i \mid I] &= m_i, \\ Var[p_i \mid I] &= \frac{b_i^2 + a_i^2 + m_i^2 - b_i a_i - b_i m_i - m_i a_i}{18}. \end{aligned} \tag{3.27}$$

As for the uniform distribution, we define $RV^{triangular}$ using equations (3.21) and (3.22) as,

$$RV_t^{triangular} = \sum_{i=1}^{1/h} \left\{ (m_i - m_{i-1})^2 + v_i + v_{i-1} - 2Min\{\sqrt{\frac{v_{i-1}}{v_i}}; 1\}\sqrt{v_i}\sqrt{v_{i-1}} \right\}, \tag{3.28}$$

where

$$v_i = \frac{b_i^2 + a_i^2 + m_i^2 - b_i a_i - b_i m_i - m_i a_i}{18}. \tag{3.29}$$

The new realized measure $RV^{triangular}$ uses the bid, ask and mid quotes. Assuming that the frictionless price has a centered triangular distribution means that the mid quotes is the most probable value for the frictionless price whereas the other values in the bid ask interval are realized with non zero probability, the least probabilities are near the edge of $[b_i, a_i]$.

Empirically, we find that $RV^{triangular}$ is less noisy than the univariate based measures RV^{bid}, RV^{ask}, and $RV^{uniform}$ as shown in the signature plots of Fig. 3.1. The forecasting performance of $RV^{triangular}$ measured by the R^2 of the Mincer-Zarnowitz regression is similar to the univariate based measures as reported in Table 3.I. However, the trader using $RV^{triangular}$ endures losses at 5 days horizon in the option trading exercise contrarily to the other univariate based measures (see Table 3.II). At the one day horizon, $RV^{triangular}$ gives less losses than the traditional realized measures.

3.4.3 Bivariate distributions

In this section, we do not assume the independence form of successive intraday prices specified by Assumption B. We rather specify the joint distribution of each successive intraday prices and use the general equation (3.20) to find the realized measure expression. We denote $\rho(h)$ the correlation between two intraday prices p_{t-1+ih} and $p_{t-1+(i-1)h}$. We assume that,

$$(i)\rho(.) \in [0,1], decreasing,$$

$$(ii)\rho(0) = 1; lim_{h\to\infty}\rho(h) = 0.$$

Assumption (i) implies that the correlation parameter decreases as the time interval between successive observations becomes larger. At the limit, (ii) assumes a zero correlation if the intraday prices are sampled at a very low frequency.

3.4.3.1 The bivariate normal distribution

In this section, we assume a joint normal distribution for successive prices. Although the normal distribution has not a bounded support, we parameterize it to have very slim tails as if the distribution has almost bounded support (coherently with Assumption A). Formally we assume,

$$\begin{bmatrix} p_i \\ p_{i-1} \end{bmatrix} | I \sim \mathscr{N}(\begin{bmatrix} m_i \\ m_{i-1} \end{bmatrix}, \begin{pmatrix} v_i & c_{i;i-1} \\ c_{i;i-1} & v_{i-1} \end{pmatrix}), \tag{3.30}$$

where,

$$c_{i,i-1} = \rho(h)\sqrt{v_i v_{i-1}}, \tag{3.31}$$

and,

$$\begin{aligned} v_i &= \lambda^2 s_i^2, \\ v_{i-1} &= \lambda^2 s_{i-1}^2. \end{aligned} \tag{3.32}$$

λ is a constant such that $P[b_i \leq p_i \leq a_i]$=0.99 and $P[b_{i-1} \leq p_{i-1} \leq a_{i-1}]$=0.99, i.e. $\lambda = 0.19$.

We define the measure $RV^{corr.Normal}$, using equations (3.20) and (3.22) by,

$$RV^{corr.Normal} = \sum_{i=1}^{1/h} (m_i - m_{i-1})^2 + v_i + v_{i-1} - 2c_{i,i-1}, \tag{3.33}$$

where the variance and covariance are given in (3.31)-(3.32).

The new measure $RV^{corr.Normal}$ is a function of the friction measure -the spread- and the correlation parameter that we specify ad hoc.

The empirical performance of the realized measure $RV^{corr.Normal}$ is similar to the traditional RV^{all} when one look at the signature plot, the Mincer-Zarnowitz regression results, and the short horizon outcome of the option trading game ; see Fig.3.1, Table 3.I, and Table 3.II, respectively. However, the trader using $RV^{corr.Normal}$ endures much less loss than the trader using RV^{all} for the long horizon.

3.4.3.2 The bivariate uniform distribution

We assume the bivariate distribution,

$$\begin{aligned} &p_i \mid I \sim \; Uniform[b_i, a_i], \\ &p_{i-1} \mid I \sim \; Uniform[b_{i-1}, a_{i-1}], \\ &Cov[p_i, p_{i-1} \mid I] = \frac{\rho(h)}{12} s_i s_{i-1}. \end{aligned}$$

Using equations (3.20) and (3.22), we define the measure $RV^{corr.Uniform}$ by

$$RV^{corr.Uniform} = \sum_{i=1}^{1/h} (m_i - m_{i-1})^2 + \frac{s_i^2}{12} + \frac{s_{i-1}^2}{12} - 2\frac{\rho(h)}{12} s_i s_{i-1}. \tag{3.34}$$

The signature plot and the forecasting results of $RV^{corr.Uniform}$ are similar to the $RV^{corr.Normal}$ as shown in Fig. 3.1 and Table 3.I. For the short horizon, $RV^{corr.Uniform}$ achieves the smallest loss compared to the measures introduced so far including the traditional measures as reported in Table 3.II. We also notice that the long horizon loss of $RV^{corr.Uniform}$

is smaller than the one for $RV^{corr.Normal}$.

3.4.3.3 The bivariate triangular distribution

We assume that successive intraday prices follow the joint distribution given by,

$$\begin{aligned}
&p_i \mid I \sim\ Triangular\{[b_i,a_i];m_i\},\\
&p_{i-1} \mid I \sim\ Triangular\{[b_{i-1},a_{i-1}];m_{i-1}\},\\
&Cov[p_i,p_{i-1} \mid I] = \frac{\rho(h)}{\sqrt{a_i-b_i}\sqrt{a_{i-1}-b_{i-1}}}[\qquad (3.35)\\
&\frac{1}{18}((m_i-b_i)^{3/2}(m_{i-1}-b_{i-1})^{3/2}+(a_i-m_i)^{3/2}(a_{i-1}-m_{i-1})^{3/2})\\
&+(\frac{\pi}{8}+\frac{4}{9})((m_i-b_i)^{3/2}(a_{i-1}-m_{i-1})^{3/2}+(a_i-m_i)^{3/2}(m_{i-1}-b_{i-1})^{3/2})].
\end{aligned}$$

Using equations (3.20) and (3.22), we define the measure $RV^{corr.Triangular}$ by,

$$RV^{corr.Triangular} = \sum_{i=1}^{1/h}(m_i-m_{i-1})^2+v_i+v_{i-1}-2Cov[p_i,p_{i-1} \mid I], \qquad (3.36)$$

where

$$v_i = \frac{b_i^2+a_i^2+m_i^2-b_ia_i-b_im_i-m_ia_i}{18}, \qquad (3.37)$$

and the covariance expression is given in (3.35).

The signature plot of $RV^{corr.Triangular}$ in Fig. 3.1 shows more bias at high frequencies than the other bivariate Uniform and Normal distributions based measures, and even the traditional realized variance RV^{all}. The forecasting performance of $RV^{corr.Triangular}$ as reported in Table 3.I is better than $RV^{corr.Normal}$ and $RV^{corr.Uniform}$ whether at short or long horizons, and in-sample or out-of-sample. The trader using $RV^{corr.Triangular}$ achieves the best profit compared to the overall realized measures of this paper for the long horizon as shown in Table 3.II.

3.5 The volume information

In order to explore the volume information, we include the intraday quoted depths in the conditioning set I.

3.5.1 The Dirac distribution

We define a weighted volume measure $RV_t^{depths.Weighted}$ as in Gatheral and Oomen (2010) by,

$$RV_t^{depths.Weighted} = \sum_{i=1}^{N} \left(\frac{V_i^B a_i + V_i^A b_i}{V_i^A + V_i^B} - \frac{V_{i-1}^B a_{i-1} + V_{i-1}^A b_{i-1}}{V_{i-1}^A + V_{i-1}^B} \right)^2 , \tag{3.38}$$

where V^B (resp. V^A) denotes the bid depth (resp. the ask depth).

The forecasting results using Alcoa data in Table 3.I show that $RV_t^{depths.Weighted}$ has the highest R^2 whether in-sample or out-of-sample, and for short or long horizons, among the other Dirac based measures RV^{all}, RV^{bid}, and RV^{ask}. The signature plot depicted in Fig. 3.1 shows evidence that $RV_t^{depths.Weighted}$ is less noisy than RV^{bid} and RV^{ask}, but more noisy than RV^{all} at high frequencies. For the option trading exercise, $RV_t^{depths.Weighted}$ is the unique realized measure that achieves profits for its user at both short and long horizons.

3.5.2 The univariate triangular distribution

We assume the price distribution,

$$P_i \mid I \sim \ Non\,Centered\,Triangular[b_i, a_i]. \tag{3.39}$$

We denote c_i the mode, or the most probable value of the distribution support $[b_i, a_i]$. The first moments are then given by,

$$\begin{aligned} E[p_i \mid I] &= \frac{a_i + b_i + c_i}{3}, \\ Var[p_i \mid I] &= \frac{b_i^2 + a_i^2 + c_i^2 - b_i a_i - b_i c_i - c_i a_i}{18}. \end{aligned} \tag{3.40}$$

Using the quoted depths, we incorporate the volume information in the mode expression. Let's denote the volume increments by $\Delta V_i^A = V_i^A - V_{i-1}^A$ and $\Delta V_i^B = V_i^B - V_{i-1}^B$. We set the mode in the following way,

$$c_i = \left\{ \begin{array}{ll} \frac{|\Delta V_i^B|}{|\Delta V_i^A| + |\Delta V_i^B|} b_i + \frac{|\Delta V_i^A|}{|\Delta V_i^A| + |\Delta V_i^B|} a_i \; if \; \Delta V_i^A \Delta V_i^B \neq 0 & \\ 0.95 b_i + 0.05 a_i & if \; \Delta V_i^A = 0; \Delta V_i^B \neq 0 \\ 0.05 b_i + 0.95 a_i & if \; \Delta V_i^A \neq 0; \Delta V_i^B = 0 \\ 0.5 b_i + 0.5 a_i & if \; \Delta V_i^A = 0; \Delta V_i^B = 0. \end{array} \right\} \tag{3.41}$$

We define $RV^{depths.Triangular}$ using equation (3.20) and by applying Proposition 2,

$$RV^{depths.Triangular}$$

$$= \sum_{i=1}^{1/h} \left\{ \left(\frac{a_i + b_i + c_i}{3} - \frac{a_{i-1} + b_{i-1} + c_{i-1}}{3}\right)^2 + v_i + v_{i-1} - 2Min\left(\sqrt{\frac{v_{i-1}}{v_i}}, 1\right) \sqrt{v_i}\sqrt{v_{i-1}} \right\}, \tag{3.42}$$

where

$$v_i = \frac{b_i^2 + a_i^2 + c_i^2 - b_i a_i - b_i c_i - c_i a_i}{18}. \tag{3.43}$$

Observe that, the non centered triangular distribution assumption for the price implies that the most probable value for the frictionless price depend on the quoted depths variations. Incorporating the bid volume and ask volume in the mode expression makes inventory control matters for volatility estimation. The variations of the quoted depths measure how severe is the friction.

The signature plot of $RV^{depths.Triangular}$ does not beat the traditional realized variance RV^{all} at high frequencies as showed in Fig. 3.1. The volume information improves the forecasting ability of the univariate triangular based measure. Indeed, $RV^{depths.Triangular}$ has higher R^2 than $RV^{triangular}$ for all horizons and both in-sample and out-of-sample (see Table 3.I). However, the bivariate triangular based measure $RV^{corr.Triangular}$ beats both $RV^{depths.Triangular}$ and $RV^{triangular}$. Therefore, we introduce in the next section a measure that is based on a bivariate Triangular distribution and also incorporates the volume information. For the option trading exercise, Table 3.II shows that $RV^{depths.Triangular}$ is better at long horizon because it causes losses at short horizon. Moreover, the profit that the trader using $RV^{depths.Triangular}$ achieves is less than the one for the trader using $RV^{depths.weighted}$ at long horizon.

3.5.3 The bivariate triangular distribution

In this section, we use the volume information and we impose a bivariate structure for successive prices. We assume the following triangular distribution for intraday prices,

$$
\begin{aligned}
& p_i \mid I \sim \ Triangular\{[b_i, a_i]; c_i\}, \\
& p_{i-1} \mid I \sim \ Triangular\{[b_{i-1}, a_{i-1}]; c_{i-1}\}, \\
& Cov[p_i, p_{i-1} \mid I] = \frac{\rho(h)}{\sqrt{a_i - b_i}\sqrt{a_{i-1} - b_{i-1}}}[\\
& \frac{1}{18}((c_i - b_i)^{3/2}(c_{i-1} - b_{i-1})^{3/2} + (a_i - c_i)^{3/2}(a_{i-1} - c_{i-1})^{3/2}) \\
& + (\frac{\pi}{8} + \frac{4}{9})((c_i - b_i)^{3/2}(a_{i-1} - c_{i-1})^{3/2} + (a_i - c_i)^{3/2}(c_{i-1} - b_{i-1})^{3/2})],
\end{aligned}
\tag{3.44}
$$

where the mode expression is given in (3.41).

Using equations (3.20) and (3.22), we define $RV^{depths.corr.Triangular}$ as

$$
\begin{aligned}
& RV^{depths.corr.Triangular} \\
& = \sum_{i=1}^{1/h} \left(\frac{a_i + b_i + c_i}{3} - \frac{a_{i-1} + b_{i-1} + c_{i-1}}{3} \right)^2 + v_i + v_{i-1} - 2Cov[p_i, p_{i-1} \mid I],
\end{aligned}
\tag{3.45}
$$

where

$$
v_i = \frac{b_i^2 + a_i^2 + c_i^2 - b_i a_i - b_i c_i - c_i a_i}{18},
$$

and the covariance is given in (3.44).

Empirically, the trader using $RV^{depths.corr.Triangular}$ has the best profit among the overall realized measures of this paper (see Table 3.II) at the long horizon. Therefore, it is important to exploit the volume information as well as a correlated structure of successive intraday prices. However, at short horizon the trader using $RV^{depths.corr.Triangular}$ endures a loss. As expected for the forecasting exercise, $RV^{depths.corr.Triangular}$ has the highest R^2 among the all the Triangular based measures $RV^{corr.Triangular}$, $RV^{Triangular}$, and $RV^{depths.Triangular}$ whether in-sample or out-of-sample and short or long horizon as depicted in Table 3.I. Finally, the signature plot of $RV^{depths.corr.Triangular}$ in Fig. 3.1 shows a more noisy measure at high frequencies than the traditional RV^{all}.

3.6 Conclusion

In this paper, we make distributional assumptions on the frictionless price and we come up with new realized measures that incorporate the spread and the quoted depths information. To assess the performance of the new realized measures, we empirically compare their forecasting ability using Mincer-Zarnowitz regression and an option trading game. For an Alcoa data sample covering 01/2009-03/2011, we find that the new realized measures beat in many cases the common robust to noise realized measures.

Figure 3.1 – Signature plots for Alcoa 01/2009-03/2011.

Alcoa	1 day		5 days	
R^2	In	Out	In	Out
RV^{all}	0.4767	0.4746	0.5105	0.4965
RV^{TS}	0.4800	0.4873	0.5133	0.5006
RV^{kernel}	0.4870	0.4934	0.5123	0.4997
RV^{pre}	0.4927	0.4959	0.5025	0.4910
RV^{inf}	0.4159	0.4101	0.5275	0.5108
RV^{sup}	0.5121	0.5095	0.5432	0.5313
RV^{bid}	0.4851	0.4827	0.5056	0.4927
RV^{ask}	0.4754	0.4728	0.5101	0.4954
$RV^{Uniform}$	0.4804	0.4781	0.5088	0.4950
$RV^{Triangular}$	0.4787	0.4765	0.5097	0.4957
$RV^{corr.Uniform}$	0.4790	0.4768	0.5096	0.4957
$RV^{corr.Triangular}$	0.5019	0.4997	0.5238	0.5116
$RV^{corr.Normal}$	0.4778	0.4756	0.5101	0.4961
$RV^{depths.weighted}$	0.4901	0.4883	0.5179	0.5045
$RV^{depths.Triangular}$	0.4862	0.4839	0.5120	0.4983
$RV^{depths.corr.Triangular}$	0.5072	0.5049	0.5275	0.5151

Table 3.I – In-sample and out-of-sample forecasting R^2.

profits/losses	H=1	H=5
RV^{all}	-0.0635	-5.8989
RV^{TS}	-0.2562	-2.3097
RV^{kernel}	-0.2520	-3.4331
RV^{pre}	-0.5743	-1.7204
RV^{inf}	-1.5702	-10.6035
RV^{sup}	-0.4364	5.7356
RV^{bid}	-0.2634	1.3735
RV^{ask}	-0.3421	1.3403
$RV^{Uniform}$	-0.0533	0.3004
$RV^{Triangular}$	-0.0816	-2.8749
$RV^{corr.Uniform}$	0.0418	-2.0785
$RV^{corr.Triangular}$	-0.1139	5.9221
$RV^{corr.Normal}$	-0.0653	-4.5714
$RV^{depths.weighted}$	0.2051	3.0643
$RV^{depths.Triangular}$	-0.0609	2.6296
$RV^{depths.corr.Triangular}$	-0.2759	7.0041

Table 3.II – Alcoa profits/losses from option trading.

CONCLUSION

La performance de la mesure et la prévision de la volatilité fondamentale ou permanente des actifs liquides dépend du traitement des frictions microstructurelles. Dans cette thèse, on améliore l'estimation et la prévision de la volatilité en explorant deux nouvelles façons d'aborder le problème du bruit microstructure. Dans le premier et troisième papier on utilise le spread et le volume pour absorber les frictions. Alors que dans le deuxième papier, on insiste sur le caractère variable en fonction du temps des frictions.

L'apport théorique de ma thèse réside dans la dérivation des distributions asymptotiques des estimateurs de volatilité. En effet, dans le premier papier on montre que l'endogénéité ne cause pas l'inconsistence des estimateurs de volatilité comme serait le cas pour les estimateurs classiques. Aussi, le fait d'avoir une nouvelle série de rendements moins contaminés par le bruit microstructure rend l'analyse beaucoup plus simple que l'approche de pre-averaging pour réduire l'impact du bruit. Le travail théorique du deuxième papier montre la flexibilité du cadre des fonctions propres de volatilité stochastique.

Empiriquement, on utilise les données d'Alcoa de la bourse de New York couvrant la période 01/2009-03/2011 et ce pour les trois papiers de la thèse. Globalement on obtient de bon résultats. Cette période contient une phase de grande volatilité correspondant à la crise financière qui a commencé en 2009. La deuxième phase, surtout en 2011 se caractérise par une faible volatilité. Plutard, il est important de tester nos modèles sur d'autres actifs et d'autres périodes de temps.

Mon agenda de recherche contient plusieurs pistes de travail. Pour le premier papier, on pense rajouter une composante de sauts dans la dynamique du prix sans frictions et en examiner l'impact sur l'approche du papier. Intuitivement, l'estimation des paramètres du bruit microstructures ne sera pas affectée car les sauts ont une taille stochastique plus petite que le rendement sans frictions. Par contre, dans l'anlyse de la volatilité, il faudrait utiliser des mesures robustes aux sauts. Il est intéressant aussi de généraliser la forme linéaire des coûts de liquidité utilisée à une forme nonparamétrique. En effet, les nonlinéarités sont mises en évidence dans les fonctions de coûts de liquidité. L'extension multivariée est importante pour la mesure des co-volatilités et les bétas. L'étude de la volatilité spot peut aussi se faire en utilisant l'approche du premier papier.

Pour le deuxième papier, il est impératif de dériver les lois asymptotiques des deux paramètres de la variance du bruit microstructure. La difficulté technique d'un tel exercice réside dans le fait qu'on fait tendre à la fois le nombre de jours vers l'infini et le pas d'échantillonnage vers zéro. On pense aussi examiner le cas des paramètres du bruit microstructure variant à l'intérieur même de la journée apporterait plus de flexibilité. L'étude empirique de ce papier ne rejette pas le modèle de la variance du bruit. On se pose la question si ce résultat reste valide si on change la série des données utilisée.

Enfin, dans le troisième papier, il est important de relaxer l'hypothèse qui contraint le prix sans frictions à être comprix entre le prix de vente et le prix d'achat.

BIBLIOGRAPHIE

Article 1

Aït-Sahalia, Y., Mykland, P. A., Zhang, L. (2005), “How often to sample a continuous-time process in the presence of market microstructure noise,” *The Review of Financial Studies*, 2, 351-416.

Aït-Sahalia, Y., Yu, J. (2009), “High frequency market microstructure noise estimates and liquidity measures,” *Annals of Applied Statistics*, 3, 1, 422-457.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2000), “Great realizations,” *Risk*, 13, 105-108. Reprinted in J. Danielsson (ed.), The Value-at-Risk Reference, London : Risk Publications, 2008.

Andersen, T.G., Bollerslev, T., Meddahi, N. (2011), “Realized volatility forecasting and market microstructure noise,” *Journal of Econometrics*, 160, 220-234.

Bandi, F., Russell, J. (2008), “Microstructure Noise, Realized Variance, and Optimal Sampling,” *The Review of Economic Studies*, 72, 2, 339-369.

Bandi, F., Russell, J., Yang, C. (2008), “Realized volatility forecasting and option pricing,” *Journal of Econometrics*, 147, 34-46.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. ,Shephard, N. (2008), “Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise,” *Econometrica*, 76, 6, 1481-1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. ,Shephard, N. (2011), “Multivariate realised kernels : Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading," *The Journal of Econometrics*, 162, 149-169.

Carrasco, M., Kotchoni, R. (2011), "Shrinkage Realized Kernels," *Revised 2011.*

Diebold, F. X., Strasser, G. H. (2008), "On the correlation structure of microstructure noise in theory and practice," *PIER working paper* 08-038.

Glosten, L. R., Harris, L. E. (1988), "Estimating the components of the bid/ask spread," *Journal of Financial Economics*, 21, 123-142.

Hansen, P. R., Lunde, A. (2006), "Realized variance and market microstructure noise," *Journal of Business and Economic Statistics*, 24, 127-161.

Hasbrouck, J. (1991), "Measuring the information content of stock trades," *The Journal of Finance*, 46, 179-207.

Hasbrouck, J. (1999), "The dynamics of discrete bid and ask quotes," *The Journal of Finance*, 54, 2109-2142.

Hautsch, N., Podolskij, M., (2010), "Pre-averaging based estimation of quadratic variation in the presence of noise and jumps : theory, implementation, and empirical evidence," Submitted.

Huang, Roger D., Stoll, Hans R., (1997), "The Components of the Bid-Ask Spread : A General Approach," *The Review of Financial Studies*, 10, 4, 995-1034.

Jacod, J., Li, Y., Mykland, P., Podolskij, M. and Vetter, M. (2009), "Microstructure noise in the continuous case : The pre-averaging approach," *Stochastic Processes and their Applications*, 119, 2249-2276.

Klotz, J. (1972), “Markov chain clustering of births by sex,” *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, 4, 173-185, Univ. of California Press.

Kalnina, I., Linton, O., (2008), “Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error,” *Journal of Econometrics*, 147, 47-59.

Kavajecz, Kenneth A. (1999), “A Specialist’s Quoted Depth and the Limit Order Book,” *The Journal of Finance*, 54, 2, 747-771.

Li, Y., Mykland, P. A.,(2007), “Are volatility estimators robust with respect to modeling assumptions ?,” *Bernoulli* 13(3), 601-622.

Podolskij, M., Vetter, M. (2009), “Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps,” *Bernoulli*, 15(3), 634-658.

Roll, R. (1984), “A simple implicit measure of the effective bid-ask spread in an efficient market,” *The journal of Finance* 39, 4, 1127-1139.

Stoll, H. R., (2000), “Friction,” *The journal of finance* 55, 4, 1479-1514.

Zhang, L., Mykland, P. A., Aït-Sahalia, Y. (2005), “A Tale of Two Time Scales : Determining Integrated Volatility With Noisy High-Frequency Data,” *Bernoulli* 12, 1019-1043.

Zhou, B. (1996), “High-Frequency Data and Volatility in Foreign-Exchange Rates,” *Journal of Business and Economic Statistics*, 14, 45-52.

Article 2

Aït-Sahalia, Y. and Mancini L. (2008), “Out of Sample Forecasts of Quadratic Variation,” *Journal of Econometrics* 147 17-33.

Aït-Sahalia, Y. and Mykland P.A., Zhang, L. (2005), “How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise,” *The Review of Financial Studies* 18 2 351-416.

Aït-Sahalia, Y., Mykland, P. A., Zhang, L. (2011), “Ultra high frequency volatility estimation with dependent microstructure noise,” *Journal of Econometrics*, 160, 160-175.

Andersen, T.G., T. Bollerslev, P.F. Christoffersen and F.X. Diebold (2006), “Volatility And Correlation Forecasting,” in G. Elliott, C.W.J. Granger and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, North-Holland.

Andersen, T.G., Bollerslev, T., Diebold, F., Labys, P. (2003), “Modeling and Forecasting Realized Volatility,” *Econometrica*, 71 2, 579-625.

Andersen, T.G., T. Bollerslev and N. Meddahi (2004), “Analytic Evaluation of Volatility Forecasts,” *International Economic Review*, 45, 1079-1110.

Andersen, T.G., T. Bollerslev and N. Meddahi (2005), “Correcting the Errors : Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities,” *Econometrica*, 73, 279-296.

Andersen, T.G., T. Bollerslev and N. Meddahi (2006), “Realized volatility forecasting and market microstructure noise,” *Working paper*. Montreal University

Andersen, T.G., Bollerslev, T., Meddahi, N. (2011), “Realized volatility forecasting and market microstructure noise,” *Journal of Econometrics*, 160, 220-234.

Bandi, F., Russell, J. (2008), “Microstructure Noise, Realized Variance, and Optimal Sampling,” *The Review of Economic Studies*, 72, 2, 339-369.

Bandi, F., Russell, J., Yang, C. (2008), “Realized volatility forecasting and option pricing,” *Journal of Econometrics*, 147, 34-46.

Bandi, F. M., Russell, J. R., Yang, C. (2010), “Realized volatility forecasting in the presence of time-varying noise,” *Working paper*.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. ,Shephard, N. (2008), “Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise,” *Econometrica*, 76, 6, 1481-1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. ,Shephard, N. (2011), “Multivariate realised kernels : Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading,” *The Journal of Econometrics*, 162, 149-169.

Hansen, P. R., Lunde, A. (2006), “Realized variance and market microstructure noise,” *Journal of Business and Economic Statistics*, 24, 127-161.

Jacod, J., Li, Y., Mykland, P., Podolskij, M. and Vetter, M. (2009), “Microstructure noise in the continuous case : The pre-averaging approach,” *Stochastic Processes and their Applications*, 119, 2249-2276.

Kalnina, I., Linton, O., (2008), “Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error,” *Journal of Econometrics*, 147, 47-59.

Meddahi, N. (2001), “An Eigenfunction Approach for Volatility Modeling,” CIRANO working paper, 2001s-70.

Patton, A.J., Sheppard, K. (2009), “Optimal combinations of realised volatility estimators,” *International Journal of Forecasting*, 25, 218-238.

Sizova, N. (2011), “Integrated variance forecasting : Model based vs. reduced form,” *Journal of Econometrics*, 162, 294-311.

Stoll, H. R., (2000), “Friction,” *The journal of finance* 55, 4, 1479-1514.

Zhang, L., P.A. Mykland and Y. Aït-Sahalia (2005), “A Tale of Two Time Scales : Determining Integrated Volatility with Noisy High-Frequency Data,” *Journal of the American Statistical Association*, 100, 1394-1411.

Article 3

Aït-Sahalia, Y. and Mancini, L. (2008), “Out of Sample Forecasts of Quadratic Variation,” *Journal of Econometrics*, 147, 17-33.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (1999), “(Understanding, Optimizing, Using and Forecasting) Realized Volatility and Correlation,” *Working paper*. Published in revised form as “Great Realizations,” *Risk*, March 2000, 105-108

Andersen, T.G., Bollerslev, T., Diebold, F., Labys, P. (2003), “Modeling and Forecasting Realized Volatility,” *Econometrica*, 71 2, 579-625.

Bandi, F., Russell, J., Yang, C. (2008), “Realized volatility forecasting and option pricing,” *Journal of Econometrics*, 147, 34-46.

Barndorff-Nielsen, O. E., Shephard, N. (2002), “Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 2, 253-280.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. ,Shephard, N. (2008), "Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise," *Econometrica*, 76, 6, 1481-1536.

Gatheral, J., Oomen, R. C. A. (2010), "Zero-intelligence realized variance estimation," *Finance and Stochastics*, 12 2, 249-283.

Hansen, P. R., Lunde, A. (2006), "Realized variance and market microstructure noise," *Journal of Business and Economic Statistics*, 24, 127-161.

Hasbrouck, J. (1999), "The dynamics of discrete bid and ask quotes," *The Journal of Finance*, 54, 2109-2142.

Jacod, J. (1994), "Limit of random measures associated with the increments of a Brownian semimartingale," *Tech. Rep. Université Paris VI*.

Jacod, J., Li, Y., Mykland, P., Podolskij, M. and Vetter, M. (2009), "Microstructure noise in the continuous case : The pre-averaging approach," *Stochastic Processes and their Applications*, 119, 2249-2276.

Kavajecz, Kenneth A. (1999), "A Specialist's Quoted Depth and the Limit Order Book," *The Journal of Finance*, 54, 2, 747-771.

Manski, C. F. (2003), "Partial Identification of Probability Distributions,"*New York : Springer-Verlag*.

Tamer, E. (2010), "Partial Identification in Econometrics," *The Annual Review of Economics*, 167-195.

Zhang, L., Mykland, P.A. and Aït-Sahalia, Y. (2005), "A Tale of Two Time Scales : Determining Integrated Volatility with Noisy High-Frequency Data," *Journal* *of the American Statistical Association*, 100, 1394-1411.

Zhang, L. (2006), "Efficient estimation of stochastic volatility using noisy observations : A multi-scale approach," *Journal of The American Statistical Association*, 100, 1394-1411.