BANK OF CANADA
BANQUE DU CANADA

Working Paper/Document de travail
2013-29

Volatility and Liquidity Costs

by Selma Chaker

www.bank-banque-canada.ca

Bank of Canada Working Paper 2013-29

August 2013

Volatility and Liquidity Costs

by

Selma Chaker

International Economic Analysis Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
schaker@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

I am grateful to Nour Meddahi for helpful discussions. I also thank Sílvia Gonçalves for feedback. I express my gratitude to Ilze Kalnina for comments and discussions about the paper's assumptions, and I thank Bruno Feunou, Tim Bollerslev, Andrew Patton, and all participants at the financial econometrics lunch group at Duke University, and Torben Andersen and all participants at the NBER-NSF time-series conference 2011. Many thanks to Jean-Sébastien Fontaine, Antonio Diez de los Rios, Gregory H. Bauer, Federico M. Bandi, Walid Chaker, Tatevik Sekhposyan and Sermin Gungor.

Abstract

Observed high-frequency prices are contaminated with liquidity costs or market microstructure noise. Using such data, we derive a new asset return variance estimator inspired by the market microstructure literature to explicitly model the noise and remove it from observed returns before estimating their variance. The returns adjusted for the estimated liquidity costs are either totally or partially free from noise. If the liquidity costs are fully removed, the sum of squared high-frequency returns – which would be inconsistent for return variance when based on observed returns – becomes a consistent variance estimator when based on adjusted returns. This novel estimator achieves the maximum possible rate of convergence. However, if the liquidity costs are only partially removed, the residual noise is smaller and closer to an exogenous white noise than the original noise. Therefore, any volatility estimator that is robust to noise relies on weaker noise assumptions if it is based on adjusted returns than if it is based on observed returns.

JEL classification: G20, C14, C51, C58
Bank classification: Econometric and statistical methods; Financial markets; Market structure and pricing

Résumé

Les prix des actifs observés à haute fréquence sont « contaminés » par des coûts de liquidité ou du bruit en raison de la présence d'effets de microstructure. S'inspirant de la littérature qui étudie la microstructure des marchés, l'auteure met au point un nouvel estimateur qui permet de modéliser explicitement le bruit à partir de ces données et de l'éliminer des rendements observés de l'actif avant d'estimer leur variance. Les rendements corrigés des coûts de liquidité estimés sont totalement ou partiellement exempts de bruit. Dans le cas où les coûts de liquidité sont entièrement retranchés, la somme des carrés des rendements à haute fréquence devient un estimateur convergent de la variance si celle-ci est calculée sur la base des rendements corrigés et non sur celle des rendements observés. Ce nouvel estimateur converge à une vitesse maximale. Toutefois, lorsqu'une partie seulement des coûts de liquidité est éliminée, le bruit résiduel est plus petit et plus proche d'un bruit blanc exogène que le bruit initial. En conséquence, un estimateur robuste de la volatilité n'exige pas d'hypothèses aussi fortes sur le bruit s'il est fondé sur les rendements corrigés plutôt que sur les rendements observés.

Classification JEL : G20, C14, C51, C58
Classification de la Banque : Méthodes économétriques et statistiques; Marchés financiers; Structure de marché et fixation des prix

1. INTRODUCTION

The advent of large intraday financial data – with a second or millisecond time stamp – has created new opportunities to measure asset return volatility-type objects that are important inputs in asset pricing, portfolio allocation and financial risk management. However, at high frequencies, observed prices are contaminated with market microstructure frictions. Demsetz (1968) and Stoll (2000) measure these frictions by the price concession paid for immediacy, referred to as liquidity costs. More recently, Aït-Sahalia and Yu (2009) relate statistical measures of the frictions to financial measures of the stock liquidity. The liquidity costs create a discrepancy between the frictionless-price process and the observed prices, resulting in the inconsistency of the realized variance – defined as the sum of the squared returns sampled at high frequency – for the return variance.

To measure volatility, the financial econometrics literature models the liquidity costs as a measurement error or noise. The problem of noise was first addressed by discarding data (Andersen *et al.* 2003; Bandi and Russell 2008). More recently, robust-to-noise volatility estimators using all the available high-frequency price data were derived (see Zhang, Mykland and Aït-Sahalia 2005 for the two time-scales estimator[1]; Barndorff-Nielsen *et al.* 2008 for the realized kernel estimator; Jacod *et al.* 2009 for the pre-averaging estimator). These robust-to-noise volatility estimators, which Diebold and Strasser (2013) describe as statistical estimators, treat noise in a fully nonparametric manner. As a consequence, the econometrician can never get rid of the measurement error. More importantly, such an approach generates rate optimal estimators that cannot beat the convergence rate achieved by the realized variance.

In this paper, we demonstrate that modelling the liquidity costs as in the market microstructure literature is a better solution, even if one misspecifies the liquidity-costs model. Specifically, we show that the realized variance based on returns adjusted for liquidity costs becomes a consistent estimator of variance if the liquidity costs are fully removed. In that case, the optimal efficiency bound for volatility estimation is reached. When the model is misspecified and the liquidity costs are partially removed, the uncaptured liquidity costs are smaller and closer to an exogenous white noise than the original liquidity costs. This results in more realistic robust-to-noise volatility estimators because they rely on less-strong assumptions.

Using simulated data, we find that the new volatility estimator outperforms the benchmark by comparing the finite-sample simulation results with those predicted by the asymptotic theory. We use the pre-averaging estimator from the statistical approach as a benchmark because it achieves the optimal rate among the robust-to-noise estimators, and also allows for non i.i.d. noise. Using real data covering 2009–2010 for Alcoa stock and performing a daily analysis, the noise is completely removed for about half of the business days of the sample. For these days, the realized variance based on adjusting high-frequency returns for liquidity costs is an error-free estimator of the daily integrated variance with the maximum possible accuracy. The noise-to-signal ratio is considerably reduced when observed returns are adjusted for liquidity costs even when the noise is partially removed.

[1]The two time-scales estimator is the first consistent estimator of volatility in the presence of noise. It is related to the work of Zhou (1996).

Our approach uses insights from the literature on market microstructure, and has two main advantages. First, by explicitly specifying the noise, this approach makes full use of the data available, including bid-ask spread and volume series, as opposed to the statistical approach where only price series are exploited. We use Roll (1984) and Glosten and Harris (1988) models precisely, to measure the liquidity costs. In the former model, a trade-direction indicator component of the trading costs captures the fixed costs of trading. In the latter model, a trading volume component of the trading costs captures the size-varying costs of providing liquidity service.

The second advantage of using this approach is that even when the liquidity costs cannot be fully removed, the residual noise – measuring the misspecification of the noise model – is less problematic than the original noise. The idea is that the explanatory variables included in the liquidity costs capture the undesirable features of the noise, namely the endogeneity with the frictionless price, the autocorrelation and the heteroskedasticity. As a result, the uncaptured liquidity costs are more likely to be free from these undesirable features, and closer to an exogenous white noise than the original noise.

The main undesirable feature of the liquidity costs is the return-noise endogeneity, which this model captures by explicitly specifying the liquidity costs driving variables. For instance, we use the trading volume as an explanatory variable of the noise, which results in nonzero return-noise correlation. Indeed, in Glosten and Harris (1988) the trading volume explains not only the liquidity costs but also drives the asymmetric information component of the efficient price. Easley and O'Hara (1987), Kyle (1985), and Glosten (1989) have theoretical models that suggest this component should increase with the quantity traded because well-informed traders maximize the returns to their perishing information. Finally, we formally test whether the explanatory variables of the liquidity costs capture the return-noise endogeneity using a Hausman specification test.

The endogeneity treatment in this paper departs from the literature. Although the noise endogeneity could be accommodated in many robust-to-noise volatility estimators such as the realized kernel and the pre-averaging estimators, it would rely on the specific parametric form of endogeneity. Alternative specifications of the endogenous noise are proposed in Barndorff-Nielsen *et al.* (2008), Kalnina and Linton (2008), and Nolte and Voev (2012), among others. However, within the statistical approach, the independence between the noise and the frictionless price is frequently assumed. In our setting, the driving variables of the liquidity costs capture the return-noise endogeneity. The only attempt in the literature that we are aware of to address the endogeneity problem using insights from market microstructute theory is by Diebold and Strasser (2013), who derive the return-noise correlation in several structural models. Our approach to capture return-noise endogeneity differs from theirs, since we model the noise term and the return-noise correlation is a by-product of the analysis. Also, compared to Diebold and Strasser (2013), we do not restrict the price volatility to be constant as they do, and we exploit quantity data and not only price data as they do.

The driving variables of the liquidity costs also capture other undesirable features: the autocorrelation and heteroskedasticity of the noise. For example, the trading volume is highly persistent because of the clustering of small-size trades. Moreover, the trading volume is heteroskedastic as a result of its U-shaped intraday pattern. Admati and Pfleiderer (1988) develop a model in which the empirical concentrated-trading patterns in the beginning and the end of

the trading day are theoretically generated. Within the statistical approach to asset return volatility, most robust-to-noise volatility estimators have richer versions that allow autocorrelation of the noise. However, not all estimators – such as the two time-scales estimator – allow for the heteroskedasticity of the noise, as is the case for the pre-averaging estimator.

The semiparametric approach used in this paper offers two main theoretical results. To measure return variance, we estimate the parameters of the liquidity costs using a price-impact regression and instrumental variables to insure against return-noise endogeneity. First, we derive the asymptotic distribution of the realized variance based on adjusted returns for the case where the liquidity costs are fully removed. Second, we derive the asymptotic distribution for the pre-averaging estimator based on adjusted returns for the case where the liquidity costs are partially removed.

The rest of this paper is organized as follows. Section 2 describes the model for market microstructure noise based on liquidity costs. In section 3, we discuss the estimation of this model and describe a test for the performance of the liquidity costs measure. In section 4, we study volatility estimation based on adjusting prices for the liquidity measure introduced in section 2. Section 5 describes a simulation exercise. Section 6 is an empirical application where we compare the estimation accuracy of the volatility estimator in this paper to the pre-averaging estimator. In section 7, we offer several conclusions.

2. THE MODEL

We introduce the liquidity costs in the context of a model that is consistent with both the standard additive price model of the high-frequency financial econometrics and several transaction-cost models from the market microstructure literature.
The standard additive model of the high-frequency financial econometrics literature is given by

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

where p_t is the observed log price, p_t^* is the log of the frictionless price and ε_t is a measurement error term summarizing the market microstructure noise generated by the trading process. The fixed interval $[0,1]$ is a day, for example. In this context, the observed price is the sum of two unobservable components, which are the frictionless price and the noise. The frictionless price p_t^* – also referred to as the true price, the efficient price or the equilibrium price – is the log of the expectation of the final value of the asset conditional on all publicly available information at time t. In a perfect market, with no trading frictions, the log-price would be p_t^*.
Within the market microstructure literature, Stoll (2000) studies various sources of noise or trading frictions. The presence of a bid-ask spread and the corresponding bounces is one source of noise. Roll (1984) provides a measure of the effective bid-ask spread based on the negative serial dependence in successive observed returns induced by trading costs. Glosten and Harris (1988) extend Roll's model by adding a trading volume component to capture the costs of providing

liquidity service. This model is nested in (1) and is given by

$$p_t = p_t^* + \underbrace{\underbrace{\underbrace{\beta_1}_{fixed\ transaction\ costs}\ q_t + \underbrace{\beta_2}_{transaction\ costs\ per\ share}\ q_t v_t}_{noise}}, \tag{2}$$

where q_t is the trade-direction indicator, which takes the value +1 if the trade is buyer-initiated and -1 if the trade is seller-initiated. For $\beta_2 = 0$, the Glosten and Harris (1988) model is reduced to the Roll (1984) model where the bid-ask spread is considered as constant.

In this paper, we extend the Glosten and Harris (1988) linear model by adding other explanatory variables in the noise. For example, we add the ask (bid) depth that specifies the maximum quantity for which the ask (bid) price applies. In Kavajecz (1999), the depths are used to capture inventory-control costs as well as asymmetric-information costs. In the market, a larger quoted depth is interpreted as an increase in liquidity. A generalized model of (2) is given by

$$p_t = p_t^* + F_t^{'}\beta, \tag{3}$$

where F is an M-vector of liquidity-cost variables. If β is known, the frictionless price p_t^* would be equal to $p_t - F_t^{'}\beta$, and would be treated as observable. However, β has to be estimated from the data.

The linear form $F_t^{'}\beta$ could be misspecified in the sense that it does not capture the entire noise ε_t. The model of this paper accounts for the misspecification of the noise term $F_t^{'}\beta$ in (3) in the following way:

$$p_t = p_t^* + \underbrace{F_t^{'}\beta + \xi_t}_{=\varepsilon_t}. \tag{4}$$

The residual noise ξ_t captures all the trading frictions that are misspecified by the $F_t^{'}\beta$ form. The magnitude of ξ_t could also be seen as a measure of the performance of the liquidity costs $F_t^{'}\beta$. If ξ_t is small, then $F_t^{'}\beta$ is a good measure of liquidity costs.

To present the model in discrete time, we introduce the following notation. We dispose of N equidistant observations at $i = 0, 1, .., N$ over [0,1]. For simplicity of notation, an intraday variable Y_i stands for $Y_{i/N}$. We 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. Using the model (4), the high-frequency returns are written as

$$r_i = r_i^* + \underbrace{X_i^{'}\beta + \Delta\xi_i}_{=\Delta\varepsilon_i}. \tag{5}$$

Next, we turn to the assumptions underlying the frictionless price and the liquidity costs. We make the standard arbitrage-free semimartingale assumption for the frictionless price. The one-dimensional price process, which is evolving in continuous time over the fixed interval [0,1], is defined on a complete probability space ($\mho$, $\mathscr{F}$, $\mathbf{P}$). We consider an information filtration, the increasing family of σ-fields $(\mathscr{F}_t)_{t\in[0,1]} \subseteq \mathscr{F}$, which satisfies the usual conditions of $\mathbf{P}$-completeness and right continuity. The prices and noise explanatory variables are included in the information

set $\mathscr{F}_t$.

Assumption 1
The frictionless price p^* follows the dynamics

$$dp_t^* = \mu_t dt + \sigma_t dW_t, \tag{6}$$

where W_t is standard Brownian motion and σ_t is a *càdlàg* volatility function, which is independent from the frictionless price (no leverage).

Assumption 1 imposes that the frictionless-return process is the sum of an adapted process of finite variation and a stochastic volatility martingale.[2] The spot volatility σ_t can exhibit non-stationarity, diurnal effects and jumps. In the high-frequency context, the drift component of (6) is of order dt, whereas the diffusion component is smaller and of order $\sqrt{dt}$. Therefore, the frictionless return is of order $\sqrt{dt}$ or, equivalently, $\mathcal{O}(1/\sqrt{N})$ using the discrete-time notation.
The object of interest in this paper is the integrated variance defined as

$$IV = \int_0^1 \sigma_u^2 du. \tag{7}$$

We make the following set of assumptions for the different components of the noise ε_t in (4). Basically, the first component of the noise $F_t^{'}\beta$ is endogenous with the frictionless return, autocorrelated and heteroskedastic whether the second component of the noise ξ_t is exogenous and identically and independently distributed (i.i.d.).

Assumption 2
(i) F_t and p_t^* are dependent.
(ii) The increments of F_t are $\mathcal{O}(1)$ and $E[F_t] = 0$.

Assumption 3
(i) ξ_t is independent from p_t^* and F_t.
(ii) ξ_t is i.i.d. and $E[\xi_t] = 0$.

Assumption 2(i) refers to the endogeneity between the liquidity-cost variables and the frictionless price. Indeed, the return-noise endogeneity is empirically evidenced and theoretically modelled. In Hansen and Lunde (2006), an empirical analysis of the Dow Jones Industrial Average stocks reveals that the noise is correlated with increments in the frictionless price. For the structural models, Diebold and Strasser (2013) derive closed-form expressions of the return-noise correlation in a variety of stylized structural models. In order to validate Assumption 2(i), we provide a Hausman specification test in section 3.3 to check whether the noise variables in F_t do capture the endogeneity between the noise ε_t and the frictionless price p_t^*.

[2] Adding a jump component to the dynamics of the frictionless price is beyond the scope of this paper. See Andersen *et al.* (2007) for an analysis of the importance of the jump component for volatility estimation and forecasting.

Assumption 2(ii) concerns the stochastic magnitude of the noise variation. The order of the increment of F_t is assumed to be $\mathcal{O}(1)$; that is, its variance does not vanish when the sample size N grows. $\mathcal{O}(1)$ is a fundamental identifying assumption; any noise explanatory variable candidate must be $\mathcal{O}(1)$. At high frequencies, the frictionless-return magnitude vanishes as a result of the semimartingale condition in Assumption 1. However, the noise component does not vanish at high frequencies. The dominance of the noise translates into the explosion of the realized variance (or the sum of squared returns) at high frequencies. The signature plot is an empirical evidence of such explosion. This plot was proposed by Andersen *et al.* (2000) and it draws the average of daily realized variances across the sampling frequency of the underlying returns. On the other hand, Awartani, Corradi and Distaso (2009) formally test that the variance of the noise is independent of the sampling frequency. This test could be seen as a test of Assumption 2(ii).

In Assumption 3, we suppose that if any noise remains after adjusting the returns, then that noise is an exogenous white noise. We argue that the liquidity-cost variables F_t should capture the undesirable features of the noise: endogeneity with the frictionless price, autocorrelation and heteroskedasticity. In the literature, the exogenous white noise assumption is made for the entire noise ε_t. This simplifying assumption is considered by Bandi and Russell (2006a, 2008), Aït-Sahalia *et al.* (2005), and Zhang *et al.* (2005).

If the frictionless return was observed, then the realized variance $\sum_{i=1}^{N} r_i^{*2}$ would be a consistent estimator of IV. However, since only noise-contaminated returns are observed, the realized variance $\sum_{i=1}^{N} r_i^2$ is inconsistent for IV. The idea of this paper is to first adjust the observed high-frequency returns for the estimated liquidity-costs $X_i^{'}\widehat{\beta}$, where $\widehat{\beta}$ is a consistent estimator of β. Second, estimate IV using the adjusted returns $r_i - X_i^{'}\widehat{\beta}$. Improved volatility estimation is due to the fact that the adjusted returns are closer to r^* and are more likely to conform to the assumptions that justify the use of model-free volatility estimators than observed returns.

3. LIQUIDITY-COST ESTIMATION

In this section, we estimate the liquidity costs. We show the consistency and the asymptotic normality of the liquidity-cost parameter estimates. To check whether the proposed liquidity-cost model is misspecified, we derive a formal econometric test. If the model is misspecified, a residual noise term should be accounted for. We also provide a test for the endogeneity between the liquidity-cost explanatory variables and the frictionless return. Indeed, if there is evidence that the estimated noise is endogenous with the frictionless return, then the residual noise is more likely to be exogenous.

The idea of the estimation is to write the price-impact regression in (5) such that all the latent variables, including the frictionless return, are in the regression's residual:

$$\underbrace{r_i}_{regressand} = \underbrace{X_i^{'}}_{regressors} \beta + \underbrace{r_i^* + \Delta\xi_i}_{residual} \quad ; \quad i = 1, ..., N. \tag{8}$$

In matrix notation, the regression is written as

$$r = X\beta + r^* + \Delta\xi, \tag{9}$$

where

$$r = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix} \quad , \quad X = \begin{pmatrix} X_1^{(1)} & \dots & X_1^{(M)} \\ \vdots & \vdots & \vdots \\ X_N^{(1)} & \dots & X_N^{(M)} \end{pmatrix} \quad , \quad \Delta\xi = \begin{pmatrix} \Delta\xi_1 \\ \vdots \\ \Delta\xi_N \end{pmatrix}. \tag{10}$$

The regression (8) cannot be estimated using ordinary least squares (OLS) if Assumption 2(i) holds. Under this assumption, the regressors are endogenous with the regression's residual. Therefore, instrumental variables are needed to consistently estimate the parameter β of (8).

3.1. *Asymptotic theory*

In this subsection, we show the consistency and the asymptotic normality of the instrumental variable estimator of β. Such results are necessary to derive the asymptotic properties of the new volatility estimator in the next section.

Let $\widehat{\beta}$ be the instrumental variable estimator of β defined by

$$\widehat{\beta} = (Z^{'}X)^{-1}Z^{'}r. \tag{11}$$

The instrument Z is the lag of the regressor X, $Z_i = X_{i-1}$. We dispose of as many instruments as regressors, usually referred to as the exactly identified case. Z is a valid instrument because it satisfies two conditions. The first condition is $E[Z_i\Delta\varepsilon_i] = 0$. Formally,

$$\begin{aligned} &E[Z_i\Delta\varepsilon_i] = E[Z_i(r_i^* + \Delta\xi_i)] = E[Z_i r_i^*] = E[X_{i-1}r_i^*] \\ &= E[X_{i-1}\int_{i-1}^{i}\mu_s ds] + E[X_{i-1}\int_{i-1}^{i}\sigma_s dW_s] \\ &= E[X_{i-1}]E[\int_{i-1}^{i}\mu_s ds] + E[X_{i-1}E[\int_{i-1}^{i}\sigma_s dW_s|\mathscr{F}_{i-1}]] \\ &= E[X_{i-1}]E[\int_{i-1}^{i}\mu_s ds] + E[X_{i-1}]E[\int_{i-1}^{i}\sigma_s dW_s] \\ &= 0. \end{aligned}$$

This result holds because Assumption 2(ii) implies $E[X_{i-1}] = 0$ and Assumption 1 implies $E[\int_{i-1}^{i}\sigma_s dW_s|\mathscr{F}_{i-1}] = E[\int_{i-1}^{i}\sigma_s dW_s]$.

The second condition to have a valid instrument is that $Z'X$ is nonsingular. This is a consequence of the persistence of the liquidity-cost variables and could be tested empirically. For example, the trading volume is a persistent variable because the small-size trades tend to be clustered.

To derive the asymptotic distribution of $\widehat{\beta}$ defined in (11), we make the following set of assumptions.

Assumption A

(i) $E[(Z_t - Z_{t-h})(F_t^{'} - F_{t-h}^{'})] = \Omega^{(t,h)}$, 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) $\frac{1}{N}\sum_{i=1}^{N}Z_i X_i^{'} \xrightarrow{P} \Omega$.

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

Assumption C $\frac{1}{N}\sum_{i=1}^{N}\sum_{\nu=-1}^{1} E\left[\Delta\xi_i \Delta\xi_{i-\nu} Z_i Z_{i-\nu}'\right] \xrightarrow{P} S.$

Assumption A concerns the regressors in (9), whereas Assumptions B and C are related to the residual of the price-impact regression. For the residual $r^* + \Delta\xi$, two cases are possible. First, the liquidity-cost explanatory variables capture all the noise ε and the remaining noise ξ is zero. Second, the liquidity-costs model $X\beta$ is misspecified and the residual noise ξ is nonzero. In the former, the residual of regression (9) is exactly the frictionless return. The heteroskedasticity of the frictionless return r^* under stochastic volatility will impact the asymptotic distribution of the price-impact regression parameters. Assumption B is useful in that case. In the latter case, the dominating regression residual term is $\Delta\xi$ because r^* is negligible. Therefore, Assumption C is necessary for an MA(1) type process; i.e., $\Delta\xi$ as residual.
We next derive the asymptotic theory for the estimator of the liquidity-cost parameters. All proofs are given in Appendix B. Convergence in probability is denoted by $\xrightarrow{P}$ or $plim(.)$, whereas convergence in law is denoted by $\xrightarrow{L}$. For mixed normal-limit distributions, we denote the stable[3] convergence as $\xrightarrow{st}$. The following proposition concerns the case where $Var[\xi] = 0$.

Proposition 1 *Under Assumptions 1-2, A, B, and* $Var[\xi] = 0$*,*

$$(i)\ \widehat{\beta} \xrightarrow{P} \beta.$$
$$(ii) N(\widehat{\beta} - \beta) \xrightarrow{L} \mathcal{N}\left((0)_{M\times 1}, V_0(\widehat{\beta})\right),$$

where $V_0(\widehat{\beta}) = \Omega^{-1}\Omega^*\Omega^{-1}$.

Consistency is then obtained with a faster rate of convergence than the usual $\sqrt{N}$. Recall the regression in that case,

$$r = X\beta + r^*. \tag{12}$$

Notice that the residual is the frictionless return, which is very small at high frequencies. On the other hand, the noise $X\beta$ is relatively big. Therefore, the regression performs well and $\widehat{\beta}$ is supra-convergent. In Stock (1987), the supra-convergence rate is obtained in a similar setting. Next, we turn to the case where $Var[\xi] \neq 0$.

Proposition 2 *Under Assumptions 1-3, A, C, and* $Var[\xi] \neq 0$*,*

$$(i)\ \widehat{\beta} \xrightarrow{P} \beta,$$
$$(ii) \sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{L} \mathcal{N}\left((0)_{M\times 1}, V_1(\widehat{\beta})\right),$$

where $V_1(\widehat{\beta}) = \Omega^{-1} S \Omega^{-1}$.

We obtain the usual $\sqrt{N}$ rate of convergence because the regression residual $\Delta\xi$ is $\mathcal{O}(1)$. The frictionless-return moments do not appear in the asymptotic variance of $\widehat{\beta}$. Indeed, the stochastic

[3]The stable convergence concept is discussed in Aldous and Eagleson (1978).

magnitude of the frictionless return is negligible compared to $\Delta\xi$.
Once β is consistently estimated, $F^{'}\widehat{\beta}$ is the liquidity-costs measure proposed in this paper. Since the variable of interest is the volatility of the frictionless return, then subtracting the liquidity-costs measure from the observed returns would decontaminate the latter from noise. Let the adjusted price and the adjusted return $\widehat{r}$ be defined, respectively, as

$$\widehat{p}_i = p_i - F_i^{'}\widehat{\beta}, \tag{13}$$

$$\widehat{r}_i = r_i - X_i^{'}\widehat{\beta}. \tag{14}$$

If $Var[\xi] = 0$, we have, using Proposition 1,

$$\widehat{r}_i = r_i^* + X_i^{'}\underbrace{(\beta - \widehat{\beta})}_{\mathcal{O}(1/N)}. \tag{15}$$

The frictionless returns are, then, the dominant term in the adjusted-return expression. However, if $Var[\xi] \neq 0$, Proposition 2 applies and

$$\widehat{r}_i = r_i^* + X_i^{'}\underbrace{(\beta - \widehat{\beta})}_{\mathcal{O}(1/\sqrt{N})} + \Delta\xi_i. \tag{16}$$

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

3.2. *Testing misspecification*

It is perhaps too strong an assumption that a few explanatory variables such as the trade-direction indicator and the signed volume can fully absorb all the noise. Therefore, we 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 no longer represent the frictionless returns and the model in (3) is misspecified. We formally test in this subsection whether the adjusted returns still have a noise component. From a market microstructure perspective, this test may be interpreted as a test for the quality of the trading-costs measure $F^{'}\widehat{\beta}$. If this is a good measure of the noise[4] ε, then the residual noise ξ should go to zero. Otherwise, the trading-costs measure does not capture all the real frictions and the term ξ is nonzero.
The null hypothesis H_0 and the alternative hypothesis H_1 are, respectively,

$$\begin{aligned} H_0 &: Var[\xi] = 0, \\ H_1 &: Var[\xi] \neq 0. \end{aligned} \tag{17}$$

The idea of the test is that the presence of the noise usually causes negative serial correlation in observed high-frequency returns. However, the argument has to apply to the adjusted returns

[4] In Bandi and Russell (2006b), the noise measure is considered as a measure of the market quality.

because the test concerns the residual noise ξ and not the original noise ε. If $Var[\xi] = 0$, the covariance between successive adjusted returns is asymptotically zero,

$$\begin{aligned} Cov[\widehat{r}_i, \widehat{r}_{i-1}] &= Cov[\underbrace{r_i^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X_i^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/N)}, \underbrace{r_{i-1}^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X_{i-1}^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/N)}] \\ &\approx Cov[r_i^*, r_{i-1}^*] \\ &= 0. \end{aligned} \tag{18}$$

If $Var[\xi] \neq 0$, the covariance between successive adjusted returns is asymptotically negative,

$$\begin{aligned} Cov[\widehat{r}_i, \widehat{r}_{i-1}] &= Cov[\underbrace{r_i^* + X_i^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{\Delta\xi_i}_{\mathcal{O}(1)}, \underbrace{r_{i-1}^* + X_{i-1}^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{\Delta\xi_{i-1}}_{\mathcal{O}(1)}] \\ &\approx Cov[\Delta\xi_i, \Delta\xi_{i-1}] \\ &= -Var[\xi] < 0. \end{aligned} \tag{19}$$

As shown above, the first-order serial covariance expression depends on whether the noise is completely absorbed (i.e., $Var[\xi] = 0$) or partially absorbed (i.e., $Var[\xi] \neq 0$). 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 nonzero serial covariance in the adjusted returns corresponds to the case where the noise is partially absorbed.
We denote by $RC(1)$ the realized autocovariance of order one for the adjusted returns

$$RC(1) = \left(\sum_{i=1}^{N} \widehat{r}_i \widehat{r}_{i-1} + \sum_{i=1}^{N} \widehat{r}_i \widehat{r}_{i+1} \right) / 2. \tag{20}$$

In the next proposition, we formally define the test statistic and give its asymptotic distribution.

Proposition 3 *Suppose Assumption 1, 2, A, and B hold. Under* H_0*,*

$$S_N \xrightarrow{d} \mathcal{N}(0,1), \tag{21}$$

where

$$S_N = \frac{\sqrt{N} RC(1)}{\sqrt{\widehat{IQ}}}, \tag{22}$$

and $\widehat{IQ}$ *is a consistent estimator of the integrated quarticity* $IQ = \int_0^1 \sigma_u^4 du$*.*

According to the proposition above, we reject H_0 at the confidence level α when $|S_N| > c_{1-\frac{\alpha}{2}}$, 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.

3.3. *Testing endogeneity*

In this section, we apply a Hausman specification test as in Hausman (1978) to formally test for the presence of endogenous liquidity costs. We define the null hypothesis as well as the alternative as

$$\begin{aligned} &H_0 : X \textit{ exogenous}, \\ &H_1 : X \textit{ endogenous}. \end{aligned} \tag{23}$$

The idea of the Hausman specification test is that under H_0, the OLS and the instrumental variable estimators of β are statistically not different. However, under the alternative H_0 the two estimators are statistically different.

Before providing the formal test statistics, we explain the source of the return-noise endogeneity. The liquidity-cost variables should capture the endogeneity between the noise and the frictionless price. For instance, in the asymmetric-information models of Glosten and Harris (1988) and Hasbrouck (1991), the trading volume captures the adverse selection in the efficient price (which is also the frictionless price). Therefore, having the volume as part of the frictionless price as well as the liquidity costs results in the endogeneity between these two components. Moreover, in Glosten and Harris (1988), the trade-direction indicator is also present in the efficient price as well as the liquidity costs (see Huang and Stoll 1997). The trade indicator is also a source of endogeneity of the noise. Diebold and Strasser (2013) examine several structural microstructure models and derive the correlation between the efficient return and the noise in each case. They show that, in some cases, the correlation depends on the bid-ask spread. Consequently, having the bid-ask spread in the noise could also capture the endogeneity between the noise and the frictionless price. In this paper, if $Var[\xi] = 0$, the covariance between the return-noise covariance is asymptotically equal to the covariance between the observable series: the adjusted return and the liquidity-costs measure. Formally, if $Var[\xi] = 0$,

$$\begin{aligned} &Cov[\widehat{r}_i, X_i^{'}\widehat{\beta}] = Cov[r_i - X_i^{'}\widehat{\beta}, X_i^{'}\widehat{\beta}] \\ &= Cov[\underbrace{r_i^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X_i^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/N)}, X_i^{'}\widehat{\beta}] \\ &\approx Cov[r_i^*, \Delta\varepsilon_i]. \end{aligned} \tag{24}$$

This result helps to provide evidence in the empirical section of this paper for the results of Diebold and Strasser (2013), who derive the sign and even bounds for the correlation between the frictionless price and the noise within several structural models.

Let $\widehat{\beta}_{OLS}$ be the OLS estimator of β defined by $\widehat{\beta}_{OLS} = (X^{'}X)^{-1}X^{'}r$. Similar to the Assumptions A, B and C, few technical assumptions are needed to derive the asymptotic distribution of the OLS estimator $\widehat{\beta}_{OLS}$, which is useful for the Hausman test.

Assumption $A^{'}$

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

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

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

Assumption $B^{'}$ $\quad \sum_{i=1}^{N} r_i^{*2} X_i X_i^{'} \xrightarrow{P} \Omega_X^*$.

Assumption $C^{'}$ $\frac{1}{N}\sum_{i=1}^{N}\sum_{\nu=-1}^{1} E\left[\Delta\xi_i \Delta\xi_{i-\nu} X_i X_{i-\nu}^{'}\right] \xrightarrow{P} S_X$.

Let the generalized inverse of a given matrix V be denoted by V^{-} . To derive the asymptotic distribution of the Hausman test statistics, it is important to distinguish between two cases. The

first case holds when all the noise is absorbed by the liquidity-cost measure, whereas the second case corresponds to the partially absorbed noise.

Proposition 4 *Under H_0:*
If $Var[\xi] = 0$, and Assumption 1, 2, A, A', B, and B' hold,

$$\widehat{SH}_0 \xrightarrow{P} \chi^2(d_0),$$

where

$$\widehat{SH}_0 = N^2 \left(\widehat{\beta} - \widehat{\beta}_{OLS}\right)' \left(V_0(\widehat{\beta}) - V_0(\widehat{\beta}_{OLS})\right)^{-} \left(\widehat{\beta} - \widehat{\beta}_{OLS}\right),$$

$d_0 = rank\left(V_0(\widehat{\beta}) - V_0(\widehat{\beta}_{OLS})\right)$ *and* $V_0(\widehat{\beta}_{OLS}) = \Omega_X^{-1}\Omega_X^*\Omega_X^{-1}$.
If $Var[\xi] \neq 0$, and Assumption 1-3, A, A', C, and C' hold,

$$\widehat{SH}_1 \xrightarrow{P} \chi^2(d_1),$$

where

$$\widehat{SH}_1 = N \left(\widehat{\beta} - \widehat{\beta}_{OLS}\right)' \left(V_1(\widehat{\beta}) - V_1(\widehat{\beta}_{OLS})\right)^{-} \left(\widehat{\beta} - \widehat{\beta}_{OLS}\right),$$

$d_1 = rank\left(V_1(\widehat{\beta}) - V_1(\widehat{\beta}_{OLS})\right)$ *and* $V_1(\widehat{\beta}_{OLS}) = \Omega_X^{-1} S_X \Omega_X^{-1}$.

From the proposition above, the Hausman test statistic differs in the two cases where $Var[\xi] = 0$ and $Var[\xi] \neq 0$. This difference is the result of the rate of convergence of the instrumental variable and the OLS estimators of β. If $Var[\xi] = 0$ this rate is $1/N$, whereas it is $1/\sqrt{N}$ in the case where $Var[\xi] \neq 0$. The asymptotic variances also differ in each case.

4. VOLATILITY ESTIMATION

Using the liquidity-costs measure derived in the previous section, we derive a novel volatility estimator in this section. The new volatility estimator is based on adjusting returns for liquidity costs. For the case where the liquidity costs are fully removed, the new estimator is the realized variance based on adjusted returns. In that case, the new estimator is a consistent volatility estimator with an optimal convergence rate. For the case where the liquidity costs are only partially absorbed, the new estimator is the pre-averaging estimator based on adjusted returns. Estimation improvement is due to relaxation of the noise underlying assumptions (endogeneity, autocorrelation and heteroskedasticity). To quantify the theoretical gain of the new estimator, we compare it to the pre-averaging estimator based on observed returns. First, we describe the pre-averaging estimator, which will serve as the benchmark. Second, we derive the asymptotic distribution of the new estimator.
To briefly summarize the idea of adjusting the high-frequency returns for liquidity costs, we compare the observed price p with the adjusted-price $\widehat{p}$, respectively, given by

$$\begin{aligned} p &= p^* + \underbrace{\varepsilon}_{noise} \\ &= p^* + \underbrace{F'\beta}_{endogenous\ noise} + \underbrace{\xi}_{exogenous\ noise}, \end{aligned} \tag{25}$$

and

$$\begin{aligned}\widehat{p} &= p - \underbrace{F^{'}\hat{\beta}}_{fitted\ noise} \\ &= p^{*} + \underbrace{F^{'}(\beta - \hat{\beta})}_{very\ small} + \underbrace{\xi}_{exogenous\ noise}\end{aligned} \quad . \tag{26}$$

If the liquidity costs are partially absorbed, the term ξ does not vanish and adjusting the returns transforms the original noise ε from endogenous, autocorrelated and heteroskedastic to a residual noise ξ that is exogenous, i.i.d. and of smaller magnitude. However, if the liquidity costs are fully removed and the term ξ vanishes, then the adjusted-price $\widehat{p}$ is asymptotically equal to the frictionless price p^{*}. In both cases, using the adjusted returns instead of observed returns improves volatility estimation, as shown in this section.

4.1. *The benchmark*

Among the existing nonparametric[5] noise-robust IV estimators, we choose the pre-averaging method of Jacod *et al.* (2009) as the benchmark, for reasons related to the precision performance as well as the noise properties. First, the authors show that the pre-averaging estimator converges to the integrated variance at the optimal rate in the presence of noise of $N^{1/4}$. Second, this approach consistently estimates the integrated quarticity that is needed in the asymptotic distribution of noise-robust volatility estimators. Third, the pre-averaging allows the market microstructure noise to be heteroskedastic. In fact, as shown in Kalnina and Linton (2008), the two time-scales estimator could be inconsistent for the integrated variance in the presence of heteroskedasticity in the noise. Finally, Hautsch and Podolskij (2013) extend the original pre-averaging method of Jacod *et al.* (2009) to allow for autocorrelated market microstructure noise.[6]

Let L_t be a given semimartingale contaminated with noise. The sum of the pre-averaged increments $[L,L]^{avg}$ is defined as

$$[L,L]^{avg} = \sum_{i=0}^{N-k} \left\{ \sum_{j=1}^{k} \phi\left(\frac{j}{k}\right) \Delta L_{i+j} \right\}^{2} , \tag{27}$$

where $\Delta L_j = L_j - L_{j-1}$, $\frac{k}{\sqrt{N}} = \theta + \mathcal{O}(N^{-1/4})$ 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.

We compare the estimator of Hautsch and Podolskij (2013), who use original returns, to the Jacod *et al.* (2009) estimator using adjusted returns. We find that using adjusted returns in the pre-averaging estimator of Jacod *et al.* (2009) achieves consistency of the integrated volatility estimator even if there is endogeneity. The pre-averaging estimator of Jacod *et al.* (2009) or Hautsch and Podolskij (2013) using the original returns is inconsistent in the presence of endogeneity.

To describe our next result, some additional notation is required. In particular, let $(F_t)_{t\geq 0}$ be a stationary q-dependent sequence, $B(q) = E[\xi^2] + E[(F^{'}\beta)^2] + 2\sum_{m=1}^{q}\rho(m)$, where $\rho(m) =$

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

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

$cov(F_t^{'}\beta, F_{t+m}^{'}\beta)$. Let $\hat{B}(q)$ be a consistent estimator of $B(q)$. The pre-averaging estimator of Hautsch and Podolskij (2013) using original prices is defined as

$$IV^{pre}(p) = \frac{12}{\theta\sqrt{N}}[p,p]^{avg} - \frac{12}{\theta^2}\hat{B}(q), \tag{28}$$

where $[p,p]^{avg}$ is given by equation (27). The $IV^{pre}(p)$ volatility estimator has three tuning parameters: θ, k and the function $\phi(.)$, which are chosen according to some optimality criteria. In the next proposition, we give the asymptotic properties of the pre-averaging estimator defined in (28), which is based on observed prices.

Proposition 5 *Suppose that Assumptions 1-3 hold. In the case* $Var[\xi] \neq 0$,
(i) in the presence of endogeneity, $IV^{pre}(p)$ *is inconsistent;*
(ii) in the absence of endogeneity,

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

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

According to Proposition 5(ii), the pre-averaging estimator is consistent when there is no endogeneity at the usual $N^{1/4}$ rate of convergence, which is the optimal rate in the presence of $\mathcal{O}(1)$ noise. However, as shown in Proposition 5(i), in the presence of endogeneity, the pre-averaging estimator based on original prices is inconsistent.

4.2. *The novel IV estimator*

In this subsection, we derive the asymptotic distribution of the new variance estimator. We define the new return variance estimator as follows. If the liquidity costs are completely removed, the new variance estimator is the realized variance based on high-frequency adjusted returns. Otherwise, if the liquidity costs are partially removed, the new variance estimator is the pre-averaging estimator computed using the adjusted returns instead of the observed returns. In the first case, we show that the new estimator is consistent for the return variance with the best possible rate of convergence. In the second case, the new estimator is robust to return-noise endogeneity, contrary to the pre-averaging estimator.
We denote by $RV(L) = \sum_{i=1}^{N}(\Delta L_i)^2$ the realized variation of a series L_i.

Theorem 1 *Under Assumptions 1, 2, A, B and* $Var[\xi] = 0$,

$$(i)\ RV(\widehat{p}) \xrightarrow{P} IV.$$
$$(ii)\ \sqrt{N}(RV(\widehat{p}) - IV) \xrightarrow{st} \mathcal{N}(0, 2\ IQ),$$

where $IQ = \int_0^1 \sigma_u^4 du$.

According to Theorem 1, if the liquidity-costs measure totally removes the noise, 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, an estimation error in $\hat{\beta}$ impacts neither the consistency nor the asymptotic

distribution of the estimator based on the adjusted returns, because this error is of a smaller order of magnitude ($\mathcal{O}(1/N)$). To compute confidence intervals[7] for the integrated volatility, a feasible estimator of the integrated quarticity is needed. We show in the proof of Theorem 1 in Appendix B that the sum of adjusted returns to the fourth power is a consistent estimator of the integrated quarticity under the assumptions of the theorem. Compared to the benchmark efficiency and underlying assumptions described in Proposition 5, the novel estimator $RV(\widehat{p})$ is robust to endogeneity and achieves the optimal rate of convergence as if there is no noise, and is written as

$$RV(\widehat{p}) = \sum_{i=1}^{N} \widehat{r}_i^2 = \sum_{i=1}^{N} \left(r_i - X_i' \widehat{\beta} \right)^2 .$$

Compared to the pre-averaging estimator, there are no tuning parameters involved in the expression of $RV(\widehat{p})$. This feature makes the new estimator easier to implement in practice than the pre-averaging estimator.

The rate of convergence of $\sqrt{N}$ obtained in Theorem 1(ii) for the estimator $RV(\widehat{p})$ is not achievable using any robust-to-noise volatility estimator. Indeed, Gloter and Jacod (2001) show that the rate of convergence of any robust-to-noise integrated volatility estimator is bounded by $N^{-1/4}$, where N is the sample size. The first consistent robust-to-noise volatility estimator of Zhang, Mykland and Aït-Sahalia (2005) achieves a convergence rate of $N^{-1/6}$. In fact, the $N^{-1/4}$ technical bound is already reached by the realized kernel estimator of Barndorff-Nielsen *et al.* (2008), as well as the pre-averaging estimator of Jacod *et al.* (2009).

Now, we treat the case where the noise is partially removed. The 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 be defined as

$$IV^{pre}(\widehat{p}) = \frac{12}{\theta\sqrt{N}} [\widehat{p}, \widehat{p}]^{avg} - \frac{6}{\theta^2 N} RV(\widehat{p}). \tag{29}$$

Theorem 2 *Suppose that Assumptions 1-3, A, C hold. In the case* $Var[\xi] \neq 0$,
(i) $IV^{pre}(\widehat{p}) \xrightarrow{P} IV + trace(\Omega_X \Omega^{-1} S \Omega^{-1})$.
(ii) $N^{1/4} \left(IV^{pre}(\widehat{p}) - trace(\widehat{\Omega}_X \widehat{\Omega}^{-1} \widehat{S} \widehat{\Omega}^{-1}) - IV \right) \xrightarrow{st} \mathcal{N}(0, \Gamma_\xi)$,
where $\Gamma_\xi = \frac{151}{140}\theta \ \widetilde{IQ} + \frac{12}{\theta} E[\xi^2] \widetilde{IV} + \frac{96}{\theta^3} E[\xi^2]^2$,
$\widetilde{IV} = plim \left(\sum_{i=1}^{N} \widetilde{r}_i^2 \right)$, $\widetilde{IQ} = plim \left(\frac{N}{3} \sum_{i=1}^{N} \widetilde{r}_i^4 \right)$, $\widetilde{r} = r^* + X'(\beta - \hat{\beta})$ *and* $\widehat{\Omega}_X$, $\widehat{\Omega}$, $\widehat{S}$ *are consistent estimators of* Ω_X, Ω *and* S, *respectively.*

Theorem 2(i) shows that the pre-averaging estimator based on adjusted prices is consistent even in the presence of endogeneity. This results from removing the estimated liquidity costs that are endogenous with the frictionless return. Observe that the asymptotic bias $trace(\Omega_X \Omega^{-1} S \Omega^{-1})$ is due to the fact that, based on their order, the frictionless returns and the estimation error of $\widehat{\beta}$ are asymptotically not distinguishable (see (16)). Theorem 2(ii) gives the asymptotic distribution of $IV^{pre}(\widehat{p})$. The rate of convergence is the same as the pre-averaging estimator based on observed returns.

[7] More accurate confidence intervals could be obtained using the bootstrap method, as in Gonçalves and Meddahi (2009).

Using Theorems 1 and 2, we define the new IV estimator by

$$\begin{aligned} IV^{new} &= RV(\widehat{p})\ if\ Var[\xi] = 0, \\ &= IV^{pre}(\widehat{p}) - trace(\widehat{\Omega}_X \widehat{\Omega}^{-1} \widehat{S} \widehat{\Omega}^{-1})\ if\ Var[\xi] \neq 0. \end{aligned} \tag{30}$$

Next, we provide a simulation exercise to examine the finite-sample properties of the noise parameters and the volatility estimators.

5. MONTE CARLO EVIDENCE

In this section, we show that the finite-sample simulation results are consistent with those predicted by the aforementioned asymptotic theory. We find that the misspecification and the endogeneity tests have a good performance. The new variance estimator is more accurate than the pre-averaging estimator benchmark.

We first describe the data-generating process for the frictionless price, the spot volatility and liquidity-cost variables. Second, we report the simulation results for the liquidity-cost estimation as well as the return variance estimation.

5.1. *The artificial data*

For the frictionless price, we use a two-factor affine stochastic volatility model, as in Andersen, Bollerslev and Meddahi (2011). Recall the frictionless-price dynamics,

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

We take a constant drift $\mu_t = \mu = 0.001$. The volatility model 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 $\kappa = 0.03$, $\theta = 0.001$ and $\sigma = 0.15$.

The vector of the noise explanatory variables is $F_t = (\ q_t \quad q_t v_t \quad q_t s_t \quad d_t^a \quad d_t^b\ )'$, which defines the trade-direction indicator, the signed volume, the signed spread, the ask depth and the bid depth, respectively. Monte Carlo experiments also require a data-generating process that provides an artificial trade indicator, trading volume, bid-ask spread and quoted depths whose time-series properties are consistent with those of the actual data. We follow Hasbrouck (1999) and generate artificial liquidity-cost variables by simulating a persistent process with an intraday U-effect.

The direction of the trade q_t is triggered by a Bernoulli process with clustering. Trades cluster since 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 that incorporates a dependence structure is given by Klotz (1972), in which he considers $q_1, q_2, ..., q_N$, 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

$$T(\alpha, \lambda) = \begin{pmatrix} \frac{1-2\alpha+\lambda\alpha}{1-\alpha} & \frac{(1-\lambda)\alpha}{1-\alpha} \\ 1-\lambda & \lambda \end{pmatrix}. \tag{31}$$

We use the parameters $\alpha = 0.55$ and $\lambda = 0.7$ to simulate the trade-direction sequence.

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

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

where ϵ^v follows a Normal distribution $\mathcal{N}(0, 0.01)$ and $\phi^v = 0.0005$. 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). We calibrate the parameters as $k_1 = 6$, $k_2^{open} = 0.5$, $k_3^{open} = 2.5$, $k_2^{close} = 0.2$ and $k_3^{close} = 3.5$.

To simulate the spread series, we follow Hasbrouck's (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

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

and 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). We calibrate the parameters as $z_1 = 0.5$, $z_2^{open} = 0.4$, $z_3^{open} = 1.5$, $z_2^{close} = 0.1$ and $z_3^{close} = 2.5$. The innovations ϵ^{c^a} and ϵ^{c^b} are independently distributed as $\mathcal{N}(0, 0.0005)$, $\phi^c = 0.001$. The tick size or minimum price variation is 0.01\$. The NYSE tick size changed from 1/16\$ to 0.01\$ on 29 January 2001. Technological innovation is indeed propelling the move in financial markets away from fractional trading and toward decimal trading.

We generate the quoted depths series using the following AR dynamics:

$$\begin{aligned}
d_i^a &= \mu^{dASK} + \phi^d(d_{i-1}^a - \mu^d) + \epsilon_i^{d^a}, \\
d_i^b &= \mu^{dBID} + \phi^d(d_{i-1}^b - \mu^d) + \epsilon_i^{d^b},
\end{aligned}$$

where ϵ^{d^a} and ϵ^{d^b} are independently distributed as $\mathcal{N}(0, 0.05)$, and $\mu^{dBID} = 5$. $\mu^{dASK} = 5.6$, $\phi^d = 0.4$.

The true parameter β is fixed as

$$\beta = (\ 8\ 10^{-4} \quad -5\ 10^{-5} \quad -0.03 \quad 5\ 10^{-5} \quad -4\ 10^{-5}\ )'.$$

We add a white noise ξ for a randomly chosen half of the intraday prices. More precisely, we take $\xi \sim \mathcal{N}(0, 8\ 10^{-8})$. We model endogeneity as in Barndorff-Nielsen *et al.* (2008) by adding the component $[0, -0.5, -0.5, -0.5, -0.5]\ r_i^*$ to the previous noise explanatory variables F_t.

5.2. *Results*

The results of the simulation show that the price-impact regression parameters – β – are estimated very accurately. Compared to the true data-generating process, both the misspecification test and the return-noise endogeneity test have good performance. Using artificial data, we find that the performance of the new volatility estimator is better than the benchmark as measured by the bias, the variance and the root mean squared error (RMSE).

We run $100,000$ replications or days. For each day, a trade occurs every 5 seconds. A business day has 6.5 working hours. For the simulation results, we report in Table 1 the bias, the relative bias (i.e., the bias in percentage terms), variance and RMSE of the interest variables for the model.

The coefficients of the liquidity-cost variables are estimated with a small RMSE ranging from $1.03\ 10^{-4}$ to $4.3\ 10^{-3}$. The misspecification test described in section 3.2 achieves an efficiency rate of 0.97% compared to the true model. For the endogeneity analysis, the Hausman test described in section 3.3 is rejected for all the days.

We compare seven volatility measures: the realized variance using high-frequency returns ($RV(p)$); the realized variance using high-frequency adjusted returns ($RV(\widehat{p})$, where the adjusted-price $\widehat{p}$ is defined in (13)); the realized variance using 40-ticks low-frequency returns ($RV^{low}(p)$); the realized variance using 40-ticks low-frequency adjusted returns ($RV^{low}(\widehat{p})$); the pre-averaging estimator based on original prices defined in (28) ($IV^{pre}(p)$); the pre-averaging estimator based on adjusted prices defined in (29) ($IV^{pre}(\widehat{p})$); and the new variance estimator defined in (30) (IV^{new}). The bias of the pre-averaging estimator in absolute values is about ten times the absolute value of the bias of the new variance estimator. This bias is due to the inconsistency of the pre-averaging estimator for the integrated variance in the presence of return-noise endogeneity. The IV^{new} has the best performance in terms of bias, variance and RMSE, as asserted by this paper's asymptotic theory. Table 1 also indicates that using the adjusted prices instead of the original prices improves the performance of the realized variance using all the data, the realized variance based on low-frequency returns and the pre-averaging estimator. This result shows that adjusting the observed returns for liquidity costs improves the traditional measures of integrated variance.

6. EMPIRICAL ANALYSIS

In this section, we test with data the performance of the model presented in section 2 as well as the performance of the new volatility estimator derived in section 4. We use Alcoa stock, listed on the NYSE. The data cover the 2009–10 period. We use five explanatory variables to capture the liquidity costs: the inferred trade-direction indicator, the trading volume, the bid-ask spread, the bid depth and the ask depth.

We find that the liquidity costs are fully removed for about half of the sample business days. For such days, the realized variance estimator based on high-frequency adjusted returns is then an error-free integrated variance estimator with optimal efficiency.

The first subsection describes the liquidity-cost estimation, and the second deals with the daily integrated variance estimation.

6.1. *Liquidity-cost estimation*

We find that the noise parameters are significant for most of the sample days except the bid-ask spread coefficient. The results of the misspecification test are that almost half the sample does not reject the linear noise model. Finally, for the return-noise endogeneity test, we find that the estimated liquidity costs are endogenous for the whole sample.

We follow the same steps as in section 3. First, we check that the liquidity-cost variables are valid candidates as noise explanatory variables. Second, we estimate the parameters of the liquidity-costs model. Third, we test for misspecification and return-noise endogeneity.

All the liquidity-cost variables that we use are observable except for the trade-direction indicator q_t. We infer the binary series q_t from observed trade and quote prices using the Lee and 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, $q_t = -1$. However, if the trade price occurs exactly at the midpoint of the bid-ask spread, then previous trades are used to determine the sign of a trade: if the trade price is higher than the previous trade price, then the trade is buyer-initiated, and vice versa. If the trade price did not change after the previous trade, the last price change should be considered instead. The trade classification requires that the trade series be matched with the quote series because in the TAQ database the two series are offered separately. We match trades and quotes by assuming a zero time lag because we use recent data. Appendix A details the data-manipulation procedure.

In Tables 2 and 3, we provide descriptive statistics to summarize the liquidity characteristics of the stock. In the year 2009, Alcoa stock was much more liquid than in 2010. The average number of transactions per day went from $4,347.2$ in 2009 to $2,806.6$ in 2010. Because of the financial crisis that started in 2008, the year 2009 is an example of abnormal times and excessive volatility regardless of whether the year 2010 is an example of a much less stressed period for financial markets. On average, there are almost as many buys as sells for each trading day. The quoted bid-ask spread is stable around one cent. The daily average size of the transactions in 2009 and 2010 is very close. However, the ask and bid depths are higher for 2010 compared to 2009.

The autocorrelation functions (ACF) of the five noise explanatory variables are plotted in Figure 1. Each plot displays the average autocorrelation across days. The first-order autocorrelation is the highest for all the variables, and the autocorrelation decays when the lag increases. However, even after 20 lags the autocorrelation does not vanish. The same pattern is observed for the estimated liquidity costs $F^{'}\widehat{\beta}$. Indeed, as shown in Figure 15, the autocorrelation of the fitted noise is about 45% at the first lag. At the 20^{th} lag, the autocorrelation of the fitted noise decreases to about 10%. Figure 2 draws the autocorrelation function of the variation of the five noise explanatory variables. It shows that, after the first lag, the autocorrelations of the noise variables increments vanish, which justifies the use of only the first lag of X as the instrumental variable to estimate β. Indeed, the higher-order lags of X are not highly correlated with X and cannot be valid instruments.

As stated earlier, the volatility signature plot of Andersen *et al.* (2000) draws the average of daily realized variances across the sampling frequency of the underlying returns. An explanatory variable is valid (i.e., $\mathcal{O}(1)$) if its quadratic variation explodes at high frequencies, as in Assumption 2(ii). The signature plot of Figure 3 illustrates the main problem of ultra-high-frequency data: the bias

due to noise contamination. After adjusting the price for the market microstructure effects, this problem is less severe, as shown in Figure 7. To formally show the bias explosion in the signature plot, suppose that ε is an exogenous white noise; then, the bias is given by

$$E\left[\sum_{i=1}^{N} r_i^2\right] - E[IV] = 2NE[\varepsilon^2]. \tag{32}$$

When N goes to infinity, the bias in (32) due to the noise also goes to infinity, which translates into the explosion of the realized variance in the signature plot.

Since q_t has a Bernoulli distribution, we know that the quadratic variation of q_t explodes at a high frequency. Figures 4, 5 and 6 use the signature plot as a visual tool to verify that the quadratic variation of the quoted bid-ask spread, the trading volume and the quoted depths explode at high frequencies. Therefore, these observables are valid noise explanatory variables.

We find that all the noise explanatory-variable coefficients except the bid-ask spread coefficient are significant at the 95% confidence level for almost all of the business days (Figures 8 to 12). The confidence intervals are computed using propositions 1 and 2. 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 a higher number of shares generates a lower cost per share. For the signed spread qs, the coefficient is mostly negative in 2009. 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 the presence of inventory-control costs. If the ask volume increases, the price rises in an attempt to elicit sales. The same is true for the bid volume.

The noise-to-signal ratio defined by $RV/2NIV^{new}$ is highly decreased if adjusted returns are used instead of original returns to compute the ratio. Figure 13 shows the time series of this ratio for 2009 and 2010.

For the misspecification test of section 3.2, we find that for 159 business days out of 252 for 2009, and 121 business days out of 252 for 2010, the test is not rejected, implying that the liquidity-cost measure absorbs all the noise in about half of the sample. Figure 14 shows the first-order realized autocovariance of the observed returns and adjusted returns ($RC(1)$ defined in (20)). The stylized fact of the negativity of the first-order autocovariance of the high-frequency returns disappears, or at least becomes much less pronounced, by adjusting the returns for liquidity costs. The graph for 2010 shows that adjusting the returns using OLS, as in section 3.3, instead of using the instrumental variable, as in section 3.1, results in positive first-order realized autocovariance. This may be due to the fact that, by using OLS, the residual noise is not an exogenous white noise. If that were the case, the first-order realized autocovariance would be negative. However, using the instrumental variable to estimate the liquidity costs results in either zero or negative first-order realized autocovariance, which is consistent with an exogenous white residual noise, as shown in (18) and (19).

Finally, Figure 16 plots the correlation between the returns and the fitted noise $X\widehat{\beta}$ using observed returns and adjusted returns. We also plot the return-noise bound derived by Diebold and Strasser (2013) (see their proposition 4). The authors find that the return-noise correlation is between $-1/\sqrt{2}$ and 0 for a one-period model of market making. In Figure 16, the return-noise correlation computed using observed returns is positive, whereas the return-noise correlation based on adjusted

returns is mostly in the interval $[-1/\sqrt{2}, 0]$, which is consistent with the theoretical result of Diebold and Strasser (2013).

6.2. *Volatility estimation*

To estimate the daily integrated variance, we use the new estimator IV^{new} defined in (30), whose properties are derived in Theorems 1 and 2. We compare IV^{new} with the benchmark estimator introduced in subsection 4.1, and find that the new volatility estimator is more precise than the benchmark estimator in 59% of the sample days.

We estimate daily integrated volatility using the original prices and the adjusted prices. For the pre-averaging estimator of Hautsch and Podolskij (2013) introduced in Proposition 5, the estimator is not necessarily positive and the authors bound it from below by zero. We have done the same in this section. Details on the asymptotic variance estimators used to compute confidence intervals are given in Appendix C.

For the days where the misspecification test of section 3.2 is not rejected we find that, for 114 business days out of 159 for 2009, and 82 business days out of 121 for 2010, the confidence interval of IV^{pre} is larger than the confidence interval for IV^{new}. This important improvement in the accuracy of the new estimator compared to the benchmark is the result of the high rate of convergence of $RV(\widehat{p})$, which is equal to IV^{new} when the misspecification test is not rejected or $Var[\xi] = 0$, as shown in Theorem 1(ii). However, for the days where the misspecification test is rejected, we find that the accuracy improvement of IV^{new} over IV^{pre} is less important. Indeed for these days, only 42 business days out of 93 for 2009, and 59 business days out of 131 for 2010, the confidence interval of IV^{pre} is larger than the confidence interval for IV^{new}. The new estimator when the misspecification test is rejected is $IV^{pre}(\widehat{p})$. Theorem 2(ii) gives the asymptotic variance of $IV^{pre}(\widehat{p})$ and it is not clear whether its variance is smaller than the asymptotic variance of $IV^{pre}(p)$ derived in Proposition 5.

Figure 17 plots $\frac{IV^{new}-IV^{pre}(p)}{IV^{pre}(p)}$ and shows that the estimators IV^{new} and $IV^{pre}(p)$ are relatively different. On average, $\frac{|IV^{new}-IV^{pre}(p)|}{IV^{pre}(p)}$ is 16.65% for 2009 and 17.10% for 2010. The relative difference $\frac{IV^{new}-IV^{pre}(p)}{IV^{pre}(p)}$ jumps for few days of the sample but remains stable for most of the days.

7. CONCLUSION

In light of the market microstructure literature that provides economic drivers for market microstructure frictions or liquidity costs, we propose a semiparametric price model that exploits a much bigger set of available trade and quote data to estimate volatility. The resulting new volatility estimator is asymptotically more accurate than the optimal efficiency bound for the purely nonparametric approach. In addition, such an estimator relies on less-strong assumptions than common nonparametric volatility estimators. These assumptions concern the endogeneity of the noise with the frictionless price, the autocorrelation and heteroskedasticity of the noise.

We derive the asymptotic theory of the new volatility estimator. Compared to the pre-averaging estimator, the new volatility estimator does not rely on the absence of an endogeneity assumption for the noise, and allows by construction for heteroskedastic and autocorrelated noise. Moreover, if

the noise is completely removed by the liquidity-cost variables considered, then the new volatility estimator is as accurate as if the frictionless return were observed. The finite-sample study, as well as the empirical analysis using Alcoa stock, confirm the theoretical results.

In this paper, we focus on integrated volatility estimation, but the approach could improve the measurement of intraday quantities such as spot volatility (see Kristensen 2010), powers of volatility, the leverage effect and integrated betas in a multivariate setting (see Christensen *et al.* 2010). These extensions would broaden the applicability of our approach to portfolio allocation, risk management and asset evaluation.

There are many possible extensions to this work. For instance, it would be interesting to allow for endogenous and non i.i.d. residual noise. Potentially, a nonlinear or an index model of liquidity costs would capture more noise than a linear one. Indeed, nonlinearities are well documented in market microstructure theory. Another extension would be to add jumps in the frictionless-price dynamics. There is evidence of jumps in the data, so accounting for discontinuities should be explored.

In addition to the estimation of volatility-type objects, this paper's approach to decontaminate high-frequency prices from liquidity costs could be used to study whether the current stylized fact of the reversal of weekly returns (see Gutierrez Jr. and Kelley 2008) is still present for returns that are adjusted for liquidity costs.

	Bias	Relative bias	Variance	RMSE
$\hat{\beta}_1$	$3.8020\ 10^{-6}$	0.0048	$4.4865\ 10^{-7}$	$6.6982\ 10^{-4}$
$\hat{\beta}_2$	$-6.6720\ 10^{-7}$	0.0013	$1.2413\ 10^{-8}$	$1.1141\ 10^{-4}$
$\hat{\beta}_3$	$1.0227\ 10^{-5}$	$-3.4091\ 10^{-4}$	$1.8843\ 10^{-5}$	0.0043
$\hat{\beta}_4$	$-1.4712\ 10^{-7}$	$-2.9424\ 10^{-4}$	$1.0740\ 10^{-8}$	$1.0364\ 10^{-4}$
$\hat{\beta}_5$	$-1.6230\ 10^{-8}$	$4.0576\ 10^{-5}$	$1.0665\ 10^{-8}$	$1.0327\ 10^{-4}$
$RV(p)$	0.0508	50.7642	$1.3328\ 10^{-6}$	0.0508
$RV(\widehat{p})$	$3.8361\ 10^{-4}$	0.3836	$1.5139\ 10^{-7}$	$5.4639\ 10^{-4}$
$RV^{low}(p)$	0.0019	1.8721	$1.2168\ 10^{-7}$	0.0019
$RV^{low}(\widehat{p})$	$2.4807\ 10^{-6}$	0.0025	$2.5131\ 10^{-8}$	$1.5855\ 10^{-4}$
$IV^{pre}(p)$	$1.1045\ 10^{-4}$	0.1105	$2.7687\ 10^{-8}$	$1.9972\ 10^{-4}$
$IV^{pre}(\widehat{p})$	$-2.3016\ 10^{-5}$	−0.0230	$1.5166\ 10^{-8}$	$1.2528\ 10^{-4}$
IV^{new}	$-1.6638\ 10^{-5}$	−0.0166	$1.2373\ 10^{-8}$	$1.1247\ 10^{-4}$

Table 1: Simulation results

Variable	Min	Max	Mean	Std
Number of transactions per day	1828	9660	4347.2	1261.8
Daily average time between transaction in seconds	2.40	10.80	5.80	1.63
Daily average trade-direction indicator	-0.44	0.54	0.08	0.12
Daily average bid-ask spread in cents	1.01	1.28	1.06	0.04
Daily average log-traded volume	5.44	6.91	6.22	0.19
Daily average log bid depth	3.76	7.08	5.64	0.64
Daily average log ask depth	3.84	7.11	5.62	0.61

Table 2: Descriptive statistics, 2009

Variable	Min	Max	Mean	Std
Number of transactions per day	729	7598	2806.6	1188.0
Daily average time between transaction in seconds	3.08	24.13	9.75	3.85
Daily average trade-direction indicator	-0.62	0.61	0.03	0.19
Daily average bid-ask spread in cents	1.01	1.09	1.03	0.01
Daily average log-traded volume	5.25	6.87	6.28	0.27
Daily average log bid depth	5.86	7.79	6.87	0.39
Daily average log ask depth	5.61	7.88	6.86	0.39

Table 3: Descriptive statistics, 2010

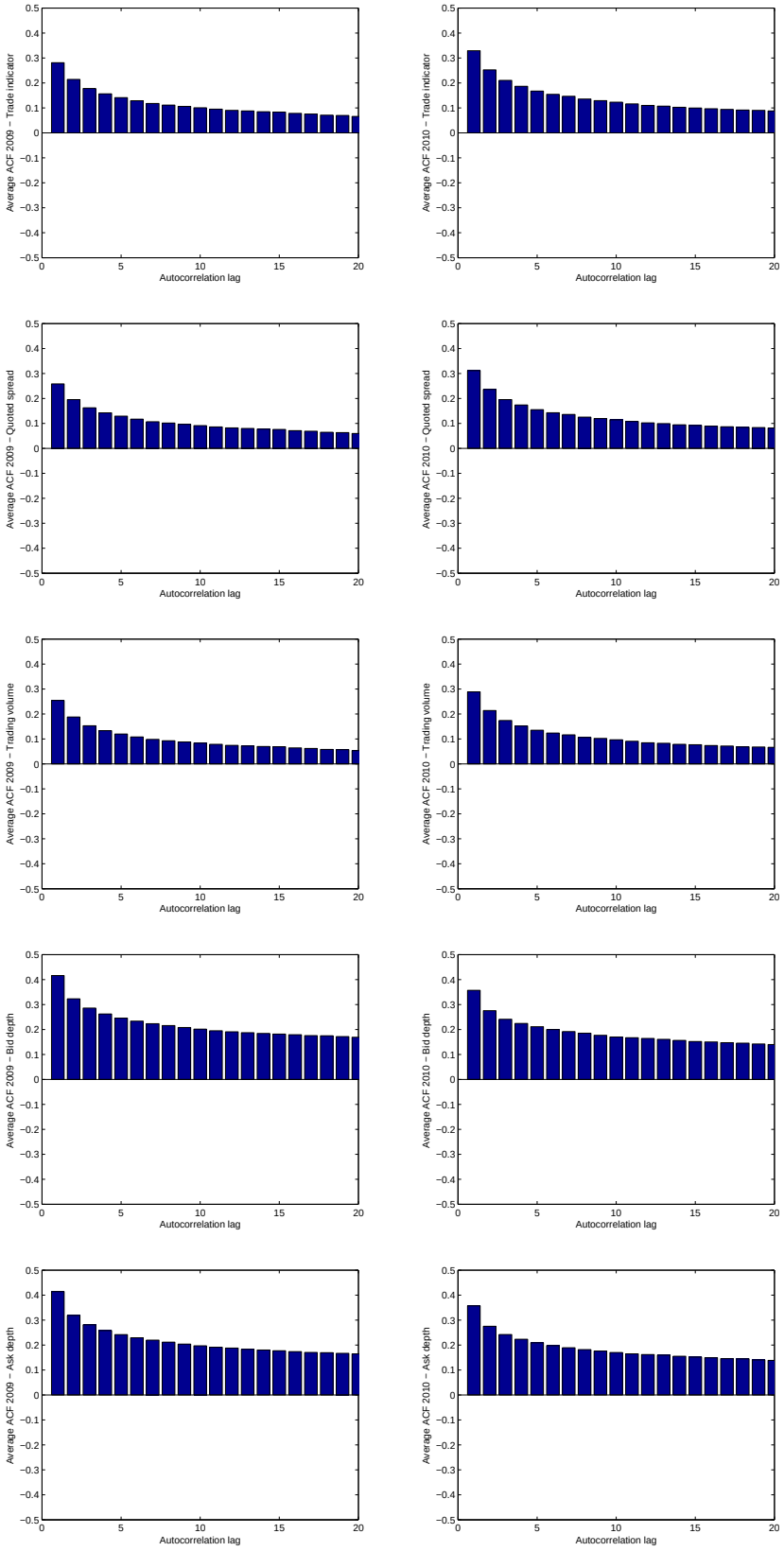
Average ACF 2009 – Trade indicator
Average ACF 2010 – Trade indicator
Average ACF 2009 – Quoted spread
Average ACF 2010 – Quoted spread
Average ACF 2009 – Trading volume
Average ACF 2010 – Trading volume
Average ACF 2009 – Bid depth
Average ACF 2010 – Bid depth
Average ACF 2009 – Ask depth
Average ACF 2010 – Ask depth
Autocorrelation lag

Figure 1: ACF

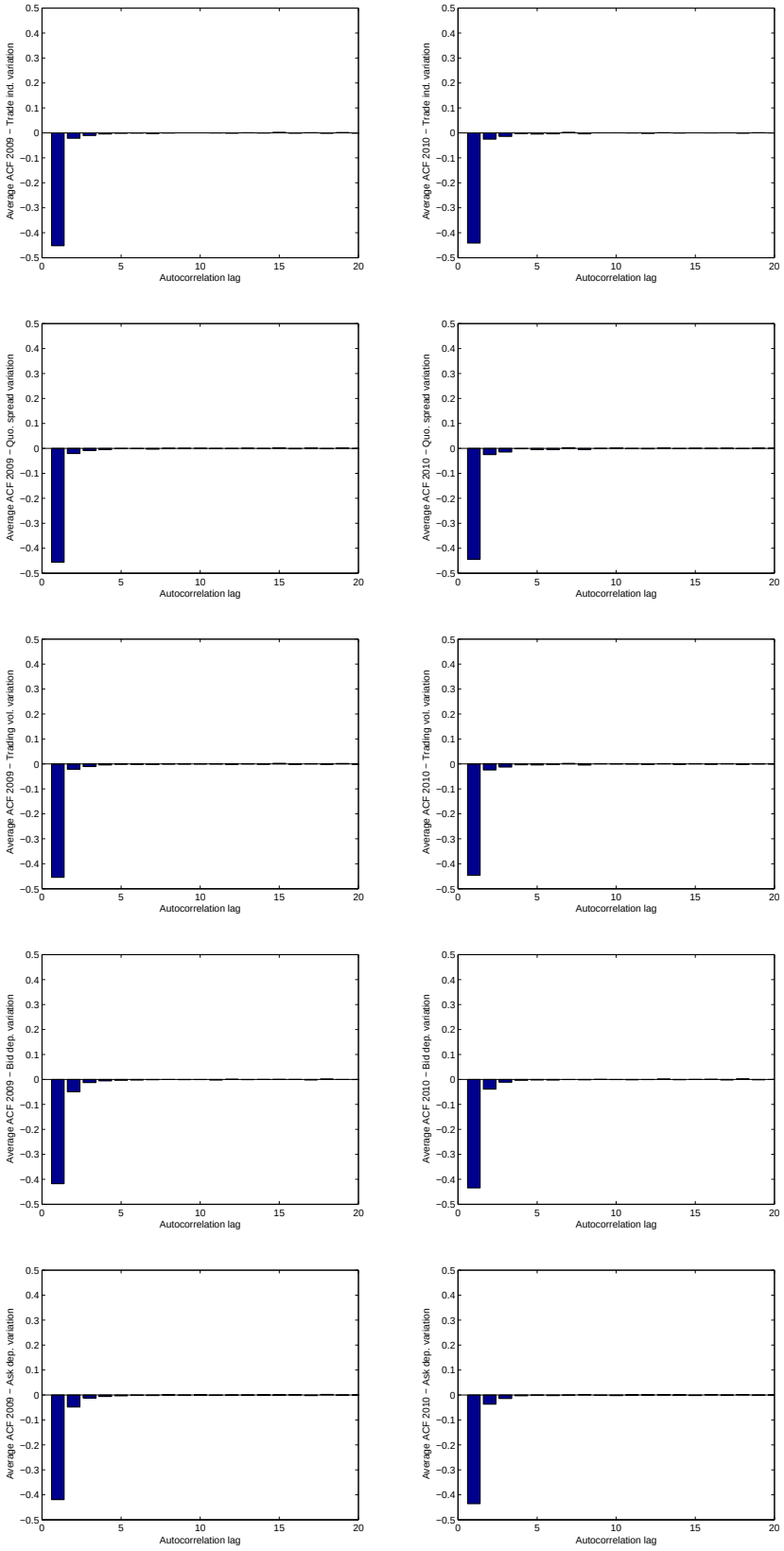

Figure 2: ACF variation

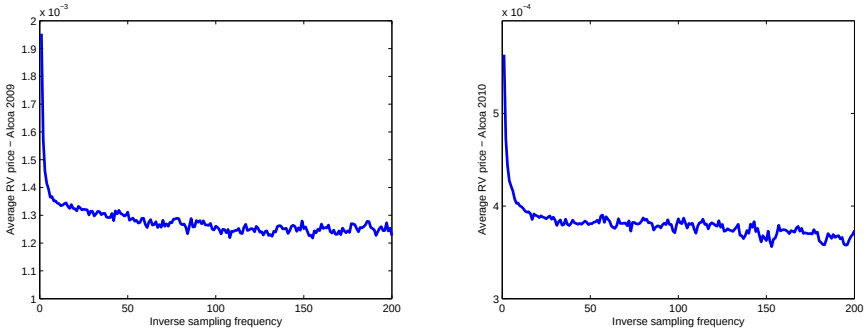

Figure 3: The trade price signature plot

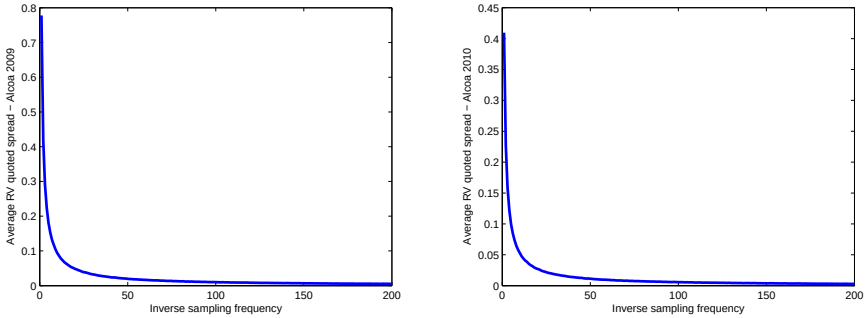

Figure 4: The quoted spread signature plot

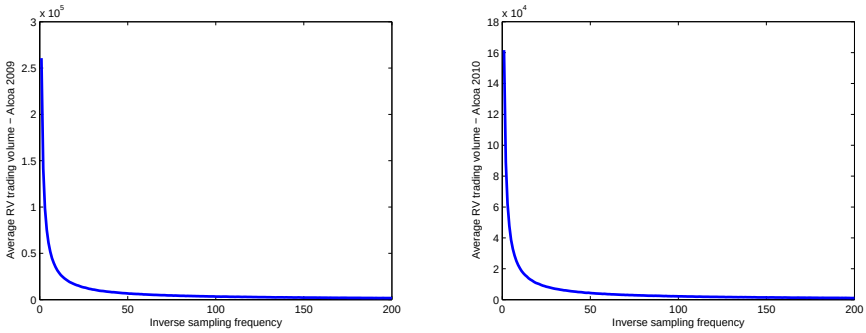

Figure 5: The trading volume signature plot

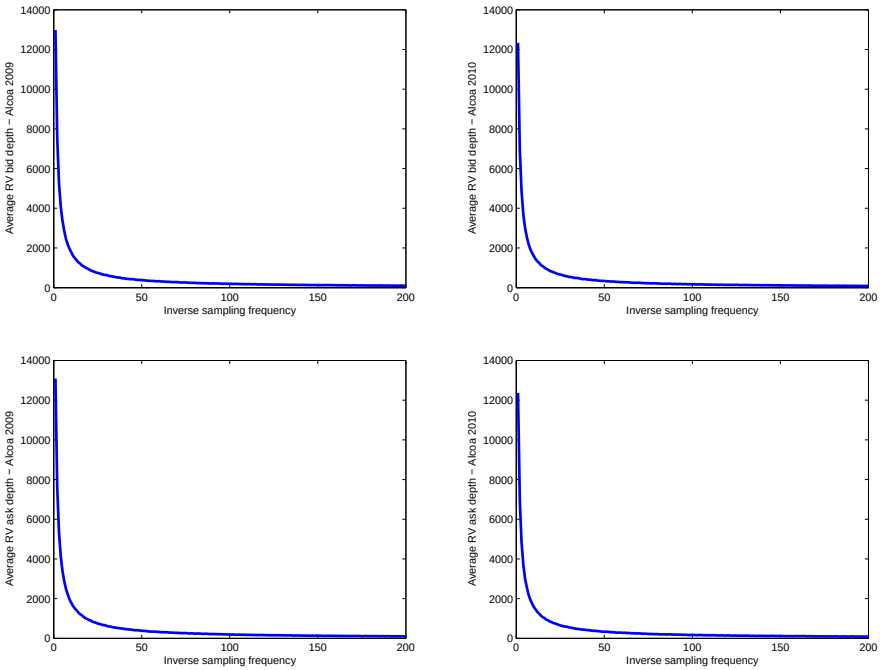

Figure 6: The quoted depths signature plot

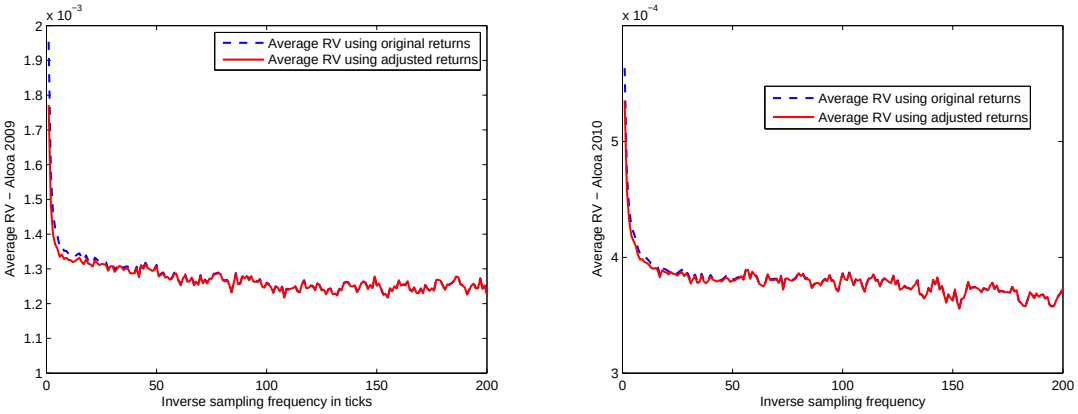

Figure 7: The original and adjusted-price signature plot

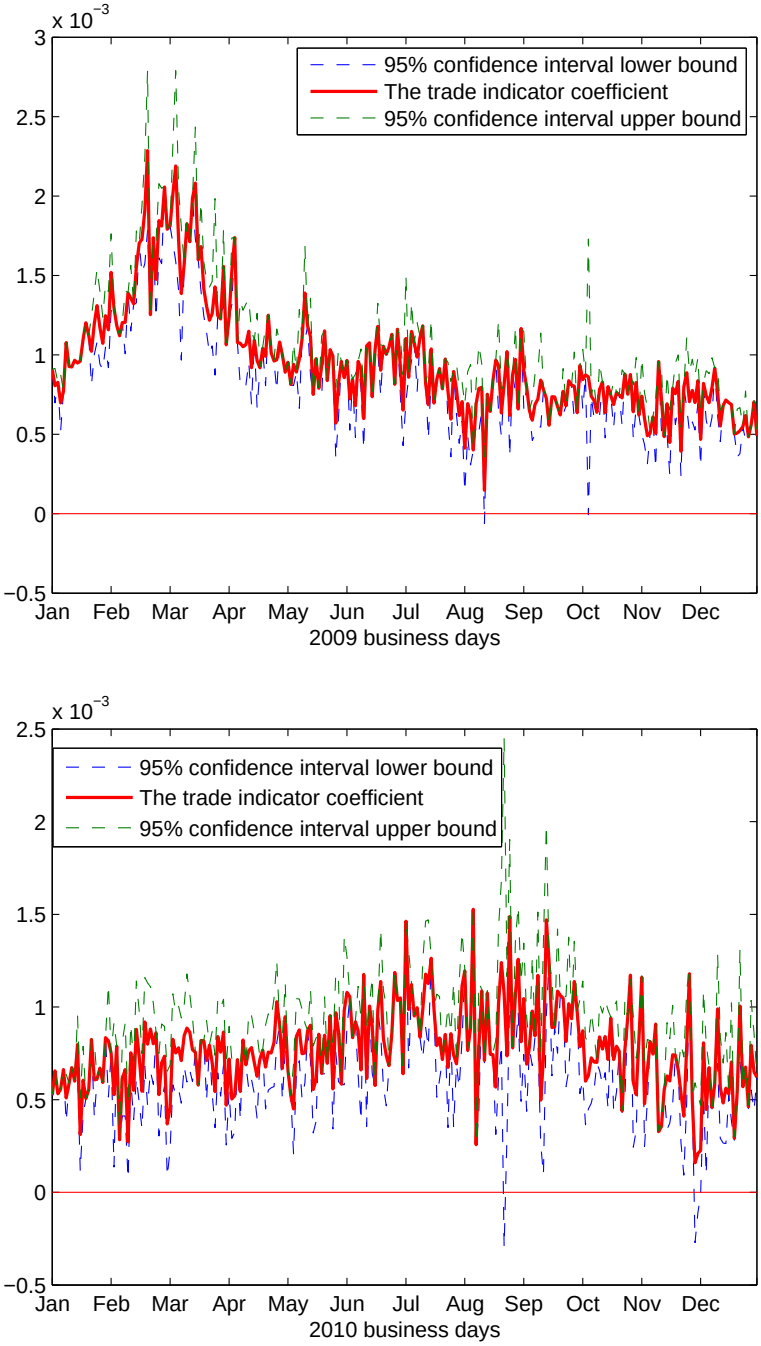

Figure 8: The trade indicator coefficient with 95% confidence interval

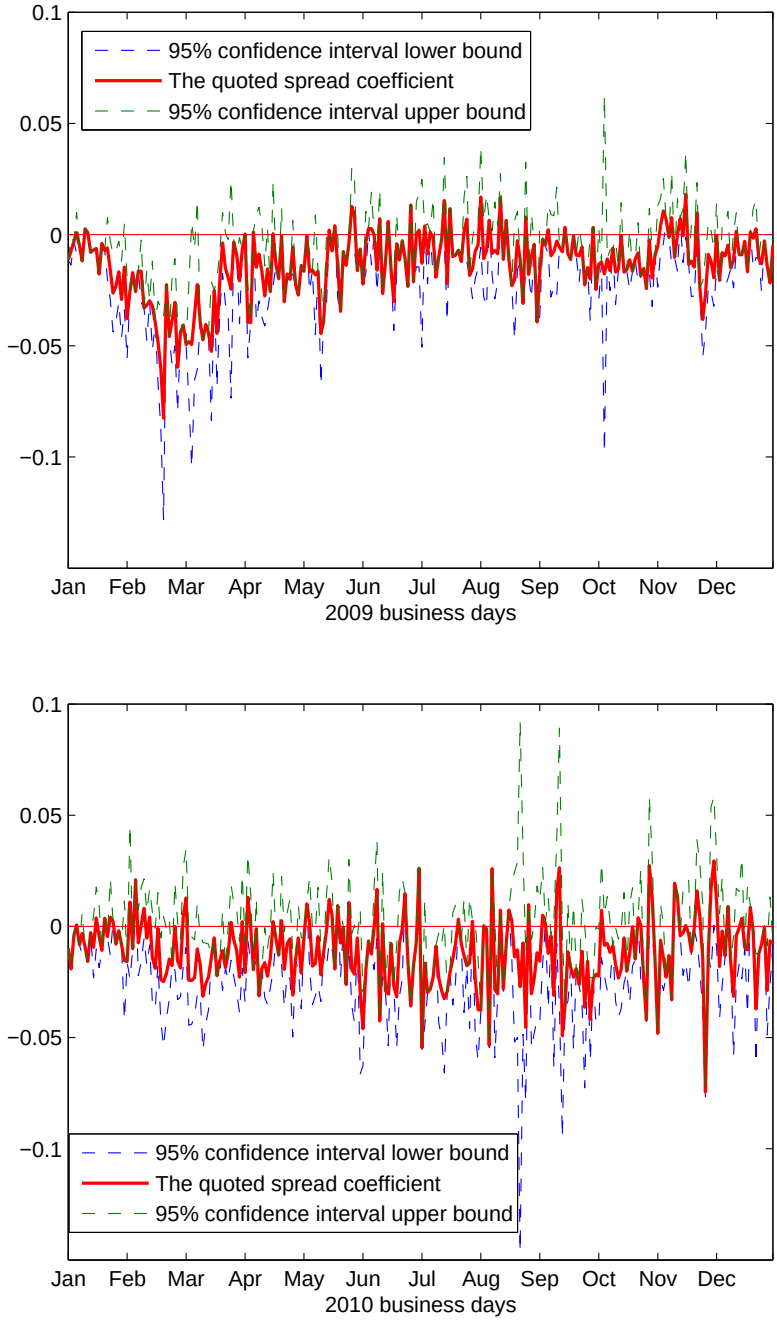

Figure 9: The quoted spread coefficient with 95% confidence interval

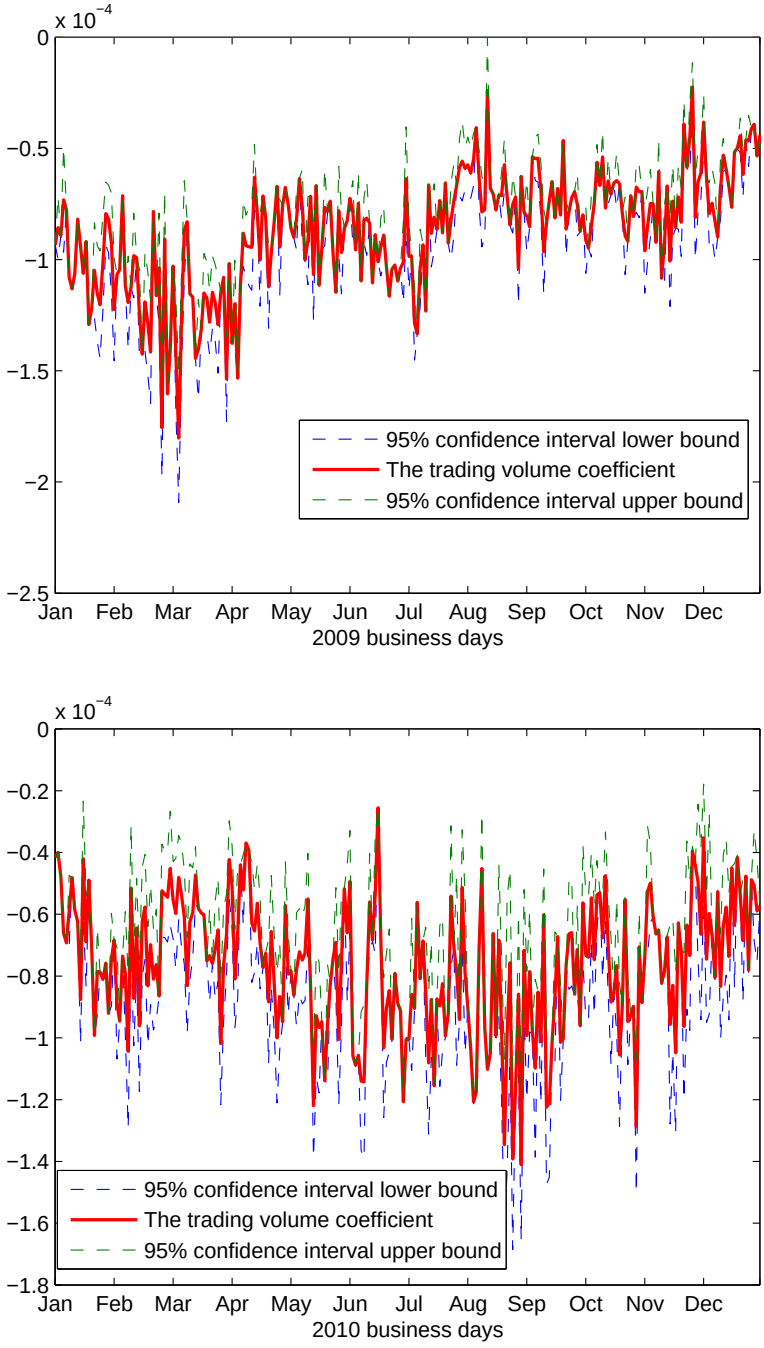

Figure 10: The trading volume coefficient with 95% confidence interval

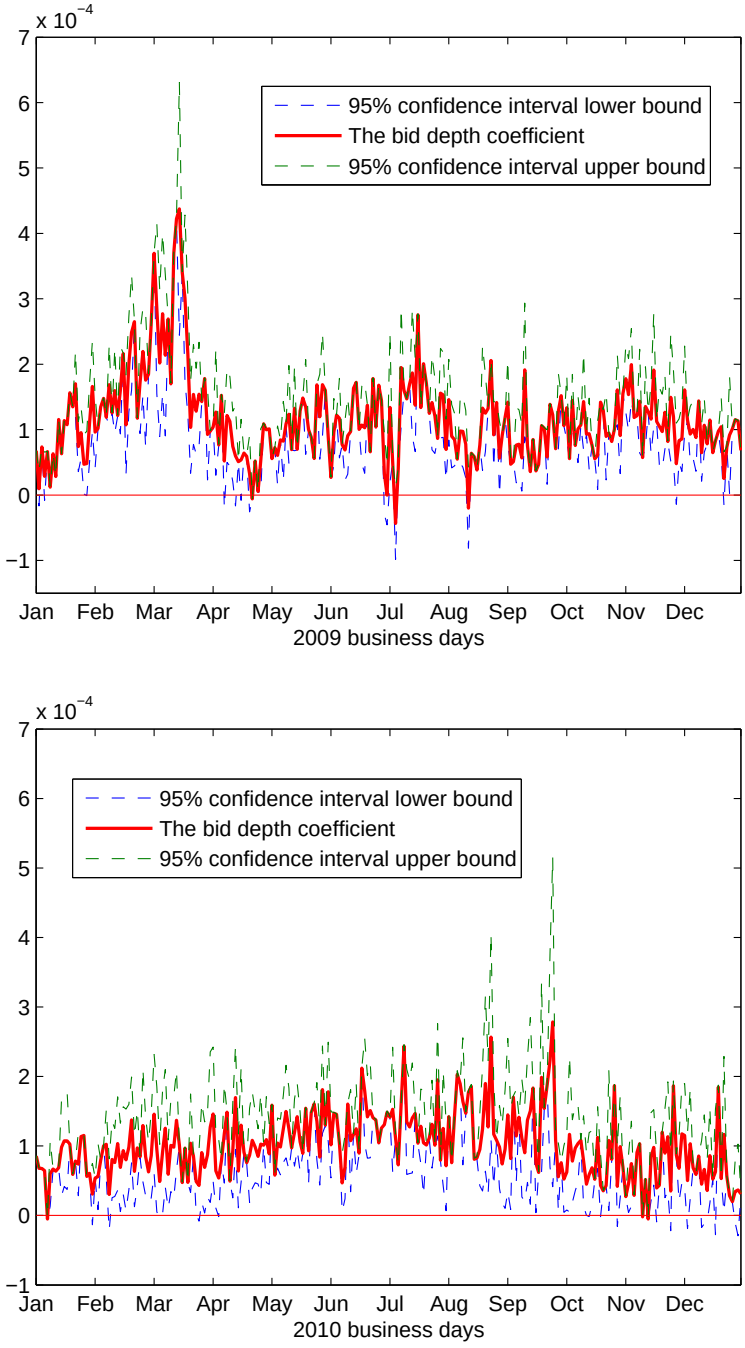

Figure 11: The bid depth coefficient with 95% confidence interval

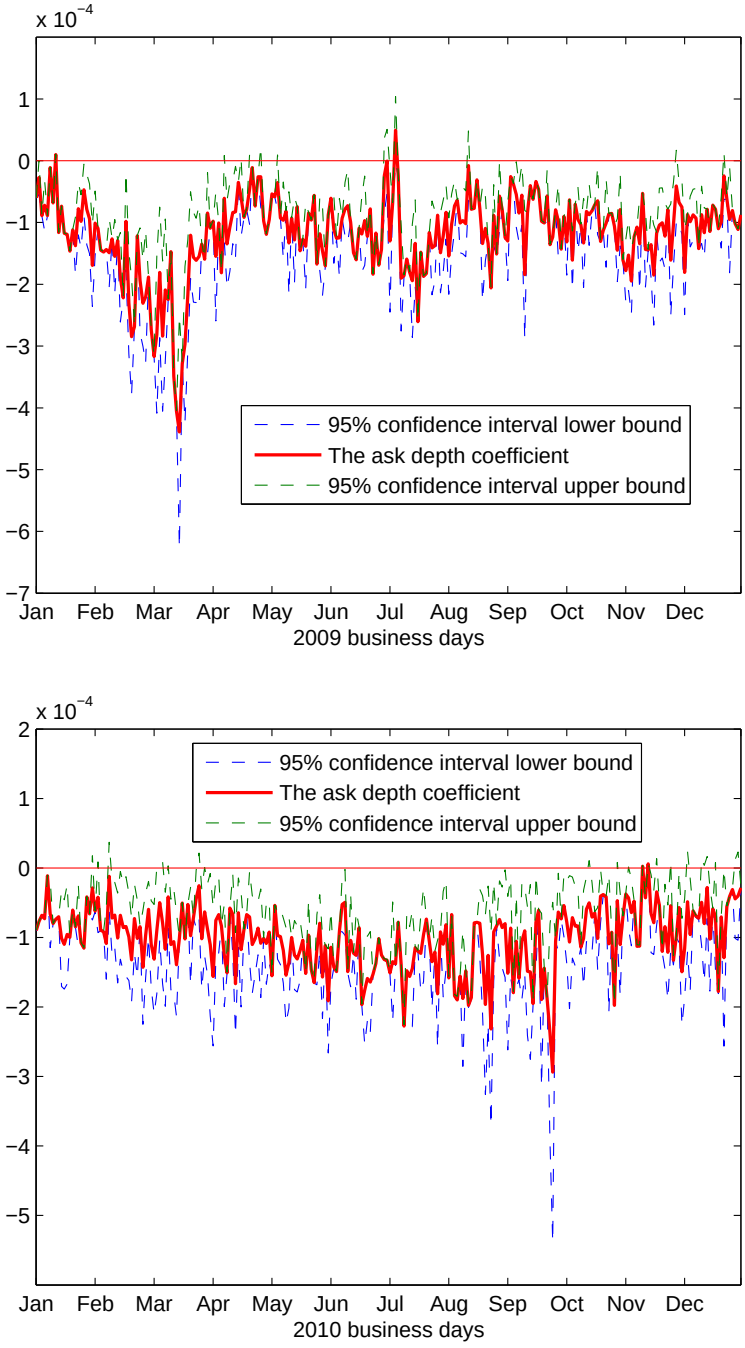

Figure 12: The ask depth coefficient with 95% confidence interval

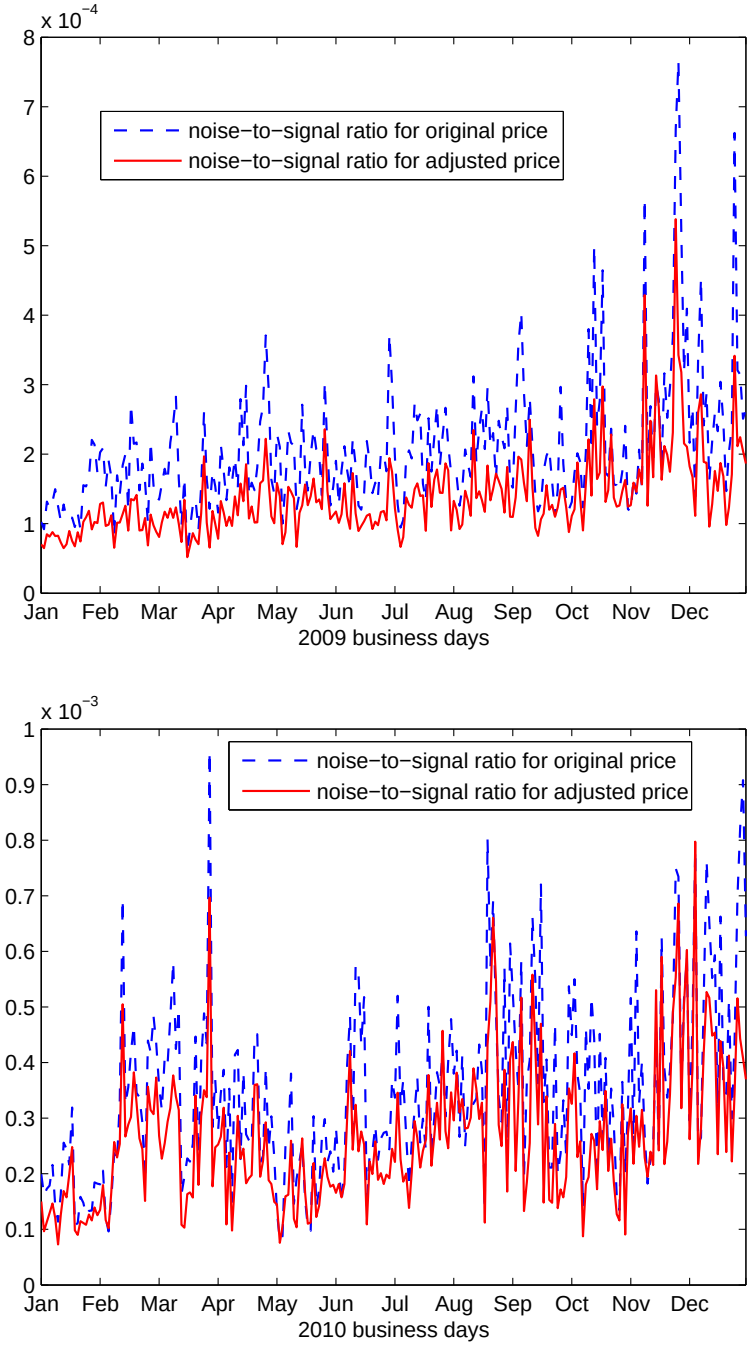

Figure 13: The noise-to-signal ratio

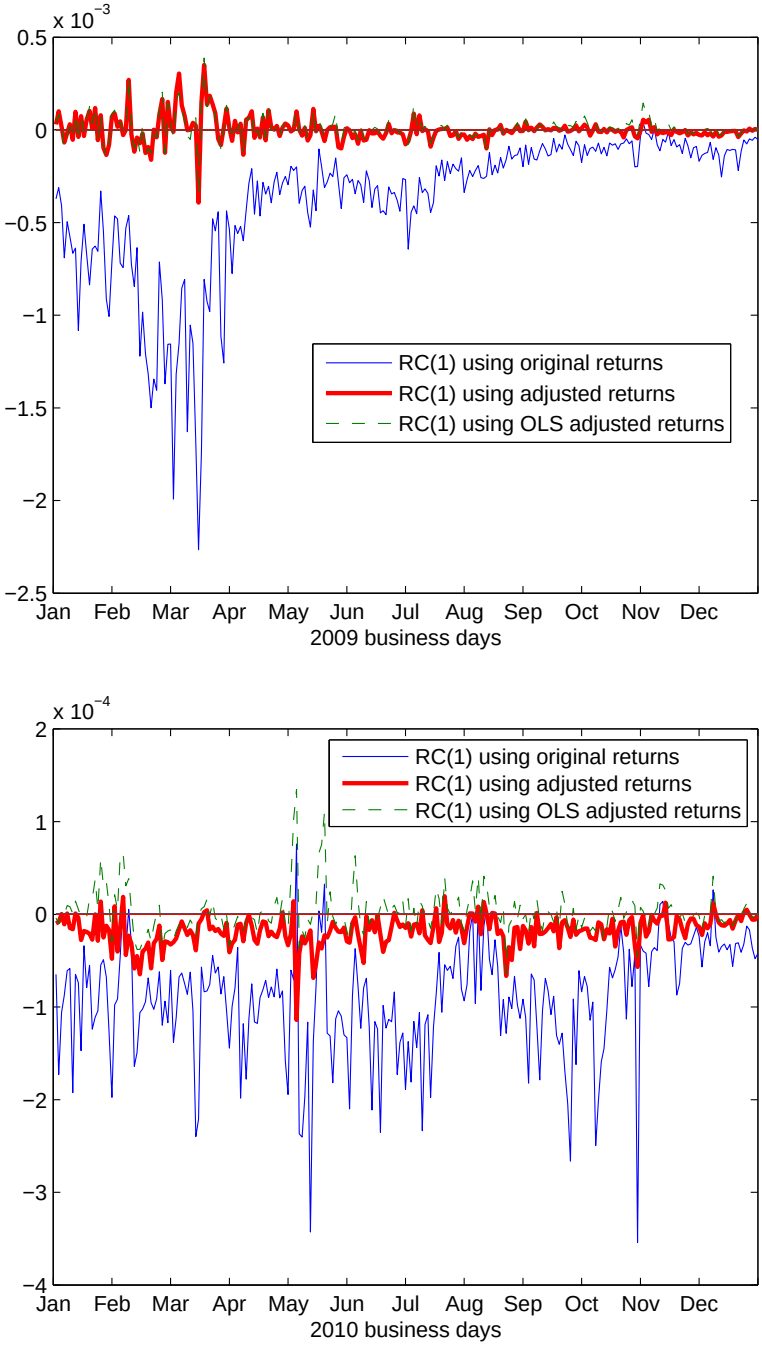

Figure 14: The first-order realized autocovariance

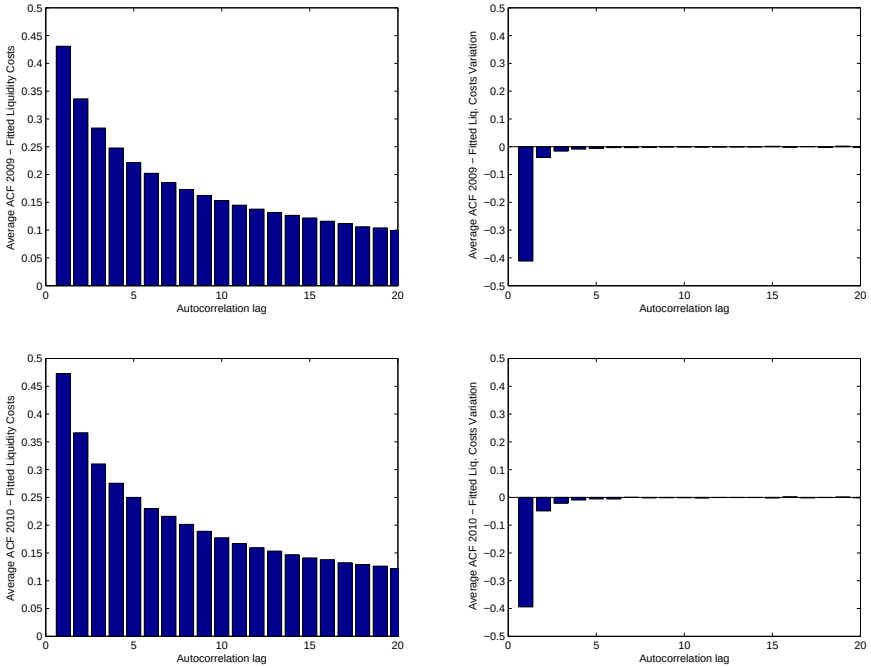

Figure 15: The Liquidity Costs ACF and the Liquidity Costs Variation ACF

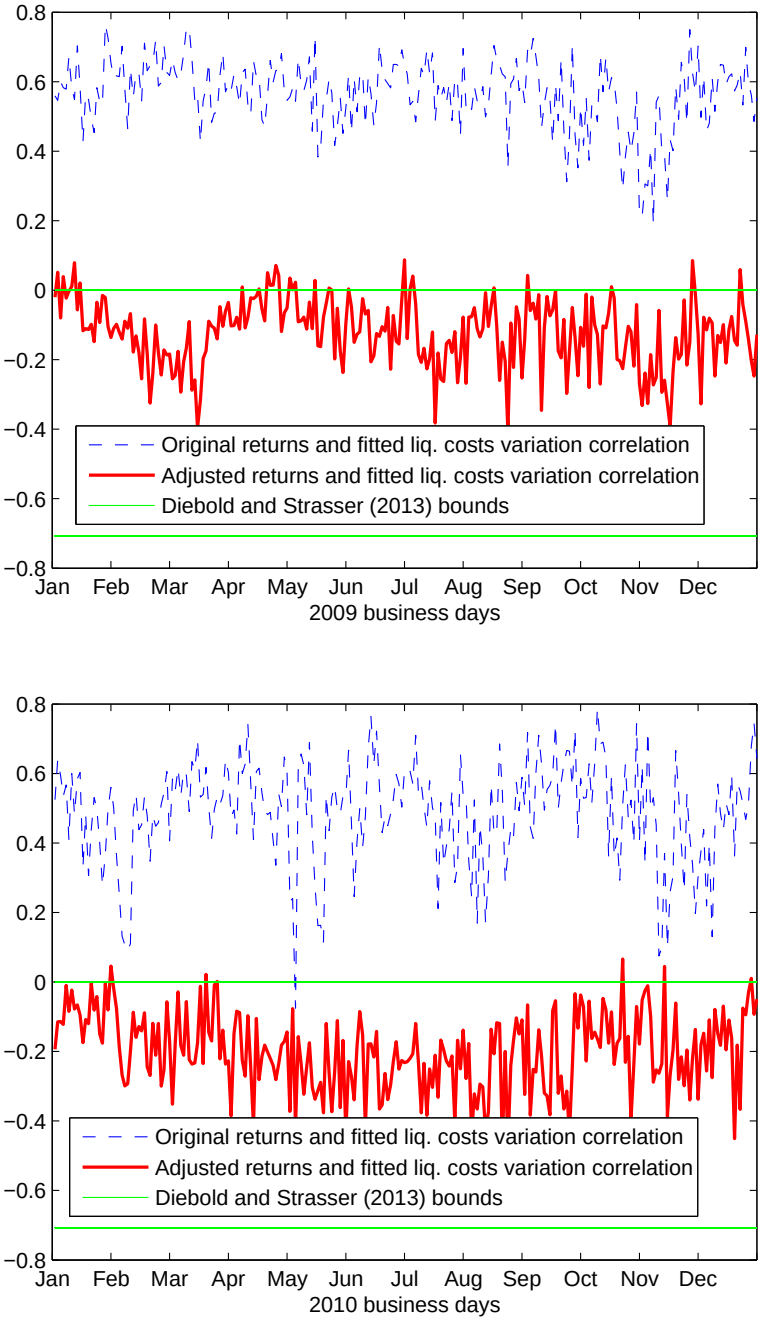

Figure 16: Return-noise correlation

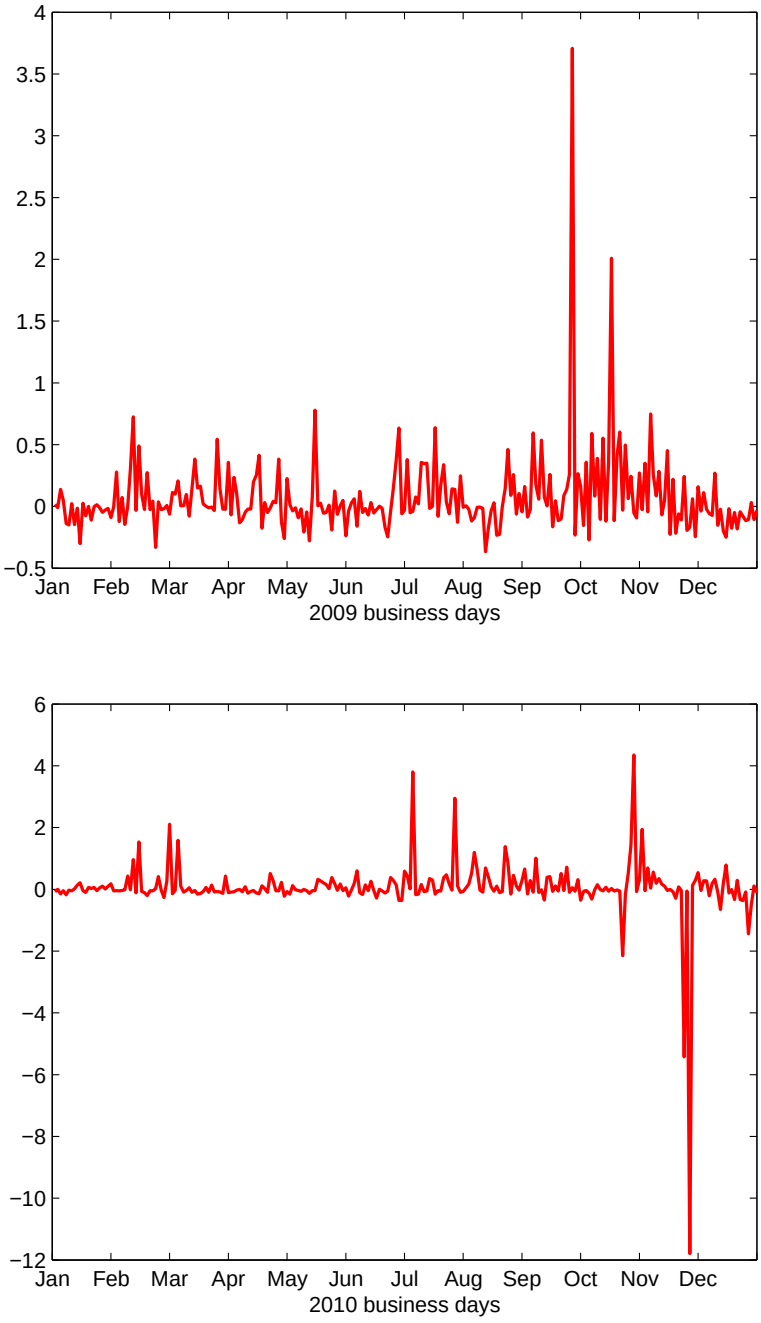

Figure 17: The daily relative difference $\frac{IV^{new}-IV^{pre}(p)}{IV^{pre}(p)}$ for 2009-2010

APPENDIX A: DATA MANIPULATIONS

As in Barndorff-Nielsen *et al.* 2008, we 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, we 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 than 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

The difference between the instrumental variable estimate $\widehat{\beta}$ and the true population parameter β is found by substituting (9) into the definition of $\widehat{\beta}$:

$$\begin{aligned}\widehat{\beta}-\beta &= (Z^{'}X)^{-1}Z^{'}r-\beta \\ &= (Z^{'}X)^{-1}Z^{'}(r^{*}+X\beta+\Delta\xi)-\beta \\ &= \underbrace{(Z^{'}X)^{-1}Z^{'}\underbrace{r^{*}}_{\mathcal{O}(1/\sqrt{N})}}_{small}+\underbrace{(Z^{'}X)^{-1}Z^{'}\underbrace{\Delta\xi}_{\mathcal{O}(1)}}_{big}.\end{aligned} \tag{B.1}$$

Assume that $Var[\xi]=0$. Then the second term in equation (B.1) vanishes. Consequently, the difference $\widehat{\beta}-\beta$ inherits the properties of the frictionless return r^{*}. Next, we derive the consistency and asymptotic distribution of $\widehat{\beta}$.
(i) Consistency:

$$\begin{aligned}(Z^{'}X)^{-1}Z^{'}r^{*} &= \left(\frac{1}{N}\sum_{i=1}^{N}Z_iX_i^{'}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}Z_ir_i^{*}\right) \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}Z_iX_i^{'}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left\{Z_i\int_{\frac{i-1}{N}}^{\frac{i}{N}}(\mu_sds+\sigma_sdW_s)\right\}\right) \\ &= \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N}Z_iX_i^{'}\right)^{-1}}_{\xrightarrow{P}\Omega^{-1}}\left(\underbrace{\frac{1}{N}\sum_{i=1}^{N}Z_i\int_{\frac{i-1}{N}}^{\frac{i}{N}}\mu_sds}_{=A}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}Z_i\int_{\frac{i-1}{N}}^{\frac{i}{N}}\sigma_sdW_s}_{=B}\right).\end{aligned} \tag{B.2}$$

For the first term, $\left(\frac{1}{N}\sum_{i=1}^{N}Z_iX_i^{'}\right)^{-1}\xrightarrow{P}\Omega^{-1}$ is obtained using Assumption A.
For the second term, $A\xrightarrow{P}0$ when $N\to\infty$ because

$$E\left[Z_i\int_{\frac{i-1}{N}}^{\frac{i}{N}}\mu_sds\right]=\underbrace{E[Z_i]}_{=0}\underbrace{\left(\int_{\frac{i-1}{N}}^{\frac{i}{N}}\mu_sds\right)}_{\xrightarrow{P}0}=0. \tag{B.3}$$

For the third term, $B \xrightarrow{P} 0$ when $N \to \infty$ because

$$\begin{aligned} E\left[Z_i \int_{\frac{i-1}{N}}^{\frac{i}{N}} \sigma_s dW_s\right] &= E\left[E\left[Z_i \int_{\frac{i-1}{N}}^{\frac{i}{N}} \sigma_s dW_s | \mathscr{F}_{i-1}\right]\right] \\ &= E\left[Z_i E\left[\int_{\frac{i-1}{N}}^{\frac{i}{N}} \sigma_s dW_s | \mathscr{F}_{i-1}\right]\right] \\ &= E\left[Z_i \underbrace{E\left[\int_{\frac{i-1}{N}}^{\frac{i}{N}} \sigma_s dW_s\right]}_{=0}\right] = 0. \end{aligned} \tag{B.4}$$

So, equations (B.2), (B.3) and (B.4) imply that $\widehat{\beta} - \beta \xrightarrow{P} 0$ when $N \to \infty$.
(ii) Asymptotic distribution:
Equation (B.1) implies that

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

We have, using Assumptions A and B,

$$\begin{aligned} &Z^{'} r^* \xrightarrow{st} \mathcal{N}((0)_{M\times 1}, \Omega^*), \\ &N^{-1} Z^{'} X \to \Omega, \\ &Var[\xi] = 0. \end{aligned} \tag{B.6}$$

Then

$$N(\widehat{\beta} - \beta) \xrightarrow{st} \mathcal{N}((0)_{M\times 1}, \Omega^{-1}\Omega^*\Omega^{-1}). \tag{B.7}$$

□

Proof of Proposition 2
From equation (B.1), we have

$$\widehat{\beta} - \beta = (Z^{'} X)^{-1} Z^{'} r^* + (Z^{'} X)^{-1} Z^{'} \Delta\xi. \tag{B.8}$$

(i) Consistency:
We have, using Assumptions 1-3,

$$\begin{aligned} &r^* \to 0, \\ &E[Z^{'} \Delta\xi] = 0. \end{aligned} \tag{B.9}$$

Then $\widehat{\beta} - \beta \to 0$.
(ii) The central limit theorem:

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta) &= \left[N^{-1} Z^{'} X\right]^{-1} \left[\sqrt{N}^{-1} Z^{'} (r^* + \Delta\xi)\right] \\ &= \left[N^{-1} Z^{'} X\right]^{-1} \left[\sqrt{N}^{-1} Z^{'} r^*\right] + \left[N^{-1} Z^{'} X\right]^{-1} \left[\sqrt{N}^{-1} Z^{'} \Delta\xi\right]. \end{aligned} \tag{B.10}$$

For the last term of the previous equation, we have, using Assumption C,

$$\sqrt{N}^{-1} Z^{'} \Delta\xi \longrightarrow \mathcal{N}((0)_{M\times 1}, S). \tag{B.11}$$

Since $Z^{'} r^*$ is bounded because it converges to 0 and $N^{-1} Z^{'} X \to \Omega$, then

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

□

Proof of Proposition 3
Recall,

$$\begin{aligned} \widehat{r}_i &= r_i - X_i^{'} \widehat{\beta} \\ &= r_i^* + X_i^{'} (\beta - \widehat{\beta}) + \Delta\xi_i. \end{aligned} \tag{B.12}$$

Under H_0,

$$\widehat{r}_i = \underbrace{r_i^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X_i^{'}(\beta - \widehat{\beta})}_{\mathcal{O}(1/N)}. \tag{B.13}$$

So the frictionless return dominates the adjusted return. Therefore, Theorem 1 of Barndorff-Nielsen *et al.* (2008) could be used to obtain that

$$\sqrt{N}\left(\sum_{i=1}^{N}\widehat{r}_i\widehat{r}_{i-1} + \sum_{i=1}^{N}\widehat{r}_i\widehat{r}_{i+1}\right) \xrightarrow{st} \mathcal{N}(0, 4IQ), \tag{B.14}$$

so $\sqrt{N}RC(1) \xrightarrow{st} \mathcal{N}(0, IQ)$.

□

Proof of Proposition 4
This proposition is a direct application of Hausman (1978). Assume that X is exogenous (H_0 holds). Then $\widehat{\beta}_{OLS}$ attains the asymptotic Cramer-Rao bound.
We distinguish the two cases $Var[\xi] = 0$ and $Var[\xi] \neq 0$ because the asymptotic variance as well as the rate of convergence of β and $\widehat{\beta}$ differ in each case.
We have, under Assumptions 1-3, A', B' and $Var[\xi] = 0$,

$$(i)\ \widehat{\beta}_{OLS} \xrightarrow{P} \beta.$$
$$(ii)N(\widehat{\beta}_{OLS} - \beta) \xrightarrow{L} \mathcal{N}\left((0)_{M\times 1}, V_0(\widehat{\beta}_{OLS})\right),$$

where $V_0(\widehat{\beta}_{OLS}) = \Omega_X^{-1}\Omega_X^*\Omega_X^{-1}$.
The proof of this result is similar to the proof of Proposition 1.
For the partially absorbed noise case, we have under Assumptions 1-3, A', C', and $Var[\xi] \neq 0$,

$$(i)\ \widehat{\beta}_{OLS} \xrightarrow{P} \beta,$$
$$(ii)\sqrt{N}(\widehat{\beta}_{OLS} - \beta) \xrightarrow{L} \mathcal{N}\left((0)_{M\times 1}, V_1(\widehat{\beta}_{OLS})\right),$$

where $V_1(\widehat{\beta}_{OLS}) = \Omega_X^{-1}S_X\Omega_X^{-1}$.
The proof of this result is similar to the proof of Proposition 2.
The matrices Ω_X, Ω_X^* and S_X are defined in Assumptions A', B' and C', respectively.
Providing the asymptotic distributions of $\widehat{\beta}_{OLS}$ and $\widehat{\beta}$ for the cases $Var[\xi] = 0$ and $Var[\xi] \neq 0$, Hausman (1978) is directly applicable to obtain the result stated in the Proposition 4.

□

Proof of Proposition 5
The pre-averaging estimator of Jacod *et al.* (2009), as well as the extended version of Hautsch and Podolskij (2013), relies on the assumption of absence of endogeneity between the frictionless price and the noise. Therefore, the result (i) is obtained.
In the case where the noise is exogenous, the pre-averaging estimator is consistent. We have

$$\begin{aligned} p &= p^* + \varepsilon \\ &= p^* + \underbrace{F'\beta}_{autocorrelated\ noise} + \underbrace{\xi}_{white\ noise} \end{aligned} \tag{B.15}$$

Under the assumption that ε is independent from p^*, we apply the pre-averaging of Hautsch and Podolskij (2013), which is robust to autocorrelated noise:

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

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

$$\begin{aligned} B(q) &= E[\varepsilon_t^2] + 2\sum_{m=1}^{q} Cov[\varepsilon_t, \varepsilon_{t+m}] \\ &= E[\left(F'_t\beta + \xi_t\right)^2] + 2\sum_{m=1}^{q} Cov[F'_t\beta + \xi_t, F'_{t+m}\beta + \xi_{t+m}] \\ &= E[\left(F'_t\beta\right)^2] + E[\xi_t^2] + 2\sum_{m=1}^{q} Cov[F'_t\beta, F'_{t+m}\beta]. \end{aligned}$$

This achieves the proof of part (ii) of Proposition 5.

□

Proof of Theorem 1
We have in the zero residual noise case,

$$\begin{aligned}\widehat{r} &= r - X\widehat{\beta} \\ &= r^* + X(\beta - \hat{\beta}),\end{aligned} \tag{B.16}$$

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

$$\widehat{r} = \underbrace{r^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X(\beta - \hat{\beta})}_{\mathcal{O}(1/N)}. \tag{B.17}$$

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

(i) $RV(\widehat{p}) \xrightarrow{P} IV$.

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

Moreover, since $\widehat{p}$ is asymptotically equal to the frictionless price p^*, then the realized quarticity $\sum_{i=1}^{N} \widehat{r}_i^4$ is a consistent estimator of the integrated quarticity IQ.

□

Proof of Theorem 2
The pre-averaging estimator using adjusted prices is a direct application of Jacod *et al.* (2009). We have,

$$\widehat{p} = p^* + \underbrace{F^{'}(\beta - \hat{\beta})}_{endogenous\ noise} + \underbrace{\xi}_{exogenous\ noise}. \tag{B.18}$$

Let $\widetilde{p}$ denote the $\mathcal{O}(1/\sqrt{N})$ of the adjusted price $\widehat{p}$

$$\widetilde{p} = p^* + F^{'}(\beta - \hat{\beta}). \tag{B.19}$$

The intuition is

$$\begin{aligned}\widehat{r} &= r^* + \underbrace{X(\beta - \hat{\beta})}_{small\ endogenous\ noise} + \underbrace{\Delta\xi}_{big\ exogenous\ noise} \\ &= \underbrace{r^*}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{X}_{\mathcal{O}(1)} \underbrace{(\beta - \hat{\beta})}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{\Delta\xi}_{\mathcal{O}(1)} \\ &= \underbrace{\widetilde{r}}_{\mathcal{O}(1/\sqrt{N})} + \underbrace{\Delta\xi}_{\mathcal{O}(1)}.\end{aligned} \tag{B.20}$$

Then $\widehat{r}$ is an $\mathcal{O}(1/\sqrt{N})$ contaminated with an i.i.d. noise. Therefore, by applying the pre-averaging estimator of Jacod *et al.* (2009), we obtain the following asymptotic distribution:

$$N^{1/4}\left(IV^{pre}(\widehat{p}) - \widetilde{IV}\right) \xrightarrow{st} \mathcal{N}(0, \Gamma_\xi), \tag{B.21}$$

where

$$\Gamma_\xi = \frac{151}{140}\theta\ \widetilde{IQ} + \frac{12}{\theta}E[\xi^2]\widetilde{IV} + \frac{96}{\theta^3}E[\xi^2]^2. \tag{B.22}$$

$$\begin{aligned}\widetilde{IV} &= plim\left(\sum_{i=1}^{N} \widetilde{r}_i^2\right), \\ \widetilde{IQ} &= plim\left(\frac{N}{3}\sum_{i=1}^{N} \widetilde{r}_i^4\right).\end{aligned} \tag{B.23}$$

Next, we turn to the asymptotic bias $IV^{pre}(\widehat{p}) - IV$. The volatility $\widetilde{IV}$ that appears in (B.21) is the limit of $\sum_{i=1}^{N} \widetilde{r}_i^2$, which is written as

$$\begin{aligned}\sum_{i=1}^{N} \widetilde{r}_i^2 &= \sum_{i=1}^{N} (r_i^* + X_i^{'}(\beta - \hat{\beta}))^2 \\ &= \sum_{i=1}^{N} (r_i^*)^2 + \sum_{i=1}^{N} (X_i^{'}(\beta - \hat{\beta}))^2 + 2\sum_{i=1}^{N} r_i^* X_i^{'}(\beta - \hat{\beta}).\end{aligned} \tag{B.24}$$

The first term of (B.24) converges to IV. For the second term,

$$\begin{aligned}
&\sum_{i=1}^{N}(X_i^{'}(\beta-\hat{\beta}))^2=\sum_{i=1}^{N}(\beta-\hat{\beta})^{'}X_iX_i^{'}(\beta-\hat{\beta})\\
&=(\beta-\hat{\beta})^{'}(\sum_{i=1}^{N}X_iX_i^{'})(\beta-\hat{\beta})\\
&=trace\left((\beta-\hat{\beta})^{'}(\sum_{i=1}^{N}X_iX_i^{'})(\beta-\hat{\beta})\right)\\
&=trace\left(\underbrace{(\frac{\sum_{i=1}^{N}X_iX_i^{'}}{N})}_{\to\Omega_X}\underbrace{N(\beta-\hat{\beta})(\beta-\hat{\beta})^{'}}_{\to\Omega^{-1}S\Omega^{-1}}\right).
\end{aligned}\tag{B.25}$$

Then

$$plim(\sum_{i=1}^{N}(X_i^{'}(\beta-\hat{\beta}))^2)=trace(\Omega_X\Omega^{-1}S\Omega^{-1}).\tag{B.26}$$

The last term of (B.24) converges to 0 because

$$\sum_{i=1}^{N}r_i^*X_i^{'}(\beta-\hat{\beta})=\underbrace{\sum_{i=1}^{N}r_i^*X_i^{'}}_{bounded}\underbrace{(\beta-\hat{\beta})}_{\to0}.\tag{B.27}$$

Using (B.24), (B.26) and (B.27), the bias $IV^{pre}(\widehat{p})-IV$ is given by $trace(\Omega_X\Omega^{-1}S\Omega^{-1})$, which proves Theorem 2(i).
For (ii), the central limit theorem of the pre-averaging estimator using adjusted prices is derived in (B.21).

□

APPENDIX C: EMPIRICAL DETAILS

Estimating the matrices S and Ω^*

For the instrumental variable estimation of the price-impact regression:
A consistent positive semidefinite estimator of the matrix S is the Newey and West (1987) estimator, which is robust to heteroskedasticity and first-order autocorrelation in the regression residuals,

$$\widehat{S}=\widehat{\Upsilon}_0+\frac{1}{2}\left(\widehat{\Upsilon}_1+\widehat{\Upsilon}_1^{'}\right),\tag{C.1}$$

where $\widehat{\Upsilon}_0=\frac{1}{N}\sum_{i=1}^{N}\widehat{r}_i^2Z_iZ_i^{'}$ and $\widehat{\Upsilon}_1=\frac{1}{N}\sum_{i=2}^{N}\widehat{r}_i\widehat{r}_{i-1}Z_iZ_{i-1}^{'}$.

$$\widehat{\Omega}^*=\frac{1}{N}\sum_{i=1}^{N}\widehat{r}_i^2Z_iZ_i^{'}.\tag{C.2}$$

For the OLS estimation of the price-impact regression:

$$\widehat{S}_X=\widehat{\Upsilon}_0+\frac{1}{2}\left(\widehat{\Upsilon}_1^{(X)}+\widehat{\Upsilon}_1^{(X)'}\right),\tag{C.3}$$

where $\widehat{\Upsilon}_0^{(X)}=\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{r}_i^{(OLS)}\right)^2X_iX_i^{'}$ and $\widehat{\Upsilon}_1^{(X)}=\frac{1}{N}\sum_{i=2}^{N}\widehat{r}_i^{(OLS)}\widehat{r}_{i-1}^{(OLS)}X_iX_{i-1}^{'}$.

$$\widehat{\Omega}_X^*=\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{r}_i^{(OLS)}\right)^2X_iX_i^{'}.\tag{C.4}$$

The pre-averaging estimator of the integrated quarticity *IQ*

The pre-averaging estimator of IQ based on the observed prices is given by

$$\begin{aligned}
IQ^{pre}(p)=&\frac{48}{\theta^2}\sum_{i=0}^{N-k}\left\{\sum_{j=1}^{k}\phi\left(\frac{j}{k}\right)r_{i+j}\right\}^4\\
&-\frac{288}{\theta^4\sqrt{N}}\sum_{i=0}^{N-2k}\left\{\sum_{j=1}^{k}\phi\left(\frac{j}{k}\right)r_{i+j}\right\}^2\widehat{B}(q)+\frac{144}{\theta^4}\widehat{B}(q)^2,
\end{aligned}\tag{C.5}$$

where $\widehat{B}(q)$ is computed using the method of Hautsch and Podolskij (2013) after estimating q with data. The pre-averaging estimator of IQ based on adjusted prices is given by

$$\begin{aligned} IQ^{pre}(\widehat{p}) = \frac{48}{\theta^2}\sum_{i=0}^{N-k}\left\{\sum_{j=1}^{k}\phi\left(\frac{j}{k}\right)\widehat{r}_{i+j}\right\}^4 \\ -\frac{144}{\theta^4 N}\sum_{i=0}^{N-2k}\left\{\sum_{j=1}^{k}\phi\left(\frac{j}{k}\right)\widehat{r}_{i+j}\right\}^2\left\{\sum_{j=i+k+1}^{i+2k}\widehat{r}_j^2\right\}+\frac{36}{\theta^4 N}\sum_{i=1}^{N-2}\widehat{r}_i^2\widehat{r}_{i+2}^2. \end{aligned} \tag{C.6}$$

Estimating the asymptotic variances Γ_ξ and $\Gamma_\varepsilon(q)$
The estimators are given by

$$\widehat{\Gamma}_\varepsilon(q) = \frac{151}{140}\theta IQ^{pre}(p) + \frac{12}{\theta}\widehat{B}(q)IV^{pre}(p) + \frac{96}{\theta^3}\widehat{B}(q)^2 \tag{C.7}$$

$$\widehat{\Gamma}_\xi = \frac{151}{140}\theta IQ^{pre}(\widehat{p}) + \frac{12}{\theta}\widehat{E}[\xi^2]IV^{pre}(\widehat{p}) + \frac{96}{\theta^3}\widehat{E}[\xi^2]^2, \tag{C.8}$$

where $\widehat{E}[\xi^2] = \frac{1}{2N}\sum_{i=1}^{N}\widehat{r}_i^2$, $IV^{pre}(p)$ and $IV^{pre}(\widehat{p})$ are given in section 4.

REFERENCES

Admati, A. R. and Pfleiderer, P. (1988), "A Theory of Intraday Patterns: Volume and Price Variability", *Review of Financial Studies*, **1**, 3–40.

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

Aït-Sahalia, Y. and Yu, J. (2009), "High-Frequency Market Microstructure Noise Estimates and Liquidity Measures", *Annals of Applied Statistics*, **3**, 422–457.

Aldous, D. G. and Eagleson, G. K. (1978), "On Mixing and Stability of Limit Theorems", *Annals of Probability*, **6**, 325–331.

Andersen, T. G., Bollerslev, T., Diebold, F. X. (2007), "Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility", *Review of Economics and Statistics*, **89**, 701–720.

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., Diebold, F. X. and Labys, P. (2003), "Modeling and Forecasting Realized Volatility", *Econometrica*, **71**, 579–625.

Andersen, T.G., Bollerslev, T. and Meddahi, N. (2011), "Realized Volatility Forecasting and Market Microstructure Noise", *Journal of Econometrics*, **160**, 220–234.

Awartani, B., Corradi, V. and Distaso, W. (2009), "Assessing Market Microstructure Effects via Realized Volatility Measures with an Application to the Dow Jones Industrial Average Stocks", *Journal of Business & Economic Statistics*, **27**, 251–265.

Bandi, F. and Russell, J. (2006a), "Separating Microstructure Noise from Volatility", *Journal of Financial Economics*, **79**, 655–692.

Bandi, F. and Russell, J. (2006b), "Full-Information Transaction Costs", *Working Paper*.

Bandi, F. and Russell, J. (2008), "Microstructure Noise, Realized Variance, and Optimal Sampling", *Review of Economic Studies*, **72**, 339–369.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008), "Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise", *Econometrica*, **76**, 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011), "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading", *Journal of Econometrics*, **162**, 149–169.

Carrasco, M. and Kotchoni, R. (2011), "Shrinkage Realized Kernels", *CIRANO Scientific Series*.

Christensen, K., Kinnerbrock, S., Podolskij, M. (2010), "Pre-averaging Estimators of the Ex-Post Covariance Matrix in Noisy Diffusion Models With Non-Synchronous Data", *Journal of Econometrics*, **159**, 116–133.

Demsetz, H. (1968), "The Cost of Transacting", *Quarterly Journal of Economics*, **82**, 33–53.

Diebold, F. X. and Strasser, G. H. (2013), "On the Correlation Structure of Microstructure Noise: A Financial Economic Approach", *Review of Economic Studies*, first published online 19 February 2013.

Easley, D. and O'Hara, M. (1987), "Price, Trade Size and Information in Securities Markets", *Journal of Financial Economics*, **19**, 69–90.

Glosten, L. R. (1989), "Insider Trading, Liquidity, and the Role of the Monopolist Specialist", *Journal of Business*, **62**, 211–235.

Glosten, L. R. and Harris, L. E. (1988), "Estimating the Components of the Bid/Ask Spread", *Journal of Financial Economics*, **21**, 123–142.

Gloter, A. and Jacod, J. (2001), "Diffusions with Measurement Errors. I. Local Asymptotic Normality", *ESAIM: Probability and Statistics*, **5**, 225–242.

Gonçalves, S. and Meddahi, N. (2009), "Bootstrapping Realized Volatility", *Econometrica*, **77**, 283–306.

Gutierrez Jr., R. C. and Kelley, E. K. (2008), "The Long-Lasting Momentum in Weekly Returns", *Journal of Finance*, **63**, 415–447.

Hansen, P. R. and Lunde, A. (2006), "Realized Variance and Market Microstructure Noise", *Journal of Business & Economic Statistics*, **24**, 127–161.

Hasbrouck, J. (1991), "Measuring the Information Content of Stock Trades", *Journal of Finance*, **46**, 179–207.

Hasbrouck, J. (1999), "The Dynamics of Discrete Bid and Ask Quotes", *Journal of Finance*, **54**, 2109–2142.

Hausman, J. A. (1978), "Specification Tests in Econometrics", *Econometrica*, **46**, 1251–1271.

Hautsch, N. and Podolskij, M. (2013), "Pre-Averaging Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence", to appear in *Journal of Business & Economic Statistics*.

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

Kalnina, I. and Linton, O. (2008), "Estimating Quadratic Variation Consistently in the Presence of Endogenous and Diurnal Measurement Error", *Journal of Econometrics*, **147**, 47–59.

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

Klotz, J. (1972), "Markov Chain Clustering of Births by Sex", *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, **4**, 173–185, Univ. of California Press.

Kristensen, D. (2010), "Nonparametric Filtering of the Realized Spot Volatility: A Kernel-Based Approach", *Econometric Theory*, **26**, 60–93.

Kyle, A. S. (1985), "Continuous Auctions and Insider Trading", *Econometrica*, **53**, 1315–1335.

Lee, C. M. C. and Ready, M. J. (1991), "Inferring Trade Direction from Intraday Data", *Journal of Finance* **26**, 733–746.

Newey, W. K., West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* **55**, 703–708.

Nolte, I., Voev, V. (2012), "Least Squares Inference on Integrated Volatility and the Relationship Between Efficient Prices and Noise", *Journal of Business & Economic Statistics* **30**, 94–108.

Podolskij, M. and Vetter, M. (2009), "Estimation of Volatility Functionals in the Simultaneous Presence of Microstructure Noise and Jumps", *Bernoulli*, **15**, 634–658.

Roll, R. (1984), "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market", *Journal of Finance* **39**, 1127–1139.

Stock, J. H. (1987), "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors", *Econometrica* **55**, 1035–1056.

Stoll, H. R. (2000), "Friction", *Journal of Finance* **55**, 1479–1514.

Zhang, L., Mykland, P. A. and 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 & Economic Statistics*, **14**, 45–52.