Working Paper 76
Business Economic Series 09
December 2009

Departamento de Economía de la Empresa
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249607

Volatility and Covariation of Financial Assets: A High-Frequency Analysis *

Alvaro Cartea [1] and Dimitrios Karyampas [2]

Forthcoming: Journal of Banking and Finance

Abstract

Using high frequency data for the price dynamics of equities we measure the impact that market microstructure noise has on estimates of the: (i) volatility of returns; and (ii) variance-covariance matrix of n assets. We propose a Kalman-filter-based methodology that allows us to deconstruct price series into the true efficient price and the microstructure noise. This approach allows us to employ volatility estimators that achieve very low Root Mean Squared Errors (RMSEs) compared to other estimators that have been proposed to deal with market microstructure noise at high frequencies. Furthermore, this price series decomposition allows us to estimate the variance covariance matrix of *n* assets in a more efficient way than the methods so far proposed in the literature. We illustrate our results by calculating how microstructure noise affects portfolio decisions and calculations of the equity beta in a CAPM setting.

Keywords: volatility estimation; high-frequency data; market microstructure theory; covariation of assets; matrix process; Kalman filter.
JEL Classification: G12, G14, C22.

* We are grateful to G. Amromin, L. Benzoni, J. Gil- Bazo, A. Justiniano, J. Navas, Z.Psaradakis, J. Penalva, E. Schwartz, and all participants at the seminars at Federal Reserve Bank of Chicago, University of Chicago, University of Toronto, University of Siena, Universidad Carlos III and Athens University of Economics and Business for useful comments. We are thankful to CEMFI (Madrid) where part of this research was undertaken, and Birkbeck College Research Committee and the Economic & Social Research Council for financial support. We are also grateful to Ike Mathur (the editor) and an anonymous referee for very useful comments.

[1] alvaro.cartea@uc3m.es
[2] d.karyampas@ems.bbk.ac.uk.

1. Introduction

Volatility of asset returns is one of the most important variables in finance. It is an important "building block" in many areas including portfolio and risk management, investment appraisal and derivatives pricing. In general, measuring volatility is not a straightforward task because it is not directly observable from the data. Focusing on financially traded assets, the answer to the question of how to estimate volatility, or quantities that use volatility as an input, will inevitably depend on modeling assumptions and compromises that will have an effect on both its estimate and estimator. For example, some of the difficulties that arise when measuring volatility stem from assumptions such as the model driving the asset price dynamics or practical issues such as the frequency or amount of data that should be employed in the estimation.

Model assumptions for asset price dynamics and the choice of data employed in the estimation of volatility are generally not independent modeling decisions. Until now, the literature has developed a large number and diverse range of volatility models that are applied in different contexts and to various financial applications. However, most of these models have been developed and tested by employing data sets that make use of a very small subset of the complete sample of available trades. In fact, the common approach has been to employ low frequency data, say one data point per trading day, when there could be thousands of intra-day observations.

Relying only on daily observations results in the discarding of a significant amount of information which in some asset classes, such as equity, can account for more than 99% of the available data. On the other hand, it is not clear whether employing as much data as possible will unequivocally improve the accuracy of the volatility estimates. The answer to the question of whether more data is preferred to less depends not only on quantity, but also on quality.

The highest resolution of stock price data is tick-by-tick data. It could be either a record of every trade or every trade and quote (including bid and ask). For a long time, the market microstructure literature has highlighted the difficulties arising from such high frequency data (see for instance Black (1986)). One of the key problems is a 'quality' issue since tick-by-tick data contains microstructure noise. In other words, tick-by-tick prices consist of the true or efficient price plus noise. Therefore, the approach of using all observations may lead to entirely different, and possibly misleading, results to those obtained if the high frequency data were to only contain the true price, i.e. no microstructure noise. Zhang et al. (2005) look into the question of how often it is optimal to sample a continuous-time diffusion process in the presence of microstructure noise.

One of the objectives of their work is to look at the delicate balance between quality and quantity of the data. They show, under a set of assumptions governing the dynamics of share prices, that estimating the volatility of the true price process uses neither all available data nor a limited subset of trades such as the daily observations previously described.

In this paper we aim to provide estimators for the variance and covariance matrices of stock returns. The main obstacle we face is to identify estimators that are not tainted by market microstructure noise. To this end, we propose a Kalman filter-based approach where we not only obtain efficient volatility estimators, but also generate efficient covariances between several assets and high frequency returns that are not affected by the microstructure noise. The Kalman-based volatility estimator we propose is a best estimator in the Cramér-Rao criterion sense, with a RMSE as small as that of the Maximum Likelihood Estimator proposed by Aït-Sahalia et al. (2005). Moreover, our proposed estimator is better, according to the RMSE metric, than the non-parametric estimators that the current literature has proposed to deal with the presence of market microstructure noise. The Kalman filter approach also offers us the ability to generate de-noised series for each asset price, and by applying common estimators of covariance to the filtered series we can build a best unbiased estimator of the covariation matrix.[1]

We illustrate the significance of our results with two examples. First, we show that the efficient frontier and the optimal weights of a portfolio differ significantly when high frequency data, rather than daily data, are used. More precisely, we show that an efficient frontier will vary in location and shape according to whether it is calculated using daily data, high frequency data with microstructure noise, or high frequency data without microstructure noise. Therefore, for a given risk target, a portfolio manager would have three different portfolio mixes to choose from depending on the data set employed, when in theory only one of these choices is correct. For example, an efficient frontier calculated with high frequency data, with microstructure noise, exhibits greater levels of risk, per level of return, than those exhibited by a frontier calculated using daily observations or a frontier calculated using high frequency data where the microstructure noise has been filtered out. Moreover, we find that the filtered high frequency efficient frontier also differs from the daily one. Finally, our results also show that measurements of log-returns will be different depending on the frequency and quality of the data used. As an example, we calculate the log-returns of the Dow Jones Industrial Average (DJIA) index constituents and find that for the majority of stocks,

[1]There are other recent studies that also employ state-space models to extract the true efficient price from prices with microstructure noise, see for instance Menkveld et al. (2007), or to model the dynamics of the volatility skew Bedendo and Hodges (2009).

daily log-returns are lower than high-frequency (filtered and unfiltered) log-returns. We find that the difference in returns is not negligible; on average daily log returns are around 2.5% lower than log-returns calculated with a high frequency data base where the microstructure noise has been filtered out.

In the second example, we compute the equity beta in the Capital Asset Pricing Model (CAPM). We show that assets' systematic and unsystematic risk are different when we take into account the additional information provided by high frequency data. We calculate equity betas for the DJIA constituents over the period January 2005 to December 2006. On average the equity betas are approximately the same whether we use daily or filtered high frequency data. However, when we look at every individual stock the differences in the equity beta estimates are frequently significant. For example, the equity beta for Eastman Kodak using daily data is 1.071, but when filtered high frequency is employed the equity beta becomes 0.61.

The rest of the paper is organized as follows. Section 2 describes the framework for stock dynamics which we use in our volatility and covariation analysis. It also summarizes the existing literature on estimating the volatility of assets and highlights the difficulties that arise due to the existence of market microstructure noise. Section 3 proposes a Kalman filter-based approach to: estimate the volatility of log-returns; construct the de-noised true path of the assets; and finally, estimate the variance-covariance matrix (VCM) between multiple assets. Section 4 considers how to consistently estimate the total covariation between a large number of financial assets. Section 5 discusses how decision making in portfolio theory and measurements of cost of capital components, such as the equity beta in the CAPM framework, are affected by the frequency of the data set employed in the calculations. Section 6 concludes.

2. Background

Engle (2000) mentions that "one measure of progress in empirical econometrics is the frequency of data use". Initially one would expect that the consistency of estimators is improved by the resolution of data employed, that is, the higher the frequency the better the consistency. This is one of the main reasons why the current literature has focused on the use of financial data at higher frequencies, for instance tick-by-tick data. The study of volatility of stock returns in a high frequency setting is perhaps the most active exponent of this line of research. Below, we summarize the existing methods currently proposed in the literature to measure realized volatility based on

high frequency data for a diffusion model, which is the framework we will use in this paper.

Following Zhang et al. (2005), we assume that the log-price of a security follows a semimartingale process defined in $(\Omega, \mathcal{F}, \mathbb{P})$ given by

$$dX_t = \underbrace{\mu(X_t;\theta)dt}_{\text{drift component}} + \underbrace{\sigma dW_t}_{\text{diffusion component}}, \tag{1}$$

where $X_t = \ln S_t$ and S_t is the price of the security, $X_0 = 0$, $\mu(X_t;\theta)$ is the drift function with θ a drift parameter, $\sigma > 0$ is the diffusion coefficient and W_t is a standard Brownian motion under the probability measure $\mathbb{P}$. For our analysis, we assume that the object of interest σ is constant through each day and estimate it using high frequency data where it is possible to sample at different resolutions; the highest possible being tick-by-tick data. In our study the resolution is such that time-intervals between observations are short enough to make the drift component in (1) negligible. This is because the drift component $\mu(X_t;\theta)dt$ is of order dt, while the order of the diffusive component σdW_t is $dt^{1/2}$. Therefore, as $dt \to 0$ the drift term is much smaller than the diffusion. Hence we can assume that

$$X_t = \sigma W_t, \qquad t \in [0, T]. \tag{2}$$

Furthermore, we assume that the observations are equally spaced, so the time interval between them is constant and equal to Δ. The observations are recorded at times $t_i = i\Delta$ with $t_N = N\Delta = T$ for $i = 0, \ldots, N$.

The assumption that data are equally spaced is arguably not optimal if we consider actual tick-by-tick data. Generally, with financial data the time intervals between consecutive trades is not deterministic or constant. This fact has lead the literature to propose models such as the Autoregressive Conditional Duration (ACD) model,[2] introduced by Engle and Russell (1998), which uses the concept of GARCH models to capture the durations between trades. Several extensions of this ACD model have also been proposed in the literature. These include, the logarithmic-ACD model proposed by Bauwens and Giot (2000), which prohibits negative durations without the additional assumptions in the ACD model; the Exponential-ACD model of Dufour and Engle (2000); and the Threshold-ACD model of Zhang et al. (2001).

[2]The volatility research uses such models in the Ultra-High-Frequency GARCH model proposed by Engle (2000), a model that incorporates the time interval parameter into the volatility process given by a GARCH representation. The Ultra-High-Frequency GARCH model gives us volatility estimates at very small frequencies.

The variance of (2) can be obtained by

$$RV_X^{(all)} = [X, X]_T^{(all)} := \sum_{i=1}^{N} (X_{t_i} - X_{t_{i-1}})^2. \tag{3}$$

$RV_X^{(all)}$ is known as the realized variance of the log-price process and is equal to the sum of the squared log-returns X_t. The notation (all) means that we use all observations in the sample. This is the most efficient way to compute the σ^2 of the process given in equation (2). Therefore, $RV_X^{(all)}$ is the unbiased and consistent estimator of the variance that drives the stochastic differential equation for the log-price X_t.

However, as mentioned above the problem we must address is that at higher frequencies the existence of microstructure error causes difficulties when estimating the true variance of the efficient price X_t. In fact, the efficient price X_t is no longer observable due to the microstructure error, see Andersen et al. (2004) and Andersen et al. (2006). By microstructure error we mean: the bid-ask bounces; differences in trade sizes; and other sources that lead to price changes related to asymmetric information held by traders, etc.

The log-price in the presence of microstructure noise may be modeled by

$$Y_t = X_t + \varepsilon_t, \tag{4}$$

and using (3), its realized variance is calculated as

$$RV_Y^{(all)} = [Y, Y]_T^{(all)} = [X, X]_T^{(all)} + 2[X, \varepsilon]_T^{(all)} + [\varepsilon, \varepsilon]_T^{(all)}. \tag{5}$$

Here, Y_t is the observed log-price, equal to the true efficient log-price X_t plus an error term ε_t that captures all types of the microstructure error described above. This representation is consistent with microstructure models such as Easley and O'Hara (1992), where the arrival of information and the timing of trades affect the true price.

The object of interest here is $RV_X^{(all)}$, which differs from the result obtained using the observed values of the log-price Y_t, due to the existence of market microstructure noise. If we assume that the ε_ts in equation (4) are i.i.d. noise, with mean zero, variance σ_ε^2, independent of W_t, and so orthogonal to the efficient price X_t ($\varepsilon_t \perp X_t$), [3] Zhang et al. (2005) show that the realized variance

[3] In section 3.2 below we discuss the case where the noise term ε_t could be correlated to the true price X_t.

obtained from Y_t is dominated by the variance of the noise term, as follows

$$\mathbb{E}\left(RV_Y^{(all)}|X_t\right) = RV_X^{(all)} + 2N\mathbb{E}(\varepsilon^2), \tag{6}$$

$$Var\left(RV_Y^{(all)}|X_t\right) = 4N\mathbb{E}\left(\varepsilon^4\right) + O_p(1). \tag{7}$$

From equations (6) and (7) we can see that there is a tradeoff between quantity and quality of data. Equation (6) shows that the higher the sampling frequency N, the larger the bias of the estimator of the realized variance of the efficient log-price will be. And from (7) we see that the same holds true for the variance of our estimator which becomes more problematic as the sampling frequency increases especially at high frequencies.

2.1. Literature Review: non-parametric high frequency volatility estimation

In practice, using the whole data set at higher frequencies is not a common alternative for estimating the realized variance of asset log-returns. The preferred approach has been to sample at relatively low frequencies in order to reduce the bias introduced by the noise. Typically, daily realized variance is measured by sampling on an infrequent basis, for instance at 5-minute or at daily intervals. Assuming that the noisy signal is of the form (4), and that there are observations every second, the estimator we obtain for the volatility component is not as biased as the one that uses all the observations; however it is clear that it might not be the most efficient estimator. It is reasonable to expect that arbitrarily discarding data is not optimal.

By sampling at five-minute intervals the realized variance of the noisy price signal becomes:

$$RV_Y^{(5min)} = [Y,Y]_T^{(5min)} = \sum_{i=1}^{N_{5min}=78} (Y_{t_i} - Y_{t_{i-1}})^2. \tag{8}$$

In general, the estimator of volatility with infrequent or sparse data, i.e. when not all data are used, is given by

$$RV_Y^{(sparse)} = [X,X]_T^{(sparse)} + 2[X,\varepsilon]_T^{(sparse)} + [\varepsilon,\varepsilon]_T^{(sparse)}, \tag{9}$$

with

$$\mathbb{E}\left(RV_Y^{(sparse)}|X_t\right) = RV_X^{(sparse)} + 2N_{(sparse)}\mathbb{E}\left(\varepsilon^2\right). \tag{10}$$

It is clear that the bias in equation (10) is smaller than in the case where all observations are used in the estimation, $RV_T^{(all)}$, because $N_{(sparse)}$ is much smaller than N. For instance, for one day $N_{(sparse)} = 78$ while $N = 23,400$ when there is an observation every second. Moreover, the variance of the sparse estimator is given by:

$$Var\left(RV_Y^{(sparse)}|X_t\right) = 4N_{(sparse)}\mathbb{E}(\varepsilon^4) + \frac{2T^2\sigma^4}{N_{(sparse)}}. \tag{11}$$

Equations (10) and (11) show the effect that the sampling frequency has on the properties of the volatility estimator. On the one hand, (10) demonstrates that sampling at low frequencies decreases the bias of the estimator. On the other hand, (11) indicates that the variance of the estimator is either increasing or decreasing in $N_{(sparse)}$. This could explain why empirical studies have found that the proxy for the daily variance, based on daily observations, is more volatile than the estimator resulting from $RV_{(sparse)}$ at higher frequencies (see Andersen and Bollerslev (1998)).

Zhang et al. (2005) propose the use of the optimal sampling frequency

$$n^*_{sparse} = \left(\frac{T^2}{4\mathbb{E}(\varepsilon^2)^2}\sigma^4\right)^{1/3} \tag{12}$$

an outcome resulting from minimizing the mean squared error of the sparse estimator. This result appears in Bandi and Russell (2008) where they also provide an estimate of the second moment of the noise term, given by

$$\widehat{\mathbb{E}(\varepsilon^2)} = \frac{1}{2N}RV_Y^{(all)}. \tag{13}$$

Yet a more efficient estimator of the realized variance, than the one given by sampling at the optimal frequency n^*_{sparse} in equation (12), is obtained by subsampling the series over different grids, computing the realized variance at each grid, and then averaging the realized variances of the grids to derive the final estimator, given by

$$RV_Y^{(avg)} = \frac{1}{K}\sum_{k=1}^{K} RV_Y^{\mathcal{G}^{(k)}}, \tag{14}$$

where $RV_Y^{\mathcal{G}^{(k)}}$ is the realized variance for each grid and K is the total number of grids. This estimator is still biased but the bias is smaller than that of the sparse estimator, see Zhang et al. (2005).

Finally, the so-called first best estimator, in the non-parametric setting, of the log-price process variance is defined by Zhang et al. (2005) as

$$TSRV_Y = RV_Y^{(avg)} - \frac{\bar{N}}{N} RV_Y^{(all)}, \tag{15}$$

which is the biased, corrected estimator of our object of interest $[X, X]_T$, with $\bar{N} = \frac{N-K+1}{K}$ and where $TSRV$ stands for Two-Scale realized volatility estimator.

The application of the TSRV has been tested in empirical studies using both Monte Carlo simulations and financial data. Extensive research can be found in Aït-Sahalia and Mancini (2006) where they compare the $TSRV$ and $RV_Y^{(sparse)}$ for forecasting integrated variance. The methodology they use to compare the two estimators is based on the concept of *encompassing regression*. That is, one regresses the true volatility on two factors, where the second factor is derived from a subset of the information set of the first. We expect the second factor to have a small effect on the total explanatory power of the model. This is the case with the two estimators used in Aït-Sahalia and Mancini (2006). The $RV_Y^{(sparse)}$ is taken from a subset of the information set of the $TSRV$ and they show that it does not add any explanatory power to the volatility forecast. Therefore, $TSRV$ is the most efficient estimator in the non-parametric framework.[4]

2.2. Literature Review: parametric approach

In order to explore the similarities between the results of the parametric and non-parametric approaches, as well as to compare them with the results of our proposed approach, we now review the parametric approach proposed by Aït-Sahalia et al. (2005). This approach is based on the classical maximum-likelihood estimation method and provides us with fully efficient estimators in the Cramér-Rao sense.

Let us start with the case where no microstructure noise is present. Following the process of equation (2) we can define the process of the log-returns as

$$r_i = \sigma(W_{t_i} - W_{t_{i-1}}). \tag{16}$$

By the definition of the Brownian motion we know that the log-returns are i.i.d. $N(0, \sigma^2\Delta)$ with

[4]Other volatility estimators proposed by the literature are briefly mentioned in the Appendix.

$\Delta = t_i - t_{i-1}$. An efficient estimator for σ^2 could be obtained by maximizing the log-likelihood

$$l(\sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2\Delta) - \frac{r'r}{2\sigma^2\Delta}, \tag{17}$$

where $r = (r_1, \ldots, r_N)'$. The maximizer of $l(\sigma^2)$ is unbiased for σ^2 with variance equal to

$$Var(\hat{\sigma}^2) = \frac{2\sigma^4\Delta}{T}.$$

This result justifies why using all available data, without microstructure noise, is preferable; the variance of the estimator decreases as the time-step Δ shrinks. Using daily observations to estimate the volatility parameter will result in a very volatile estimator compared to the one that incorporates the information contained in the high frequency data. Therefore, in the absence of microstructure noise, the use of an estimator based on subsets of the available data will most certainly lead us to inaccurate conclusions.

When microstructure noise is present the framework does not change significantly. We assume that the log-price is given by equation (4) and that the noise term is normally distributed, $\varepsilon_{t_i} \sim N(0, \sigma_\varepsilon^2)$. In this case, the log-returns follow a moving average process of order one, $MA(1)$, since log-returns are given by

$$\begin{aligned} r_i &= \sigma(W_{t_i} - W_{t_{i-1}}) + \varepsilon_{t_i} - \varepsilon_{t_{i-1}} \\ &= \zeta_i + \eta\zeta_{i-1}, \end{aligned} \tag{18}$$

where the ζ_i's are i.i.d. $N(0, \gamma^2)$. Empirical evidence indicates that it is not always the case that stock returns exhibit the $MA(1)$ structure in the dynamics of returns, something that requires alterations in the likelihood function.[5] Here we assume that the stock returns follow an $MA(1)$ structure.

Using the relations

$$\gamma^2(1+\eta^2) = \sigma^2\Delta + 2\sigma_\varepsilon^2, \tag{19}$$

$$\gamma^2\eta = -\sigma_\varepsilon^2, \tag{20}$$

[5]Cases where the $MA(1)$ structure is not present requires a different specification of the microstructure noise term allowing for covariation between the noise ε_t and the asset price X_t, or the autocorrelation of the microstructure noise (relation between ε_t and its past values). Section 3.2 below describes how the Kalman filtering is applied in the case where the microstructure noise is correlated with the price process.

we can estimate σ^2, as well as the variance of the microstructure error, by estimating γ^2 and η. Since the log-returns follow an $MA(1)$ the log-likelihood function is given by

$$l(\eta, \gamma^2) = -\ln \det(V)/2 - N/2 \ln(2\pi\gamma^2) - \frac{1}{2\gamma^2} r' V^{-1} r, \tag{21}$$

where $\gamma^2 V$ is the covariance matrix of the returns and

$$V = \begin{pmatrix} 1+\eta^2 & \eta & 0 & \dots & 0 \\ \eta & 1+\eta^2 & \eta & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \eta \\ 0 & \cdots & 0 & \eta & 1+\eta^2 \end{pmatrix}.$$

An interesting feature that should be mentioned is that wrongly specifying the distribution of the noise term will not alter the asymptotic properties of $\hat{\sigma}^2$. The $\hat{\sigma}^2$ is fully efficient and we can also obtain an estimate for $\hat{\sigma}^2_\varepsilon$ which can be of interest and can lead us to further lines of research.

3. Efficient volatility estimators: filtering microstructure noise

Our aim here is to present an estimation of the realized variance of a diffusion process based on an unobserved components model. We will compare our results with the parametric and non-parametric estimators discussed above. The idea is motivated by the linear filtering problem where we have a Brownian motion plus other noise. Consider the case

$$\begin{aligned} dX_t &= \sigma dW_t, \\ Y_t &= X_t + \varepsilon_t \end{aligned} \tag{22}$$

where X_t is the unobserved true log-price, Y_t is the observed noisy price, σ is constant and W_t is a standard Brownian motion, independent of the Gaussian stochastic process ε_t. To filter out the microstructure noise from the observed high frequency data Y_t, we employ the discrete time 1-dimensional Kalman-Bucy filter and denote the de-noised series by $\widetilde{X}_t$, see Øksendal (2007) and Durbin and Koopman (2001).

Therefore, instead of working with the observable series Y_t, our approach is to work with $\widetilde{X}_t$ which we take as a proxy for the true efficient log-price X_t. Our method yields four interesting

results. First, it allows us to obtain consistent estimators for both σ^2 and the variance of ε_t. Second, it allows us to generate the state series, i.e. the true efficient log-price. Third, the extraction of the efficient log-price will allow us to better estimate the covariance matrix of several assets in a more efficient way than the method employed in the literature to date. Finally, we also obtain an estimate of the high frequency return which is not affected by the microstructure noise.

If we assume that we observe transactions at discrete times t_i, for $i = 1 \ldots, N$ where N is the total number of log-price observations, we can write the log-price process Y_t in discrete time as

$$Y_{t_i} = X_{t_i} + \varepsilon_{t_i},$$

where, over a time step $\Delta = t_i - t_{i-1}$,

$$\varepsilon_{t_i} \sim N(0, \sigma_\varepsilon^2), \tag{23}$$

and since X_t is driven by Brownian motion it can be represented, in discrete time, by the random walk

$$X_{t_{i+1}} = X_{t_i} + \eta_{t_i}, \text{ with } \eta_{t_i} \sim N(0, \sigma_\eta^2). \tag{24}$$

Our objective is to consistently estimate σ_η^2, which will give us the estimator for the variance of the true efficient price σ^2 and which is equal to σ_η^2/Δ.

Next, we use the filtering method by updating our information every time a new observation Y_{t_i} is brought into the information set $\mathcal{I}_{t_{i-1}} = \{Y_{t_1}, \ldots, Y_{t_{i-1}}\}$. We assume that the distribution of X_{t_i}, conditional on $\mathcal{I}_{t_{i-1}}$, is given by $N(\bar{X}_{t_i}, P_{t_i})$ where $\bar{X}_{t_i} = \mathbb{E}(X_{t_{i-1}}|\mathcal{I}_{t_{i-1}})$ and $P_{t_{i+1}} = Var(X_{t_{i+1}}|\mathcal{I}_{t_i})$. Given $\bar{X}_{t_i}$ and P_{t_i}, our aim is to calculate the next periods' values, $\bar{X}_{t_{i+1}}$ and $P_{t_{i+1}}$, when Y_{t_i} is observed. This can be easily done by using the usual regression theory as follows. First note that

$$\bar{X}_{t_{i+1}} = \mathbb{E}(X_{t_{i+1}}|\mathcal{I}_{t_i}) = \mathbb{E}(X_{t_i} + \eta_{t_i}|\mathcal{I}_{t_i}) = \mathbb{E}(X_{t_i}|\mathcal{I}_{t_i}) \qquad \text{since} \qquad \mathbb{E}(\eta_{t_i}|\mathcal{I}_{t_i}) = 0, \tag{25}$$

and

$$P_{t_{i+1}} = Var(X_{t_{i+1}}|\mathcal{I}_{t_i}) = Var(X_{t_i} + \eta_{t_i}|\mathcal{I}_{t_i}) = Var(X_{t_i}|\mathcal{I}_{t_i}) + \sigma_\eta^2. \tag{26}$$

Now define $u_{t_i} = Y_{t_i} - \bar{X}_{t_i}$ and $Var(u_{t_i}) = F_{t_i}$. Using the law of iterated expectations and the fact that $\mathbb{E}(u_{t_i}|\mathcal{I}_{t_{i-1}}) = 0$ we see that the unconditional expectation of u_{t_i} is equal to zero and

independent from Y_{t_i}. We also have that[6]

$$\mathbb{E}(X_{t_i}|\mathcal{I}_{t_i}) = \mathbb{E}(X_{t_i}|\mathcal{I}_{t_{i-1}}, u_{t_i}) = \mathbb{E}(X_{t_i}|\mathcal{I}_{t_{i-1}}) + Cov(X_{t_i}, u_{t_i})\, Var(u_{t_i})^{-1} u_{t_i}. \tag{27}$$

It is easy to show that the covariance between u_{t_i} and the efficient log-price X_{t_i} is equal to $Cov(X_{t_i}, u_{t_i}) = \mathbb{E}(\,Var(X_{t_i}|\mathcal{I}_{t_{i-1}})) = P_{t_i}$ and $Var(u_{t_i}) = F_{t_i} = Var(X_{t_i}|\mathcal{I}_{t_{i-1}}) + Var(\varepsilon_{t_i}) = P_{t_i} + \sigma_\varepsilon^2$. Therefore, equation (27) can be rewritten as

$$\mathbb{E}(X_{t_i}|\mathcal{I}_{t_i}) = \bar{X}_{t_i} + K_{t_i} u_{t_i}, \tag{28}$$

where $K_{t_i} = P_{t_i}/F_{t_i}$. Combining equations (25) and (28) we get the recursive relation

$$\bar{X}_{t_{i+1}} = \bar{X}_{t_i} + K_{t_i} u_{t_i},$$

which constitutes our Kalman filter together with the following equations:

$$\begin{aligned} u_{t_i} &= Y_{t_i} - \bar{X}_{t_i}, & K_{t_i} &= P_{t_i}/F_{t_i}, \\ P_{t_{i+1}} &= P_{t_i}(1 - K_{t_i}) + \sigma_\eta^2, & F_{t_i} &= P_{t_i} + \sigma_\varepsilon^2. \end{aligned} \tag{29}$$

We can estimate the parameters σ_ε^2 and σ_η^2 in two ways. The first approach is based on the maximization of the likelihood function which is equal to the joint density of our observations $\{Y_{t_1}, \ldots, Y_{t_N}\}$. The loglikelihood function is given by

$$l(\theta) = -\,(N/2)\ln(2\pi) - 1/2 \sum_{t=1}^{N} \left(\ln F_{t_i} + \frac{u_{t_i}^2}{F_{t_i}} \right) \tag{30}$$

where $\theta = (\sigma_\eta^2, \sigma_\varepsilon^2)'$ is the unknown parameter vector. Usually, it is more convenient to maximize with respect to the log values of σ_ε^2 and σ_η^2. Therefore, maximizing over $\phi_\varepsilon = \ln \sigma_\varepsilon^2$ and $\phi_\eta = \ln \sigma_\eta^2$, the estimator for the variance of the process in the stochastic differential equation (1) is given by

$$\hat{\sigma}^2 = \exp(\hat{\phi}_\eta^2)/\Delta \tag{31}$$

where $\hat{\phi}_\eta^2$ is the MLE of ϕ_η^2.

Alternatively, the second approach, and the one we use here, estimates the vector $\theta = (\sigma_\eta^2, \sigma_\varepsilon^2)'$

[6] Using Lemma 1 in Section B of the Appendix.

using the Expectation-Maximization (EM) algorithm proposed by Dempster et al. (1977).[7] Since the EM algorithm converges to a local maximum, instead of a global one, the choice of initial values in the algorithm is important. In the simulations section below, we use the EM algorithm and use as initial values: σ^2 resulting from the $TSRV$; and σ_ε^2 resulting from (13). More details about the algorithm can be found in Harvey (1990).

With the estimates $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\eta^2$ and the state series of the log-price given by the Kalman recursive equations (29) we can obtain the de-noised series of observed log-price process, which we denote $\widetilde{X}_t$. This is an important step in our procedure since we can estimate the variance σ^2 of the true price process X_t by calculating $[\widetilde{X}, \widetilde{X}]_T^{(all)}$. This allows us to obtain an estimator of the true variance which is numerically close to the $\hat{\sigma}^2$ (given in (31)), as they have the same RMSE value. Moreover, it is very useful to base calculations on $\widetilde{X}_t$ when we face the high dimensionality problem which results from estimating the VCM of several assets. For instance, the de-noised series $\widetilde{X}_t$ can be used to derive efficient estimates of the VCM of asset returns in a straightforward way that will be described below.

The estimators for both microstructure error and efficient price series variances are fully efficient in the Cramér-Rao sense criterion because they result from the maximum likelihood estimation given by equation (30). Hence, this property makes our proposed approach ideal for empirical high frequency analysis. It provides a great deal of flexibility in efficiently estimating the VCM between assets, when we face high-dimensionality problems, and in calculating the de-noised high-frequency returns.

To summarize, there are three points where our approach can enhance the financial high frequency analysis. First, we can estimate the variance of the microstructure noise. Second, we can estimate the variance of the efficient series X_t. (Note that these two estimators are as efficient as those proposed in Aït-Sahalia et al. (2005)). Third, we can generate the state series $\widetilde{X}_t$, which is a proxy of the real price process X_t. This third point is the one that connects the findings of the univariate with the multivariate case. It allows us to generate efficient estimates of the VCM when a large number of assets is present. The estimates we obtain play a fundamental role when we look at applications such as risk and portfolio management and, in addition to calculating VCMs, we also achieve de-noised high frequency returns.

[7] The EM algorithm has been used in the literature to estimate unknown parameters in unobserved components models, such as the one described by equation (23) (for instance, see Watson and Engle (1983)).

3.1. Simulation results

At this point we use equation (4) to simulate noisy price paths so that we can assess the properties of the estimators of the volatility of X_t, when microstructure noise is present. We draw comparisons from two perspectives. First we compare six volatility estimators, including our Kalman-based estimator, by looking at the RMSE for each one. The estimators we compare are: [I] $RV_T^{(all)}$, [II] $RV_T^{(sparse)}$, [III] $RV_T^{(avg)}$, [IV] $TSRV$, [V] MLE and [VI] Kalman. By comparing the RMSEs of all estimators we provide evidence that, in terms of efficiency, the Kalman-based estimator we propose is better than the non-parametric estimators ([I] to [IV]) and is close to the MLE estimator.

The second method of comparison allows us to consider the validity of calculating estimates of volatility of log-returns using the de-noised price series $\widetilde{X}_t$. As it was argued above, the classical and straightforward estimators of volatility fail because microstructure noise, at the high frequency level, swamps the true volatility estimates. The classical estimators, for instance $RV_T^{(all)}$, are designed to obtain efficient and unbiased estimates, but are not designed to cope with microstructure noise. Our Kalman methodology however, not only delivers an estimate of the volatility of the true efficient log-returns when the data contain microstructure noise, but also enables us to generate a proxy for the true efficient log-price series. Hence, using $\widetilde{X}_t$ as the log-price series we propose to estimate the volatility of the true log-returns by calculating $[\widetilde{X}, \widetilde{X}]_T^{(all)}$. Secondly, by looking at its RMSE, we can gauge how efficient this estimator is relative to estimators [I] to [VI]. We label this approach as the 'Filtered estimator'.

Before examining the comparison of high-frequency estimators it is important to stress that, in terms of efficiency, estimators based on daily data could perform worse than the high frequency estimators [I] to [VI]. Figure 1 helps us to illustrate this point and further justifies the analysis that follows. It presents the difference between the MLE fully efficient estimator, denoted by $\hat{\sigma}_{MLE}$, and the volatility estimator based on daily observations, denoted by $\hat{\sigma}_{daily}$. The latter is calculated by employing (3) and using daily closing prices, for British Petroleum (BP) over 2003. The $\hat{\sigma}_{daily}$ is much more volatile than $\hat{\sigma}_{MLE}$. And this is consistent with the results presented by Andersen and Bollerslev (1998), where the realized volatility based on intra-daily data is less volatile than the volatility based on closing prices. In the same figure we see that $\hat{\sigma}_{daily}$ could give us estimates over a range $[0\%, 75\%]$ for the level of volatility expressed on a yearly basis while $\hat{\sigma}_{MLE}$ has a maximum value of 35.46% which seems more plausible for the period under examination. We repeat the exercise for MSFT (not depicted in the figure) and find that the results for $\hat{\sigma}_{daily}$ are worse over

the same period. In some cases, the volatility estimate is more than 70% while the $\hat{\sigma}_{MLE}$ varies from 10% to 40%. Crucial to these findings is that for daily observations we only take into account the 252 observations for each working day in 2003, as opposed to the total population of $623,759$ observations available for the whole year, i.e. estimates based on daily data employ less than 0.05% of the total information.

We now focus on the comparison of the six high frequency volatility estimators using Monte Carlo simulations to produce price paths using equation (4). We do this to examine in more detail which high frequency estimator is the most efficient for empirical research. In the simulations, each day has $23,400$ observations, equivalent to the total number of seconds in one trading day, and we assume that $\sigma_\varepsilon^2 = 1 \times 10^{-6}$, $\sigma^2 = 0.09$, $\Delta = 1/23,400$ and the starting value of the asset is $S_0 = 30$, so $X_0 = \ln S_0 = 3.4012$. We implement this procedure $1,000$ times and vary the value for σ_ε^2 to verify the robustness of the results. The different values of σ_ε^2 do not alter the main findings which we describe below.

The results of our first comparison of the different estimators are summarized in Figure 2 where we can see that the Kalman-based approach exhibits the lowest RMSE of the six estimators. From Figure 2 it is clear that the $RV_T^{(all)}$ applied to the noisy signal Y_t yields the worst estimator of the true σ^2 (its value, on average, is equal to 0.1370). This estimator has the largest RMSE; a result in agreement with Zhang et al. (2005). Therefore, naively using all observations at high frequencies is undesirable if the objective is to obtain consistent estimators of volatility. The microstructure noise, which is present and relevant at these frequencies, leads us to an estimator with an excessively high RMSE.

The sparse sampling estimator [II], where we sample every 5 minutes, is again very misleading relative to the true variance. Recall that this estimator does not take the whole data set into account and this is why the estimate of the true value of the variance of the log-returns is very poor.

The next estimator we employ is the $RV_T^{(avg)}$, where we recall that it uses all available observations. Figure 2 denotes the RMSE of this estimator by RV_{grid}. We can observe that this estimator has a smaller RMSE than the previous two, but we know from the theoretical analysis in Section 2 that we can do better by using the $TSRV$ estimator as shown in Figure 2 .

Finally, the result we highlight is that the Kalman filter approach also has a small RMSE, which is very close to the RMSE of the MLE estimator. The Kalman estimator is fully efficient in the

maximum likelihood estimation method sense and it can be used to find best estimators for the variance of the efficient log-price, given the existence of microstructure noise.

The results of our second method of comparison are shown in Figure 3. Here we see more clearly the difference between the RMSE of the $TSRV$, which is the 1st best non-parametric estimator in Zhang et al. (2005), and the RMSEs of the MLE and Kalman based estimators. The $TSRV$ exhibits greater RMSE than that of the MLE and Kalman. Furthermore, the figure also shows the RMSE of the Filtered estimator $[\widetilde{X}, \widetilde{X}]_T^{(all)}$ which is as small as that of the MLE estimator (its mean value over the total number of simulations is equal to 0.0902). This is an important result because it is a strong indicator that $\widetilde{X}_t$ is actually a good proxy for the real log-price.

3.2. Extension

Before going into the high-dimensional volatility estimation problem we briefly analyze the case where the microstructure noise is correlated with the price process and discuss the assumption of constant volatility throughout the trading day.

There are both empirical evidence and theoretical models, see for instance Glosten (1987) and Hansen and Lunde (2006), where the microstructure process ε_t could be correlated with X_t. The Kalman approach provides ample flexibility in this setting and it has to be modified to estimate the extra parameter responsible for the correlation between the two processes. In this case the filtering equation is modified in the following way. First assume that

$$\mathbb{E}(\eta_t \varepsilon_s) = \begin{cases} G_t, & t = s \\ 0, & t \neq s. \end{cases} \tag{32}$$

Now the variance of the u_{t_i}, F_{t_i} in equation (29), has an extra component coming from the correlation of the price and microstructure noise. This gives the following relation:

$$Var(u_{t_i}) = F_{t_i} = Var(X_{t_i}|\mathcal{I}_{t_{i-1}}) + Var(\varepsilon_{t_i}) + 2Cov(X_{t_i}, \varepsilon) = P_{t_i} + \sigma_\varepsilon^2 + 2G_{t_i}. \tag{33}$$

And, since

$$Cov(X_{t_i}, u_{t_i}) = Cov(X_{t_i}, X_{t_i} + \varepsilon_{t_i} - \bar{X}_{t_i}) = P_{t_i} + G_{t_i} \tag{34}$$

we can write

$$K_{t_i} = \frac{P_{t_i} + G_{t_i}}{F_{t_i}}. \tag{35}$$

Therefore the equation for $P_{t_{i+1}}$, also known as *Riccati equation*, becomes

$$P_{t_{i+1}} = P_{t_i} - \frac{(P_{t_i} + G_{t_i})^2}{F_{t_i}} + \sigma_\eta^2. \tag{36}$$

The unknown parameters σ_η^2, σ_ε^2 and G_{t_i} can be estimated by the maximum likelihood approach described above, see equation (30). The likelihood function (30) is now given by modified updating equations described in this subsection (equations (33)-(36)). An alternative method to estimate the parameters is to modify the system in such a way that the new one will have uncorrelated disturbances. For further details we refer to Harvey (1990) and Chan et al. (1984).

We note that as in Aït-Sahalia et al. (2005) we also assume that volatility is constant throughout the day; an assumption supported by the findings in Oomen (2006). However, it is a well known fact that volatility exhibits an intra-day pattern: higher during the opening and closing hours ($\approx$ 9.30 - 11.00am and $\approx$ 15.00 - 16.00pm respectively) and lower during the rest of the trading day. In principle it is possible to incorporate this volatility U-pattern in a state-space setup by allowing, for example, for a smooth trend in volatility (see Harvey (1990) for more details). However, unless one puts restrictions on such a specification the volatility will be stochastic, and this does not fit into our overall framework. Alternatively, one might think of specifying volatility as a nonlinear deterministic function of time capable of producing this U-shape. An idea on how this could be implemented could be found in Wood et al. (1985).

4. Estimating the variance-covariance matrix of multiple assets in the presence of microstructure noise

The univariate volatility estimators can be extended to the case of multiple assets. Of course, for a portfolio or a risk manager the object that plays an important role is not just the volatility of one asset, but the VCM of the assets in the portfolio. In the case of several assets, the above results in a univariate set-up can be used to obtain the diagonal elements of the VCM by using one of the above estimators. Since we examine the case where the volatility is constant through the day, but can change from day to day, the best way of estimating the diagonal elements of the VCM,

which we denote by Σ_t, is by using the MLE or the Filtered estimator, that is, to work with the de-noised log-price series $\widetilde{X}_t$. Furthermore, we also have to estimate the off-diagonal elements of Σ_t, not a straightforward task given to the large number of those elements.

The market is now given by an $\mathcal{F}_t^{(n)}$-adapted $(n+1)$-dimensional Itô process

$$X(t) = (X_0(t), X_1(t), \ldots, X_n(t)) \text{ with } 0 \leq t \leq T,$$

where $X_0(t)$ is the log-risk-free asset,

$$dX_j(t) = \mu_j dt + \sum_{k=1}^{n} \sigma_{jk} dW_k(t), \text{ for } j = 1, \ldots, n, \tag{37}$$

and $W_k(t)$ is an n-dimensional standard Brownian motion.

The aim is to estimate the VCM given by

$$\Sigma_t = Var(X_j(t)) = \mathbb{E}(X_j(t)X_j(t)') = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{pmatrix}, \text{ when } \mu_j = 0, \forall j. \tag{38}$$

A crucial problem when using financial data is the so-called *non-synchronized* data problem. This is the difference between the frequency of transactions for different assets. It is clear that a very liquid asset, for instance Microsoft (MSFT), has, on average, observations every 2 seconds, while BP, or other less liquid assets, have fewer observations throughout the trading day. Moreover, these observations may occur at completely different times. In our analysis we are using minutely observations, taking the last observation of each minute for each asset.

One of the assumptions in our analysis is that volatility is constant on a daily basis. This assumption allows us to use the more efficient estimator, MLE or Kalman, instead of the $TSRV$ non-parametric estimator. The stochastic volatility case can be explored by using the $TSRV$ estimator but the complexity increases for the off-diagonal elements of the Σ_t defined in (38) due to multiple grids that appear for each element of the two assets.

In the case where microstructure noise is not present, the efficient estimators for the diagonal

elements of Σ_t are given by

$$\sigma_j^2 = RV_{X_j}^{(all)} = \sum_{i=1}^{N}(X_j(t_i) - X_j(t_{i-1}))^2, 1 \leq j \leq n,$$

and for the off-diagonal covariances of the assets j_1 and j_2

$$\sigma_{j_1 j_2} = \sum_{i=1}^{N}(X_{j_1}(t_i) - X_{j_1}(t_{i-1}))(X_{j_2}(t_i) - X_{j_2}(t_{i-1})), \text{ for } j_1 \neq j_2, \tag{39}$$

see Hayashi and Yoshida (2005) and Zhang (2006).

But again, for higher frequencies, the microstructure noise also affects the estimators for the covariances of assets. This can be seen in the following equation, which gives the covariance between two assets in the presence of microstructure noise.

$$\begin{aligned}
\sigma_{j_1 j_2} &= \sum_{i=1}^{N}(Y_{j_1}(t_i) - Y_{j_1}(t_{i-1}))(Y_{j_2}(t_i) - Y_{j_2}(t_{i-1})) \\
&= \sum_{i=1}^{N}(X_{j_1}(t_i) - X_{j_1}(t_{i-1}))(X_{j_2}(t_i) - X_{j_2}(t_{i-1})) \\
&\quad + \sum_{i=1}^{N}(X_{j_1}(t_i) - X_{j_1}(t_{i-1}))(\varepsilon_{j_2}(t_i) - \varepsilon_{j_2}(t_{i-1})) \\
&\quad + \sum_{i=1}^{N}(X_{j_2}(t_i) - X_{j_2}(t_{i-1}))(\varepsilon_{j_1}(t_i) - \varepsilon_{j_1}(t_{i-1})) \\
&\quad + \sum_{i=1}^{N}(\varepsilon_{j_1}(t_i) - \varepsilon_{j_1}(t_{i-1}))(\varepsilon_{j_2}(t_i) - \varepsilon_{j_2}(t_{i-1})).
\end{aligned} \tag{40}$$

We are interested in estimating the first term in equation (40). However, the remaining terms in (40) make the covariance estimator biased-inefficient. In the literature, we can find estimates of the whole matrix using the $TSRV$ for the diagonal elements and co-variation using data from an infrequent time interval of 5 minutes. These estimates are not the best, as shown in the simulations section above. Here we use the proposed method of Kalman filtering to derive efficient estimates, with very small RMSEs, for both the diagonal, and the off diagonal elements of the VCM.

Using the same arguments as in the univariate case, a covariation estimate based on sparse sampling would not be the most efficient one because we arbitrarily discard information from both

assets to estimate their covariation. Therefore, we propose to use a framework where we choose the most efficient estimators in order to obtain the efficient VCM in the presence of noise. Our assertion is that we can do better than the non-parametric estimators (see for instance Wang et al. (2007)), in terms of efficiency, by using the Kalman-based estimator and the state series $\widetilde{X}_j(t)$, as defined and computed for the univariate case in Section 3.

The framework we propose is a two-step procedure. First, we apply the Kalman filter to the n assets to generate the filtered $\widetilde{X}_j(t)$ series ($j = 1, \ldots, n$). The second step applies standard estimators to $\widetilde{X}_j(t)$.

In this way we calculate the off-diagonal elements of Σ_t using

$$\sigma_{j_1,j_2} = \sum_{i=1}^{N} (\widetilde{X}_{j_1}(t_i) - \widetilde{X}_{j_1}(t_{i-1}))(\widetilde{X}_{j_2}(t_i) - \widetilde{X}_{j_2}(t_{i-1})), \tag{41}$$

which delivers the efficient realized co-volatility between assets j_1 and j_2 in the presence of microstructure error. Furthermore, the diagonal elements of the VCM can be efficiently estimated by applying either the MLE or the Filtered estimator. This estimate of Σ_t is the most efficient we can obtain and it is more efficient than the estimators proposed in the literature.

Once we have obtained an efficient estimation of the VCM, i.e. a measure of $\widehat{\Sigma}_t$ not affected by the microstructure noise, we use it to study the effect that microstructure noise has on quantities and parameters used in finance, such as portfolio theory and calculations of the cost of equity in the CAPM.

Instead of the two-step approach we propose here for the estimation of the VCM, one could also perform the filtering and estimation in one step, see Menkveld et al. (2007). We use the simpler two-step approach because of the dimensionality of the problem (we deal with a VCM for 28 assets) and because in our framework we have assumed that: the microstructure noise is not correlated with the price innovations, it is not autoregressive, and microstructure noise is not correlated across different stocks.

5. Empirical results: portfolio efficient frontier and equity beta

In this section we compare how the estimate of parameters used in portfolio theory and in the CAPM can differ depending on the resolution of data employed in the estimation. In particular,

we show how the location and shape of the efficient frontier of a portfolio of assets is affected by the use of low and high frequency data (with and without microstructure noise). Furthermore, we calculate equity betas for a number of firms and see how the frequency and quality of the data impinges on the results.

5.1. Efficient frontier

To calculate the efficient frontier for a portfolio of n assets, we need the $n \times n$ VCM and the returns of the assets. Depending on the data we employ, there are many ways of calculating the VCM and the returns used in the applications below. Some of the choices are: filtered and unfiltered high frequency and low frequency data. To differentiate the four cases that arise from the data used to calculate the VCMs, we denote them by:

- $\widehat{\Sigma}_{t,daily}$ the VCM resulting from the traditional approach, where daily observations are employed.
- $\widehat{\Sigma}_{t,filtered}$, and refer to it as the "Filtered-Kalman", the VCM resulting from working with de-noised price series $\widetilde{X}_j(t)$,
- $\widehat{\Sigma}_{t,HF}$, and refer to it as the "unfiltered high frequency",
- Finally, $\widehat{\Sigma}_{t,5min}$, and refer to it as the "high frequency 5-min midquotes", the VCM resulting from using high frequency midquotes at a 5-min frequency.

Similarly, the calculation of returns can also be done by either using daily closing prices or employing the high frequency data set.

We present two examples. The first one calculates the efficient frontiers and optimal portfolio decisions with two assets and focuses on the difference in results arising from using $\widehat{\Sigma}_{t,daily}$, $\widehat{\Sigma}_{t,filtered}$ and returns calculated on daily and filtered high frequency basis. The second example uses the constituents of the DJIA and compares efficient frontiers and optimal portfolio decisions as in the first example, but includes the cases arising from using unfiltered or noisy high frequency data, i.e. employ $\widehat{\Sigma}_{t,HF}$ in the calculations, and the 5-min midquote data.

Example 1: Portfolio management with 2 assets

In this example we look at MSFT and Intel Corporation (INTC) over the period March 1 to June 30 2007. Our data set consists of tick-by-tick and daily closing prices for both assets. The former

data set is obtained from NYSE Trades and Quotes (TAQ) and the latter from the Centre for Research in Security Prices (CRSP) database.

A practical problem we face when using the high frequency data is that, although very seldom, there are one minute intervals throughout the day where there are no trades. In these cases we do not have 390 observations per asset per day, that is an observation for every minute from 9.30am to 4.00pm. To overcome this difficulty we prepare the data by filling the missing observations with the price of the preceding observed trade. For example, if over the period 9:35am to 9:36am there was no trade, we assume that there was a trade and take the price of the trade used for the slot 9:33am to 9:34am.[8] MSFT and INTC are very liquid assets so the number of missing values is very small relative to the total number of observations.[9]

To understand the effect of the high frequency data on the VCM and returns, we first show two efficient frontiers, one based on $\widehat{\Sigma}_{t,filtered}$, which uses data every minute, that is $33,150$ observations for each asset, and the other one $\widehat{\Sigma}_{t,daily}$, and in both cases the returns vector was calculated using daily observations. We then repeat the exercise with the same two VCMs, but use the returns vector resulting from employing the filtered high frequency data set.

Following the formulas for the realized variance and the covariation of two assets, we have

$$\widehat{\Sigma}_{t,daily} = \begin{pmatrix} 0.029255 & 0.014126 \\ 0.014126 & 0.045429 \end{pmatrix}, \tag{42}$$

while the estimator of the covariance matrix, based on filtered high frequency data, is equal to

$$\widehat{\Sigma}_{t,filtered} = \begin{pmatrix} 0.033651 & 0.012790 \\ 0.012790 & 0.046540 \end{pmatrix}. \tag{43}$$

It is worth highlighting the difference in the diagonal elements of the VCMs. Not only do the diagonal elements of the matrix differ, but also the off-diagonal elements of the two matrices are different. This is a crucial result that will have an effect on the trade-off between return and risk. We know that the estimator based on daily data is not efficient because we discard a great amount of information, whereas $\widehat{\Sigma}_{t,filtered}$ is more efficient and will give us a better view of the true

[8] When there is more than one trade within a minute, as is usually the case in liquid stocks, we take as observation the last trade before the end of that minute.

[9] When we look at a portfolio whose constituents are those of the DJIA index, the number of missing values is greater for less liquid assets, but still small enough so that the results are not affected. In fact the case where most data was missing only accounted for 350 of the total sample of 33,150.

covariation between assets. Recall that the covariation is calculated according to equation (41), where we have filtered out the microstructure noise and we have used data at higher frequencies in order to get better estimators for all the elements of the VCM.

Figure 4 depicts the efficient frontiers resulting from daily and filtered high frequency data together with returns calculated from daily observations. It is clear that the efficient frontier generated from the filtered high frequency data differs from that based on daily data, a direct effect of the estimates of the covariance matrix, and crucially, as a result of the disparity in the covariance between MSFT and INTC. This has an impact on the decisions that a portfolio manager should take. We see that returns are overestimated for the same level of risk. For instance, with a 17% risk, we face 42% return based on daily data, while the filtered high frequency data yields a return under 41%.

If we take into account the returns that we get from the filtered high frequency data, i.e. $r_t = \widetilde{X}_t - \widetilde{X}_{t-1}$, the filtered high frequency efficient frontier takes the form shown in Figure 5. Compared to Figure 4, we have an upward parallel movement of the efficient frontier based on data at higher frequencies. This is because when filtered high frequency data is employed to calculate returns we get: $r_{MSFT} = 16.67\%$ and $r_{INTC} = 58.67\%$, whereas, on the other hand, the return based on daily data yields $r_{MSFT} = 14.39\%$ and $r_{INTC} = 57.64\%$.

Now we give an example of how these differences in returns per level of risk, arising from the granularity of the data, will affect the decisions and strategies taken by the management of a financial institution. Suppose that the management aims at a minimum 30% return by investing in a portfolio of two assets: MSFT and INTC. Therefore the optimization problem reduces to minimizing the portfolio's risk, keeping the return fixed, by changing the weights of the two assets in the portfolio.

Formally, the manager solves the following optimization problem:

$$\begin{array}{rcccl} \min & \sigma_p^2 & = & w'\Sigma w & \text{w.r.t. } w \\ \text{subject to} & r_p & = & w'r_i \geq 30\%, & \text{for } i = 1, 2, \end{array} \tag{44}$$

where r_p is the return of the portfolio, calculated with either daily or filtered high frequency data according to the case we examine, σ_p is the portfolio risk and $w' = (w_1, w_2)$ are the portfolio weights of the two assets. We assume that short selling is not permitted, that is $w \geq 0$.

We exemplify the difference in results depicted in Figure 5 by looking at four portfolios with

different return-risk targets, see Table 1. For instance, Portfolio I shows that a 30% return target can be achieved by bearing 15.66% (resp. 15.64%) volatility when daily (resp. filtered high frequency) data are used. Similarly, if the return target is 50% as in Portfolio IV, the volatility borne by the investor employing daily data is 18.92% as opposed to 18.56% when filtered high frequency is employed. Note that in all cases the portfolio risk when 5-minute midquote data are used is higher than the Filtered-Kalman. This result is in line with the Monte Carlo findings above, where it is shown that the estimates from the sparse sampling approach are biased by the microstructure noise.

Furthermore, when using a high frequency analysis, in portfolio IV we see that by investing 20.65% in MSFT and 79.35% in INTC produces a return of 50.00% return with volatility 18.56%. The computations of the efficient frontier for daily data have shown that the same weights produce the same risk but the return is equal to 48.7% which is lower than the high frequency return. The importance of this result is clear because using daily analysis in a large portfolio of a fund would lead to a loss of 1.3% return by investing in assets in a non-optimal way. The combination of the above weights leads the manager to take a smaller 'true risk' (the high frequency variance of 18.92%) than the one she had computed 19.15%.

Example 2: Portfolio management with DJIA stocks

The 2×2 case described above is interesting, but one must ask three further questions. First, can the message be extended to larger portfolios? Second, what happens to the efficient frontier when a relatively large number of assets are included? Third, what is the impact of employing high frequency data but without filtering the microstructure noise? From equation (40) we see that the impact of the microstructure noise in the 2×2 asset case is present and cannot be neglected; an undesirable effect also present in the general $n \times n$ case. Hence, to illustrate the impact of the microstructure noise in higher dimensional portfolios we solve the same portfolio program as above, but with the DJIA constituents as the investment opportunity set.[10] The dataset examined here refers to the same time period as the one used above, i.e. March 1 to June 30 2007.

Table 2 shows the stocks used in our example. The table presents the volatility and the return for each stock based on daily data and on filtered, unfiltered and 5-min midquote data. We use 28 out of the total 30 constituents of the DJIA because there are two stocks for which our data base

[10]Note that for simplicity we have not introduced a risk-free asset here.

does not have a clean record.

From Table 2 it is clear that there are differences both in returns and in the volatilities for all assets. The differences between the filtered high frequency data and daily data (as well as noisy high frequency data) are considerable. For example, for Alcoa we see that the difference in the estimates of mean returns and risk can be of up to 300 basis points for the returns and of up to 600 basis points for the risk estimates. These differences are not as pronounced when we compare the results using the filtered high frequency data with those of the 5-min midquotes. For example, the difference in the risk estimate of Alcoa using the filtered high frequency data and the 5-min midquotes is of around 16 basis points. Although this difference may seem small compared to the difference obtained when daily or noisy high frequency data are employed, below we calculate the investment efficient frontier using the constituents of the DJIA and show that the relatively small difference in basis points that we find on a per asset basis in Table 2 (for the filtered and 5-min midquotes) becomes of economic significance when filtered high frequency data or 5-min midquotes are employed in the calculations of the efficient frontier.

Proceeding as in the 2×2 case we calculate the 28×28 VCM. Tables 5 to 8 in the appendix show that the filtered high frequency correlations are different from those obtained when daily data are used. This difference in the elements of the correlation matrices has an effect on the efficient frontiers.Figure 6 depicts four efficient frontiers that were calculated using $\widehat{\Sigma}_{t,daily}$ with daily returns; $\widehat{\Sigma}_{t,filtered}$ with filtered high frequency returns; $\widehat{\Sigma}_{t,HF}$ with unfiltered high frequency returns; and $\widehat{\Sigma}_{t,5min}$ with 5-min midquotes. It is interesting to point out that the difference in the efficient frontier, between the daily and filtered high frequency frontiers, is similar to that obtained in the 2×2 case. The unfiltered high frequency efficient frontier gives, for a given return target, portfolios with much higher volatility levels than the other two frontiers. For instance, aiming at a 30% return the efficient frontier based on daily data has a risk of around 8.5%, with filtered high frequency data we get a volatility almost equal to 9.4%, and with noisy high frequency observations the volatility is much higher, close to 10.7%.

Further study of Figure 6 shows that the daily and filtered high frequency efficient frontiers cross each other at close to the 10% risk mark. For lower levels of risk, daily data underestimate the risks borne by an investor targeting returns below 4.5%. On the other hand, for risk levels greater than 10% we see that the portfolio return based on daily data underestimates the returns that can be obtained by certain combinations of the underlying asset. To examine this point further, we compare these two efficient frontiers in Figure 7 by plotting the difference in returns for the

same level of risk. To obtain this difference we fit the two efficient frontiers to ratios of polynomial functions so that we can generate return values for same levels of volatility.[11] The figure clearly shows that for risk levels above 10% the returns obtained using the filtered high frequency analysis are higher than those based on daily data. The portfolio manager therefore underestimates the return of her portfolio for a great range of volatility levels, something that, in addition to the results in Table 1 regarding the optimal portfolio weights, shows the importance of high frequency analysis in determining decision making and in particular the use of filtered high frequency data.

Figure 8 shows the efficient frontiers calculated when the filtered HF and the 5-min midquotes returns are employed. We see that for levels of portfolio risk below 16% the frontier calculated with filtered data is to the left of the frontier based on 5-min midquotes. This result is in agreement with the Monte Carlo analysis shown above where the estimates of the volatility based on the Kalman filter are more efficient than those based on the sparse sampling approach (every 5-minutes) when dealing with microstructure noise. The difference in the level of risk (for a given level of return) between the two frontiers is attributed to the bias of the microstructure noise that has not been sufficiently filtered out by the sparse sampling approach. Furthermore, Figure 9 shows the difference in returns for the same level of risk when filtered high frequency data and 5-min midquotes are employed; we observe that the difference is also of economic significance.

5.2. Equity beta

Now we focus on the computation of equity betas based on both daily and high frequency observations. We look at DJIA stocks over the period 03 January 2005 to 29 December 2006. We implement the CAPM framework for 24 stocks for which we have a clean record over this period. We compare the equity betas based on daily observations, which are extensively used by the literature, with the equity betas we obtained from the filtered and unfiltered high frequency analysis.

Data are obtained from both CRSP and TAQ databases, as well as Fama and French Research Portfolios and Factors database. In order to compute the equity beta, we perform the following regression for all stocks

$$r_{i,t} = r_{rf,t} + \beta_i(r_{Mkt,t} - r_{rf,t}) + e_{i,t} \tag{45}$$

where $r_{i,t}$ is the return for each stock i, $r_{rf,t}$ the risk-free rate given by the One-Month Treasury Bill rate, $r_{Mkt,t}$ the return on the market portfolio and $e_{i,t}$ an error term. The equity beta could also

[11] See appendix for details.

be computed as $\beta_i = Cov(r_i, r_{Mkt})/Var(r_{Mkt})$. We implement equation (45) in order to examine not only the differences in betas, but also the possible incongruities in the explanatory power of the regressions given by the R^2.

It is common in empirical research to estimate simultaneous equations by the Seemingly Unrelated Regressions model (SUR) in order to allow possible correlation between the errors. The CAPM can be modeled as a SUR model with identical regressors (i.e. the difference between the return of the market portfolio and the risk-free rate is the common regressor). In this case, we know that the generalized least squares of the SUR model is equivalent to the ordinary least squares. Therefore, we would obtain the same results by implementing a SUR model for every single regression. More details about this result can be found in Greene (2008).

We sample the tick-by-tick data at every minute and perform the analysis described above to filter the noisy data and extract the efficient log-price $\widetilde{X}_t$ for each stock. The returns for the filtered high frequency analysis are calculated by $r_t = \widetilde{X}_t - \widetilde{X}_{t-1}$, while the return for the noisy high frequency data is equal to $r_t^* = Y_t - Y_{t-1}$. The same approach is applied to the DJIA index tick-by-tick data to calculate the market return r_{Mkt}. We implement the CAPM regressions using the DJIA index as the market portfolio. The total number of daily observations we have is 503, while for the high frequency (every minute) data set we have $196,170$ for each stock. Therefore, daily observations account for less than 0.3%, i.e. $503/196,170$ of the minute-by-minute data set.

Table 3 presents the results. It is clear that equity betas show significant differences depending on whether we use the filtered high, unfiltered high, or low frequency information sets. We can observe that high frequency equity betas could either be higher or lower than those resulting from the daily analysis. For instance, HPQ's beta is higher in the filtered high frequency analysis (1.044) than the daily measure (0.870), while for EK we have a smaller equity beta (0.618) with higher frequencies than the daily beta (1.071).

We also see crucial differences in the explanatory power of the regressions. Some stocks exhibit a smaller R^2 in the filtered high frequency analysis than the measure we obtain when daily data are used. One interpretation is that there is more diversifiable risk for this asset than the daily data suggests. On the other hand, there are assets where the filtered high frequency analysis indicates that the asset carries more systematic risk than a low frequency study would suggest, see for instance HPQ. But on average, the R^2 is lower for the filtered high frequency data case.

Finally, from Table 3 we observe that for all stocks the unfiltered high frequency equity betas

are greater than the filtered high frequency betas. Moreover, apart from three cases (JNJ, KO and XOM), the unfiltered high-frequency R^2s are lower than those of the filtered high frequency analysis which, as expected, could be taken as an indication that the filtered data is less noisy than the unfiltered one.

6. Conclusions

The present paper focuses on the high frequency volatility and covariation analysis of financial assets in the presence of market microstructure noise. In our study we assume that the true efficient log-price follows an arithmetic Brownian motion and that the noisy log-price is the true log-price plus noise. We review several realized variance estimators that have been proposed in the literature and, based on the Kalman filter methodology, we establish a framework that allows us to: obtain efficient estimators for the variance of the price process; calculate the variance of the microstructure noise; and generate the filtered series of the log-price process. Our estimator is compared to others in the literature and we use Monte Carlo simulations to show that it achieves very low RMSE relative to the other benchmarks, some of which have been designed to handle the presence of microstructure noise at high frequency levels.

The fact that our approach allows us to generate the true efficient price series, or in other words, that we are able to filter out the microstructure noise from the high frequency data, is important in a univariate set up, but perhaps more important in a multivariate setting. Working with filtered high frequency price series allows us to employ best estimators when calculating, for example, variance-covariance matrices between several financial assets. Therefore, a direct consequence of being able to work with data without microstructure noise makes our proposed variance-covariance estimators more efficient than several estimators proposed in the literature.

The estimation of the covariation between assets is a crucial element in both theoretical and empirical research in finance. Most applications, for instance portfolio and risk management, rely on being able to accurately measure the volatility and covariation of assets. Regarding portfolio management applications, we present two examples to show how the quality and quantity of data used to estimate the different components of portfolio analysis have an important effect on the overall result. We look at a small portfolio of two very liquid traded assets (MSFT and INTC) and a large portfolio that includes stocks of the DJIA index. We compute and compare the efficient frontiers we obtain from daily, noisy high frequency and filtered high frequency data. Evidence

shows that there are crucial differences between the efficient frontiers. These differences lead to different portfolio management decisions in terms of the risk-reward tradeoff or the portfolio weights required to achieve a certain risk-reward target.

Moreover, we examine the CAPM equity beta estimates that we calculate from the three data sets: daily, filtered high frequency and noisy high frequency data. Again, there is evidence that the filtered high frequency estimates differ significantly from the daily estimates. The beta estimates we obtain from the high frequency data are robust to the bid-ask bounce and other sources of microstructure noise, as we have filtered the data. Moreover, the R^2 values in the CAPM regressions for filtered high frequency data are smaller than those obtained using daily observations.

Appendix

A. Volatility estimators

Here we briefly mention some alternative estimators that have been proposed in the literature to account for finite small samples, autocorrelation in the noise term and discontinuities of the log-price process.

A further adjustment of the $TSRV$ has been introduced by Zhang et al. (2005) in order to achieve unbiasedness in finite samples, formally described as

$$TSRV_Y^{(adj)} = \left(1 - \frac{\bar{N}}{N}\right)^{-1} TSRV. \tag{46}$$

Hansen and Lunde (2004) propose an unbiased estimator that takes into account the possible autocorrelation, say of order q in its general case, of the microstructure noise. The realized variance is given by

$$RV_Y^{(AC(q))} = \sum_{i=1}^{N} r^2(i) + 2\sum_{k=1}^{q} \frac{N}{N-k} \sum_{i=1}^{N-k} r(i)r(i+k) \tag{47}$$

where $r(i)$ is the log-return.

The case where jumps are incorporated in the price path is examined by Barndorff-Nielsen and Shephard (2004), who proposed the *bi-power realized variation* and *multi-power variation*. Other approaches have been the *Threshold realized variance* of Mancini and Renò (2006) and the *wavelet realized volatility* of Fan and Wang (2007).

Interesting studies where the price process is a pure jump process are those of Oomen (2005) and Oomen (2006). In both approaches the price process follows a compound Poisson process. Microstructure noise is incorporated in the process as the proposed variance estimator uses observations at the highest frequency level, at transaction level. Properties of volatility estimators under alternative sampling schemes are also provided.

The existence of discontinuities in the price process is clearly of interest. Aït-Sahalia and Jacod (2006) propose a statistic which can be used to test for the existence of jumps in high frequency data. The empirical results that they demonstrate prove the existence of jumps in the price processes, emphasizing the importance of taking jumps into account when estimating realized variances. In this paper we do not incorporate jumps in our log-price path, something that could be a point for

further research.

B. Lemma 1

The following lemma has been used to find the expression for the conditional expectation of the efficient log-price X_t in Section 3.

$$\mathbb{E}(x|y,z) = \mathbb{E}(x|y) + \Sigma_{xz}\Sigma_{zz}^{-1}z, \tag{48}$$

$$Var(x|y,z) = Var(x|y) - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}^{'}. \tag{49}$$

C. Difference in returns for the same level of risk: Figure 7

Once we calculated the efficient frontiers, one using filtered high frequency observations and the other daily data, we are interested in quantifying the difference between the return per level of risk. The approach we follow here is to fit both efficient frontiers to equations where we can then generate the returns for a series of risk levels. The function we fit is the ratio of two polynomials.

Regarding the efficient frontiers based on daily data and filtered high frequency data we use a rational model with a quadratic numerator and a 5th degree polynomial in the denominator:

$$f(x) = \frac{p_1x^2 + p_2x + p_3}{x^5 + q_1x^4 + q_2x^3 + q_3x^2 + q_4x + q_5}$$

The power of the fit for this model using daily data is 99.86% and the RMSE is 0.005. The results for the coefficients of the model are presented in Table 4.

In the case of filtered high frequency data the R^2 is 99.87% and the RMSE is 0.005. Figure 7 depicts the difference between these two fitted frontiers.

D. Correlation matrix for DJIA-Daily observations

E. Correlation matrix for DJIA-High Frequency observations

References

Aït-Sahalia, Y. and Jacod, J.: 2006, Testing for jumps in a discretely observed process, *Annals of Statistics* .

Aït-Sahalia, Y. and Mancini, L.: 2006, Out of sample forecasts of quadratic variation, *manuscript Princeton University* .

Aït-Sahalia, Y., Mykland, P. and Zhang, L.: 2005, How often to sample a continuous-time process in the presence of market microstructure noise, *Review of Financial Studies* **18**(2), 351–416.

Andersen, T. and Bollerslev, T.: 1998, Answering the skeptics: Yes, standard volatility models do provide accurate forecasts, *International Economic Review* **39**(4), 885–905.

Andersen, T., Bollerslev, T. and Meddahi, N.: 2004, Analytic evaluation of volatility forecasts, *International Economic Review* **45**, 1079–111.

Andersen, T., Bollerslev, T. and Meddahi, N.: 2006, Realized volatility forecasting and market microstructure noise. Working Paper.

Bandi, F. and Russell, J.: 2008, Microstructure noise, realized variance, and optimal sampling, *Review of Economic Studies* **75**(2), 339–369.

Barndorff-Nielsen, O. and Shephard, N.: 2004, Power and bipower variation with stochastic volatility and jumps, *Journal of Financial Econometrics* **2**(1), 1–37.

Bauwens, L. and Giot, P.: 2000, The logarithmic acd model: An application to the bid-ask quote process of three nyse stocks, *Annales d'Economie et de Statistique* **60**, 117–149.

Bedendo, M. and Hodges, S.: 2009, The dynamics of the volatility skew: A kalman filter approach, *Journal of Banking & Finance* **33**(6), 1156–1165.

Black, F.: 1986, Noise, *Journal of Finance* **41**(3), 529–543.

Chan, S., Goodwin, G. and Sin, K.: 1984, Convergence properties of the riccati difference equation in optimal filtering of nonstabilizable systems, *Automatic Control, IEEE Transactions on* **29**(2), 110–118.

Dempster, A., Laird, N. and Rubin, D.: 1977, Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society* **39**(1), 1–38.

Dufour, A. and Engle, R.: 2000, Time and the price impact of a trade, *The Journal of Finance* **55**(6), 2467–2498.

Durbin, J. and Koopman, S.: 2001, *Time Series Analysis by State Space Methods*, Oxford University Press.

Easley, D. and O'Hara, M.: 1992, Time and the process of security price adjustment, *Journal of Finance* **47**(2), 577–605.

Engle, R.: 2000, The econometrics of ultra-high-frequency data, *Econometrica* **68**(1), 1–22.

Engle, R. and Russell, J.: 1998, Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica* **66**(5), 1127–1162.

Fan, J. and Wang, Y.: 2007, Multi-scale jump and volatility analysis for high-frequency financial data, *Journal of American Statistical Association* **102**(480), 1349–1362.

Glosten, L.: 1987, Components of the bid-ask spread and the statistical properties of transaction prices, *Journal of Finance* **42**(5), 1293–1307.

Greene, W.: 2008, *Econometric analysis*, Prentice Hall Upper Saddle River, NJ.

Hansen, P. and Lunde, A.: 2004, An unbiased measure of realized variance, *Working paper, Social Science Research Network* .

Hansen, P. and Lunde, A.: 2006, Realized variance and market microstructure noise, *Journal of Business and Economic Statistics* **24**, 127–218.

Harvey, A.: 1990, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Hayashi, T. and Yoshida, N.: 2005, On covariance estimation of non-synchronously observed diffusion processes, *Bernouli* **11**(2), 359–379.

Mancini, C. and Renò, R.: 2006, Threshold estimation of jump-diffusion models and interest rate modeling, *Manuscript, University of Florence and University of Siena* .

Menkveld, A., Koopman, S. and Lucas, A.: 2007, Modeling around-the-clock price discovery for cross-listed stocks using state space methods, *Journal of Business and Economic Statistics* **25**(2), 213–225.

Øksendal, B.: 2007, *Stochastic differential equations: an introduction with applications*, Springer-Verlag.

Oomen, R.: 2005, Properties of bias-corrected realized variance under alternative sampling schemes, *Journal of Financial Econometrics* **3**(4), 555–577.

Oomen, R.: 2006, Properties realized variance under alternative sampling schemes, *Journal of Business & Economic Statistics* **24**(2), 219–237.

Wang, Y., Yao, Q., Li, P. and Zou, J.: 2007, High dimensional voalatility modeling and analysis for high-frequency financial data, *working paper, merlot.stat.uconn.edu* .

Watson, M. and Engle, R.: 1983, Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models, *Journal of Econometrics* **23**(3), 385–400.

Wood, R., McInish, T. and Ord, J.: 1985, An investigation of transactions data for nyse stocks, *Journal of Finance* **40**(3), 723–739.

Zhang, L.: 2006, Estimating covariation: Epps effect, microstructure noise, *manuscript* .

Zhang, L., Mykland, P. and Aït-Sahalia, Y.: 2005, A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* **100**, 1394–1411.

Zhang, M., Russell, J. and Tsay, R.: 2001, A nonlinear autoregressive conditional duration model with applications to financial transaction data, *Journal of Econometrics* **104**(1), 179–207.

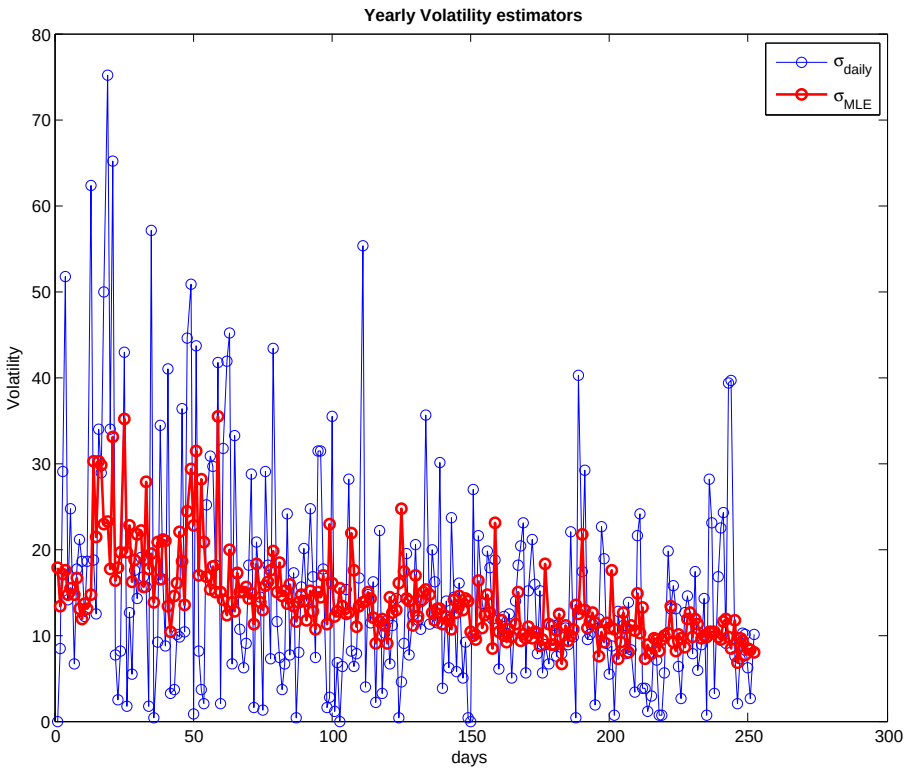

Figure 1. Difference between a volatility estimator based on daily observations and one based on high frequency data. A wide range of values of $\hat{\sigma}_{daily}$ seem implausible, for instance, we get volatility estimates from 0% or up to 75%. The MLE estimator based on minute-by-minute observations, on the other hand, has a less volatile path over the year.

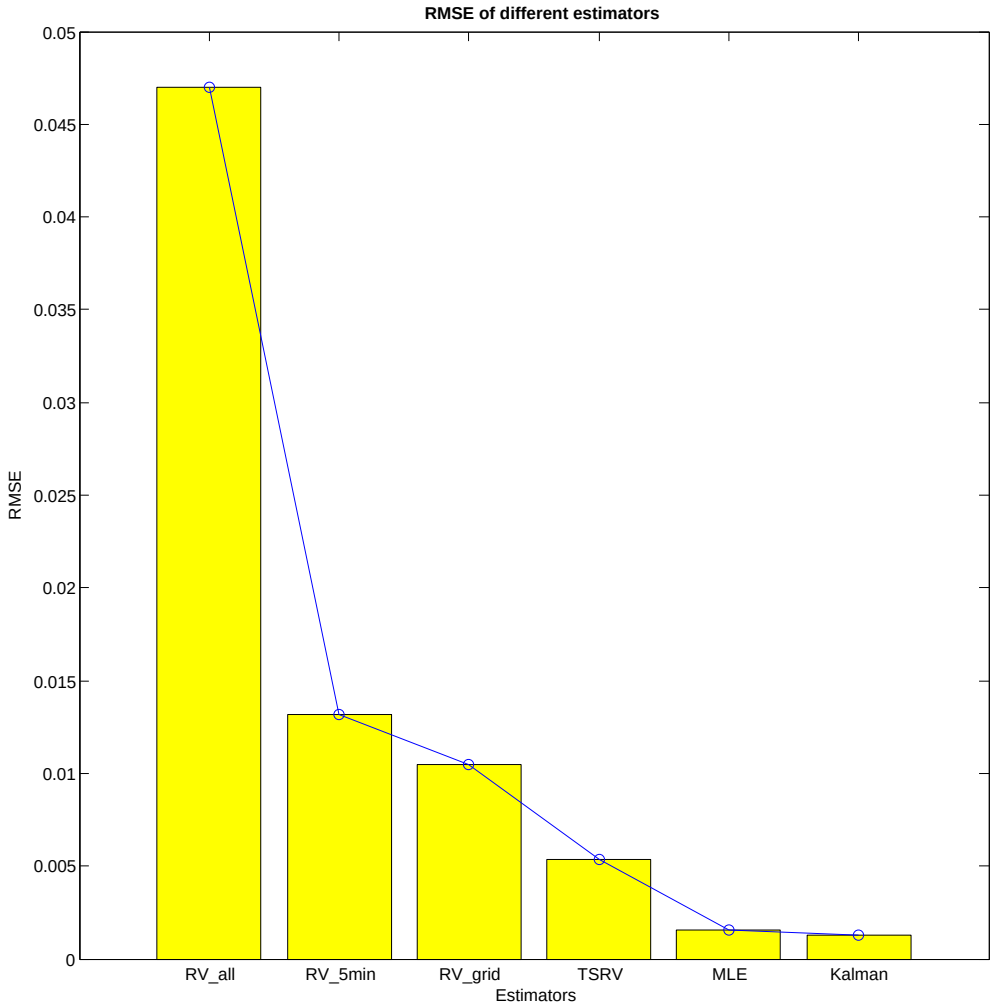

Figure 2. RMSEs of the estimators (non-parametric, parametric and the proposed Kalman) for $1,000$ simulations, with $\sigma_\varepsilon^2 = 1 \times 10^{-6}$, $\sigma^2 = 0.09$, $\Delta = 1/23,400$ and the starting value of the asset is $S_0 = 30$.

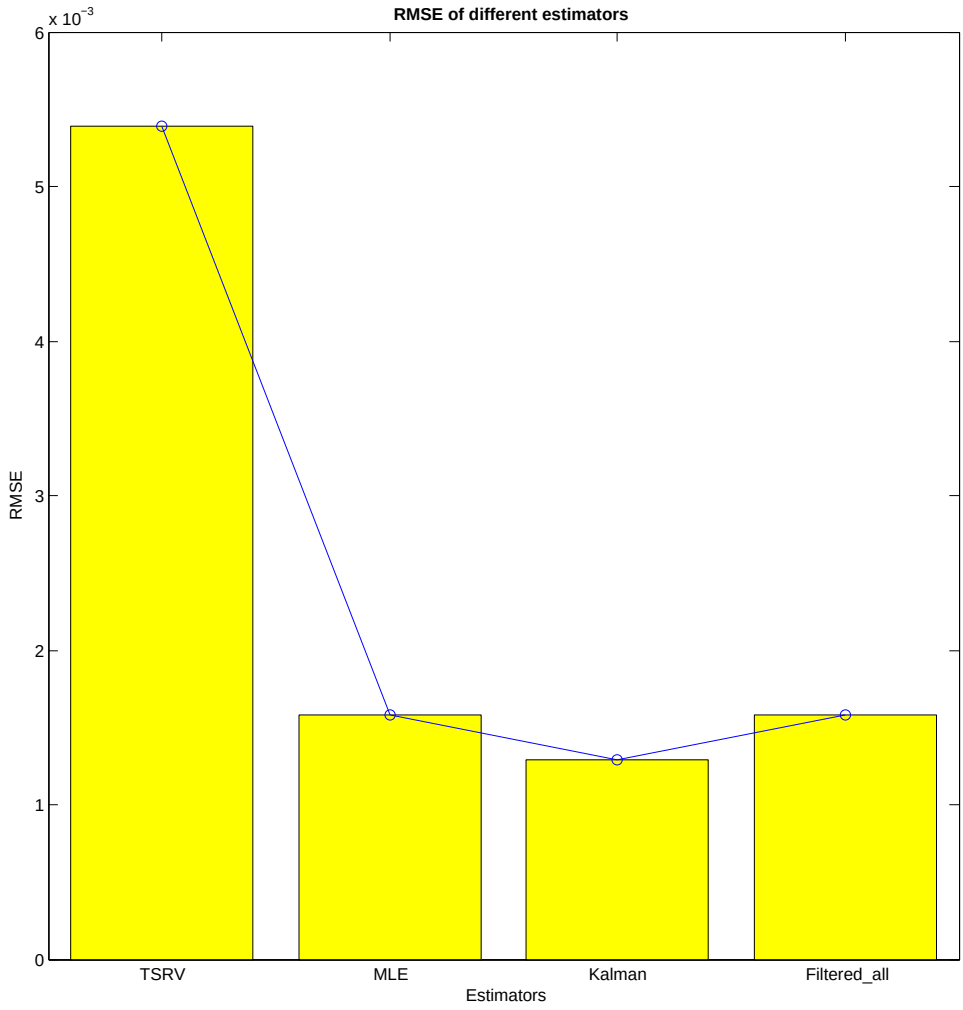

Figure 3. This figure shows that the RMSE of MLE estimator and that of the Kalman filter almost coincide. $TSRV$, on the other hand, has a greater RMSE than the other two estimators.

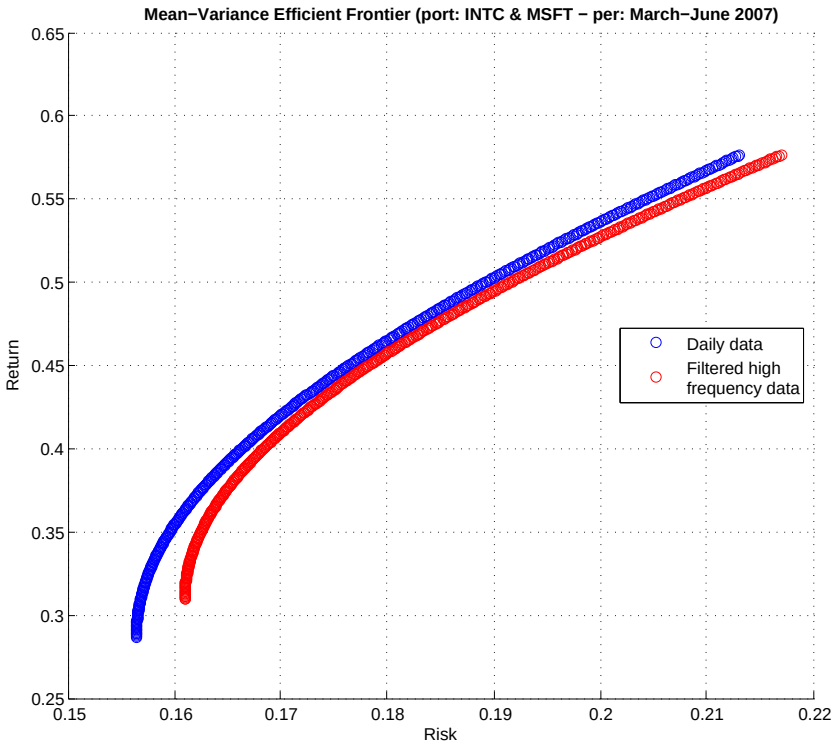

Figure 4. Efficient frontiers based on different estimators of the VCM of MSFT and INTC that use daily and filtered high frequency data. For both frontiers the returns are calculated using daily observations.

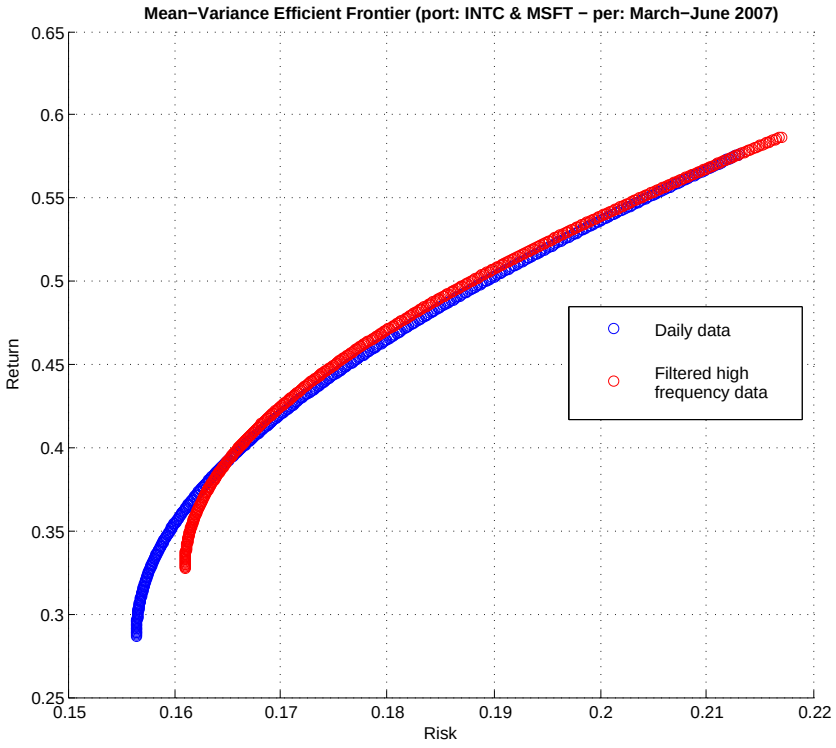

Figure 5. Efficient frontiers based on different estimators of the VCM of MSFT and INTC that use daily and filtered high frequency data. For the 'daily' frontier returns are calculated using daily data. For the 'filtered high frequency frontier' returns are calculated using filtered high frequency data.

Table 1: Optimization Results

	Case	Return	Risk	Weights	
				MSFT	INTC
Portfolio I	Daily data	30.00%	**15.66%**	63.90%	36.09%
	5-min Midquote data	30.53%	**15.65%**	66.92%	33.08%
	Filtered High Frequency data	30.55%	**15.64%**	66.94%	33.06%
Portfolio II	Daily data	35.00%	**15.96%**	52.35%	47.65%
	5-min Midquote data	35.00%	**15.89%**	54.35%	45.65%
	Filtered High Frequency data	35.00%	**15.80%**	56.35%	43.65%
Portfolio III	Daily data	40.00%	**16.63%**	40.79%	59.21%
	5-min Midquote data	40.00%	**16.52%**	43.42%	56.58%
	Filtered High Frequency data	40.00%	**16.38%**	44.45%	55.55%
Portfolio IV	Daily data	50.00%	**18.92%**	17.66%	82.34%
	5-min Midquote data	50.00%	**18.69%**	20.22%	79.78%
	Filtered High Frequency data	50.00%	**18.56%**	20.65%	79.35%

Table 2: Return-Volatility for Daily and High Frequency (HF) data

Asset		Daily		Filtered HF		Noisy HF		5-min HF	
		Return	Risk	Return	Risk	Return	Risk	Return	Risk
AA	Alcoa Inc	59.40%	27.53%	63.33%	30.45%	63.33%	36.43%	63.33%	30.59%
AXP	American Express	23.61%	18.25%	25.17%	19.35%	25.14%	21.25%	24.93%	19.27%
BA	Boeing Company	27.11%	15.20%	30.92%	17.34%	30.98%	19.27%	30.92%	17.34%
C	Citigroup	1.23%	18.31%	7.62%	18.99%	7.68%	23.51%	7.92%	18.99%
CAT	Caterpillar Inc	60.97%	17.70%	60.01%	21.71%	60.10%	24.30%	59.68%	21.72%
DD	Du Pont De Nemours	-0.18%	17.52%	1.75%	20.40%	1.99%	24.01%	1.75%	20.72%
DIS	Walt Disney	-2.19%	14.83%	3.41%	16.64%	3.59%	20.34%	3.41%	16.34%
EK	Eastman Kodak	43.67%	29.05%	49.10%	32.25%	49.22%	41.45%	49.01%	33.41%
GE	General Electric	26.87%	13.56%	29.29%	15.22%	29.29%	39.09%	29.53%	15.26%
GM	General Motors	54.32%	32.34%	56.12%	29.50%	55.98%	32.48%	56.12%	30.00%
HD	Home Depot Inc	-0.76%	17.58%	1.33%	22.19%	1.42%	24.76%	1.13%	22.52%
HON	Honeywell	59.27%	19.22%	64.04%	20.50%	64.31%	25.77%	64.25%	20.56%
HPQ	Helett-Packard	40.93%	14.57%	43.37%	18.18%	43.46%	21.73%	43.41%	18.25%
IBM	Int. Business Machines	39.49%	16.80%	47.14%	16.72%	47.17%	20.18%	46.87%	16.67%
INTC	Intel Corp	57.64%	21.31%	58.67%	21.70%	59.06%	25.08%	59.18%	21.28%
IP	International Paper	26.24%	17.89%	26.65%	19.11%	26.65%	25.40%	26.42%	19.36%
JNJ	Johnson & Johnson	-4.01%	11.62%	-3.35%	14.79%	-3.50%	26.72%	-3.56%	14.77%
JPM	J. P. Morgan	-4.61%	18.68%	-3.47%	18.82%	-3.53%	21.31%	-3.35%	18.80%
KO	Coca-Cola	35.19%	12.09%	35.34%	14.11%	35.22%	28.72%	35.16%	14.07%
MCD	McDonalds	42.19%	15.59%	45.71%	19.12%	45.84%	21.83%	45.66%	19.28%
MO	Minesota Mng Mfg	-55.31%	45.88%	-51.14%	49.26%	-51.38%	50.81%	-51.08%	49.19%
MRK	Merck	37.22%	21.20%	36.73%	22.30%	37.03%	28.00%	36.61%	22.67%
MSFT	Microsoft	14.39%	17.20%	16.66%	18.45%	17.08%	20.86%	16.78%	18.29%
PG	Procter & Gamble	-11.97%	11.40%	-10.11%	14.11%	-9.90%	15.68%	-10.05%	13.91%
T	AT & T Corp	36.79%	18.29%	40.62%	19.46%	40.47%	23.21%	39.96%	19.66%
UTX	United Technologies	25.64%	12.66%	27.78%	15.79%	27.96%	17.96%	28.08%	15.92%
WMT	Wal-Mart Stores	1.38%	16.40%	1.87%	17.27%	1.99%	19.33%	1.87%	17.35%
XOM	Exxon Mobil	50.05%	18.42%	49.69%	18.26%	49.72%	20.89%	49.69%	18.09%
	Average value	**24.45** %		**26.94%**		**27.01%**		**26.91%**	

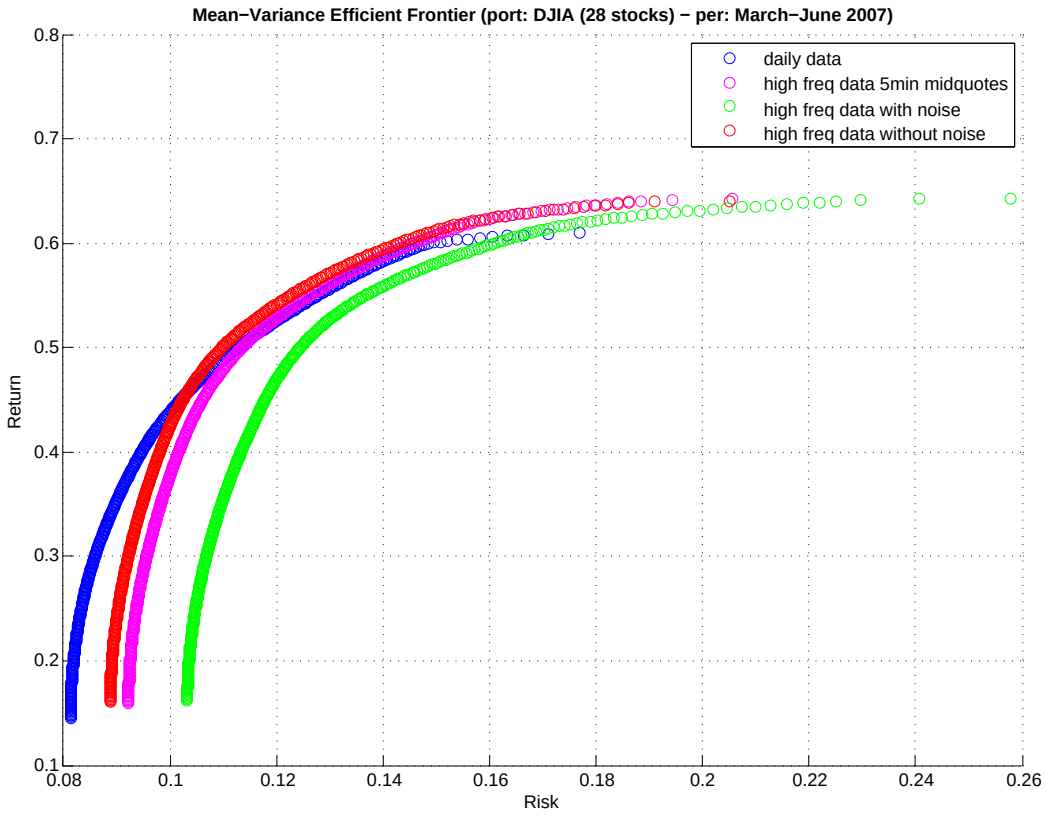

Figure 6. Efficient frontiers based on the different estimators of the VCM using 28 constituents of the DJIA.

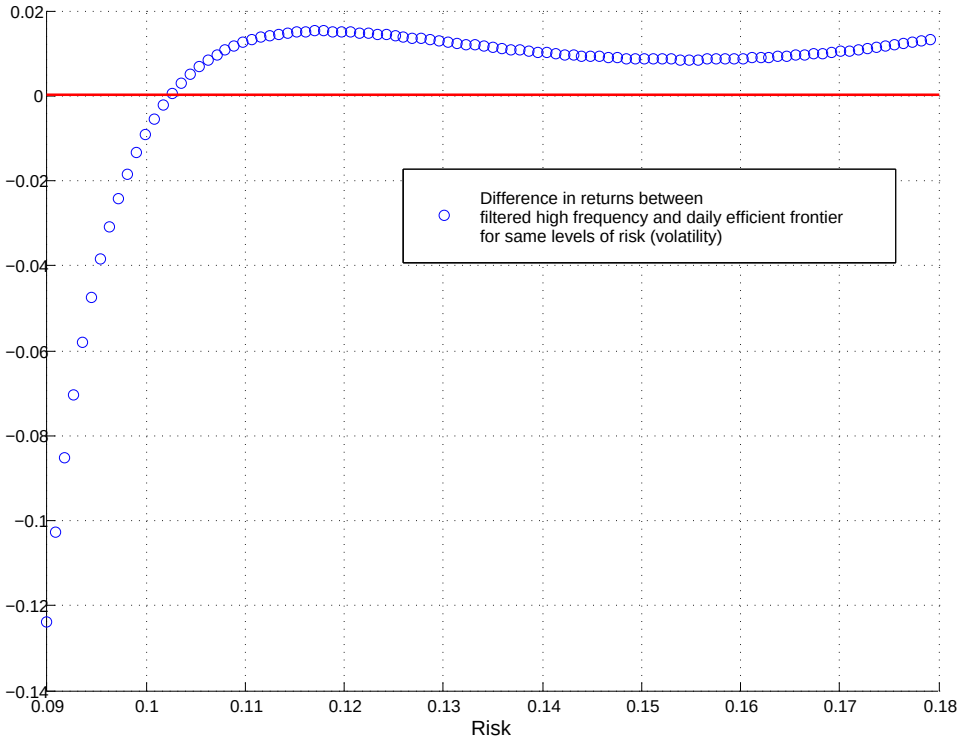

Figure 7. Difference in the returns of the two efficient frontiers based on daily and filtered high frequency data of Figure 6.

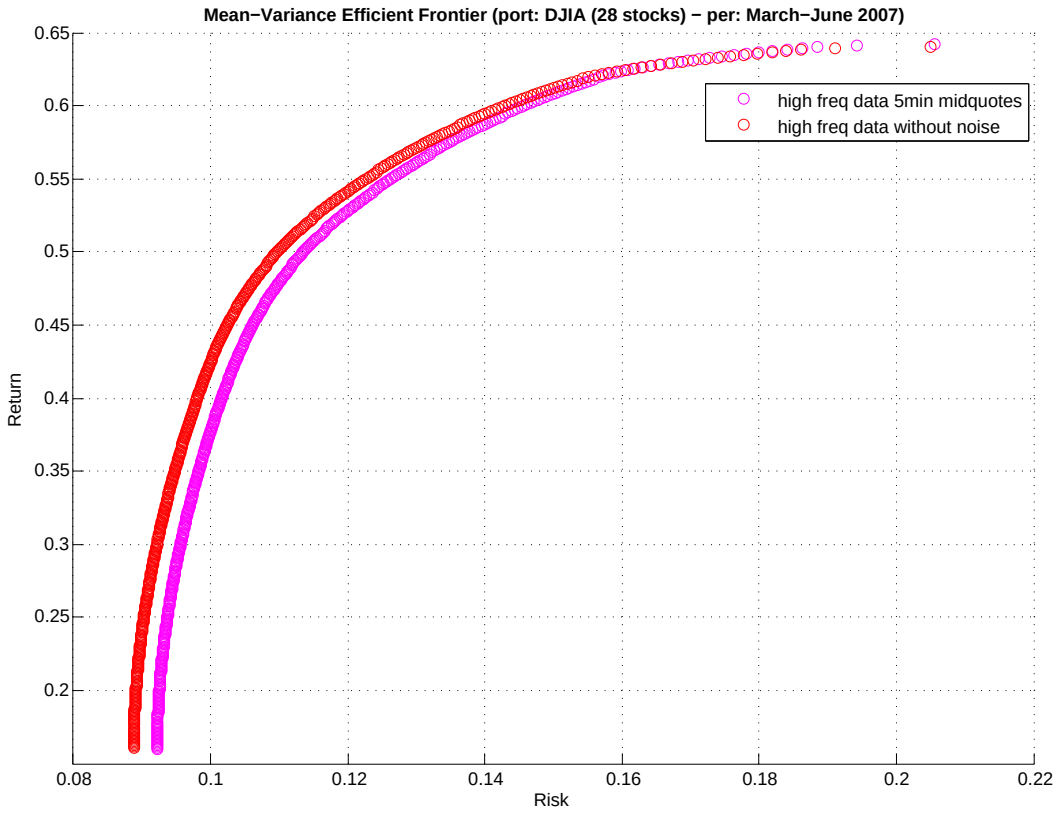

Figure 8. 5-min midquote and filtered HF efficient frontiers.

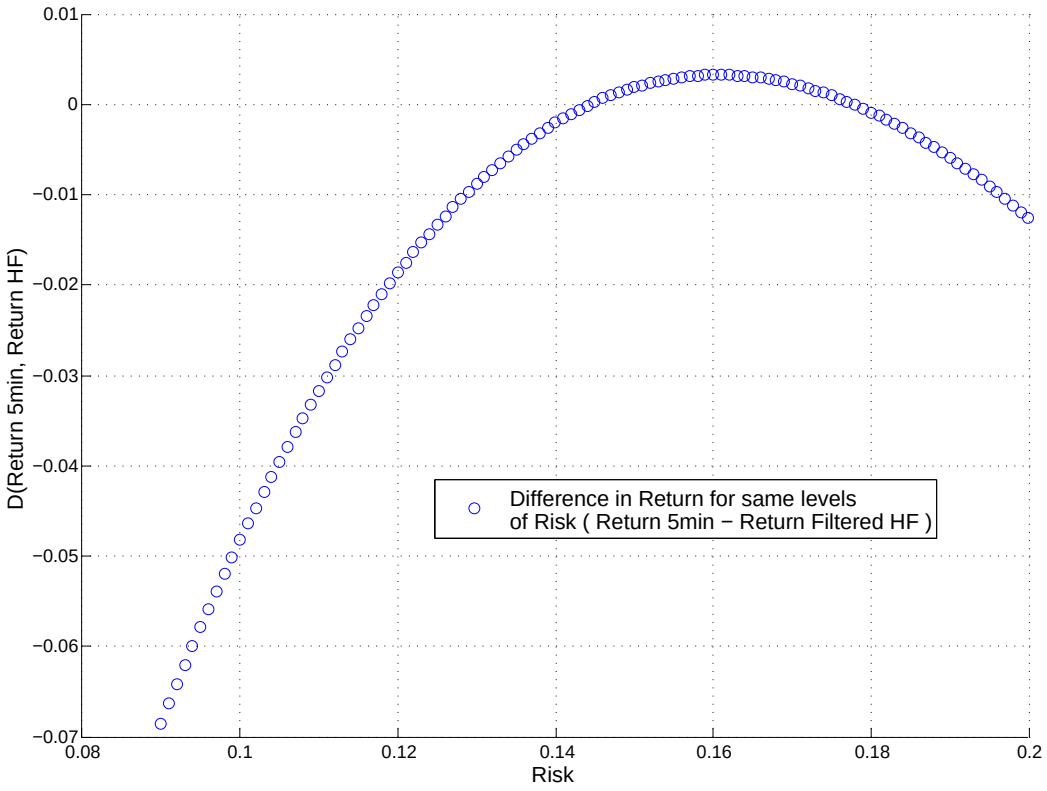

Figure 9. Difference in the returns, per level of risk, of the two efficient frontiers based on 5 min midquotes and filtered high frequency data of Figure 8.

Table 3: Equity betas for daily and high frequency(HF) data-DJIA index (market portfolio)

Asset	Daily Data		Filtered HF Data		Noisy HF Data	
	Beta coefficient	R^2	Beta coefficient	R^2	Beta coefficient	R^2
AA	1.093	21.40%	1.191	14.47%	1.242	12.81%
AXP	0.941	26.93%	0.782	16.14%	0.828	15.24%
BA	1.077	24.87%	1.054	21.79%	1.122	0.87%
DD	0.953	30.15%	0.978	20.84%	1.048	17.77%
DIS	0.850	22.43%	0.762	12.88%	0.847	11.31%
EK	1.071	14.21%	0.618	3.90%	0.646	3.61%
GE	0.780	37.10%	0.747	23.79%	0.803	18.73%
GM	1.197	8.44%	1.029	5.84%	1.066	5.39%
HD	1.107	31.88%	1.022	18.22%	1.079	16.06%
HON	1.050	31.29%	1.134	21.59%	1.218	19.09%
HPQ	0.870	11.21%	1.044	12.60%	1.107	12.22%
IBM	0.839	28.53%	0.958	26.11%	1.019	25.84%
INTC	1.231	27.57%	1.191	18.09%	1.289	1.93%
IP	1.066	27.21%	0.838	10.15%	0.856	9.95%
JNJ	0.529	17.89%	0.603	15.90%	0.665	16.08%
JPM	0.948	39.77%	0.893	24.19%	0.937	6.30%
KO	0.630	32.10%	0.411	5.33%	0.751	16.43%
MCD	1.016	25.29%	0.902	15.33%	0.956	10.36%
MO	0.696	16.79%	0.772	14.99%	0.807	13.71%
MRK	0.799	12.19%	0.816	10.19%	0.874	9.69%
MSFT	0.773	19.54%	0.796	14.65%	0.888	12.79
PG	0.665	24.07%	.768	18.86%	0.825	18.39%
WMT	0.799	25.95%	0.783	15.96%	0.828	13.57%
XOM	0.987	22.72%	0.940	14.65%	0.977	15.87%
Average value	0.915	**24.15%**	0.876	**15.79%**	*0.945*	***12.67%***

Table 4: Fitting results

	Coefficient value
DAILY FIT	
p_1	307.100
p_2	-36.870
p_3	0.993
q_1	206.200
q_2	48.770
q_3	346.100
q_4	-21.130
q_5	-0.257
R^2	0.9986
RMSE	0.005
HIGH FREQUENCY FIT	
p_1	187.700
p_2	-3.713
p_3	-1.067
q_1	173.500
q_2	48.940
q_3	186.900
q_4	14.390
q_5	-2.345
R^2	0.9987
RMSE	0.005

Table 5: Correlation coefficients (Daily Data)

AA	AXP	BA	C	CAT	
1.0000	0.3532	0.4141	0.3322	0.2228	AA
	1.0000	0.4765	0.7100	0.5151	AXP
		1.0000	0.4848	0.4482	BA
			1.0000	0.4148	C
				1.0000	CAT

DD	DIS	EK	GE	GM	
0.4403	0.3974	0.2613	0.3421	0.1842	AA
0.6007	0.5590	0.3685	0.4731	0.2627	AXP
0.3317	0.2531	0.1801	0.2796	0.1418	BA
0.5502	0.5487	0.3566	0.5303	0.2917	C
0.3531	0.4142	0.2302	0.4085	0.2479	CAT
1.0000	0.5322	0.2012	0.4289	0.3076	DD
	1.0000	0.3238	0.4909	0.3123	DIS
		1.0000	0.2632	0.1529	EK
			1.0000	0.2676	GE
				1.0000	GM

HD	HON	HPQ	IBM	INTC	
0.0609	0.1808	0.2500	0.2461	0.2686	AA
0.4088	0.5713	0.4120	0.3768	0.5111	AXP
0.1763	0.3699	0.2762	0.1904	0.2815	BA
0.3759	0.3875	0.4081	0.4492	0.5533	C
0.2717	0.6038	0.4193	0.3054	0.3030	CAT
0.4657	0.4346	0.3411	0.4405	0.4333	DD
0.3373	0.3792	0.5872	0.3926	0.3943	DIS
0.3321	0.3511	0.1440	0.3390	0.3271	EK
0.2577	0.3352	0.3066	0.3662	0.2840	GE
0.2092	0.2090	0.1975	0.1915	0.4237	GM
1.0000	0.3186	0.1572	0.3696	0.3745	HD
	1.0000	0.3988	0.5003	0.4775	HON
		1.0000	0.3170	0.2640	HPQ
			1.0000	0.4409	IBM
				1.0000	INTC

IP	JNJ	JPM	KO	MCD	
0.5520	0.1469	0.3465	0.2350	0.1507	AA
0.4970	0.4036	0.7495	0.4487	0.4464	AXP
0.3721	0.2456	0.6075	0.3942	0.2653	BA
0.4968	0.4989	0.8257	0.5026	0.3706	C
0.3705	0.3332	0.5500	0.3736	0.2543	CAT
0.5647	0.2665	0.4895	0.3814	0.3296	DD
0.5649	0.3428	0.5191	0.4057	0.2721	DIS
0.2568	0.2495	0.2945	0.0972	0.1373	EK
0.4517	0.3051	0.4812	0.2635	0.2554	GE
0.3597	0.1508	0.2911	0.2623	0.3891	GM
0.3291	0.5043	0.3534	0.4584	0.2017	HD
0.3320	0.3423	0.4885	0.3853	0.2943	HON
0.2668	0.2530	0.4345	0.2607	0.2785	HPQ
0.3877	0.3650	0.3762	0.4976	0.2420	IBM
0.4110	0.4090	0.5780	0.4442	0.3858	INTC
1.0000	0.2507	0.5114	0.3942	0.3839	IP
	1.0000	0.5153	0.5478	0.3990	JNJ
		1.0000	0.5553	0.4900	JPM
			1.0000	0.2731	KO
				1.0000	MCD

Table 6: Correlation coefficients (Daily Data)

MO	MRK	MSFT	PG	T	
0.1020	0.2274	0.2716	0.1435	0.3284	AA
0.2494	0.5022	0.5965	0.3350	0.4004	AXP
0.1486	0.1188	0.3466	0.2406	0.3576	BA
0.2026	0.3298	0.5146	0.5456	0.4828	C
0.1755	0.2001	0.4953	0.1975	0.3064	CAT
0.1902	0.2701	0.4171	0.3245	0.3092	DD
0.1200	0.3671	0.4821	0.2469	0.4524	DIS
−0.0206	0.1797	0.2739	0.0249	0.1115	EK
0.1107	0.2500	0.5460	0.3379	0.4125	GE
0.0244	0.2981	0.1561	0.2134	0.3217	GM
0.0641	0.1447	0.3391	0.2024	0.1968	HD
0.1306	0.2788	0.5570	0.2225	0.3087	HON
0.0862	0.2665	0.4726	0.1296	0.2415	HPQ
0.0141	0.1830	0.4715	0.3340	0.1926	IBM
0.1143	0.3082	0.3845	0.2265	0.2843	INTC
0.0532	0.2532	0.2787	0.2014	0.4532	IP
0.1632	0.3091	0.3993	0.3982	0.3231	JNJ
0.1952	0.3793	0.5604	0.4434	0.4440	JPM
−0.0087	0.2878	0.4892	0.3697	0.3993	KO
0.1952	0.3102	0.3572	0.1216	0.3136	MCD
1.0000	−0.0545	0.1902	0.1146	0.0757	MO
	1.0000	0.2766	0.2235	0.2962	MRK
		1.0000	0.3236	0.3050	MSFT
			1.0000	0.4258	PG
				1.0000	T

UTX	WMT	XOM	
0.4047	0.1950	0.4675	AA
0.5462	0.4297	0.5897	AXP
0.5231	0.2002	0.4320	BA
0.4921	0.3411	0.5642	C
0.4449	0.3359	0.5398	CAT
0.6051	0.4440	0.4055	DD
0.3909	0.4194	0.6415	DIS
0.2993	0.2430	0.3544	EK
0.3466	0.2587	0.4561	GE
0.4237	0.2625	0.3589	GM
0.4365	0.4918	0.2124	HD
0.4363	0.4516	0.5035	HON
0.3769	0.2895	0.4689	HPQ
0.3273	0.1910	0.4400	IBM
0.5716	0.2512	0.4445	INTC
0.5041	0.4807	0.5344	IP
0.3070	0.3213	0.3838	JNJ
0.5848	0.3680	0.6010	JPM
0.4276	0.4028	0.4766	KO
0.5014	0.2621	0.2588	MCD
0.1735	0.0102	0.0687	MO
0.2321	0.2600	0.4031	MRK
0.4685	0.3251	0.5475	MSFT
0.2853	0.1968	0.2756	PG
0.4305	0.3525	0.4541	T
1.0000	0.4375	0.4174	UTX
	1.0000	0.4146	WMT
		1.0000	XOM

Table 7: Correlation coefficients (High frequency Data)

AA	AXP	BA	C	CAT	
1.0000	0.2534	0.2549	0.2588	0.2531	AA
	1.0000	0.3040	0.4031	0.3534	AXP
		1.0000	0.3167	0.2878	BA
			1.0000	0.3492	C
				1.0000	CAT

DD	DIS	EK	GE	GM	
0.2771	0.2651	0.1006	0.2481	0.1344	AA
0.3137	0.3096	0.1577	0.3270	0.2101	AXP
0.2911	0.2795	0.1009	0.2955	0.1933	BA
0.2773	0.3255	0.1428	0.3429	0.2197	C
0.2350	0.2846	0.1399	0.2977	0.2047	CAT
1.0000	0.2490	0.1253	0.2654	0.1667	DD
	1.0000	0.1230	0.2943	0.1817	DIS
		1.0000	0.1572	0.0643	EK
			1.0000	0.1971	GE
				1.0000	GM

HD	HON	HPQ	IBM	INTC	
0.1859	0.2297	0.2242	0.2218	0.2312	AA
0.2532	0.3262	0.2805	0.3162	0.2865	AXP
0.2247	0.3078	0.2305	0.3111	0.2809	BA
0.2521	0.2776	0.2904	0.3273	0.3156	C
0.2473	0.4117	0.2525	0.2797	0.2845	CAT
0.1953	0.2492	0.2314	0.2770	0.2429	DD
0.2159	0.2641	0.2933	0.2703	0.2658	DIS
0.1219	0.1397	0.1046	0.1339	0.1016	EK
0.2516	0.2900	0.2726	0.2933	0.2808	GE
0.1523	0.1834	0.1459	0.2024	0.2191	GM
1.0000	0.2238	0.2111	0.2255	0.2408	HD
	1.0000	0.2709	0.2864	0.2603	HON
		1.0000	0.2731	0.2774	HPQ
			1.0000	0.2800	IBM
				1.0000	INTC

IP	JNJ	JPM	KO	MCD	
0.2758	0.1377	0.2614	0.1967	0.1715	AA
0.2828	0.2414	0.4257	0.2609	0.2889	AXP
0.2624	0.2089	0.3017	0.2691	0.2263	BA
0.3035	0.2816	0.4702	0.2971	0.2469	C
0.3093	0.2346	0.3603	0.2464	0.2672	CAT
0.2809	0.1632	0.2933	0.2208	0.2173	DD
0.2814	0.2237	0.3297	0.2427	0.2057	DIS
0.0916	0.1357	0.1402	0.0985	0.1228	EK
0.3127	0.2422	0.3275	0.2724	0.2256	GE
0.1896	0.1681	0.1962	0.2097	0.1777	GM
0.2382	0.2594	0.2751	0.2198	0.1672	HD
0.3106	0.1974	0.2866	0.2376	0.2817	HON
0.2356	0.1822	0.2897	0.2082	0.1971	HPQ
0.3120	0.2589	0.2635	0.2883	0.2170	IBM
0.2901	0.2252	0.3410	0.2373	0.2191	INTC
1.0000	0.2214	0.3212	0.2718	0.2331	IP
	1.0000	0.2500	0.2797	0.2029	JNJ
		1.0000	0.2817	0.2785	JPM
			1.0000	0.1952	KO
				1.0000	MCD

Table 8: Correlation coefficients (High Frequency Data)

MO	MRK	MSFT	PG	T	
0.0495	0.1545	0.2194	0.1819	0.2368	AA
0.0726	0.2632	0.2982	0.3181	0.3346	AXP
0.0724	0.1659	0.2573	0.2243	0.2849	BA
0.0756	0.2222	0.3128	0.2902	0.3439	C
0.0797	0.2084	0.2601	0.2612	0.2876	CAT
0.0439	0.2238	0.2390	0.2141	0.2863	DD
0.1010	0.2004	0.2728	0.2643	0.2605	DIS
0.0035	0.0848	0.1269	0.1078	0.1303	EK
0.0446	0.2442	0.3210	0.2820	0.3352	GE
0.0687	0.1620	0.1921	0.1827	0.2076	GM
0.0684	0.1470	0.2520	0.2031	0.2266	HD
0.0525	0.2725	0.2549	0.2411	0.2871	HON
0.0200	0.2124	0.2817	0.2310	0.2735	HPQ
0.0389	0.2412	0.2985	0.2830	0.3017	IBM
0.0763	0.1769	0.3194	0.2489	0.3092	INTC
0.0592	0.2879	0.2563	0.2603	0.2976	IP
0.0710	0.2148	0.2219	0.2028	0.2301	JNJ
0.1021	0.2311	0.3197	0.3240	0.3643	JPM
0.0228	0.2029	0.2344	0.2520	0.2857	KO
0.2130	0.2275	0.2112	0.2100	0.2774	MCD
1.0000	−0.0121	0.0672	0.0649	0.0129	MO
	1.0000	0.1857	0.1988	0.2550	MRK
		1.0000	0.2607	0.3121	MSFT
			1.0000	0.3012	PG
				1.0000	T

UTX	WMT	XOM	
0.2696	0.1886	0.2844	AA
0.3723	0.2762	0.3601	AXP
0.3549	0.2326	0.3004	BA
0.3777	0.2690	0.3671	C
0.4484	0.2385	0.3402	CAT
0.2879	0.2153	0.2829	DD
0.3261	0.2487	0.3068	DIS
0.1098	0.1533	0.1396	EK
0.3227	0.2810	0.3153	GE
0.1984	0.1487	0.2066	GM
0.2700	0.2414	0.2421	HD
0.3830	0.2422	0.3105	HON
0.2755	0.2081	0.3011	HPQ
0.3144	0.2655	0.3327	IBM
0.3209	0.2526	0.3036	INTC
0.3254	0.2677	0.3197	IP
0.2318	0.1912	0.2692	JNJ
0.4048	0.2873	0.3731	JPM
0.2547	0.2309	0.3014	KO
0.2719	0.2018	0.2853	MCD
0.0694	0.0400	0.0525	MO
0.2157	0.2100	0.2619	MRK
0.2853	0.2336	0.3053	MSFT
0.2922	0.2529	0.2879	PG
0.3345	0.2683	0.3619	T
1.0000	0.2656	0.3303	UTX
	1.0000	0.2851	WMT
		1.0000	XOM