Estimation of Asset Volatility and Correlation Over Market Microstructure Noise in High-Frequency Data

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:14398547">http://nrs.harvard.edu/urn-3:HUL.InstRepos:14398547</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Estimation of Asset Volatility and Correlation over Market Microstructure Noise in High-Frequency Data

Roman Yevstihnyeyev

April 1, 2015

Advised by Professor Neil Shephard

Harvard University
I would like to thank my thesis advisor, Neil Shephard, for his extraordinarily helpful guidance throughout the process. I am also grateful to Mark Glickman for his helpful comments, and to Aaron Perez for helping me when hardware technical issues came up. Last but not least, I am grateful to my parents and my wife for their continual inspiration.
Abstract

Accurate measurement of asset return volatility and correlation is an important problem in financial econometrics. The presence of market microstructure noise in high-frequency data complicates such estimations. This study extends a prior application of a model-based volatility estimator with autocorrelated market microstructure noise to estimation of correlation. The model is applied to a high-frequency dataset including a stock and an index, and the results are compared to some existing models. This study supports previous findings that including an autocorrelation factor produces an estimator potentially less vulnerable to market microstructure noise, and finds that the same is true about the extended correlation estimator that is introduced here.
# Contents

1 Introduction ................................................. 4

2 Data Set ...................................................... 6
   2.1 Data Source ........................................... 6
   2.2 Data Cleaning and Limitations ......................... 6

3 Describing the Price Process .............................. 7
   3.1 Brownian Motion ....................................... 7
   3.2 Market Microstructure Noise ............................ 9

4 Moment-based Estimators ................................... 9
   4.1 Realized Volatility .................................... 9
   4.2 Realized Correlation .................................. 10

5 Model-based Estimators ..................................... 11
   5.1 Univariate Process .................................... 11
   5.2 Autocorrelated Microstructure Noise .................. 12
   5.3 Bivariate Process (BIV) ............................... 15
   5.4 Bivariate Process with Autocorrelated Noise (BIVAC) . 16

6 Validation of Model Implementation ...................... 18
   6.1 Independent noise ..................................... 18
   6.2 Autocorrelated noise .................................. 23
   6.3 Bivariate: Independent Noise ......................... 23
   6.4 Bivariate: Autocorrelated Noise ....................... 32

7 Data Analysis ................................................ 32
   7.1 Optimization Issues ................................... 32
   7.2 R, Rcpp and RcppArmadillo ............................ 35
   7.3 Results and Discussion ................................ 36
1 Introduction

The rapid development of financial markets in recent years, in combination with the financial crisis of 2007-2009, has made risk measurement of financial assets a particularly relevant and important topic. While Kim (2014) points out the importance of measuring volatility, it is no less important to also measure correlations between assets. Asset correlation is relevant in a number of ways: 1) Optimal asset allocation in Modern Portfolio Theory, formulated by Markowitz (1952), whose fundamental ideas about finding the appropriate risk-return trade-off were still relevant fifty years later, e.g. Fabozzi, Gupta, and Markowitz (2002), and are frequently used today, e.g. Bolder (2015), requires asset correlations as inputs. 2) In systemic risk modeling, assets of large systemically important financial institutions are analyzed by financial regulators to determine how much risk they are holding and to determine how much capital those institutions should hold to cushion their potential impact on the larger financial system in case of financial distress. Such risk is inherently dependent on asset correlations (Lehar 2005); for example, during the financial crisis of 2007-2009, many financial assets became significantly more correlated than before, increasing the riskiness of many portfolios and therefore financial institutions. As discussed in Huang et al. (2009), these developments prompted proposals of stress-test metrics that involve measuring how large banks’ portfolios would perform if asset correlations were to undergo a dramatic change.

Measuring observed volatilities and correlations in bivariate time series is complicated by the presence of a so-called market microstructure noise. This noise is caused by bid-ask bounces, liquidity constraints, impact of block trades,
short-term momentum effects, and other auto-correlation effects. Such noise is especially noticeable in high-frequency data. The implication is that the observed volatilities and correlations may not represent values that we are interested in, so that there is a need for approaches to separate the noise from the underlying price movements.

Following up on Kim’s work (2014), we reproduce his analysis of volatility estimators for a longer period of time (several years instead of several days), as well as extend the application of the auto-correlated market microstructure noise model to bivariate case to estimate asset correlation, where we consider a stock and a broad market index through parts of 2013 and 2014 for comparison. Kim bases his analysis on five-day data samples, so that looking at longer time horizons provides an extended test of Kim’s findings about various model-based and moment-based estimators. Extending Kim’s model to bivariate time series provides an additional contribution.

This study will be organized in the following sections. Section 2 will discuss the data set used in the study. Section 3.1 will describe how log price time series can be modeled using Brownian motion, and Section 3.2 will incorporate market microstructure noise to this model. Section 4 will introduce and discuss moment-based estimators: Realized Volatility and Realized Correlation. Section 5 will introduce univariate and bivariate model-based estimators, as well as outline their properties and derivation of Kalman-filter based estimations. Section 6 will describe numeric optimization used during the estimation. Section 7 will describe some issues related to stability and speed of optimization that came up, and how we have resolved them using log-transformation and Rcpp/RcppArmadillo R libraries, as well as describe and discuss the results of running the analysis on data. Section 8 will discuss conclusions and suggest possible directions for the future work.
2 Data Set

2.1 Data Source

The data sets used in this study come from the Wharton Research Data Services (WRDS), a web-based business data research service from the Wharton School of the University of Pennsylvania. We use second-by-second intraday trades of Apple stock (AAPL) and of an exchange-traded fund closely tracking the S&P 500 index (SPY), which are a part of the New York Stock Exchange Trade and Quote (NYSE TAQ) sub-database on WRDS. We chose AAPL because it has a significant market capitalization ($718.66 billion as of March 26, 2015, according to Google Finance) and high liquidity (average daily trading volume of 57.57 million shares as of March 26, 2015), both suggesting the stock’s importance in the markets. We chose SPY, a highly liquid accurate proxy of the S&P 500 index, as a second asset because analyzing stock’s covariance/correlation with a broad index is closely related to the concept of beta, a key component in the capital asset pricing model (CAPM) (e.g. see Fama and French (2004)) and its influential modifications, such as the Fama-French three-factor model (Fama and French, 1992). Beta is also widely used in practice by investors for calculating hedges against market exposure, a so-called beta-hedging strategy (Jorion, 2007).

The time period used for both assets is Dec 2, 2013 - Sep 30, 2014, a relatively recent post-financial crisis time interval of about ten months.

2.2 Data Cleaning and Limitations

The WRDS database does not provide time series of prices: sometimes there are multiple trades at different prices recorded within the same second, since many trades take less than one second to execute, and for some seconds there are no trades recorded at all. To create the second-by-second time series required for the analysis, we follow the data cleaning process of backfilling and averaging.
described in Kim (2014). First, we limit our attention to standard NYSE trading hours: between 9:30am and 4pm, since most of the trading happens during this period. All after-hours trading activity is therefore ignored due to its more erratic nature. For each second with trades at different prices, we take the price average (unweighted by volume). For each second during the trading hours with no recorded trades, we "backfill" by using the value we have recorded for the previous second.

The particular choices during data cleaning, required for constructing the time series to use as an input to our models, lose some information (in case of averaging) and introduce some non-information (in case of backfilling, since we are interpolating the price without seeing any actual trades). Another limitation of the data comes from its relatively low frequency: in high-frequency trading industry, it is common to perform analysis at micro- and even nano-second scale.

3 Describing the Price Process

3.1 Brownian Motion

After performing the data cleaning, an intraday data for a particular asset is a time series of prices

\[ Y_t : 1 \leq t \leq 23400, t \in \mathbb{Z} \]

since there are \( n = 23400 \) seconds in a trading day: \( Y_1 \) corresponds to the time interval 9:30:00 - 9:30:00:999am, and \( Y_{23400} \) to 3:59:59 - 3:59:59:999pm. Since it is not possible to distinguish when exactly during the second a trade took place, \( Y_1 \) in the time series will be obtained from prices labeled 9:30:00am, and \( Y_{23400} \) – from prices labeled 3:59:59pm.

It is common in the literature to assume (e.g. Xiu (2010)) that the intraday
log price time series \((y_t = \log Y_t)\) follows a scaled Brownian motion without drift:
\[
y_t = y_1 + \sigma B \left( \frac{t - 1}{n - 1} \right), \quad t \in [1, n]
\]
where \(B(s) \sim \mathcal{N}(0, s)\) is standard Brownian motion. We have normalized Brownian motion here so that the daily log price change (from open to close) has variance \(\sigma^2\):
\[
y_{23400} - y_1 = \sigma B \left( \frac{23399}{23399} \right) \sim \mathcal{N}(0, \sigma^2)
\]
While this is a continuous-time model, our time series are discrete; an equivalent discrete representation – snapshots of the continuous process spaced at equal time intervals – is
\[
y_{t+1} = y_t + \eta_t, \quad \eta_t \sim \mathcal{N} \left( 0, \frac{\sigma^2}{n - 1} \right), \quad 1 \leq t \leq n - 1, \quad t \in \mathbb{Z}
\]
In a bivariate time series of log prices,
\[
y_t = y_1 + \sigma B \left( \frac{t - 1}{n - 1} \right), \quad t \in [1, n]
\]
where
\[
B(s) \sim \mathcal{N}(0, sI)
\]
with \(I\) a 2x2 identity matrix, and we pick the matrix \(\sigma\) such that
\[
\Sigma = \begin{pmatrix} \sigma_1^2 & \gamma \\ \gamma & \sigma_2^2 \end{pmatrix} = \sigma \sigma^T,
\]
so that
\[
\text{Cov}(y_{23400} - y_1) = \sigma \frac{n - 1}{n - 1} I \sigma^T = \Sigma
\]
and therefore
\[
y_{23400} - y_1 \sim \mathcal{N}(0, \Sigma)
\]
Here \(\gamma\) is the daily covariance between the two assets.

The corresponding bivariate snapshots can then be described by the following discrete process:
\[
y_{t+1} = y_t + \eta_t, \quad \eta_t \sim \mathcal{N} \left( 0, \frac{\Sigma}{n - 1} \right), \quad 1 \leq t \leq n - 1, \quad t \in \mathbb{Z}
\]
\(\eta_t\) here also has the meaning of a log return \(r_{t+1} = y_{t+1} - y_t\).
3.2 Market Microstructure Noise

So far, the models provided do not account for the market microstructure noise described in the introduction. As Xiu (2010) describes, we can incorporate the market microstructure noise by introducing an additional measurement equation, in which observed price $y_t$ is different from the underlying price $\mu_t$ by the measurement error $\epsilon_t$, representing the noise:

$$y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2_{\epsilon})$$

In the bivariate case, we assume that the microstructure noise is uncorrelated, so that

$$\Sigma_\epsilon = \begin{pmatrix} \sigma^2_{\epsilon 1} & 0 \\ 0 & \sigma^2_{\epsilon 2} \end{pmatrix}$$

Letting $\sigma^2_{\eta} = \frac{\sigma^2}{n-1}$, the underlying price transition equation, as described in the previous section, is

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2_{\eta})$$

4 Moment-based Estimators

Realized Volatility and Realized Correlation are standard moment-based estimators of asset volatility and correlation, but they are biased in the presence of market microstructure noise. However, at longer time horizons, as the impact of the noise declines, both can be used as baseline measurements for asset return volatility and correlation.

4.1 Realized Volatility

Realized Volatility of the log price time series $y_1, \ldots, y_n$ is defined as the square root of the Realized Variance (RV), defined as

$$\text{RV} = \sum_{t=1}^{n-1} (y_{t+1} - y_t)^2 = \sum_{t=1}^{n-1} r^2_{t+1}$$
Hansen and Lunde (2006) have shown that Realized Variance is an unbiased estimator of $\sigma^2$ in absence of market microstructure noise (MMN), but has positive bias when MMN is present:

$$E[RV] = \sigma^2 + 2(n - 1)\sigma^2_{\epsilon}$$

### 4.2 Realized Correlation

In the bivariate case, where

$$y_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix}$$

a concept of Realized Covariance (RC) can be defined as follows:

$$RC = \sum_{t=1}^{n-1} (y_{1,t+1} - y_{1,t})(y_{2,t+1} - y_{2,t}) = \sum_{t=1}^{n-1} r_{1,t+1} \cdot r_{2,t+1}$$

Here

$$r_t \sim \mathcal{N} \left(0, \frac{\Sigma}{n - 1} \right)$$

with

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \gamma \\ \gamma & \sigma_2^2 \end{pmatrix}$$

In absence of MMN,

$$E[RC] = \sum_{t=1}^{n-1} E[r_{1,t+1}r_{2,t+1}] = \sum_{t=1}^{n-1} \frac{\gamma}{n - 1} = \gamma$$

where we have used the fact that $E[r_{1,t+1}] = E[r_{2,t+1}] = 0$. Therefore, RC is an unbiased estimator of covariance.

With MMN,

$$E[RC] = \gamma$$

which is unbiased; however, Realized Correlation combines Realized Covariance and Realized Volatility:

$$\text{Realized Correlation} = \frac{RC}{\text{Realized Volatility of 1st Asset} \cdot \text{Realized Volatility of 2nd Asset}}$$
As a consequence of Realized Volatilities being present in the denominator, market microstructure noise causes negative bias for the Realized Correlation estimator.

5 Model-based Estimators

To take into account the shortcomings of the moment-based estimators in presence of MMN, a number of model-based estimators have been proposed. The estimators we consider in this analysis are based on the maximum likelihood estimation procedure. To specify the likelihood functions, we use Kalman filter to find the relevant conditional distributions of the form $y_t|y_1, \ldots, y_{t-1}, \theta$, where $\theta$ is a vector of parameters to be estimated. The derivations in this section follow Blitzstein and Morris (2013), Harvey (1990), and Kim (2014).

5.1 Univariate Process

As described in Section 3, we assume that stock prices can be modeled via Brownian motion, so second-by-second price transitions can be modeled by the following transition equation:

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2_\eta)$$

We also assume that the market microstructure noise is independent between each second, and normally distributed, which gives us the measurement equation:

$$y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2_\epsilon)$$

We then use maximum likelihood estimation to estimate $\sigma^2_\eta, \sigma^2_\epsilon$. Implicitly, we are assuming these variances stay constant throughout the time series (one trading day). To do the estimation, we need an intermediate step of finding a predictive distribution for $\mu_{t+1}|y_t, \ldots, y_1$. We use a recursive Kalman filter derivation to
find it. Assume we know

\[ \mu_t | y_{t-1}, \ldots, y_1, \sigma^2_\eta, \sigma^2_\epsilon \sim N(a_t, P_t) \]

Then

\[ \mu_{t+1} | y_t, \ldots, y_1, \sigma^2_\eta, \sigma^2_\epsilon \sim N(a_t + K_t v_t, P_t - K_t^2 F_t + \sigma^2_\eta) \]

where

\[ v_t = y_t - a_t \]
\[ F_t = P_t + \sigma^2_\epsilon \]
\[ K_t = P_t F_t^{-1} \]

From this, we can then also find

\[ y_{t+1} | y_t, \ldots, y_1, \sigma^2_\eta, \sigma^2_\epsilon \sim N(a_t + K_t v_t, P_t - K_t^2 F_t + \sigma^2_\eta + \sigma^2_\epsilon) \]

This then allows us to express the likelihood as

\[ f(y_t, \ldots, y_1 | \sigma^2_\eta, \sigma^2_\epsilon) = f(y_t | y_{t-1}, \ldots, y_1, \sigma^2_\eta, \sigma^2_\epsilon) f(y_{t-1} | y_{t-2}, \ldots y_1, \sigma^2_\eta, \sigma^2_\epsilon) \ldots f(y_1 | \sigma^2_\eta, \sigma^2_\epsilon) \]

which we can maximize numerically using optim function in R.

### 5.2 Autocorrelated Microstructure Noise

We can now introduce dependence structure in the measurement noise term (market microstructure noise) to closer reflect the idea that, empirically, returns, and therefore log returns, are negatively correlated even when measured more than one period apart, as documented for a very similar TAQ dataset by Kim (2014).

The returns are defined as

\[ R_t = \frac{Y_t}{Y_{t-1}} \]

so that the log returns are

\[ r_t = \log Y_t - \log Y_{t-1} = y_t - y_{t-1} \]
In particular, in the previous model with i.i.d $\epsilon_t$,

$$\text{Cov}(r_t, r_{t-1}) = \text{Cov}(y_t - y_{t-1}, y_{t-1} - y_{t-2}) = \text{Cov}(\eta_{t-1} + \epsilon_t - \epsilon_{t-1}, \eta_{t-2} + \epsilon_{t-1} - \epsilon_{t-2})$$

$$= -\text{Var}(\epsilon_{t-1}) = -\sigma^2_\epsilon$$

$$\text{Cov}(r_t, r_{t-i}) = \text{Cov}(y_t - y_{t-1}, y_{t-i} - y_{t-i-1}) = \text{Cov}(\eta_{t-1} + \epsilon_t - \epsilon_{t-1}, \eta_{t-i-1} + \epsilon_{t-i} - \epsilon_{t-i-1})$$

$$= 0$$

where $i > 1$.

Now consider the following extension:

$$y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2_\epsilon)$$

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2_\eta)$$

$$\epsilon_{t+1} = \phi \epsilon_t + w_t, \quad w_t \sim \text{i.i.d } \mathcal{N}(0, \sigma^2(1 - \phi^2))$$

with $|\phi| < 1$. The particular form of variance for $w_t$ ensures that the variance of $\epsilon_t$ stays the same as in the independent noise model:

$$\text{Var}(\epsilon_1) = \sigma^2_\epsilon$$

Now, assuming $\text{Var}(\epsilon_t) = \sigma^2_\epsilon$,

$$\text{Var}(\epsilon_{t+1}) = \phi^2 \text{Var}(\epsilon_t) + \text{Var}(w_t) = \phi^2 \sigma^2_\epsilon + \sigma^2(1 - \phi^2) = \sigma^2_\epsilon$$

which inductively shows $\text{Var}(\epsilon_t) = \sigma^2_\epsilon$ for all $t$.

In this model,

$$\text{Cov}(\epsilon_t, \epsilon_{t-i}) = \phi \text{Cov}(\epsilon_{t-1}, \epsilon_{t-i}) = \phi^i \sigma^2_\epsilon \quad \text{for } i \geq 1.$$ 

So

$$\text{Cov}(r_t, r_{t-1}) = \text{Cov}(y_t - y_{t-1}, y_{t-1} - y_{t-2}) = \text{Cov}(\eta_{t-1} + \epsilon_t - \epsilon_{t-1}, \eta_{t-2} + \epsilon_{t-1} - \epsilon_{t-2})$$

$$= \phi \sigma^2_\epsilon - \phi^2 \sigma^2_\epsilon - \sigma^2_\epsilon + \phi \sigma^2_\epsilon = -(1 - \phi)^2 \sigma^2_\epsilon$$

$$\text{Cov}(r_t, r_{t-i}) = \phi^{i-1} \text{Cov}(r_t, r_{t-1})$$
To observe the negative covariance structure documented by Kim (2014), we therefore need to restrict $\phi$ to be non-negative.

The model can be written in the matrix form as

\[
y_t = Z \alpha_t, \quad Z = \begin{pmatrix} 1 & 1 \end{pmatrix}, \quad \alpha_t = \begin{pmatrix} \mu_t \\ \epsilon_t \end{pmatrix}
\]

\[
\alpha_{t+1} = T \alpha_t + V_t, \quad T = \begin{pmatrix} 1 & 0 \\ 0 & \phi \end{pmatrix}, \quad V_t = \begin{pmatrix} \eta_t \\ w_t \end{pmatrix}
\]

Also let

\[
\Omega = \text{Cov}(V_t) = \begin{pmatrix} \sigma^2_{\eta} & 0 \\ 0 & \sigma^2_{\epsilon}(1 - \phi^2) \end{pmatrix}
\]

To estimate parameters using MLE, we can use Kalman filter again to formulate the likelihood function. Assume we know

\[
\alpha_t|y_{t-1}, \ldots, y_1, \sigma^2_\epsilon, \sigma^2_\eta, \phi \sim \mathcal{N}(a_t, P_t)
\]

Then

\[
\alpha_{t+1}|y_t, \ldots, y_1, \sigma^2_\epsilon, \sigma^2_\eta, \phi \sim \mathcal{N}(Ta_t + K_tv_t, (T - K_tZ)P_tT' + \Omega)
\]

where

\[
v_t = y_t - Za_t
\]

\[
K_t = TP_tZ'(ZP_tZ')^{-1}
\]

So

\[
y_{t+1}|y_t, \ldots, y_1, \sigma^2_\epsilon, \sigma^2_\eta, \phi \sim \mathcal{N}(Za_{t+1}, ZP_{t+1}Z')
\]
5.3 Bivariate Process (BIV)

This is an extension of the univariate model with independent noise to include two stock prices (or indices) and consider their covariance.

\[ y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d } \mathcal{N}(0, \Sigma_\epsilon), \quad \Sigma_\epsilon = \begin{pmatrix} \sigma^2_{\epsilon_1} & 0 \\ 0 & \sigma^2_{\epsilon_2} \end{pmatrix} \]
\[ \mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \text{i.i.d } \mathcal{N}(0, \Sigma_\eta), \quad \Sigma_\eta = \begin{pmatrix} \sigma^2_{\eta_1} & \gamma \\ \gamma & \sigma^2_{\eta_2} \end{pmatrix} \]

Here we need to estimate five parameters: microstructure noise variances \( \sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2} \), price transition variances \( \sigma^2_{\eta_1}, \sigma^2_{\eta_2} \), and the price transition covariance \( \gamma \). For notational convenience, group all parameters into \( \theta = \{ \sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \sigma^2_{\eta_1}, \sigma^2_{\eta_2}, \gamma \} \). An implicit assumption is that the microstructure noise is uncorrelated for the two stocks.

The derivation of the likelihood also stays the same as in the univariate case, so that assuming

\[ \mu_t|y_{t-1}, \cdots, y_1, \theta \sim \mathcal{N}(a_t, P_t), \]

then

\[ \mu_{t+1}|y_t, \cdots, y_1, \theta \sim \mathcal{N}(a_t + K_tv_t, P_t - (K_tK'_t)F_t + \Sigma_\eta) \]

where

\[ v_t = y_t - a_t \]
\[ F_t = P_t + \Sigma_\epsilon \]
\[ K_t = P_tF_t^{-1} \]

From this, we can then also find

\[ y_{t+1}|y_t, \cdots, y_1, \theta \sim \mathcal{N}(a_t + K_tv_t, P_t - (K_tK'_t)F_t + \Sigma_\eta + \Sigma_\epsilon) \]

This then allows us to express the likelihood as

\[ f(y_t, \cdots, y_1|\theta) = f(y_t|y_{t-1}, \cdots, y_1, \theta) f(y_{t-1}|y_{t-2}, \cdots y_1, \theta) \cdots f(y_1|\theta) \]
To do this optimization, we can either optimize directly using optim in R, or do the two-stage estimation. In the two-stage estimation, we first estimate $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \sigma^2_{\eta_1}, \sigma^2_{\eta_2}$ based on the univariate model for each stock (using the optimization procedure for the univariate case), and then plug in those values into the likelihood function to optimize on correlation $\rho = \frac{2}{\sigma_{\eta_1}\sigma_{\eta_2}}$ Both approaches produce acceptable – although slightly different – results, but the two-stage estimation takes significantly less time to run in code, so we are performing estimations on empirical data using two-stage estimation.

5.4 Bivariate Process with Autocorrelated Noise (BIVAC)

Bivariate process with autocorrelated noise (BIVAC) is a model combining the autocorrelated noise approach and bivariate analysis of previous sections. The model is

\[
y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon), \quad \Sigma_\epsilon = \begin{pmatrix} \sigma^2_{\epsilon_1} & 0 \\ 0 & \sigma^2_{\epsilon_2} \end{pmatrix}
\]

\[
\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \text{i.i.d } \mathcal{N}(0, \Sigma_\eta), \quad \Sigma_\eta = \begin{pmatrix} \sigma^2_{\eta_1} & \gamma \\ \gamma & \sigma^2_{\eta_2} \end{pmatrix}
\]

\[
\epsilon_{t+1} = \phi \epsilon_t + w_t, \quad w_t \sim \text{i.i.d } \mathcal{N}(0, \Sigma_w), \quad \Sigma_w = \begin{pmatrix} (1 - \phi^2_1)\sigma^2_{\epsilon_1} & 0 \\ 0 & (1 - \phi^2_2)\sigma^2_{\epsilon_2} \end{pmatrix}
\]

with

\[
\phi = \begin{pmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{pmatrix}
\]

so that

\[
\text{Cov}(\epsilon_{t+1}) = \phi \text{Cov}(\epsilon_t) \phi' + \text{Cov}(w_t) = \begin{pmatrix} \phi^2_1\sigma^2_{\epsilon_1} & 0 \\ 0 & \phi^2_2\sigma^2_{\epsilon_2} \end{pmatrix} + \begin{pmatrix} (1 - \phi^2_1)\sigma^2_{\epsilon_1} & 0 \\ 0 & (1 - \phi^2_2)\sigma^2_{\epsilon_2} \end{pmatrix}
\]

\[
= \begin{pmatrix} \sigma^2_{\epsilon_1} & 0 \\ 0 & \sigma^2_{\epsilon_2} \end{pmatrix} = \Sigma_\epsilon
\]
so that the market microstructure noise covariance matrix stays constant over
time. To be able to use Kalman filter to derive the likelihood, we re-write the
model in the following way:

\[ y_t = Z\alpha_t, \quad Z = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \alpha_t = \begin{pmatrix} \mu_t \\ \epsilon_t \end{pmatrix} \]

\[ \alpha_{t+1} = T\alpha_t + V_t, \quad T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \phi_1 & 0 \\ 0 & 0 & 0 & \phi_2 \end{pmatrix}, \quad V_t = \begin{pmatrix} \eta_t \\ w_t \end{pmatrix} \]

and let

\[ \Omega = \text{Cov}(V_t) = \begin{pmatrix} \Sigma_\eta & 0 \\ 0 & \Sigma_w \end{pmatrix} \]

The Kalman derivation stays unchanged from the univariate autocorrelated noise
model, except we now have seven parameters: \( \theta = \{\sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, \sigma_{\eta_1}^2, \sigma_{\eta_2}^2, \gamma, \phi_1, \phi_2\} \).
Assume

\[ \alpha_t|y_{t-1}, \cdots, y_1, \theta \sim \mathcal{N}(a_t, P_t) \]

Then

\[ \alpha_{t+1}|y_t, \cdots, y_1, \theta \sim \mathcal{N}(Ta_t + K_tv_t, (T - K_tZ)P_tT' + \Omega) \]

where

\[ v_t = y_t - Za_t \]

\[ K_t = TP_tZ'(ZP_tZ')^{-1} \]

So

\[ y_{t+1}|y_t, \cdots, y_1, \theta \sim \mathcal{N}(Za_{t+1}, ZP_{t+1}Z') \]

and we can again optimize the likelihood in two ways, as described in Section 5.3.
It is also the case here that the two-stage estimation runs significantly faster, so
we use it during the empirical estimation.
6 Validation of Model Implementation

6.1 Independent noise

Using the univariate model with independent noise, we estimate the parameters to use as a starting point for running simulations to show correctness of implementation and characteristics of the estimator. Choosing AAPL second-by-second time series for the trading hours of May 28, 2013, we find the 1-second estimates to be

\[ \hat{\sigma}_\epsilon = 4.26 \cdot 10^{-5} \]
\[ \hat{\sigma}_\eta = 9.64 \cdot 10^{-5} \]

To ensure that optimization and inference are stable, we now simulate a sequence of 23400 observations following the univariate model with independent noise with these parameters, and use our code to again determine maximum likelihood estimates. Figures 1, 2 contain the two likelihood graphs showing what happens if we keep one of these constant at their MLE value, and vary the other, with vertical dotted lines showing the true and the estimated values.

Since the log likelihoods evaluated at the MLE and true values are within 3 of each other, we can be more confident that the implementation works properly. This heuristic assessment is due to the fact that \( 2 \cdot (\text{loglik}(\hat{\theta}_{\text{MLE}}) - \text{loglik}(\theta_{\text{TRUE}})) \) is approximately \( \chi^2(1) \)-distributed by the likelihood ratio test, so that seeing a value greater than or equal to three only occurs with a low probability: for \( X \sim \chi^2(1) \),

\[ P(X \geq 6) = 0.0143 \]

To ensure the MLE estimates are not biased, we simulate the process (of length 23,400) with the same parameters for 1000 iterations, and then plot the histograms of \( \hat{\sigma}_{\epsilon,\text{MLE}} \) and \( \hat{\sigma}_{\eta,\text{MLE}} \), with dotted lines showing their true values (that is, the ones used in the simulation). See Figures 3, 4.

The distributions are fairly symmetric around the true values.
Figure 1: Log likelihood on a simulated data (n = 23,400) using univariate model with independent noise, with respect to $\sigma_\epsilon$. The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of $\sigma_\epsilon$. 
Figure 2: Log likelihood on a simulated data (n = 23,400) using univariate model with independent noise, with respect to \( \sigma_\eta \). The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of \( \sigma_\eta \).
Figure 3: Histogram of estimated \( \hat{\sigma}_{\epsilon, MLE} \) from 1000 simulations of univariate process with independent noise with true parameters \( \sigma_\epsilon, \sigma_\eta \). The true value of \( \sigma_\epsilon \) is marked by the vertical dotted line.
Figure 4: Histogram of estimated $\hat{\sigma}_\eta, MLE$ from 1000 simulations of univariate process with independent noise with true parameters $\sigma_\epsilon, \sigma_\eta$. The true value of $\sigma_\eta$ is marked by the vertical dotted line.
6.2 Autocorrelated noise

We perform the same validation checks as for the independent noise, except here we have $\phi$, one extra parameter to estimate. It is important to remember that we limit $\phi$ to be non-negative during the optimization. The results for AAPL second-by-second on May 28, 2013 are

$$\hat{\sigma}_\epsilon = 4.93 \cdot 10^{-5}$$
$$\hat{\sigma}_\eta = 9.38 \cdot 10^{-5}$$
$$\hat{\phi} = 0.153$$

We simulate a sequence of 23400 observations using these parameters. For each variable, we vary it while keeping all others at their MLE values, to see the shape of the likelihood. The results are shown in Figures 5, 6, 7, with dotted lines showing the true and the estimated values:

The distances are within 3 from the true values, which, according to the previously discussed heuristic, indicates no serious concern for bias.

To see what the sampling distributions of the MLE estimates look like, we run multiple simulations using parameters obtained for AAPL on May 28, 2013. The histograms after 1,000 iterations of 23,400 time steps each are shown in Figures 8, 9, 12.

The distributions are fairly symmetric around the true values.

6.3 Bivariate: Independent Noise

In the two-stage estimation, we only need to confirm $\rho$ is estimated properly, since the first stage entirely relies on the univariate estimation, which we have investigated in the previous two sections. Estimating the parameters for May 28,
Figure 5: Log likelihood on a simulated data \((n = 23,400)\) using univariate model with autocorrelated noise, with respect to \(\sigma_\varepsilon\). The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of \(\sigma_\varepsilon\).
Figure 6: Log likelihood on a simulated data (n = 23,400) using univariate model with autocorrelated noise, with respect to $\sigma_\eta$. The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of $\sigma_\eta$. 
Figure 7: Log likelihood on a simulated data (n = 23,400) using univariate model with autocorrelated noise, with respect to $\phi$. The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of $\phi$. 
Figure 8: Histogram of estimated $\hat{\sigma}_\epsilon, MLE$ from 1000 simulations of univariate process with autocorrelated noise with true parameters $\sigma_\epsilon, \sigma_\eta, \phi$. The true value of $\sigma_\epsilon$ is marked by the vertical dotted line.
Figure 9: Histogram of estimated $\hat{\sigma}_\eta, MLE$ from 1000 simulations of univariate process with autocorrelated noise with true parameters $\sigma_\epsilon, \sigma_\eta, \phi$. The true value of $\sigma_\eta$ is marked by the vertical dotted line.
Figure 10: Histogram of estimated $\hat{\phi}_{MLE}$ from 1000 simulations of univariate process with autocorrelated noise with true parameters $\sigma_c, \sigma_\eta, \phi$. The true value of $\phi$ is marked by the vertical dotted line.
Figure 11: Log likelihood on a simulated data \((n = 23,400)\) using bivariate model with independent noise, with respect to \(\rho\). The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of \(\rho\).

2013 for the \((AAPL, SPY)\) pair:

\[
\hat{\sigma}_{\epsilon,AAPL} = 4.26 \cdot 10^{-5}
\]
\[
\hat{\sigma}_{\epsilon,SPY} = 6.21 \cdot 10^{-5}
\]
\[
\hat{\sigma}_{\eta,AAPL} = 9.64 \cdot 10^{-5}
\]
\[
\hat{\sigma}_{\eta,SPY} = 4.24 \cdot 10^{-5}
\]
\[
\hat{\rho}_{AAPL,SPY} = 0.328
\]

We generate a sequence of \(n = 23,400\) observations with these parameters. The log-likelihood as we vary \(\rho\) from the true value is shown in Figure 11. Here the log-likelihoods are again within three of each other, which is an encouraging
Figure 12: Histogram of estimated $\hat{\rho}_{MLE}$ from 180 simulations of bivariate process with independent noise with true parameters $\hat{\sigma}_{\epsilon,AAPL}, \hat{\sigma}_{\eta,AAPL}, \hat{\sigma}_{\epsilon,SPY}, \hat{\sigma}_{\eta,AAPL}, \hat{\sigma}_{\eta,SPY}, \hat{\rho}_{AAPL,SPY}$. The true value of $\rho$ is marked by the vertical dotted line.

result based on the previously discussed heuristic.

To have an idea about the sampling distribution of $\rho$, we simulate the data with the same parameters 180 times and make a histogram of $\hat{\rho}_{MLE}$. See Figure 12.

The distribution looks generally symmetric around the true value.
6.4 Bivariate: Autocorrelated Noise

The results for May 28, 2013 for (AAPL, SPY) are:

\[
\hat{\sigma}_\epsilon,\text{AAPL} = 4.93 \cdot 10^{-5} \\
\hat{\sigma}_\epsilon,\text{SPY} = 6.21 \cdot 10^{-5} \\
\hat{\sigma}_\eta,\text{AAPL} = 9.38 \cdot 10^{-5} \\
\hat{\sigma}_\eta,\text{SPY} = 4.24 \cdot 10^{-5} \\
\hat{\phi}_{\text{AAPL}} = 0.153 \\
\hat{\phi}_{\text{SPY}} = 0 \\
\hat{\rho}_{\text{AAPL,SPY}} = 0.338
\]

Using these values, the likelihood for estimation of \( \rho \) is shown in Figure 13.

Here the log-likelihoods are again within three of each other, which is an encouraging result based on the previously discussed heuristic.

We now run 180 simulations with parameters estimated above as true values. The histogram of the sampling distribution of \( \rho \) is shown in Figure 14.

7 Data Analysis

7.1 Optimization Issues

In my initial implementation, for AAPL on May 28, 2013, the daily variance estimate \( \sigma_\eta^2 \) and the market microstructure noise \( \sigma_\epsilon^2 \) (based on 1-sec data) were dependent on the number of observations since beginning of day that we take into account, as well as on the starting point of the optimization (see Table 1).

Since this is unstable and heavily dependent on initial point, we transform the data to optimize over \( \log(10000 \times \sigma_\epsilon) \) and \( \log(10000 \times \sigma_\eta) \) The results are significantly more stable now. See Table 2.

Including times at the beginning and at the end of the day increases volatility,
Figure 13: Log likelihood on a simulated data (n = 23,400) using bivariate model with autocorrelated noise, with respect to $\rho$. The two vertical dotted lines show the MLE value (at the peak of log-likelihood) and the true value of $\rho$.

<table>
<thead>
<tr>
<th>$(\sigma_0, \sigma_\eta)$</th>
<th>$n$</th>
<th>$\sigma_\epsilon^2$</th>
<th>$\sigma_\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.01, 0.01)</td>
<td>3000</td>
<td>$0.5 \cdot 10^{-8}$</td>
<td>$117 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.001, 0.01)</td>
<td>3000</td>
<td>$0.37 \cdot 10^{-8}$</td>
<td>$86.1 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.1, 0.05)</td>
<td>3000</td>
<td>$1 \cdot 10^{-16}$</td>
<td>$62.4 \cdot 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 1: Estimated parameters $\sigma_\epsilon^2$, $\sigma_\eta^2$ are very sensitive to choice of starting optimization point $(\sigma_0, \sigma_\eta)$. 
Figure 14: Histogram of estimated $\hat{\rho}_{MLE}$ from 180 simulations of bivariate process with autocorrelated noise with true parameters $\hat{\sigma}_\epsilon^{AAPL}, \hat{\sigma}_\eta^{AAPL}, \hat{\sigma}_\epsilon^{SPY}, \hat{\sigma}_\eta^{AAPL}, \hat{\sigma}_\eta^{SPY}, \hat{\phi}_AAPL, \hat{\phi}_SPY, \hat{\rho}_{AAPL,SPY}$. The true value of $\rho$ is marked by the vertical dotted line.

<table>
<thead>
<tr>
<th>$(\sigma_{\epsilon 0}, \sigma_{\eta 0})$</th>
<th>$n$</th>
<th>$\sigma^2_{\epsilon}$</th>
<th>$\sigma^2_{\eta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.01, 0.01)</td>
<td>3000</td>
<td>$0.66 \cdot 10^{-8}$</td>
<td>$3.25 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.001, 0.01)</td>
<td>3000</td>
<td>$0.66 \cdot 10^{-8}$</td>
<td>$3.25 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.1, 0.05)</td>
<td>3000</td>
<td>$0.66 \cdot 10^{-8}$</td>
<td>$3.25 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.1, 0.05)</td>
<td>10000</td>
<td>$0.29 \cdot 10^{-8}$</td>
<td>$2.15 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>(0.1, 0.05)</td>
<td>23400</td>
<td>$0.2 \cdot 10^{-8}$</td>
<td>$5.81 \cdot 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 2: Estimated parameters $\sigma^2_{\epsilon}, \sigma^2_{\eta}$ are no longer sensitive to choice of starting optimization point $(\sigma_{\epsilon 0}, \sigma_{\eta 0})$. They do depend on how much of the trading day we include in estimation.
as we would expect, because much of the trading (and price movements) occur during the beginning and the end of the trading day. See Kim (2014).

7.2 R, Rcpp and RcppArmadillo

When we first implemented the Kalman filtering and the optimization for independent noise in R, the univariate model-based estimation code was slow for the number of observations $n > 10,000$, so we have had to optimize the running times to make analyses at longer time horizons feasible. See Table 3.

We have used two R libraries to speed up the code. The first one, Rcpp, allows to incorporate C++ code, which runs significantly faster, into R workflow. The second one, RcppArmadillo, provides a rich range of linear algebra C++ functions to ease implementation of vector and matrix manipulations standard in R-based analyses. Both libraries are discussed in some depth in Eddelbuettel (2013).

After speeding the code up by using these two libraries to re-implement the computationally-intensive Kalman filter calculations, we got the runtimes to be orders of magnitude smaller. See Table 4.

We have implemented similar speed-ups for the bivariate models as well.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>3.9</td>
</tr>
<tr>
<td>5000</td>
<td>35</td>
</tr>
<tr>
<td>10000</td>
<td>125</td>
</tr>
<tr>
<td>23400</td>
<td>742</td>
</tr>
</tbody>
</table>

Table 3: Computation time becomes large quickly as we extend the length of the time series.
Time (s) | n
---|---
0.007 | 1000
0.026 | 5000
0.045 | 10000
0.118 | 23400

Table 4: Computation time has improved significantly after using Rcpp and RcppArmadillo.

7.3 Results and Discussion

For AAPL and SPY between Dec 2, 2013 and Sep 30, 2014, we first chart separately the assets’ average estimated volatilities ($\sigma_\eta$), obtained by taking the square root of the averages of daily $\sigma_\eta$ estimates from Realized Volatility, Bivariate model (Xiu), and Bivariate with AC model (Kim). On the x-axis is the interval, in seconds, between sampled prices used in the estimation for each method. See Figures 15, 16.

We consider the realized volatility measure found using hourly intervals as the desired estimate, only minimally affected by the market microstructure noise. These estimates are shown by dotted lines in the graphs above. Consistent with Kim’s findings, we see that at all sampling intervals, especially the shortest ones, the Realized Volatility estimates are significantly higher than the other two. Of those two, the auto-correlated result generally provides lower volatility estimates than the independent-noise one. The implication is that the higher-frequency auto-correlated result provides volatility estimates that are closer to the longer-interval measure of volatility we are interested in.

We now plot the average correlations found using the three correlation estimators: Realized Correlation, Bivariate (Xiu), Bivariate with AC (Kim). See Figure 17.

We similarly consider the realized correlation measure found using hourly
Figure 15: Comparison of the three AAPL volatility estimators, with dotted line representing the average of values obtained using Realized Volatility on data sampled hourly. The daily estimates for each sampling interval are averaged.
Figure 16: Comparison of the three SPY volatility estimators, with dotted line representing the average of values obtained using Realized Volatility on data sampled hourly. The daily estimates for each sampling interval are averaged.
Figure 17: Comparison of the three correlation estimators, with dotted line representing the average of values obtained using Realized Correlation on data sampled hourly. The daily estimates for each sampling interval are averaged.
intervals as the desired estimate, shown in the graph by the dotted line. Here the average correlation increases as the time interval goes up likely due to the Epps’ effect – the covariance is underestimated due to lack of trading in some assets during the shorter time intervals (see Epps (1979)). We here again find that at higher-frequency intervals (less than 100 seconds), \( \hat{\rho}_{BIVAC} \), the Bivariate with AC estimate, is above \( \hat{\rho}_{BIV} \), the Bivariate estimate, which is in turn above \( \hat{\rho}_{RC} \), the Realized Correlation estimate. For the shorter time intervals (less than 10 seconds), \( \hat{\rho}_{BIVAC} \) seems to provide results that are the closest to the desirable ones; its behavior at longer time intervals becomes more erratic likely due to insufficient number of data points for a reliable estimation. This is likely due to the fact that the volatility estimates from the BIVAC model are lower, and therefore closer to the longer-term Realized result, than in the other two models, so that the resulting correlation estimate is higher and also closer to the longer-term Realized estimate.

To compare all three estimators through time, we plot the one-second-based estimates for two volatilities and correlation throughout the time period. See Figures 18, 19, 20.

Based on these graphs, BIVAC is the most steady volatility estimator of the three, and its correlation estimates are consistently higher than for the other two estimators.

To look closer at the properties of the market microstructure noise autocorrelation at various intervals, we also plot average \( \phi_{AAPL} \), \( \phi_{SPY} \). See Figures 21, 22.

We observe that for both AAPL and SPY, the estimates of the autocorrelation factor \( \phi \) peak near the sampling interval of 50-100 seconds.
Figure 18: Comparison of the three daily AAPL volatility estimators through time, for the period between Dec 2, 2013 and Sep 30, 2014.
Figure 19: Comparison of the three daily SPY volatility estimators through time, for the period between Dec 2, 2013 and Sep 30, 2014.
Figure 20: Comparison of the three daily (AAPL, SPY) correlation estimators through time, for the period between Dec 2, 2013 and Sep 30, 2014.
Figure 21: Average estimates of noise autocorrelation factor $\phi_{AAPL}$ by sampling interval.
Figure 22: Average estimates of noise autocorrelation factor $\phi_{SPY}$ by sampling interval.
8 Conclusions and Future Work

This has been an exploratory analysis to get an idea of what the application of the autocorrelated factor extension model to the bivariate high-frequency time series data would yield. Using the NYSE TAQ high-frequency data set containing AAPL and SPY intraday trading data for about 10 consecutive months, we have found that the resulting estimators for both volatility and correlation produces results that are closer to those generated by moment-based estimates in lower-frequency data, when the market microstructure noise largely cancels out. This finding suggests that the autocorrelation factor estimator is less affected by the market microstructure noise in high-frequency data than some other estimators without the autocorrelation factor.

Future work may include analyzing how well the volatilities and covariances estimated using various parametric and non-parametric methods predict future volatilities and covariances. The prediction capability looks promising because the model-based higher-frequency estimates are using more of the available data than the standard moment-based lower-frequency estimates. The bivariate model itself can also be extended to apply to more than two time series at a time. With the models discussed in this analysis, while it is possible to compute the covariance matrix entries for assets in a portfolio pairwise, the resulting covariance matrix will not necessarily be positive semi-definite.

Another future direction could possibly be to investigate alternative models for volatility and correlation, with some potentially approaching estimation from a Bayesian perspective for comparison.

9 References

course in probability.


Google Finance.


Wharton Research Data Services. "NYSE Trade and Quote (TAQ),” Wharton School of the University of Pennsylvania.