

RICE UNIVERSITY

Estimating Realized Covariance using High Frequency Data

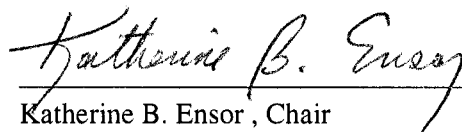
by

Lada Maria Kyj

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY

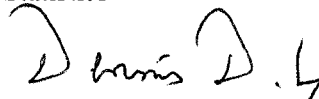
APPROVED, THESIS COMMITTEE:



Katherine B. Ensor, Chair
Professor of Statistics



Barbara Ostdiek
Associate Professor of Management and
Statistics



Dennis Cox
Professor of Statistics

Houston, Texas

May, 2008

UMI Number: 3309900

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3309900

Copyright 2008 by ProQuest LLC.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Estimating Realized Covariance using High Frequency Data

by

Lada Maria Kyj

Assessing the economic value of increasingly precise covariance estimates is of great interest in finance. We present a realized tick-time covariance estimator that incorporates cross-market tick-matching and intelligent sub-sampling. These features of the estimator offer the potential for improved performance in the presence of asynchronicity and market microstructure noise. Specifically, tick-matching preserves information when arrival structures are asynchronous, and intelligent sampling and averaging across sub-samples reduces microstructure-induced noise and estimation error. We compare the performance of this estimator with prevailing methodologies in a simulation study and by assessing out-of-sample volatility-timing portfolio optimization strategies. We demonstrate the benefits of tick time over calendar time, optimal sampling over ad-hoc sampling, and sub-sampling over sampling. Results show that our estimator has smaller mean squared error, smaller bias, and greater economic utility than prevailing methodologies. Our proposed optimized tick-time estimator improves upon both prevailing calendar-time methods and ad-hoc sampling schemes in tick time. Empirical results indicate substantial gains; approximately 70 basis points improvement against the 5 minute calendar time sampling scheme; approximately 80 basis points against optimally sampled calendar time; and 30 basis points against

tick time sampled every 5th tick. Both simulation and empirical results indicate that tick time is the better sampling scheme for portfolios with illiquid securities.

Asset allocation is inherently a high dimensional problem and estimated realized covariance matrices fail to be well-conditioned in high dimensions. As a result, the portfolios constructed are far-from optimal. Factor modeling offers a solution to both the growing computational complexity and conditioning of the covariance matrices. We find that risk averse investors would be willing to pay up to 30 basis points annually to switch from the best performing exponentially smoothed portfolio to the best performing single-index portfolio. As the number of assets increases, portfolio allocation using the single-index model is better able to replicate the benchmark index. For high-dimensional allocation problems, factor models are a more natural setting for employing realized covariance estimators.

Dedication

Dedicated to my grandparents Lydia and Vasyl Kyj and Maria and Evhen Lozynskyj.

Acknowledgments

I would like to express my deep and sincere gratitude to my thesis committee and especially my co-advisors. My co-advisor Dr. Katherine Ensor was instrumental in my decision to remain at Rice following my bachelor's degree and pursue a Ph.D. in statistics. I am thankful for her support over the years. Furthermore, I am very grateful to have had the opportunity to work under the supervision of Dr. Barbara Ostdiek and have her guidance as my co-advisor.

I would like to acknowledge the financial support I received from the Vigre program funded by the National Science Foundation under Grant DMS - 02400588, and the Jesse H. Jones Graduate School of Management. The data for this thesis was in part provided by the Center for Computational Finance and Economic Systems (CoFES) at Rice University. Computing support was provided in part by the Rice Computational Research Cluster funded by the National Science Foundation under Grant CNS-0421109, and a partnership between Rice University, AMD and Cray. I would also like to acknowledge the support of Army Research Office Grant R-14610 for my co-advisor Dr. Katherine Ensor.

My classmates have played a substantial role in my academic success and the assistance they provided at critical times is greatly appreciated. My years as a graduate student have been joyous and I attribute this to my wonderful friends, who have helped me maintain a balance in life.

Finally, I would like to express a special appreciation for the love and support of my family. My mother and father have encouraged me at every juncture of this journey and my brothers, Oles and Evhen, have always looked out for me.

Contents

Abstract	ii
Dedication	iv
Acknowledgments	v
List of Figures	ix
List of Tables	xiv
1 Introduction	1
1.1 Context	4
1.1.1 Calendar-Time Realized Covariance Estimation	4
1.1.2 Tick-Time Realized Covariance Estimation	7
1.1.3 Economic Assessment	10
1.1.4 High Dimensional Estimation	11
1.2 Contribution	12
2 The Economic Value of Cross-Market Tick-Matching Realized Covariance	14
2.1 Introduction	14
2.1.1 Model Framework	18
2.2 Data	19
2.3 Method	22
2.3.1 Tick-Time Covariance Estimation	22
2.3.2 Cross-Market Tick-Matching	24
2.3.3 Properties of Proposed Estimator	24

2.3.4	Optimal Sampling	27
2.3.5	Parameter Estimation	28
2.4	Simulation	28
2.4.1	Simulation Design	29
2.4.2	Simulation Results	30
2.5	Application	32
2.5.1	Data Analysis	32
2.5.2	Volatility Timing	35
2.5.3	Isolating A Single Hypothesis	36
2.5.4	Economic Value Results	37
2.6	Conclusion	43
3	Parsimonious Realized Portfolio Selection using High-Frequency Data	45
3.1	Introduction	45
3.2	Methods	47
3.2.1	Review of Realized Estimators	47
3.2.2	Ill-Conditioned Covariance Matrices	51
3.2.3	Rolling Estimators	52
3.2.4	Single-Index Model	55
3.2.5	Assessment Criteria	57
3.3	Empirical Analysis	59
3.3.1	Data: Dow Jones Industrial Average	59
3.3.2	Computation	61

3.3.3	Results	62
3.4	Conclusion	69
4	Derivation and Simulation of Cross-Market Tick-Matching Estimator	71
4.1	Derivation of Cross-Market Tick-Matching	72
4.2	Simulation	78
4.2.1	Simulating Brownian Motion	81
4.2.2	Dynamic Arrival Rate	82
4.2.3	Brownian Motion with Jumps	83
4.3	Results	84
5	Conclusion	90
5.1	Future Work	91
5.1.1	Tracking Error Minimization	91
5.1.2	Time-of-day Covariance Estimation	92
5.1.3	Generalized Market Microstructure Noise	92
5.1.4	Covariance Estimation in the Presence of Jumps	92
A	Filtering	94
A.1	Quotes	94
A.2	Trades	95
A.3	Quote Trade Matching	95

List of Figures

1.1	Size of Trades and Quotes (TAQ) database of the New York Stock Exchange (NYSE) contains all recorded trades and quotes on NYSE, AMEX, NASDAQ, and the regional exchanges from 2000 to 2006. Derived from Yan (2007)	2
1.2	Illustration of sub-sampling. Two sub-grids, top and bottom, and two estimates are calculated for each set. Finally, the top and bottom estimates are averaged.	6
2.1	Stock prices from January 1, 2002 to December 31, 2006 for Exxon Mobil Corp. (XOM), Occidental Petroleum Corp. (OXY) and The J.M. Smucker Company (SJM). Prices have been adjusted for stock splits.	21
2.2	Noise-to-Signal Ratio. XOM has the lowest noise-to-signal ratio. All three stocks show a decrease in noise-to-signal ratio corresponding to an increase in quote frequency.	23
2.3	Visualization of Cross-Market Tick-Matching. The slower process (A), represented by squares, is used as the base. The observations of the faster process (B), represented by circles, are matched to the corresponding observations of (A). By construction this estimator has overlapping intervals, which are represented by the grey rectangles.	25

2.4	Bias of realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. Tick-time estimators are unbiased for low noise-to-signal settings. For low quote frequency settings, calendar-time estimators display a downward bias. As the quote frequency increases, the ad-hoc tick time estimator displays a positive bias.	30
2.5	Mean Squared Error of realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. Tick-time estimators have smallest MSE, but the difference relative to calendar-time estimates is less pronounced as the noise-to-signal ratio increases.	31
2.6	Mean Squared Error of sub-sampled realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. The sub-sampled estimators have much smaller MSE than the MSE presented in Figure 2.5.	31
2.7	Tick-time Variance Sampling Frequency	33
2.8	Tick-time Covariance Sampling Frequency	33
2.9	Estimated σ	34
2.10	Estimated ρ	34
2.11	A realization of portfolio weights using a volatility-timing strategy. Presents results for CT5, CTO-S, TT5-S, and TTO-S.	40
3.1	Partition of Covariance Matrix into sub-matrices of dimension 5. Each sub-matrix is run independently. The computation is accelerated further by computing each year independently. This allows for parallel computation in $\frac{1}{105}$ of the time necessary when using a serial technique.	62

4.1	Aggregated Tick-Time Estimation. Grey rectangles represent a traditional Hayashi Yoshida tick-time estimator; the sum of the product of non-empty overlapping intervals. The dashed lines are a visual representation of aggregation scheme, which matches the ticks of the faster process (B) to the corresponding ticks of the slower process (A).	72
4.2	Number of quotes observed per minute. X-axis is the time of day, with the market opening at 9:30 and closing at 16:00. The Y-axis is the number of quotes observed per minute. A diurnal effect is observed with more quotes at the start and conclusion of trading day, and comparatively fewer quotes during the middle of the day.	80
4.3	Realization of Bivariate Brownian Motion. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously.	82
4.4	Realization of Bivariate Brownian Motion with Dynamic Arrival Rates. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously. The dynamic arrival rate is evidenced by the relatively fewer observations during the middle of the day.	83

4.5	Realization of Bivariate Brownian Motion with Jumps. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously. We see a large jump in the grey process within the dashed rectangle.	85
4.6	Bias of realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. The x-axis is the noise-to-signal ratio and the y-axis is the bias of the estimators. The optimized tick-time estimator, represented by the black line, is the least biased in all three frequency settings.	86
4.7	Mean Squared Error of realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. Tick-time estimators have the smallest MSE, but the difference relative to the calendar-time estimators is less pronounced as the noise-to-signal ratio increases.	87
4.8	Mean Squared Error of sub-sampled realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. The sub-sampled estimators have much smaller MSE than in Figure 4.7.	87

4.9	Bias of realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. The x-axis is the noise-to-signal ratio and the y-axis is the bias of the estimators. The bias of the optimized tick-time estimator, represented by the black line, increases as the noise-to-signal ratio increases.	88
4.10	Mean Squared Error of realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. Tick-time estimators have the smallest MSE for low noise-to-signal ratios, but are no longer the smallest as the noise-to-signal ratio increases.	88
4.11	Mean Squared Error of sub-sampled realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. The sub-sampled estimators have much smaller MSE than in Figure 4.10.	89
A.1	Example of limitation of quotes only filtering and how trade quote matching mitigates problem. Data is IBM on March 11-12, 2002.	95
A.2	Criteria for Filtering Quotes with respect to Trade	96

List of Tables

2.1	Summary statistics of stock returns from January 1, 2002 to December 31, 2006. Table reports market capitalizations of stocks, annualized mean μ and standard deviation σ of returns, noise-to-signal ξ/σ ratios for the returns, and correlation of daily returns.	21
2.2	Annual averages of daily admissible quotes. The number of quotes increases over time for all three stocks. SJM is relatively less frequently quoted than XOM or OXY.	22
2.3	Realized covariance estimators considered in simulation. Calendar time is abbreviated as (CT) and tick time is abbreviated as (TT).	29
2.4	Summary of Open-to-Close Returns. Presents annualized mean and standard deviations of open-to-close returns, estimated Sharpe ratio, μ/σ , and correlation matrix. The last line presents an equally weighted portfolio. . .	35
2.5	Summary results for portfolio containing XOM, OXY, and SJM. Target return is $\mu_P = 7.5\%$. Presents annualized means, standard deviations, and Sharpe ratios for portfolio returns using 8 different covariance estimators. .	38
2.6	Annualized Basis Points for portfolio containing XOM, OXY, and SJM. Target return is $\mu_P = 7.5\%$. Investor risk aversion increases with γ	38
2.7	Summary results for disaggregated portfolio pairs. Target return is $\mu_P = 7.5\%$. Presents annualized means, standard deviations, and Sharpe ratios for portfolio returns using 8 different covariance estimators.	41

2.8	Annualized Basis Points for disaggregated portfolio pairs. Target return is $\mu_P = 7.5\%$. Investor risk aversion increases with γ	42
3.1	Composition of Dow Jones Industrial Average from January 1, 2002 to December 31, 2006.	60
3.2	Optimal weight parameters for 8 different realized covariance estimators. Weights are smallest for ad hoc calendar-time estimators. Optimal weight parameters for variance only estimators are larger than corresponding covariance estimators.	63
3.3	Performance Characteristics of Global Minimum Variance (GMV) portfolios using 5 stocks.	66
3.4	Performance Characteristics of Global Minimum Variance (GMV) portfolios using the 30 stocks within the Dow Jones Industrial Average.	67
3.5	Performance Characteristics of maximum Sharpe Ratio Minimum Variance (SRMV) portfolios using the 30 stocks of the Dow Jones Industrial Average where $\mu_P = 5\%$	68
4.1	Realized covariance estimators considered in simulation. Calendar time is abbreviated as (CT) and tick time is abbreviated as (TT).	81

Chapter 1

Introduction

Assessing the economic value of increasingly precise covariance estimates is of great interest in finance. Adoption of high-frequency data provides opportunities for better inference of market behavior, but at the same time necessitates the development of more sophisticated computational methods. We present a realized tick-time covariance estimator that incorporates cross-market tick-matching and intelligent sub-sampling. These features of the estimator offer the potential for improved performance in the presence of asynchronous observations and market microstructure noise. We compare the performance of this estimator with prevailing methodologies in a simulation study and by assessing out-of-sample volatility-timing portfolio optimization strategies. Results show that our estimator has smaller mean squared error, smaller bias, and greater economic utility than prevailing methodologies. For high-dimensional allocation problems we address the problem of ill-conditioned covariance matrices by considering the performance in the settings of rolling regression and factor models. We conclude that factor models are a more natural setting for employing realized covariance estimators.

High-frequency financial data are observations recorded at the finest time scale, often at the transaction level. These data sets have been used to study market microstructure and realized covariance. Market microstructure is concerned with characterizing the moment-to-moment structure of exchanges. Realized covariance estimation utilizes high frequency data and provides considerable efficiency gains in measuring variance, which is an essential statistic for trading decisions, risk measurement, and portfolio optimization. Two problems

emerge with the adoption of high frequency data: 1) market microstructure effects contaminate the underlying process, and 2) observations are asynchronous and irregularly spaced. Our proposed estimator addresses both of these problems.

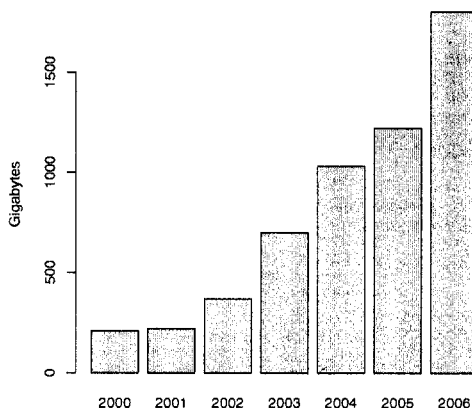


Figure 1.1 Size of Trades and Quotes (TAQ) database of the New York Stock Exchange (NYSE) contains all recorded trades and quotes on NYSE, AMEX, NASDAQ, and the regional exchanges from 2000 to 2006. Derived from Yan (2007)

High frequency databases have been available for over two decades, and the analysis of this data continues to evolve. Wood (2000) chronicles the development of this class of data; from the Fitch data in the early 80's, to ISSM data, to the release of the Trades and Quotes (TAQ) database in 1993. The TAQ database contains all recorded trades and quotes on the New York Stock Exchange (NYSE), AMEX, NASDAQ, and the regional exchanges from 1992 to present. Figure 1.1, derived from Yan (2007), shows the growth in the TAQ database and that the number of quotes has changed dramatically in recent years. As confirmed in a study by Hansen and Lunde (2006) properties of market microstructure have changed substantially over (2000-2004). This coincides with the NYSE's phased transition from fractional to decimal pricing completed in February 2001.

Market microstructure describes the features of the interaction between buyers and sellers of a financial asset. A survey by Madhavan (2000) identifies three sources of market frictions which result in departures from the latent price process assumption: 1. trading frictions, 2. private information, and 3. alternative trading structures. Economic theory of market microstructure is discussed in O'Hara (1995) and Hasbrouck (2007) provides a discussion of empirical findings. Although market microstructure effects capture the impact of trading mechanisms on the price formation process, from a statistical perspective we can view market microstructure effects as "observation error", the deviation of the observed values from the hypothesized latent price process. In the remainder of this thesis we refer to the market microstructure as MM noise.

Estimating covariance is challenging as: 1) the continuous time covariance process is observed only at discrete times, 2) asynchronicity introduces the so called Epps effect, where the realized covariance estimator is biased towards zero as the sampling frequency increases, and 3) in high dimensional settings, the realized covariance matrix is ill-conditioned. Numerous methods have been proposed to solve the problem of discrete observations of continuous time processes. See Sorensen (2004) or Fan (2005) for survey papers. The consideration of multiple processes, arriving asynchronously, and contaminated with market microstructure noise further complicates the estimation problem. See Renò (2003) for a detailed analysis of the Epps effect. This problem is more pronounced for sparse data, such as generated by less actively traded assets.

Andersen, Bollerslev, Diebold, and Labys (2001) developed nonparametric estimators, termed *realized* estimators, which estimate the covariance matrix from the sum of squares or cross-products of returns without consideration for market microstructure noise. Only

recently have works on realized covariance begun to address the role market-microstructure plays in covariance estimation. The estimation of realized covariance of asynchronous observations has largely relied upon calendar-time methods of synchronization. Hayashi and Yoshida (2005) developed a tick matching covariance estimator that does not rely on ad-hoc synchronization and this has generated a number of open questions in the estimation of the realized covariance. Examining the performance of these estimators has developed into an active research area. As evidenced by Bandi, Russell, and Zhu (2008), Bandi and Russell (2006), Zhang, Mykland, and Ait-Sahalia (2005), Voev and Lunde (2007), Griffin and Oomen (2006), most studies have been limited to low dimensional settings ($p \leq 3$). Higher dimensional studies, such as de Pooter, Martens, and van Dijk (2006), have addressed the ill-conditioned covariance matrix problem by employing rolling regression techniques.

1.1 Context

1.1.1 Calendar-Time Realized Covariance Estimation

The discretely observed price process $p(t_i)$ is a function of both the latent price process and the market microstructure effects. Andersen, Bollerslev, Diebold, and Labys (2001), introduced nonparametric estimators of the quadratic variation and covariation, as they estimate the unobserved quantities from the sum of squares or cross-products of returns. This methodology first creates a homogenous time series by sampling the observations at equally spaced intervals. Specifically, we allow for m equally spaced intraday observations within the daily time span of 0 as the start of the day and T as the terminal time of the day.

We define return of asset A as:

$$r_{A(m)}(t_i) \equiv p_A(t_i) - p_A(t_i - \frac{1 * T}{m}) \quad t_i = \frac{1 * T}{m}, \frac{2 * T}{m}, \dots, T. \quad (1.1)$$

The resulting realized variance and realized covariance are of the form:

$$\hat{V}_A(m) = \sum_{i=1}^m r_{A(m)}^2(\frac{i}{m}T) \quad (1.2)$$

and

$$\hat{C}_{AB}(m) = \sum_{i=1}^m r_{A(m)}(\frac{i}{m}T) \times r_{B(m)}(\frac{i}{m}T). \quad (1.3)$$

Asymptotic distribution theory for these estimators is developed in Barndorff-Nielsen and Shephard (2002, 2004a), who show that they are consistent estimators that converges to the true variance and covariance at a rate of \sqrt{m} .

Realized covariance estimators require synchronous observations and this is achieved by interpolating prices onto an ad-hoc common sampling grid, i.e every 5 minutes. This creates an artificial dependency on the choice of the granularity of the sampling grid. Popular interpolation scheme include previous tick and linear interpolation. Dacorogna, Gencay, Muller, Olsen, and Pictet (2001) provide a survey of these methods and highlight inherent problems. The basic issue is a balance between the bias introduced by asynchronicity and the reduction in variance.

Eliminating dependency on ad-hoc gridding schemes motivates the discussion of finding optimal sampling frequencies. Bandi and Russell (2006) consider optimal sampling and present a technique for separating the microstructure noise from the realized covariance. They present an optimal sampling frequency which is a function of the signal-to-noise ratio. This suggests that a variant of signal-to-noise ratio should be used to sample the series, and hence different assets will should be sampled at different frequencies.

Zhang, Mykland, and Ait-Sahalia (2005) present a similar optimization of the sampling frequency methodology for calculating realized volatility in the presence of microstructure noise. Moreover, where sampling throws away most of the data, Zhang, Mykland, and Ait-Sahalia (2005) introduce sub-sampling as a method for preserving the richness of high frequency data.

For variance reduction Zhang, Mykland, and Ait-Sahalia (2005) advocate sub-sampling and averaging. This is carried out by dividing the time domain grid into K non-overlapping subgrids, and then average of the estimates calculated over the different subgrids. Figure 1.2 illustrates the intuition behind sub-sampling. This example includes two subgrids, top and bottom. Estimates are calculated for both the top and bottom sets, and then the two estimates are averaged. Zhang, Mykland, and Ait-Sahalia (2005) present Monte Carlo evidence of the performance of different estimation strategies and attribute substantial bias reduction to sparse sampling. They attribute variance reduction to sparse sampling and/or sub-sampling.

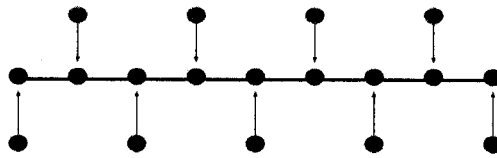


Figure 1.2 Illustration of sub-sampling. Two sub-grids, top and bottom, and two estimates are calculated for each set. Finally, the top and bottom estimates are averaged.

Hansen and Lunde (2006) discuss optimal sampling of realized variance in the context of different noise structures. Specifically, their data analysis rejects the notion that noise is independent and identically distributed (i.i.d.), and instead they advocate for autocorrelated noise that is dependent on price. Under this noise specification Hansen and Lunde (2006)

advocate a bias-correcting kernel estimator introduced by Zhou (1996) that incorporates the first order autocovariance. This proposed kernel estimator is shown to reduce the bias due to noise and motivates the discussion of lead-lag estimators. Examples of kernel-type realized covariance estimators include Griffin and Oomen (2006), Voev and Lunde (2007) and Hansen, Large, and Lunde (2006).

Bandi and Russell (2005) extends their univariate work and translate their methodology into optimal sampling for covariance. The Bandi and Russell optimal sampling frequency can be understood as the ratio of the second moment of the cross product of price returns over the squared second moment of the cross product of the noise process. De Pooter, Martens, and van Dijk (2006) is a closely related study, where realized covariances are not only optimally sampled, they are also sub-sampled and then bias corrected. De Pooter, Martens, and van Dijk (2006) also consider lead lag covariances to control the bias. Zhang (2006) provides a comprehensive discussion of the Epps effect and provides an optimal sampling scheme that balances the trade-off between bias and the numerous sources of error.

1.1.2 Tick-Time Realized Covariance Estimation

Sampling schemes emerge as another important issue in determining optimal sampling frequencies. Ait-Sahalia, Mykland, and Zhang (2005) considers non-uniform sampling intervals, including randomly spaced intervals. Their chief finding is that once the noise structure is modeled, then it is optimal to sample as often as possible, independent of arrival structure. Oomen (2006) compares different time scales, calendar time (CT), and tick time sampling (TT), and finds that TT provides optimal results with respect to the MSE criteria.

He credits this to the location of the sampling points that generate the information set. Hansen and Lunde (2006) also consider different sampling schemes, and accordingly find favorable results using tick time instead of calendar time.

Hayashi and Yoshida (2005), HY hereafter, introduce a *cumulative covariance estimator* that is a transaction time estimator and does not require any artificial grids. The HY estimator is the sum of the cross product of returns of non-empty overlapping intervals. In the absence of noise this estimator is unbiased and consistent, moreover the HY estimator is consistent when observations are asynchronous. This estimator is similar to the estimator introduced by de Jong and Nijman (1997). The de Jong estimator regresses the cross-covariance on the length of the overlapping interval, and is limited to a discrete time model with stationary increments. Hayashi and Yoshida claim that their estimator should display properties of unbiasedness and consistency even if Brownian motions are replaced with Levy processes, as the key properties in their proofs are independence of increments and finiteness of moments. The main contribution of this estimator is correcting for the bias introduced by asynchronicity. This advance can be understood as intelligently matching ticks across assets. The introduction of MM noise renders this estimator inconsistent and dependent on the noise structure, it may also be biased. An additional issue with this estimator is the computational intensity.

For the sake of computational efficiency Palandri (2006) develops an aggregation scheme for the HY estimator and shows that this scheme preserves the informational content of the original HY estimator. Voev and Lunde (2007) also propose an aggregated HY estimator where the ticks of the faster process are matched to the ticks of the slower process.

Griffin and Oomen (2006) compare the bias and variance of a calendar-time realized covariance estimator and the HY estimator in the presence of asynchronous observations arriving with a Poisson process. The microstructure noise is assumed to be independent and identically distributed (i.i.d.) and independent of the efficient price process. Griffin and Oomen conclude that under i.i.d. noise both estimators are unbiased but inconsistent, optimal sampling frequencies are determined by the slowest process, and that relative performance is a function of the noise-to-signal ratio. For low noise-to-signal ratios the HY is ideal, but for high noise-to-signal ratios an optimally sampled calendar-time realized covariance estimator may perform better. Griffin and Oomen suggest sampling both processes with the same sampling frequency, which is counter to the approach of this thesis.

Moreover the Voev and Lunde (2007) study examines the HY estimator where the MM noise is contemporaneously and serially correlated with the price process. An extensive simulation study is conducted where the MSE of the covariance estimators of different arrival structures are compared under different noise structures. In the absence of noise, the original HY estimator is optimal with respect to MSE. The introduction of i.i.d. noise makes an ad-hoc sub-sampled HY optimal. In the case of stochastic correlation and correlated noise, the ad-hoc sub-sampled HY and an ad-hoc sub-sampled kernel type HY estimators are most competitive. Results suggest that the greatest gains are achieved by sub-sampling. The study suggests optimizing the sub-sampling frequency as a topic of interest.

1.1.3 Economic Assessment

Andersen, Bollerslev, Diebold, and Labys (2001) find that assets realized covariances display persistence and this slow rate of decay is preserved even under temporal aggregation. These findings motivate the use of volatility-timing strategies for assessing the performance of different variance and covariance estimators.

Fleming, Kirby, and Ostdiek (2001, 2003) popularized volatility-timing as an objective methodology for assessing the economic “value-added” of alternative estimators. “Value-added” is understood in a portfolio framework as the amount a risk-averse investor would be willing to pay to capture the observed gains in portfolio performance. Specifically, investors follow a volatility-timing strategy where the portfolio weights vary only with changes in estimates of the conditional covariance matrix of daily returns. Markowitz mean-variance (MV) optimization is the standard theoretical framework for optimal portfolio construction followed in this research. A quadratic utility function is used to differentiate between portfolios generated using different estimators. The incremental value of using the one estimator instead of another is calculated by finding the constant which makes the two utilities equal.

Bandi and Russell (2006) evaluate their optimally sampled realized variance estimator using the volatility-timing strategy presented in Fleming, Kirby, and Ostdiek (2003). They find that a risk-averse investor is willing to pay between 25 and 300 basis points per year to switch from fixed intervals to optimally sampled intervals. Furthermore, Bandi, Russell, and Zhu (2008) evaluate a related optimally sampled covariance estimator, and find that an investor would be willing to pay around 80 basis points a year to change from fixed

intervals to optimal intervals to achieve this superior level of performance.

1.1.4 High Dimensional Estimation

Markowitz mean-variance (MV) optimization is the standard theoretical framework for optimal portfolio construction. MV optimization requires covariance matrices to be not only invertible, but also well-conditioned. Realized covariance estimation has emerged as a viable candidate for covariance estimation. This class of estimators employs high-frequency data and provides more precise estimates. In a low dimensional setting, it has been shown that realized covariance estimators provide utility gains over traditional GARCH approaches based on daily data for risk averse investor following a MV optimization strategy.

Estimation of high dimensional covariance matrices is computationally expensive. As the number of dimension increases, the number of operations increases quadratically. Realized covariance estimates also become numerically ill-conditioned due to small effective sampling sizes. As a result the inversion of the matrix, a necessary step in mean-variance optimization, becomes problematic. This is a paradox of high frequency data, at first glance it appears as there is “too much data”, but once asynchronicity and market microstructure effects are acknowledged, then once again we are confronted with excessive sampling errors.

Previous literature has addressed imprecise covariance matrix estimates by imposing more structure on the covariance matrix. The single-factor model is one possible alternative. It has the advantage of reducing the dimension of the estimation problem, but also a disadvantage of possibly not capturing all of the covariation and thereby resulting in biased

estimates. In addition, Andreou and Ghysels (2002) suggest rolling regression as a more structured estimator. We will examine the performance of realized covariance estimators in both of these settings.

1.2 Contribution

This thesis has addressed realized covariance estimation using high frequency data, an issue that is of great interest in finance. Market microstructure effects and asynchronous observations necessitate the development of more sophisticated computational methods. We present a realized tick-time covariance estimator that incorporates cross-market tick-matching and optimal sub-sampling. The bias and variance properties of this Cross-Market Tick-Matching Estimator are derived. We compare the performance of this estimator with prevailing methodologies in a simulation study and by assessing out-of-sample volatility-timing portfolio optimization strategies. Results show that our estimator has smaller mean squared error, smaller bias, and greater economic utility than prevailing methodologies. For high-dimensional allocation problems, factor models are a more natural setting for employing realized covariance estimators. We consider parsimonious realized portfolio selection and conclude that the Cross-Market Tick-Matching Estimator provides as good as or better levels of utility within a single-factor model setting as any other competing estimator when using a fully estimated covariance matrix. The computational efficiency of the factor-model is a very desirable feature and suggests a practical application for our proposed estimator. The plan of the thesis is as follows:

Chapter 2 develops and assesses an optimally sub-sampled tick-time realized covariance estimator. The properties of this proposed estimator are assessed in a simulation study.

In addition, a volatility-timing framework, as popularized by Fleming, Kirby, and Ostdiek (2001, 2003), is used to empirically assess the economic value of this proposed estimator with respect to prevailing methodologies. We assess a small portfolio with only three assets and conclude that our proposed estimator performs well in the presence of infrequently quoted assets and provides risk averse investors with utility gains.

Chapter 3 considers a parsimonious method for portfolio allocation of a high dimensional problem. The realized covariance estimator suffers the curse of dimensionality as the number of dimensions increases. In particular the realized covariance matrix becomes ill-conditioned and fails to provide optimal portfolio weights. We compare the performance of exponentially weighted covariance matrices against covariance matrices generated by a single-index model. The single-index model has an attractive feature of being far less computationally intensive. Using the Dow Jones Industrial Average, we find that the single-index model is a plausible alternative to estimating the full covariance matrix as it provides similar utility gains for a risk averse investor.

Chapter 4 presents the derivation and properties of the Cross-Market Tick-Matching estimator as well as simulated results. We find that tick-time estimators perform very well in the presence of dynamic arrival rates. We also show a limitation of tick-time estimation as it performs poorly in the presence of large noise-to-signal disturbances further contaminated by asynchronous jumps. This motivates future research into accommodating for jumps in realized covariance estimation.

Finally, Chapter 5 summarizes the finding within this thesis and outlines four avenues for future research.

Chapter 2

The Economic Value of Cross-Market Tick-Matching Realized Covariance

2.1 Introduction

Asset prices can be understood as discrete time observations of underlying continuous time price processes. Realized covariance is a non-parametric estimate of the covariance structure obtained by summing the squares of returns, for variance elements, and cross-products of returns, for covariance elements. Adoption of high-frequency data is desirable under the assumption of continuous time price processes, because more frequent sampling mitigates the estimation error caused by discrete observations. Noise from market microstructure effects, however, can be an important component of high frequency returns. As a result of these competing factors, the sampling frequency can greatly influence the performance of realized covariance estimators. Sparse sampling provides insufficient information improvement over daily data, but sampling too frequently can cause market microstructure effects to dominate the covariance estimates. Hence, the adoption of high-frequency data provides opportunities to better model and understand market behavior, but it requires the development of more sophisticated methods.

Past literature largely focuses on estimating volatility and addresses many of the concerns raised by microstructure noise contamination of the univariate series. Andersen, Bollerslev, Diebold, and Labys (2001) first proposed realized variance estimation using ad-hoc calendar-time sampling. Recent literature on realized volatilities includes works by Dacorogna, Gencay, Muller, Olsen, and Pictet (2001), Andreou and Ghysels (2002), Bandi and Russell (2006), Ghysels, Santa-Clara, and Valkanov (2006), and Hansen and Lunde

(2006). Asymptotic distribution theory for these estimators is developed in Barndorff-Nielsen and Shephard (2002, 2004a). A limitation of calendar-time sampling is the need for synchronous observations across markets. Under the standard methodology this is achieved by interpolating prices onto an ad-hoc common sampling grid i.e, every 5 minutes. This causes realized covariance estimates to be vulnerable to the so called “Epps effect” (Epps 1979), where the covariation estimate converges to zero as the sampling grid gets finer. (See Renò (2003) and Zhang (2006).)

Eliminating dependency on ad-hoc calendar-time grid intervals motivates determining the optimal sampling frequency by minimizing the mean squared error (MSE) of the realized covariance estimator. The optimal sampling frequency is a function of the return series’s signal-to-noise ratio, implying that different assets should be sampled at different frequencies. Recent literature on optimal sampling frequencies for calendar-time estimators includes works by Bandi and Russell (2006), Zhang, Mykland, and Ait-Sahalia (2005), and Griffin and Oomen (2006). Sampling, however, considers only a small portion of the data. Zhang, Mykland, and Ait-Sahalia (2005) advocate sub-sampling and averaging as a technique for exploiting the richness of high frequency data. They divide the time domain grid into K non-overlapping subgrids and average the estimates over the K different subgrids to calculate the final estimate.

When sampling high-frequency data, choosing tick time or calendar time is another important issue. Operating in tick time samples the price process according to changes in the level of market activity. As a result, tick-time sampling offers superior location of the sampling points that generate the information set. Oomen (2006) finds that tick-time sampling yields lower MSE in the univariate setting. Recent literature on tick-time estimation

includes Hansen and Lunde (2006), Garcia and Meddahi (2006), Ait-Sahalia, Mykland, and Zhang (2005), and Griffin and Oomen (2008). Hayashi and Yoshida (2005) and Corsi (2006) develop a tick-time covariance estimator (HY hereafter) that does not rely on ad-hoc synchronization, but leaves open how to determine the optimal sampling frequency. Griffin and Oomen (2006) and Voev and Lunde (2007) evaluate the performance of the HY estimator against prevailing calendar-time estimators. Griffin and Oomen (2006) determine an optimal sampling frequency and sample both processes independently at this frequency; a setup that fails to address asynchronicity. Voev and Lunde (2007) introduce a cross-market tick-matching algorithm, but they implement it with ad-hoc tick-time sampling.

We address the implementation issues introduced by Hayashi and Yoshida (2005) by presenting an optimally sub-sampled realized covariance estimator via a cross-market tick-time algorithm. Working in transaction time rather than calendar time, we can construct a Cross-Market Tick-Matching (CMTM) estimator that better addresses non-synchronous price observations. We determine the optimal sampling frequency for the CMTM estimator with respect to the mean squared error (MSE) criterion. Finally, we exploit the richness of high frequency data by generating multiple sampling sets (via sub-sampling) and averaging the resulting estimators. Our proposed estimator differs from existing methods as the level of trading activity determines the optimal sampling frequency. We also provide empirical analysis that highlights the advantages of operating in tick time when estimating covariances for portfolios containing less active stocks.

This chapter provides three primary contributions: 1) derivation of a Cross-Market Tick-Matching covariance estimator, 2) optimal sub-sampling of this CMTM estimator, and 3) quantification of the incremental contribution of each enhancement to the covariance

estimator. In addition to conducting simulation experiments, we use the volatility-timing strategies popularized by Fleming, Kirby, and Ostdiek (2001, 2003) to objectively assess the economic value of the CMTM estimator with respect to prevailing methodologies. In this setting, we demonstrate the benefits of tick time over calendar time, optimal sampling over ad-hoc sampling, and sub-sampling over sampling.

As demonstrated in Bandi and Russell (2006), de Pooter, Martens, and van Dijk (2006), and Fleming, Kirby, and Ostdiek (2003), calendar-time covariance estimators using high frequency data have provided considerable efficiency gains in measuring return volatility, an essential statistic for trading decisions, risk measurement, and portfolio optimization. Our proposed optimized tick-time estimator improves upon the prevailing calendar-time methods and upon ad-hoc sampling schemes in tick time. We realize approximately 70 basis points improvement against the five-minute calendar time sampling scheme, approximately 80 basis points against an optimally sampled calendar-time estimator, and 30 basis points against a tick-time estimator sampled every fifth tick. Both simulation and empirical results indicate that tick time is the better sampling scheme for portfolios with less active securities.

This chapter is organized as follows. Section 2.1.1 outlines the theoretical framework for realized covariance estimators. Section 2.2 presents our data. Our proposed estimator is presented in Section 2.3. Section 2.4 discusses the Monte Carlo simulation results. Section 2.5 presents the results of the volatility-timing trading strategy. Section 2.6 concludes and suggests avenues for future research. The Appendix A.1 provides details of the data filtering technique.

2.1.1 Model Framework

To begin, let $X_t = [x_{(1,t)}, x_{(2,t)}, \dots, x_{(n,t)}]$ be an n -dimensional vector of log price processes define as:

$$X_t = \int_0^t \mu(s)ds + \int_0^t \Theta(s)dW(s). \quad (2.1)$$

Here, $\mu(s)$ is the n -dimensional vector of drift coefficients, $\Theta(s)$ is a $n \times n$ matrix satisfying $\Theta(s)\Theta(s)' = \Sigma(s)$, and W is an n -dimensional Brownian motion. The *instantaneous covariance*

$$\Sigma_{A,B}(s) = \begin{bmatrix} \sigma_A^2(s) & \sigma_A \sigma_B \rho(s) \\ \sigma_A \sigma_B \rho(s) & \sigma_B^2(s) \end{bmatrix}, \quad (2.2)$$

yields the cumulative covariance between processes A and B, such that for all $t < \infty$:

$$\int_0^t \Sigma_{A,B}(s)ds < \infty. \quad (2.3)$$

In the continuous time setting, we can assume that $\mu = 0$ because the estimate of μ is small relative to the estimation error. This assumption holds for sufficiently small increments in the discrete time setting. We are interested in estimating the quadratic or *cumulative* covariation between process A and process B which is defined as:

$$V_A = \int_0^t \Sigma_A(s)ds \quad (2.4)$$

$$C_{A,B} = \int_0^t \Sigma_{A,B}(s)ds. \quad (2.5)$$

The presumed latent price process fails to fully capture market microstructure effects. Hence the discretely observed price process $P(t_i)$ is a function of both the latent price process $X(t_i)$ and the market microstructure effects $u(t_i)$, which are treated as “observation error” such that the price of asset A is observed as:

$$p_A(t_i) = x_A(t_i) + u_A(t_i), \quad i = 1, 2, \dots, n. \quad (2.6)$$

Hence, returns are written as

$$r_A(t_i) = \Delta p_A(t_i) = \Delta x_A(t_i) + \Delta u_A(t_i). \quad (2.7)$$

Define the noise terms to be contemporaneously correlated such that:

$$\begin{bmatrix} u_A \\ u_B \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \xi_A^2 & \xi_{A,B} \\ \xi_{A,B} & \xi_B^2 \end{bmatrix} \right) \quad (2.8)$$

and assume that noise is uncorrelated with the latent price process $X(t_i)$. In addition, the noise is assumed to be uncorrelated with non-contemporaneous noise terms.

2.2 Data

High frequency databases have been available for over two decades and the analysis of this data continues to evolve. Wood (2000) chronicles the development of this class of data from the Fitch data in the early 1980's, to the Institute for the Study of Securities Markets (ISSM) data, to the release of the Trade and Quote (TAQ) database in 1993. Academic research in both market microstructure and covariance modeling rely heavily on high frequency data. Market microstructure literature is concerned with characterizing the moment-to-moment structure of exchanges affecting trade and quote dynamics. This includes market structure for transactions, agents facing asymmetric information, and operational constraints such as clearing and inventory costs. Economic theory of market microstructure is discussed in O'Hara (1995) and Hasbrouck (2007) provides a discussion of empirical findings. Although market microstructure effects capture the impact of trading mechanisms on the price formation process, from a statistical perspective microstructure effects act as observation error, which we label "MM noise." Only recently has realized covariance research begun to address the role market microstructure plays in estimation. (See

Bandi, Russell, and Zhu (2008), Bandi and Russell (2006), Zhang, Mykland, and Ait-Sahalia (2005), Voev and Lunde (2007), and Griffin and Oomen (2006).) This research relies on estimating optimal sampling frequency as a function of the volatility-to-noise ratio of a given time series of returns. Our realized covariance estimator is a function of the estimate of MM noise.

The characteristics of high-frequency data continue to evolve. Yan (2007) shows nearly 8-fold increase in the number of trade and quote observations in the TAQ database in the six years following 2000. Hansen and Lunde (2006) observe that in addition to the increased number of quotes, the characteristics of these quotes has also changed dramatically in recent years. A dramatic structural change was completed in February 2001 when the New York Stock Exchange (NYSE) finished the transition from fractional to decimal pricing. (See Goldstein et al. (2008).) This in turn led to a reduction in market makers' rents, changing the nature of price discovery. Traditionally, regional exchanges competed with the NYSE by offering competitive quotes, cheaper executions, and anonymity. Now quotes posted on the NYSE are more often alone at the National Best Bid and Offer (NBBO) (Goldstein et al. 2008). To avoid this shift in market microstructure properties we limit our sample to the period after decimalization, namely post 2001.

In this study we examine the covariance structure between Exxon Mobil Corp. (XOM), Occidental Petroleum Corp. (OXY), and The J. M. Smucker Company (SJM) over the period from of January 2002 to December 2006. Appendix A.1 details the data filtering criteria. Figure 2.1 shows the observed price processes over this time horizon and Table 2.1 presents summary statistics. In Table 2.1, the reported market capitalization is calculated on January 2002, the beginning of our study; the annualized mean μ and volatility σ are for

close-to-close returns, and the noise-to-signal ratio ξ/σ , captures the market microstructure effects. These stocks exhibit different market capitalization, quote frequencies, and noise-to-signal characteristics, and there are different levels of return correlation across the three pairs. As seen in Table 2.1 the energy stocks XOM and OXY are strongly correlated, but SJM (a food products manufacturer) is weakly correlated with the energy stocks.

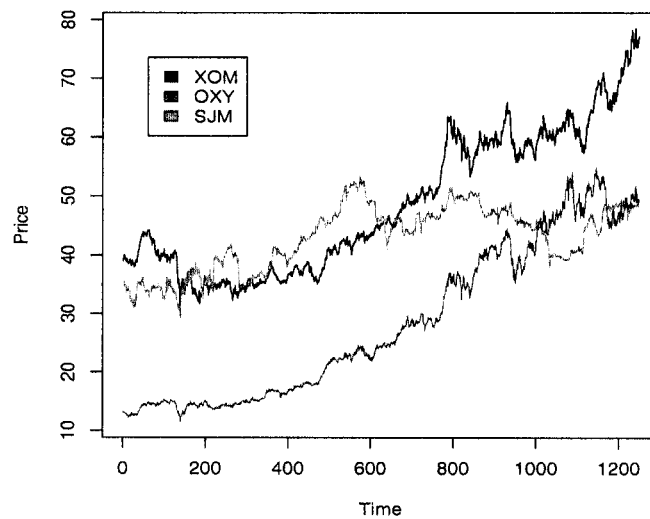


Figure 2.1 Stock prices from January 1, 2002 to December 31, 2006 for Exxon Mobil Corp. (XOM), Occidental Petroleum Corp. (OXY) and The J.M. Smucker Company (SJM). Prices have been adjusted for stock splits.

Table 2.1 Summary statistics of stock returns from January 1, 2002 to December 31, 2006. Table reports market capitalizations of stocks, annualized mean μ and standard deviation σ of returns, noise-to-signal ξ/σ ratios for the returns, and correlation of daily returns.

Ticker	Market Cap	μ	σ	ξ/σ	ρ		
					XOM	OXY	SJM
XOM	267B	.13	.184	.022	1	0.71	0.15
OXY	9.7B	.26	.204	.029		1	0.10
SJM	0.79B	.07	.166	.062			1

Table 2.2 Annual averages of daily admissible quotes. The number of quotes increases over time for all three stocks. SJM is relatively less frequently quoted than XOM or OXY.

Ticker	2002	2003	2004	2005	2006
XOM	1578	2424	2439	4247	5374
OXY	820	1496	2533	4616	6386
SJM	388	770	933	912	1230

Table 2.2 shows that the number of admissible quotes (see Appendix A.1 for definition) has increased substantially over time for all three stocks. XOM, selected because it is one of the largest stocks on the NYSE by market capitalization, is a very actively quoted security. In contrast, SJM is a relatively less actively quoted security with approximately a quarter of the activity observed in XOM. OXY, another energy stock, was selected because it has a strong correlation with XOM. Over the five year sample period, the quote activity for OXY increases from about half the XOM level of activity to surpassing it. Figure 2.2 illustrates the evolution of the noise-to-signal ratios over time. We see that the noise-to-signal ratios over this period decreased for each stock, dramatically so for OXY and SJM. A portion of this decrease can be explained by the increase in quote activity over the period. The different levels of quote activity and noise-to-signal characteristics offer important contrasts for assessing the properties of realized covariance estimates.

2.3 Method

2.3.1 Tick-Time Covariance Estimation

Hayashi and Yoshida (2005) introduce a tick-time *cumulative covariance estimator*:

$$HY := \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} \Delta p_A(I^i) \Delta p_B(J^j) 1_{\{I^i \cdot J^j \neq \emptyset\}}. \quad (2.9)$$

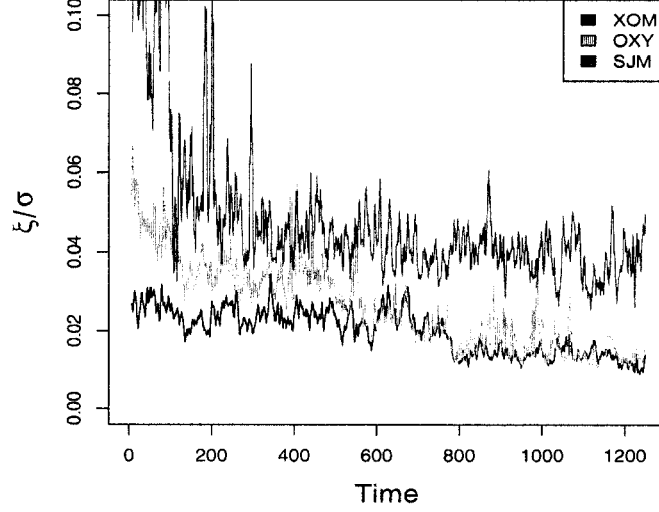


Figure 2.2 Noise-to-Signal Ratio. XOM has the lowest noise-to-signal ratio. All three stocks show a decrease in noise-to-signal ratio corresponding to an increase in quote frequency.

Define T as the terminal time and $\Pi^A = \{t_i\}_{i=0,1,2,\dots,M_A}$ and $\Pi^B = \{t_j\}_{j=0,1,2,\dots,M_B}$ to be the sets of observation times for processes A and B respectively, where $0 \leq t_i \leq T$ and $0 \leq t_j \leq T$ for all i, j . For process A, the interval of observation times I^i is defined as $(t_{i-1}, t_i]$. For process B, the interval of observation times J^j is defined as $(t_{j-1}, t_j]$. The intervals satisfy the following conditions:

Condition 1

1. (I^i) and (J^j) are independent of p_A and p_B .
2. As $n \rightarrow \infty$, $E[\max_i |I^i| \vee \max_j |J^j|] = o(1)$.

The HY estimator is similar to the estimator introduced by de Jong and Nijman (1997) and Corsi (2006). This estimator is the sum of the product of any pair of overlapping intervals.

In the absence of noise, calendar-time estimators are biased and inconsistent, but the HY estimator is not. This improvement can be attributed primarily to the cross-market

tick-matching which corrects for the bias introduced by asynchronicity. Never the less, the introduction of market microstructure noise renders this estimator inconsistent and, dependent on the noise structure, it may also be biased.

An additional issue with this estimator is computational complexity. For the sake of computational efficiency Palandri (2006) and Voev and Lunde (2007) develop aggregation schemes for the HY estimator. These aggregation schemes preserve the informational content of the original HY estimator. In particular, Voev and Lunde propose an aggregated HY estimator where the ticks of the faster process are matched to the ticks of the slower process.

2.3.2 Cross-Market Tick-Matching

Building on this research, we develop an aggregate tick-time estimator by matching the ticks of the faster process (B) to the slower process (A) by: $p_B(t_i^\wedge) = p_B(\min\{t_j \in \Pi^B : t_j \geq \max\{t_i \in I^i\}\})$ and $p_B(t_i^\vee) = p_B(\max\{t_j \in \Pi^B : t_j \leq \min\{t_i \in I^i\}\})$. This Cross-Market Tick-Matching estimator (CMTM) is shown in Figure 2.3. Using the slower process (A) as the base arrival process we sample with respect to M_A the number of intervals between observations for process (A).

This allows us to write:

$$CMTM = \sum_{i=1}^{M_A} \Delta p_A(I^i)(p_B(t_i^\wedge) - p_B(t_i^\vee)) = \sum_{i=1}^{M_A} r_A(I^i)r_B(t_i^\vee, t_i^\wedge]. \quad (2.10)$$

2.3.3 Properties of Proposed Estimator

The expected value of our proposed estimator is

$$E = \rho\sigma_A\sigma_B + 2\pi \sum_{i=1}^{M_A} \xi_{AB}. \quad (2.11)$$

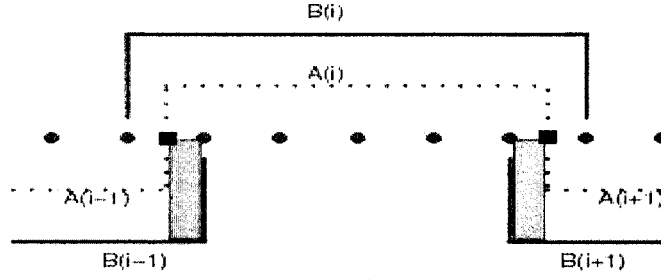


Figure 2.3 Visualization of Cross-Market Tick-Matching. The slower process (A), represented by squares, is used as the base. The observations of the faster process (B), represented by circles, are matched to the corresponding observations of (A). By construction this estimator has overlapping intervals, which are represented by the grey rectangles.

We define π as the probability of synchronous observations. We see that the noise accumulation is influenced by the extent of asynchronicity. (See Chapter 4 for derivation.)

Consistent with the original HY estimator, this estimator is unbiased in either the absence or in the presence of i.i.d. noise. Contemporaneously correlated noise does render the estimator biased. The magnitude of the bias is a function of the number of observations of the slower process multiplied by the probability that the price pairs are synchronously observed.

The variance of our proposed estimator is

$$\begin{aligned}
 V = & \sum_{i=1}^{M_A} (\rho\sigma_A\sigma_B)^2 l(I^i)^2 + \sigma_A^2 l(I^i) \sigma_B^2 l(t_i^\vee, t_i^\wedge]) \\
 & + 2\rho\sigma_A\sigma_B l(I^i \cap (t_{i-1}^\vee, t_{i-1}^\wedge]) \rho\sigma_A\sigma_B l(I^{i-1} \cap (t_i^\vee, t_i^\wedge]) \\
 & + 2\xi_A^2 \sum_{i=1}^{M_A} \sigma_B^2 l(I^i) + 2\xi_B^2 \sum_{i=1}^{M_A} \sigma_A^2 l(I^i) + \sum_{i=1}^{M_A} 2(1+\pi)\xi_A^2 \xi_B^2 + \sum_{i=1}^{M_A-1} 2\pi\xi_A^2 \xi_B^2 \\
 & + 4\pi \sum_{i=1}^{M_A} \xi_{AB}^2 + 2 \sum_{i=1}^{M_A-1} (2\pi - \pi^2) \xi_{AB}^2 + 4\pi \sum_{i=1}^{M_A} \sigma_A\sigma_B\rho\xi_{AB}.
 \end{aligned} \tag{2.12}$$

(See Chapter 4 for derivation.) From Condition 1 it follows that the sum of squared obser-

vation intervals converges to zero:

$$\sum_{i=1}^n |I^i|^2 \rightarrow 0 \text{ in probability as } n \rightarrow \infty. \quad (2.13)$$

This ensures that the variance due to discretization converges to zero as the number of observations goes to infinity. We define l as the length of the interval. Rewriting Equation 2.12 the variance of our proposed estimator reduces to:

$$\begin{aligned} V = & \frac{1}{M_A} \{ (\rho \sigma_A \sigma_B)^2 (1 + 2\alpha_1 \alpha_2) + \sigma_A^2 \sigma_B^2 (1 + \alpha_0) \} \\ & + 2\xi_A^2 \sigma_B^2 + 2\xi_B^2 \sigma_A^2 + \sum_{i=1}^{M_A} 2(1 + \pi) \xi_A^2 \xi_B^2 + \sum_{i=1}^{M_A-1} 2\pi \xi_A^2 \xi_B^2 \\ & + 4\pi \sum_{i=1}^{M_A} \xi_{AB}^2 + 2 \sum_{i=1}^{M_A-1} (2\pi - \pi^2) \xi_{AB}^2 + 4\pi \sum_{i=1}^{M_A} \sigma_A \sigma_B \rho \xi_{AB}. \end{aligned} \quad (2.14)$$

The first line in Equation 2.14 represents the error due to discretization. Namely, as M_A goes to infinity, the first line goes to 0. The remaining lines show the error contributions due to market microstructure effects. The terms in the final three lines increase with M_A rendering the estimator inconsistent. The i.i.d. noise setting is obtained by letting $\xi_{AB} = 0$.

In this context, α_i are a measure of the overlap in the CMTM estimator. In the second element, $1 + \alpha_0$ represents $l(t_i^\vee, t_i^\wedge)/l(I^i)$, where l is the length of the intervals. Thus, α_0 represents the estimated overlap of contemporaneous intervals. In similar fashion, α_1 is $l(t_{i-1}^\vee, t_i^\wedge)/l(I^i)$ and α_2 is $l(t_i^\vee, t_i^\wedge)/l(I^{i-1})$, each capturing the estimated overlap with the adjacent intervals. Consistent with Voev and Lunde (2007), this allows us to express the bounds of asynchronicity in terms of α_i .

2.3.4 Optimal Sampling

The MSE is stated as a function of sampling frequency k by replacing M_A with M_A/k and $l(I^i)$ with $k \times l(I^i)$:

$$\begin{aligned} MSE(k) = & \frac{k}{M_A} \{ (\rho\sigma_A\sigma_B)^2 (1 + 2\frac{\alpha_1\alpha_2}{k^2}) + \sigma_A^2\sigma_B^2 (1 + \frac{\alpha_0}{k}) \} \\ & + 2\xi_A^2\sigma_B^2 + 2\xi_B^2\sigma_A^2 + \frac{M_A}{k} 2(1 + \pi)\xi_A^2\xi_B^2 + \frac{M_A - 1}{k} 2\pi\xi_A^2\xi_B^2 + 4\pi\frac{M_A}{k}\xi_{AB}^2 \\ & + 2\frac{M_A - 1}{k} (2\pi - \pi^2)\xi_{AB}^2 + 4\pi\frac{M_A}{k}\sigma_A\sigma_B\rho\xi_{AB} + 4\pi^2\frac{M_A^2}{k^2}\xi_{AB}^2. \end{aligned} \quad (2.15)$$

Taking the first derivative of the MSE with respect to the sampling frequency k and setting equal to zero yields:

$$g_0 + \frac{g_1}{k} + \frac{g_2}{k^2} + \frac{g_3}{k^3} = 0, \quad (2.16)$$

where $g_0 = \frac{1}{M_A} \{ (\rho\sigma_A\sigma_B)^2 + \sigma_A^2\sigma_B^2 \}$, $g_3 = -8\pi^2 M_A^2 \xi_{AB}^2$, $g_1 = 0$, and $g_2 =$ remaining terms.

The optimal solution is approximated by selecting:

$$k \approx \left(\frac{-g_3}{g_0} \right)^{1/3} \approx \left(\frac{8\pi^2 M_A^3 \xi_{AB}^2}{\rho^2 \sigma_A^2 \sigma_B^2 + \sigma_A^2 \sigma_B^2} \right)^{1/3}. \quad (2.17)$$

The second derivative is positive at this value, verifying that the optimization yields a minimum for the MSE . This approximation is very precise for low noise-to-signal levels and low frequencies. At higher frequencies the approximation tends to be slightly below the exact k , but in the high frequency setting individual ticks are not as informative as in the low frequency setting. The corresponding sampling frequency for the tick-time estimated variance is found by simple substitution as:

$$k \approx \left(\frac{8M_A^3 (\xi_A^2)^2}{2\sigma_A^4} \right)^{1/3}. \quad (2.18)$$

2.3.5 Parameter Estimation

We estimate the underlying parameters as follows:

- M_A is the number of intervals between observations for the slow process (A).
- λ_A is the arrival intensity of process A.
- ξ_A is estimated as $1/M_A \sum_i^{M_A} r_A^2(I^i)$, using all the observations.
- ξ_{AB} is estimated as $1/n^* \sum_j^{n^*} r_{A,j} r_{B,j}$, where $n^* = \frac{T}{\delta}$ such that $\delta = 2(\max\{\lambda_A, \lambda_B\})$.
- The overlapping intervals are estimated as:

$$\begin{aligned}\alpha_0 &\approx \lambda_B/\lambda_A \\ \alpha_1 &\approx \frac{1}{2}\lambda_B/\lambda_A \\ \alpha_2 &\approx \frac{1}{2}\lambda_B/\lambda_A.\end{aligned}$$

- The Barndorff-Nielsen and Shephard (2004a) quarticity estimates, with a sampling frequency of 15 minutes, are used to obtain more stable estimates for the denominators in Equations 2.17 and 2.18.

We find that it is necessary to smooth the ξ_i estimates to derive reasonable sampling frequencies. We employ the rolling smoothing technique advocated in Andreou and Ghysels (2002) with the following recursion: $\tilde{\xi}_{A,t}^2 = 0.9\xi_{A,t}^2 + 0.1\tilde{\xi}_{A,t-1}^2$.

2.4 Simulation

We present Monte Carlo simulation results comparing the efficiency of our proposed sub-sampled CMTM estimator. Each simulation consists of 5000 iterations. We evaluate

the relative performance of the covariance estimators under contemporaneously correlated noise structures. Three different arrival regimes, in terms of expected duration between arrivals, are examined: 1.) low-low (30:30), 2.) low-high (30:10), and 3.) high-high (10:10). These arrival regimes are representative of the subset of the TAQ quote data that we are studying. In order to consider the marginal contribution of tick matching, optimal sampling, and sub-sampling, we consider the five estimators presented in Table 2.3.

Table 2.3 Realized covariance estimators considered in simulation. Calendar time is abbreviated as (CT) and tick time is abbreviated as (TT).

Symbol	Estimator	Abbreviation
○	5 minute calendar time	(CT5)
◇	15 minute calendar time	(CT15)
△	calendar time optimally sampled	(CTO)
×	5th tick estimator	(TT5)
+	optimally sampled CMTM estimator	(TTO)

2.4.1 Simulation Design

The latent process is a bivariate Brownian motion. This process is observed at non-synchronous Poisson times. The generator is a two step process. First the latent process is generated using the Euler scheme suggested by Higham (2001):

$$\begin{bmatrix} \Delta x_A \\ \Delta x_B \end{bmatrix} = \sqrt{\Delta t} \begin{bmatrix} \Theta^T \end{bmatrix} \begin{bmatrix} z_A \\ z_B \end{bmatrix} \quad (2.19)$$

where Θ^T is the Cholesky factorization of the covariance matrix such that $\Theta\Theta^T = \Sigma$. Σ is a 2×2 matrix. Z represents the Brownian motion and $z_i \sim N(0, 1)$. The inter-arrival times are exponential so arrival times are determined by the cumulative sum of the generated vector of exponential random variables and rounded to the nearest integer

value. The resulting values determine which elements of the latent process will be observed. Asynchronicity is achieved by simulating two different arrival time vectors. Finally, market microstructure effects are added to the selected latent values to create observed values ($p_A(t_i) = x_A(t_i) + u_A(t_i)$).

Specifically, we consider $\sigma_{1,1} = \sigma_{2,2} = 1$ and $\sigma_{1,2} = 0.9$. We generate our realizations of the processes using $\Delta t = 1$ second. We assume that $\sigma_{1,1}$ and $\sigma_{2,2}$ are known and focus exclusively on estimating the covariance.

2.4.2 Simulation Results

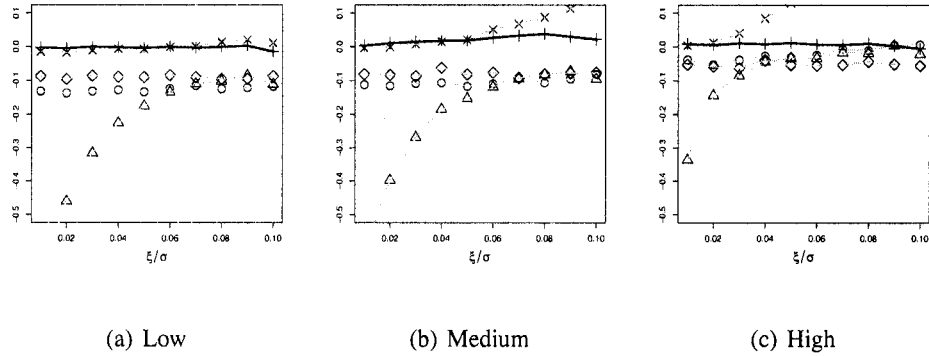


Figure 2.4 Bias of realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. Tick-time estimators are unbiased for low noise-to-signal settings. For low quote frequency settings, calendar-time estimators display a downward bias. As the quote frequency increases, the ad-hoc tick time estimator displays a positive bias.

Figure 2.4 presents the bias of the covariance estimators under the three different arrival regimes. Figures 2.5 and 2.6 show the associated mean squared error (MSE). We see that under the low arrival regime, the tick-matching estimators are the least biased. In contrast, for very low noise-to-signal ratios the calendar-time optimal sampling methodology performs very poorly. This is due to the Epps effect induced by the large degree of asynchronicity. For the ad-hoc estimators (5 minute and 15 minute), the 15 minute estima-

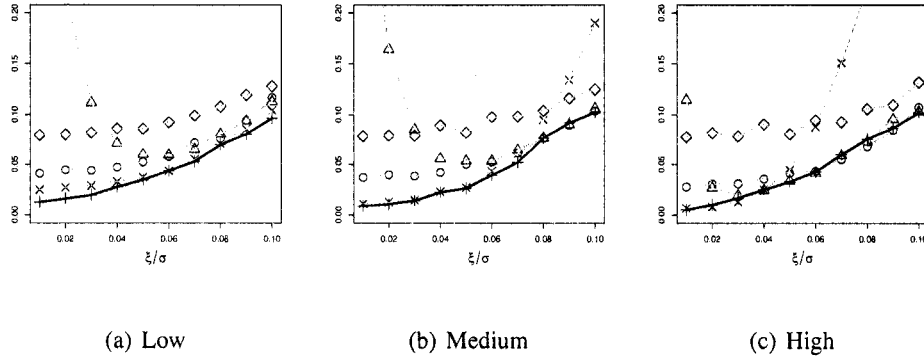


Figure 2.5 Mean Squared Error of realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. Tick-time estimators have smallest MSE, but the difference relative to calendar-time estimates is less pronounced as the noise-to-signal ratio increases.

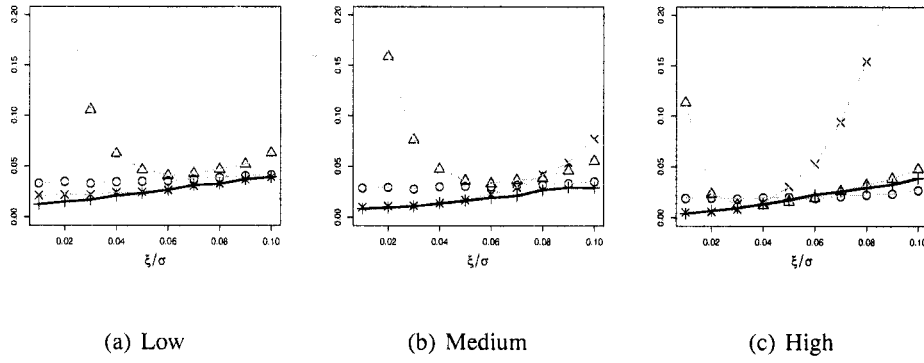


Figure 2.6 Mean Squared Error of sub-sampled realized covariance estimators, where $\rho = 0.9$ and symbols are defined in Table 2.3. The sub-sampled estimators have much smaller MSE than the MSE presented in Figure 2.5.

tor is less biased, because it mitigates the effect of asynchronicity. The medium and high arrival regimes show bias reductions for all the calendar methods, with the CTO improving the most dramatically. This reflects the reduction in asynchronicity with more frequent observations.

We also see the benefit of sampling in tick time. In contrast to calendar time, tick-time estimators demonstrate a positive bias that increases with the noise-to-signal ratio and with arrival frequency. This is consistent with additive noise and with decreased asynchronicity

as frequency increases. The contrast in the bias of the TT5 and TTO estimators indicates the benefits of optimal sampling in the tick-time setting. Note that for this range of noise-to-signal values, the optimally sampled tick-time estimator is always less biased than the calendar-time alternatives.

In Figures 2.5 and 2.6 we see that for low noise-to-signal ratios the tick-time estimators have the smallest MSE, and indeed the sub-sampled TTO estimator has the smallest MSE overall. As the noise level increases, the tick-time methods and the CT5 and CTO methods provide similar results. The optimally sampled calendar-time estimator is non-competitive for low noise, and low frequency due to the large bias. As the frequency and noise increase, this estimator begins to outperform the ad-hoc calendar-time estimators. Figure 2.6 shows that sub-sampling reduces the MSE with greatest benefits in high noise settings.

2.5 Application

Andersen, Bollerslev, Diebold, and Labys (2001) find that realized covariance displays strong persistence and is characterized by a slow decay of autocorrelation which is preserved even under temporal aggregation. These findings motivate the use of volatility-timing strategies for assessing the performance of different covariance estimators.

2.5.1 Data Analysis

Figure 2.7 displays the sampling frequency for the tick-time variance estimator and Figure 2.8 shows the sampling frequency for our proposed cross-market tick-matching covariance estimator. The variance sampling frequency follows the day-to-day quote activity and for OXY and SJM shows an increase that corresponds with the increased quote activity in these two stocks. The covariance sampling frequency is largely determined by the quote

activity of the less actively quoted asset. This can be seen easily in Figure 2.8 as panels (b) and (c) both have SJM as the less actively quoted asset and have similar sampling frequencies, but the frequencies are much lower than for the actively traded pair (panel (a)).

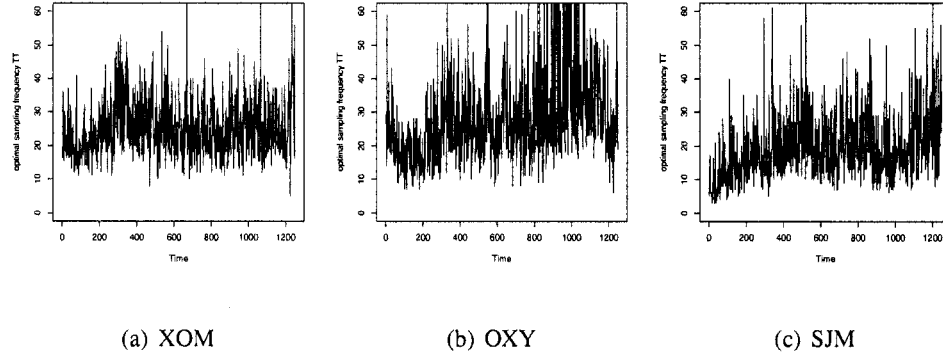


Figure 2.7 Tick-time Variance Sampling Frequency

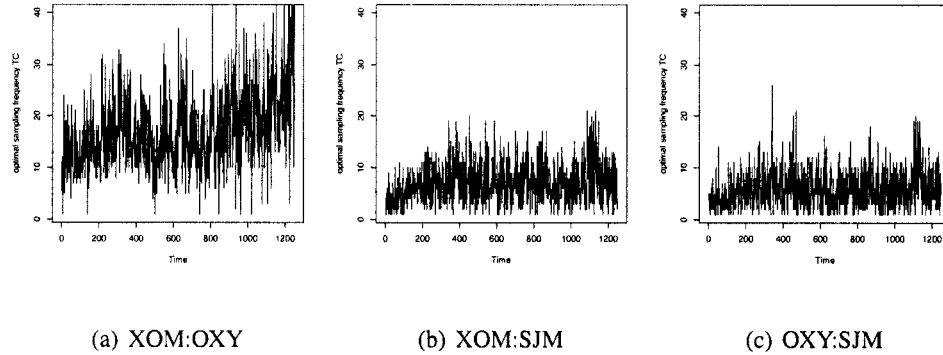


Figure 2.8 Tick-time Covariance Sampling Frequency

Figure 2.9 displays the volatility estimates of XOM using CT5, CT15, and TTO-Sub, and Figure 2.10 shows the correlation estimates between XOM and SJM using these three techniques. The ad-hoc calendar-time methods display larger variations in day-to-day volatility estimates. In contrast, the tick tick optimal sampling technique appears much more stable.

Table 2.4 presents the returns, volatilities, and estimated Sharpe ratios (μ/σ), a mea-

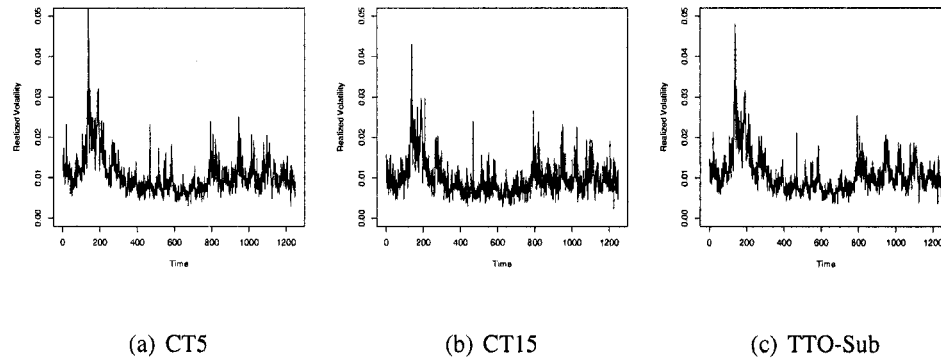


Figure 2.9 Estimated σ

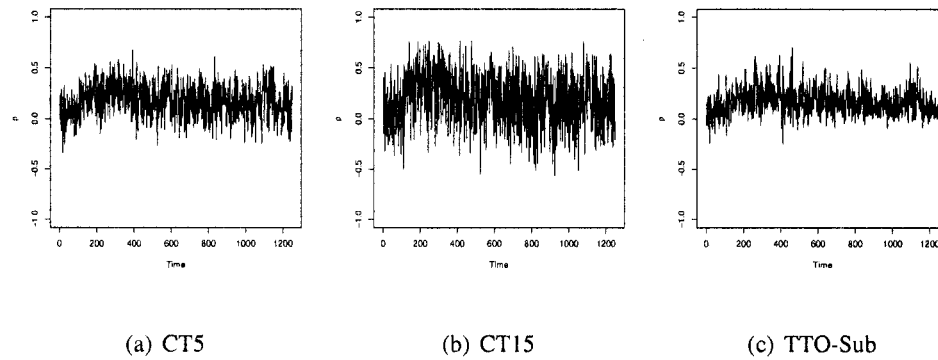


Figure 2.10 Estimated ρ

sure of reward to risk, generated by following a strategy of taking positions at the market open and exiting positions at the market close. The final row presents the results of an equally weighted portfolio. The open-to-close returns of XOM and OXY are close to zero and even negative, which is consistent with the results presented in Cliff, Cooper, and Gulen (2007). Table 2.4 also shows the correlations of open-to-close returns, and we see higher correlations between SJM and the energy stocks (XOM and OXY) than in the close-to-close correlations presented in Table 2.1.

Table 2.4 Summary of Open-to-Close Returns. Presents annualized mean and standard deviations of open-to-close returns, estimated Sharpe ratio, μ/σ , and correlation matrix. The last line presents an equally weighted portfolio.

	μ	σ	μ/σ	ρ		
				XOM	OXY	SJM
XOM	-2.83%	17.80%	-0.14	1	0.71	0.28
OXY	3.10%	20.20%	0.16		1	0.22
SJM	21.50%	16.85%	1.29			1
Equal	7.26%	14.29%	0.52			

2.5.2 Volatility Timing

Fleming, Kirby, and Ostdiek (2001, 2003) popularized volatility-timing as an objective methodology for assessing the economic “value-added” of alternative estimators. The value-added is measured in a portfolio framework as the fee (as a percent of assets) a risk-averse investor would be willing to pay to capture the gains in portfolio performance made possible by using a given covariance estimator. The assessment is made based on the investor’s utility improvement which is converted into a management fee and expressed in basis points (i.e. 0.01%). Specifically, investors follow a volatility-timing strategy where the portfolio weights vary only with changes in estimates of the conditional covariance matrix of daily returns. Define F_{t-1} to be the day $t - 1$ information set. Conditional volatility is minimized using risky asset weights:

$$w_t = \frac{\mu_p \Sigma_t^{-1} \mu_t}{\mu_t' \Sigma_t^{-1} \mu_t} \quad (2.20)$$

where μ_p is the target expected return on the portfolio, $\mu_t \equiv E[R_t | F_{t-1}]$ is the vector of conditional means, and $\Sigma_t \equiv E[(R_t - \mu_t)(R_t - \mu_t)' | F_{t-1}]$. These weights are the result of standard mean-variance portfolio optimization.

The realized daily utility generated by the returns of this portfolio (R_{pt}) for an investor with γ relative risk aversion is

$$U(R_{pt}) = W_0 \left((1 + R_f + R_{pt}) - \frac{\gamma}{2(1 + \gamma)} (1 + R_f + R_{pt})^2 \right). \quad (2.21)$$

R_f is the risk-free rate and W_0 is the amount of wealth that is invested. The incremental value of using the second estimator instead of the first is calculated by finding the constant Δ , such that $\sum_{t=1}^T U(R_{p1t}) = \sum_{t=1}^T U(R_{p2t} - \Delta)$. This can be thought of as the basis point fee an investor would be willing to pay to access the second estimator instead of the first.

2.5.3 Isolating A Single Hypothesis

Recall that in Equation 2.20, the optimal allocation is a function of the targeted portfolio expected return, the conditional means, and the conditional covariance. In order to test only the performance of the covariance estimators, we must control for misidentification of the expected returns. Engle and Colacito (2006) suggest that in the multivariate setting, the relationship between returns (i.e. angle in a bivariate setting) is more critical than the relative magnitudes. Chopra and Ziemba (1993) consider the impact of specifying the functional form of the utility function and parameter estimation of the means and covariance matrix. They find that portfolio optimization is invariant to the utility function, and correct specification of the means is ten times more important than specification of the covariance matrix. Within the covariance matrix, they find that specification of the variance elements is twice as important covariance elements.

We isolate the impact of our proposed covariance estimator by controlling the mean estimates through a bootstrap procedure. We sample with replacement blocks of 50 returns to bootstrap 5000 different realizations of the return process. (We use block resampling to

preserves the dependence structure in the return series. See Politis (2003).) We then take the mean return of these realizations and in this fashion obtain a distribution of possible expected returns. We consider the performance of the volatility-timing strategies using this set of 5000 bootstrapped expected returns.

We find that overnight returns are very noisy and bring additional estimation challenges. For this analysis, therefore, we constrain our investor to taking positions at the market open and liquidating at market close, thereby eliminating overnight returns from our analysis. We refer readers to Hansen and Lunde (2005) for a detailed discussion of the estimation of an entire day's realized variance by including the squared overnight returns.

2.5.4 Economic Value Results

Table 2.5 compares the mean, standard deviation, and Sharpe ratios of the volatility-timing portfolios using the 5 and 15 minute calendar-time (CT5) and (CT15), optimally sampled calendar-time (CTO), CMTM sampled every 5 ticks (TT5), and optimally sampled CMTM (TTO) covariance estimators. We also consider sub-sampled versions of the optimally sampled calendar-time (CTO-S), CMTM sampled every 5 ticks (TT5-S), and optimally sampled CMTM (TTO-S) covariance estimators. We set the target annual return at $\mu = 7.5\%$, slightly more than the return on an equally weighted portfolio as presented in Table 2.4. We set the risk free rate $R_f = 0$. From Table 2.5 we see that the calendar-time methods have smaller returns than the tick-time methods. The TTO-S estimator has the greatest return with one of the smallest standard deviations. The optimal sampling estimators result in larger standard deviations for both calendar-time and tick-time estimators. Sub-sampling over the optimal sampling frequency, however, results in a portfolio stan-

dard deviation for the tick-time estimator that is less than that for the sub-sampled ad-hoc estimators.

Table 2.5 Summary results for portfolio containing XOM, OXY, and SJM. Target return is $\mu_P = 7.5\%$. Presents annualized means, standard deviations, and Sharpe ratios for portfolio returns using 8 different covariance estimators.

	CT5	CT15	CTO	CTO-S	TT5	TT5-S	TTO	TTO-S
μ_R	6.20%	5.84%	6.19%	6.16%	6.53%	6.62%	6.56%	6.90%
σ_R	5.61%	5.75%	5.85%	5.84%	5.71%	5.71%	5.82%	5.64%
Sharpe Ratio	1.11	1.02	1.06	1.06	1.15	1.16	1.13	1.22

Table 2.6 shows the annualized basis point fees that a risk averse investor is willing to pay to switch to the TTO-S method from traditional methods. Outliers, more than 500 basis points away from the median, were removed and the filtered means are presented. We consider the performance of our proposed estimator relative to calendar-time estimates with ad-hoc and optimal sampling techniques and relative to ad-hoc tick-time sampling. This allows us to investigate the benefits of using tick time rather than calendar time, the benefits of optimal sampling, and the contribution due to sub-sampling.

Table 2.6 Annualized Basis Points for portfolio containing XOM, OXY, and SJM. Target return is $\mu_P = 7.5\%$. Investor risk aversion increases with γ .

Method	γ_1	γ_5	γ_{10}
CT5:TTO-S	70.20	69.52	68.68
CT15:TTO-S	107.54	110.13	113.36
CTO:CTO-S	-2.81	-2.55	-2.24
CTO:TTO	37.18	38.02	39.07
CTO-S:TTO-S	75.49	80.13	85.94
TT5:TTO	2.41	-0.16	-3.39
TT5-S:TTO-S	27.71	30.19	32.04
TTO:TTO-S	35.50	39.55	44.63

We see that the TTO-S estimator's greater returns with lower volatility translate into greater utility over the ad-hoc sampling techniques. The TTO-S estimator generates nearly 70 additional basis points against the 5 minute calendar-time estimator, and around 110 against the 15 minute estimator, regardless of the assumed risk-aversion parameter. The optimally sampled tick-time estimator generates an additional 40 basis points over its calendar-time counterpart. Moreover, the tick-time estimator benefits more from sub-sampling than its calendar-time counterpart. The sub-sampled optimally sampled tick-time estimator (TTO-S) provides 80 basis points gain over CTO-S and 30 basis points gain over TT5-S. This demonstrates the benefits of tick time over calendar time, optimal sampling over ad-hoc sampling, and sub-sampling over sampling.

Figure 2.11 shows the weights of the positions taken using this volatility timing strategy where the expected returns are set by the means reported in Table 2.4. To control for outliers in the portfolio weights, the trading strategies are implemented with the constraint that $|w_i| \leq 300\%$. (Empirical investigation indicates that the qualitative assessment is not sensitive to weight constraints.) As expected, the sign and magnitude of each of the weights depends on the expected return. The day-to-day changes in the weights are driven by the covariance estimates.

To better identify why the TTO-S estimator provides utility gains we further disaggregate the results by considering the three pairs of XOM:OXY, XOM:SJM, and OXY:SJM. Table 2.7 shows that the tick-time estimators consistently provide superior reward for risk in any setting. For portfolios with the less actively quoted stock (SJM), the tick-time estimators generally result in the largest portfolio returns. For XOM:SJM, TTO-S provides the best Sharpe ratio, and for OXY:SJM, TT5-S and TTO-S provide superior Sharpe ratios.

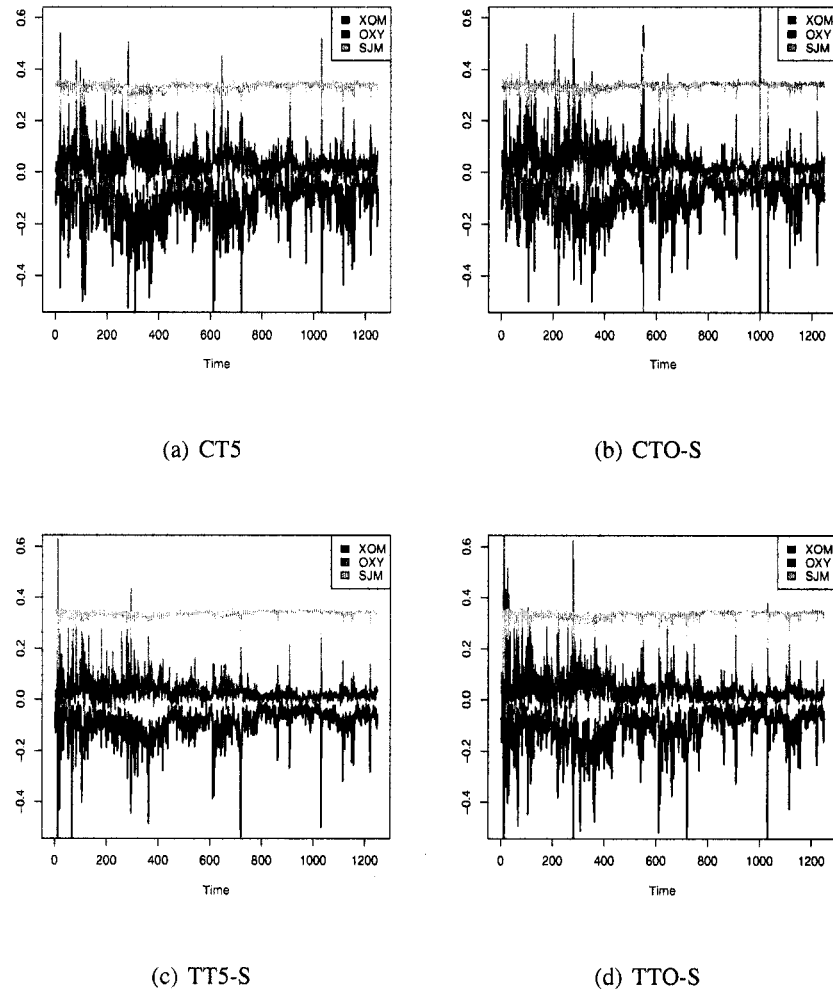


Figure 2.11 A realization of portfolio weights using a volatility-timing strategy. Presents results for CT5, CTO-S, TT5-S, and TTO-S.

Table 2.8 decomposes the impact of calendar-time vs. tick-time sampling and subsampling across the security pairs. For the XOM:OXY pair the TTO-S estimator slightly underperforms the static methods. This may indicate that actively quoted stocks need a more generalized noise structure, such as the autocorrelated noise structure discussed in Hansen and Lunde (2006). These results are consistent with the expectations from the simulation experiment. TTO provides 6 to 46 basis point gain over CTO, and the substantial

Table 2.7 Summary results for disaggregated portfolio pairs. Target return is $\mu_P = 7.5\%$. Presents annualized means, standard deviations, and Sharpe ratios for portfolio returns using 8 different covariance estimators.

	CT5	CT15	CTO	CTO-S	TT5	TT5-S	TTO	TTO-S
XOM:OXY								
μ_R	2.35%	2.15%	1.94%	1.96%	1.81%	1.89%	1.96%	2.20%
σ_R	20.42%	20.28%	21.25%	21.23%	21.62%	21.67%	20.80%	20.73%
μ_R/σ_R	0.115	0.106	0.0911	0.0925	0.0838	0.0871	0.0942	0.106
XOM:SJM								
μ_R	7.07%	6.88%	7.34%	7.34%	7.32%	7.35%	7.37%	7.56%
σ_R	6.16%	6.23%	6.24%	6.24%	6.18%	6.17%	6.22%	6.18%
μ_R/σ_R	1.15	1.10	1.18	1.18	1.18	1.19	1.18	1.22
OXY:SJM								
μ_R	6.99%	6.62%	6.74%	6.75%	7.22%	7.28%	7.08%	7.18%
σ_R	7.01%	7.15%	7.06%	7.06%	7.06%	7.06%	7.08%	7.03%
μ_R/σ_R	0.997	0.926	0.955	0.956	1.02	1.03	1.00	1.02

difference between the three risk aversion levels indicates that CTO is a more volatile estimator than TTO. We also show that optimizing in tick time provides 12 to 48 basis point gains over every 5th tick sampling. Finally, sub-sampling the optimized tick-time estimator provides an additional 30 to 40 basis points gain. The TTO-S estimator provides 30-90 basis point gains over CTO-S and 40-100 basis point gains over TT5-S. This suggests that tick-time sampling alone does not provide the utility gains, but optimizing the sampling frequency plays a non trivial role. For XOM:SJM we see approximate gains of 20 basis points against the CTO and the TT5. TTO-S provides 50 and 70 basis point gains against the CT5 and CT15 estimators respectively. In the OXY:SJM pair, we see gains of 20, 60, and 33 basis points when using TTO in place of CT5, CT15, and CTO respectively. We see a loss of 15 basis points when comparing TT5 against TTO. This loss is between 10

Table 2.8 Annualized Basis Points for disaggregated portfolio pairs. Target return is $\mu_P = 7.5\%$. Investor risk aversion increases with γ .

Method		γ_1	γ_5	γ_{10}
XOM:OXY				
	CT5:TTO-S	-19.75	-28.29	-35.97
	CT15:TTO-S	2.89	2.84	1.37
	CTO:CTO-S	2.62	4.53	6.80
	CTO:TTO	6.33	26.28	46.78
	CTO-S:TTO-S	32.10	59.94	87.15
	TT5:TTO	11.53	29.32	48.64
	TT5-S:TTO-S	40.28	70.97	102.85
	TTO:TTO-S	28.99	34.62	39.67
XOM:SJM				
	CT5:TTO-S	49.25	48.76	48.16
	CT15:TTO-S	68.92	70.11	71.61
	CTO:CTO-S	-0.24	-0.25	-0.28
	CTO:TTO	3.13	3.69	4.40
	CTO-S:TTO-S	22.50	24.08	26.05
	TT5:TTO	4.99	3.94	2.62
	TT5-S:TTO-S	21.50	21.28	21.01
	TTO:TTO-S	19.13	20.12	21.37
OXY:SJM				
	CT5:TTO-S	19.14	18.51	17.72
	CT15:TTO-S	57.29	60.52	64.58
	CTO:CTO-S	0.98	0.99	1.00
	CTO:TTO	33.61	33.04	32.33
	CTO-S:TTO-S	42.70	43.52	44.54
	TT5:TTO	-14.71	-15.41	-16.29
	TT5-S:TTO-S	-10.04	-9.31	-8.39
	TTO:TTO-S	10.08	11.47	13.21

to 8 basis points when considering sub-sampled estimators. The loss against the TT5 can be partially explained by the fact that the mean value of the estimated optimal sampling frequency for the tick-time estimator is approximately five. This indicates that smoothing the optimal sampling frequency may be beneficial.

Results suggest that calendar-time sub-sampling is not effective when less actively quoted stocks introduce asynchronicity. This is in sharp contrast to the large gains realized by the sub-sampled tick-time estimators in this case. These results suggest that tick-time is the more natural setting for covariance sub-sampling.

2.6 Conclusion

We contribute to the discussion of realized covariance by deriving an optimal sampling frequency for a cross-market tick-matching estimator. Furthermore, we demonstrated the potential benefits of this new estimator via a simulation study that shows tick matching and sub-sampling providing substantial MSE reduction and we estimate the economic value-added of this estimator.

Unlike Griffin and Oomen (2006) we first match the ticks and then sample with respect to the slow process. Unlike Voev and Lunde (2007), who use tick matching but have an ad-hoc sampling scheme, we calculate an optimal sampling frequency with respect to MSE. Finally, we provide the first economic value added assessment of cross-market tick-matching estimators in a volatility-timing framework. We find that our optimally sub-sampled cross-market tick-matching estimator provides 70 basis point gains over the 5 minute calendar time strategy, approximately 110 basis point gains over the 15 minute calendar time strategy, and 80 basis point gain over a corresponding optimally sub-sampled

calendar time strategy. The proposed estimator provides the greatest contribution in the presence of less actively quoted securities.

As further research, we are currently assessing the performance for a higher dimensional portfolio and examining the ability of the estimator to identify extreme events. Finally, we recognize that Hansen and Lunde (2006) have shown that noise is correlated with the price process and that the noise process may be time dependent. These noise generalizations are beyond the scope of this study but they are clear avenues of future research.

Chapter 3

Parsimonious Realized Portfolio Selection using High-Frequency Data

3.1 Introduction

Markowitz mean-variance (MV) optimization is the standard theoretical framework for optimal portfolio construction (See Chan, Karceski, and Lakonishok (1999), Jagannathan and Ma (2003), and references therein). MV optimization requires covariance matrices to be not only invertible, but also well-conditioned. Michaud (1989) points out that this procedure maximizes the effects of errors in the input assumptions and as a result practical implementation is problematic. Britten-Jones (1999) examined the sampling error of the weights of mean-variance efficient portfolios and found them to be very large.

Realized covariance estimation has emerged as a viable candidate for covariance estimation. This class of estimators employs high-frequency data and provides more precise estimates. In a low dimensional setting, Fleming, Kirby, and Ostdiek (2003) have shown that realized covariance estimates provide utility gains over implied covariance estimates for risk averse investor following a MV optimization strategy. Realized covariance literature has focused on improving these estimators by using techniques such as cross-market tick-matching (See Kyj, Ensor, and Ostdiek (2008), Corsi (2006), Hayashi and Yoshida (2005), Voev and Lunde (2007), Griffin and Oomen (2006)), optimal sampling (See Bandi and Russell (2006), Bandi, Russell, and Zhu (2008), Oomen (2006), de Pooter, Martens, and van Dijk (2006), and sub-sampling (See Zhang, Mykland, and Ait-Sahalia (2005), Voev and Lunde (2007), Kyj, Ensor, and Ostdiek (2008)).

Estimation of high dimensional covariance matrices is computationally expensive. There

are $p \times (p + 1)/2$ operations, where p is the number of assets considered. Realized covariance estimates also become numerically ill-conditioned due to sampling error. As a result the inversion of the matrix, a necessary step in mean-variance optimization, becomes problematic. Ledoit and Wolf (2003) suggest that the number of observations, n , needs to be at least ten times the number of dimensions, p . In the case of five minute calendar time sampled realized covariance estimation, we have an effective sample size of $n = 78$ and this rule of thumb is exceeded when considering more than 7 dimensions. This is a paradox of high frequency data, at first glance it appears as there is "too much data", but once asynchronicity and market microstructure effects are acknowledged, then once again we are confronted with errors due to small sample size.

Previous literature has addressed imprecise covariance matrix estimates by imposing more structure on covariance matrix. Variants of shrinkage are employed to mitigate ill-conditioned matrices. Fleming, Kirby, and Ostdiek (2003) and de Pooter, Martens, and van Dijk (2006) use "rolling" estimators, Bandi, Russell, and Zhu (2008) use ARFIMA forecasting, Jagannathan and Ma (2003) use non-negative constraints, and Ledoit and Wolf (2003) use shrinkage toward market estimate. Bauer and Vorkink (2007) employ the matrix logarithm function suggested by Kawakatsu (2006) to ensure positive definiteness. This amounts to an exponential transformation of the eigenvalues.

Factor models have the advantage of providing a parsimonious representation of the information and have been shown to offer utility gains over strategies employing full sample covariance matrices. Chan, Karceski, and Lakonishok (1999) consider both the Sharpe ratio and tracking error as portfolio performance diagnostics and find that a three factor model is adequate for selecting the minimum-variance portfolio, but argue that more factors are

necessary for minimizing tracking error volatility. Jagannathan and Ma (2003) compare the performance of a non-negative sample covariance using daily level data against factor models and shrinkage estimators. Indeed, the non-negativity constraint is a special form of the shrinkage. They show that the single factor model performs very well when the number of observation is not much greater than the number of dimensions. Han (2006) discusses the importance of factor models in high dimensional settings. Bollerslev and Zhang (2003) find that within the context of high-frequency data, factor models systematically outperform monthly rolling regression-based estimates. Fan, Fan, and Lv (2007) provides a theoretical understanding of the factor modeling of high dimensional covariance matrices. They find that the factor model is a more consistent estimate for the inverse of the covariance matrix.

This chapter sets out to compare a number of competing realized covariance techniques and their utility gains. We argue that in the presence of linear versus quadratic growth in computational complexity and ill-conditioned matrices, factor models are a more natural setting for comparing covariance estimators. Indeed, we find that that a single-index model can provide similar levels of utility as a fully estimated realized covariance matrix that has been smoothed via exponential weights.

3.2 Methods

3.2.1 Review of Realized Estimators

The discretely observed price process $p(t_i)$ is a function of both the latent price process $x(t_i)$ and the market microstructure effects $u(t_i)$, which are treated as “observation error”

such that the price of asset A is observed as:

$$p_A(t_i) = x_A(t_i) + u_A(t_i), \quad i = 1, 2, \dots, n. \quad (3.1)$$

Hence, returns are written as

$$r_A(t_i) = \Delta p_A(t_i) = \Delta x_A(t_i) + \Delta u_A(t_i). \quad (3.2)$$

We define the noise terms to be contemporaneously correlated such that:

$$\begin{bmatrix} u_A \\ u_B \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \xi_A^2 & \xi_{A,B} \\ \xi_{A,B} & \xi_B^2 \end{bmatrix} \right) \quad (3.3)$$

and assume that noise is uncorrelated with x_t and uncorrelated with non-contemporaneous own and cross-market noise terms.

Andersen, Bollerslev, Diebold, and Labys (2001) first proposed realized variance estimation using ad-hoc calendar-time sampling. Calendar-time sampling requires synchronous observations across markets and this is achieved by interpolating prices onto an ad-hoc common sampling grid i.e, every 5 minutes. We construct m equally spaced intraday observations, and these calendar time realized estimators can be written as:

$$\widehat{V}_A(m) = \sum_i^m r_{A,(m)}^2(i/m) \quad (3.4)$$

and

$$\widehat{C}_{AB}(m) = \sum_i^m r_{A,(m)}(i/m) \times r_{B,(m)}(i/m). \quad (3.5)$$

Eliminating dependency on ad-hoc calendar-time grid intervals motivates determining the optimal sampling frequency by minimizing the mean squared error (MSE) of the realized covariance estimator. The optimal sampling frequency is a function of the return

series's signal-to-noise ratio. Zhang, Mykland, and Ait-Sahalia (2005) and Bandi and Russell (2005) develop optimal sampling schemes in the calendar time setting. The optimal sampling frequencies for the variance and covariance are given in Equations 3.6 and 3.7 respectively.

$$m^* = \left(\frac{2\sigma_A^4}{2(\xi_A^2)^2} \right)^{1/3} \quad (3.6)$$

$$m^* = \left(\frac{\rho^2 \sigma_A^2 \sigma_B^2 + \sigma_A^2 \sigma_B^2}{(\xi_{AB})^2} \right)^{1/3} \quad (3.7)$$

For variance reduction, Zhang, Mykland, and Ait-Sahalia (2005) advocate sub-sampling and averaging as a technique for exploiting the richness of high frequency data. They divide the time domain grid into K non-overlapping subgrids and average the estimates over the K different subgrids to calculate the final estimate.

When sampling high-frequency data, choosing tick time or calendar time is another important issue. The interpolation schemes in calendar time estimation causes realized covariance estimates to be vulnerable to the so called “Epps effect” (Epps 1979), where the covariation estimate converges to zero as the sampling grid gets finer. Operating in tick time samples the price process according to changes in the level of market activity. As a result, tick-time sampling offers superior location of the sampling points that generate the information set. Working in transaction time rather than calendar time, Kyj, Ensor, and Ostdiek (2008) construct a Cross-Market Tick-Matching (CMTM) estimator that better addresses asynchronous price observations. For asset A, the interval of observation times I^i is defined as $(t_{i-1}, t_i]$. For asset B, the interval of observation times J^j is defined as $(t_{j-1}, t_j]$. This matches the ticks of the faster asset (B) to the slower asset (A) by: $p_B(t_i^\wedge) = p_B(\min\{t_j \in \Pi^B : t_j \geq \max\{t_i \in I^i\}\})$ and $p_B(t_i^\vee) = p_B(\max\{t_j \in \Pi^B : t_j \leq$

$\min\{t_i \in I^i\}$). Using the slower asset(A) as the base arrival process we sample with respect to M_A the number of intervals between observations for asset(A). This allows us to write:

$$CMTM = \sum_i^{M_A} \Delta p_A(I^i)(p_B(t_i^\wedge) - p_B(t_i^\vee)) = \sum_i^{M_A} r_A(I^i)r_B(t_i^\vee, t_i^\wedge]. \quad (3.8)$$

Kyj, Ensor, and Ostdiek (2008) determine the optimal sampling frequency for the CMTM estimator with respect to the mean squared error (MSE) criterion. We define π as the probability of synchronous observations. The optimal solution is approximated by selecting:

$$k \approx \left(\frac{8\pi^2 M_A^3 \xi_{AB}^2}{\rho^2 \sigma_A^2 \sigma_B^2 + \sigma_A^2 \sigma_B^2} \right)^{1/3}. \quad (3.9)$$

The corresponding sampling frequency for the tick-time estimated variance is found by simple substitution as:

$$k \approx \left(\frac{8M_A^3 (\xi_A^2)^2}{2\sigma_A^4} \right)^{1/3}. \quad (3.10)$$

Section 3.3 will compare the performance of volatility-timing portfolios using the 5 and 15 minute calendar-time (CT5) and (CT15), optimally sampled calendar-time (CTO), CMTM sampled every 5 ticks (TT5), and optimally sampled CMTM (TTO) covariance estimators. We also consider sub-sampled versions of the optimally sampled calendar-time (CTO-S), CMTM sampled every 5 ticks (TT5-S), and optimally sampled CMTM (TTO-S) covariance estimators. These estimators will be considered using both rolling estimators and single index modeling discussed in Sections 3.2.3 and 3.2.4 respectively.

3.2.2 Ill-Conditioned Covariance Matrices

Many applied financial problems require a covariance matrix estimator that is not only invertible, but also well-conditioned. The true covariance matrix is well-conditioned, but estimators may not be due to sampling error. The sample covariance matrix is a consistent estimate of the true covariance matrix as $\frac{p}{n} \rightarrow 0$, but when $\frac{p}{n} \rightarrow c$ it may be ill-conditioned. When the sample covariance matrix is not consistent it is due to the accumulation of a large number of small errors off the diagonal. Michaud (1989) points out that within the MV context, ill-conditioned covariance estimates results in exaggerated estimation error. This problem has been identified as a barrier to practitioner adoption of the MV framework.

A positive definite matrix is necessary for matrix inversion, an essential step in the MV framework. The following three tests are necessary and sufficient conditions for a symmetric matrix A to be positive definite:

1. $x^T A x > 0$ for all nonzero vectors x .
2. All the eigenvalues of A satisfy $\lambda_i > 0$.
3. All the upper left submatrices A_k have positive determinants.

The relationship between positive definiteness and invertibility is understood via the eigenvalues. The determinant is defined as: $\det(A) = \prod_{i=1}^p \lambda_i$. A matrix is invertible when the $\det(A) \neq 0$. Hence the second test ensures that a positive definite matrix is invertible.

A well-conditioned operator is defined as having the property that all small perturbations of x lead to only small changes in $f(x)$. The condition number of a matrix A is defined as: $\kappa(A) = \|A\|_F \|A^{-1}\|_F$, where $\|\cdot\|_F$ is the Frobenius norm and can be ex-

pressed as: $\|A\|_F = \sqrt{\text{tr}(AA^T)}$. We note that $\|A\|_F^2 = \sum_{i=1}^p \sum_{j=1}^p A_{ij}^2$, and in our case A is symmetric so $\|A\|_F^2 = \sum_{i=1}^p \lambda_{A_i}^2$ where λ_{A_i} are the eigenvalues of A . In this setting, the condition number can be interpreted as the eccentricity of the ratio of eigenvalues. An ill-conditioned matrix is close to being non-invertible.

The definitions above indicate that the relative magnitude of eigenvalues of realized covariance matrices plays a very prominent role in MV asset allocation. Positive definiteness requires the eigenvalues be positive, and the well-conditioned property imposes an additional requirement of proportional eigenvalues. Imposing more structure can help mitigate the imprecision of realized covariance matrices. The two methods discussed in Sections 3.2.3 and 3.2.4 have been shown to be consistent estimators of the true covariance matrix and are well-conditioned.

3.2.3 Rolling Estimators

Conditional heteroskedasticity is a well know property of financial time series. In the presence of conditional heteroskedasticity, estimation of the covariance matrix requires a delicate balance between considering a sufficiently large number of observations to obtain an unbiased and consistent estimate, and adaptive enough to accommodate changes in in the covariance structure. Rolling realized covariance estimation attempts to balance the statistical power obtained using a large sample against potential problems caused by heteroskedasticity. Foster and Nelson (1996) demonstrate that exponential weighting minimizes the asymptotic MSE of rolling variance estimators. In a related empirical study, Andreou and Ghysels (2002) confirmed that exponentially weighted rolling estimators provided MSE efficiency gains for realized covariance estimators.

Rolling estimation is a common feature in realized covariance applications (See Fleming, Kirby, and Ostdiek (2003), Bandi, Russell, and Zhu (2008), de Pooter, Martens, and van Dijk (2006), Bandi and Russell (2006)). We follow the framework outlined in Fleming, Kirby, and Ostdiek (2003) and construct exponentially weighted rolling covariance estimators for Σ_t where:

$$\tilde{\Sigma}_t = (1 - \alpha)\tilde{\Sigma}_{t-1} + \alpha\hat{C}_{t-1}. \quad (3.11)$$

Equation 3.11 resembles a GARCH(1,1) and this relationship motivates adopting the GARCH estimation technique for selecting the optimal decay parameters. As noted by, Kawakatsu (2006) and Tse and Tsui (2002), the multivariate GARCH models suffer from the “curse of dimensionality” in that the number of parameter estimates grows quadratically, and equally important, high dimensional GARCH models have difficulty maintaining positive definiteness of the covariance matrix. Moreover, Andersen, Bollerslev, Diebold, and Labys (2003) state that the realized covariance matrix will also display difficulty maintaining positive definiteness as the dimensions increase. Imposing some simplification is necessary and Fleming, Kirby, and Ostdiek (2001) suggest an adjustment, where only one α is estimated for the entire covariance matrix, as opposed to a different $\alpha_{i,j}$ for every element of the covariance matrix.

We estimate α by means of maximum likelihood estimation as outlined in Tse and Tsui (2002). Fleming, Kirby, and Ostdiek (2001) and de Pooter, Martens, and van Dijk (2006) both examine the possible look-ahead bias due to fitting the decay parameter α using all available data and conclude that the method used to evaluate the performance of the different portfolios are robust to decay parameter estimation. In fact, Fleming, Kirby, and

Ostdiek (2001) show that volatility timing is more effective when estimates are smoothed more than is optimal according to the asymptotic MSE criterion.

Proposed by Ledoit and Wolf (2004), shrinkage seeks to minimize the expected quadratic loss $E[\|\tilde{\Sigma}_T - \Sigma\|_F^2]$ where $\tilde{\Sigma}_T(\alpha) = (1 - \alpha)G + \alpha\hat{\Sigma}_T$, $\alpha \in [0, 1]$, and G is some matrix. Rolling estimation may be viewed as shrinkage in the time domain. It is known that the true covariance matrix Σ_t is positive definite and better conditioned than estimates of the covariance matrix estimates and this motivates shrinkage toward the true covariance matrix. Ledoit and Wolf (2004, 2003) applied shrinkage in a spatial setting to reduce the dimensionality. Sancetta (2008) applied shrinkage to time series dependent observations.

Finally, we discuss the conditioning of rolling estimators. In this setting, $\tilde{\Sigma}_t$ is assumed to be a consistent estimator of Σ_t . Rolling estimation shrinks the realized covariance estimator towards the more consistent rolling estimator and thereby provides better conditioned covariance matrices. First, we examine the positive definiteness of this estimator:

$$\begin{aligned} a^T \tilde{\Sigma}_t a &= a^T (1 - \alpha) \tilde{\Sigma}_{t-1} a + a^T \alpha \hat{C}_{t-1} a \\ &= (1 - \alpha) \underbrace{a^T \tilde{\Sigma}_{t-1} a}_{(1)} + \alpha \underbrace{a^T \hat{C}_{t-1} a}_{(2)} \end{aligned} \quad (3.12)$$

For $p > m$, where p is the dimension of the matrix, and m is the number of intraday observations sampled, \hat{C}_{t-1} will not be of full rank and hence (2) is not always positive definite. Using the consistency result from Foster and Nelson (1996), we can see that the quantity (1) is positive definite in expectation. Andreou and Ghysels (2002) state that rolling the realized estimator accelerates the convergence to the true covariance matrix. Hence, the rolling realized covariance estimator should be better conditioned.

3.2.4 Single-Index Model

Factor models capture data of high dimensionality using a parsimonious set of common factors. In finance, factor models are invaluable in simplifying the estimation of the covariance matrix. Assuming conditional independence of contemporaneous returns of a large number p of assets given a small number of K factors, one dramatically reduces the number of parameters needed to estimate to capture the cross-sectional dependence between returns. Chamberlain and Rothschild (1983) discuss the relationship between factor structure and asset pricing. Chan, Karceski, and Lakonishok (1999) and Jagannathan and Ma (2003) both show that factor models can reduce the variance of optimal mean-variance portfolios. Bollerslev and Zhang (2003) employ high-frequency data in a multi-factor model and find improved asset pricing predictions when compared with conventional monthly rolling estimates. We aim to assess realized covariance estimation within a single-index model.

The single-index model introduced by Sharpe (1963) states that:

$$r_{A,t} = \alpha_A + \beta_A r_{M,t} + \epsilon_{A,t} \quad (3.13)$$

In this setting r_A is the return of the individual stock, r_M is the return of the market represented by the Index, and α represents the non-systemic risk, and β represents the comovement with the systemic risk that is represented by the Index. We assume that $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and that residuals ϵ_A are uncorrelated to market returns. The resulting covariance matrix is:

$$\Phi = \sigma_{M,M}^2 \beta \beta^T + D \quad (3.14)$$

From ordinary least squares regression we know that $\beta_A = \frac{\sigma_{A,M}}{\sigma_M^2}$. We can estimate this

value by computing the realized covariance and realized variance of the index. The matrix D represents the diagonal matrix of $\delta_{i,i} = VAR(\epsilon)$, the residual variances. As stated in Mardia, Kent, and Bibby (1979), it is natural to set $\delta_{i,i} = s_{i,i}$. Where s_i^2 is the estimate of σ_i^2 , the estimated Φ can be written as:

$$\hat{\phi}_{i,j} = \begin{cases} s_i^2 & \text{if } i=j \\ s_0^2 bb^T & \text{if } i \neq j \end{cases} \quad (3.15)$$

The assumption from Equation 3.14 is that all the covariation between stocks is captured by the covariation estimated by the market index. Intuitively, the single factor is trying to capture all the systemic risk. This is similar to the shrinkage toward identity matrix estimator proposed by Ledoit and Wolf (2004). The covariance matrix of the single-factor model is always positive definite.

$$a^T \Phi a = a^T (\sigma_M^2 \beta \beta^T + D) a = \sigma_M^2 \underbrace{a^T \beta \beta^T a}_{>0} + \underbrace{a^T D a}_{>0}. \quad (3.16)$$

It is easy to see that for all nonzero vector a , $a^T D a > 0$ as D is a diagonal matrix of positive values. Likewise, $\beta^T a = v$, is a scalar, and as a result $a^T \beta \beta^T a = v^2 > 0$.

Having established positive definiteness, we now consider the conditioning of this covariance matrix. Fan, Fan, and Lv (2007) state that the major advantage of factor models is in the estimation of the inverse of the covariance matrix. They show that when $K = o(p)$, where K is the number of factors, the inverse of the factor model covariance matrix converges to the true inverse covariance faster than the inverse of the sample covariance matrix. Again, this faster convergence implies that the factor model is better conditioned than the full matrix.

3.2.5 Assessment Criteria

We consider an array of methods for portfolio allocation assessment. Chan, Karceski, and Lakonishok (1999) and Jagannathan and Ma (2003) advocate the use of Global Minimum Variance (GMV) due to problematic nature of μ_t the vector of expected returns. Indeed, portfolios constructed using the historical mean as an estimate for expected returns fail to outperform a naive assumption of $\mu_t = 0$. The GMV circumvents this issue as it does not depend on estimates of μ_t and solves the following optimization problem:

$$\begin{aligned} \min_{w_t} \quad & w_t' \Sigma_t w_t \\ \text{s.t.} \quad & w_t' j = 1 \end{aligned} \tag{3.17}$$

Where Σ is the covariance matrix and j is a unitary vector of length p , the GMV weights are given as:

$$w_{t,GMV} = \frac{\Sigma_t^{-1} j}{j' \Sigma_t^{-1} j} \tag{3.18}$$

We also consider a portfolio that minimizes variance given a set target return as outlined in de Pooter, Martens, and van Dijk (2006). We choose this construction because like the GMV it requires that the investor be fully invested in the market and allows for loose comparison. We circumvent the previously mentioned concerns regarding estimating μ_t by bootstrapping a distribution of possible realizations from historical returns. This simulation approach was suggested in Jorion (1992). The Sharpe Ratio Minimum Variance (SRMV) is a linear combination of GMV portfolio weights and Maximum Sharpe Ratio (MSR) weights for a target return μ_P . The Sharpe ratio is defined as: $\frac{\mu_P}{\sigma_P}$. The optimization

problem is given as:

$$\begin{aligned} \min_{w_t} \quad & w_t' \Sigma_t w_t \\ \text{s.t.} \quad & w_t' \mu_t = \mu_P \quad \text{and} \quad w_t' j = 1 \end{aligned} \quad (3.19)$$

where we define $\mu_t = E[r_t]$ and μ_P is the target return for the portfolio. The weights of the MSR portfolio are given as:

$$w_{t,MSR} = \frac{\Sigma_t^{-1} \mu_t}{j' \Sigma_t^{-1} \mu_t} \quad (3.20)$$

and the weights for the target return portfolio are the given by:

$$w_{t,SRMV} = \frac{\mu_{t,MSR} - \mu_P}{\mu_{t,MSR} - \mu_{t,GMV}} w_{t,GMV} + \frac{\mu_P - \mu_{t,GMV}}{\mu_{t,MSR} - \mu_{t,GMV}} w_{t,MSR} \quad (3.21)$$

where we define $\mu_{t,MSR} = w_{t,MSR}' r_t$ and $\mu_{t,GMV} = w_{t,GMV}' r_t$.

Tracking Error (TE) is a measure of imperfect replication of a given benchmark portfolio. Let b be a vector of benchmark returns, and r_P be a vector of portfolio returns, where both vectors are of length N , then the TE is given in Equation 3.22.

$$TE = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (r_{P,i} - b_i)^2} \quad (3.22)$$

Minimum tracking error portfolios are of great interest as in practice it is often necessary to construct portfolios using a subset of all available stocks due to transaction costs or liquidity constraints. Jorion (2003) outlines implementing portfolio optimization with tracking error constraints.

Fleming, Kirby, and Ostdiek (2001, 2003) popularized volatility-timing as an objective methodology for assessing the economic “value-added” of alternative estimators. The value-added is measured in a portfolio framework as the fee (as a percent of assets) a risk-averse investor would be willing to pay to capture the gains in portfolio performance made

possible by using a given covariance estimator. The assessment is made based on the investor's utility improvement which is converted into a management fee and expressed in basis points (i.e. 0.01%). Specifically, investors follow a volatility-timing strategy where the portfolio weights vary only with changes in estimates of the conditional covariance matrix of daily returns.

The realized daily utility generated by this portfolio (R_{pt}) for an investor with γ relative risk aversion is

$$U(R_{pt}) = W_0 \left((1 + R_f + R_{pt}) - \frac{\gamma}{2(1 + \gamma)} (1 + R_f + R_{pt})^2 \right). \quad (3.23)$$

R_f is the risk-free rate and W_0 is the amount of wealth that is invested. Misidentification of the utility function is not a great concern as Chopra and Ziemba (1993) state that several different utility functions result in similar portfolio allocations for similar levels of risk aversion. The incremental value of using the second estimator instead of the first is calculated by finding the constant Δ , such that $\sum_{t=1}^T U(R_{p1t}) = \sum_{t=1}^T U(R_{p2t} - \Delta)$. This can be thought of as the basis point fee an investor would be willing to pay to access the second estimator instead of the first.

3.3 Empirical Analysis

3.3.1 Data: Dow Jones Industrial Average

We consider the the covariance structure of the Dow Jones Industrial Average (DJIA) over the period from January 2002 to December 2006. The DJIA is a price weighted index of 30 blue-chip stocks representative of major US industrial corporations. It is is a scaled average, as the divisor is adjusted to accommodate for stock splits and other corporate actions which influence the relationship between price and market capitalization.

Table 3.1 Composition of Dow Jones Industrial Average from January 1, 2002 to December 31, 2006.

Company Name	Symbol	Active (DAYMONTHYEAR)
Alcoa Inc.	AA	
Altria Group	MO	
Amer Express	AXP	
Amer International Group	AIG	08042004-31122006
AT&T Corp.	T	01012002-07042004 & 22112005-31122006
Boeing Co.	BA	
Caterpillar Inc.	CAT	
Citigroup	C	
Coca-Cola Co	KO	
dupont(e.i.)denemours	DD	
Eastman Kodak	EK	01012002-07042004
Exxon Mobil	XOM	
Genl Electric	GE	
Genl Motors	GM	
Hewlett-Packard Co.	HP	
Home Depot	HD	
Intel Corp.	INTC	
Intl Bus. Machines	IBM	
Intl Paper	IP	01012002-07042004
J.P. Morgan Chase	JPM	
Johnson & Johnson	JNJ	
Mcdonald's Corp	MCD	
Merck & Co	MRK	
Microsoft Corp	MSFT	
3M Co	MMM	
Pfizer Inc.	PFE	08042004-31122006
Procter & Gamble	PG	
SBC Communications Inc.	SBC	01012002-21112005
United Technologies Corp.	UTX	
Verizon Communications Inc.	VZ	08042004-31122006
Wal-Mart Stores	WMT	
Walt Disney Co.	DIS	

We consider a single factor model with the DJIA as the sole factor. We estimate the covariance of the returns of the 30 components with the DJIA futures contract traded on the Chicago Board of Trade, using the symbol DJ. These futures contracts are pit traded and only changes in transaction prices are recorded. The stock price data was obtained from the TAQ database and was filtered according to the trade-quote matching technique suggested in Lee and Ready (1991) and Henker and Wang (2006). Appendix A.3 provides details. The futures data was obtained from TickData Inc.

We limit our data set to observations posted from 10:00am EST to 4:00pm EST. The first 30 minutes of every trading day are omitted to eliminate the effects of the opening call auction. As stated in Hansen and Lunde (2005), overnight returns are very noisy and bring additional estimation challenges. Our risk averse investors take positions at market open and liquidate at market close, thereby eliminating overnight returns from our analysis.

3.3.2 Computation

For 30 assets we need to estimate 465 elements every single day for each covariance estimation technique. This is a computational challenge. We employed “Ada” - the Cray XD1 Cluster at Rice University to help speed up the computation. We parallelized the process by decomposing the covariance matrix into 36 sub-blocks each of dimension $p = 5$. Then due to the symmetry, we only considered the upper triangle which contained 21 blocks. Figure 3.1 shows how the covariance matrix was partitioned. We further decomposed the computation, by partitioning the data set into years. We run $21 \times 5 = 105$ parallel processes to perform the necessary estimations. This allowed us to run the computation in approximately $\frac{1}{105}$ the computing time required if we were to run this using serial code.

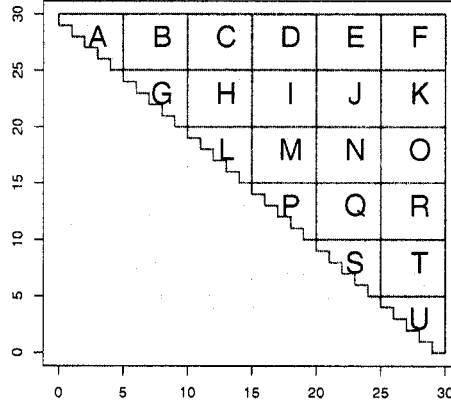


Figure 3.1 Partition of Covariance Matrix into sub-matrices of dimension 5. Each sub-matrix is run independently. The computation is accelerated further by computing each year independently. This allows for parallel computation in $\frac{1}{105}$ of the time necessary when using a serial technique.

3.3.3 Results

In Table 3.2 we present the optimal decay parameter estimates for the entire covariance matrix α_{Σ} , as well as for just the diagonal elements α_D . Recall from Section 3.2.3 that the exponential smoothing is of the form:

$$\tilde{\Sigma}_t = (1 - \alpha)\tilde{\Sigma}_{t-1} + \alpha\hat{C}_{t-1}. \quad (3.24)$$

The α parameter represents the degree of local persistence. In general, tick-time methods display greater local persistence as more weight is put on recent lags, as $TT5 = .0967$ and $TTO = .0709$. In contrast, the α of the calendar-time estimator sampled every 15 minutes, $CT15 = .0320$, puts less weight on recent lags and includes a larger window of values to average. These results are consistent with Foster and Nelson (1996); as the sampling intervals get finer, the rolling regression estimator includes a growing number of observations generated over a shrinking period of time. The α_D estimates are presented to

show that the diagonal elements display a greater degree of persistence. The comparison suggests that recent lags play a greater role in volatility estimation, whereas covariance is a longer memory process.

Table 3.2 Optimal weight parameters for 8 different realized covariance estimators. Weights are smallest for ad hoc calendar-time estimators. Optimal weight parameters for variance only estimators are larger than corresponding covariance estimators.

Model	CT5	CT15	CTO	CTOS	TT5	TT5S	TTO	TTOS
α_{Σ}	.0519	.0320	.0814	.0809	.0967	.1035	.0709	.0762
α_D	.2326	.1720	.2641	.2571	.2845	.2886	.2369	.2597

In Table 3.3 we consider the performance of a full realized covariance matrix of dimension 5 against the exponentially weighted rolling covariance matrix, “Rolling” hereafter, and a single-index model, “Factor” hereafter. Characteristics include the resulting portfolio returns: mean, the standard deviation (St. Dev), the Sharpe ratio (SR), tracking error (TE), correlation to market (ρ_M), and basis points gains relative to and equally weighted portfolio for investors with risk aversion $\gamma = 1$ and $\gamma = 10$. In panel (a) we include present the DJIA futures (DJIA), and in panel (b) a naive, equally weighted portfolio (Equal). The five stocks considered are: American Express (AXP), 3M Corporation (MMM), Merck (MRK), Proctor and Gamble (PG), and Walmart (WMT). These stocks were selected as they are included in the DJIA for the entire duration of the study and they are representative of the entire DJIA spanning the financial, conglomerates, health care, consumer goods, and services sectors respectively. Panels (c-e) present the results of the full covariance matrix, the rolling covariance matrix, and the single-index model. We see that the rolling estimators obtained the smallest minimum variances. The full covariance matrix performed

poorly, displaying high variances, low Sharpe ratios and negative utility gains relative to the equally weighted portfolio. The ad-hoc calendar time methods performed best in the rolling covariance matrix setting, but were outperformed by the other estimators in the factor model setting. The Sharpe ratios are generally lower in the factor models than the rolling covariance matrices. Of interest, sub-sampling failed to produce smaller portfolio volatilities in rolling and single-factor covariance matrix settings. As anticipated, the tracking errors are lower in the factor models than in the full or rolling matrices. The single-factor model shows utility gains of 50 to 150 basis points over the equally weighted portfolio. The ad-hoc calendar-time estimators show larger gains in the rolling covariance matrix setting. Table 3.3 shows that for even low dimensional settings, imposing structure on the realized covariance estimators can generate utility for risk averse investors.

We consider the performance characteristics of the GMV portfolio in a higher dimensional setting using the 30 stocks that make up the Dow Jones Industrial Average Index. In Table 3.4 we show the results. The full covariance matrix is not presented as it is ill-conditioned at this dimensionality. Instead we focus on the performance of the rolling covariance matrix and the single-factor model using the different realized covariance estimators. In the rolling covariance matrix setting, the CT5 method obtains the smallest minimum variance. The TTO-S is the second smallest minimum variance. In the single-index setting, all realized estimators outperform an equally-weighted portfolio. The lowest minimum variances are obtained by the CT5, CTO-S, and TTO-S. Of interest, the CT15, TT5, and TT5-S have the largest minimum variances. The correlation to market statistic shows that the portfolios generated by the single-index model are more closely correlated to the benchmark of the DJ futures. In both the single-index model and the rolling covariance

matrix the CT15 estimator has the highest Sharpe ratio. In particular, the exceptionally large Sharpe ratios of the rolling CT5 and CT15 relative to the equally weighted portfolio appear unrealistic. A limitation of the GMV methodology is that it only considers one realization for a given optimized portfolio. Jorion (1992) suggests bootstrapping possible realizations and evaluating the distribution of MV portfolio returns as a better method of assessment.

We bootstrap 150 possible realizations for the expected mean μ_t and implement the SRMV portfolio allocation method. The number of realizations to bootstraps was determined by convergence diagnostics. The results are presented in Table 3.5. By construction the variances of the portfolio returns are larger than in Table 3.4. The bootstrapping addressed the concern with the Sharpe ratios, and we see that we no longer have Sharpe ratios that are many times larger than the benchmark equally weighted portfolio.

We see that the portfolio returns using the ad hoc calendar-time rolling estimators have the highest Sharpe ratios, but also may have large negative utility losses. The CT5 estimation technique using the rolling estimator results in 76 basis point gains for the investor with mild risk aversion ($\gamma = 1$). This is a much smaller than the 378 basis point gain reported for CT5 in Panel (c) of Table 3.4. A more risk averse investor with $\gamma = 10$ obtains utility gains of 433 annual basis points. Again, this is smaller than the 767 previously reported. We see that the rolling CT5, CT15, and CTOS outperforms the equally weighted portfolio with respect to Sharpe ratio. The other methods fail to outperform the equally weighted portfolio. The large variances in the portfolio returns and large negative basis point losses suggest that the rolling estimators may be ill-conditioned and provide poor portfolio allocation strategies.

Table 3.3 Performance Characteristics of Global Minimum Variance (GMV) portfolios using 5 stocks.

Model	Mean	St. Dev.	SR	TE	ρ_M	$bp_{\gamma=1}$	$bp_{\gamma=10}$
(a) Market							
DJIA	0.0037	0.1424	0.0260	0.0000	1.0000		
(b) Benchmark							
Equal	0.1258	0.1401	0.8980	0.0774	0.8499	0	0
(c) Full							
CT5	0.1159	0.1317	0.8799	0.0914	0.7806	-87	15
CT15	0.0924	0.1383	0.6679	0.1096	0.6952	-332	-310
CTO	0.1174	0.1308	0.8974	0.0835	0.8164	-72	42
CTOS	0.1151	0.1309	0.8792	0.0836	0.8161	-94	18
TT5	0.1208	0.1306	0.9250	0.0854	0.8077	-37	79
TT5S	0.1178	0.1303	0.9042	0.0846	0.8112	-66	53
TTO	0.1155	0.1322	0.8738	0.0892	0.7914	-92	5
TTOS	0.1039	0.1311	0.7923	0.0870	0.8008	-207	-97
(d) Rolling α_Σ							
CT5	0.1282	0.1260	1.0170	0.0822	0.8191	43	211
CT15	0.1330	0.1241	1.0720	0.0880	0.7902	94	285
CTO	0.1215	0.1294	0.9389	0.0785	0.8373	-28	101
CTOS	0.1214	0.1295	0.9378	0.0785	0.8375	-29	99
TT5	0.1239	0.1281	0.9667	0.0800	0.8304	-3	142
TT5S	0.1240	0.1283	0.9668	0.0800	0.8300	-2	141
TTO	0.1256	0.1280	0.9814	0.0799	0.8307	15	160
TTOS	0.1241	0.1281	0.9686	0.0796	0.8319	0	144
(e) Factor							
CT5	0.1280	0.1324	0.9668	0.0779	0.8417	33	127
CT15	0.1242	0.1324	0.9380	0.0805	0.8307	-6	89
CTO	0.1293	0.1316	0.9831	0.0776	0.8425	47	152
CTOS	0.1280	0.1317	0.9722	0.0776	0.8424	34	137
TT5	0.1300	0.1320	0.9843	0.0771	0.8446	53	152
TT5S	0.1294	0.1319	0.9808	0.0771	0.8447	48	147
TTO	0.1288	0.1320	0.9753	0.0785	0.8390	41	140
TTOS	0.1253	0.1318	0.9514	0.0780	0.8407	7	109

Table 3.4 Performance Characteristics of Global Minimum Variance (GMV) portfolios using the 30 stocks within the Dow Jones Industrial Average.

Model	Mean	St. Dev.	SR	TE	ρ_M	$bp_{\gamma=1}$	$bp_{\gamma=10}$
(a) Market							
DJIA	0.0037	0.1424	0.0260	0.0000	1.0000		
(b) Benchmark							
Equal	0.0320	0.1440	0.2225	0.0516	0.9351	0	0
(c) Rolling α_Σ							
CT5	0.0655	0.1100	0.5958	0.0869	0.7923	378	767
CT15	0.1387	0.1197	1.1590	0.1204	0.5897	1098	1387
CTO	-0.0901	0.2906	-0.3101	0.2725	0.3686	-1540	-4431
CTOS	0.1019	0.2013	0.5061	0.1771	0.5132	599	-294
TT5	0.0283	0.1154	0.2452	0.0680	0.8814	0	333
TT5S	0.0247	0.1154	0.2144	0.0678	0.8823	-36	298
TTO	0.0186	0.1164	0.1593	0.0722	0.8631	-99	224
TTOS	0.0237	0.1147	0.2066	0.0699	0.8739	-46	295
(d) Factor							
CT5	0.0403	0.1290	0.3121	0.0539	0.9259	102	287
CT15	0.0486	0.1296	0.3754	0.0562	0.9190	185	363
CTO	0.0420	0.1292	0.3251	0.0536	0.9266	119	302
CTOS	0.0405	0.1289	0.3139	0.0537	0.9265	105	289
TT5	0.0399	0.1296	0.3080	0.0534	0.9273	98	276
TT5S	0.0410	0.1297	0.3157	0.0528	0.9290	109	284
TTO	0.0418	0.1293	0.3230	0.0536	0.9267	117	298
TTOS	0.0396	0.1289	0.3072	0.0537	0.9266	96	281

Table 3.5 Performance Characteristics of maximum Sharpe Ratio Minimum Variance (SRMV) portfolios using the 30 stocks of the Dow Jones Industrial Average where $\mu_P = 5\%$

Model	Mean	St. Dev.	SR	TE	ρ_M	$bp_{\gamma=1}$	$bp_{\gamma=10}$
(a) Market							
DJIA	0.0037	0.1424	0.0260	0.0000	1.0000		
(b) Benchmark							
Equal	0.0320	0.1440	0.2225	0.0516	0.9351	0	0
(c) Rolling α_Σ							
CT5	0.0356	0.1130	0.3154	0.0863	0.7963	76	433
CT15	0.1045	0.2666	0.3918	0.2627	0.4922	-1527	-1490
CTO	0.0551	0.3796	0.1452	0.3387	0.7098	-9489	-1251
CTOS	0.0340	0.1316	0.2500	0.0840	0.8212	34	164
TT5	0.0149	0.1170	0.1273	0.0684	0.8791	-136	180
TT5S	0.0116	0.1171	0.0992	0.0682	0.8799	-169	146
TTO	0.0113	0.1965	0.0573	0.1549	0.7953	-1246	-459
TTOS	0.0101	0.1178	0.0857	0.0720	0.8649	-186	120
(d)Factor							
CT5	0.0386	0.1297	0.2976	0.0553	0.9218	85	261
CT15	0.0375	0.1696	0.2208	0.1007	0.8737	-354	-93
CTO	0.0407	0.1298	0.3133	0.0551	0.9225	105	280
CTOS	0.0397	0.1296	0.3067	0.0552	0.9224	97	273
TT5	0.0398	0.1301	0.3064	0.0550	0.9229	97	268
TT5S	0.0408	0.1302	0.3135	0.0545	0.9244	107	276
TTO	0.0410	0.1299	0.3160	0.0552	0.9223	109	282
TTOS	0.0391	0.1296	0.3015	0.0552	0.9223	90	266

In Panel (d) of Table 3.5 we show the results using the single-index model. With the exception of the CT15, all of the methods outperform the equally weighted portfolio with respect to Sharpe ratio. Employing any of these strategies generates approximately 100 basis point gains over an equally weighted portfolio for an investor with risk aversion $\gamma = 1$ and approximately 270 basis points for an investor with risk aversion $\gamma = 10$. We can conclude that a risk averse investor can obtain the same level of utility when using a realized volatility estimation technique in a single-factor model as from the best estimation technique in a smoothed covariance matrix setting. The single-index model CTO, TTO, and TT5S estimators all have Sharpe ratios that match the CT5 estimator in the rolling covariance matrix setting. The factor models have greater variance and so a very risk averse investor $\gamma = 10$ will still be willing to pay an additional 150 basis points to use the CT5 rolling covariance matrix over any single-factor model.

A rolling realized covariance matrix provides minimized variance in portfolio returns. When performance is assessed in terms of expected returns and variance, then single-index models offer a computationally convenient alternative. Our results suggest that the single-factor model performs on par with more computationally intensive methods of estimating the entire covariance matrix. Moreover we see that the returns using the single-index model are more strongly correlated with the benchmark index, and this encourages future assessment of minimal tracking error portfolios.

3.4 Conclusion

In this chapter we have compared a number of realized covariance techniques and the characteristics of the Markowitz mean-variance optimized portfolios they generate. We

compared the performance of two technique for improving the conditioning of matrices; exponentially weighted rolling realized covariance matrix and a single-index model. We argue that in the presence of linear versus quadratic growth in computational complexity and ill-conditioned matrices, factor models are a more natural setting for comparing covariance estimators.

In the future we will consider alternatives to the rolling estimator such as the shrinkage toward market estimator proposed in Ledoit and Wolf (2003), the exponential matrices as suggested in Kawakatsu (2006), and the regularization of large covariance matrices as suggested by Bickel and Levina (2008). The first two methods are computationally intensive, requiring estimation of all the covariance elements. The third method offers a computationally parsimonious alternative.

From a financial perspective, the strong performance of the factor model in this setting motivates further exploiting the computational efficiency of this model. The next step is to consider a portfolio which minimizes tracking error. Our objective would be to develop portfolios with minimized tracking error which incur minimal trading costs by holding relatively few assets. This would be of great interest for the finance community where many fund managers are assessed according to their ability to track indices.

Chapter 4

Derivation and Simulation of Cross-Market Tick-Matching Estimator

Tick time estimation offers an advantage of not rely upon any artificial grids. Traditional calendar-time estimator are subject to a downward bias introduced by interpolating asynchronous observations onto a synchronous sampling grid. Figure 4.1 offers a graphical interpretation of the Hayashi and Yoshida (2005) tick-time estimator, which is simply the sum of the product of overlapping intervals. Palandri (2006) and Voev and Lunde (2007) both observe that due to computational efficiency it is advantageous to develop an aggregated version of this tick-time estimator. As seen by the dashed rectangles in Figure 4.1, aggregation preserves the informational content of the original tick-time estimator.

Cross-Market Tick Matching (CMTM) can be written as:

$$\begin{aligned}
 CMTM &= \sum_{i=1}^{M_A} \Delta p_A(I^i) (p_B(t_i^\wedge) - p_B(t_i^\vee)) \\
 &= \sum_{i=1}^{M_A} r_A(I^i) r_B(t_i^\vee, t_i^\wedge]
 \end{aligned} \tag{4.1}$$

where T is defined as the terminal time and $\Pi^A = \{t_i\}_{i=0,1,2,\dots,M_A}$ and $\Pi^B = \{t_j\}_{j=0,1,2,\dots,M_B}$ are the sets of observation times for processes A and B respectively, where $0 \leq t_i \leq T$ and $0 \leq t_j \leq T$ for all i, j . For process A, the interval of observation times I^i is defined as $(t_{i-1}, t_i]$. For process B, the interval of observation times J^j is defined as $(t_{j-1}, t_j]$. Cross-Market Tick Matching (CMTM) is an aggregate tick time estimator that matches the ticks of the faster process (B) to the slower process (A) by: $p_B(t_i^\wedge) = p_B(\min\{t_j \in \Pi^B : t_j \geq \max\{t_i \in I^i\}\})$ and $p_B(t_i^\vee) = p_B(\max\{t_j \in \Pi^B : t_j \leq \min\{t_i \in I^i\}\})$.

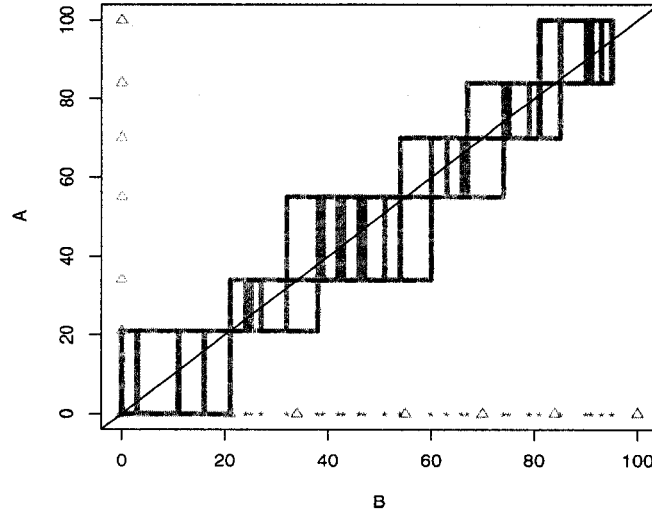


Figure 4.1 Aggregated Tick-Time Estimation. Grey rectangles represent a traditional Hayashi Yoshida tick-time estimator; the sum of the product of non-empty overlapping intervals. The dashed lines are a visual representation of aggregation scheme, which matches the ticks of the faster process (B) to the corresponding ticks of the slower process (A).

4.1 Derivation of Cross-Market Tick-Matching

The model framework is defined in Section 2.1.1. Building on this framework we discuss the derivation of the CMTM estimator. We begin by restating the changes in the price processes in terms the latent price processes and the MM noise. Then we expand the terms and decompose the product into four elements: 1) D_1 the cross-product of latent price processes, 2) D_2 the cross product of the MM noise of A and latent price process B, 3) D_3 the cross product of the MM noise of B and latent price process A, and 4) D_4 the

cross-product of MM noises.

$$\begin{aligned}
CMTM &= \sum_{i=1}^{M_A} \Delta p_A(I^i) (p_B(t_i^\wedge) - p_B(t_i^\vee)) \\
&= \sum_i^{M_A} (\Delta x_A(I^i) + \Delta u_{A,i}) (x_B(t_i^\wedge) - x_B(t_i^\vee) + \Delta u_{B,i}) \\
&= \sum_{i=1}^{M_A} \left(\underbrace{\Delta x_A(I^i) (x_B(t_i^\wedge) - x_B(t_i^\vee))}_{D1} + \underbrace{\Delta u_{A,i} (x_B(t_i^\wedge) - x_B(t_i^\vee))}_{D2} \right. \\
&\quad \left. + \underbrace{\Delta u_{B,i} \Delta x_A(I^i)}_{D3} + \underbrace{\Delta u_{A,i} \Delta u_{B,i}}_{D4} \right) \\
&= \sum_{i=1}^{M_A} (D1_i + D2_i + D3_i + D4_i)
\end{aligned}$$

We calculate the expected value of the CMTM estimator. First we use the fact that the expectation operator is linear, allowing us to rewrite this as the expectation of the four components.

$$\begin{aligned}
E(CMTM) &= E \left(\sum_{i=1}^{M_A} (D1_i + D2_i + D3_i + D4_i) \right) \\
&= E \left(\sum_{i=1}^{M_A} D1_i + \sum_{i=1}^{M_A} D2_i + \sum_{i=1}^{M_A} D3_i + \sum_{i=1}^{M_A} D4_i \right) \\
&= \sum_{i=1}^{M_A} E(D1_i) + \sum_{i=1}^{M_A} E(D2_i) + \sum_{i=1}^{M_A} E(D3_i) + \sum_{i=1}^{M_A} E(D4_i) \\
&= \sum_{i=1}^{M_A} \rho \sigma_A \sigma_B l(I^i \cap [t_i^\vee, t_i^\wedge]) + \sum_{i=1}^{M_A} E(\Delta u_{A,i} \Delta u_{B,i}) \\
&= \rho \sigma_A \sigma_B + 2\pi \sum_{i=1}^{M_A} \xi_{A,B}.
\end{aligned}$$

The operator $l(\cdot)$ is the length of the interval and π represents the probability of synchronous observations. The expectation of $D2$ and $D3$ are both zero by the assumption that the latent price process and MM noise are independent. Note, the noise accumulation in

the expected value is influenced by the extent of asynchronicity.

$$\begin{aligned}
 VAR(CMTM) &= VAR\left(\sum_{i=1}^{M_A} (D1_i + D2_i + D3_i + D4_i)\right) \\
 &= VAR\left(\sum_{i=1}^{M_A} D1_i + \sum_{i=1}^{M_A} D2_i + \sum_{i=1}^{M_A} D3_i + \sum_{i=1}^{M_A} D4_i\right) \\
 &= VAR\left(\sum_{i=1}^{M_A} D1_i\right) + VAR\left(\sum_{i=1}^{M_A} D2_i\right) + VAR\left(\sum_{i=1}^{M_A} D3_i\right) \\
 &\quad + VAR\left(\sum_{i=1}^{M_A} D4_i\right) + 2COV\left(\sum_{i=1}^{M_A} D2_i, \sum_{i=1}^{M_A} D3_i\right)
 \end{aligned} \tag{4.2}$$

Only one covariance term is included, the other terms equal zero due to the assumption of independence of MM noise with the latent price process. $VAR(D_1)$ is the variance of CMTM in the absence of noise and is a generalization of the result in Voev and Lunde (2007).

$$\begin{aligned}
 VAR\left(\sum_{i=1}^{M_A} D1_i\right) &= \sum_{i=1}^{M_A} ((\rho\sigma_A\sigma_B l(I^i))^2 + \sigma_A^2 l(I^i)\sigma_B^2 l((t_i^\vee, t_i^\wedge])) \\
 &\quad + 2\rho\sigma_A\sigma_B l(I^i \cap (t_{i-1}^\vee, t_{i-1}^\wedge])\rho\sigma_A\sigma_B l(I^{i-1} \cap (t_i^\vee, t_i^\wedge]))
 \end{aligned} \tag{4.3}$$

The second term, presented in Equation 4.4, considers the cross product of the latent price process and the MM noise. The first line follows due to the definition of variance and the expectation of $D_2 = 0$ due to the assumption of independence between the latent price process and the MM noise. The second line follows due to the linearity of the expectations operator. As illustrated in Figure 2.3, the overlap of the intervals is limited to the adjacent intervals. We consider the interval itself i , the preceding interval $i - 1$, and the following

interval $i + 1$.

$$\begin{aligned}
 VAR(\sum_{i=1}^{M_A} D2_i) &= E\left(\sum_{i=1}^{M_A} D2_i - E(\sum_{i=1}^{M_A} D2_i)\right)^2 = E\left(\left(\sum_{i=1}^{M_A} D2_i\right)^2\right) \quad (4.4) \\
 &= E\left\{\sum_{i=1}^{M_A} \sum_{i^*=1}^{M_A} D2_i D2_{i^*}\right\} = \sum_{i=1}^{M_A} \sum_{i^*=1}^{M_A} E\{D2_i D2_{i^*}\} \\
 &= \sum_{i=1}^{M_A} \sum_{i^*=i-1}^{i+1} E\{\Delta u_{A,i} r_B(t_i^\vee, t_i^\wedge) \Delta u_{A,i^*} r_B(t_{i^*}^\vee, t_{i^*}^\wedge)\}
 \end{aligned}$$

When $i^* = i$ then due to independence between the noise and efficient price process:

$$E\{(\Delta u_{A,i})^2\} E\{(r_B(t_i^\vee, t_i^\wedge))^2\} = 2\xi_A^2 \sigma_B^2 l(t_i^\vee, t_i^\wedge)$$

When $i^* = i - 1$ then,

$$E\{(\Delta u_{A,i} \Delta u_{A,i-1})\} E\{(r_B(t_i^\vee, t_i^\wedge) r_B(t_{i-1}^\vee, t_{i-1}^\wedge))\} = -\xi_A^2 \sigma_B^2 l((t_i^\vee, t_i^\wedge] \cap (t_{i-1}^\vee, t_{i-1}^\wedge]).$$

We have an overlap on both sides, hence by symmetry,

$$\begin{aligned}
 VAR(\sum_{i=1}^{M_A} D2_i) &= \sum_i^{M_A} (2\xi_A^2 \sigma_B^2 l(t_i^\vee, t_i^\wedge)) \\
 &\quad - \sum_{i=1}^{M_A-1} (\xi_A^2 \sigma_B^2 l((t_i^\vee, t_i^\wedge] \cap (t_{i-1}^\vee, t_{i-1}^\wedge]) + \xi_A^2 \sigma_B^2 l((t_i^\vee, t_i^\wedge] \cap (t_{i+1}^\vee, t_{i+1}^\wedge])) \\
 &= \sum_{i=1}^{M_A} 2\xi_A^2 \sigma_B^2 l(I^i)
 \end{aligned}$$

The calculations for $VAR(\sum_{i=1}^{M_A} D3_i)$ is simple due to the independence between the efficient price process and noise. Note, the Δu_B is defined by the A process and may not include the same error term, and as a result the MM error due to overlap terms do not play any role.

$$VAR(\sum_{i=1}^{M_A} D3_i) = \sum_{i=1}^{M_A} \sum_{i^*=i-1}^{i+1} E\{\Delta u_{B,i} \Delta r_A(I^i) \Delta u_{B,i^*} \Delta r_A(I^{i^*})\} = \sum_{i=1}^{M_A} 2\xi_B^2 \sigma_A^2 l(I^i) \quad (4.5)$$

The variance of the MM error terms is computed by rewriting in the variance as a covariance term. Then, as the covariance is a bilinear operator, we can move the covariance operator inside the double sum. The second line restates the covariance as a sum of diagonal and off-diagonal elements. Finally, recall that only adjacent intervals can overlap and have any contribution. Hence, we only consider covariance terms when $i^* = i + 1$.

$$\begin{aligned}
VAR\left(\sum_{i=1}^{M_A} D4_i\right) &= COV\left(\sum_{i=1}^{M_A} D4_i, \sum_{i^*=1}^{M_A} D4_{i^*}\right) \\
&= \sum_{i=1}^{M_A} \sum_{i^*=1}^{M_A} COV(D4_i, D4_{i^*}) \\
&= \sum_{i=1}^{M_A} COV(D4_i, D4_i) + 2 \sum_{i=1}^{M_A-1} \sum_{i^*=i+1}^{M_A} COV(D4_i, D4_{i^*}) \\
&= \sum_{i=1}^{M_A} VAR(D4_i) + 2 \sum_{i=1}^{M_A-1} COV(D4_i, D4_{i+1})
\end{aligned} \tag{4.6}$$

Now, using the definition of variance, we have

$$VAR(D4_i) = E(D4_i^2) - E(D4_i)^2. \tag{4.7}$$

We already know from the expectation calculation that $E(D4_i) = 2\pi\xi_{A,B}$. The first term is determined by Equation 4.8. This result is obtained by first writing out the difference formulas. Recall that the observations of process (B) are not indexed according to process (A). Our notation from the price process follows to the MM noise. Specifically, $u_B(i^\wedge) = u_B(\min\{t_j \in \Pi^B : t_j \geq \max\{t_i \in I^i\}\})$ and $u_B(i^\vee) = u_B(\max\{t_j \in \Pi^B : t_j \leq \min\{t_i \in I^i\}\})$. We then expand the multiplication and identify non-contemporaneous terms as “NC”. According to our assumption of the MM noise process, non-contemporaneous noise terms are uncorrelated. We expand our multiplication once more, and then take the expectation

of the resulting three terms. We use the fact that $E\{x_A^2 x_B^2\} = \sigma_A^2 \sigma_B^2 + 2\sigma_{AB}^2$.

$$\begin{aligned}
E(D4_i^2) &= E\{(\Delta u_{A,i} \Delta u_{B,i})^2\} \\
&= E\{((u_{A,i+1} - u_{A,i})(u_{B,(i+1)^\wedge} - u_{B,i^\vee}))^2\} \\
&= E\{(u_{A,i+1} u_{B,(i+1)^\wedge} - \underbrace{u_{A,i+1} u_{B,i^\vee}}_{NC} - \underbrace{u_{A,i} u_{B,(i+1)^\wedge}}_{NC} + u_{A,i} u_{B,i^\vee})^2\} \\
&= E\{(u_{A,i+1} u_{B,(i+1)^\wedge})^2 + 2u_{A,i+1} u_{B,(i+1)^\wedge} u_{A,i} u_{B,i^\vee} + (u_{A,i} u_{B,i^\vee})^2\} \\
&= 2\pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] + 2\xi_A^2 \xi_B^2 + 4\pi^2 \xi_{AB}^2
\end{aligned} \tag{4.8}$$

This results in

$$VAR(D4_i) = \underbrace{2\pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] + 2\xi_A^2 \xi_B^2 + 4\pi^2 \xi_{AB}^2}_{E(D4_i^2)} - \underbrace{4\pi^2 \xi_{AB}^2}_{E(D4_i)^2}. \tag{4.9}$$

We return to the final term in Equation 4.6.

$$\begin{aligned}
COV(D4_i, D4_{i+1}) &= E[D4_i D4_{i+1}] - E(D4_i)E(D4_{i+1}) \\
&= E[D4_i, D4_{i+1}] - E(D4_i)E(D4_{i+1}) \\
&= \pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] + 3\pi^2 \xi_{AB}^2 - 4\pi^2 \xi_{AB}^2 \\
&= \pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] - \pi^2 \xi_{AB}^2
\end{aligned} \tag{4.10}$$

Hence,

$$VAR(\sum_{i=1}^{M_A} D4_i) = \sum_{i=1}^{M_A} 2\pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] + \sum_{i=1}^{M_A} 2\xi_A^2 \xi_B^2 + \sum_{i=1}^{M_A-1} 2\pi[2\xi_{AB}^2 + \xi_A^2 \xi_B^2] - \sum_{i=1}^{M_A-1} 2\pi^2 \xi_{AB}^2 \tag{4.11}$$

Finally, we return to Equation 4.2 and determine $COV(\sum_{i=1}^{M_A} D2_i, \sum_{i=1}^{M_A} D3_i)$ as:

$$\begin{aligned}
 COV(\sum_{i=1}^{M_A} D2_i, \sum_{i=1}^{M_A} D3_i) &= \sum_{i=1}^{M_A} COV(D2_i, D3_i) \\
 &= \sum_{i=1}^{M_A} (E(D2_i D3_i) - E(D2_i)E(D3_i)) \\
 &= \sum_{i=1}^{M_A} E(\Delta u_{A,i} r_{B,i} \Delta u_{B,i} r_{A,i}) \\
 &= \sum_{i=1}^{M_A} E(r_A r_B) E(\Delta u_{A,i} \Delta u_{B,i}) \\
 &= \sum_{i=1}^{M_A} (\sigma_A \sigma_B \rho) * (2\pi \xi_{A,B})
 \end{aligned} \tag{4.12}$$

In the second line, we recall that the expectations of $D2_i$ and $D3_i$ are both equal to 0, and then the rest follows.

Putting all these pieces together we obtain the variance of the CMTM estimator.

$$\begin{aligned}
 VAR(CMTM) &= \sum_i^{M_A} (\rho \sigma_A \sigma_B)^2 l(I^i)^2 + \sigma_A^2 l(I^i) \sigma_B^2 l(t_i^\vee, t_i^\wedge]) \\
 &\quad + 2\rho \sigma_A \sigma_B l(I^i \cap (t_{i-1}^\vee, t_{i-1}^\wedge]) \rho \sigma_A \sigma_B l(I^{i-1} \cap (t_i^\vee, t_i^\wedge]) \\
 &\quad + 2\xi_A^2 \sum_i^{M_A} \sigma_B^2 l(I^i) + 2\xi_B^2 \sum_i^{M_A} \sigma_A^2 l(I^i) + \sum_i^{M_A} 2(1 + \pi) \xi_A^2 \xi_B^2 \\
 &\quad + \sum_i^{M_A-1} 2\pi \xi_A^2 \xi_B^2 + 4\pi \sum_i^{M_A} \xi_{AB}^2 + 2 \sum_i^{M_A-1} (2\pi - \pi^2) \xi_{AB}^2 \\
 &\quad + 4\pi \sum_i^{M_A} \sigma_A \sigma_B \rho \xi_{AB}.
 \end{aligned} \tag{4.13}$$

4.2 Simulation

Simulation studies of tick-time covariance estimators include Griffin and Oomen (2006) and Voev and Lunde (2007). We consider scenarios not previously considered: 1. Dynamic Arrival Rates, and 2. Brownian Motion plus Jump Process.

The first scenario is largely motivated by the financial duration modeling literature. Engle and Russell (1998) show strong evidence of deterministic time-of-day effects in the duration times of financial transaction data. Exploratory data analysis confirms that arrival rates are not constant, but rather display this well known diurnal (or U-shaped) pattern. Barndorff-Nielsen and Shephard (2002) note that the impact of diurnal effects on realized volatility should not be ignored. Figure 4.2 shows the estimated quote intensity as a function of time-of-day. A diurnal effect is observed with more quotes at the start and conclusion of trading day, and comparatively fewer quotes during the middle of the day. This intensity pattern can be modeled as a quadratic function. Dynamic arrival rates, which display a diurnal pattern, are a more accurate characterization of the price process. As a result, we will examine the impact on realized covariance estimators when observations arrive in a diurnal fashion.

The second scenario examines the impact of asynchronous jump processes on the realized estimators. Barndorff-Nielsen and Shephard (2004a) show that the calendar-time variance and covariance estimators are not robust to the inclusion of jump processes. Furthermore, Barndorff-Nielsen and Shephard (2004b) introduce realized bipower variation (BPV),

$$BPV(t) \sum_{i=1}^m |r_i|^1 |r_{i+1}|^1, \quad (4.14)$$

as a robust alternative to realized covariance estimation.

Andersen, Bollerslev, and Diebold (2005) build on the results in Barndorff-Nielsen and Shephard (2004b) and provide an empirical analysis of variance and jump estimation using the bipower variation methodology. Jumps are found to be prevalent across asset classes,

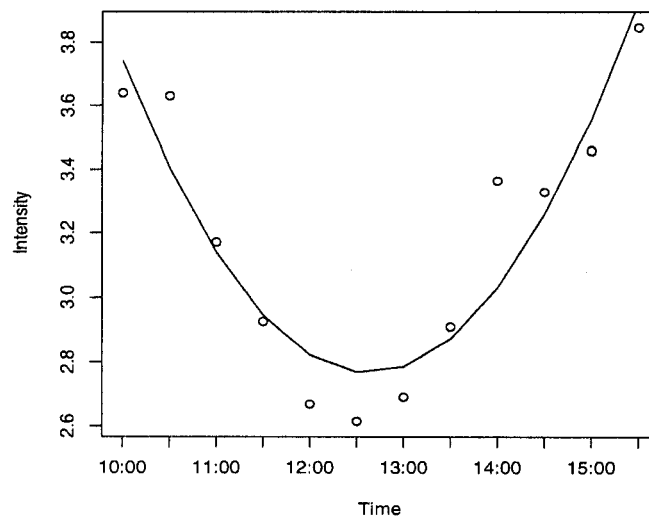


Figure 4.2 Number of quotes observed per minute. X-axis is the time of day, with the market opening at 9:30 and closing at 16:00. The Y-axis is the number of quotes observed per minute. A diurnal effect is observed with more quotes at the start and conclusion of trading day, and comparatively fewer quotes during the middle of the day.

but less persistent than the variation of the continuous process. An interesting finding in Barndorff-Nielsen and Shephard (2006) is that the jump process is better identified as the sampling interval decreases, this is in contrast to the sparse sampling advocated for diffusions contaminated by observational noise. In a related study, Oomen (2006) also considers the price process as a pure jump process. Optimal sampling is identified as a future avenue of research.

We present Monte Carlo simulation results comparing the efficiency of our proposed sub-sampled CMTM estimator. Each simulation consists of 2500 iterations. We evaluate the relative performance of the covariance estimators under contemporaneously correlated noise structures. Three different arrival regimes, in terms of expected duration between arrivals, are examined: 1.) low-low (30:30), 2.) low-high (30:10), and 3.) high-high

(10:10). These arrival regimes are representative of the subset of the TAQ quote data that we are studying. In order to consider the marginal contribution of tick matching, optimal sampling, and sub-sampling, we consider the five estimators presented in Table 4.1.

Table 4.1 Realized covariance estimators considered in simulation. Calendar time is abbreviated as (CT) and tick time is abbreviated as (TT).

Symbol	Estimator	Abbreviation
◦	5 minute calendar time	(CT5)
◊	15 minute calendar time	(CT15)
△	calendar time optimally sampled	(CTO)
×	5th tick estimator	(TT5)
+	optimally sampled CMTM estimator	(TTO)

4.2.1 Simulating Brownian Motion

The latent process is a bivariate Brownian motion. This process is observed at non-synchronous Poisson times. First the latent process is generated using the Euler scheme suggested by Higham (2001):

$$\begin{bmatrix} \Delta x_A \\ \Delta x_B \end{bmatrix} = \sqrt{\Delta t} \begin{bmatrix} \Theta^T \end{bmatrix} \begin{bmatrix} z_A \\ z_B \end{bmatrix} \quad (4.15)$$

where Θ^T is the Cholesky factorization of the covariance matrix such that $\Theta\Theta^T = \Sigma$. Σ is a 2×2 matrix. Z represents the Brownian motion and $z_i \sim N(0, 1)$. The inter-arrival times are exponential so arrival times are determined by the cumulative sum of the generated vector of exponential random variables and rounded to the nearest integer value. The resulting values determine which elements of the latent process will be observed. Non-synchronicity is achieved by simulating two different arrival time vectors. Finally, contemporaneously correlated microstructure effects are added to the selected latent values to create observed

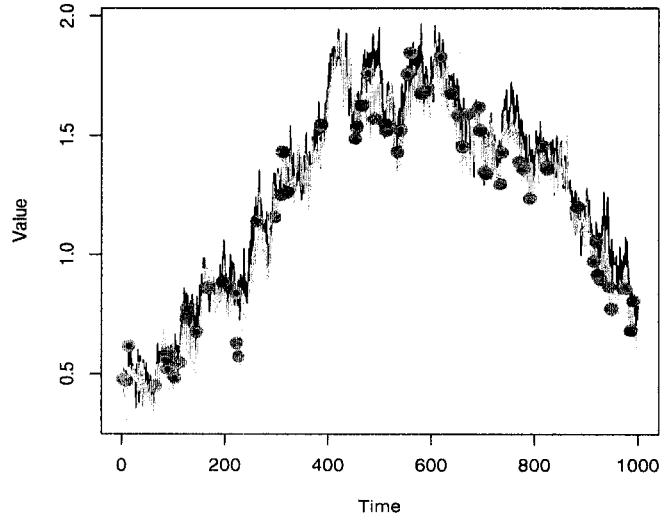


Figure 4.3 Realization of Bivariate Brownian Motion. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously.

values $(p_{A,t} = x_{A,t} + u_{A,t})$. Specifically, we consider $\sigma_{1,1} = \sigma_{2,2} = 1$ and $\sigma_{1,2} = 0.9$. We generate our realizations of the processes using $\Delta t = 1$ second. We assume that $\sigma_{1,1}$ and $\sigma_{2,2}$ are known and focus exclusively on estimating the covariance.

Figure 4.3 shows a realization of this process. The lines represent the Euler realization of the continuous time process when $\Delta t = 1$ second, and the points are the observations recorded at Poisson times. It is easy to see from the figure that the observations are asynchronous, yet obtained from correlated latent processes.

4.2.2 Dynamic Arrival Rate

We again model the underlying price process as a bivariate Brownian motion as given in Section 4.2.1. The arrival intensities are determined according to time-of-day dynamics. Let T be the terminal time in seconds, then the arrival intensity follows a diurnal pattern

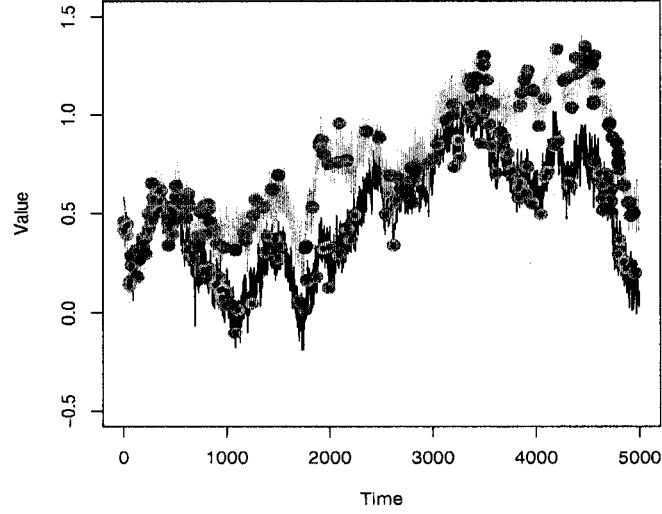


Figure 4.4 Realization of Bivariate Brownian Motion with Dynamic Arrival Rates. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously. The dynamic arrival rate is evidenced by the relatively fewer observations during the middle of the day.

given by:

$$\lambda(t) = a * (t - T/2)^2 + b/3 \quad \text{where} \quad a = b * (T/2)^{-2}. \quad (4.16)$$

In this equation b is the intensity rate at the start of the trading day. Figure 4.4 shows a realization of this process. The lines represent the Euler realization of the continuous time process when $\Delta t = 1$ second, and the points are the observations determined by the dynamic arrival rate. Observations are less frequent during the middle of the day. Also asynchronicity is more pronounced in the middle of the day.

4.2.3 Brownian Motion with Jumps

We also examine whether realized covariance estimates are robust in the presence of jumps. Barndorff-Nielsen and Shephard (2004a) have shown that calendar-time estimators

are not robust in the presence of jumps, and in turn we consider the performance of tick-time estimators. Following Glasserman (2004) this process can be generated by:

$$\begin{bmatrix} \Delta x_A \\ \Delta x_B \end{bmatrix} = \sqrt{\Delta t} \begin{bmatrix} \Theta^T \end{bmatrix} \begin{bmatrix} z_A \\ z_B \end{bmatrix} + \text{Jumpsize} \begin{bmatrix} B_A & 0 \\ 0 & B_B \end{bmatrix} \begin{bmatrix} J_A \\ J_B \end{bmatrix} \quad (4.17)$$

where B_i is distributed as a *Bernoulli*(q_i), and J_i is distributed as *lognormal*(0,1). In our simulations the probability of a jump occurring is $q = 0.0005$ for both processes, and $\text{Jumpsize} = 0.1$. The parameters of the bivariate Brownian motion are given in Section 4.2.1. Figure 4.5 shows a realization of this process. The lines represent the Euler realization of the continuous time process when $\Delta t = 1$ second, and the points are the observations. A jump in the grey process occurred within the dashed rectangle.

4.3 Results

Figure 4.6 presents the bias of the covariance estimators of a bivariate Brownian motion with dynamic arrival rates under the three different arrival regimes. Figures 4.7 and 4.8 show the associated mean squared error (MSE) plots. We see that under the low arrival regime (panel a), the tick-matching estimators are almost unbiased. In contrast, for very low noise-to-signal ratios the calendar-time optimal sampling methodology performs very poorly and displays a large negative bias. A comparison of the ad-hoc estimators (5 minute and 15 minute) shows that the 15 minute estimator is less biased. The 5 minute estimator performs worse than in constant arrival rate simulation in Figure 2.4. The large negative biases of the calendar time methods is due to the Epps effect induced by the greater degree of asynchronicity during the mid-day. The medium and high arrival regimes show bias re-

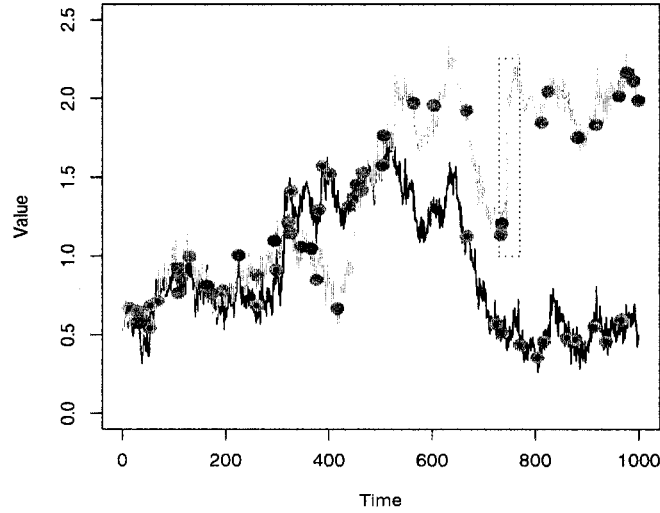


Figure 4.5 Realization of Bivariate Brownian Motion with Jumps. The lines represent the underlying price processes generated using an Euler scheme with step size $\Delta t = 1$ second. The points are the observations recorded at Poisson times. The two processes are strongly correlated, with $\rho = 0.9$, but observed asynchronously. We see a large jump in the grey process within the dashed rectangle.

ductions for all the calendar methods, with the CTO improving the most dramatically. This reflects the reduction in asynchronicity with more frequent observations. Even in the high arrival regime the optimal tick-time estimator is always less biased than the calendar-time alternatives for all noise-to-signal levels.

We also see the benefit of sampling in tick time. In contrast to calendar time, the ad-hoc tick-time estimator demonstrates a positive bias that increases with the noise-to-signal ratio and with arrival frequency. The difference in the biases of the TT5 and TTO estimators indicates that the benefits of optimal sampling in the tick-time setting persist even in a dynamic arrival rate setting.

In Figures 4.7 and 4.8 we see that for low noise-to-signal ratios the tick-time estimators have the smallest MSE, and the sub-sampled TTO estimator has the smallest MSE

overall. As the noise-to-signal ratio increases, the ad-hoc tick-time estimator becomes non competitive in the medium and high frequency settings. In the high frequency, high noise-to-signal setting the CT5, CTO, and TTO methods provide similar results. The optimally sampled calendar-time estimator is non-competitive for low noise, low frequency due to the large bias. Figure 4.8 shows that sub-sampling reduces the MSE for all estimators with the benefits increasing with the noise-to-signal ratio.

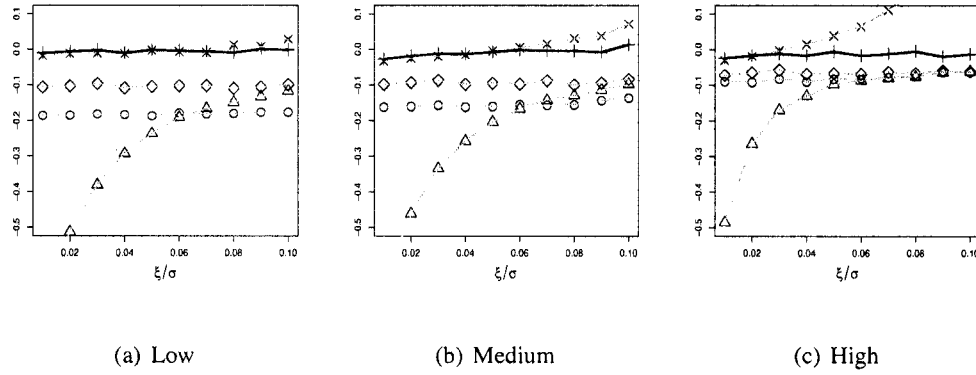


Figure 4.6 Bias of realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. The x-axis is the noise-to-signal ratio and the y-axis is the bias of the estimators. The optimized tick-time estimator, represented by the black line, is the least biased in all three frequency settings.

Figure 4.9 presents the bias of the covariance estimators of a bivariate Brownian motion with jumps under the three different arrival regimes. Figures 4.10 and 4.11 show the associated mean squared error (MSE) plots. For low noise-to-signal settings, the tick-time estimators display small biases relative to the other realized estimators. The results suggest that in the absence of noise tick-time estimation is robust to jumps. Barndorff-Nielsen and Shephard (2006) state that jump identification requires frequent observation of the process, and the low noise-to-signal setting will suggest the most frequent sampling. As the noise-to-signal ratio increases, our optimal sampling frequency increases, hence we are less able

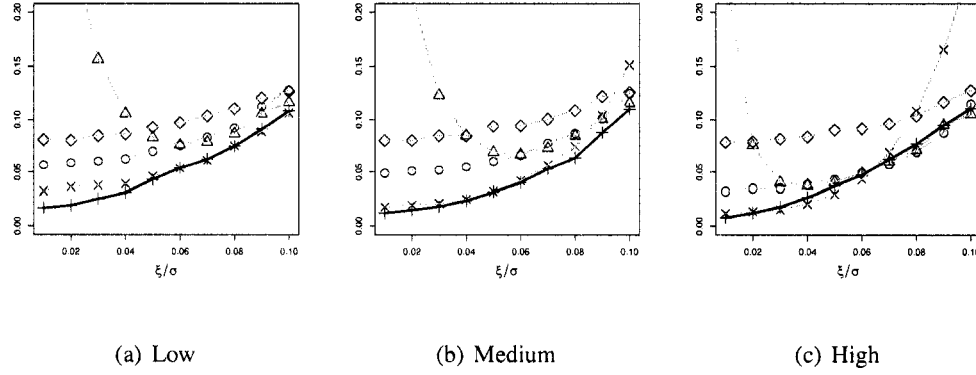


Figure 4.7 Mean Squared Error of realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. Tick-time estimators have the smallest MSE, but the difference relative to the calendar-time estimators is less pronounced as the noise-to-signal ratio increases.

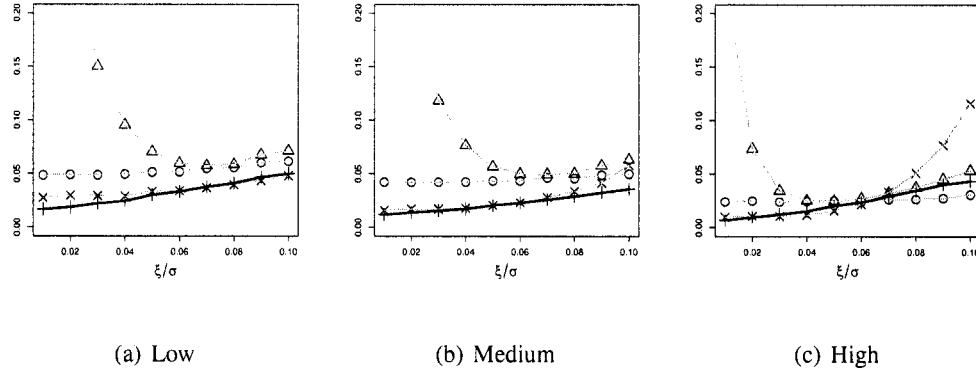


Figure 4.8 Mean Squared Error of sub-sampled realized covariance estimators under the Dynamic Arrival Rate, where $\rho = 0.9$ and symbols are defined in Table 4.1. The sub-sampled estimators have much smaller MSE than in Figure 4.7.

to isolate the jump component and obtain contaminated realized covariance estimates. We see that for all signal-to-noise levels the optimal tick-time estimator is less biased or equal to the ad-hoc tick-time method. The calendar-time estimators are biased for low noise-to-signal ratios and suggest that calendar time realized covariance estimators are not robust to jumps. These results are consistent with Barndorff-Nielsen and Shephard (2004b).

In Figures 4.10 and 4.11 we see that in the low and medium frequency settings the tick-

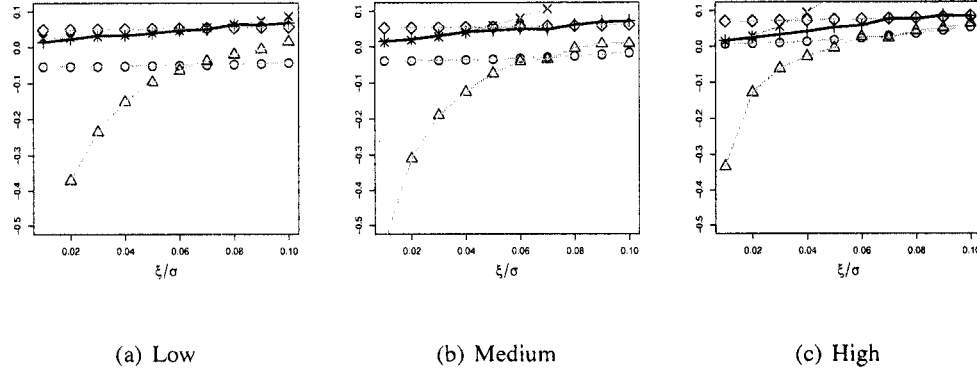


Figure 4.9 Bias of realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. The x-axis is the noise-to-signal ratio and the y-axis is the bias of the estimators. The bias of the optimized tick-time estimator, represented by the black line, increases as the noise-to-signal ratio increases.

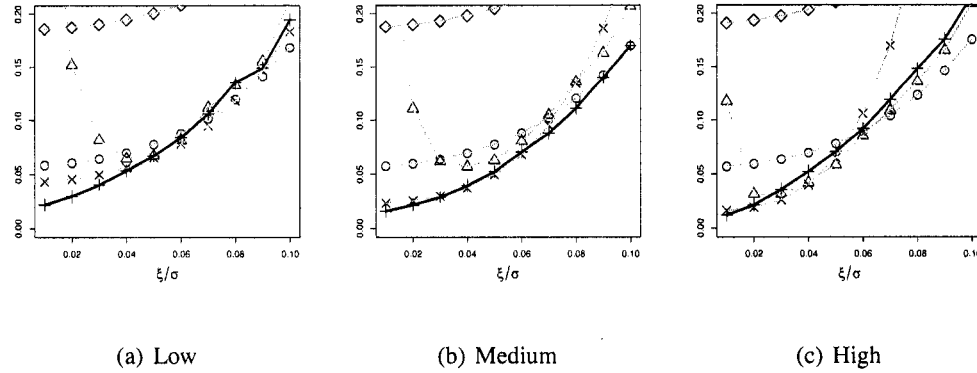


Figure 4.10 Mean Squared Error of realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. Tick-time estimators have the smallest MSE for low noise-to-signal ratios, but are no longer the smallest as the noise-to-signal ratio increases.

time estimators have the smallest MSE for low noise-to-signal ratios. The CT15 estimator performs very poorly in all settings. The optimally sampled calendar-time estimator is non-competitive for low noise, slow frequency due to the large negative bias. The CT5 estimator appears relatively competitive for noise-to-signal ratios $\geq .5$. Overall, the MSE is much higher than for the bivariate Brownian motion presented in Figures 2.5 and 2.6 and the dynamic arrival rate presented in Figures 4.7 and 4.8. Again sub-sampling reduces

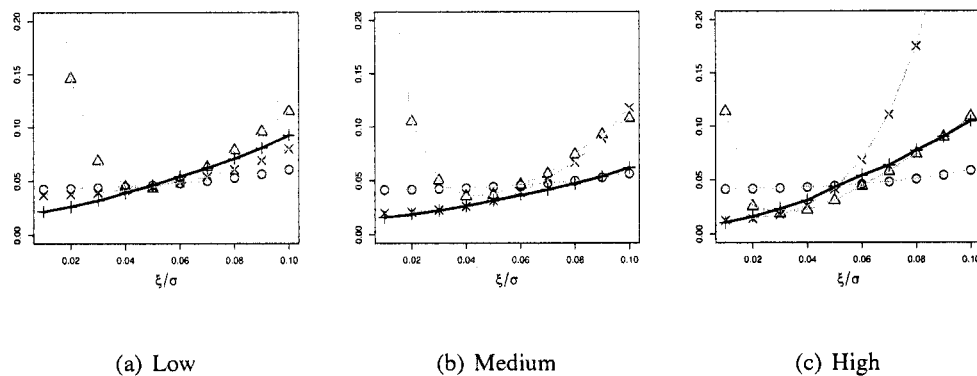


Figure 4.11 Mean Squared Error of sub-sampled realized covariance estimators with Jumps, where $\rho = 0.9$ and symbols are defined in Table 4.1. The sub-sampled estimators have much smaller MSE than in Figure 4.10.

the MSE with greatest benefits in high noise-to-signal settings as shown in Figure 4.11.

Chapter 5

Conclusion

This thesis presents a realized tick-time covariance estimator that incorporates cross-market tick-matching and intelligent sub-sampling. Results show that our estimator has smaller mean squared error, smaller bias, and greater economic utility than prevailing methodologies. Assessing the economic value of increasingly precise covariance estimates is of great interest in finance. We compare the performance of this estimator with prevailing methodologies in a simulation study and by assessing out-of-sample volatility-timing portfolio optimization strategies. For high-dimensional allocation problems we address the problem of ill-conditioned covariance matrices by considering the performance in the settings of rolled regression and factor models. We conclude that factor models are a more natural setting for employing realized covariance estimators.

This thesis has demonstrated that tick-time estimators can play an important role in realized covariance estimation. In Chapter 2 we showed the tick-time covariance estimators provide smaller mean squared error in the presence of less frequently quoted processes. In Chapter 3 we show the role that tick-time covariance estimators can play in high dimensional covariance estimation. Specifically, when a factor model structure is imposed on the covariance matrix, the optimally sampled tick-time estimator provides a risk averse investor with the better than or equal to level of utility as any exponentially smoothed full realized covariance matrix.

This work should be understood as an empirical validation of a vast body of literature. Many studies have provided simulation results, or at best low dimensional empirical anal-

ysis. This work is distinguished in that it assesses a computationally challenging problem. Moreover, this study considers the practical challenges of employing high frequency data in a high dimensional setting. It demonstrates that factor modeling dramatically reduces the dimensionality of the problem and allows for efficient estimation of high dimensional covariance matrices.

The framework for this estimator is sufficiently general to be applicable in a larger spectrum of problems. Essentially any estimation of a multivariate diffusion process, whose observations are asynchronous and contaminated by error, can benefit from this proposed methodology. Such fields include: stochastic kinetic biochemical modeling (Golightly and Wilkinson 2006), dynamics of mechanical devices (Cao and Pope 2003), and real estate valuation (Gelfand, Banerjee, Sirmans, Tu, and Ong 2007), to list a few.

5.1 Future Work

The results presented in Chapters 2 through 4 uncovered many interesting avenues for future research. The data analysis in Chapter 2 motivates further exploration into generalized market microstructure characterizations. Likewise, the encouraging single-factor model results in Chapter 3 warrant further investigation into the practical use of this methodology. The following section outlines four avenues for future research.

5.1.1 Tracking Error Minimization

The strong performance of the factor model in the multivariate setting motivates further exploiting the computational efficiency of this model. The next step is to consider a portfolio which minimizes tracking error. The objective would be to develop portfolios with minimized tracking error while incurring minimal trading costs by holding relatively

few assets. This would be of great interest for the finance community where many fund managers are assessed according to their ability to track indices.

5.1.2 Time-of-day Covariance Estimation

The U-shaped pattern of financial quote activity with respect to time of day is well known. Currently, realized variance and covariance estimators have focused on obtaining estimates of daily variation. Given the fact that arrival rates are dynamic during the day, we should also consider how variance and covariance change during the day. A hierarchical Bayesian framework may allow for more dynamic realized covariance estimation. A possible application could be to help determine optimal timing for trade execution. Hasbrouck (2004) has examined within day price characteristics in a Markov Chain Monte Carlo setting.

5.1.3 Generalized Market Microstructure Noise

Generalizing the market microstructure noise assumptions to allow for autocorrelation is a natural extension to this work. This idea is supported by Hansen and Lunde (2006), which provides empirical results suggesting that market microstructure noise displays autocorrelation and is correlated with the latent price process. The autocorrelation assumption would result in different optimal sampling frequencies for both calendar and tick time estimators. It is of great interest to compare the estimation performance under the two different noise assumptions.

5.1.4 Covariance Estimation in the Presence of Jumps

The simulation results in Chapter 4 show that tick-time realized covariance estimation is robust to jumps at very low noise-to-signal settings. As the noise-to-signal ratio

increases, tick-time estimators are no longer robust to jumps. It would be interesting to compare the performance of bipower variation in the calendar-time against tick-tick time setting. Oomen (2006) has shown promising results for tick-time estimators in the univariate setting.

Appendix A

Filtering

In February 2001 the NYSE completed a phased transition from fractional to decimal pricing. This led to a reduction in market makers' rents and thereby changed the nature of price discovery. Traditionally, regional exchanges competed with the NYSE by offering competitive quotes, cheaper executions, and anonymity. Now quotes posted in the NYSE have become more genuine price discovery signals as under decimalization the NYSE is now more often alone at the National Best Bid and Offer (NBBO) Goldstein et al. (2008).

A.1 Quotes

We limit the data to quotes posted on the NYSE due to this exchange's dominance in setting the NBBO Goldstein et al. (2008). We define admissible quotes according to the following filtering criteria (See, e.g. Yan (2007), Hansen and Lunde (2006) Chordia et al. (2005), Cliff et al. (2007)):

1. Remove quotes with mode of (4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28)
2. Remove quotes with negative prices and/or negative/zero volumes
3. Remove bid ask pairs where the bid ask spread is more than 10% of the bid.
4. Remove midquote prices that are more than 10% change from prevailing price.
5. Remove midquote prices that are the same as the prevailing price (i.e. depth revisions)

This final criteria addresses depth quotes. Note, for stocks listed on the NASDAQ, we limit our data to quotes posted on the NASDAQ.

A.2 Trades

We define admissible trades according to the following filtering criteria (See, e.g. Yan (2007), Chordia et al. (2005), Cliff et al. (2007)):

1. Remove trades with condition of ("O", "Z", "B", "T", "L", "G", "W", "J", "K")
2. Remove trades with negative prices and/or negative/zero volumes
3. Remove trades with correction not in (0, 1, 2).
4. Remove trade reversals.

A.3 Quote Trade Matching

Filtering based only on quotes may at times admit unreasonable quote values. By considering quote relative to trades, the Figure A.1 motivates the need for quote trade matching. We can see that the opening bid ask quotes on March 12 are 160.4 and 160.48 respec-

DATE	PRICE	qtime	bid	ofr	ex
20020311	104.9	15:59:41	104.9	104.92	N
20020311	104.92	15:59:42	104.9	104.92	N
20020311	104.98	15:59:49	104.9	104.98	N
20020311	104.95	15:59:49	104.6	104.98	N
20020311	105.01	16:00:03	104.9	105.01	N
20020311	105.01	16:00:03	104.9	105.01	N
20020311	105.1	16:00:03	104.9	105.01	N
20020311	105.24	16:00:36	105	105.24	N
20020311	105.24	16:00:46	105	105.25	N
20020311	105.24	16:00:46	105	105.25	N
20020311	105.24	16:01:25	105.1	105.25	N
20020312	106.45	9:33:51	160.4	160.48	N
20020312	106.45	9:33:51	160.4	160.48	N
20020312	106.5	9:34:34	160.4	160.5	N
20020312	106.5	9:34:47	160.4	160.55	N
20020312	106.5	9:34:47	160.4	160.55	N
20020312	106.5	9:35:24	106.5	160.6	N
20020312	106.5	9:35:49	106.5	160.6	N
20020312	106.5	9:35:49	106.5	160.6	N
20020312	106.6	9:36:17	106.5	160.6	N
20020312	106.6	9:36:40	106.6	160.7	N
20020312	106.6	9:37:01	106.6	160.7	N
20020312	106.6	9:37:23	106.6	106.8	N
20020312	106.6	9:37:33	106.6	106.8	N
20020312	106.8	9:37:42	106.6	106.8	N

Figure A.1 Example of limitation of quotes only filtering and how trade quote matching mitigates problem. Data is IBM on March 11-12, 2002.

tively. The previous quotes were 105.1 and 105.25. The corresponding opening trade was

at 106.45. Hence the trade suggests that the opening quotes are invalid. This suggests a need for using a filtering criteria that is dependent on trade prices. We employ the quote trade matching developed in Lee and Ready (1991) and Henker and Wang (2006). To address the problem identified in Figure A.1 we impose an additional criteria of requiring the trade to be within a spread above the offer and a spread below the bid as illustrated in Figure A.2.

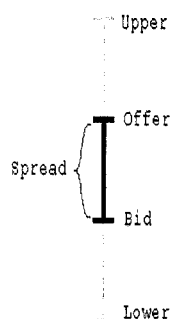


Figure A.2 Criteria for Filtering Quotes with respect to Trade

We define admissible quotes according to the following filtering criteria:

1. Remove quotes with mode of (4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28)
2. Remove quotes with negative prices and/or negative/zero volumes
3. Remove bid ask quote pairs where the bid ask spread is more than 10% of the bid.
4. Remove trades with condition of ("O", "Z", "B", "T", "L", "G", "W", "J", "K")
5. Remove trades with negative prices and/or negative/zero volumes
6. Remove trades with correction not in (0, 1, 2).
7. Quote time = Quote time + 1

8. Match trades to quotes
9. Keep if $\text{Trade} \in (\text{bid} - \text{spread}, \text{ofr} + \text{spread})$
10. Remove midquote prices that are the same as the prevailing price (i.e. depth revisions)

Bibliography

- Ait-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*, 351–416.
- Andersen, T., Bollerslev, T., and Diebold, F. (2005), "Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility," Working Paper University of Pennsylvania.
- Andersen, T., Bollerslev, T., Diebold, F., and Labys, P. (2001), "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.
- (2003), "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.
- Andreou, E. and Ghysels, E. (2002), "Rolling-Sample Volatility Estimators: Some New Theoretical, Simulation, and Empirical Results," *Journal of Business and Economic Statistics*, 363–376.
- Bandi, F. and Russell, J. (2005), "Realized Covariation, Realized Beta, and Microstructure Noise," Working Paper University of Chicago.
- (2006), "Separating Microstructure Noise from Volatility," *Journal of Financial Economics*, 79, 655–692.
- Bandi, F., Russell, J., and Zhu, Y. (2008), "Using High-Frequency Data in Dynamic Portfolio Choice," *Econometric Reviews*, forthcoming.
- Barndorff-Nielsen, O. and Shephard, N. (2002), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, Ser. B.*, 64, 253–280.
- (2004a), "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72, 885–925.
- (2004b), "Power and Bipower Variation with Stochastic Volatility and Jumps," *Journal of Financial Econometrics*, 2, 1–37.
- (2006), "Econometrics of Testing for Jumps in Financial Econometrics Using Bipower Variation," *Journal of Financial Econometrics*, 4, 1–30.
- Bauer, G. H. and Vorkink, K. (2007), "Multivariate Realized Stock Market Volatility," Working Papers 07-20, Bank of Canada.
- Bickel, P. J. and Levina, E. (2008), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36, 199–227.

- Bollerslev, T. and Zhang, B. Y. B. (2003), "Measuring and modeling systematic risk in factor pricing models using high-frequency data," *Journal of Empirical Finance*, 10, 533–558.
- Britten-Jones, M. (1999), "The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights," *Journal of Finance*, 54, 655–671.
- Cao, R. and Pope, S. B. (2003), "Numerical integration of stochastic differential equations: weak second-order mid-point scheme for application in the composition PDF method," *Journal of Computational Physics*, 185, 194–212.
- Chamberlain, G. and Rothschild, M. (1983), "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304.
- Chan, L., Karceski, J., and Lakonishok, J. (1999), "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model," *The Review of Financial Studies*, 12, 937–974.
- Chopra, V. and Ziemba, W. (1993), "The effect of error in means, variances and covariances on optimal portfolio choice," *Journal of Portfolio Management*, 19, 6–11.
- Chordia, T., Sarkar, A., and Subrahmanyam, A. (2005), "An Empirical Analysis of Stock and Bond Market Liquidity," *The Review of Financial Studies*, 18, 85–129.
- Cliff, M., Cooper, M., and Gulen, H. (2007), "Return Differences between Trading and Non-trading Hours: Like Night and Day," Working Paper The University of Utah.
- Corsi, F. (2006), "Realized Correlation Tick-by-Tick," Working Paper University of Lugano.
- Dacorogna, M., Gencay, R., Muller, U., Olsen, R., and Pictet, O. (2001), *An Introduction to High-Frequency Finance*, San Diego, CA: Academic Press.
- de Jong, F. and Nijman, T. (1997), "High Frequency Analysis of lead-lag relationships between financial markets," *Journal of Empirical Finance*, 4, 259–277.
- de Pooter, M., Martens, M., and van Dijk, D. (2006), "Predicting the Daily Covariance Matrix of S&P100 Stocks Using Intraday Data - but which frequency to use?" *Econometric Reviews*, forthcoming.
- Engle, R. and Colacito, R. (2006), "Testing and Valuing Dynamic Correlations for Asset Allocation," *Journal of Business and Economic Statistics*, 238–253.
- Engle, R. F. and Russell, J. R. (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.
- Epps, T. (1979), "Comovement in Stock Prices in the Very Short Run," *Journal of the American Statistical Association*, 291–298.

- Fan, J. (2005), "A Selective Overview of Nonparametric Methods in Financial Econometrics," *Statistical Science*, 20, 317–337.
- Fan, J., Fan, Y., and Lv, J. (2007), "High Dimensional Covariance Matrix Estimation Using a Factor Model," Tech. rep., Princeton University.
- Fleming, J., Kirby, C., and Ostdiek, B. (2001), "The Economic Value of Volatility Timing," *Journal of Finance*, 329–352.
- (2003), "The Economic Value of Volatility Timing Using "Realized" Volatility," *Journal of Financial Economics*, 473–509.
- Foster, D. and Nelson, D. (1996), "Continuous record asymptotics for rolling sample variance estimators," *Econometrica*, 64, 139–174.
- Garcia, R. and Meddahi, N. (2006), "Comment on 'Realized Variance and Market Microstructure Noise'," *Journal of Business and Economic Statistics*, 24, 184–192.
- Gelfand, A. E., Banerjee, S., Sirmans, C., Tu, Y., and Ong, S. E. (2007), "Multilevel modeling using spatial processes: Application to the Singapore housing market," *Computational Statistics & Data Analysis*, 51, 3567–3579.
- Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006), "Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies," *Journal of Econometrics*, 59–95.
- Glasserman, P. (2004), *Monte Carlo Methods in Financial Engineering*, New York: Springer.
- Goldstein, M., Shkilko, A., Ness, B. V., and Ness, R. V. (2008), "Inter-Market Competition for NYSE-listed Securities under Decimals," *Journal of Financial Markets*, forthcoming.
- Golightly, A. and Wilkinson, D. J. (2006), "Bayesian Sequential Inference for Stochastic Kinetic Biochemical Network Models," *Journal of Computational Biology*, 13, 838–851.
- Griffin, J. and Oomen, R. (2006), "Covariance Measurement in the Presence of Non-synchronous Trading and Market Microstructure Noise," Working Paper University of Warwick.
- (2008), "Sampling Returns for Realized Variance Calculations: Tick Time or Transaction Time," *Econometric Reviews*, forthcoming.
- Han, Y. (2006), "Asset Allocation with a High Dimensional Latent Factor Stochastic Volatility Model," *The Review of Financial Studies*, 19, 237–271.
- Hansen, P. R., Large, J., and Lunde, A. (2006), "Moving Average-Based Estimators of Integrated Variance," Working Paper of Nuffield College.
- Hansen, P. R. and Lunde, A. (2005), "A Realized Variance for the Whole Day Based on Intermittent High-Frequency Data," *Journal of Financial Econometrics*, 3, 525–554.

- (2006), “Realized Variance and Market Microstructure Noise,” *Journal of Business and Economic Statistics*, 24, 127–161.
- Hasbrouck, J. (2004), “Liquidity in the Futures Pit: Inferring Market dynamics from incomplete data,” *Journal of Financial and Quantitative Analysis*, 39, 305–326.
- (2007), *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, New York: Oxford University Press.
- Hayashi, T. and Yoshida, N. (2005), “On Covariance Eestimation of Non-synchronously Observed Diffusion Processes,” *Bernoulli*, 11, 359–379.
- Henker, T. and Wang, J.-X. (2006), “On the Importance of Timing Specifications in Market Microstructure Research,” *Journal of Financial Markets*, 9, 162–179.
- Higham, D. (2001), “An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations,” *SIAM Review*, 43, 525–546.
- Jagannathan, R. and Ma, T. (2003), “Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps,” *Journal of Finance*, 58, 1651–1683.
- Jorion, P. (1992), “Portfolio Optimization in Practice,” *Financial Analysts Journal*, 48, 68–74.
- (2003), “Portfolio Optimization with Tracking-Error Constraints,” *Financial Analysts Journal*, 59, 70–82.
- Kawakatsu, H. (2006), “Matrix Exponential GARCH,” *Journal of Econometrics*, 134, 95–128.
- Kyj, L., Ensor, K., and Ostdiek, B. (2008), “Economic Value of Optimally Sub-sampled Realized Covaraince of Asynchronous and Noisy High-Frequency Data,” Technical Report TR2008-01, Rice University, Statistics Department.
- Ledoit, O. and Wolf, M. (2003), “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection,” *Journal of Empirical Finance*, 10, 603–621.
- (2004), “A well-conditioned estimator for large-dimensional covariance matrices,” *Journal of Multivariate Analysis*, 88, 365–411.
- Lee, C. and Ready, M. (1991), “Inferring Trade Direction from Intraday Data,” *Journal of Finance*, 46, 733–747.
- Madhavan, A. (2000), “Market Microstructure: A Survey,” *Journal of Financial Markets*, 3, 205–258.
- Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

- Michaud, R. O. (1989), "The Markowitz Optimization Enigma: Is Optimized Optimal," *Financial Analysts Journal*, 45, 31–42.
- O'Hara, M. (1995), *Market Microstructure Theory*, Blackwell.
- Oomen, R. (2006), "Properties of Realized Variance Under Alternative Sampling Schemes," *Journal of Business and Economic Statistics*, 24, 219–237.
- Palandri, A. (2006), "Consistent Realized Covariance for Asynchronous Observations Contaminated by Market Microstructure Noise," Working Paper University of Copenhagen.
- Politis, D. (2003), "The Impact of Bootstrap Methods on Time Series Analysis," *Statistical Science*, 18, 219–230.
- Rendò, R. (2003), "A Closer Look at the Epps Effect," *International Journal of Theoretical and Applied Finance*, 6, 87–102.
- Sancetta, A. (2008), "Sample Covariance Shrinkage for High Dimensional Dependent Data," *Journal of Multivariate Analysis*, 99, 949–967.
- Sharpe, W. (1963), "A Simplified Model for Portfolio Analysis," *Management Science*, 9, 277–293.
- Sorensen, H. (2004), "Parametric Inference for Diffusion Processes Observed at Discrete Points in Time: A Survey," *International Statistical Review*, 72, 337–354.
- Tse, Y. and Tsui, A. (2002), "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations," *Journal of Business & Economics Statistics*, 20, 351–362.
- Voev, V. and Lunde, A. (2007), "Integrated Covariance Estimation Using High-Frequency Data in the Presence of Noise," *Journal of Financial Econometrics*, 68–104.
- Wood, R. (2000), "Market Microstructure Research Databases: History and Projections," *Journal of Business and Economic Statistics*, 18, 140–145.
- Yan, Y. (2007), "Introduction to TAQ," presentation, 2007 WRDS Users Conference.
- Zhang, L. (2006), "Estimating Covariation: Epps Effect, Microstructure Noise," Working Paper University of Illinois at Chicago.
- Zhang, L., Mykland, P., and Ait-Sahalia, Y. (2005), "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 1394–1411.
- Zhou, B. (1996), "High-Frequency Data and Volatility in Foreign-Exchange Rates," *Journal of Business and Economic Statistics*, 45–52.