Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università degli Studi di Messina
GRINS

2nd Workshop on Sustainable Finance Spoke 4 - GRINS
Università Ca' Foscari Venezia
2-3 December 2024

Building a realized volatility database: some preliminary results

A. Insana

ForVARD - Forecasting Volatility And Risk Dynamics
University of Messina

Outline

ForVARD - University of Messina - Department of Economics

Project coordinator: E. Otranto
Research staff: M. B. Donato, M. Milasi, F. Spagnolo
Junior researchers: G. Cruciani, A. Insana
Consultant: F. Cipollini (University of Firenze - DiSIA)
Special guest: G. M. Gallo

Research purpose

- Construct an openly accessible dataset containing realized volatility and covariance measures.
- Employ classical and modern models to forecast volatility and covariance matrices.
- Offer methodologies for building sustainable portfolios.

ForVARD - Forecasting Volatility And Risk Dynamics

- **Task 1.1:** Selection of realized volatility measures.
  Identifying the tick-by-tick dataset to acquire for subsequent processing.
- **Task 1.2:** Selection of univariate and multivariate models.
- **Task 1.3:** Selection of approaches for sustainable portfolio management.

- **Task 2.1:** Collection and cleaning of the data.
- **Task 2.2:** Data aggregation and quality check. Derivation of realized measures.
- **Task 2.3:** Realized library and updating.

ForVARD - Forecasting Volatility And Risk Dynamics

- **Task 1.1:** Selection of realized volatility measures.
  Identifying the tick-by-tick dataset to acquire for subsequent processing.
- **Task 1.2:** Selection of univariate and multivariate models.
- **Task 1.3:** Selection of approaches for sustainable portfolio management.

- **Task 2.1:** Collection and cleaning of the data.
- **Task 2.2:** Data aggregation and quality check. Derivation of realized measures.
- **Task 2.3:** Realized library and updating.

Forecasting Volatility And Risk Dynamics

Understanding and accurately modelling the dynamics of asset return volatility has significant implications in:

- **Asset Allocation**
- **Risk Management**
- **Derivative Pricing**

By modelling these dynamics, we can make more informed decisions, optimise our strategies and ultimately drive success in the financial markets.

High-Frequency Data on Realized Volatility

- Huge amounts of data and data errors
- Microstructure noise:
  - Asynchronous nature of tick data
  - Bid-ask bounce
  - Infrequent trading
  - Price discreteness
- Data aggregation and quality check:
  - Building series at regular intraday intervals (i.e. 1-minute or 5-minute) and aggregating traded volumes for further processing.
  - Perform the aggregation operation to construct the daily time series of realized variances and covariances.
  - Perform regular operations of data quality control from the most granular level to the daily series.

Kibot dataset

Data package:

- Intraday data and up to 60 years of daily data for all available US stocks and ETFs, Futures, Forex, Indexes and OTC BB stocks.
- Over 15 years (since 2009) of tick with bid/ask and one second data.
- Over 26 years (since 1998) of one minute adjusted and unadjusted.

They provide different files for data:

- **tickbidask**: tick files (with bid/ask prices)
- **tick**: tick files (without bid/ask)
- **tick_sec**: 1, 5 or 10 second files
- **minute, #**: 1, 3, 5, ...
- **bidask_min**: 1, 5, 15, 30 min.

Download Data

- **Input.**

```
start_date = "MM/DD/YYYY"
end_date = "MM/DD/YYYY"
start_time_str = '09:30:00.000'
end_time_str = '16:00:00.000'
symbols = ['MSFT', 'AAPL', 'GOOGL', 'KO', 'FLR']
should_process_data = True
```

(for the aggregation on the same timestamps data)

- **Output.**
  A folder with the name of the `ticker` containing:
  - the daily files `YYYY_MM_DD.txt` with `Time, Price, Volume, Trades`.
  - A summary of the overnight values is stored `overnight.txt`
  - The adjustment information `adjustment.txt` relative to the selected period.
  - A `.log` file containing download information.

Downloading times and storage

We consider a list of 5 stocks from 01/01/2023 to 01/01/2024.

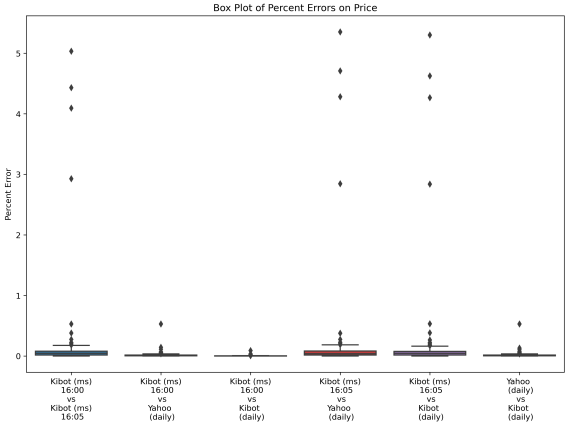
	All data	Aggregated
AAPL	4.07 GB	2.31 GB
MSFT	2.71 GB	1.53 GB
GOOGL	2.32 GB	1.19 GB
KO	879 MB	513 MB
FLR	126 MB	64 MB
Total storage	10.105 GB	5.607 GB
Computational Time[1]	2.49 hr	2.59 hr

[1] PC: Surface Pro 7, Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz, RAM 16.0 GB.

Kibot Data

Price analysis on MSFT series (1y)

	Kibot (ms) 16:00	Kibot (ms) 16:05	Kibot (daily)	Yahoo (daily)
count	250	250	250	250
mean	311.04	311.16	311.04	311.03
median	321.71	321.71	321.73	321.72
std	41.28	41.19	41.28	41.29
min	219.10	219.10	219.10	219.16
25%	279.75	280.07	279.74	279.68
50%	321.71	321.71	321.73	321.72
75%	334.81	334.82	334.80	334.79
max	380.62	380.57	380.62	380.62

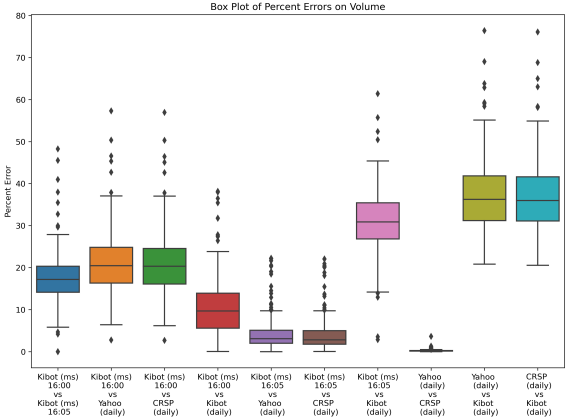


Kibot Data

Volume analysis on MSFT series (1y)

	Kibot (ms) 16:00	Kibot (ms) 16:05	Kibot (daily)	Yahoo (daily)	CRSP (daily)
count	250	250	250	250	250
mean	2.18e+07	2.64e+07	2.04e+07	2.77e+07	2.76e+07
median	1.94e+07	2.41e+07	1.84e+07	2.51e+07	2.50e+07
std	8.66e+06	9.90e+06	8.32e+06	1.06e+07	1.06e+07
min	9.22e+06	9.22e+06	6.81e+06	1.02e+07	1.01e+07
25%	1.63e+07	2.01e+07	1.50e+07	2.12e+07	2.12e+07
50%	1.94e+07	2.41e+07	1.84e+07	2.51e+07	2.50e+07
75%	2.40e+07	2.86e+07	2.22e+07	2.99e+07	2.99e+07
max	6.23e+07	6.71e+07	5.57e+07	7.85e+07	7.78e+07

The Kibot dataset does not include unfiltered ”odd lot” transactions or transactions smaller than 100 shares.



Outlier detection

The sheer volume of data and the presence of irregular time intervals make high frequency data prone to errors and outliers.

Sources of outliers:

- human input errors
- system glitches (bugs in the data feed)
- market anomalies
- delayed trade reporting on block trades

Identifying Outliers:

- Exchange-provided information (e.g., trade conditions)
- Data-driven techniques [Brownlees and Gallo, 2006, Barndorff-Nielsen et al., 2009]
  - Rely on ad hoc tuning parameters
  - Risk of over- or under-cleaning the data

Outlier detection

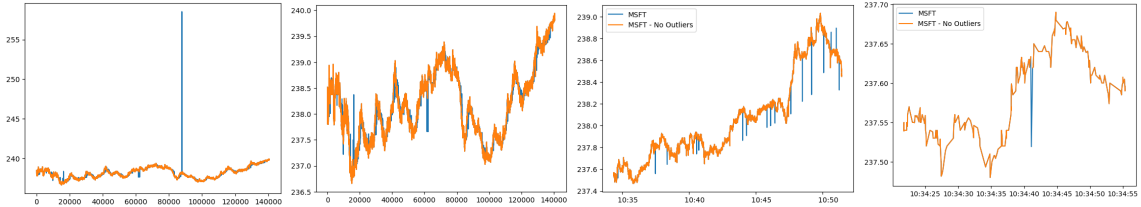
For the outlier detection we use the procedure of [Brownlees and Gallo, 2006]. They remove outlier price p_i if

$$|p_i - \bar{p}_i(k)| \geq 3\, s_i(k) + \gamma$$

- K are the neighborhood, the observations around i.
- $\bar{p}_i(k)$ is the δ-trimmed sample mean of a neighborhood k observations around i.
- $s_i(k)$ is the δ-trimmed sample standard deviation of the same neighborhood.
- γ is a granularity parameter to avoid zero variances from sequences of equal prices.

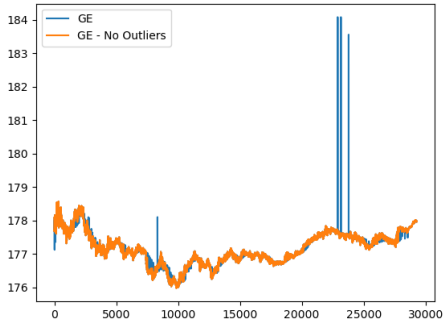
Outlier detection

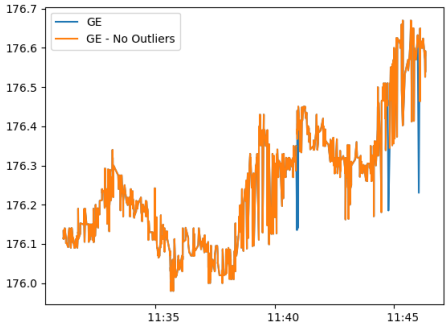
data: Kibot tickms
ticker: MSFT
date: December 30, 2022
 # observations: 140,653
Parameter choices: $k = 60,\ \delta = 0.1,\ \gamma = 0.06 \rightarrow$ 434 outliers



Outlier detection

data: Kibot tickms
ticker: GE
date: November 20, 2024
 # observations: 29,269
Parameter choices: $k = 60,\ \delta = 0.1,\ \gamma = 0.06 \rightarrow$ 181 outliers





Adaptive outlier detection

Problem: Find a methodology that suits different types of data (stocks, indices, ETFs, ...) with different levels of trading activity.

- Some rules for choosing the parameters k and γ in the [Brownlees and Gallo, 2006] approach.
- Adapting the methods of [Brownlees and Gallo, 2006] to price changes.
- Consider a new cleaning approach to price changes.

Future Steps

- Understand whether the presence of outliers affects the realized volatility measure.
- Understand the impact of official market closures on realized volatility measures.
- Implement different methods for realized volatility (and realized covariances).
- Create an open database of volatility measures.

Appendix and References

Forecasting Volatility And Risk Dynamics Models

Parametric Volatility Models

- Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [Bollerslev, 1986]
- GJR-GARCH [Glosten et al., 1993]

Realized Volatility Models

- Realized Volatility (RV) (High Frequency data) [Andersen and Bollerslev, 1998], [Andersen et al., 2003]
- Bipower variation (BV) [Barndorff-Nielsen and Shephard, 2004]
- Realized kernel volatility [Barndorff-Nielsen et al., 2008]
- Heterogeneous AutoRegressive (HAR) [Corsi, 2009]
- HAR-GARCH [Corsi et al., 2008]
- Multiplicative Error Model (MEM) [Engle and Gallo, 2006]

Forecasting Volatility And Risk Dynamics Models

Early Covariance Models

- MGARCH
  - VECH [Bollerslev et al., 1988]
  - DCC [Engle, 2002]
  - BEKK [Engle and Kroner, 1995]
- Conditional Correlation Models [Engle, 2002]. To model R:
  - Constant Conditional Correlation (CCC) [Bollerslev, 1990]
  - A time-varying Conditional Correlation (TVCC) [Tse and Tsui, 2002]
  - DCC [Engle, 2002]

Models for Realized Covariance Matrices

- Realized quadratic covariation [Andersen et al., 2011]
- Realized BiPower Covariation [Barndorff-Nielsen and Shephard, 2004]
- VECH-HAR [Chiriac and Voev, 2011]
- DRD (DCC+HAR) [Oh and Patton, 2016]
- The Conditional Autoregressive Wishart (CAW) Models [Golosnoy et al., 2012]

References I

- Andersen, T. G. and Bollerslev, T. (1998).
  Answering the skeptics: Yes, standard volatility models do provide accurate forecasts.
  International Economic Review, 39(4):885–905.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003).
  Modeling and forecasting realized volatility.
  Econometrica, 71(2):579–625.
- Andersen, T. G., Bollerslev, T., and Meddahi, N. (2011).
  Realized volatility forecasting and market microstructure noise.
  Journal of Econometrics, 160(1):220–234.
  Realized Volatility.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008).
  Designing realized kernels to measure the ex post variation of equity prices in the presence of noise.
  Econometrica, 76(6):1481–1536.

References II

- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009).
  Realized kernels in practice: trades and quotes.
  The Econometrics Journal, 12(3):C1–C32.
- Barndorff-Nielsen, O. E. and Shephard, N. (2004).
  Power and Bipower Variation with Stochastic Volatility and Jumps.
  Journal of Financial Econometrics, 2(1):1–37.
- Bollerslev, T. (1986).
  Generalized autoregressive conditional heteroskedasticity.
  Journal of Econometrics, 31(3):307–327.
- Bollerslev, T. (1990).
  Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model.
  The review of economics and statistics, pages 498–505.

References III

- Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988).
  A capital asset pricing model with time-varying covariances.
  Journal of political Economy, 96(1):116–131.
- Brownlees, C. and Gallo, G. (2006).
  Financial econometric analysis at ultra-high frequency: Data handling concerns.
  Computational Statistics Data Analysis, 51(4):2232–2245.
  Nonlinear Modelling and Financial Econometrics.
- Chiriac, R. and Voev, V. (2011).
  Modelling and forecasting multivariate realized volatility.
  Journal of Applied Econometrics, 26(6):922–947.
- Corsi, F. (2009).
  A simple approximate long-memory model of realized volatility.
  Journal of Financial Econometrics, 7(2):174–196.

References IV

- Corsi, F., Mittnik, S., Pigorsch, C., and Pigorsch, U. (2008).
  The volatility of realized volatility.
  Econometric Reviews, 27(1-3):46–78.
- Engle, R. (2002).
  Dynamic conditional correlation-a simple class of multivariate garch models.
  Journal of Business and Economic Statistics, 20(3):339–350.
- Engle, R. F. and Gallo, G. M. (2006).
  A multiple indicators model for volatility using intra-daily data.
  Journal of econometrics, 131(1-2):3–27.
- Engle, R. F. and Kroner, K. F. (1995).
  Multivariate simultaneous generalized arch.
  Econometric theory, 11(1):122–150.

References V

- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993).
  On the relation between the expected value and the volatility of the nominal excess return on stocks.
  The journal of finance, 48(5):1779–1801.
- Golosnoy, V., Gribisch, B., and Liesenfeld, R. (2012).
  The conditional autoregressive wishart model for multivariate stock market volatility.
  Journal of Econometrics, 167(1):211–223.
- Oh, D. H. and Patton, A. J. (2016).
  High-dimensional copula-based distributions with mixed frequency data.
  Journal of Econometrics, 193(2):349–366.
  The Econometric Analysis of Mixed Frequency Data Sampling.
- Tse, Y. K. and Tsui, A. K. C. (2002).
  A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations.
  Journal of Business & Economic Statistics, 20(3):351–362.