Computational Challenges in HF Volatility Estimation

Aleksey Zemnitskiy
aleksey.zemnitskiy@snowfallsystems.com
Startup founded in 2010

Interdisciplinary research related to HF data

All employees are involved in Research & Development

Office in New York City

Timeline of the ICE-9 HF Analytics Platform

- **2010 June**
  - Company is established

- **2011 February**
  - Project ICE-9 is initiated

- **2012 January**
  - 1st Milestone: ~70 model tests, Java-to-R interface

- **2012 February**
  - ~150 model tests, Demo of remote connectivity

- **2012 June**
  - 2nd Milestone: On-demand subscription with FAST/FIX 5.0 protocol

- **2012 July**
  - Presentation at the Stevens Institute of Technology

- **2012 October**
  - ICE-9 website redesign
Outline

- Integrated Variance
- Computational Challenges
- Historical Data Analysis
Part 1

Integrated Variance
Consider this popular model for the price of a financial asset:

$$dX_t = \mu_t X_t dt + \sigma_t X_t dW_t$$

where $\mu_t$ is a drift process, $\sigma_t$ is a volatility process and $W_t$ is a Std. BM. We assume that $X_t$ is observed at time points $t_i = i/N$ with $i = 0,\ldots,N$.

The object of interest is the integrated variance (IV), i.e. the amount of variation at time point $t$ accumulated over a past time interval $T$.

$$IV = \int_0^T \sigma_s^2 ds$$

Notice, that as any cumulative measure defined on a set of non-negative values, IV is more affected in percentage terms by earlier observations.
With the increase in sampling frequency market microstructure frictions cannot be ignored anymore.

For most practical purposes microstructure noise contamination could be often modeled as additive

\[ Y_t = X_t + u_t \]

At moderately high frequencies noise could be assumed white, as the sampling frequency goes up, noise becomes more dependent on the “efficient” market price and serial correlation of the noise becomes stronger.
# Comparison of IV Estimators

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Unbiased</th>
<th>Consistent</th>
<th>Jump Robust</th>
<th>Time dependence noise</th>
<th>Endogenous noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td>RV</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Subsampling</td>
<td>TSRV</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>TSRV&lt;sub&gt;adjust&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>MSRV</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Kernel</td>
<td>KRV&lt;sub&gt;FT&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>KRV&lt;sub&gt;NFT&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Range</td>
<td>RRV</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>MRV</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>MRV&lt;sub&gt;adjust&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Quantile</td>
<td>QRV</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>QRV&lt;sub&gt;mq&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Duration</td>
<td>DRV</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
Latest research on nonparametric estimators of integrated variance is focused on endogeneous noise models that arise at ultra-high frequencies.

There is a trend to design estimators that are robust to price jumps.

We would study MRV more in the next section.
Outline

- Integrated Variance
- **Computational Challenges**
- Historical Data Analysis
Part 2

Computational Challenges
Modulated Realized Variance is a consistent estimator of integrated variance introduced in Podolskij and Vetter (2009) with convergence speed of $n^{1/4}$.

- Efficient price $X_t$ is a continuous stochastic volatility semi-martingale
- Microstructure noise is additive zero mean iid random variable
- Noise is independent of the price process realization (exogeneous noise)
- Noise variance is of order $O(1)$

**Modulated Realized Variance**

$$MRV(Y)_n = \frac{c_1 c_2 MBV(Y, 2, 0)_n - \nu_2 \hat{\omega}^2}{\nu_1} \to IV$$

$$MBV(Y, r, l)_n = n^{(r+1)/4-1/2} \sum_{m=1}^{M} |\bar{Y}_m^{(K)}| r |\bar{Y}_{m+1}^{(K)}|, \quad r, l \geq 0$$

$$Y_m^{(K)} = \frac{1}{n/M - K + 1} \sum_{i=(m-1)n/M}^{mn/M - K} (Y_{(i+K)/n} - Y_{i/n})$$
Simulation Design

Efficient price process
GBM model with no drift, Heston model in volatility

Microstructure Noise
Additive normal iid random variables

Sample size
n = 23,400 (length of a trading day in seconds at US stock markets)

Simulation Environment
Ubuntu Linux 64-bit, 8 GB RAM, 2.6 GHz Core i5
There are several ways to compute this sum. We will consider two different approaches.
Step 1: Cumulative Summation

\[ S^m_n = \sum_{i=n}^{m} f_i = S_m - S_n \]

\[ S^4_1 = \sum_{i=4}^{4} f_i = S_4 - S_1 \]

\[
MRV: \sum_{i=n}^{m} (Y_{i+K} - Y_i) = \left( \sum_{i=0}^{m} Y_{i+K} - \sum_{i=0}^{n} Y_{i+K} \right) - \left( \sum_{i=0}^{m} Y_i - \sum_{i=0}^{n} Y_i \right) = (S_m - S_n) - (S_m^* - S_n^*)
\]
Timing Results: Fast Summation

Sequential MRV computation time per part
(fast summation, no rolling window)

- Part number: 2, Time (ms): 150
- Part number: 4, Time (ms): 200
- Part number: 6, Time (ms): 250
- Part number: 8, Time (ms): 300
- Part number: 10, Time (ms): 300

Fast Summation
2413.66 ms

Sequential MRV computation time per part
(slow summation, no rolling window)

- Part number: 2, Time (ms): 50
- Part number: 4, Time (ms): 100
- Part number: 6, Time (ms): 150
- Part number: 8, Time (ms): 200
- Part number: 10, Time (ms): 250

Slow Summation
4806.97 ms
2-Dimensional Concurrency

Modulated Realized Variance

\[ MRV(Y)_n = \frac{c_1 c_2 MBV(Y, 2, 0)}{\nu_1} - \nu \hat{\omega}^2 \rightarrow IV \]

1. \[ MBV(Y, r, l)_n = n^{(r+1)/4-1/2} \sum_{m=1}^{M} | Y_m^{(K)} | r | Y_{m+1}^{(K)} |, \ r, l \geq 0 \]

\[ Y_m^{(K)} = \frac{1}{n/M-K+1} \sum_{i=(m-1)n/M}^{mn/M-K} (Y_{i+K}/n - Y_i/n) \]

2. \[ \hat{\omega}^2 = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - Y_{i-1})^2 \]

Modulated Bipower Variation

Noise Variance

These two estimators are computed independently of each other before they enter the final equation.

In a 2-Dimensional concurrency approach we compute them in parallel at each step and then compute resulting MRV value in the end.
Step 2: Concurrent Estimator Design

Single Thread: MRV Estimator

Thread 1:
- MBV
- Other

Thread 2:
- Noise Variance

Algorithm time

Benefits from Concurrency

Thread Management time

Step 2: Concurrent Estimator Design
Timing Results: 2D Concurrency

Sequential MRV computation time per part (fast summation, no rolling window)

Concurrent MRV computation time per part (fast summation, no rolling window)

Sequential
2413.66 ms

2D Concurrent
10191.52 ms
We introduce MRV estimator computed in a rolling window to assign more weight on latest price observations.

The trade-off for a rolling window estimate is a constant bias caused by a limited sample size.
Step 3: **Linear vs Circular FIFO Buffer**

- **Linear FIFO Buffer**
  - Shows a sequence of entries with a clear head and tail.

- **Circular FIFO Buffer**
  - Illustrates circular growth with a head and tail.
  - Demonstrates the concept of free space and circular growth.

Diagram comparison highlights the differences between linear and circular buffer structures.
Timing Results: Rolling Window

Sequential MRV computation time per part
(fast summation, rolling window: 5850, circular buffer)

![Circular Buffer](2331.426 ms)

Sequential MRV computation time per part
(fast summation, rolling window: 5850, array buffer)

![Array Buffer](1980.64 ms)
Outline

- Integrated Variance
- Computational Challenges
- Historical Data Analysis
Part 3

Historical Data Analysis
Impact of Reporting on Variance Estimates

- Two quarterly reports of GOOG: Q4 2008 and Q1 2009
- GOOG quote frequency - 30 sec
- Removed gaps and jumps in data
- Used MRV estimator with rolling window of 20000 obs
- Removed daily seasonality and applied log transform
- Variance estimates have two unit roots
- BIC criterion yields ARIMA(2,2,1), $R^2 = 99\%$
- To highlight reporting impact we reduce to ARIMA(1,2,1) as AR(2) coef is not always significant
- Earnings information was taken from http://investor.google.com/earnings/
Example 1: GOOG, 2009-Q1

Close Price

GOOG (freq: 30 sec, 16 April 2009 – day of quarterly report)
Example 1: GOOG, 2009-Q1

Close Price without Gaps

GOOG (with removed overnight gaps) – 1Q 2009
Example 1: GOOG, 2009-Q1

Close Price without Gaps and Jumps

GOOG (with removed overnight gaps and other jumps) – 1Q 2009
**Example 1: GOOG, 2009-Q1**

**MRV Estimator**

Mrv – 1Q 2009

![Graph showing MRV estimator for GOOG, 2009-Q1]
Example 1: GOOG, 2009-Q1

First and second order autocorrelations

ACF – 1Q 2009

Estimators
- ACF(1)
- ACF(2)

Days (8 April – 17 April)
# Model Coefficients

## ARIMA(1,2,1)

<table>
<thead>
<tr>
<th>Results</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>0.24 (0.13)</td>
<td>0.41 (0.1)</td>
<td>0.53 (0.09)</td>
<td>0.61 (0.05)</td>
<td>0.48 (0)</td>
<td>0.71 (0.04)</td>
<td>0.75 (0.03)</td>
</tr>
<tr>
<td>MA</td>
<td>0.15 (0.13)</td>
<td>0.01 (0.1)</td>
<td>-0.15 (0.1)</td>
<td>-0.13 (0.06)</td>
<td>-0.2 (0)</td>
<td>-0.24 (0.06)</td>
<td>-0.05 (0.04)</td>
</tr>
<tr>
<td>R²(%)</td>
<td>14</td>
<td>18</td>
<td>17</td>
<td>26</td>
<td>9</td>
<td>31</td>
<td>53</td>
</tr>
</tbody>
</table>
Example 2: **GOOG, 2008-Q4**

Close Price

GOOG (freq: 30 sec, 22 January 2009 – day of quarterly report)
Example 2: GOOG, 2008-Q4

Close Price without Gaps

GOOG (with removed overnight gaps) – 4Q 2008

Days (15 January – 26 January)

Close price

0 2 4 6
Example 2: GOOG, 2008-Q4

Close Price without Gaps and Jumps

GOOG (with removed overnight gaps and other jumps) – 4Q 2008
Example 2: GOOG, 2008-Q4

MRV Estimator

Mrv - 4Q 2008

Variance, $10^{-4}$

Days (15 January - 26 January)
Example 2: **GOOG, 2008-Q4**

First and second order autocorrelations

**ACF – 4Q 2008**

Estimators
- ACF(1)
- ACF(2)

Days (15 January – 26 January)
### ARIMA(1,2,1)

<table>
<thead>
<tr>
<th>Results</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AR</strong></td>
<td>-0.13 (0.06)</td>
<td>-0.15 (0.05)</td>
<td>-0.15 (0.05)</td>
<td>-0.23 (0.06)</td>
<td>-0.16 (0.06)</td>
<td>-0.35 (0.06)</td>
<td>-0.62 (0.04)</td>
</tr>
<tr>
<td><strong>MA</strong></td>
<td>-0.56 (0.05)</td>
<td>-0.63 (0.04)</td>
<td>-0.65 (0.04)</td>
<td>-0.4 (0.06)</td>
<td>-0.5 (0.05)</td>
<td>-0.25 (0.05)</td>
<td>-0.24 (0.05)</td>
</tr>
<tr>
<td><strong>R²(%)</strong></td>
<td>32</td>
<td>38</td>
<td>40</td>
<td>30</td>
<td>31</td>
<td>29</td>
<td>55</td>
</tr>
</tbody>
</table>
What About Residues?

Andrey demonstrating residuals plot in R

Aleksey Zemnitskiy   37 of 43   www.snowfallsystems.com
ICE-9 HF Analytics Platform

What is ICE-9?

- Spot and accumulated volatility indicators
- Designed for real-time streaming
- High throughput, low latency with sub-millisecond revisions
- Tools to simulate market prices and microstructure effects
- Comprehensive documentation

Local Deployment

- Supplied as a Java class library (JAR)
- Pure Java for cross-platform portability
- Supports thread safe design
- Easy-to-use API

Plugins

- Interface to R software
<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Unbiased</th>
<th>Consistent</th>
<th>Jump Robust</th>
<th>Time dependence noise</th>
<th>Endogenous noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td>RV</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Subsampling</td>
<td>TSRV</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Subsampling</td>
<td>TSRV &lt;sub&gt;adj&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Subsampling</td>
<td>MSRV</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Kernel</td>
<td>KRV&lt;sub&gt;FT&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Kernel</td>
<td>KRV&lt;sub&gt;NFT&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Range</td>
<td>RRV</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Range</td>
<td>MRV</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Range</td>
<td>MRV &lt;sub&gt;adj&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Quantile</td>
<td>QRV</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Quantile</td>
<td>QRV&lt;sub&gt;mq&lt;/sub&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Demand</td>
<td>DRV</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
ICE-9 has one of the most powerful stochastic simulation systems available on the JVM:

- discrete and descretely sampled continuous processes
- jump and volatility models
- market microstructure effects
- arbitrary number of correlated processes
- simulations in multiple dimensions
Subscription with FAST/FIX 5.0

- Subscription-based delivery allows clients to request real-time updates of volatility on a per-security basis.
- Clients may customize the feed to receive only the symbols they need.
- FAST (FIX adapted for streaming) uses binary encoding and compression to transport high volumes of data with ultra-low latency.
- FAST is an open format and has recognized industry support.

Diagram:
- ICE-9 SERVER
- High Frequency and Quant Trading
- Real-time Risk Management
- Volatility and Variance Derivatives Pricing
Download R code for the plots

http://www.snowfallsystems.com/download/public/demo_NY.zip
Thank you for attention!

Contacts
Snowfall Systems, Inc.
330 East 46th Street, suite 6S
New York, 10017 NY
Tel: (347) 993-1766
Email: info@snowfallsystems.com