What is High Frequency Data?
Likelihood Connection

Financial High Frequency Data

Per Mykland

University of Chicago, October 2012
Outline

1. What is High Frequency Data?
   - The data
   - Basic statistical inference

2. Likelihood Connection
   - General Connection
   - Some Applications
What is High Frequency Data?

Likelihood Connection

The data
Basic statistical inference

High Frequency Data

- In our case: financial prices, and/or volumes
- Intra-day:
  - transactions tick-by-tick, from TAQ, Reuters, etc
  - quotes - bid, ask - same sources
  - limit order books, harder to get but more information
  - stocks, bonds, futures, currencies, ...
  - low latency data
- Close to continuous observation:
  - Up to several observations per second
### Example of Transaction Data (medium density data)

<table>
<thead>
<tr>
<th>Time</th>
<th>Stock</th>
<th>Price</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRK 20050405 9:41:37</td>
<td>32.69</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:42</td>
<td>32.68</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:43</td>
<td>32.69</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:44</td>
<td>32.68</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:48</td>
<td>32.69</td>
<td>2900</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:48</td>
<td>32.68</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:48</td>
<td>32.68</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:51</td>
<td>32.68</td>
<td>4200</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:52</td>
<td>32.69</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:53</td>
<td>32.68</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:41:57</td>
<td>32.69</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:03</td>
<td>32.67</td>
<td>2500</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:04</td>
<td>32.69</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:05</td>
<td>32.69</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:15</td>
<td>32.68</td>
<td>3500</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:17</td>
<td>32.69</td>
<td>800</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:17</td>
<td>32.68</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>MRK 20050405 9:42:17</td>
<td>32.68</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

**Merck excerpt**

*April 4, 2005*

Total of 6302 Merck transactions on that day

On same day:

80982 Microsoft (MSFT) transactions

**Four years later:**

On April 6, 2009:

63846 Merck transactions

144842 MSFT transactions

---

**Likelihood Connection**

The data

Basic statistical inference

---

**Mykland**

Financial High Frequency Data
Evolution of Data Size

# of Merck transactions, first Monday in April

<table>
<thead>
<tr>
<th>Year</th>
<th># of Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>20000</td>
</tr>
<tr>
<td>2005</td>
<td>40000</td>
</tr>
<tr>
<td>2010</td>
<td>60000</td>
</tr>
</tbody>
</table>

Mykland

Financial High Frequency Data
Evolution of Data Size

# of Merck transactions, first Monday in April

![Graph showing the evolution of data size from 1995 to 2010. The x-axis represents the year, and the y-axis represents the log of the number of Merck transactions. The data shows a significant increase in transactions from 2005 to 2010.]
CME at Midnight

<table>
<thead>
<tr>
<th>Time</th>
<th>CentiSec</th>
<th>Quantity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:42</td>
<td>47</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:04</td>
<td>69</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:04</td>
<td>80</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:05</td>
<td>64</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:06</td>
<td>56</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:32</td>
<td>09</td>
<td>20</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:32</td>
<td>09</td>
<td>34</td>
<td>150150</td>
</tr>
<tr>
<td>00:01:52</td>
<td>24</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:02:32</td>
<td>03</td>
<td>10</td>
<td>150150</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>1</td>
<td>150175</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>1</td>
<td>150175</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>1</td>
<td>150175</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>5</td>
<td>150175</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>1</td>
<td>150175</td>
</tr>
<tr>
<td>00:02:58</td>
<td>43</td>
<td>1</td>
<td>150175</td>
</tr>
<tr>
<td>00:03:42</td>
<td>75</td>
<td>1</td>
<td>150150</td>
</tr>
<tr>
<td>00:03:43</td>
<td>20</td>
<td>1</td>
<td>150150</td>
</tr>
</tbody>
</table>

E-mini SP500 Futures
May 3, 2007

Total of 62659 trades on that day
<table>
<thead>
<tr>
<th>Time</th>
<th>CentiSec</th>
<th>Quantity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00:00</td>
<td>25</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:00</td>
<td>25</td>
<td>29</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:00</td>
<td>45</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:00</td>
<td>87</td>
<td>10</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>73</td>
<td>50</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>87</td>
<td>37</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>88</td>
<td>463</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>48</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>2</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>3</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>95</td>
<td>2</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>98</td>
<td>1</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>98</td>
<td>4</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>98</td>
<td>2</td>
<td>150850</td>
</tr>
<tr>
<td>10:00:01</td>
<td>98</td>
<td>3</td>
<td>150850</td>
</tr>
</tbody>
</table>

E-mini SP500 Futures
May 3, 2007

Total of 62659 trades on that day
What is High Frequency Data?
Likelihood Connection

The data
Basic statistical inference

Turning Data into Knowledge

- Modern quantitative finance uses high frequency constructions in stochastic processes:
  - to price assets, underlying and derivative
  - to construct trading strategies

- The high frequency data are the empirical realization of the same processes

- The data open a new angle on quantitative finance:
  - better estimators and models
  - well crafted daily summaries (relationship to sufficiency)
  - combination with longer horizon macroeconomic data
  - a complement to cross-sectionally based (implied) quantities
  - unification of econometrics, risk mgmt, and quantititative finance?
  - a new way of having fun with semimartingales
Direct Impact of Estimating Intraday or "Spot" Quantities

- Asset management, portfolio optimization
- Empirical or conservative options hedging
- Risk management
- Early detection of abrupt changes in market conditions
- Better trade execution
- Input to longer run models

This is relevant both from the public and private perspective.
Natural to use same model as in quantitative finance: the \( \text{Itô} \) process:

\[
\log \text{ securities price: } X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s
\]

\( B_t \) is Brownian motion; \( \mu_t \) and \( \sigma_t \) can be random processes

Model can also include jumps (different but related results)

High frequency data formalism:

- Up to several transactions per second, sampling times 
  \( 0 = t_0 < t_1 < \ldots < t_n = T \)
- Time period of analysis \([0, T]\): one day (5 min, 2 weeks)
What is High Frequency Data?

Likelihood Connection

The data

Basic statistical inference

Quantities that can be Estimated in In Data from One Day

- Classical target: Integrated volatility:
  \[ \langle X, X \rangle = \int_0^T \sigma_t^2 \, dt = \lim_{\Delta t \to \infty} \sum_{t_i+1 \leq T} (X_{t_{i+1}} - X_{t_i})^2 \]

- Other powers of volatility: \[ \int_0^T \sigma_t^p \, dt \]

- Leverage effect: \[ \langle \sigma^2, X \rangle_T \], or corresponding correlation

- Volatility of volatility \[ \langle \sigma^2, \sigma^2 \rangle_T \]

- Regression of one process on another, integrated alphas and betas, ANOVA

- Same quantities, but instantaneously

- Nonparametric trading strategies

- Liquidity; time to execution
The Classical case: Realized Volatility (RV) as Measure of Integrated Volatility

High frequency data: possibility to estimate $\langle X, X \rangle_T$ very precisely

Usual estimator: $RV = \sum_{0 < t_{i+1} \leq T} (X_{t_{i+1}} - X_{t_i})^2$ “realized volatility”

- consistent as $\Delta t \to 0$ (stochastic calculus)
- widely used (Andersen, Bollerslev, many others)
- convergence rate $n^{1/2}$, asymptotically mixed normal, with estimable variance (Barndorff-Nielsen & Shephard, Jacod & Protter, M & Z)
Microstructure Noise, and The Hidden Semimartingale model ("Nugget Effect")

- observed log stock price: $Y_{t_i} = X_{t_i} + \epsilon_i$
- $X_t$ is latent log price, semimartingale, say, Ito process
  \[ dX_t = \mu_t dt + \sigma_t dB_t \]
- $\epsilon_i$ is stationary or iid, or similar

In financial data
- more realistic model because of microstructure
- small deviations from semimartingale model allowable because it may not be possible to take advantage of these for arbitrage
- noise need not mess up options hedging
What is High Frequency Data?

Likelihood Connection

The data

Basic statistical inference

Theology vs. Data: RV vs Sampling Interval

dependence of estimated volatility on number of subgrids
What is High Frequency Data?
Likelihood Connection

Basic statistical inference

RV as One Samples More Frequently

dependence of estimated volatility on sampling frequency

volatility of AA, Jan 4, 2001, annualized, sq. root scale

sampling frequency (seconds)

Mykland
Financial High Frequency Data
Can We Learn Anything from Parametric Inference?

Thought Experiment:
- What if one pretended that $\sigma_t$ is constant over blocks of $M$ sampling times $t_i$?
- One possibility: parametric inference for each block, then aggregate results across blocks

Does this give estimators that are
- Consistent?
- Efficient?
- Look different than current estimators?

Important agendas: multivariate case, elusive univariate quantities (such as leverage effect, volatility of volatility)

Sufficiency can inform summaries
A statistical “risk neutral” (equivalent martingale) measure

- Can without much loss of generality assume $\mu_t = 0$
- Results shown for $\mu_t = 0$ carry over to general drifts $\mu_t$ using Girsanov’s Theorem and stable convergence
- Measure change commutes with stable convergence
- Same phenomenon: cannot consistently estimate $\int_0^T \mu_t \, dt$

To be precise:

- True probability $Q$:
  \[ dX_t = \mu_t \, dt + \sigma_t dB^Q_t \]
- Statistical risk neutral probability $P$:
  \[ dX_t = \sigma_t \, dB_t \]
An approximate model

- Sampling times $G_n = \{0 = t_{n,0} < t_{n,1} < \ldots < t_{n,n} = T\}$
- Break points $\tau_{n,i}$: Every $M$’th observation, or other subset of $G_n$
- Approximate model $P_n$: $dX_t = \sigma_{\tau_{n,i-1}} dB_t$ for $t \in (\tau_{n,i-1}, \tau_{n,i}]$
- For our results need: Asymptotic decoupling delay (ADD)

$$K(t) = \lim_{n \to \infty} \sum_i \sum_{t_{n,j} \in (\tau_{n,i-1}, \tau_{n,i}] \cap [0,t]} (t_{n,j} - \tau_{n,i-1})$$

$$= \frac{(M - 1)t}{2}$$ in equidistant case

Under $P_n$, for $\tau_i < t_j \leq \tau_{i+1}$, conditionally on $\mathcal{F}_{\tau_{n,i-1}}$:
- $\Delta X_{t_j} = X_{t_j} - X_{t_j-1}$ are independent normal vectors,
- $\Delta X_{t_j} \sim N(0, \sigma_{\tau_{n,i-1}}^T \sigma_{\tau_{n,i-1}} \Delta t_j)$
The main approximation result

Suppose: $\sigma_t^2$ is itself an Itô process:

$$d\sigma_t^2 = \nu_t dt + \xi_t dW_t$$

Assume $\nu_t, \xi_t$ are locally bounded; also $\inf \sigma_t^2 > 0$ (smallest eigenvalue $> 0$ in the case of matrix process)

**Theorem** (subject to regularity conditions):

- $dP$ and $dP_n$ are mutually absolutely continuous as probabilities on $(X_{t_0}, ..., X_{t_n})$
- $\frac{dP}{dP_n}(X_{t_0}, ..., X_{t_n}) \to \exp\{\eta Z - \frac{1}{2} \eta^2\}$ as $n \to \infty$, in stably law under $P_n$, where $Z$ is $N(0,1)$, independent of $F_T$, and where

$$\eta^2 = \frac{3}{8} \int_0^T \sigma_t^{-6} (\langle \sigma^2, X_t \rangle_t')^2 dt + \frac{1}{2} \int_0^T \sigma_t^{-4} \langle \sigma^2, \sigma^2 \rangle_t' dK(t)$$

related to leverage effect related to # in each block

In other words, $P_n$ and $P$ are contiguous (Le Cam)
What is High Frequency Data?

Likelihood Connection

General Connection

Some Applications

Implications of contiguity for Estimation

- $\theta$: quantity to be estimated, say $\theta = \int_0^T \sigma_t^4 dt$
- Suppose under $P_n$: $n^{1/4}(\hat{\theta}_n - \theta) \rightarrow \mathcal{N}(0, d^2)$, stably in law
- Then under $P$: $n^{1/4}(\hat{\theta}_n - \theta) \rightarrow \mathcal{N}(b, d^2)$ stably in law
- Convergence is stable in law; more generally: $n^\alpha(\hat{\theta}_n - \theta)$
- For this: weak additional assumptions

Effect of approximation is therefore:

- Possible asymptotic bias $b = d \eta \times P_n$-asymptotic covariance between $n^{1/4}(\hat{\theta}_n - \theta)$ and $\log \frac{dP}{dP_n}$
- Consistency, order of convergence, and asymptotic variance $d^2$ all stay the same
Example: Estimating Integrals of Powers of Volatility

Target ("parameter"): \( \theta = \int_0^T \sigma_t^p \, dt \)

"Moment" estimator: \( \hat{\theta}_n = c_p \frac{n}{T} \sum_{j=1}^n |\Delta X_t|^p \)

Under our approximation, for \( \tau_i < t_j \leq \tau_{i+1} \):
\[ \Delta X_{t_j} = X_{t_j} - X_{t_{j-1}} \text{ are iid } \mathcal{N}(0, \sigma_{\tau_i}^2 \Delta t) \]

UMVU estimator of \( \sigma_{\tau_i}^2 \Delta t \):
\[ \hat{\sigma}_{\tau_i}^2 \Delta t = \frac{1}{M} \sum_{t_j \in (\tau_{i-1}, \tau_i]} \Delta X_{t_j}^2. \]

The UMVU estimator of \( \sigma_{\tau_i}^p \) is proportional to \( (\hat{\sigma}_{\tau_i}^2)^{p/2} \)

Improved efficiency for \( \theta \) under \( P_n \). Carries over to \( P \)
Asymptotic Relative Efficiency of Estimators

M=20

M=100

Mykland

Financial High Frequency Data
What is High Frequency Data?
Likelihood Connection
General Connection
Some Applications

ANOVA (Analysis of Variance/Variation)

The setting:
- Processes: $X_t$ (can be multivariate), $Y_t$ are observed at times $t_j$
- The two processes are related by

$$dY_t = \sum_{i=1}^{p} f_s^{(i)} dX_s^{(i)} + dZ_t, \text{ with } \langle X^{(i)}, Z \rangle_t = 0 \text{ for all } t \text{ and } i$$

- Problem: estimate $\langle Z, Z \rangle_T$ based on the data ($Z$ (2001))
- Three interpretations of $Z$:
  - idiosyncratic component of price of a stock
  - hedging error of an option
  - estimation or model error

Simple solution: regular regression and ANOVA in each block of $M$ observations. Can use traditional regression theory.

$$\langle \hat{Z}, \hat{Z} \rangle_T = \frac{M}{M - p} \sum \text{(RSS in block } i)$$
Properties of estimator $\langle Z, Z \rangle_T$

- Asymptotic bias $b = 0$
- Asymptotic variance for $n^{1/2}(\langle \hat{Z}, Z \rangle_T - \langle Z, Z \rangle_T)$:
  
  \[ 2 \frac{M}{M - p} \int_0^T (\langle Z, Z \rangle_t')^2 dt \]

- Similar to results for $p = 1$ when $M \to \infty$ with $n$
**Integrated Beta**

- Same setting as for ANOVA: Regression equation:
  \[
  dY_t = \sum_{i=1}^{p} f_s^{(i)} dX_s^{(i)} + dZ_t, \quad \text{with } \langle X^{(i)}, Z \rangle_t = 0 \text{ for all } t \text{ and } i
  \]

- Problem: estimate \( \int_0^T f_t^{(i)} dt \)

- Same solution as for ANOVA: regular regression in each block of \( M \) observations. Can use traditional regression theory.

- Asymptotic variance of \( n^{1/2} (\int_0^T f_t dt - \int_0^T f_t dt) \) is
  \[
  \frac{MT}{M - p - 1} \int_0^T \langle Z, Z \rangle_t' (\langle X, X \rangle_t')^{-1} dt.
  \]
Preaveraging and local MA(1)-ness

midquote May 3 2007

1-min averaged data

Mykland

Financial High Frequency Data
What is High Frequency Data?

Likelihood Connection

General Connection

Some Applications

Best Bid Process

Best Bid on May 3 2007

1-min Averaged Best Bid

Mykland

Financial High Frequency Data
Best Ask Process

Best Ask on May 3 2007

1-min Averaged Best Ask
What is High Frequency Data?

Likelihood Connection

General Connection

Some Applications

Before and After

midquote series for 05–03 trading session

pre–averaged 15 sec series

pre–averaged 1 min series

pre–averaged 5 min series

Index

log price

log price

log price

log price

0 20000 40000 60000 80000
11.918 11.921 11.924
midquote series for 05–03 trading session
time
log price
0 1000 2000 3000 4000 5000
11.918 11.921 11.924
pre–averaged 15 sec series
Index
log price
0 50 100 150 200 250
11.918 11.921 11.924
pre–averaged 1 min series
Index
log price
0 20000 40000 60000 80000
11.918 11.921 11.924
pre–averaged 5 min series
Index
log price
0 50 100 150 200 250
11.918 11.921 11.924
pre–averaged 5 min series

Mykland
FlInancial High Frequency Data
### The Limits to Inference

<table>
<thead>
<tr>
<th>“parameter”</th>
<th>Effective Sample Size in Different Situations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Microstructure Absent</td>
</tr>
<tr>
<td>volatility</td>
<td>daily</td>
</tr>
<tr>
<td>regression</td>
<td>$O_p(n)$</td>
</tr>
<tr>
<td>ANOVA</td>
<td>$O_p(n^{1/2})$</td>
</tr>
<tr>
<td>leverage effect</td>
<td>daily</td>
</tr>
<tr>
<td>vol of vol</td>
<td>$O_p(n^{1/2})$</td>
</tr>
<tr>
<td>vol of vol</td>
<td>$O_p(n^{1/4})$</td>
</tr>
</tbody>
</table>

$n$ is daily number of transactions/quotes;
Effective sample size = rate of convergence$^2$
### The Limits to Inference

<table>
<thead>
<tr>
<th>“parameter”</th>
<th>Rate of Convergence in Different Situations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Microstructure Absent</td>
</tr>
<tr>
<td></td>
<td>Daily</td>
</tr>
<tr>
<td>volatility regression</td>
<td>$O_p(n^{1/2})$</td>
</tr>
<tr>
<td>ANOVA</td>
<td>$O_p(n^{1/4})$</td>
</tr>
<tr>
<td>leverage effect</td>
<td>$O_p(n^{1/4})$</td>
</tr>
<tr>
<td>vol of vol</td>
<td>$O_p(n^{1/4})$</td>
</tr>
</tbody>
</table>
Conclusions

- A vast amount of data
- How to turn into knowledge?
- How to turn into regulation? risk management? trading?
- Financial prices have “error"
- Connection to parametric inference
Political High Frequency Data: Romney on Intrade, 23 Oct 2012