Experimental Finance

IEOR – Columbia University

Mike Lipkin, Alexander Stanton
Outline

• Why are we teaching this?
• Laboratory Focus
• IVY and LiveVol Databases
• Initial Setup and Using the Database

Why?

- **Theory:** Conventional Finance
  - Theories show mis-pricings of the market
  - Theories follow mathematically rigor and are complete
  - Theories determine the “real” price

- **Practice:** Guys Like Mike
  - Classical theory doesn’t hold in practice
  - There are plenty of opportunities to make (and lose) money through strategies that exploit this gap. The trick is using intuition, observation and plenty of back-testing, both to find them and to minimize losses when (not if) they occur.

- **Implementation:** The Lab
  It does not matter whether you favor theory or games playing – without rigorous testing, the process will be painful. Unexpected problems, even insufficient, bad, late and unexpected data all make for a fun time.

Be Good at All Three Modes
Why?

- We’ll be taking a look at how conventional theory can be at odds with real-life trading
- Back-testing and exploring (un)known occurrences. Can predictions be made?
- We’ll be looking at how practical constraints affect an analyst’s ability to evaluate theories, monitor signals and execute trading strategies
- Data latencies, computation times and execution costs all impact an analyst’s ability to do what may be theoretically (or intuitively) possible

Your ability to blend theory, practice and the right amount of skepticism will drive your PNL
A few practical considerations:

- Execution costs (including the fact that a round trip on the CBOE must be under 10ms or you will get fined)

- Regulations require market makers to provide liquidity. The solution is to quote options very wide in order not to get hit, but what if things move really fast, e.g. during a flash crash?

- What happens when your RT link goes down for 2 seconds, 1 minute, 1 hour?

- Strategy pre-calculation across entire market take hours to run nightly. What happens when the nightly feed goes down, or everything but the dividend data comes in, or the admin crashes the server by pouring coffee on it, or….

- You’re bidding $2.50 because you think it’s worth $2.60. Everyone else is bidding $2.30. Who is right?

Some you can do something about, others very little
Interested in Mean Reversion strategies?

- A strategy often determines a theoretical relationship between two measurable data, e.g. between instruments and/or baskets of instruments

\[ Y_t = f(X_t) + \varepsilon_t \]

Where \( \varepsilon \) is the error/deviation/premium used to trigger a strategy

- The trigger levels are usually implemented gradually, i.e. new positions are put on (off) with increasing (decreasing) \( \varepsilon \), up to a defined maximum exposure limit

- This results in a strategy with an adaptive threshold and direction-dependent parameters for positive and negative \( \varepsilon \) swings (we all know up does not equal down and we are optimizing for P/L)

- When things are going really well we double down. This all sounds great...
At some point after 2001, this will be heard: “Let’s just hang on to this a little longer…”
Laboratory Focus

- Data extraction and manipulation
- How to interact with very large data sets
- In industry, you have to be a good mechanic: Knowledge of data sources, matlab, excel, python, scripting and other glue is critical to getting things done.
- 20% is coding the algorithm
  80% is data checking and error-handling
- You are free to use any analytical tool (Excel, matlab, SPSS, Python)

Our data repository is the IVY database by OptionMetrics and LiveVol for intraday data
Some Statistics

• US market data traffic is larger than all traffic rates for major equity markets worldwide

• OPRA (Option Price Reporting Authority) peak processing capabilities are:
  – 100 MPS (1996)
  – 360,000 MPS (2006)
  – 4,000,000 MPS (2011 – weekly’s introduced)
  – 22,000,000 MPS (2015)

• The options market is doubling every year

• Quote-to-trade ratio is above 1000 to one

• Average share holding period: 5yrs (1970), 7mo (2007), 22s supposedly (2012)
Some IVY Statistics

**IVY:**
- Close to 17 years of data
- 500 GB+ of data
- ~139 Thousand security names (incl. name changes)
- ~33 Million security prices
- ~642 Million option prices

**Note that these are only close prices. Looking at intraday prices involves significant scaling and performance consideration**

- LiveVol: 1TB+ for 1 month of data
  We have SP100 stocks + SPY ETF
Companies that use Livevol

- Barclays
- Morgan Stanley
- Citibank
- Bank of America
- CBOE
Companies that use IVY

IVY is used by many large companies, such as:

- Lehman Brothers
- Bear Stearns

And many non-fictitious companies:

- Bloomberg Research
- Parallax Fund
- Gargoyle Strategic Investments
J. Gatheral, I. Matic, Ivan, R. Radoicic, D. Stefanica – “Tighter Bounds for Implied Volatility” August 7th, 2017


Y. Yang, Y. Zheng, T.M. Hospedales – “Gated Neural Networks for Option Pricing: Rationality by Design” November 30th, 2016

H. Kraft, A. Schmidt, "Systemic Risk in the Financial Sector: What Can We Learn from Option Markets?"


A. Fodor, S. Gokkaya, "Implied Volatility and Flotation Costs of SEOs"

K. Zhang, "Option Pricing and Calibration with Time-Changed Levy Processes"

• 1996/01 – 2013/08
• All equity/index securities (AMEX, NYSE, NASDAQ, OTC)
• Close prices for stocks and options (including volume and open interest)
• History of name/ticker/CUSIP changes
• Distributions and splits
• Industry groups
• Dividends
• Interest rates
• Option data: implied Vols, historical Vols, Deltas etc.
• Daily Updates Available in production environment
• Running on Microsoft SQL (i.e. MS SQL)
How Do We Use All This Data?

• In production (whether live trading or back testing a strategy), the balance between analytic calculations, real-time strategy calculation and trade execution is paramount.

• With ever increasing data, what is calculated on the data store vs. analytic servers becomes a critical consideration:
  – Network bandwidth (looping through every row in the an option price table is very expensive)
  – Stats/Math packages are great at calculating, but cannot handle extremely large data sets
  – Database servers are bad at complex calculations. Only rudimentary functions are available (count, average, sum, max, min etc.)
General Strategy

1) Let the database serve data, and the analysis engines do the crunching. Have distributed clients do complex algo calculations.

2) Have the server do as much as possible to reduce the size of the dataset being passed to the execution engine or analytics package, including pre-computation and data segmentation.

Seems obvious… not so easy in practice as we’ll find out
Some Basics - Databases

- Hierarchical and network models (the beginnings of databases)
- Relational Databases (MS SQL, Oracle, DB2, MySQL, Informix)
- Object Oriented Databases
- “NoSQL” databases (MongoDB, CouchDB, Hadoop)
- Real Time in-memory databases (K)

All have their place, however Relational Database Management Systems (RDBMS) are currently the norm for handling and exchanging data until very particular problems are better handled by a specific model.
Four Basic Building Blocks:

- **Tables:**
  Information is split into multiple tables in order to minimize redundant information storage and to maximize performance.

- **Indices/Keys:**
  Lookup tables that organizing data in an efficient manner for sorting – without indices large data sets are not possible.

- **Views:**
  Aggregation of table information into one virtual table.

- **Stored procedures:**
  Similar to functions in C – can pass data by reference or by value. Have the added capability of returning a query result.
Getting started - initial problem set

- Don’t just jump in
- Read the course package
- Read the IVY reference very carefully
- Understand the IVY securityID
- There are many gotchas:
  - Delisted stocks
  - Exchange flags
  - Stocks that only started listing options well into the data set
  - Stock name changes
  - Stock splits, mergers etc.
Databases

You must obey namespaces and specify the database:

Select <column> from <database>.<owner>.<table>

**XFDATA** Database contains historical data

**XF** contains your tables/views/functions/storedProcedures (empty right now apart from a few helper tables)

Select securityID from xfddata.dbo.security

Select date from xfddata.dbo.lv_minute_options_calcs

Select signal from xf.ivyuser.ss2892_myalgo
Ivy Tables

Major tables of interest

- Security (and perhaps Security_Name)
- Security_Price
- Distribution
- Option_Price_<DATE>  (e.g. Option_Price_2010_03)
- Option_Price_View - UNION view of all 180 Option_Price monthly tables + the daily downloaded data in Option_Price (empty)
## Ivy Tables: Option_Price

<table>
<thead>
<tr>
<th>Datatype</th>
<th>Length</th>
<th>Field Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>integer</td>
<td>-</td>
<td>Security ID</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>-</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>char</td>
<td>5</td>
<td>Root</td>
<td></td>
</tr>
<tr>
<td>char</td>
<td>2</td>
<td>Suffix</td>
<td></td>
</tr>
<tr>
<td>integer</td>
<td>-</td>
<td>Strike Price</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>-</td>
<td>Expiration Date</td>
<td></td>
</tr>
<tr>
<td>char</td>
<td>1</td>
<td>Call/Put Flag</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Best Bid</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Best Offer</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>-</td>
<td>Last Trade Date</td>
<td></td>
</tr>
<tr>
<td>integer</td>
<td>-</td>
<td>Volume</td>
<td>Volume</td>
</tr>
<tr>
<td>integer</td>
<td>-</td>
<td>Open Interest</td>
<td>Open Interest</td>
</tr>
<tr>
<td>char</td>
<td>1</td>
<td>Special</td>
<td>Special Settlement Flag</td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Implied Volatility</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Delta</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Gamma</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Vega/Kappa</td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>-</td>
<td>Theta</td>
<td></td>
</tr>
<tr>
<td>int</td>
<td>-</td>
<td>OptionID</td>
<td>OptionID</td>
</tr>
<tr>
<td>int</td>
<td>-</td>
<td>Adjustment Factor</td>
<td></td>
</tr>
</tbody>
</table>
Other IVY Tables

• EXCHANGE
• OPTION_INFO (exercise style, AM/PM settlement)
• INDEX_DIVIDEND
• STD_OPTION_PRICE (interpolated options)
• OPTION_VOLUME
• VOLATILITY_SURFACE
• HISTORICAL_VOLATILITY
• ZERO_CURVE
Useful Tables

Useful tables not part of IVY:

- Expirations
- optionTradeDates
- SP500

Connecting to the database:

- Write SQL and interact with the database in code using matlab drivers, C, Java or use clients such as LINQPad, TOAD, etc.
- However, a tool that allows you to inspect the database at it’s fundamental level is highly recommended. This gives you insight into not only the data, but indexes, foreign keys, types, defaults and definitions.
Microsoft SQL Management Studio (Express version is free)
Table Columns & Indexes
SELECT Queries

**Basic Format**

```
SELECT <columns>
FROM <database>.<owner>.<table>
WHERE <condition>
ORDER BY <columns>
```

e.g. “SELECT ticker FROM security WHERE securityID=1”
result: “MSFT”

e.g. “SELECT securityID FROM security WHERE ticker like “MSFT””
result: 1
WHERE clause can use several operators:

\[
\begin{align*}
&> \\
&< \\
&= \\
&\text{AND, OR, LIKE, BETWEEN, IS NULL, IN, NOT}
\end{align*}
\]

e.g. 
```
SELECT strategy, ticker, date, closePrice, PnL
FROM myStrategies
WHERE strategy LIKE 'stratA%' AND
date BETWEEN '2013-01-02' AND '2013-02-01'
ORDER BY date, ticker
```

Returns all strategy PnL and closePrices in the date range for the strategies that begin with the letter A (e.g. “stratA1”, “stratA2”, “stratA2b”, “stratA2bFixed” etc.), ordered first by date, then ticker.
Conventions and Hints:

- **When creating system objects, prepend your UNI in order to differentiate from others.**
- Name your columns with an alias when using built-in functions (MAX, MIN etc.)
- Submit problem sets using “Lastname, firstname – Problem Set N”
- ZIP the file
- Highlight a table and use <ALT>-F1 to get information on it.
- Use the **STOP** button
- SQL Analyzer has a solid help reference for T-SQL
- IVY Reference located on the server
Get Started…

…and don’t forget the STOP button