Challenges in Computational Finance and Financial Data Analysis

James E. Gentle
Department of Computational and Data Sciences
George Mason University
jgentle@gmu.edu
http:\\mason.gmu.edu/~jgentle

Outline

- **Financial data**

Outline

- **Financial data**

- **Mining financial data**

Outline

- **Financial data**
- **Mining financial data**
- **Why we're interested**

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data
- Volatility patterns

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data
- Volatility patterns
- Text analysis

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data
- Volatility patterns
- Text analysis

Data

I consider whatever can be encoded and stored in the computer to be "data".

That is,

"information" is data;

"knowledge" is data;

a computer program is data;

text documents are data;

images are data.

Financial Data

Financial data include

- balance sheet and earnings statement data
- officers and directors
- news items relative to activities of the company or of its competiters
    - etc. etc. etc.
- stock prices
- trading volume
    - etc. etc. etc.

Data on Trades of Financial Assets

I will limit the discussion to data relating to trades of publicly-traded financial assets, or securities.

A security may be a share in a corporation, it may be an option on a number of shares, it may be a bond, it may be share in a portfolio of other securities, and so on.

There are approximately 2,800 different securities (corporate shares or portfolio shares) traded on the NY Stock Exchange.

Each trading day on the NYSE, approximately 2 billion individual shares are traded in approximately 6 million trades for a total of approximately 75 billion dollars.

By most measures, the NYSE is the largest market, but there are several others in the US, including the NASDAQ, at which securities similar to those on the NYSE are traded, and various commodities and futures markets.

Data on Trades of Financial Assets

The primary data are the multivariate time series of price and volume of every trade for each security. In the US, this may be 20×10^6 bivariate time-stamped points (price and volume) daily.

This is not extremely large as datasets go nowadays.

And unlike the case in the physical sciences, the amount of data does not depend on the number of experiments the scientist is able to do or on the number of sensors or satellites that are deployed to collect the data.

Additional data describe activities of companies or other news items that may affect the price. This rather amorphous set of data is quite huge.

Outline

- Financial data
- **Mining financial data**
- Why we're interested
- The data generating process
- Stylized facts about financial data
- Volatility patterns
- Text analysis

Data Mining and Knowledge Discovery

In the early 1980s it was discovered that when the winner of the Super Bowl was a team from the old American Football League, the market went up for the rest of the year.

Who would have expected such a relationship?

It could have been discovered by mining of large and disparate datasets.

It is *knowledge* discovery! (It actually happened.)

It is interesting!

Unfortunately, it is worthless.

Data mining and knowledge discovery must be kept in context.

Data Mining and Knowledge Discovery: The "January Effect"

Several years ago, it was discovered that there are anomalies in security prices during the first few days of January.

The year after the discovery, the anomalies disappeared (although they're still being discussed).

Duh!
In the field of finance there is an interesting variation on the uncertainty prinicple.

The market is efficient!
(If you believe that, you probably believe the tooth fairly is what makes the market efficient.)

If there was a systemic reason for the January effect, might that cause result in a cyclic, but attenuated anomaly?

Technical Analysis: A Venerable Application of Data Mining

"Technical analysis" (as distinguished from "fundamental analysis") is based only on price data.

The assumption is that

future price changes

are related to patterns of

past price changes.

"Momentum" — or just a random walk?

"Head and shoulders" — or just a random walk?

"Broadening Top" — or just a random walk?

What happens **after** one of these quaint patterns?

Outline

- Financial data
- Mining financial data
- **Why we're interested**
- The data generating process
- Stylized facts about financial data
- Volatility patterns
- Text analysis

Why Are We Interested in This Kind of Data?

Understanding of the data can help regulators ensure that the trades are "fair". Most markets now have in place diagnostic programs that identify suspicious trading activity.

The programs are rather primitive. (They work by detecting anomalous data; but to do that we need good models of non-anomalous data.)

The ability to mine the potentially relevant text data is lacking.

Orderly markets are desirable. Understanding the large volatility swings would help preserve confidence in the markets.

Outline

- Financial data
- Mining financial data
- Why we're interested
- **The data generating process**
- Stylized facts about financial data
- Volatility patterns
- Text analysis

Pricing Models

A stochastic model of the price of a stock may view the price as a random variable that depends on previous prices and some characteristic parameters of the particular stock.

For example, in discrete time:

$$S_{t+1} = f(S_t, \mu, \sigma)$$

where t indexes time,
μ and σ are parameters,
and f is some function that contains a random component.

The randomness in f may be assumed to reflect all variation in the price that is not accounted for in the model.

Pricing Models

The model

$$S_{t+1} = f(S_t, \mu, \sigma)$$

is usually given one of two forms, either a time series model, such as a GARCH model, or a stochastic diffusion model driven by Brownian motion.

A simple form of the latter type of model, is geometric Brownian motion,

$$\mathrm{d}S(t) = \mu S(t)\mathrm{d}t + \sigma S(t)\mathrm{d}B(t),$$

in which μ and σ are constants, characteristic of the particular stock being modeled.

Use of this model, although a somewhat crude approximation, led to a revolution in the pricing of derivative assets.

Pricing Models

There are several aspects of observational data that indicate that the simple geometric Brownian motion model does not describe the data generating process very well.

One approach would be to substitute some other distribution for the Gaussian. Another would be to superimpose some kind of jump process.

Whatever kind of model may work best, it is clear that a key component of the model standard deviation of the rate of return (the σ in the geometric Brownian motion model).

The is what financial analysts call risk or volatility.

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- **Stylized facts about financial data**
- Volatility patterns
- Text analysis

Rates of return do not fit a Gaussian distribution well.

• Heavy tails. The frequency distribution of rates of return decrease more slowly than $\exp(-x^2/2)$**.**

• Asymmetry in rates of return. Rates of return are slightly negatively skewed. (Because traders react more strongly to negative information than to positive information.)

• Asymmetry in lagged correlations. Coarse volatility predicts fine volatility better than the other way around.

• Aggregational normality.

• Quasi long range dependence.

• Seasonality.

• Custering of volatility.

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data
- **Volatility patterns**
- Text analysis

Volatility

Volatility is the standard deviation of the rate of return.

A sample standard deviation can usually be used to estimate a model standard deviation. The problem is that it is not constant.

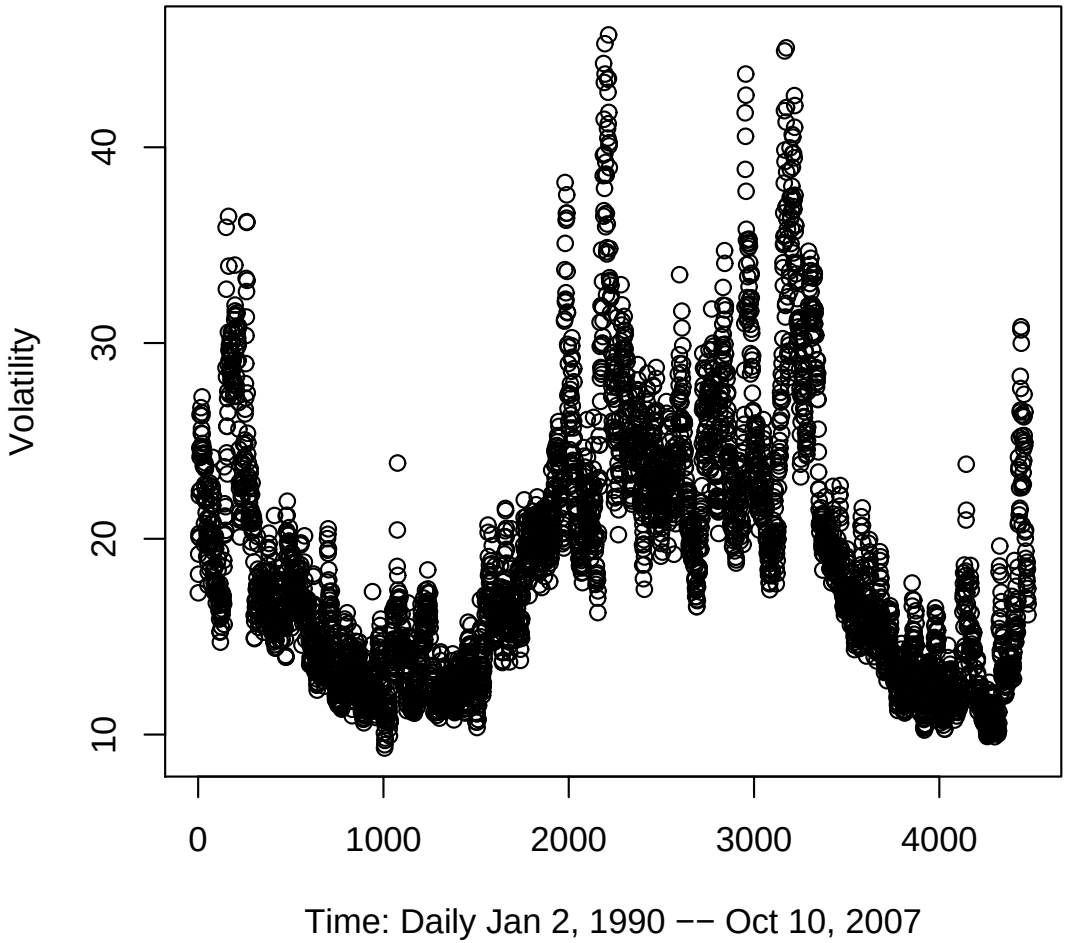
Developing a meaningful way to measure volatility in such streaming data is a very interesting research project.

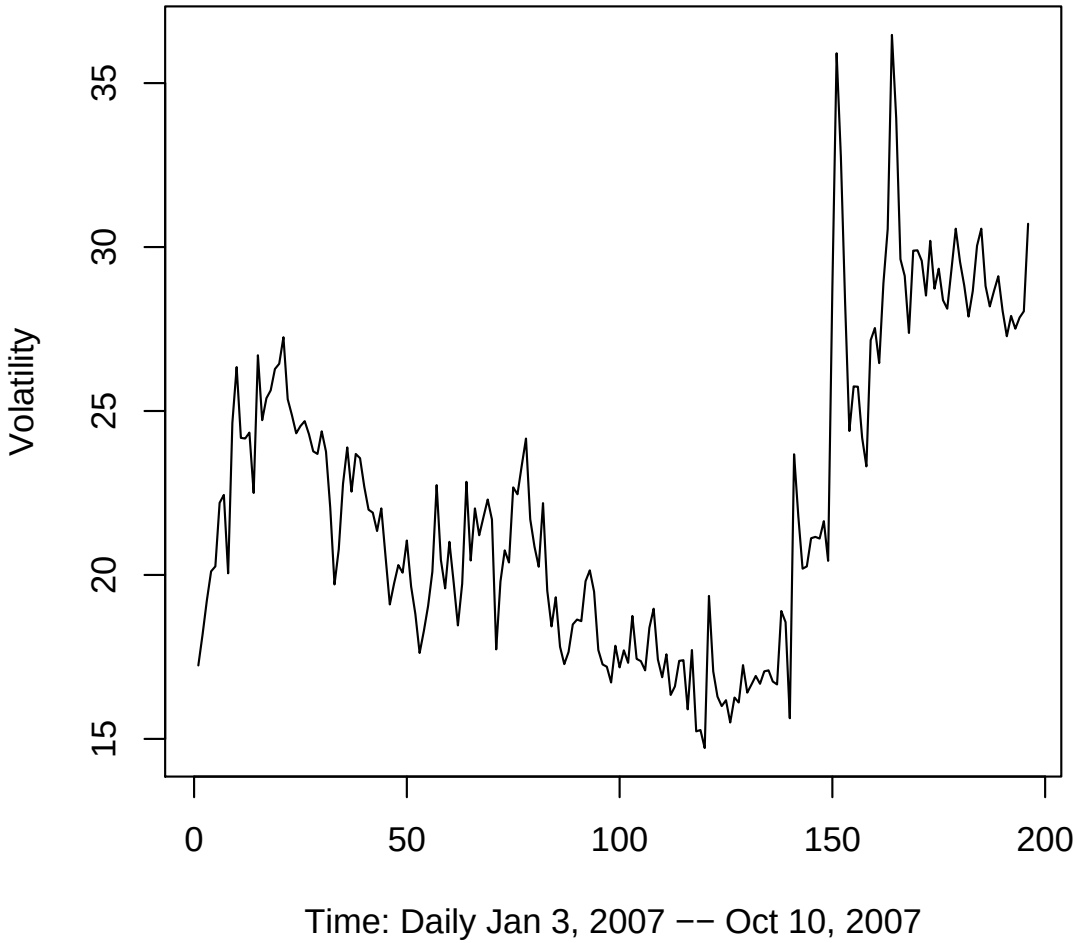
The study of volatility, including meaningful ways to measure it, should be a fruitful area for cyber-enabled discovery.

A Surrogate for Volatility

In the meantime, those who study volatility use the volatility implied by a modified Black-Scholes formula applied to options on the S&P500.

It's called the VIX ("volatility index"). Just like other indexes, you can trade futures on it.


Volatility
10
20
30
40
0
1000
2000
3000
4000
Time: Daily Jan 2, 1990 −− Oct 10, 2007


Volatility
15
20
25
30
35
0
50
100
150
200
Time: Daily Jan 3, 2007 −− Oct 10, 2007

Volatility Clustering

What is the meaning of the clusters of volatility?

If we look at the volatility of individual securities, we find a similar clustering.

Are volatilities of individual securities positively correlated? (Yes, even if their prices are negatively correlated.) How do you measure correlation of standard deviations?

Can increases in volatility of some securities indicate future increased volatility in the index?

Can volatility be related to the derivatives market?

Can volatility be related to global markets?

Volatility patterns suggest constrained clustering.

Volatility Clustering

Can this swarming behavior be understood? Are there leading indicators of it?

Is the most fruitful approach to seek explanations in basic human nature?

or, perhaps are there exogenous economic events that trigger volatility increases?

or, can an accumulation of various analysts discussions or touts predict increased volatility, perhaps beginning in one sector.

Outline

- Financial data
- Mining financial data
- Why we're interested
- The data generating process
- Stylized facts about financial data
- Volatility patterns
- **Text analysis**

Text Mining

There are thousands of documents related to financial assets generated daily.

These come in a variety of forms and from a variety of sources.

Developing some taxonomy of relevant documents would be a useful exercise.

An initial approach would be to limit the catalogue to a small number of documents from a few large financial research houses, and develop methods for relating their content to asset prices.

Data Mining of Financial Data

Financial data presents a number of challenges for mining.

Much of the data mining in this area has yielded only meaningless relationships.

Meaningful progress must come from an integrated exploration of data from a wide range of sources, both price/volume data from multiple markets and text data from a variety of sources.