Prediction of Stock Price Movement from Options Data

Charmaine Chia (cchia@stanford.edu)

Background

An **option** is a contract that gives the buyer the *right* to buy or sell an underlying stock at an agreed upon **strike price K**, during a certain period of time.

Q: Can we predict how the underlying stock price will move from **time series data** on the options market?

Examples of options variables:

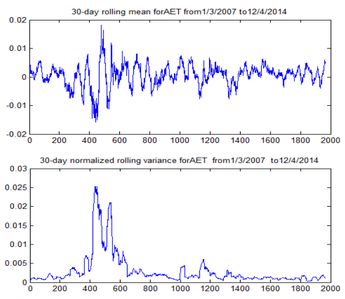
- Volume of puts & calls traded
- Put-call parity (PCP) deviance
- Implied volatility (IV)
  → Perceived future volatility of stock
  → Used to calculate options price
- IV spread & skew

The Data

- Time series data of 57 healthcare companies from 1/3/2007 to 12/4/2014
  - Data from successive time points is not i.i.d
  - Not necessarily normally distributed
- From the stock price data, we can obtain daily returns:

$$Return = \frac{Close\ price\ today\ - Close\ price\ yesterday}{Close\ price\ yesterday}$$

- From the options data, we get 39 options market-related variables, often highly correlated with each other.
- Time series data can be smoothed by applying either
  - Simple Moving Average (SMA)
  - Exponential Moving Average (EMA)
  - Gaussian moving filter

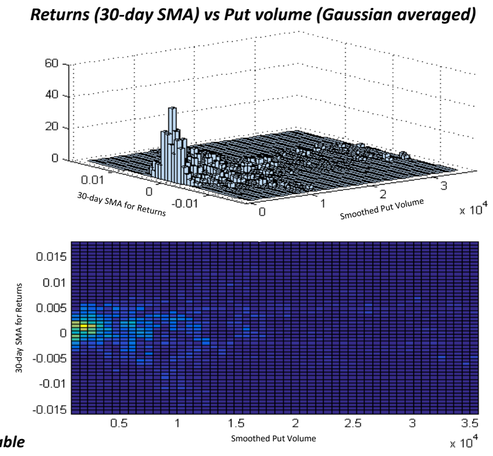


Initial Approach

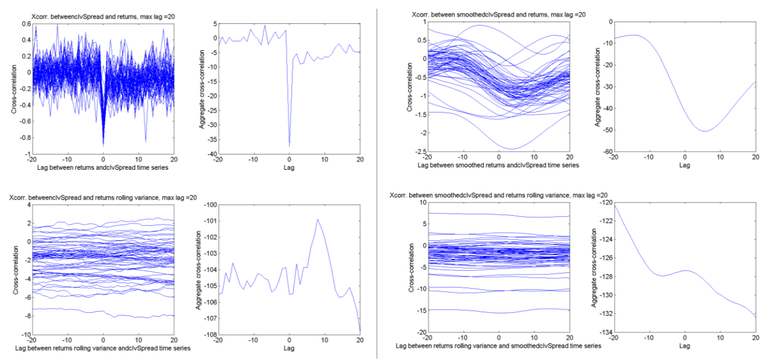
Q: Are there any obvious correlations between the returns data and individual options variables?

- Scatter plots of (Options variable(t), Returns(t))
- Cross-correlation of time series of
  - Raw Returns vs Options variable
  - Smoothed Returns vs Options variable
  - Returns rolling variance vs Options variable

Returns (30-day SMA) vs Put volume (Gaussian averaged)



Overlay of 57 Cross-correlation plots for time series Raw & Smoothed Returns vs an Options Variable

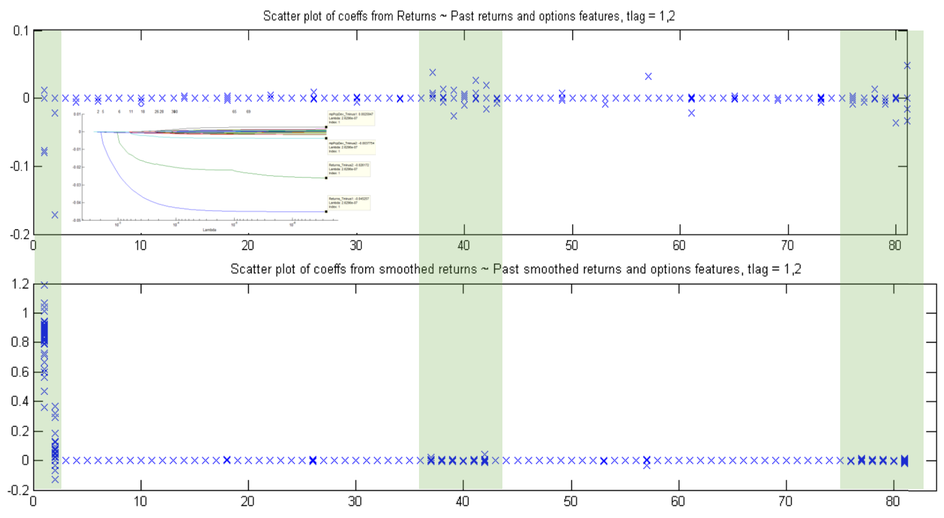


Based on model that there might be a lag between stimulus (due to options variable) and effect (on returns) → useful for regression

However, the xcorr plots are hard to interpret. Results from raw and smoothed variables also differ.

Linear regression with regularization (elastic net, α = 0.5)

- Returns (t) ~ Returns (t-1) + Returns (t-2) + Options variables (t-1) + Options variables (t-2)
- Not expected to work well, since Returns are determined by far more factors



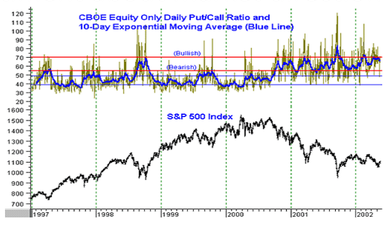
- Hard to compare relative importance of variables since they are not standardized
- No clear non-zero predictor emerges from regression on raw returns
- Past returns and Variables 37-42 possibly useful

Classification problem

- Can we even predict the direction of stock movement (i.e. + or – return)?
- Does a linear model accurately capture the relationship between the signal and effect?
- What variables contain should we use?

Try:

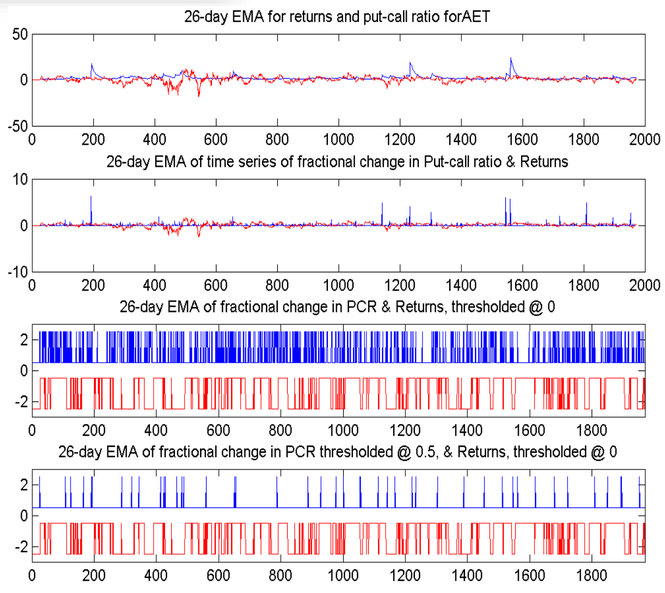
- Put-to call ratio → indication of trader sentiment about market direction



- 26-day EMA of Returns and PCR→ capture trends with less random jumps

T = 26

$$EMA_n = P_n \frac{2}{T+1} + EMA_{n-1}\left(1 - \frac{2}{T+1}\right)$$

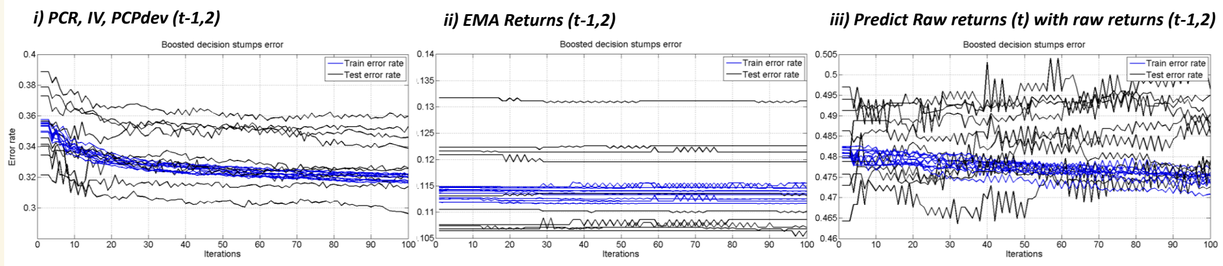


- Second plot shows the PCR data processed to get daily **fractional change in PCR**.
- Possible (negative) correlation between position of thresholded spikes in PCR and returns?

Decision stumps boosting

- Decision stump: $\phi_{j,s}(x) = sign(x_j - s) = \begin{cases} 1 & if\ x_j \geq s \\ -1 & if\ otherwise \end{cases}$
- x_j consist of past returns, PCR and options variables
- 10-fold CV: training data drawn from 52 companies, testing data from remaining 5 companies
- Effect of omitting certain variables tested

Prediction error for 3 experiments using different X variables. 100 iterations, 10-fold CV



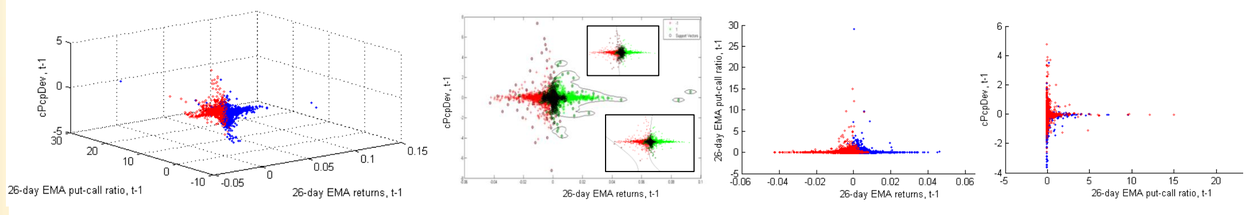
Summary of average training error for different combinations of X variables used in decision stumps

	A) Returns only	B) PCR	C) IV + PCPdev	B) + C): PCR + IV + PCPdev
w EMA Returns	11.5%	11.7%	11.5%	11.5%
w/o EMA Returns	48.5% (Raw returns)	38.2%	33.3%	33.0%

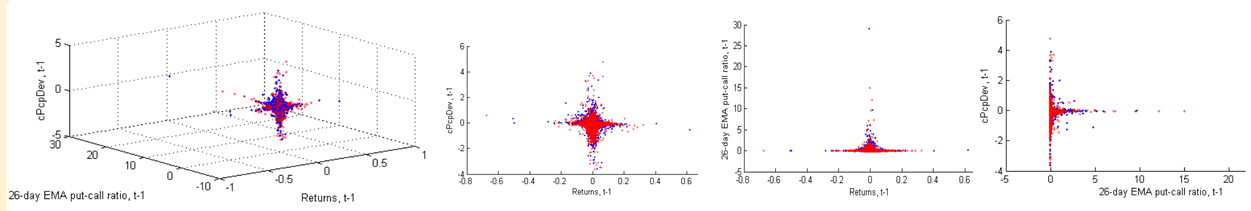
- Raw returns are too noisy to make predictions on / with
- EMA return (t-1) is the most predictive variable
- PCR, IV and PCPdev contain some signal
- Raw return can be extracted from EMA return prediction

Other classification schemes

- One problem with decision stumps boosting is that the way hypotheses are picked is somewhat arbitrary and difficult to intuitively explain
- Nonetheless, it informed us on which variables were more significant. We can plot the data along the top 3 variables from each class (A, B, C):



(Contrast this to the scatter plots for classifications thresholded on raw returns):



➔ Region of overlap of data points labelled +1 and -1. Unlikely to be resolved effectively using a linear learning algorithm.

➔ Tried: SVM with different kernels (linear, quadratic, RBF)
Convergence obtained only when the KKT violation level is increased to 0.2 ~ 0.4

Future work

- Model to predict the *magnitude* of returns, given prediction of whether it will be + or –
  ➔ MSE of 7.18e-06 on full dataset
- MSE of elastic net regression ranged from 6.66e-06 to 1.82e-05 (as the lambda is decreased)
- Also try: Regression trees, SVM regression