**MTH9898 Assignment A**

Due Feb 29th.2016

**STUDENT: YUE WANG**

**Initial data understanding:**

I used python to visualize the data pattern for both price and volume of the small data input, (apply a non-parallel fashion) to statistically test some potential methodologies of recognizing the pattern and separating the noise (given the smaller testing input set represents the larger testing data set). The traditional three sigma deviation from price and volume applied, the underlying assumption is to leverage the central limit theorem, the i.i.d. price/volume series’ means will converge to a normal distribution, which implies the possibility of tail events happens should diminish as it moves far away from mean. By applying this methodology, the 10K, 100K, 1000K both applied, it turned out a terrible filter as almost 50% data input being identified as noise for 100k and 1000k. By plotting things out, I found the above mentioned techniques serve terrible purpose, so the assumption of normality is forgone. From the plot, we do see price jump frequently and spike in a short period of time, which testify the poor performance of central limit theorem. Intuitively, the volume could vary a lot from time to time, so it doesn’t really make sense from a practical perspective to apply central limit theorem. For volume, there is no good way filter out outliers as market orders or potentially the hidden orders are all possible, the only reasonable treatment I proposed is to exclude any trading volume that is below 0.

**Methodology:**

1. **Price and Volume filter:** This is an adaptive exercise, I tuned different parameters to make it work on different scales, it turns out the price volatility set not to exceed 25% between consecutive observations and volumes should not be negative.
2. **Date filter:** after passed price and volume filter, the current data also needs to be tested for adjacency, if the date are in the same date for the observation before it and after it, then I test how close each observations are for the two neighbor tests, if time lag is within 5 seconds then it passed, if not I will test a front and back sliding windows of 10 each, say if a data is not 5s away from it prior or its later observation, I will test the prior 10 observations as front sliding windows [n-10, n-9…n-1], if 8 or above are 30s within the current observation, and also later sliding windows[n+1, n+2,… n+10], if 8 or above are 30s within in order to passes as a valid observation.
3. **Normality test:** I used Jarque-Bera test to test the normality of return data, by setting a local variables and keep tracking the running total of four moments separately. The ideas is the testing statistics are assumed to followed a chi-square distribution given a large data set.

**Result**:

The noise for all level inputs are about roughly 2% of the input, (10k, 100k, 1000k) and Jarque Bera Test at 99% confidence interval indicates the non-normality of the returns (it doesn't follow a normal distribution).