Empirical Invariance in Stock Market and Related Problems

Chii-Ruey Hwang
Institute of Mathematics, Academia Sinica, Taipei, TAIWAN

1 Introduction

We study stock market data from an empirical point of view without assuming any model by looking at simple attributes. Our approach is to describe these attributes using as little information as possible. The raw data set comes from the Wharton Research Data Services (WRDS). I would like to thank the College of Management, National Taiwan University, especially Prof. Shing-yang Hu for granting my access to the WRDS in November 2007.

This is an ongoing research with Lo-bin Chang, Shu-Chun Chen, Alok Goswami, Fushing Hsieh, Max Palmer, and Jun-Ying Chen who manages our data set. In this note, I just give a sketch of what we did in [2] and [3]. [4] contains further statistical analysis on highly volatile periods. Related work may be found in [5]. [6] is a book for general audience, I found that some of the points made there are still valid now.

Let the discrete time series of one particular stock price be denoted by $\{S(t_i), i = 0, ..., n\}$ with $t_i - t_{i-1} = \delta$. The return process is defined by $\{X(t_i) = \frac{S(t_i)-S(t_{i-1})}{S(t_{i-1})}, i = 1, ..., n\}$. Let $\{V(t_i), i = 1, ..., n\}$ be the corresponding volume process, where $V(t_i)$ denotes the cumulative volume for the time period $(t_{i-1}\ \ t_i]$.

Mark the time point t_i 0 if $X(t_i)$ falls in a certain percentile of the returns, say the upper ten percentile, otherwise 1. The return process thus turns into a $0-1$ process with $m = n/10$ zeros. This $0-1$ process is divided into $m+1$ sections consisting of runs of 1s. $V(t_i)$ is marked similarly. The empirical distribution of the length of runs of 1s, the waiting time of hitting a certain percentile, which plays the key role in our analysis. The empirical distributions are considered for different stocks, different time units, different years from the markets.

Note that for any increasing function of $X(t_i)$ or $V(t_i)$, we still have exactly the same $0-1$ process. For example the logarithmic return $\log \frac{S(t_i)}{S(t_{i-1})}$ is just $\log(X(t_i) + 1)$.

Consider two distributions $F(x)$ and $G(x)$, and take $F(x)$ as the baseline distribution, then the ROC curve is defined as the curve of $(F(x), G(x))$ for all $x \in (-\infty, \infty)$. Mathematically this ROC curve of $F(x)$ and $G(x)$ is defined as

$$R(t|F \Rightarrow G) = G(F^{-1}(t))\ t \in [0, 1],$$

where $F^{-1}(t)$ is the quantile function corresponding to $F(x)$.

One may use the following two criteria to measure the closeness of these two distributions.
The ROC area:

$$\int_0^1 \mid G(F^{-1}(t)) - t \mid dt,$$

The Kolmogorov-Smirnov distance ($Sup - norm$):

$$Sup_x \mid F(x) - G(x) \mid .$$

We consider companies in S&P500 list which varies slightly each year. The tables in the next page give us a glimpse of an empirical invariance, using IBM as the base line, for each company we calculate the ROC area and the KS distance of the above mentioned empirical distribution w.r.t. that of IBM. Then calculate the mean, variance, and extremal values for each year.

We do find an empirical invariance for the real stock prices. And the outliers have financial implications. When the returns follow a Lévy process, we prove the invariance distribution being geometric. The invariance property for the fractional Brownian motion is yet to be proved. However both invariances are different to each other and are different from the one from the real data empirically.

An empirical invariance is also established for the volume. The theoretical counterpart is yet to be proposed. The relationship between the price and volume is under investigation.

IBM	1998		2005		2006		2006	
(5min)	0-10%	90-100%	0-10%	90-100%	0-10%	90-100%	0-5%	95-100%
ROC	0.0297 ± 0.0198	0.0234 ± 0.0186	0.0099 ± 0.0071	0.0105 ± 0.0071	0.0141 ± 0.0094	0.0134 ± 0.0095	0.0158 ± 0.0093	0.0148 ± 0.0108
	(0.0030, 0.2039)	(0.0022, 0.1993)	(0.0019, 0.0761)	(0.0020, 0.0712)	(0.0021, 0.1108)	(0.0017, 0.1041)	(0.0045, 0.0993)	(0.0037, 0.0982)
K-S	0.0587 ± 0.0333	0.0506 ± 0.0341	0.0237 ± 0.0123	0.0245 ± 0.0120	0.0307 ± 0.0166	0.0275 ± 0.0153	0.0386 ± 0.0164	0.0363 ± 0.0181
	(0.0106, 0.3109)	(0.0094, 0.3102)	(0.0072, 0.1140)	(0.0088, 0.1034)	(0.0080, 0.1847)	(0.0052, 0.1668)	(0.0135, 0.1634)	(0.0125, 0.1612)
(1min)			0-10%	90-100%	0-10%	90-100%	0-5%	95-100%
ROC			0.0203 ± 0.0149	0.0185 ± 0.0146	0.0203 ± 0.0159	0.0161 ± 0.0150	0.0197 ± 0.0169	0.0155 ± 0.0149
			(0.0018, 0.0915)	(0.0010, 0.0864)	(0.0009, 0.0930)	(0.0015, 0.0821)	(0.0015, 0.1231)	(0.0027, 0.1167)
K-S			0.0409 ± 0.0283	0.0365 ± 0.0287	0.0404 ± 0.0322	0.0324 ± 0.0296	0.0381 ± 0.0328	0.0323 ± 0.0282
			(0.0065, 0.1734)	(0.0040, 0.1637)	(0.0044, 0.1961)	(0.0051, 0.1947)	(0.0063, 0.2448)	(0.0080, 0.2340)

IBM	2001		2002	
(5min)	0-10%	90-100%	0-10%	90-100%
ROC	0.0393 ± 0.0199	0.0243 ± 0.0216	0.0182 ± 0.0096	0.0178 ± 0.0090
	(0.0053, 0.2521)	(0.0021, 0.2669)	(0.0030, 0.0822)	(0.0034, 0.0670)
K-S	0.0707 ± 0.0308	0.0487 ± 0.0339	0.0393 ± 0.0158	0.0402 ± 0.0169
	(0.0155, 0.4047)	(0.0073, 0.4166)	(0.0123, 0.1451)	(0.0110, 0.1412)
(1min)	0-10%	90-100%	0-10%	90-100%
ROC	0.0225 ± 0.0195	0.0196 ± 0.0191	0.0147 ± 0.0104	0.0134 ± 0.0098
	(0.0022, 0.2223)	(0.0029, 0.2425)	(0.0017, 0.1018)	(0.0026, 0.1004)
K-S	0.0436 ± 0.0307	0.0383 ± 0.0303	0.0316 ± 0.0209	0.0280 ± 0.0193
	(0.0062, 0.3611)	(0.0071, 0.3836)	(0.0057, 0.2152)	(0.0056, 0.2028)

2 Mathematical Framework and Discussions

For each stock the empirical distribution of the waiting time to hit the upper (and lower) ten percentile of the returns is considered. Most of the empirical distributions are close to each other under two different comparison criteria, ROC area and Kolmogorov-Smirnov distance. Comparisons are done across stocks, years, different time units. This may be regarded as an empirical invariance. IBM is used as the base line through most of our study with no particular reason. One may pick other base line for comparison.

We have analyzed the actual trade price data for 2006, 2005, 2002, 2001, 1998 and the cumulative volume data for each 30 seconds for 2005. A possible invariance of the correlation between price and volume is yet to be addressed. The analysis of attributes of the ask and bid prices seems very challenging, but unfortunately this data set is not available in WRDS.

We carry out a similar empirical analysis when the returns are finite sequence of i.i.d. random variables, e.g. from a Lévy process. The corresponding empirical distributions which are the same as those from finite sequence of exchangeable random variables converge completely to a geometric distribution. For the fractional Brownian motions we only have the empirical study.

More precisely, the stock price $S(t)$ follows

$$S(t) = S(0)\exp Z(t),$$

where $Z(t)$ is a Lévy process or

$$S(t) = S(0)\exp(\mu t - \frac{\sigma^2}{2}t^{2H} + \sigma B^H(t)),$$

where B^H is a fractional Brownian motion with parameter H.

A fractional Brownian motion with parameter H in (0, 1) is a continuous-time Gaussian process B^H starting at zero with mean zero and covariance function

$$E(B^H(s)B^H(t)) = \frac{1}{2}(|s|^{2H} + |t|^{2H} - |s-t|^{2H}).$$

For any non-overlapping intervals $(t_0, t_1)\cdots(t_{n-1}, t_n)$, $Z(t_1)-Z(t_0), \cdots, Z(t_n)-Z(t_{n-1})$ are independent. And the distribution of $Z(t) - Z(s)$ depends only on $t - s$.

Note that the empirical distributions derived from a Lévy process converge a.s. to a geometric distribution. This is our main theorem. Detailed proof is in [3]. This is a kind of law of large numbers. What are the corresponding Kolmogorov theorem (rate of convergence) and Donsker's theorem (central limit theorem)?

What is the corresponding limiting distribution for the fractional Brownian motion? Most importantly, what is that invariance in the real market and what are the dynamics behind this invariance financially and mathematically?

The entropy of the empirical distribution of the waiting time from the real data is smaller than that from the i.i.d. case. Very high reject rate is observed for the hypothesis testing of entropy. For the countable case with fixed mean the geometric distribution maximizes the entropy. It is reasonable that the entropy calculated from one-year data for each stock is smaller.

But for a small fixed $n$, the entropy of the empirical distribution of the waiting time from the i.i.d. returns is a random variable. What sort of optimization problem is it to justify our observation?

3 References

[1] Athreya, K. B. (1994) Entropy maximization, IMA Preprint 1231.

[2] Chang, Lo-Bin, Shu-Chun Chen, Fushing Hsieh, Chii-Ruey Hwang, Max Palmer (2008) An empirical invariance for the stock price, in preparation.

[3]Chang, Lo-Bin, Alok Goswami, Fushing Hsieh, Chii-Ruey Hwang (2008) An invariance property for the empirical distributions of occupancy problems with application to finance, manuscript.

[4]Fushing Hsieh, Chii-Ruey Hwang (2008) Statistical finance with high frequency data: Non-parametric volatility decoding and predictions, and signature-phase coherence among return, volume and trading number, first draft.

[5]Geman, Stuart(2008) Rare events in financial markets, the Ninth Annual Bahadur Memorial Lectures, May 5, 2008.

[6]Lowenstein, Roger (2000) *When Genius Failed: The Rise and Fall of Long-Term Capital Management*, Random House, NY.