Empirical Invariance in Stock Market and Related Problems

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

WSAF09
June 29 - July 3, 2009
Empirical Analysis

Introduction and Discussions

Mathematical Results

Remarks

References
This is a joint work with Lo-bin Chang, Shu-Chun Chen, Alok Goswami, Fushing Hsieh, Max Palmer. The raw data consists of the actual trade price and volume of the intraday transactions data (trades and quotes) of companies in S&P500 list from 1998 to 2007 and part of 2008. The return process is analyzed first at five-minute, one-minute, and 30-second intervals for a whole year.
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 did try OMC (OMNICOM GP INC) which has a common stock price. The result is the same.

June 26, 2009: IBM (105.68), OMC (31.78); September 19, 2008: IBM (118.85), OMC (41.90).
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
\[ X(t_i) = \frac{S(t_i) - S(t_{i-1})}{S(t_{i-1})}, \quad i = 1, ..., n. \]
Let \( \{ V(t_i), i = 1, ..., n \} \) and \( \{ U(t_i), i = 1, ..., n \} \) be the corresponding volume process and frequency process, where \( V(t_i) \) and \( U(t_i) \) denote the cumulative volume and number of transactions (frequency) for the time period \( (t_{i-1}, t_i] \).
Mark the time point \( t_i \) 1 if \( X(t_i) \) falls in a certain percentile of the returns, say the upper ten percentile, otherwise 0. The return process thus turns into a \( 0 - 1 \) process with \( m = n/10 \) ones. This \( 0 - 1 \) process is divided into \( m + 1 \) sections consisting of runs of 0s. \( V(t_i) \) and \( U(t_i) \) are marked similarly. The empirical distribution of the length of runs of 0s, the waiting time of hitting a certain percentile, 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 $V(t_i)$, $U(t_i)$), we still have exactly the same $0 \rightarrow 1$ process. For example the logarithmic return $\log \left( \frac{S(t_i)}{S(t_{i-1})} \right)$ is just $\log(X(t_i) + 1)$. 
One may use the following two criteria to measure the closeness of two distributions.

The ROC area:

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

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

$$Sup_x | F(x) - G(x) |.$$
Entropy: $- \sum p_i \log p_i$.

Volatile period is defined hierarchically: if the length of runs of 0s falls in say upper ten percentile, denote that period $1^*$, otherwise $0^*$; repeat the same procedure for the length of runs of $0^*$s and denote the period in the upper ten percentile $1^@$. We may regard $1^@$ the volatile period.

Graphs and tables.
We study 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.

We do find an empirical invariance for the real stock prices. What are the mathematics and financial dynamics driving this invariance are still not clear. And 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.
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 $H$. 
However both invariances are different to each other and are different from the one from the real data empirically.

Empirical invariance is also observed for the volume process and the frequency process. The theoretical counterpart is yet to be proposed. The volatile periods of the return, the volume and the frequency are highly correlated.
A Lévy process is a continuous-time stochastic process $Z(t)$ with stationary independent increments.

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

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

$H = \frac{1}{2}$ is the Brownian motion.
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 \(t - s\).

These two processes are generalizations of the Brownian motion, one keeps the stationary independent increments and the other one the stationary Gaussian increments.
For each stock the empirical distribution of the waiting time to hit the upper (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.
We carry out a similar analysis when the returns are finite sequence of i.i.d. random variables, e.g. from 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 $G(x)$. This is a law of large numbers, but the limit is already an invariance. How about the corresponding Kolmogorov theorem and Donsker’s theorem?

$$\sqrt{n} \sup_x | P_n(x) - G(x) | \quad \text{converges weakly to}$$

$$\sup_x | B(G(x)) | ,$$

$$\sqrt{n}(P_n(x) - G(x)) \quad \text{converges weakly to} \quad B(G(x)),$$

where $B(t)$ is a Brownian bridge $(W(t) - tW(1))$ in $[0, 1]$?
For the fractional Brownian motions we only have the empirical study.

The invariances from the models are different from the market one.

It is interesting to observe that the stock prices of most of the outliers have specific financial meaning.
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 fixed small $n$, 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?

Other percentiles, the overlapping of the time points falling say in the low 10 percentile of the returns of each stock with those of IBM and comparisons across different years, different percentiles are also studied.
Consider $n$ objects arranged in a row. Suppose $m$ of the objects are selected at random, with each of the $\binom{n}{m}$ possible selections having the same probability. If we describe a particular selection by dubbing each selected object as a 1 and each unselected object as a 0, each selection gives an $n$-long binary sequence with $m$ many 1s and $n - m$ many 0s.
Let now $Y_1^n, \ldots, Y_{m+1}^n$ denote the lengths of the $m + 1$ runs of 0s thus obtained. We include the 0-runs, possibly of zero length, before the first 1 and after the last 1. This gives a sequence of $m + 1$ non-negative integer valued random variables, which are clearly not independent, because $Y_1^n + \cdots + Y_{m+1}^n = n - m$. 
Thus, the possible values of the random vector $(Y_1^n, \ldots, Y_{m+1}^n)$ are vectors $(l_1, \ldots, l_{m+1})$ of non-negative integers with $l_1 + \cdots + l_{m+1} = n - m$ and, for each such vector,

$$P(Y_1^n = l_1, \ldots, Y_{m+1}^n = l_{m+1}) = \frac{1}{\binom{n}{m}},$$

since the event on the rhs corresponds precisely to selecting the $(l_1 + 1)$th, $(l_1 + l_2 + 2)$th, $\ldots$, and the $(l_1 + \cdots + l_m + m)$th objects among the $n$ objects.
It is an easy consequence of this that the random variables $Y^n_1, \ldots, Y^n_{m+1}$ are exchangeable. This is because, for any permutation $\pi$ on $\{1, \ldots, m+1\}$, the event
\[
\{ Y^n_{\pi(1)} = l_1, \ldots, Y^n_{\pi(m+1)} = l_{m+1} \}
\]
is the same as
\[
\{ Y^n_1 = l_{\pi^{-1}(1)}, \ldots, Y^n_{m+1} = l_{\pi^{-1}(m+1)} \},
\]
and this last event has the same probability as the event
\[
\{ Y^n_1 = l_1, \ldots, Y^n_{m+1} = l_{m+1} \},
\]
both equal to $1/\binom{n}{m}$. 
If now $n \to \infty$, we get a triangular array where each row consists of a finite sequence of random variables that are exchangeable but not independent. We show that if $n$ and $m$ both go to infinity in such a way that $m/n \to p$ for some $p \in (0, 1)$, then the random variables become asymptotically independent. Moreover, the limiting common distribution is geometric with parameter $p$. 
Lemma 1
If $n \to \infty$ and $m \to \infty$ in such a way that $m/n \to p \in (0, 1)$, then, for any $k \geq 1$,

$$(Y_1^n, \ldots, Y_k^n) \xrightarrow{d} (Y_1, \ldots, Y_k),$$

where $Y_1, \ldots, Y_k$ are independent and identically distributed random variables having the geometric distribution with parameter $p$. 
For each $n$, we consider the probability histogram generated by the random variables $Y_1^n, \ldots, Y_{m+1}^n$. We get a (random) probability distribution on non-negative integers, given by the probability mass functions

$$
\theta_n(l)(\omega) = \frac{1}{m+1} \sum_{i=1}^{m+1} 1\{Y_i^n(\omega) = l\} , \ l = 0, 1, \ldots .
$$
The next theorem says that these probability distributions converge, with probability 1, to the geometric distribution with parameter $p$. In other words, the empirical distribution from each row of the triangular array converges almost surely to the geometric distribution.

**Theorem 2**

If $n \to \infty$ and $m \to \infty$ in such a way that $m/n \to p \in (0, 1)$, then,

$$P \left( \lim_{n \to \infty} \theta_n(l) = p(1 - p)^l, \ l = 0, 1, \ldots \right) = 1.$$
Using Scheffe’s Theorem and the fact that all the distributions involved are concentrated on non-negative integers, and denoting the empirical distribution in the $n$th row by $P_n$, then one has

**Corollary**

$P_n$ converges, with probability 1, to the geometric distribution $G$ with parameter $p$, in total variation as well as in Kolmogorov distance. Moreover, the convergence $\theta_n(l) \to p(1 - p)^l$ holds uniformly in $l$ with probability 1.
Given prices of a stock at equal intervals of time, if we consider the times of occurrences of extreme values for the returns over successive time intervals, we end up selecting a certain subset of a fixed proportion from the set of all time points. Our result simply says that if under the assumed model for a stock price, the returns over successive time intervals have an exchangeable joint distribution, then all selections are equally likely. This should be obvious. We elaborate it only for the sake of completeness.
Let $\alpha, \beta \geq 0$ with $0 < \alpha + \beta < 1$. From a data set consisting $n$ points $(x_1, x_2, \ldots, x_n)$, we want to choose those that form the lower $100\alpha$-percentile and those that form the upper $100\beta$-percentile. To avoid trivialities, let us assume the size $n$ of the data set is strictly larger than $n > (1 - \alpha - \beta)^{-1}$. In case the data points are all distinct, we have an unambiguous choice. Indeed, we may arrange the data points in the (strictly) decreasing order as $x(1) < x(2) < \cdots < x(n)$. 
If now \( k \) and \( l \) are integers satisfying

\[
\frac{k}{n} \leq \alpha < \frac{k + 1}{n} \leq \frac{l - 1}{n} < 1 - \beta \leq \frac{l}{n},
\]

then \((x_1, \ldots, x_k)\) will form the lower 100\( \alpha \)-percentile and \((x_l, \ldots, x_n)\) will form the upper 100\( \beta \)-percentile. In case the data points are not all distinct, we may have more than one possible choices for the \( k \) among the \( n \) data points that form the lower 100\( \alpha \)-percentile or for the \( n - l + 1 \) that form the upper 100\( \beta \)-percentile.
In such cases, our prescription is to pick one among the possible choices with equal probability for each. Thus, we will always end up selecting exactly $k + n - l + 1$ from the $n$ data points with $k$ of them forming the lower $100\alpha$-percentile and remaining $n - l + 1$ forming the upper $100\beta$-percentile. The next theorem considers the case when the data points consist of $n$ random variables with an exchangeable joint distribution. This result provides the required connecting link between stock price data and the limiting results in the earlier theorems.
Theorem 3
If $X_1, \ldots, X_n$ are random variables with an exchangeable joint distribution, then any one of the $\binom{n}{k+n-l+1}$ possible choices can occur with equal probability as the set of points constituting the lower $100\alpha$- and upper $100\beta$-percentiles.
Sketch of Proofs

Sketch of Proof of Lemma 1: (Feller, Volume I) We have to prove that, for every choice of \( k \) non-negative integers \( l_1, \ldots, l_k \),

\[
P(Y_1^n = l_1, \ldots, Y^n_k = l_k) \longrightarrow \prod_{i=1}^{k} [p(1 - p)^{l_i}].
\]

The left hand side clearly equals

\[
\binom{n - l_1 - \cdots - l_k - k}{m - k} / \binom{n}{m}.
\]
Denoting  
\[ s_1 = l_1 + 1, s_2 = l_1 + l_2 + 2, \ldots, s_k = l_1 + \cdots + l_k + k, \]
this last expression can be written as  
\[ \prod_{i=1}^{k} \frac{(n - s_i)}{(m - i)} \bigg/ \frac{(n - s_i + l_i + 1)}{(m - i + 1)} \]. It, therefore, suffices for us to prove that, for each \( i = 1, 2, \ldots, k, \)

\[ \frac{(n - s_i)}{(m - i)} \bigg/ \frac{(n - s_i + l_i + 1)}{(m - i + 1)} \longrightarrow p(1 - p)^l_i. \]
Sketch of Proof of Theorem 2: First of all, it is enough to prove that for each \( l = 0, 1, 2 \ldots \),

\[
P \left( \lim_{n \to \infty} \theta_n(l) = p(1 - p)^l \right) = 1.
\]

In fact, it is enough to do this only for \( l = 1, 2, \ldots \), since the case \( l = 0 \) will then automatically follow.
Secondly, since \( E(1_{\{Y^n_i = l\}}) = P(Y^n_i = l) \longrightarrow p(1 - p)^l \) by

\[
\frac{1}{m + 1} \sum_{i=1}^{m+1} \left[ 1_{\{Y^n_i = l\}} - E(1_{\{Y^n_i = l\}}) \right] \longrightarrow 0, \text{ a.s.}
\]

This follows from

\[
\sum_{n=1}^{\infty} \frac{1}{(m + 1)^4} E \left| \sum_{i=1}^{m+1} \left[ 1_{\{Y^n_i = l\}} - E(1_{\{Y^n_i = l\}}) \right] \right|^{4} < \infty.
\]
To compute the fourth moments in the summands, let
\[ Z^n_i = 1_{\{Y^n_i = l\}}. \] Then \( E(Z^n_i) = P(Y^n_i = l) \), which by the exchangeability of the random variables \( \{ Y^n_i, 1 \leq i \leq m + 1 \} \) is also equal to \( P(Y^n_1 = l) = p_1 \), say. Further, exchangeability of \( \{ Y^n_i, 1 \leq i \leq m + 1 \} \) implies exchangeability of \( \{ Z^n_i, 1 \leq i \leq m + 1 \} \) as well,

\[
E \left| \sum_{i=1}^{m+1} \left[ 1_{\{Y^n_i = l\}} - E(1_{\{Y^n_i = l\}}) \right] \right|^4
\]
Empirical Invariance in Stock Market and Related Problems

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

Empirical Analysis

Introduction and Discussions

Mathematical Results

Remarks

References

\[
E \left| \sum_{i=1}^{m+1} (Z_i^n - p_1) \right|^4 = S_1 + S_2 + S_3 + S_4.
\]

\[
S_1 = (m + 1)E(Z_1^n - p_1)^4,
\]

\[
S_2 = m(m + 1)E \left[ (Z_1^n - p_1)^3(Z_2^n - p_1) \right]
\]

\[
+ \frac{m(m + 1)}{2}E \left[ (Z_1^n - p_1)^2(Z_2^n - p_1)^2 \right],
\]
$$S_3 = \frac{m(m - 1)(m + 1)}{2} E((Z^n_1 - p_1)^2(Z^n_2 - p_1)(Z^n_3 - p_1)),$$

$$S_4 = \binom{m + 1}{4} E((Z^n_1 - p_1)(Z^n_2 - p_1)(Z^n_3 - p_1)(Z^n_4 - p_1)).$$
Thus the series becomes

\[ \sum_{n=1}^{\infty} \frac{1}{(m + 1)^4} [S_1 + S_2 + S_3 + S_4]. \]

Noting that \( |Z_i^n - p_1| \leq 1 \) and \( m \sim np \) with \( p \in (0, 1) \), it is clear that both the series

\[ \sum_{n=1}^{\infty} \frac{S_1}{(m + 1)^4} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{S_2}{(m + 1)^4} \]

are convergent.
To now show convergence of the two series

\[ \sum_{n=1}^{\infty} \frac{S_3}{(m + 1)^4} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{S_4}{(m + 1)^4}, \]

we introduce some notations. In analogy with the notation

\[ p_1 = P(Z_1^n = 1) = p_1, \] let us denote

\[ p_0 = P(Z_i^n = 0) = 1 - p_1. \] Similarly, let

\[ p_{11} = P(Z_1^n = 1, Z_2^n = 1), \quad p_{10} = P(Z_1^n = 1, Z_2^n = 0), \]
\[ p_{01} = P(Z_1^n = 0, Z_2^n = 1), \quad p_{00} = P(Z_1^n = 0, Z_2^n = 0). \]
One can express the last three in terms of $p_1$ and $p_{11}$. Indeed,

$$p_{10} = p_{01} = p_1 - p_{11},$$

and

$$p_{00} = 1 - 2p_1 + p_{11}.$$

We can likewise define $p_{ijk}$ and $p_{ijkh}$ for all $i, j, k, h \in \{0, 1\}$ and express all the $p_{ijk}$ in terms of $p_1, p_{11}, p_{111}$ and all the $p_{ijkh}$ in terms of $p_1, p_{11}, p_{111}, p_{1111}$. 
Indeed,

\[ p_{110} = p_{101} = p_{011} = p_{11} - p_{111}, \]
\[ p_{100} = p_{010} = p_{001} = p_{1} - 2p_{11} + p_{111}, \]
\[ p_{000} = 1 - 3p_{1} + 3p_{11} - p_{111}; \]
\[ p_{1110} = p_{1101} = p_{1011} = p_{0111} = p_{111} - p_{1111}, \]
\[ p_{1100} = p_{1010} = p_{1001} = p_{0011} = p_{0101} = p_{0110} \]
\[ = p_{11} - 2p_{111} + p_{1111} \]
\[ p_{1000} = p_{0100} = p_{0010} = p_{0001} \]
\[ = p_{1} - 3p_{11} + 3p_{111} - p_{1111}, \]
\[ p_{0000} = 1 - 4p_{1} + 6p_{11} - 4p_{111} + p_{1111}. \]
In the expression

\[ E \left[ (Z_1^n - p_1)^2 (Z_2^n - p_1)(Z_3^n - p_1) \right] \]

\[ = \sum_{i,j,k \in \{0,1\}} (i - p_1)^2 (j - p_1)(k - p_1) p_{ijk}, \]

if one uses the above formulas for the \( p_{ijk} \) and simplifies, one gets

\[ E \left[ (Z_1^n - p_1)^2 (Z_2^n - p_1)(Z_3^n - p_1) \right] \]

\[ = p_1^3 - 3p_1^4 - 2p_1p_{11} + 5p_1^2 p_{11} + p_{111} - 2p_1 p_{111}. \]
\[ E \left( (Z_1^n - p_1)(Z_2^n - p_1)(Z_3^n - p_1)(Z_4^n - p_1) \right) = -3p_1^4 + 6p_1^2p_{11} - 4p_1p_{111} + p_{1111}. \]

Indeed, these simplifications were achieved by using symbolic simplification program in Mathematica. Then a lengthy and delicate asymptotic analysis leads to the proof.
1. To maximize the entropy of discrete probabilities \( \{p_k\}_{k \geq 0} \) with \( p_0 > 0 \) and a fixed expectation,
\[
\log p_k / p_0 = \lambda k.
\]
For the infinite case, the maximizer is the geometric distribution, the invariance in Theorem 2.
Consider $n$ objects arranged in a row. Suppose $m$ of the objects are selected according to some probability law. If we describe a particular selection by dubbing each selected object as a 1 and each unselected object as a 0, each selection gives an $n$-long binary sequence with $m$ many 1s and $n - m$ many 0s. In our study, we have $n$ returns, objects are selected if they fall say in the upper ten percentile of the returns.
Let now $Y_1^n, \ldots, Y_{m+1}^n$ denote the lengths of the $m+1$ runs of 0s thus obtained. We include the 0-runs, possibly of zero length, before the first 1 and after the last 1. This gives a sequence of $m+1$ non-negative integer valued random variables, which are clearly not independent, because $Y_1^n + \cdots + Y_{m+1}^n = n - m$. 
2. For finite $n$, the random vector $(Y_1^n, \ldots, Y_{m+1}^n)$ are vectors $(l_1, \ldots, l_{m+1})$ of non-negative integers with $l_1 + \cdots + l_{m+1} = n - m$. The corresponding empirical distribution always has expectation $(n - m)/(m + 1)$ for any joint distribution. Apparently the case $P(Y_1^n = l_1, \ldots, Y_{m+1}^n = l_{m+1}) = \frac{1}{\binom{n}{m}}$ should enjoy the maximum entropy property in some suitable formulation, but how to formulate it?
3. On the other hand, if the logarithm of the stock price is a fractional Brownian motion (FBM) and look at the returns at equally spaced n time points, these are stationary. Simulation shows that the limit is nonrandom, but what is it?

Of course one may consider the discrete time versions directly, namely the returns have stationary independent increments or stationary Gaussian (from FBM).
4. The empirical distribution $P_n$ of $(Y_1^n, \ldots, Y_{m+1}^n)$ with $P(Y_1^n = l_1, \ldots, Y_{m+1}^n = l_{m+1}) = \frac{1}{\binom{n}{m}}$ converges to the geometric distribution $G$ a.s. This is already an invariant theorem. What are the corresponding Kolmogorov theorem (rate of convergence) and Donsker’s theorem (central limit theorem)? i.e.

$$\sqrt{n} \text{Sup}_x | P_n(x) - G(x) |$$ converges weakly to

$$\text{Sup}_x | B(G(x)) |,$$

$$\sqrt{n}(P_n(x) - G(x))$$ converges weakly to $B(G(x))$, where $B(t)$ is a Brownian bridge $(W(t) - tW(1))$ in $[0, 1]$?


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

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

But what are those invariants exactly?
Thank You!