Stochastic system: a study of three examples

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

July 14, 2007
Beijing Normal University

Exploring stock dynamics via hierarchical segmentation:An empirical invariance

Nyström method and large random matrices

Spectral gap of antisymmetrically perturbed Laplacian on N-torus

US Market: IBM, Ford, AA, ABMD, SPX.

Japanese Market: Toyota, Takeda, Canon, Mizuho, NTTDoCoMo.

Taiwanese Market: Formosa Plastics, United Microelectronics, Cathay Financial Holdings, Taiwan Semi-conductor, Mega Financial Holdings.

Most data consist of 2-year stock prices at 5-minute intervals. SPX is for 8 years, AA is tic-to-tic data for 4 years.

Brownian motion (Black-Scholes Model), i.i.d. uniform r.v.'s.

We look into the distributions of duration of the collection of volatile periods of different stocks and use ROC curve to compare the distributions. We discover that many US stocks share very similar such distribution shape. We call this an invariance property of stock dynamics. This invariance property is also seen in the Japanese market. In a computer simulation study, the simulated Black-Scholes model (or Garch model) based stock prices time series shows a different 'invariance'.

Receiver Operating Characteristic (ROC)

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

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

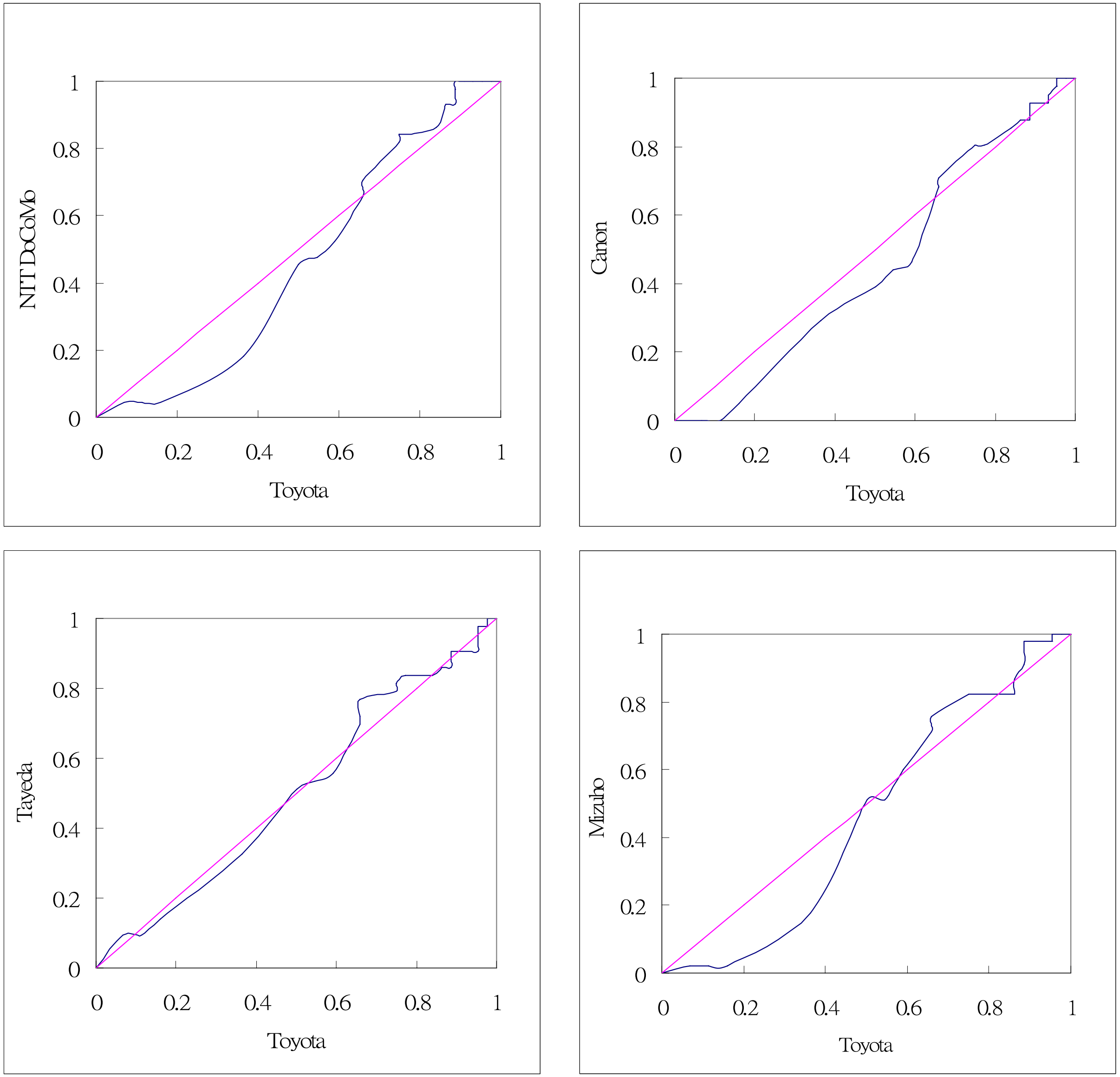
where $F^{-1}(w)$ is the quantile function corresponding to $F(x)$.

Note that this is a model-free approach. The invariance property of the stock dynamics may come from the 'herding phenomenon' but not the market structure. We do not know how to formulate the limiting laws.

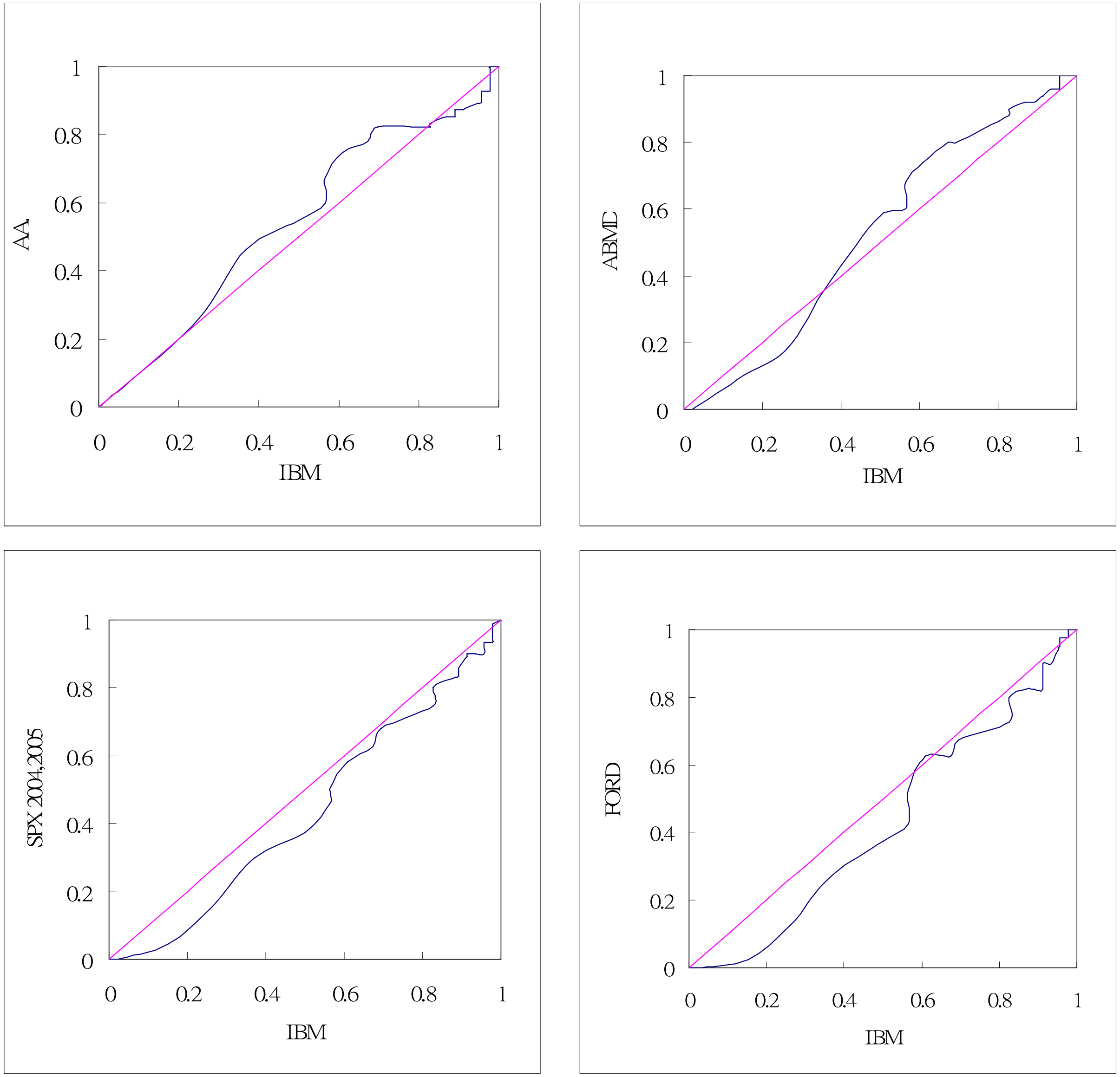
This is an ongoing joint work with Fushing Hsieh, Lo-Bin Chang and Ritu Sen, plus Max Palmer of Stuart Frankel & Co and FlexTrade Systems.

The hierarchical segmentation idea came from our previous study: Hsieh, H, Lee, Lan, Horng (2006) Testing and mapping non-stationarity in animal behavioral processes: A case study on an individual female bean weevil, J. Theoretical Biology,238, Issue 4, 805-816.

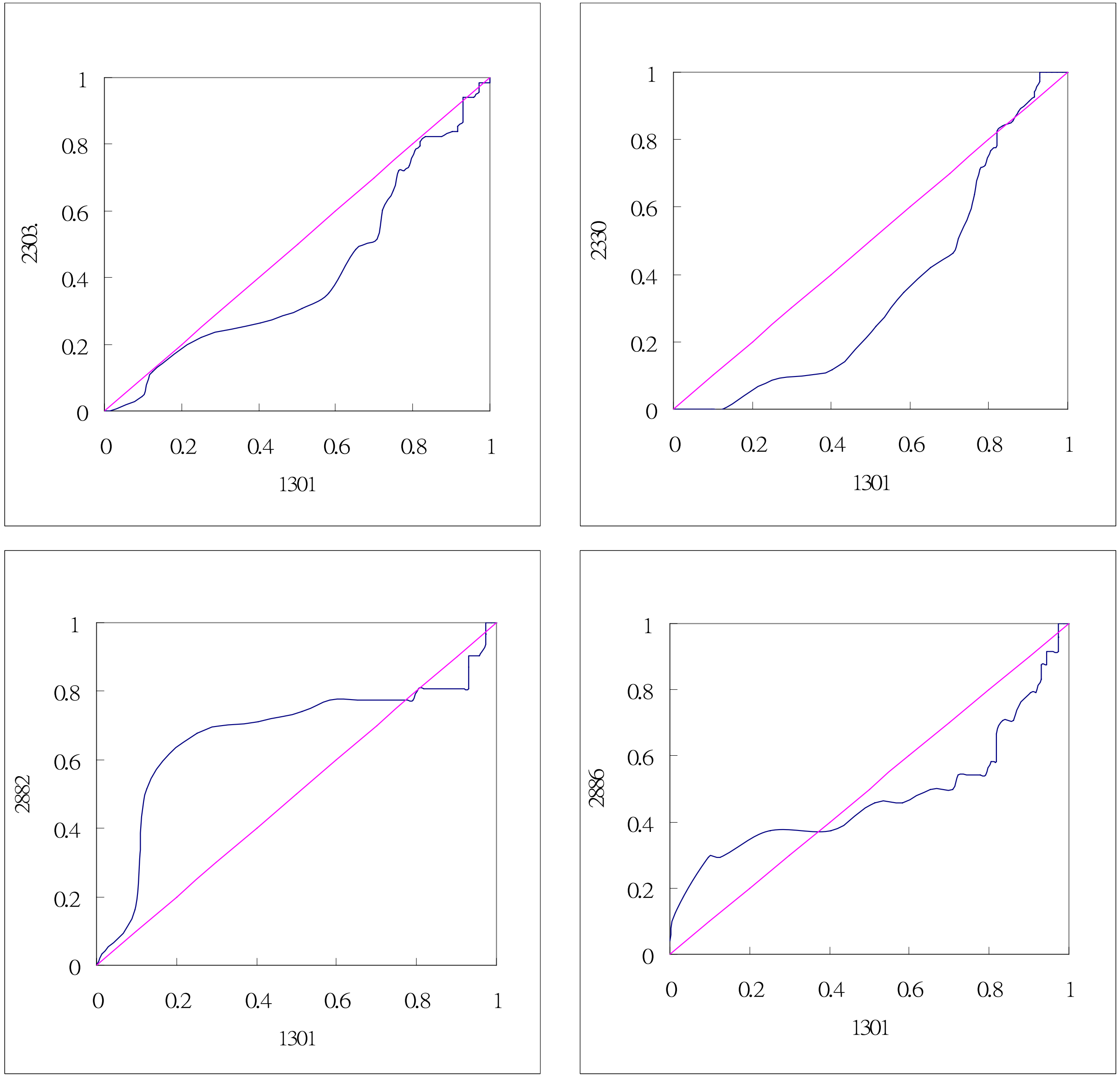
First we take a look at figures.



ROC Curve of Japanese stocks



ROC Curve of US stocks



ROC Curve of Taiwanese stocks,

1301: Formosa Plastics Corporation, 2303:United Microelectronics Corporation, 2882:Cathay Financial Holdings, 2330:Taiwan Semi-conductor Manufacturing Company, 2886:Mega Financial Holdings

Hierarchical segmentation

Let the discrete time series of one particular stock price be denoted by $\{S(t_i), i = 0, ..., n\}$ with $t_i - t_{i-1} = \delta$, with the return process defined by $\{X(t_i) = \frac{S(t_i)-S(t_{i-1})}{S(t_{i-1})}, i = 1, ..., n\}$.

1. [Defining extreme event:]Consider the cumulative histogram H_0 of the time series $\{X(t_i), i = 1, ..., n\}$. Let $H_0^{-1}(p_1)$ and $H_0^{-1}(p_2)$ be the p_1-th and p_2-th sample quantiles of H_0 with a chosen $p_1 \geq 0.9$ and $p_2 \leq 0.1$.

We defined that an extreme event is observed at time t_i if $X(t_i) \geq H_0^{-1}(p_1)$ or $X(t_i) \leq H_0^{-1}(p_2)$. Such an extreme event is coded with symbol 0. While an $X(t_i)$ falling in the non-extreme event category is coded with code 1. According to this symbolic coding scheme, the discrete time series $\{X(t_i), i = 1, ..., n\}$ is transformed into a 0-1 digital string and is denoted by $C_0 = \{c_1, c_2, ...c_n\}$.

2. [Recurrence distribution of extreme event:]Based on digital string C_0 we can approximate the recurrence distribution of extreme events (code 0's) by first constructing a sequence of spacing (i.e. counting the number of 1's) between successive extreme events (code 0's), and denote this sequence as $S_* = \{s_1, s_2, ...s_{n^*}\}$.

Then its corresponding cumulative histogram as an empirical version of recurrence distribution of extreme events is denoted by H_*. Upon a preselected constant $0.5 < p^* < 1$,say 0.9, the sequence S_* is further coded into a $0^* - 1^*$ symbolic string in the following coding scheme: s_i is coded with symbol 0^* if $s_i \geq H_*^{-1}(p^*)$, otherwise is coded by symbolic code 1^*.

It is clear that the symbolic code 0^* marks a relative lengthy consecutive $1-$codes segment on digital string C_0. Correspondingly 0^* marks a segment on the time series $X(t)$ that is "starving" of extreme event. The resultant symbolic code sequence is denoted as C_*.

3. [Segment of aggregating extreme events:]Upon the sequence C_*, the sequence spacing of 0^* codes is in turn denoted as $S_{@} = \{s_1^*, s_2^*, ...s_{n^{@}}^*\}$, and its corresponding cumulative histogram $H_{@}$. With a preselected constant, again say $p^{@} = 0.9$, the sequence $S_{@}$ is further coded into a $0^{@} - 1^{@}$ symbolic string: s_i^* is coded with symbol $0^{@}$ if $s_i^* \geq H_{@}^{-1}(p^{@})$, otherwise $1^{@}$. The resultant symbolic code sequence is denoted as $C_{@}$.

Characteristically symbolic code $0^{@}$ marks a segment on S_* in that contains clustering of 1^* codes, which corresponding to a segment of a cluster of consecutive relative-small s_i's on S_*. In this manner code $0^{@}$ maps out a segment of aggregating code 0 on C_0, that is, a segment of aggregating extreme events on time series $X(t)$.

Let $D_{@}(S)$ be the empirical distribution of the length of segments (on the time scale of time series $S(t)$) corresponding to all symbolic $0^{@}$ codes in $C_{@}$. We discuss an empirical invariance among a collection of distributions $D_{@}(S)$ in stock markets.

Invariance property in volatile period via ROC curve.

We demonstrate the invariant phenomenon among a collection of distributions $D_{©}(S)$ within the US and Japanese stock markets. However this invariance property **may in general not be** observed crossing the market boundary, or even within a smaller market, such as, the Taiwanese one. Of course there are exceptions.

Mathematically speaking, only two identical distributions could possibly give rise to the straight diagonal line segment as their ROC curve. In real world such special ROC curve is not possible to be seen out of two empirical distributions of $D_{@}(S)$ of two real stocks. For the purpose of exploring possible empirical patterns underlying stock dynamics, we demonstrate the following empirical evidences of invariance by raw eyesight visualization, that is, we restrain ourself from making a precise definition of invariance.

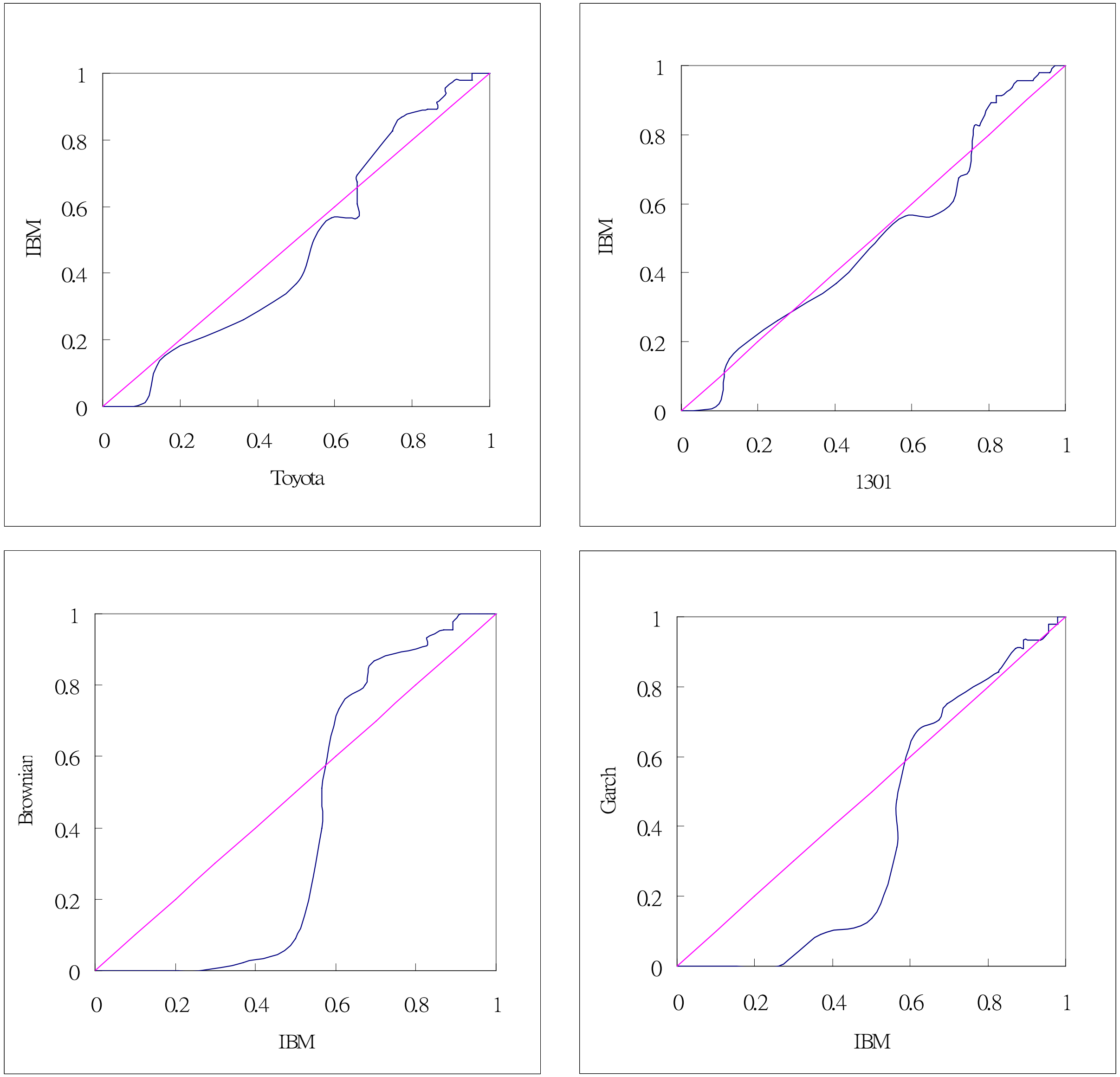
$$dS_t = aS_t dt + \sigma S_t dW_t,$$

$$\frac{S(t_{i+1}) - S(t_i)}{S(t_i)} = a\delta + \sigma(W(t_{i+1} - W(t_i)).$$

We have i.i.d. normal r.v.'s.

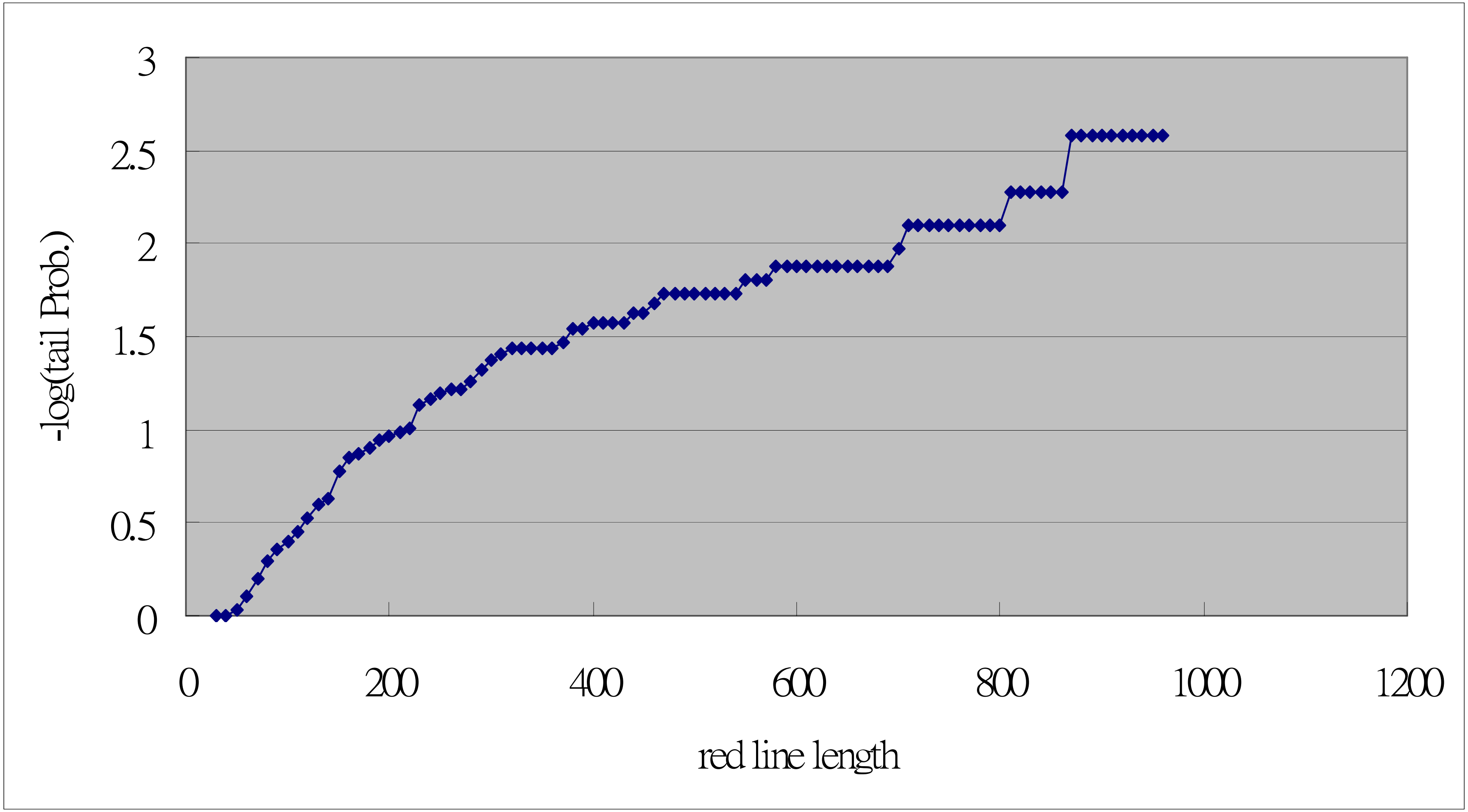
Using the same coding scheme for simulated data from i.i.d. normal (and uniform) r.v.'s. The resulting empirical distributions of the length of segments corresponding to all symbolic $0^{@}$ codes in $C_{@}$ seem close to deterministic.

What are possible formulations for invariance laws?

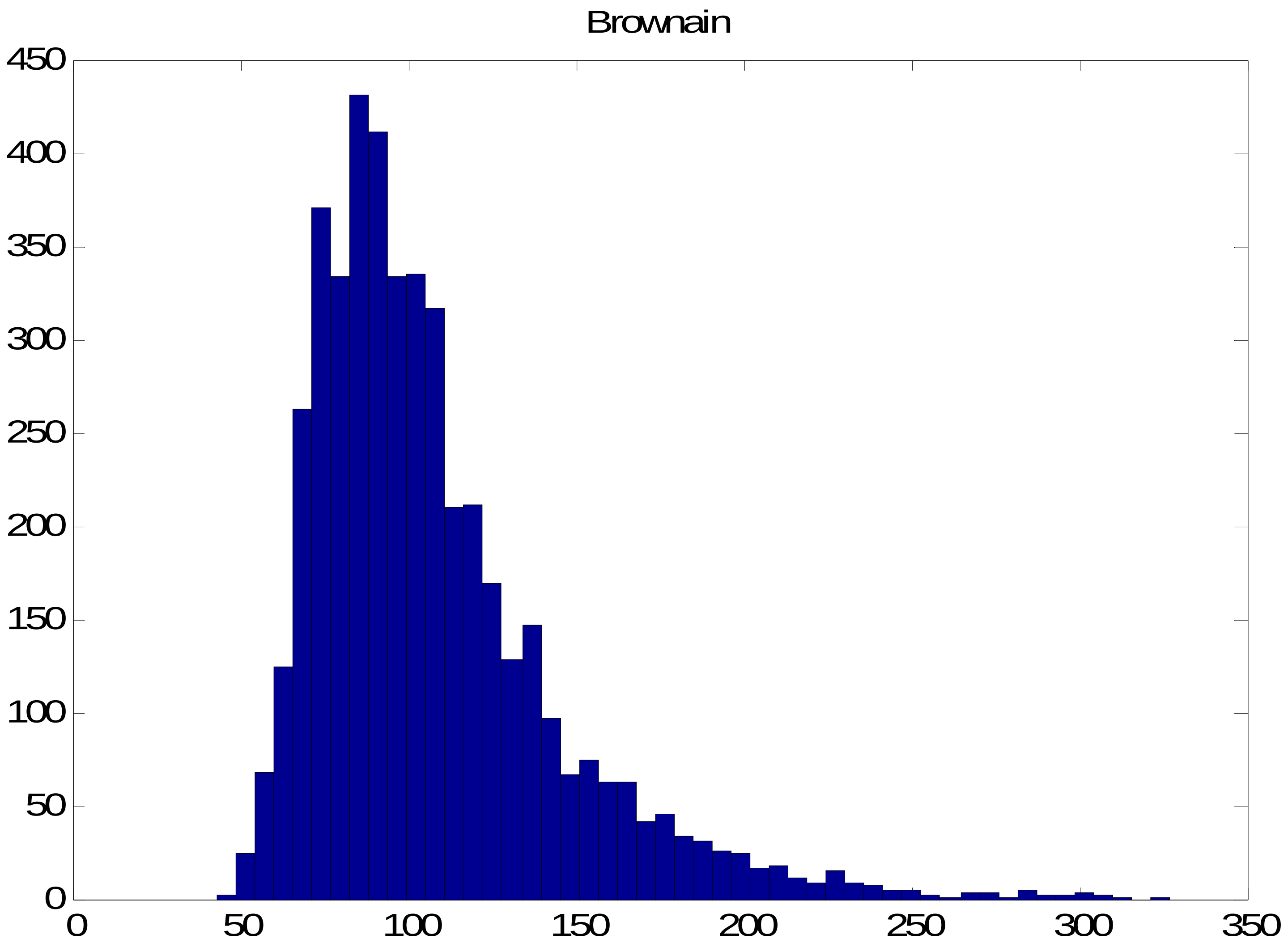


ROC Curve of IBM, Toyota, 1301, Brownian, and Garch model

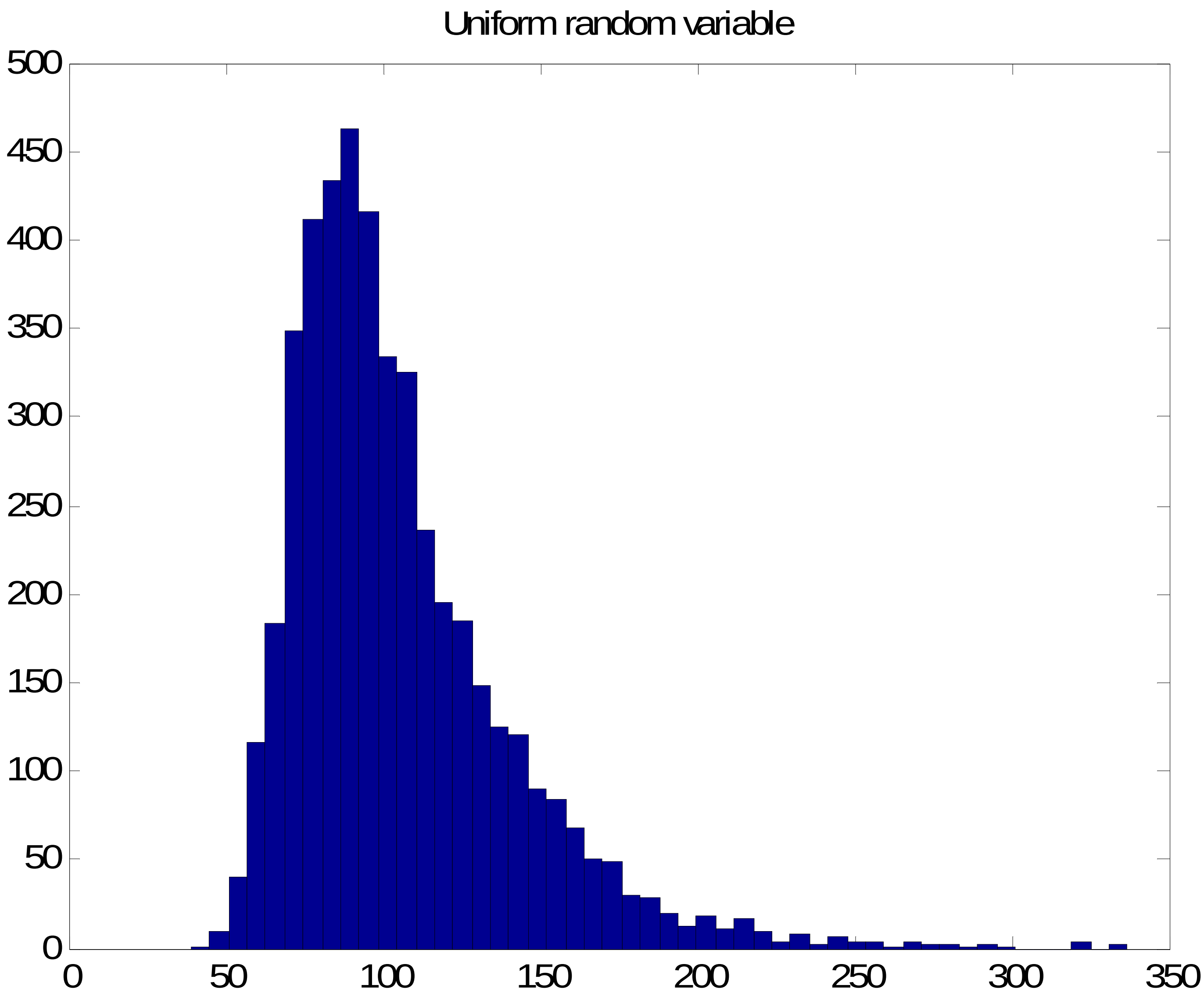
SPX

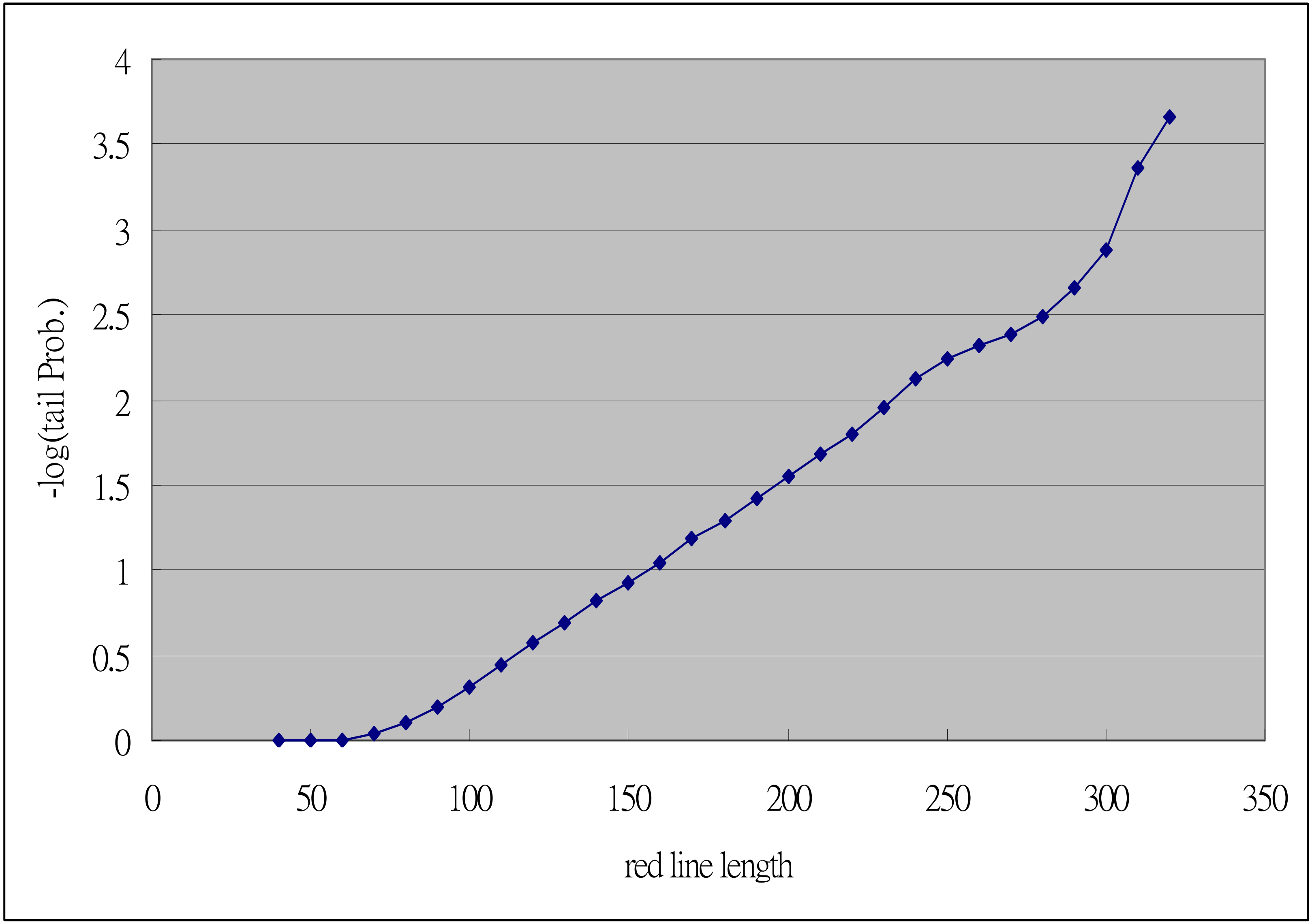


3
2.5
2
1.5
1
0.5
0
0
200
400
600
800
1000
1200
-log(tail Prob.)
red line length

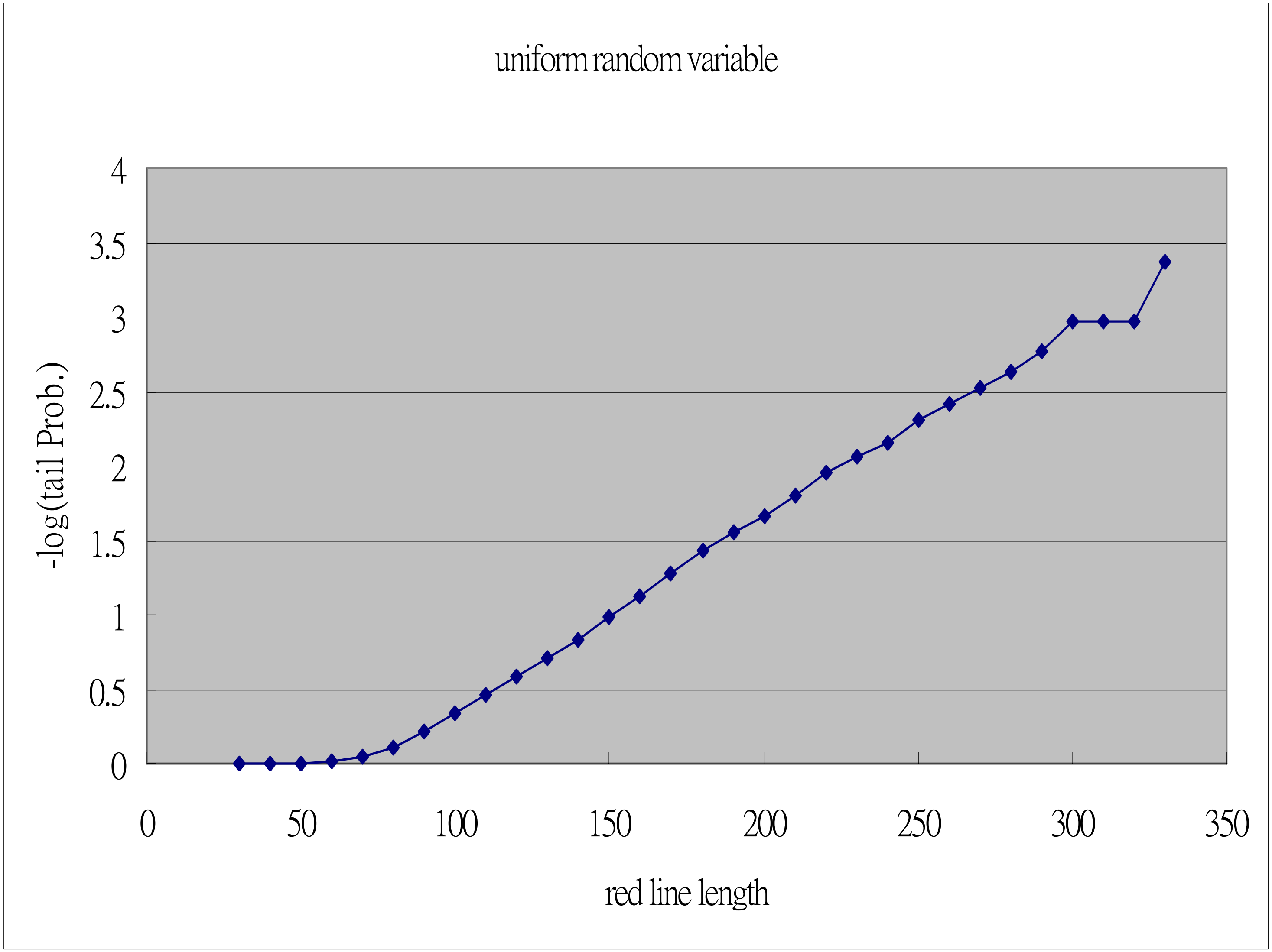


Brownain
450
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350


Uniform random variable
0
50
100
150
200
250
300
350
400
450
500
0
50
100
150
200
250
300
350

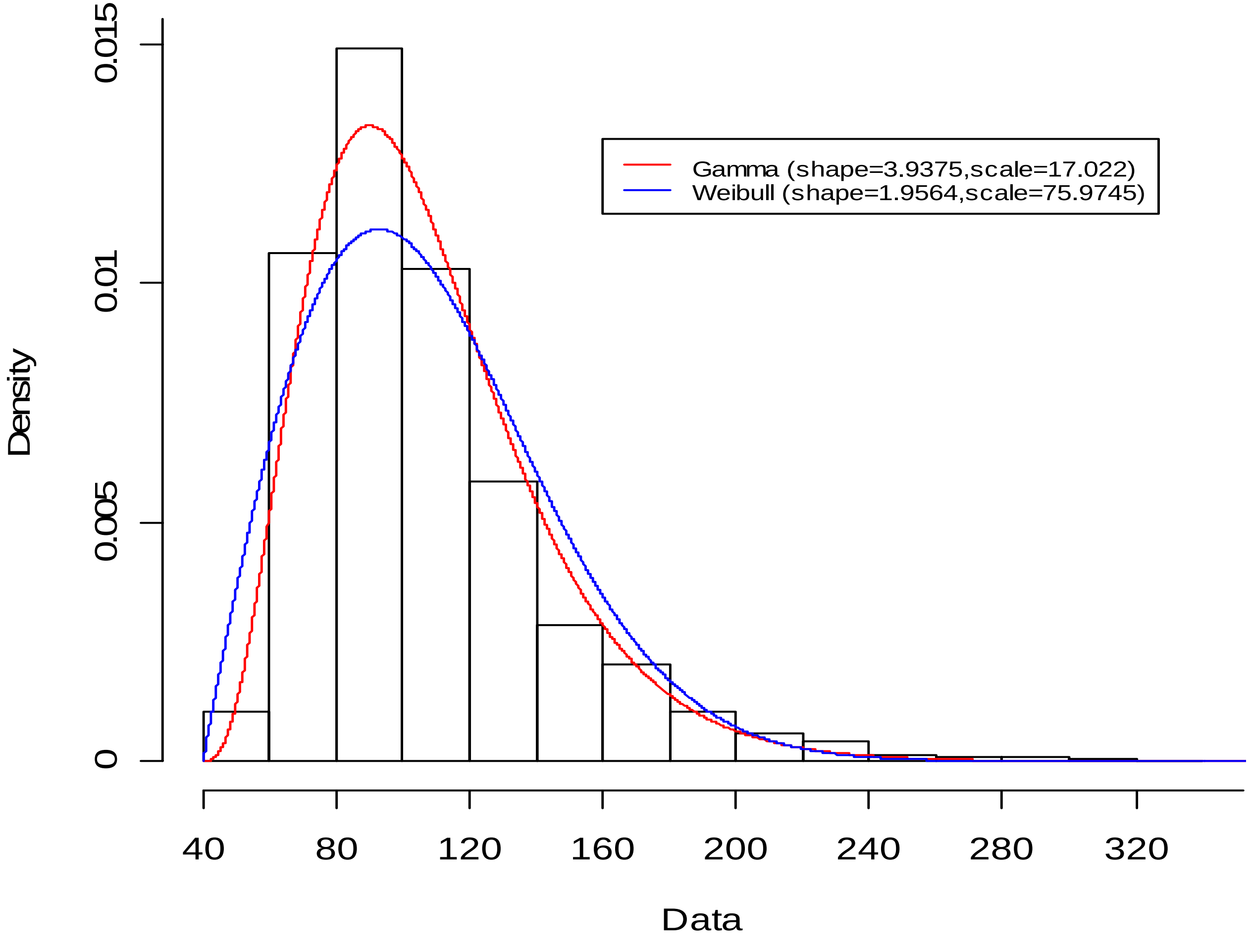


4
3.5
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250
300
350
-log(tail Prob.)
red line length

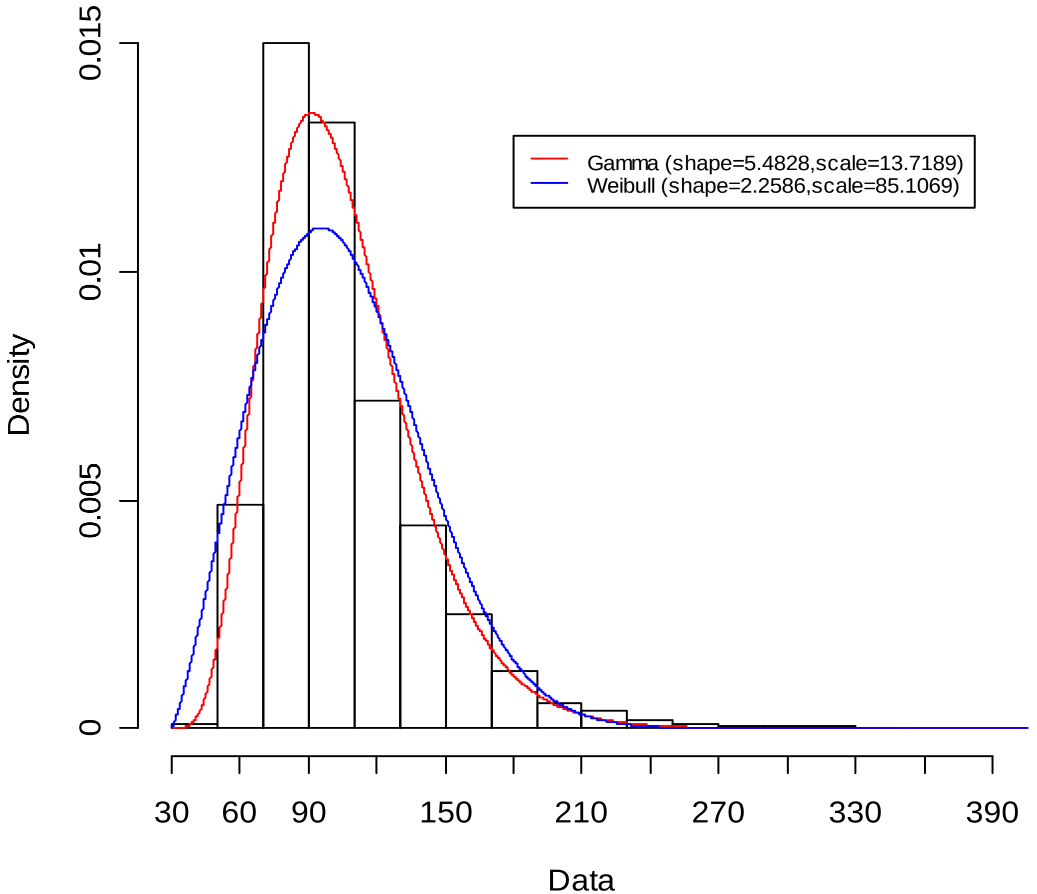


uniform random variable
-log(tail Prob.)
0
0.5
1
1.5
2
2.5
3
3.5
4
0
50
100
150
200
250
300
350
red line length

Brownian Motion

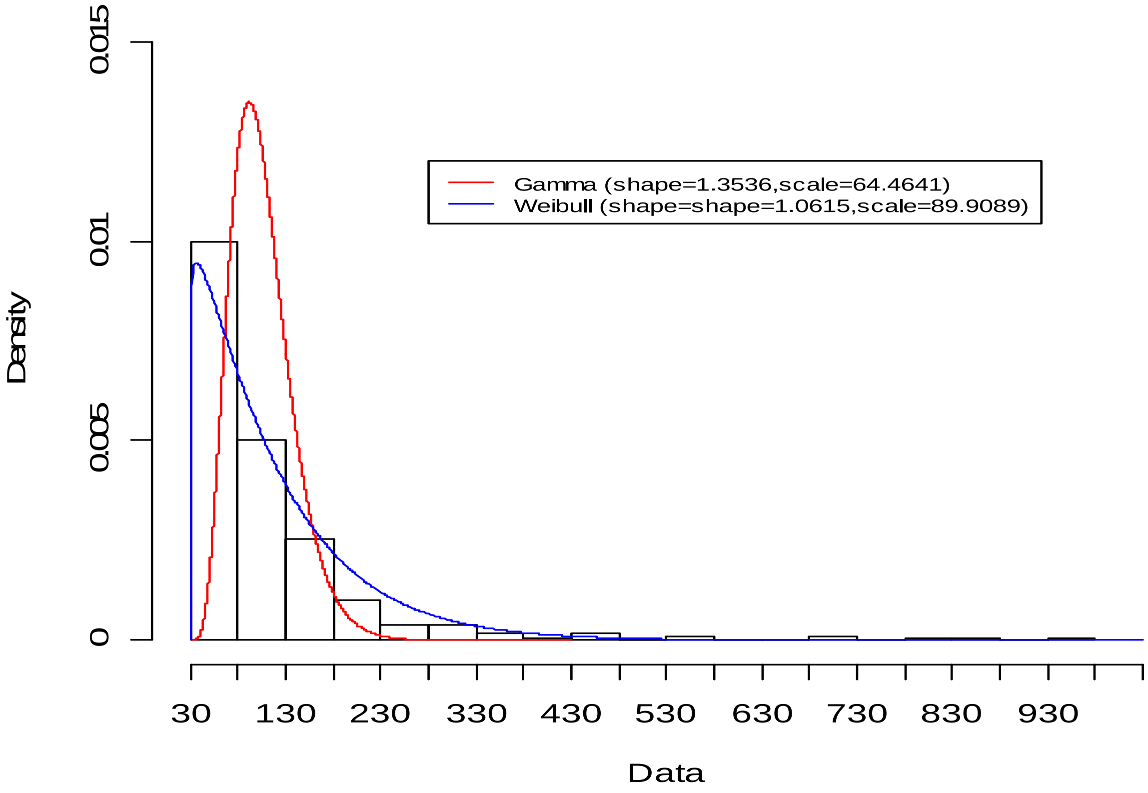


Uniform



Gamma (shape=5.4828,scale=13.7189)
Weibull (shape=2.2586,scale=85.1069)
Density
0
0.005
0.01
0.015
30
60
90
150
210
270
330
390
Data

SPX

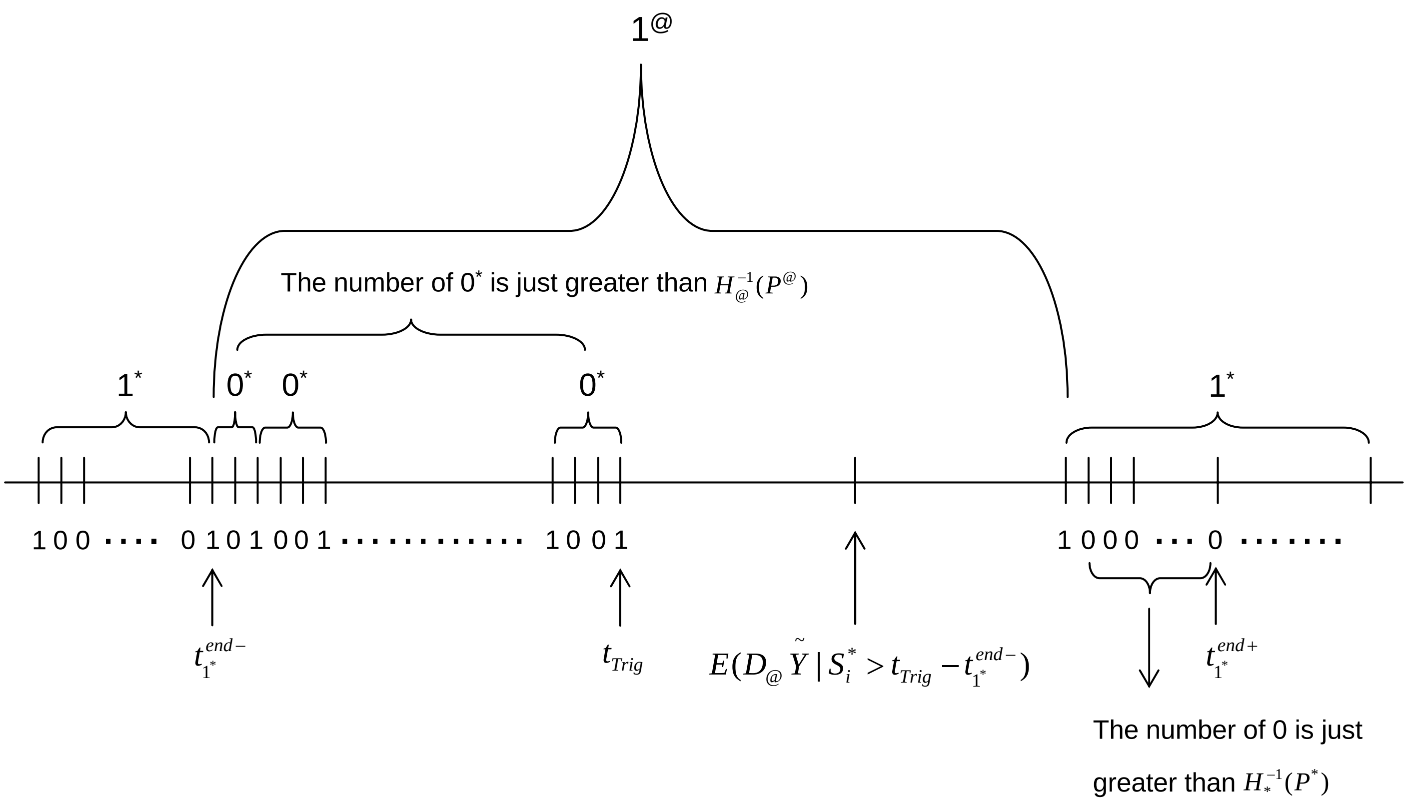


Density
0.015
0.01
0.005
0
Gamma (shape=1.3536,scale=64.4641)
Weibull (shape=shape=1.0615,scale=89.9089)
30
130
230
330
430
530
630
730
830
930
Data

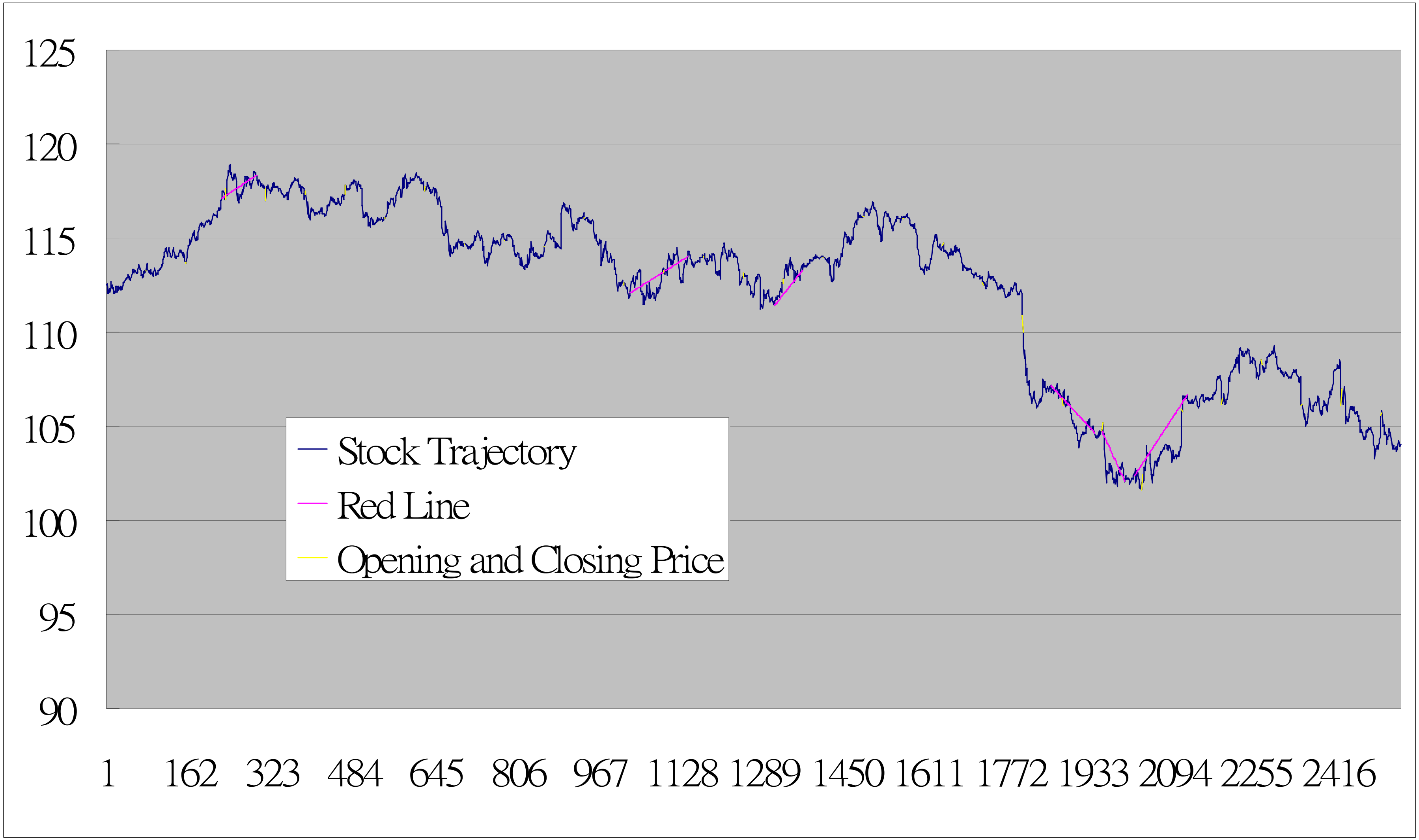
Black-Scholes Model
Geometric Brownian Motion

$$S_{t+\Delta t} = S_t e^{(r - \frac{1}{2}\sigma^2)\Delta t + \sigma\sqrt{\Delta t} Z}, \; Z \sim N(0,1)$$

$$\frac{S_{t+\Delta t} - S_t}{S_t} = e^{(r - \frac{1}{2}\sigma^2)\Delta t + \sigma\sqrt{\Delta t} Z} - 1$$



$1^{@}$
The number of 0^* is just greater than $H_{@}^{-1}(P^{@})$
1^*
0^*
0^*
0^*
1^*
1 0 0 .... 0 1 0 1 0 0 1 .... 1 0 0 1
1 0 0 0 ... 0 ....
$t_{1^*}^{end-}$
t_{Trig}
$E(D_{@}\tilde{Y} \mid S_i^* > t_{Trig} - t_{1^*}^{end-})$
$t_{1^*}^{end+}$
The number of 0 is just
greater than $H_*^{-1}(P^*)$



IBM stock (6/1/2001~7/19/2001)

Let M_n be an $n \times n$ positive-definite matrix. For large n to find the first few leading eigenvalues and the corresponding eigenvectors is difficult. The following approach may be used in practice. Pick m columns from M_n , for demonstration, say the first m column.

$$M_n = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{matrix} m \\ n-m \end{matrix}$$
$$\qquad\qquad m \quad n-m$$

It is easy to find reason and counter-examples to refute to the above approach. But what is the rationale behind this practice and what is a reasonable mathematical framework to justify this? Let $K(x, y)$ be a positive definite function on $E \times E$, and F be a probability distribution on E. Assume that

$$\int K^2(x, y) dF(x) dF(y) < \infty,$$

Define

$$\begin{aligned}\widehat{M}_n &= \begin{pmatrix} M_{11} \\ M_{21} \end{pmatrix} M_{11}^{-1} (M_{11} \; M_{12}) \\ &= \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{21} M_{11}^{-1} M_{12} \end{pmatrix}\end{aligned}$$

One instead solves the eigenvalue and eigenvector problem for $\widehat{M}_n$. Note that $\widehat{M}$ has only rank m. Usually it works in practice. Why?

$$K(x,y) = \sum_{k=1}^{\infty} \lambda_k \eta_k(x) \eta_k(y).$$

Let $X_1, \cdots, X_k, \cdots$ be iid r.v.'s with distribution F. And define

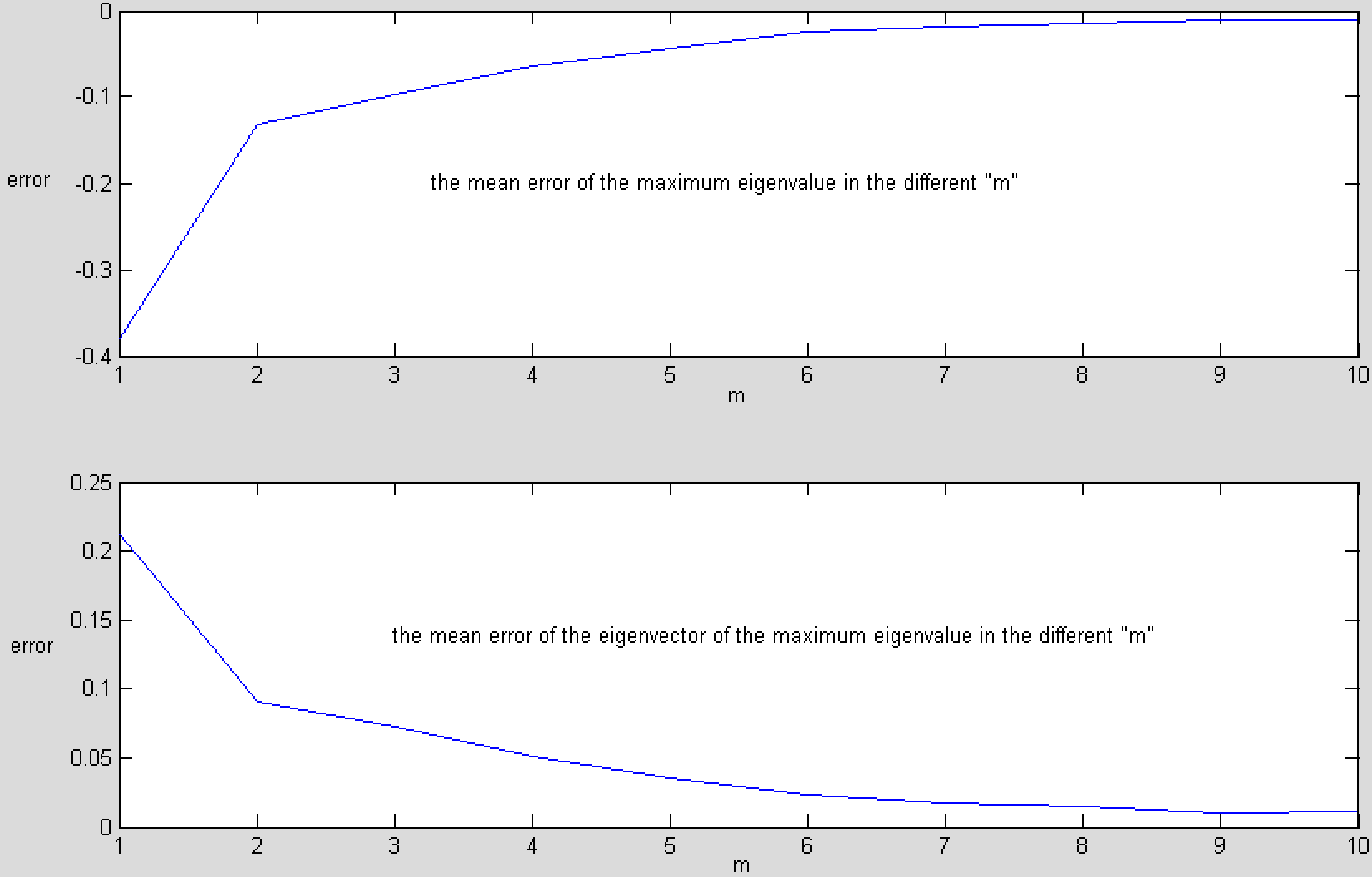
$$M_n = n^{-1} \left(K(X_i, X_j)\right)_{i,j=1}^{n}.$$

$\widehat{M}_n$ is defined correspondingly.

For example $K(x,y) = min(s,t), E = [0 \ \ 1]$ and $X_i's$ are i.i.d. uniform. The simulation results for the comparison of the largest eigenvalues (corresponding eigenvectors) of M_n and $\widehat{M}_n$ are shown is the following figures, $n = 100, m = 1, ..., 10$.

With regularity assumptions on $K(x,y)$, we show that the leading eigenvalues (and the corresponding eigenvectors) of M_n and $\widehat{M}_n$ are close to each other in probabilistic sense.

Other possible formulations?



Errors of maximum eigenvalues and their corresponding eigenvectors.
(The random matrix is 100 by 100 (n=100) and m is from 1 to 10.)

Bai, Z.-D., L.-B. Chang, S.-Y. Huang, C.-R. Hwang (2007) On the Nyström method and large random matrices, first draft.

For Nyström method, one may consult Rasmussen, C. E., C. K. I. Williams (2006) Gaussian Processes for Machine Learning, the MIT Press, 2006.

Consider the linear operator on $L^2(T^N)$, denoted by $\mathcal{H}$, in the form

$$L_{kC} \doteq \Delta + kC \cdot \nabla,$$

where $C : \mathbb{R}^N \to \mathbb{R}^N$ is periodic, smooth and divergence free, and k is a nonzero integer.

In our analysis, C is of some specific form. And for each C, the spectral gap of L_{kC} is analyzed as k goes to infinity. kC is considered as a perturbation and the multiplying constant $k \in \mathbb{N} \cup \{0\}$ its magnitude. The spectral gap of L_{kC} is defined as

$$\lambda(kC) \doteq -\sup\{ \text{ real part of } \mu,$$

$$\mu \text{ is in the spectrum of } L_{kC}, \mu \neq 0\}.$$

Note that $\lambda(kC)$ is positive.

Assume

$$C(x) = p \cos(q \cdot x)$$

$$= \begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix} \cos(q_1 x_1 + q_2 x_2 + \dots q_N x_N),$$

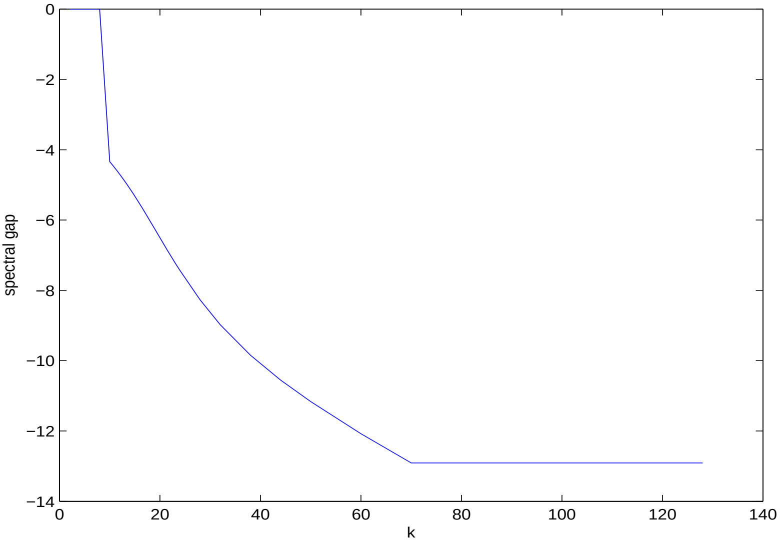
for $x = (x_1, ..., x_N) \in \mathbb{R}^N$, where $p, q \in \mathbb{Z}^N$ satisfy $p \cdot q = 0$.

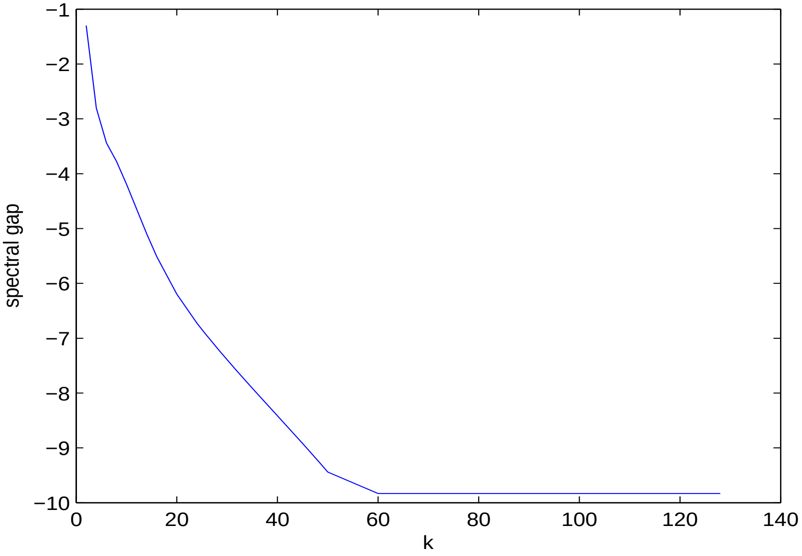
The main theorem is:

$$\lim_{k\to\infty} \lambda(kC) = \inf\{|\hat{m}|^2 : \hat{m} \cdot p = 0, \hat{m} \in \mathbb{Z}^N - \{0\}\}.$$

In particular for $p = (1, M, M^2, ..., M^{N-1})$ with $M \in \mathbb{Z}_+$, $\lim_{k\to\infty} \lambda(kC) = (M^2 + 1)$. Hence the spectral gap can be pushed to ∞.

H., H.-M. Pai (2005) Blowing up spectral gap of Laplacian on N-torus by antisymmetric perturbations, manuscript.


0
−2
−4
−6
−8
−10
−12
−14
0
20
40
60
80
100
120
140
k
spectral gap


spectral gap
k
−1
−2
−3
−4
−5
−6
−7
−8
−9
−10
0
20
40
60
80
100
120
140

In a general set up Franke et al.(2007) showed that $\lim_{k\to\infty} \lambda(kC)$ has a variational form expression involved all the eigenfunctions of $C \cdot \nabla$.

Can one push the spectral gap to ∞ in S^2?

Existence of C in T^2 or S^2 such that $\lim_{k\to\infty} \lambda(kC) = \infty$?

Franke, B., H., S.-J. Sheu and H.-M. Pei (2007) The behaviour of the spectral gap under growing drift, manuscript.

H., Shu-Yin Hwang-Ma, Shuenn-Jyi Sheu. (1993) Accelerating Gaussian Diffusions. *Ann. Appl. Probab.* 897-913.

H., Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. (2005) Accelerating Diffusions. *Ann. Appl. Probab.* 1433-1444.

Berestycki H., Hamel F., Nadirashvili N. (2005) Elliptic eigenvalue problems with large drift and applications to nonlinear propagation phenomena, *Commun. Math. Phys.*, 253, 451-480.

Constantin P., Kislev A., Ryzhik L., Zlatos A.: Diffusion and mixing in fluid flows, to appear in *Ann. Math.*.

Thank You!