Statistical analysis of time series: Gibbs measures and chains with complete connections

Rui Vilela Mendes

Statistical analysis of time series data

- Time series

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

$X \in Y$: the *state space*
$Y^{\mathbb{Z}}$: the *path space*

Statistical analysis of time series data

- Time series

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

$X \in Y$: the *state space*
$Y^{\mathbb{Z}}$: the *path space*

- **Statistical properties: (3 levels)**
  (1) Expectation values of the observables
  (2) Probability measures on state space Y
  (3) Probability measures on path space Y^Z

Statistical analysis of time series data

- Time series

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

  $X \in Y$: the *state space*
  $Y^{\mathbb{Z}}$: the *path space*
- **Statistical properties: (3 levels)**
  (1) Expectation values of the observables
  (2) Probability measures on state space Y
  (3) Probability measures on path space Y^Z
- *Level 1, 2 and 3- statistical indicators.*
  (*Mean partial sums*, *empirical measures* (pdf's) and *empirical process*)

Statistical analysis of time series data

- Time series

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

  $X \in Y$: the *state space*
  $Y^{\mathbb{Z}}$: the *path space*
- **Statistical properties: (3 levels)**
  (1) Expectation values of the observables
  (2) Probability measures on state space Y
  (3) Probability measures on path space Y^Z
- *Level 1, 2 and 3- statistical indicators.*
  (*Mean partial sums*, *empirical measures* (pdf's) and *empirical process*)
- **Analysis and reconstruction of the process:**
  Purpose: To extract *Grammar* and *measure*

Statistical analysis of time series data

- Time series

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

  $X \in Y$: the *state space*
  $Y^{\mathbb{Z}}$: the *path space*
- **Statistical properties: (3 levels)**
  (1) Expectation values of the observables
  (2) Probability measures on state space Y
  (3) Probability measures on path space Y^Z
- *Level 1, 2 and 3- statistical indicators.*
  (*Mean partial sums*, *empirical measures* (pdf's) and *empirical process*)
- **Analysis and reconstruction of the process:**
  Purpose: To extract *Grammar* and *measure*
- Examples:
  Hydrodynamic turbulence
  Market fluctuations:
  (there are analogies but the statistical indicators are different)

Statistical analysis of time series data

- **Working hypothesis**: statistical methods are an appropriate tool to describe and reconstruct the market fluctuation process
  Related to modern view of the *efficient market* (expected value of abnormal returns is zero - Fama)
  Opposite view: behavioral component must always be included
  However: Behavioral trends not inconsistent with statistical description if the different reaction times of the market components as well as secondary reactions are taken into account (Olsen et al.)

Statistical analysis of time series data

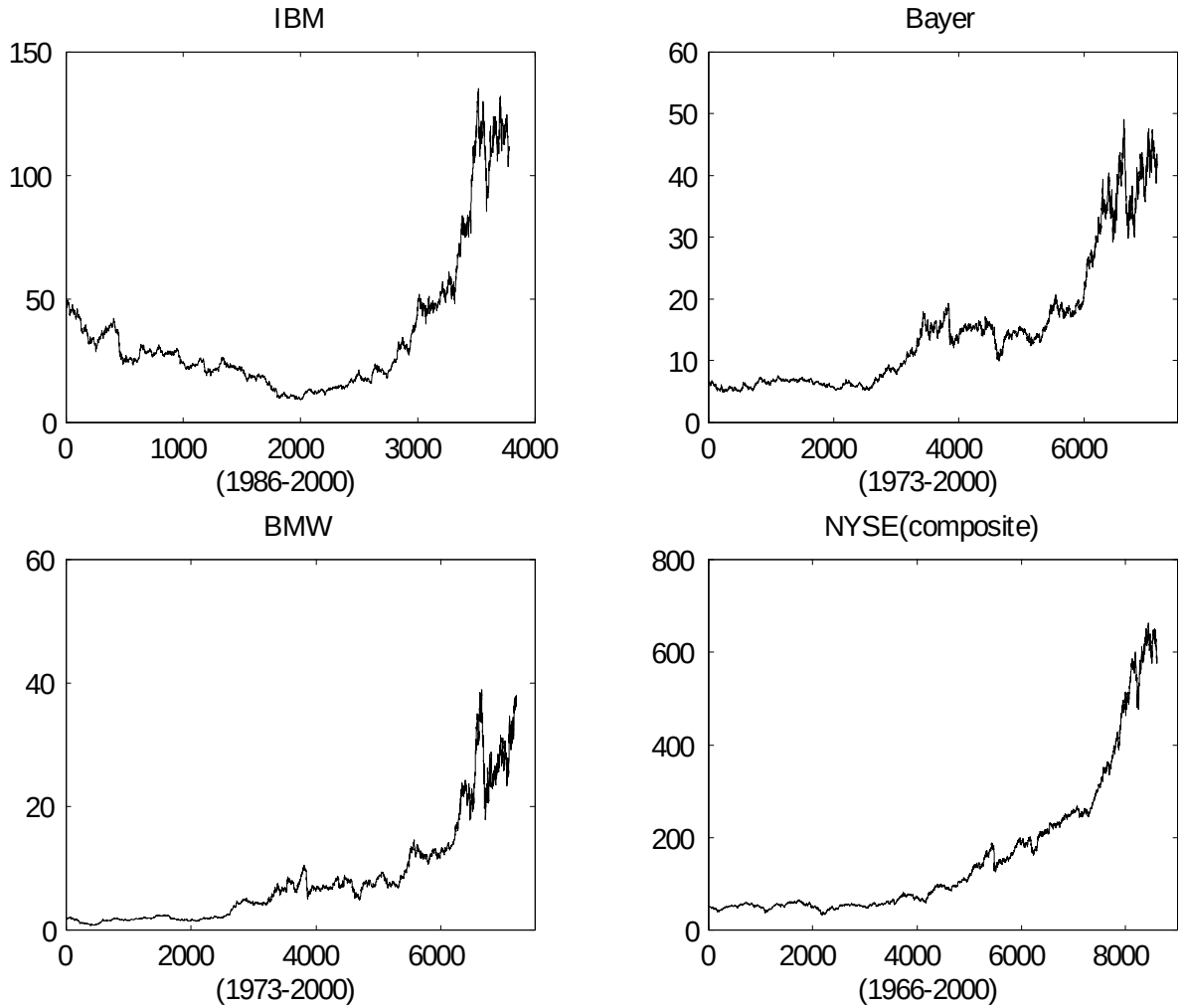
- **Working hypothesis**: statistical methods are an appropriate tool to describe and reconstruct the market fluctuation process
  Related to modern view of the *efficient market* (expected value of abnormal returns is zero - Fama)
  Opposite view: behavioral component must always be included
  However: Behavioral trends not inconsistent with statistical description if the different reaction times of the market components as well as secondary reactions are taken into account (Olsen et al.)
- Application of statistical tools requires:
  (i) Stationary or asymptotically stationary process
  (ii) Typical samples

Statistical analysis of time series data

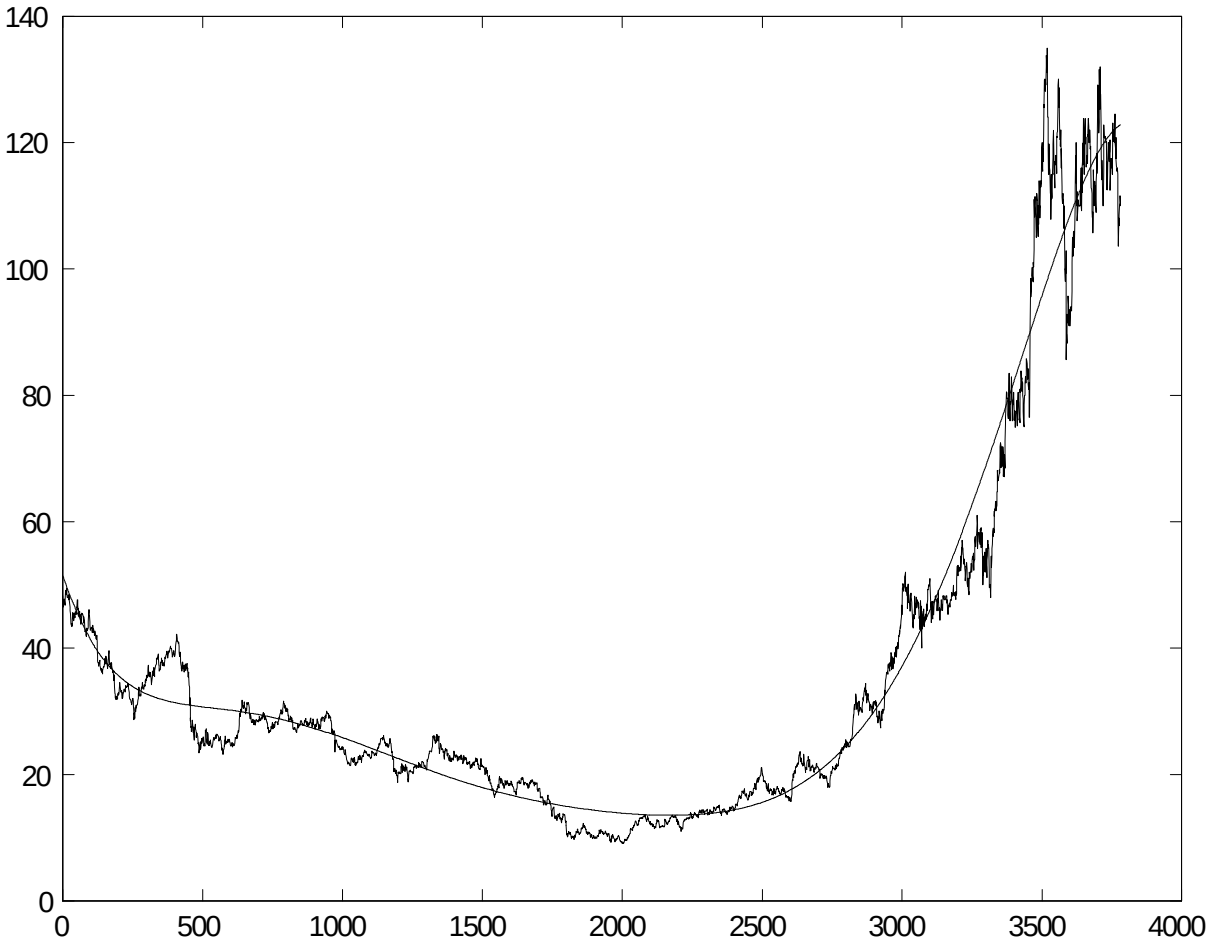
- **Working hypothesis**: statistical methods are an appropriate tool to describe and reconstruct the market fluctuation process
  Related to modern view of the *efficient market* (expected value of abnormal returns is zero - Fama)
  Opposite view: behavioral component must always be included
  However: Behavioral trends not inconsistent with statistical description if the different reaction times of the market components as well as secondary reactions are taken into account (Olsen et al.)
- Application of statistical tools requires:
  (i) Stationary or asymptotically stationary process
  (ii) Typical samples
- Stocks as experimental probes revealing the mechanisms of the market process
  (i) $\Rightarrow$ preprocessing of the data
  High-frequency versus low-frequency data
  Complexity versus statistics trade-off

Statistical analysis of time series data

Daily data $p(t)$

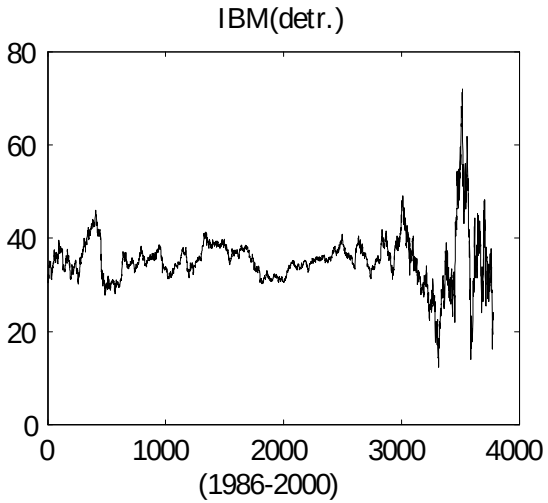


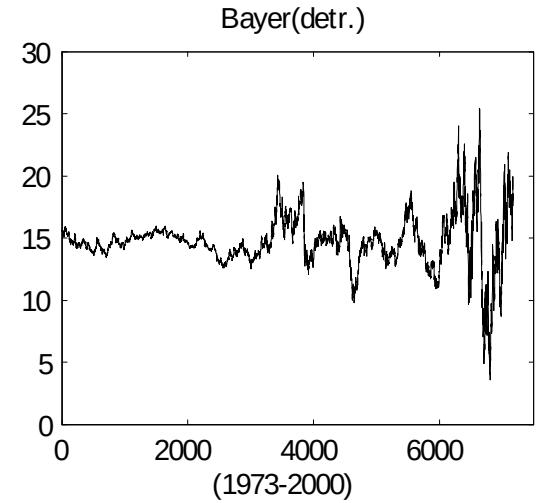
Detrending by a polynomial

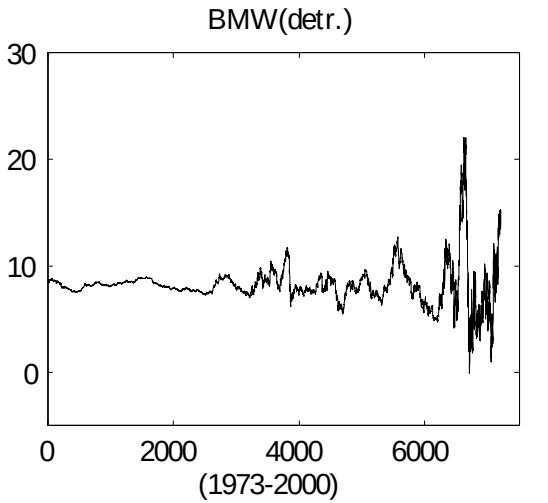


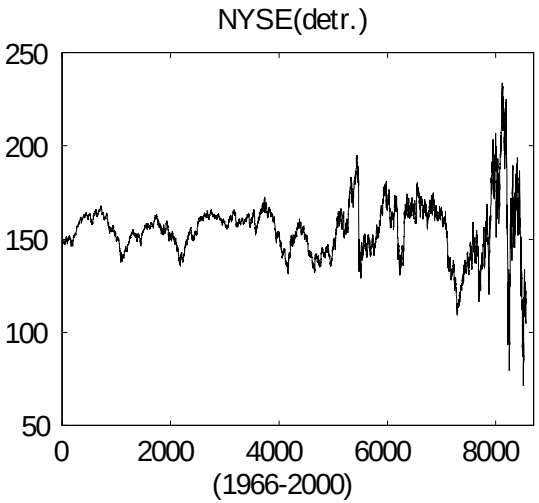
Statistical analysis of time series data

Detrended data $p(t) - q(t)$



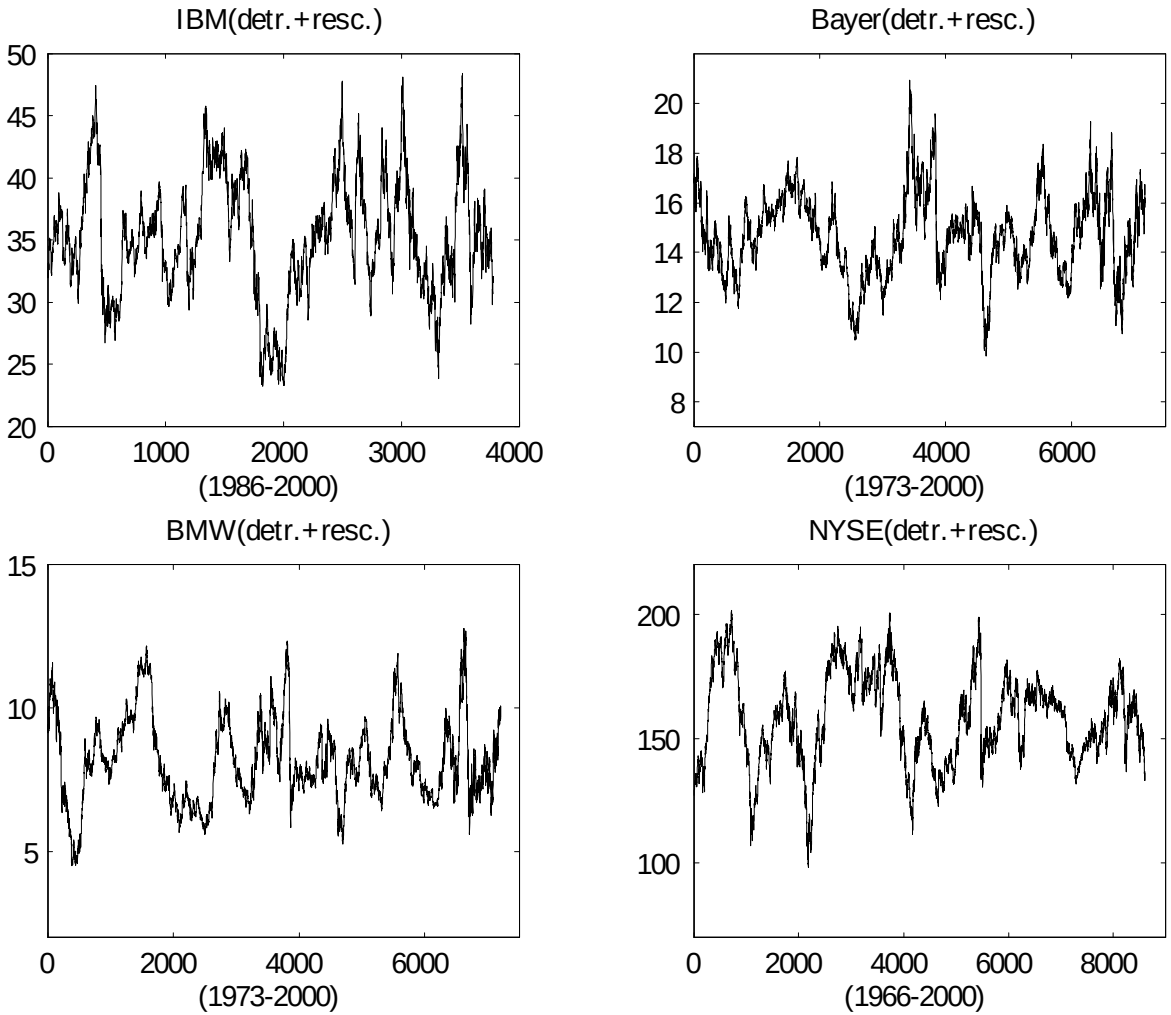






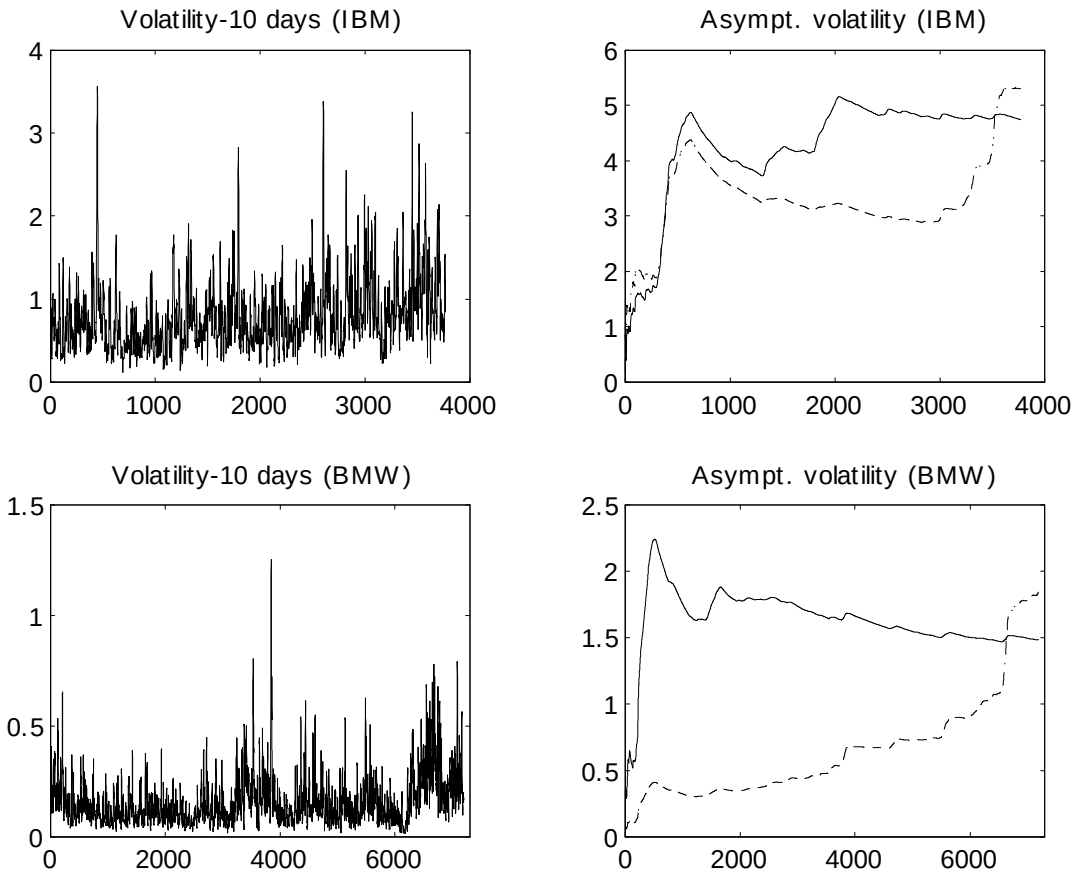
Statistical analysis of time series data

Detrended and rescaled data $x(t) = (p(t) - q(t)) \frac{<p(t)>}{q(t)}$



Statistical analysis of time series data

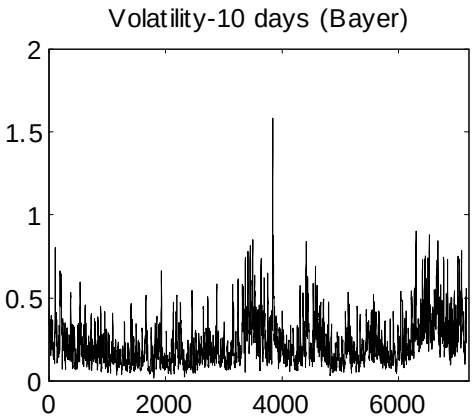
Ten-days window volatility and comparison of asymptotic volatility for the rescaled and non-rescaled data (IBM and BMW)

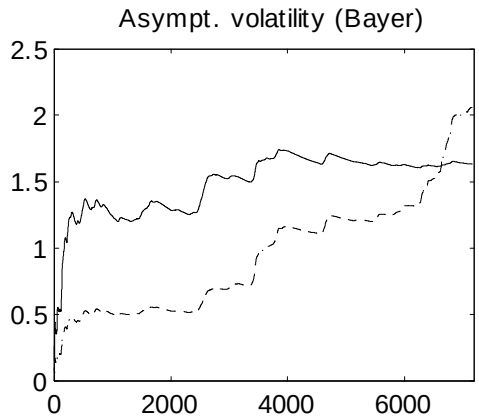


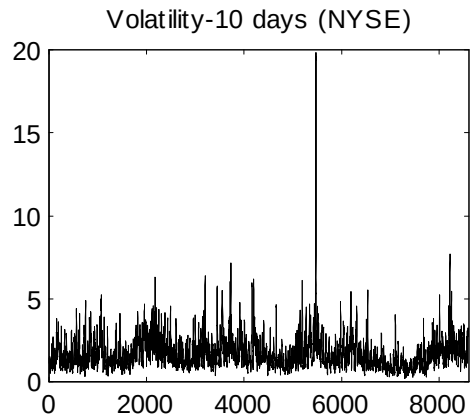
Local non-stationarity versus asymptotic stationarity

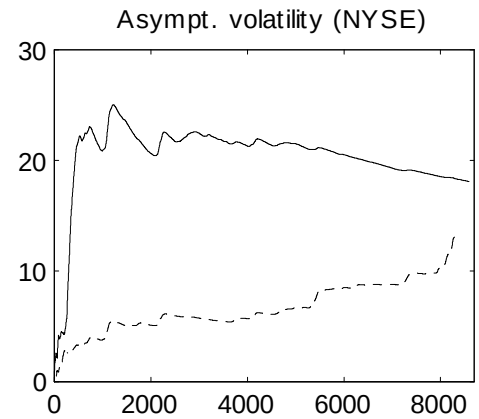
Statistical analysis of time series data

Ten-days window volatility and comparison of asymptotic volatility for the rescaled and non-rescaled data (Bayer and NYSE)









Also direct test of stationarity computing the entropies of multi-symbol words in the first and second half of the samples

Statistical indicators

- $n-$days return

$$r(t,n) = \log p(t+n) - \log p(t)$$

Statistical indicators

- $n-$days return

$$r(t,n) = \log p(t+n) - \log p(t)$$

- (i) *Maximum* (over t) of $r(t,n)$

$$\delta(n) = \max_t \{r(t,n)\}$$

Statistical indicators

- $n-$days return

$$r(t,n) = \log p\,(t+n) - \log p\,(t)$$

- (i) *Maximum* (over t) of $r(t,n)$

$$\delta\,(n) = \max_{t} \{r(t,n)\}$$

- (ii) *Moments* of the distribution of $|r(t,n)|$

$$S_q(n) = \langle |r(t,n)|^q \rangle$$

Statistical indicators

- $n-$days return

$$r(t,n) = \log p\,(t+n) - \log p\,(t)$$

- (i) *Maximum* (over t) of $r(t,n)$

$$\delta\,(n) = \max_t \{r(t,n)\}$$

- (ii) *Moments* of the distribution of $|r(t,n)|$

$$S_q(n) = \langle |r(t,n)|^q \rangle$$

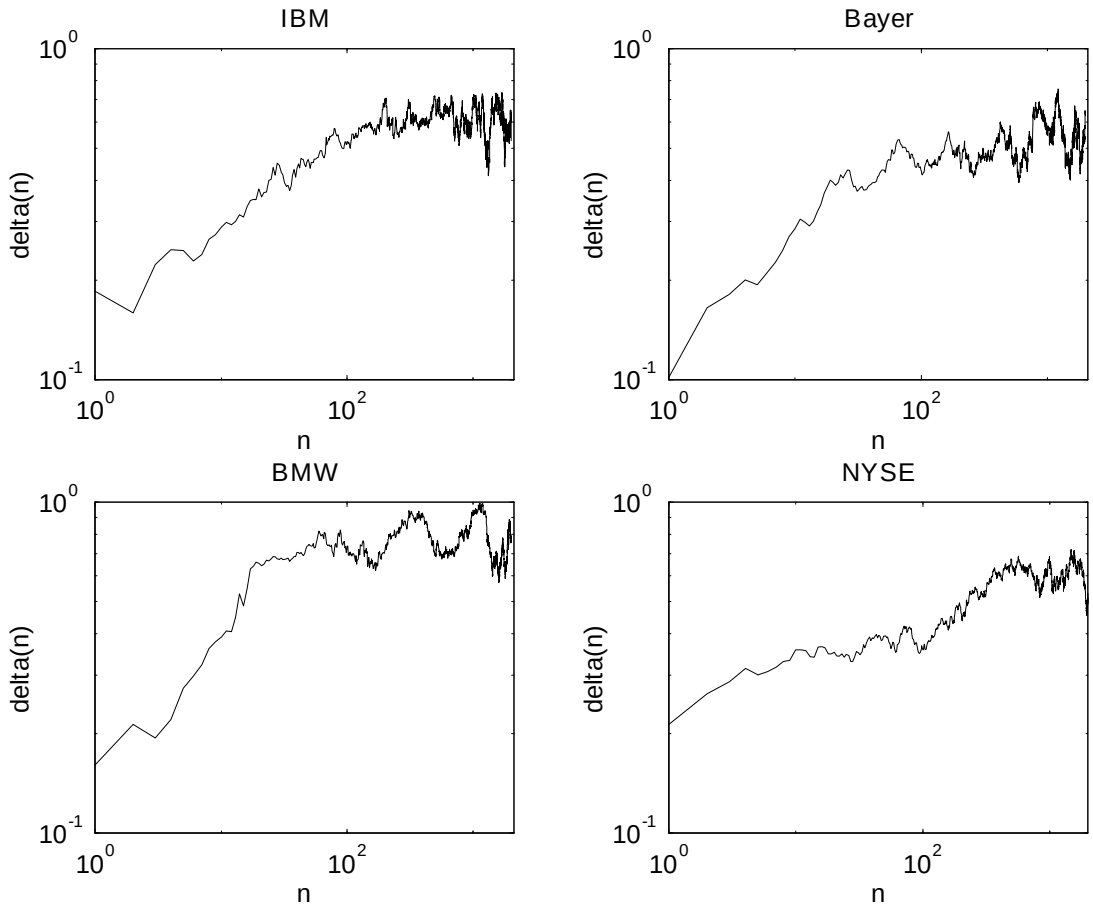
- (iii) If in some range ($n = 2$ to $n = 60$)

$$S_q(n) \sim n^{\chi(q)}$$

$\chi(q)$ is the *scaling exponent*

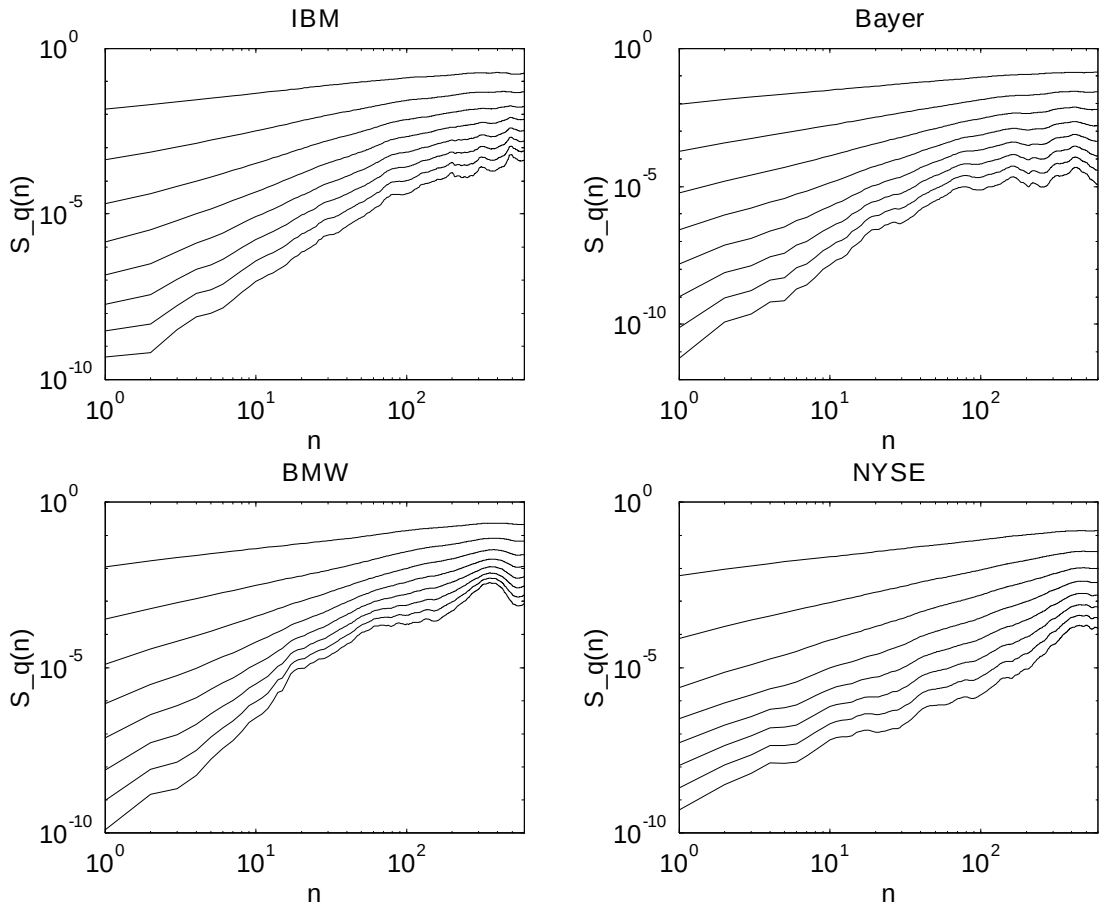
Statistical indicators

Maximum $\delta(n)$ of log-prices differences



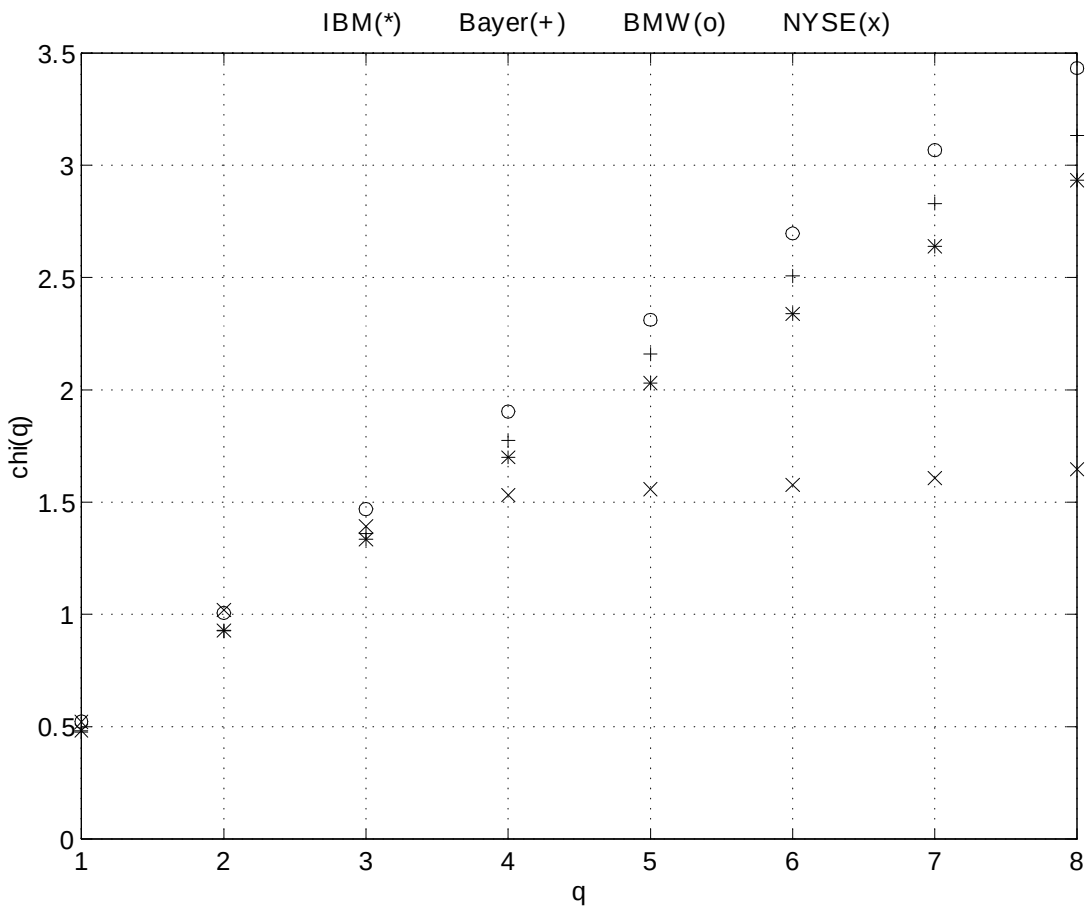
Statistical indicators

Moments of the $|r(t,n)|$ distribution



Statistical indicators

Scaling exponent $\chi(q)$



Statistical indicators

- Main conclusions:
  (a) $\delta(n)$ is log-concave and probably asymptotically constant for large r
  (b) $S_q(n)$ is a log-concave function of n with an inertial range
  (c) The scaling law $\chi(q)$ is an increasing concave function of q
  (d) $\chi(1)$ in the scaling region ($n = 2$ to $n = 60$) is close to 0.5
  (e) Scaling properties of NYSE somewhat different from the others

Statistical indicators

- Main conclusions:
  (a) $\delta(n)$ is log-concave and probably asymptotically constant for large r
  (b) $S_q(n)$ is a log-concave function of n with an inertial range
  (c) The scaling law $\chi(q)$ is an increasing concave function of q
  (d) $\chi(1)$ in the scaling region ($n = 2$ to $n = 60$) is close to 0.5
  (e) Scaling properties of NYSE somewhat different from the others

- Properties (a) to (c) are shared by the turbulence data, but with different values for the statistical indicators (in turbulence data $\chi(1) = \frac{1}{3}$, here $\chi(1) \approx 0.5 \Rightarrow$ essentially uncorrelated signal for $n \geq 2$)

Statistical indicators

- Main conclusions:
  (a) $\delta(n)$ is log-concave and probably asymptotically constant for large r
  (b) $S_q(n)$ is a log-concave function of n with an inertial range
  (c) The scaling law $\chi(q)$ is an increasing concave function of q
  (d) $\chi(1)$ in the scaling region ($n = 2$ to $n = 60$) is close to 0.5
  (e) Scaling properties of NYSE somewhat different from the others

- Properties (a) to (c) are shared by the turbulence data, but with different values for the statistical indicators (in turbulence data $\chi(1) = \frac{1}{3}$, here $\chi(1) \approx 0.5 \Rightarrow$ essentially uncorrelated signal for $n \geq 2$)

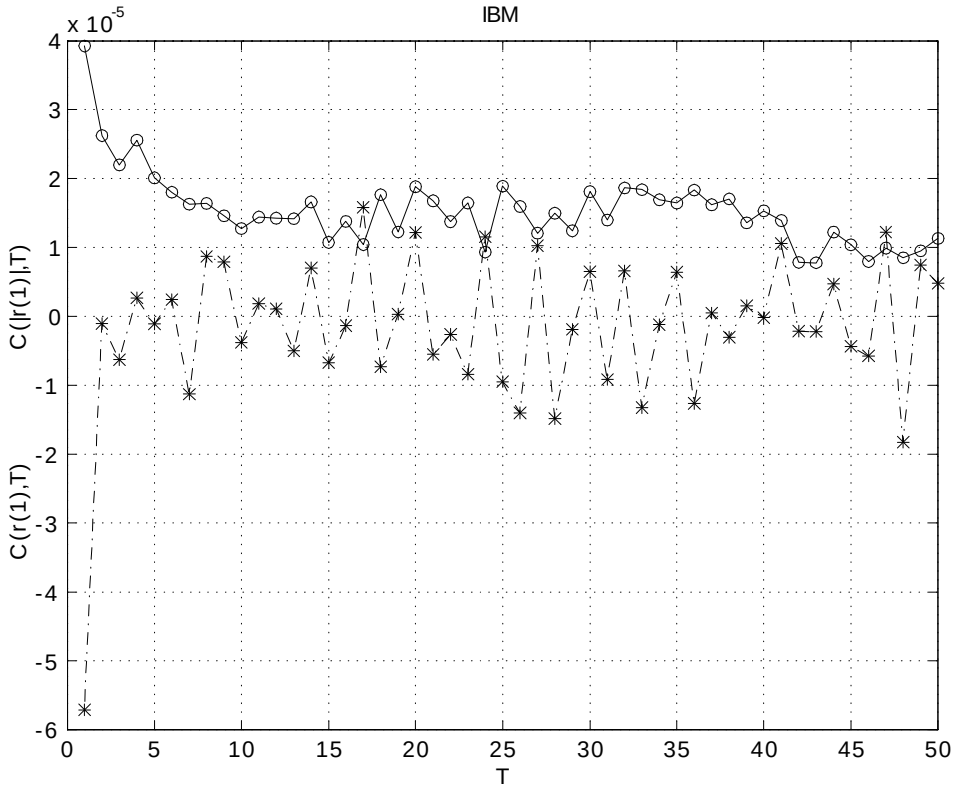
- The behavior of the statistical indicators $\delta(n)$, $S_q(n)$ and $\chi(q) \Longrightarrow$ If the process is a topological Markov chain the transitions allowed by the transition matrix T must lie inside a strictly convex domain around the diagonal of T

Statistical indicators

Correlation function of one-day returns ($\star$) and its absolute value (o)

$$C\left(r(1), T\right) = \left\langle r\left(t+T, 1\right) r\left(t, 1\right)\right\rangle$$

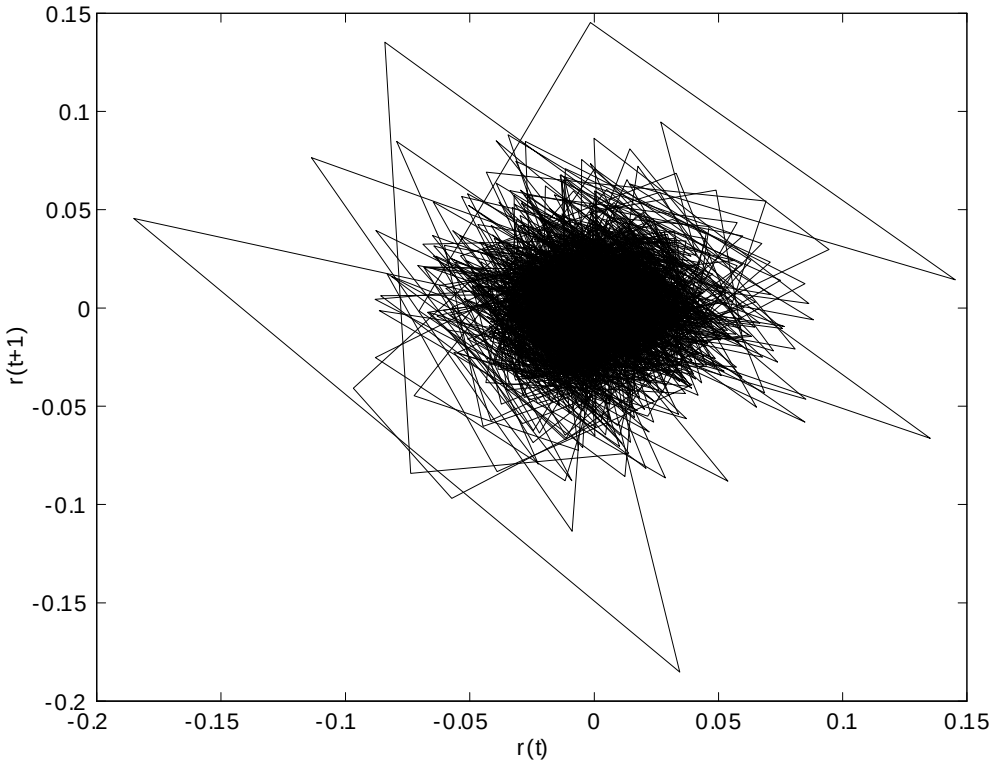
$$C\left(\left|r(1)\right|, T\right) = \left\langle \left|r\left(t+T, 1\right)\right| \left|r\left(t, 1\right)\right|\right\rangle$$



Statistical indicators

Dynamics of one-day returns

$$r(t,1) \longrightarrow r(t+1,1)$$



Looking for a Gibbs measure

Why Gibbs measures are "natural"

- Events $\{X_i\}$
  p_i = probability of event X_i

Looking for a Gibbs measure

Why Gibbs measures are "natural"

- Events $\{X_i\}$
  p_i = probability of event X_i
- Constraints:
  (i) Normalization, $\sum_i p_i = 1$
  (ii) Expectation value of known observables, $\sum_i p_i F_k(X_i) = C_k$

Looking for a Gibbs measure

Why Gibbs measures are "natural"

- Events $\{X_i\}$
  p_i = probability of event X_i
- Constraints:
  (i) Normalization, $\sum_i p_i = 1$
  (ii) Expectation value of known observables, $\sum_i p_i F_k(X_i) = C_k$
- Maximum entropy principle (to maximize the uncertainty about what is not known) = use the most unbiased estimate

$$S = -\sum_i p_i \log p_i + \lambda_0 \sum_i p_i + \sum_k \lambda_k \sum_i p_i F_k(X_i)$$

$$\frac{\partial S}{\partial p_i} = 0 \qquad \Longrightarrow \qquad -\log p_i - 1 + \lambda_0 + \textstyle\sum_k \lambda_k F_k(X_i) = 0$$

$$p_i = \exp\left(-1 + \lambda_0 + \sum_k \lambda_k F_k(X_i)\right)$$

with $\lambda_0, \lambda_1, \cdots$ obtained from the constraints

Looking for a Gibbs measure

- Coding by a finite alphabet Σ
  Space Ω of orbits $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$
  Dynamical law: a shift σ

$$\sigma\omega = i_2 \cdots i_k \cdots$$

Looking for a Gibbs measure

- Coding by a finite alphabet Σ
  Space Ω of orbits $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$
  Dynamical law: a shift σ

$$\sigma\omega = i_2 \cdots i_k \cdots$$

- Grammar: set of allowed sequences in Ω

Looking for a Gibbs measure

- Coding by a finite alphabet Σ
  Space Ω of orbits $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$
  Dynamical law: a shift σ

$$\sigma\omega = i_2 \cdots i_k \cdots$$

- Grammar: set of allowed sequences in Ω
- Sequences which coincide on the first n symbols: $n-$*cylinder* (or $n-$block) denoted by $[i_1 i_2 \cdots i_n]$
  Probability measures over the cylinders

Looking for a Gibbs measure

- Coding by a finite alphabet Σ
  Space Ω of orbits $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$
  Dynamical law: a shift σ

$$\sigma\omega = i_2 \cdots i_k \cdots$$

- Grammar: set of allowed sequences in Ω
- Sequences which coincide on the first n symbols: $n-cylinder$ (or $n-$block) denoted by $[i_1 i_2 \cdots i_n]$
  Probability measures over the cylinders
- **Gibbs measure**

$$c_1 \leq \frac{\mu\left([i_1(\omega) i_2(\omega) \cdots i_n(\omega)]\right)}{\exp\left(-nP + (S_n\phi)(\omega)\right)} \leq c_2$$

  $(S_n\phi)(\omega) = \sum_{k=0}^{n-1} \phi\left(\sigma^k\omega\right)$, ϕ being Hölder continuous function on Ω (the *potential)*
  $P(\phi, G)$: a function depending on potential and grammar (the *pressure* of ϕ)

Looking for a Gibbs measure

- Relation to the entropy

$$h(\mu) = \lim_{n\to\infty} \frac{H_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i_1 \cdots i_n} \mu\left([i_1 i_2 \cdots i_n]\right) \log \mu\left([i_1 i_2 \cdots i_n]\right)$$

Looking for a Gibbs measure

- Relation to the entropy

$$h(\mu) = \lim_{n\to\infty} \frac{H_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i_1 \cdots i_n} \mu\left([i_1 i_2 \cdots i_n]\right) \log \mu\left([i_1 i_2 \cdots i_n]\right)$$

- Variational principle: for each potential and grammar, $\sup_\eta \left\{ h(\eta) + \int \phi d\eta \right\}$ over all $\sigma-$invariant measures η is reached only for the Gibbs measure μ and equals the pressure

$$P(\phi, G) = h(\mu) + \int \phi d\mu$$

Looking for a Gibbs measure

- Relation to the entropy

$$h(\mu) = \lim_{n\to\infty} \frac{H_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i_1 \cdots i_n} \mu\left([i_1 i_2 \cdots i_n]\right) \log \mu\left([i_1 i_2 \cdots i_n]\right)$$

- Variational principle: for each potential and grammar, $\sup_\eta \left\{ h(\eta) + \int \phi d\eta \right\}$ over all $\sigma-$invariant measures η is reached only for the Gibbs measure μ and equals the pressure

$$P(\phi, G) = h(\mu) + \int \phi d\mu$$

- Potential may be chosen such that $P = 0$ (*normalized potential*). Then

$$\phi(\omega) = \lim_{n\to\infty} \log \frac{\mu\left([i_1(\omega) \cdots i_n(\omega)]\right)}{\mu\left([i_2(\omega) \cdots i_n(\omega)]\right)}$$

(Practical use hindered by poor statistics of large blocks)

Looking for a Gibbs measure

- **Gibbs measures for finite range potentials**
  (finite range potentials approximate uniformly any Hölder continuous potential)

Looking for a Gibbs measure

- **Gibbs measures for finite range potentials**
  (finite range potentials approximate uniformly any Hölder continuous potential)
- Property of range$-r$ potentials: for all values $i_1 i_2 \cdots i_n$ with $n \geq r$

$$\mu\left([i_1 \cdots i_n]\right) = \frac{\mu\left([i_1 \cdots i_r]\right) \times \cdots \times \mu\left([i_{n-r+1} \cdots i_n]\right)}{\mu\left([i_2 \cdots i_r]\right) \times \cdots \times \mu\left([i_{n-r+1} \cdots i_{n-1}]\right)} \tag{1}$$

$\Longrightarrow$

$$h(\mu) = -\sum_{i_1 \cdots i_k} \mu\left([i_1 \cdots i_k]\right) \log \frac{\mu\left([i_1 \cdots i_k]\right)}{\mu\left([i_1 \cdots i_{k-1}]\right)} = H_k - H_{k-1}$$

for all $k \geq r$ if $r > 1$. If $r = 1$
$h(\mu) = H_1 H_k = -\sum_{i_1 \cdots i_k} \mu\left([i_1 \cdots i_k]\right) \log \mu\left([i_1 \cdots i_k]\right)$

Looking for a Gibbs measure

- **Gibbs measures for finite range potentials**
  (finite range potentials approximate uniformly any Hölder continuous potential)
- Property of range$-r$ potentials: for all values $i_1 i_2 \cdots i_n$ with $n \geq r$

$$\mu\left(\left[i_1 \cdots i_n\right]\right) = \frac{\mu\left(\left[i_1 \cdots i_r\right]\right) \times \cdots \times \mu\left(\left[i_{n-r+1} \cdots i_n\right]\right)}{\mu\left(\left[i_2 \cdots i_r\right]\right) \times \cdots \times \mu\left(\left[i_{n-r+1} \cdots i_{n-1}\right]\right)} \tag{1}$$

$\Longrightarrow$

$$h\left(\mu\right) = -\sum_{i_1 \cdots i_k} \mu\left(\left[i_1 \cdots i_k\right]\right) \log \frac{\mu\left(\left[i_1 \cdots i_k\right]\right)}{\mu\left(\left[i_1 \cdots i_{k-1}\right]\right)} = H_k - H_{k-1}$$

for all $k \geq r$ if $r > 1$. If $r = 1$
$h\left(\mu\right) = H_1 H_k = -\sum_{i_1 \cdots i_k} \mu\left(\left[i_1 \cdots i_k\right]\right) \log \mu\left(\left[i_1 \cdots i_k\right]\right)$

- $\Longrightarrow$ **criterium to find the range of the potential**: range of the potential found when $H_k - H_{k-1}$ tends to a constant.
  Once the range is found, the potential may be constructed from the empirical weights $\widetilde{\mu}\left(\left[i_1 \cdots i_k\right]\right)$.

Looking for a Gibbs measure

- Another consequence of (1) is that for $k > r$

$$\mu\left([i_1 \cdots i_{k+1}]\right) = \frac{\mu\left([i_1 \cdots i_k]\right)\mu\left([i_2 \cdots i_{k+1}]\right)}{\mu\left([i_2 \cdots i_k]\right)}$$

Looking for a Gibbs measure

- Another consequence of (1) is that for $k > r$

$$\mu\left(\left[i_1 \cdots i_{k+1}\right]\right) = \frac{\mu\left(\left[i_1 \cdots i_k\right]\right)\mu\left(\left[i_2 \cdots i_{k+1}\right]\right)}{\mu\left(\left[i_2 \cdots i_k\right]\right)}$$

- **Application to the market fluctuations**:
  Five-symbols code $\Sigma = \{-2, -1, 0, 1, 2\}$ for

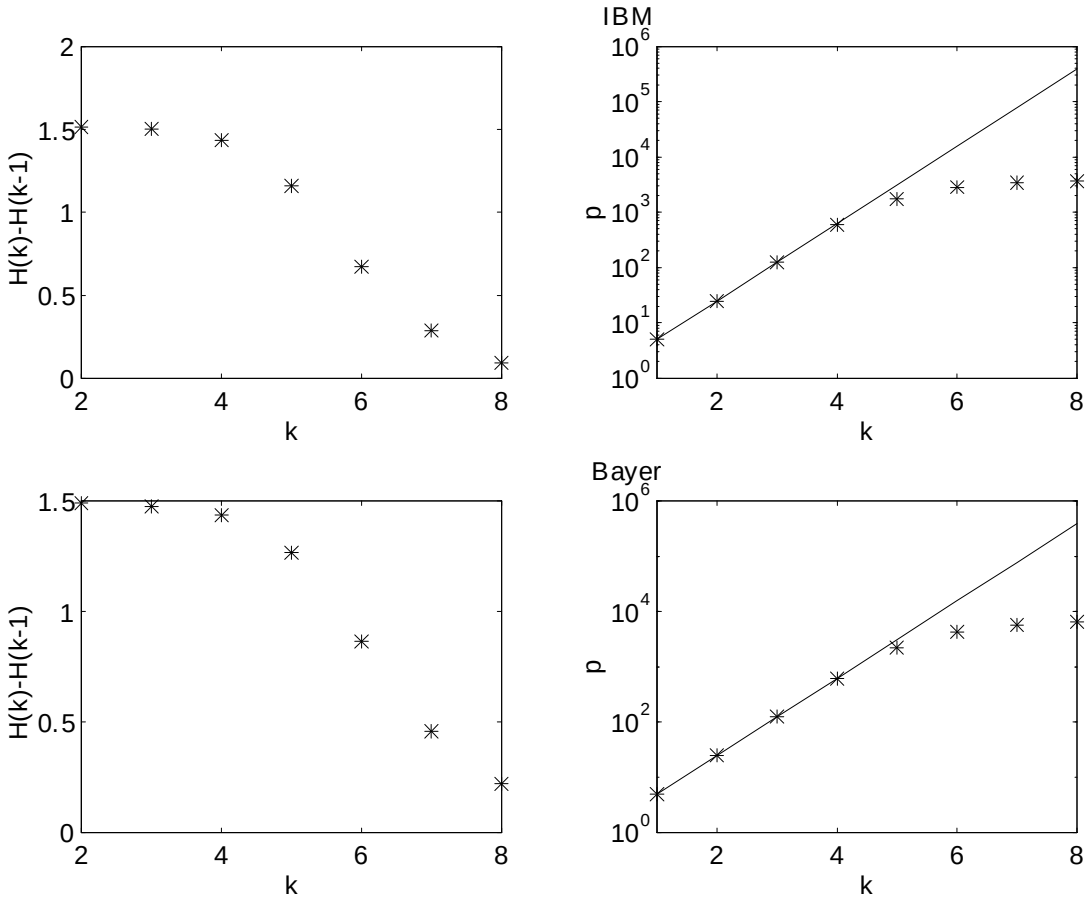
$$r(t) = \log p(t+1) - \log p(t)$$

Average $\overline{r(t)}$ and standard deviation $s = \sqrt{\overline{\left(r^2(t) - \overline{r(t)}^2\right)}}$

$$\begin{aligned}
\left(r(t) - \overline{r(t)}\right) > s &\iff 2 \\
s \geq \left(r(t) - \overline{r(t)}\right) > \tfrac{s}{3} &\iff 1 \\
\tfrac{s}{3} \geq \left(r(t) - \overline{r(t)}\right) > -\tfrac{s}{3} &\iff 0 \\
-\tfrac{s}{3} \geq \sigma\left(r(t) - \overline{r(t)}\right) > -s &\iff -1 \\
-s \geq \left(r(t) - \overline{r(t)}\right) &\iff -2
\end{aligned}$$

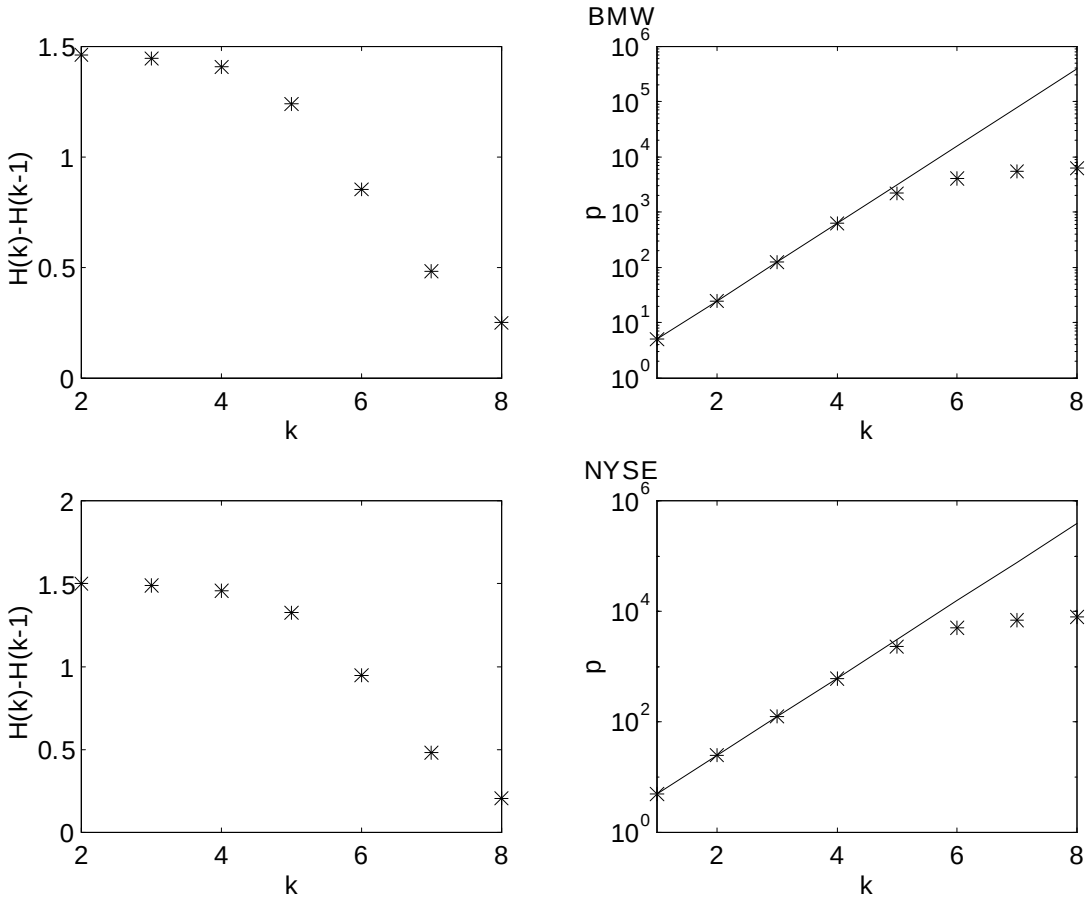
Looking for a Gibbs measure

$H_k - H_{k-1}$ and the number of occurring blocks of size k (IBM and Bayer)



Looking for a Gibbs measure

$H_k - H_{k-1}$ and the number of occurring blocks of size k (BMW and NYSE)



Looking for a Gibbs measure

- Behavior of $H_k - H_{k-1}$ quite different from hydrodynamic turbulence data

Looking for a Gibbs measure

- Behavior of $H_k - H_{k-1}$ quite different from hydrodynamic turbulence data
- To check whether a short-range potential is reliable :
  For successively higher k estimate

$$\mu_e\left([i_1 \cdots i_{k+1}]\right) = \frac{\widetilde{\mu}\left([i_1 \cdots i_k]\right)\widetilde{\mu}\left([i_2 \cdots i_{k+1}]\right)}{\widetilde{\mu}\left([i_2 \cdots i_k]\right)}$$

  then compare with the empirically observed $\widetilde{\mu}\left([i_1 \cdots i_{k+1}]\right)$

Looking for a Gibbs measure

- Behavior of $H_k - H_{k-1}$ quite different from hydrodynamic turbulence data
- To check whether a short-range potential is reliable :
  For successively higher k estimate

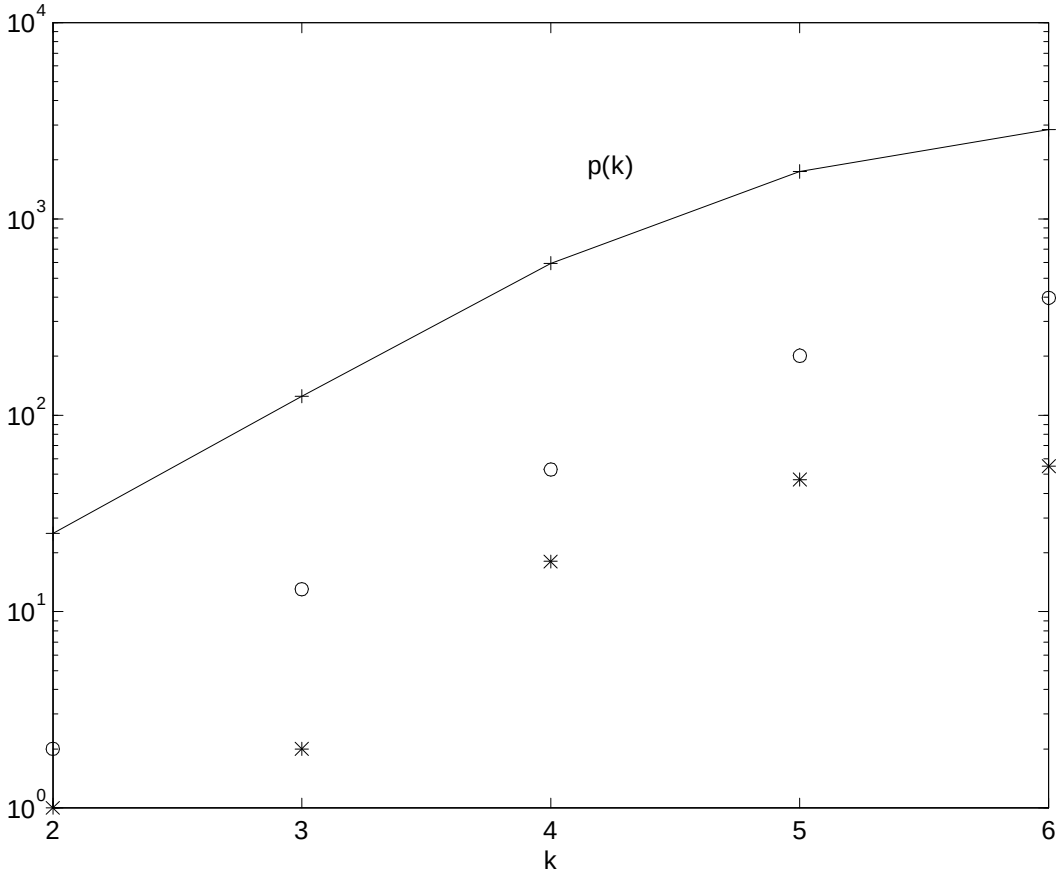
$$\mu_e\left(\left[i_1 \cdots i_{k+1}\right]\right) = \frac{\widetilde{\mu}\left(\left[i_1 \cdots i_k\right]\right)\widetilde{\mu}\left(\left[i_2 \cdots i_{k+1}\right]\right)}{\widetilde{\mu}\left(\left[i_2 \cdots i_k\right]\right)}$$

  then compare with the empirically observed $\widetilde{\mu}\left(\left[i_1 \cdots i_{k+1}\right]\right)$
- Standard deviation of the relative positive errors

$$\varepsilon_k = \max\left(0, \frac{\widetilde{\mu}\left(\left[i_1 \cdots i_{k+1}\right]\right) - \mu_e\left(\left[i_1 \cdots i_{k+1}\right]\right)}{\frac{1}{2}\left(\widetilde{\mu}\left(\left[i_1 \cdots i_{k+1}\right]\right) + \mu_e\left(\left[i_1 \cdots i_{k+1}\right]\right)\right)}\right)$$

Looking for a Gibbs measure

Underestimation errors one (o) and two (*) standard deviations away from the mean and the total number $p(k)$ of observed blocks



Looking for a Gibbs measure

- Large number of large deviation errors
  Correspond to blocks involving 2 and -2
  Large deviations misrepresented by empirically constructed measure

Looking for a Gibbs measure

- Large number of large deviation errors
  Correspond to blocks involving 2 and -2
  Large deviations misrepresented by empirically constructed measure
- ?? $\Rightarrow$ non-Gibbsian measure
  or
  ?? $\Rightarrow$ Gibbsian measure with long-range potential (sharp rise of $H_k - H_{k-1}$ at $k = 2$ followed by a very slow increase)

Looking for a Gibbs measure

- Large number of large deviation errors
  Correspond to blocks involving 2 and -2
  Large deviations misrepresented by empirically constructed measure
- ?? $\Rightarrow$ non-Gibbsian measure
  or
  ?? $\Rightarrow$ Gibbsian measure with long-range potential (sharp rise of $H_k - H_{k-1}$ at $k = 2$ followed by a very slow increase)
- Large deviation analysis applied to the calculation of H_k consistent with this hypothesis

Looking for a Gibbs measure

- Large number of large deviation errors
  Correspond to blocks involving 2 and -2
  Large deviations misrepresented by empirically constructed measure
- ?? $\Rightarrow$ non-Gibbsian measure
  or
  ?? $\Rightarrow$ Gibbsian measure with long-range potential (sharp rise of $H_k - H_{k-1}$ at $k = 2$ followed by a very slow increase)
- Large deviation analysis applied to the calculation of H_k consistent with this hypothesis
- In any case it needs an approach suited to deal with long-memory processes

Chains with complete connections

- **Chain with complete connections** (CCC)

Chains with complete connections

- **Chain with complete connections** (CCC)

1. $\forall a_i \in \Sigma$

$$P\left(X_1 = a_1, \cdots, X_n = a_n\right) > 0$$

Chains with complete connections

- **Chain with complete connections** (CCC)

1. $\forall a_i \in \Sigma$

$$P\left(X_1 = a_1, \cdots, X_n = a_n\right) > 0$$

2. The limit

$$\begin{aligned} & \lim_{m \to \infty} P\left(X_0 = a_0 | X_j = a_j, -m \leq j \leq -1\right) \\ = \quad & P\left(X_0 = a_0 | X_j = a_j, j \leq -1\right) \end{aligned}$$

exists $\forall a_i, j \leq -1$

Chains with complete connections

- **Chain with complete connections** (CCC)

1. $\forall a_i \in \Sigma$
$$P\left(X_1 = a_1, \cdots, X_n = a_n\right) > 0$$
2. The limit
$$\begin{aligned} & \lim_{m \to \infty} P\left(X_0 = a_0 | X_j = a_j, -m \leq j \leq -1\right) \\ = \quad & P\left(X_0 = a_0 | X_j = a_j, j \leq -1\right) \end{aligned}$$
exists $\forall a_i, j \leq -1$
3. There is a sequence $\left(\gamma_m\right)_{m \geq 1}$ with $\lim_{m \to \infty} \gamma_m = 0$, such that for all $\left\{a_j, b_j \in \Sigma, j \leq -1\right\}$ with $a_j = b_j$ for $-m \leq j \leq -1$
$$\left| \left(\frac{P\left(X_0 = a_0 | X_j = a_j, j \leq -1\right)}{P\left(X_0 = a_0 | X_j = b_j, j \leq -1\right)} - 1 \right) \right| \leq \gamma_m$$

Looking for a Gibbs measure

- **Chain with complete connections and summable decay** (CCCSD) CCC with $\sum \gamma_m < \infty$

Looking for a Gibbs measure

- **Chain with complete connections and summable decay** (CCCSD) CCC with $\sum \gamma_m < \infty$
- Conditions 1. and 2. implicitly assumed for the pre-processed data Decays γ_m estimated from a typical sample of the process. From the empirical probabilities

$$P\left(a_0 \middle| a_1 \cdots a_m A\right) = \frac{P\left(a_0 a_1 \cdots a_m A\right)}{P\left(a_1 \cdots a_m A\right)}$$

A a block of arbitrary length

$$g\left(a_0 a_1 \cdots a_m\right) = \left(\frac{\max_A P\left(a_0 \middle| a_1 \cdots a_m A\right)}{\min_A P\left(a_0 \middle| a_1 \cdots a_m A\right)} - 1\right)$$

$$\gamma_m = \max_{a_0 a_1 \cdots a_m} g\left(a_0 a_1 \cdots a_m\right)$$

If the statistics for long blocks is poor $\Rightarrow$ large fluctuations in γ_m

Looking for a Gibbs measure

- **Chain with complete connections and summable decay** (CCCSD) CCC with $\sum \gamma_m < \infty$
- Conditions 1. and 2. implicitly assumed for the pre-processed data Decays γ_m estimated from a typical sample of the process. From the empirical probabilities

$$P\left(a_0 \middle| a_1 \cdots a_m A\right) = \frac{P\left(a_0 a_1 \cdots a_m A\right)}{P\left(a_1 \cdots a_m A\right)}$$

A a block of arbitrary length

$$g\left(a_0 a_1 \cdots a_m\right) = \left(\frac{\max_A P\left(a_0 \middle| a_1 \cdots a_m A\right)}{\min_A P\left(a_0 \middle| a_1 \cdots a_m A\right)} - 1\right)$$

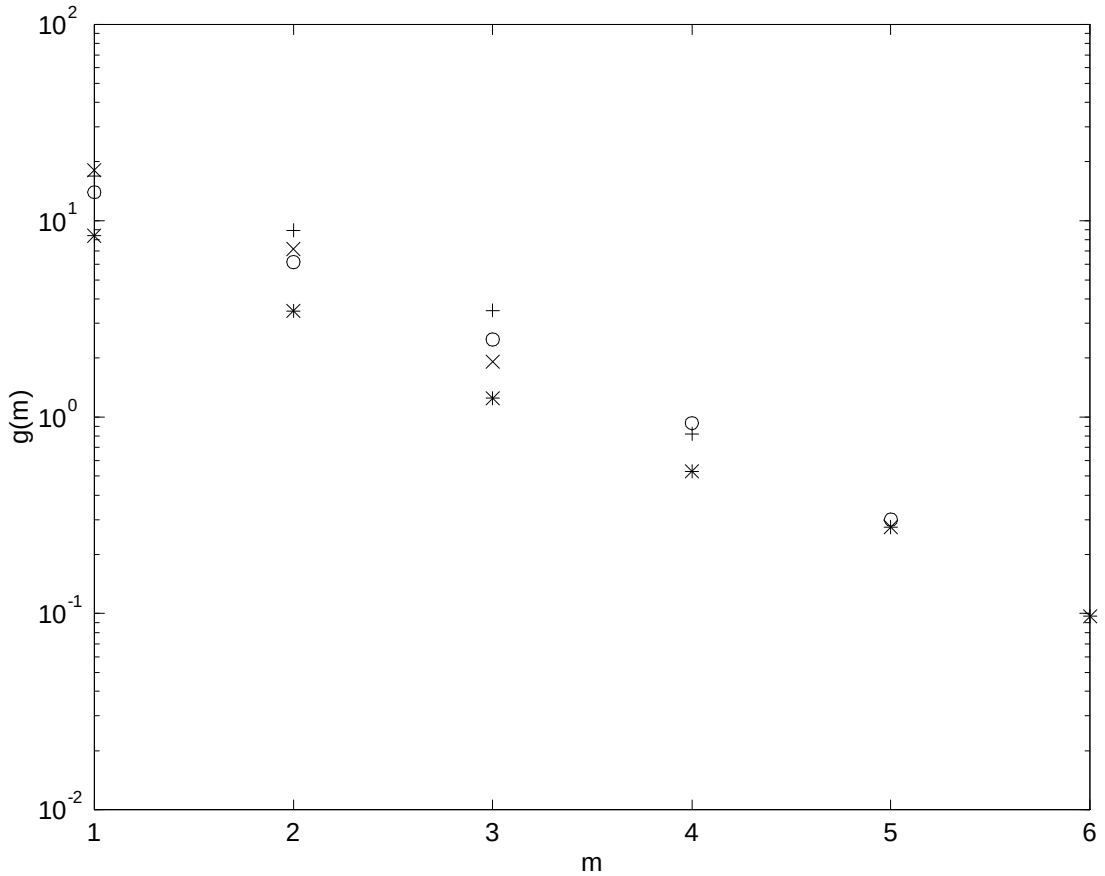
$$\gamma_m = \max_{a_0 a_1 \cdots a_m} g\left(a_0 a_1 \cdots a_m\right)$$

If the statistics for long blocks is poor $\Rightarrow$ large fluctuations in γ_m

- Better estimate of the decay behavior with $\overline{g(m)} = \overline{g\left(a_0 a_1 \cdots a_m\right)}$, the average being taken over all sets $a_0 a_1 \cdots a_m$ of size m.

Chains with complete connections

$g(m)$ computed using A blocks of length 5 to 8 $(\times, +, o, *)$



Chains with complete connections

- Result compatible with exponential decay
  $\Rightarrow$ summability of the γ_m's

Chains with complete connections

- Result compatible with exponential decay
  $\Rightarrow$ summability of the γ_m's
- **A CCC-process with summable decays is the $\overline{d}-$limit of its Markov approximations of order k**

Chains with complete connections

- Result compatible with exponential decay
  $\Rightarrow$ summability of the γ_m's
- **A CCC-process with summable decays is the $\overline{d}-$limit of its Markov approximations of order k**
- *$\overline{d}-$distance*
  Coupling between two processes $X = \{X_n\}$ and $Y = \{Y_n\}$ is another process $\left\{\widetilde{X}_n, \widetilde{Y}_n\right\}$ over $\Sigma \times \Sigma$ such that the marginal probabilities of $\widetilde{X}$ and $\widetilde{Y}$ coincide with those of X and Y

$$\overline{d}(X, Y) = \inf\left\{P\left(\widetilde{X}_0 \neq \widetilde{Y}_0\right) : \left\{\widetilde{X}_n, \widetilde{Y}_n\right\} \quad \begin{array}{l}\text{is a stationary} \\ \text{coupling of X and Y}\end{array}\right\}$$

Chains with complete connections

- Result compatible with exponential decay
  $\Rightarrow$ summability of the γ_m's
- **A CCC-process with summable decays is the $\overline{d}-$limit of its Markov approximations of order k**
- *$\overline{d}-$distance*
  Coupling between two processes $X = \{X_n\}$ and $Y = \{Y_n\}$ is another process $\left\{\widetilde{X}_n, \widetilde{Y}_n\right\}$ over $\Sigma \times \Sigma$ such that the marginal probabilities of $\widetilde{X}$ and $\widetilde{Y}$ coincide with those of X and Y

$$\overline{d}(X, Y) = \inf\left\{P\left(\widetilde{X}_0 \neq \widetilde{Y}_0\right) : \left\{\widetilde{X}_n, \widetilde{Y}_n\right\} \quad \begin{array}{l}\text{is a stationary} \\ \text{coupling of X and Y}\end{array}\right\}$$

- $\overline{d}-$distance tending to zero does mean that the processes will coincide after a certain time

Chains with complete connections

- Result compatible with exponential decay
  $\Rightarrow$ summability of the γ_m's
- **A CCC-process with summable decays is the $\overline{d}-$limit of its Markov approximations of order k**
- *$\overline{d}-$distance*
  Coupling between two processes $X = \{X_n\}$ and $Y = \{Y_n\}$ is another process $\left\{\widetilde{X}_n, \widetilde{Y}_n\right\}$ over $\Sigma \times \Sigma$ such that the marginal probabilities of $\widetilde{X}$ and $\widetilde{Y}$ coincide with those of X and Y

$$\overline{d}(X, Y) = \inf\left\{P\left(\widetilde{X}_0 \neq \widetilde{Y}_0\right) : \left\{\widetilde{X}_n, \widetilde{Y}_n\right\} \quad \begin{array}{l}\text{is a stationary} \\ \text{coupling of X and Y}\end{array}\right\}$$

- $\overline{d}-$distance tending to zero does mean that the processes will coincide after a certain time
- **Perfect simulation** always understood in the $\overline{d}-$distance sense. It does not mean **perfect prediction** (Means that a process is constructed with the same conditional probabilities of the original one)

Chains with complete connections

- **Simulation scheme by the sequence of canonical Markov approximations of finite order k ($k-$CMA)**
  $k-$CMA of a process X is a Markov chain $Y^{(k)}$ of order k with conditional probabilities $P^{(k)}$

$$P^{(k)}\left(a_0 \middle| a_1 \cdots a_k\right) = P\left(a_0 \middle| a_1 \cdots a_k\right) = \sum_A P\left(a_0 \middle| a_1 \cdots a_k A\right)$$

Chains with complete connections

- **Simulation scheme by the sequence of canonical Markov approximations of finite order k ($k-$CMA)**
  $k-$CMA of a process X is a Markov chain $Y^{(k)}$ of order k with conditional probabilities $P^{(k)}$

$$P^{(k)}\left(a_0|a_1\cdots a_k\right) = P\left(a_0|a_1\cdots a_k\right) = \sum_A P\left(a_0|a_1\cdots a_k A\right)$$

- For a CCC X with summable decays

$$\overline{d}\left(X, Y^{(k)}\right) \leq C\gamma_k \tag{2}$$

The property of the Markov approximation, essential for the approximation result (2), is

$$\inf_A P\left(a_0|a_1\cdots a_k A\right) \leq P^{(k)}\left(a_0|a_1\cdots a_k\right) \leq \sup_A P\left(a_0|a_1\cdots a_k A\right) \tag{3}$$

meaning that for Markov approximation schemes, other than the canonical one, Eq.(2) holds provided (3) is satisfied

Looking for a Gibbs measure

- **For the market fluctuation data**:
  $\leq k-$Markov approximation:
  i) Empirical transition probabilities $\widetilde{P}\left(a_0|a_1\cdots a_m\right)$ inferred from the probability of blocks of order $m+1$. up to m_{Max}
  ii) $\leq k-$Simulation: look at the current block $\left(a_1\cdots a_k\right)$ and use the $k-$empirical probability to infer the next state a_0. If that block has not appeared in the trainig data, use the $k-1$ sized block $a_2\cdots a_k$ and the $k-1$ order empirical probabilities

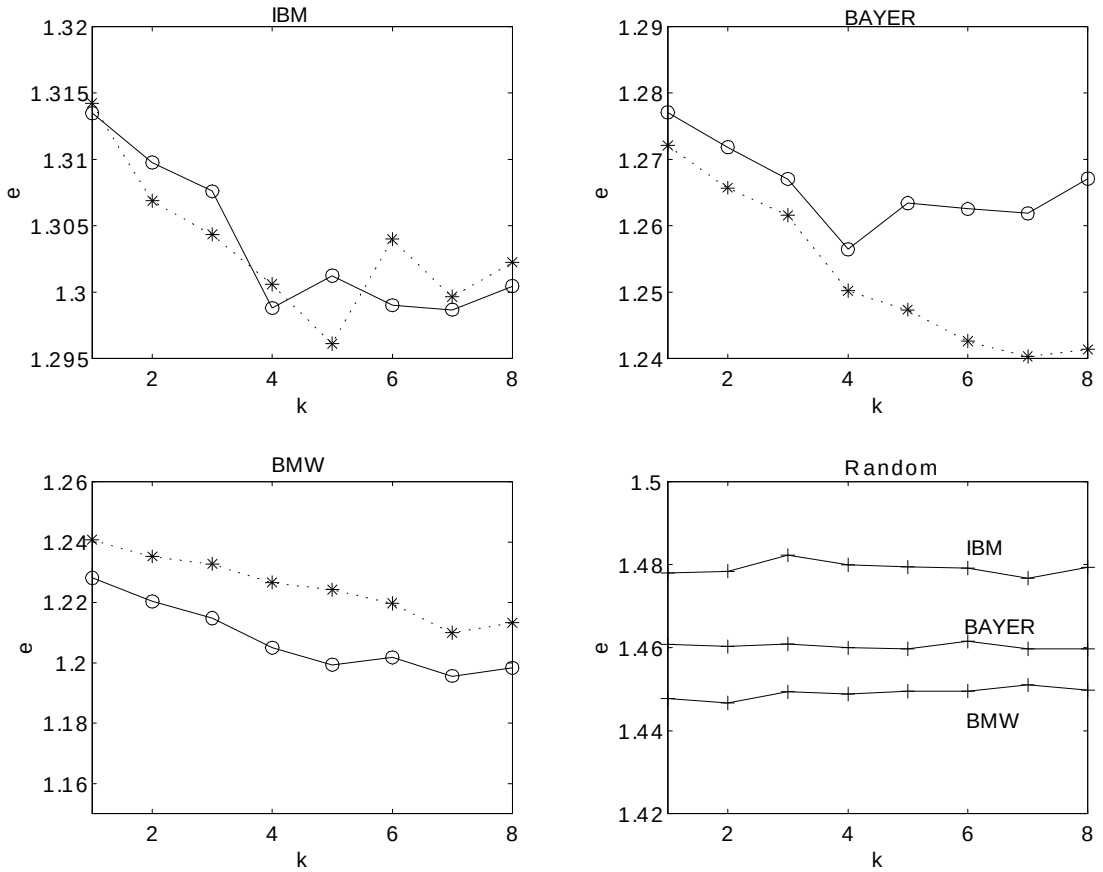
Looking for a Gibbs measure

- **For the market fluctuation data**:
  $\leq k-$Markov approximation:
  i) Empirical transition probabilities $\widetilde{P}\left(a_0|a_1\cdots a_m\right)$ inferred from the probability of blocks of order $m+1$. up to m_{Max}
  ii) $\leq k-$Simulation: look at the current block $\left(a_1\cdots a_k\right)$ and use the $k-$empirical probability to infer the next state a_0. If that block has not appeared in the trainig data, use the $k-1$ sized block $a_2\cdots a_k$ and the $k-1$ order empirical probabilities
- Averaged squared error

$$e^2=\left\langle\left(\widetilde{a}_0-a_0\right)^2\right\rangle$$

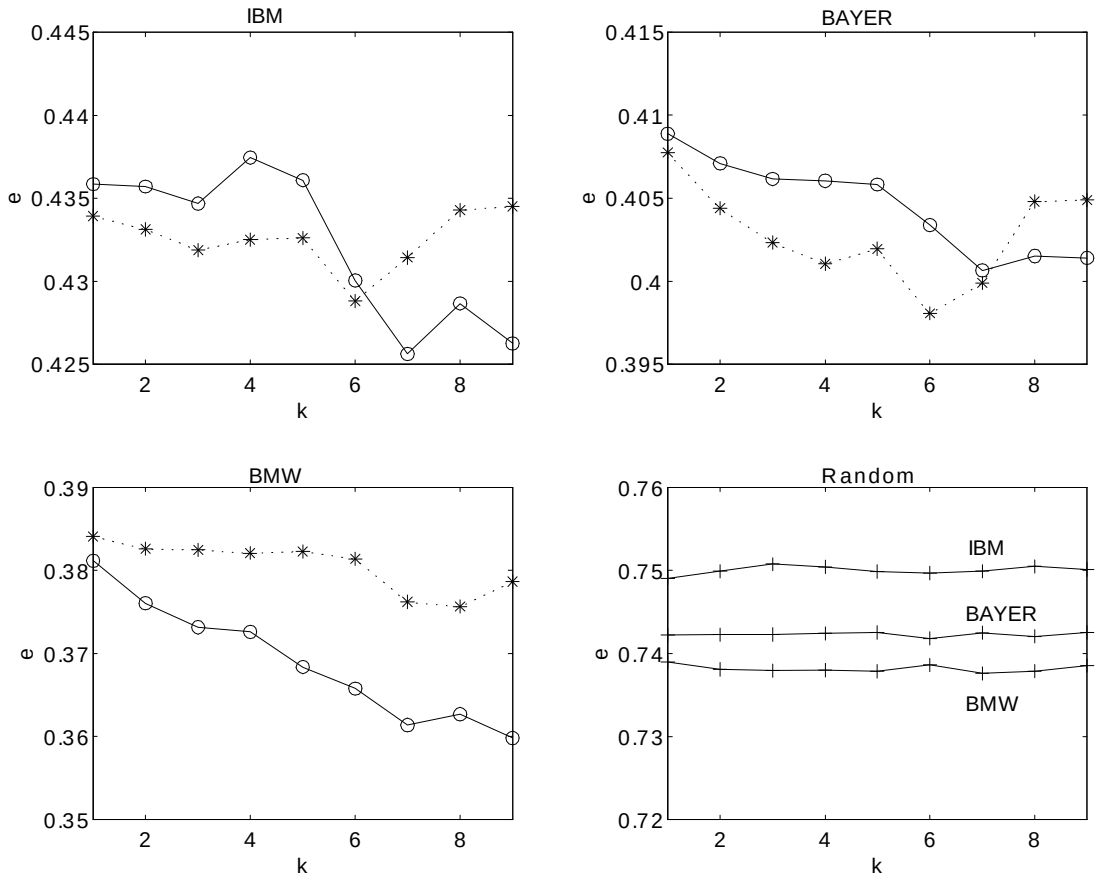
Chains with complete connections

The past predicting the future (o) and the future predicting the past (*), compared to random choice



Chains with complete connections

The past predicting the future (o) and the future predicting the past (*), compared to random choice



Chains with complete connections

Main conclusions:

- Average prediction better than random choice.
- Main improvement results from correct accounting of the two-symbol probabilities ($k = 1$)
- Small (but consistent) improvement using past information up $k = 4$ or 5. No significant improvement using higher order approximations
- Bulk of data represented by a short-memory process. Nevertheless there is evidence for a small long-memory component that is captured by the higher-order Markov approximations
- There is a maximum $k = k_m$ that should be used for the simulation process
- Inter-companies prediction: improvement coming from the one-symbol probabilities (as compared to random choice) is obtained. For the long-memory component the behavior is company-dependent.

Statistical analysis of market data

General conclusions

1. Bulk of the market fluctuation process is a short-memory process. In addition it has a small long-memory component associated with the large fluctuations of the returns.
2. Existence of the long-memory component suggests the *chains with complete connections and summable decays* as a framework
3. Although the decays may be exponentially converging, the lack of accurate data for long blocks prevents an accurate description by a finite range Gibbs potential.
4. The sequence of empirical based $\leq k-$Markov approximations seems the most unbiased simulation of the process. Eventual convergence in the $\overline{d}-$distance sense, because of summable decays.
5. For the future: high frequency market data. Beware of the possibly multi-scale and multi-component nature of the processes.

References

J.-R. Chazottes, E. Floriani and R. Lima; *Relative entropy and identification of Gibbs measures in dynamical systems*, J. of Stat. Phys. 90 (1998) 697-725.

E. F. Fama; *Market efficiency, long-term returns and behavioral finance*, Journal of Financial Economics 49 (1998) 283-306.

R. B. Olsen, M. M. Dacorogna, U. A. Müller and O. V. Pictet; *Going back to basics - Rethinking market efficiency*, Olsen \& Associates discussion paper RBO.1992-09-07.

D. Ruelle; *Thermodynamic formalism*, Encyclopedia of Mathematics and its Applications 5, Addison-Wesley 1978.

O. Onicescu and G. Mihoc; *Le comportment asymptotique des chaines à liaisons complètes*; Disq. Math. Phys. 1 (1940) 61-62.

M. Iosifescu and S. Grigorescu; *Dependence with complete connections and its applications*, Cambridge U. P., Cambridge 1990.

D. S. Ornstein and B. Weiss; *How sampling reveals a process*, The Annals of Prob. 18 (1990) 905-930.

References

X. Bressaud, R. Fernandez and A. Galves; *Speed of $\overline{d}-$convergence for Markov approximations of chains with complete connections. A coupling approach*, Stoch. Proc. and Appl. 83 (1999) 127-138.

F. Comets, R. Fernandez and P. A. Ferrari; *Processes with long memory: Regenerative construction and perfect simulation*, arXiv:math.PR/0009204.

RVM, R. Lima and T. Araújo; *A process-reconstruction analysis of market fluctuations,* Int. Journal of Theoretical and Applied Finance, 5 (2002) 797