Working Paper 97-27
Statistics and Econometrics Series 14
March 1997

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

INFORMATION-THEORETIC ANALYSIS OF SERIAL DEPENDENCE AND COINTEGRATION

F.M. Aparicio and A. Escribano.*

Abstract

This paper presents a wider characterization of memory in time series and of cointegration in terms of information-theoretic statistics such as the entropy and the mutual information between pairs of variables. This suggests a new methodology for exploratory data analysis and for testing the hypothesis of long-memory and of the existence of a cointegrating relationship. We illustrate the performances of the new techniques with some simulation experiments, and finally apply them to the analysis of the relationship between pairs of financial time series from a foreign exchange-rate and a stock return markets.

Keywords:
Information-theoretic statistics, long-memory, cointegration, nonlinearity, causality.

*Aparicio, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. C/ Madrid, 128 28903 Madrid. Spain. Ph: 34-1-624.98.71, Fax: 34-1-624.98.49, e-mail: aparicio@est-econ.uc3m.es; Escribano, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. This research, initiated at the Department of Economics (Aparicio and Granger, 1995) and the Institute for Nonlinear Science of the University of California and San Diego, was partially supported by the Spanish DGICYT under Grant PB95-0298. We are grateful to Ignacio Peña and to Z. Ding for kindly providing the foreign exchange-rate and the stock return series, respectively.

1 Introduction

Many economic time series exhibit important random changes in their mean behaviour. These series are sometimes said to be *integrated*, since it is possible to simulate the most important features in their patterns with sums of an increasing number of weakly-dependent random variables. Integrated series can be expressed in terms of the *unobserved components model*, where one of the components is a *stochastic trend*. The fact that remote shocks have a persistent influence on the levels of these series is known as the *long-memory* or the *extended-memory* property, depending on whether this influence is linear or not (Granger, 1995 [10]).

In some cases, the accumulated changes in mean behaviour may be correlated accross series. In the context of macroeconomics and finance, certain models suggest the presence of economic or social forces preventing two or more series from drifting too far apart from each other. Pairs of series which exhibit a common long-memory component or stochastic trend are said to be *cointegrated.* The concept of *cointegration* was coined by Granger (1981 [7]). and later on developed by Engle and Granger. (1987 [4]). Well-known examples of cointegrating relationships can be found between income and expenditure. prices of a particular good in different markets, interest rates in different parts of a country. etc.

Underlying the idea of cointegration is the existence of a long-run *equilibrium* (i.e. a deterministic relationship that holds on the average for the levels) between two integrated variables. $x_t. y_t$. A strict (linear) equilibrium exists when for some $a \neq 0$. one has $y_t = ax_t$. This unrealistic situation is replaced. in practice, by that of a (linear) cointegrating relationship, in which the equilibrium error $z_t = y_t - ax_t$ is different from zero but fluctuates around this value much more frequently than the individual series (i.e. z_t is mean-reverting), while the size of these fluctuations could be much smaller.

It turns out that many apparently non-cointegrated series may have a *nonlinear equilibrium.* Unfortunately, conventional cointegration tests tend to have low power when nonlinearity enters in the relationship between the variables. It is therefore important to investigate new methods capable of detecting equilibriums others than linear, and of rejecting the linear cointegration assumption when false.

There have been attempts to address this problem. For example, Hallman (1990) [13] proposed to apply standard non-cointegration tests (*unit-root* tests) to the ranks rather than to the levels of the series in order to robustize these tests against mononotonic nonlinear transformations of cointegrated variables. However, this strategy could not cope with more complex types of nonlinearities in the relationship. Moreover, Hallman's approach relies on an assumption of invariance of the distributional properties of the conventional tests when applied to the ranks.

Granger and Hallman (1991) [11] proposed estimating the nonlinear transformations using a nonparametric technique known as the *Alternate Conditional Expectation* (ACE) algorithm (Breiman and Friedman, 1985 [2]). This was followed by a standard cointegration test applied on the transformed variables obtained using the ACE estimates. Further, these estimates also allowed the possibility of testing the hypothesis of linearity in cointegration. However, the estimation and the inference properties of ACE estimates rely on the stationarity and ergodicity of the series, properties which exclude integrated variables. Moreover, as remarked by these authors, it is not yet clear how nonparametric estimators of the transformations affect the distribution of the standard cointegration test statistics.

The previous difficulties call for a new characterization of cointegration which could be used to test this hypothesis in a general context (i.e. where nonlinearity is allowed), and without requiring prior estimation of the nonlinearities.

In this paper, we review the concepts of *mean-reversion*, *short* and *long memory*, and *cointegration*, and introduce a new characterization of these properties using *information-theoretic* ideas. This will lead us to proposing some new schemes for exploratory data analysis and for testing the hypothesis of long-memory and of cointegration between two long-memory time series. Although the focus of this paper is on the univariate case, these ideas can be readily applied in a multivariate context.

The rest of the paper is structured as follows. Section 2 introduces a general framework for analyzing mean-reversion, short(long)-memory, and cointegration, in order to deal with nonlinearity. Section 3 presents the information-theoretic tools to be used later. In particular, we introduce the definitions of *entropy* and *mutual information* for random variables and for stochastic processes. In section 4, we propose an interpretation of dependence in and among time series using the previous tools, which lead us to a more general definition of long-memory and cointegration. In section 5 we turn the previous characterization into exploratory tests of long-memory and of cointegration. Sections 6 and 7 present for our cointegration analysis, some simulations results, and a real-world experiment on financial data from a stock and a foreign exchange-rate market. Finally, section 8 gives a concise summary of the paper.

2 Towards a general characterization of memory and cointegration

There are important drawbacks with the standard definitions of long memory and of cointegration when dealing with non-Gaussian time series, and with pairs of series which are nonlinearly related. In the first case, the trouble is that the autocorrelation function (ACF) fails to capture the higher-order dependencies in the data. In the second, that series which do not appear to be "aligned" in their mean behaviour could be cointegrated after being nonlinearly transformed. In fact, what we need is a different measure of serial dependence, and to reformulate the cointegration concept in terms of the latter.

2.1 A general characterization of memory in time series

The standard characterization of memory in a time series x_t is given in terms of its ACF, say $\rho_x(\tau, t) = cov(x_t, x_{t-\tau})/var(x_t)$, which we consider to be generally dependent on a time index, so as to allow for some heterogeneity.

Definition 1 *A process x_t is said to be* **mean-reverting** *if* $\forall t \lim_{\tau \to \infty} \rho_x(\tau, t) = 0$.

Intuitively, the process x_t is mean-reverting if $x_t - E(x_t)$ changes sign with nonzero probability. When the process is not mean-reverting, its memory span is necessarily larger since $\lim_{\tau \to \infty} \rho_x(\tau, t) > 0$, and thus any two infinitely distant variables from the process are still correlated (persistent behaviour).

However, even for a mean-reverting process, the memory span can be very large in the sense that its ACF decays very slowly as τ grows. This motivates the distinction between *short* and *long memory*.

Definition 2 *A process x_t is said to be* **short-memory** *if* $\forall t$ $\exists b_t < \infty$ *such that* $\sum_{\tau > 0} \rho_x(\tau, t) = b_t$.

Definition 3 *A process x_t is said to be* **long-memory** *if* $\forall t$ $\sum_{\tau > 0} \rho_x(\tau, t) = \infty$.

Definition 4 *A time series of x_t is said to be* **integrated of order d**, *in short $x_t \sim I(d)$, if $\sum_{\tau>0} \rho_x(\tau,t) = \infty$, $\forall t$, and d is the smallest positive real number such that $\sum_{\tau>0} \rho_z(\tau,t) < \infty$, $\forall t$, with $z_t = (1-B)^d x_t$.*

The parameter d which appears in this latter definition serves to quantify the memory in the series.

The previous characterization of memory in terms of the ACF is adequate for Gaussian series, since all the dependence structure is captured by its second order moments. With non-Gaussian time series, in particular, nonlinear time series, the ACF cannot provide a full account of the serial dependence structure. A first attempt to establish a general characterization of memory in a non-Gaussian context was due to Granger and Terasvirta (1993 [12]). They proposed a general definition of mean-reversion in terms of the conditional distribution function of the process. Let X_t denote the *r.v.* at time t from a time series of a stochastic process x_t, and let $F_h(x) = P(X_{t+h} \leq x|I_t)$ represent the conditional distribution function of the *r.v.* X_{t+h} given its h-horizon past. $I_t = \mathcal{F}_x^{-\infty,t}$, where $\mathcal{F}_x^{-\infty,t}$ denotes the σ-field generated by the *r.v.*'s $X_t, X_{t-1}, \cdots$.

Definition 5 *A process x_t has* **no extended-memory** *if $\lim_{h-\infty} F_h(x)$ does not depend on the conditioning past, I_t.*

As a consequence, for any Borel sets C_1, C_2 and for any integer k such that $P(X_{t-k} \in C_2) > 0$, we would have

$$\lim_{h-\infty} \mid P(x_{t+h} \in C_1 | x_{t-k} \in C_2) - P(x_{t+h} \in C_1) \mid = 0 \tag{1}$$

This property reminds the concept of *ϕ-mixing*, since it means that the dependence among temporarily nonoverlapping blocks of *r.v.*'s from the process vanishes in the limit, when the temporal distance between the blocks becomes infinite.

A major shortcoming of this definition is that it cannot be easily checked in practice. In the sequel, we propose a straightforward generalization of the memory concept for time series, based on conditions which can be easily tested. For this, we only need a measure of serial dependence which generalizes the ACF. Suppose $i_x(\tau,t)$ is this new serial dependence measure that captures the higher-order dependence structure in the series [1]. A most general characterization of mean-

[1] We will later on propose a useful candidate for this measure based on the mutual information concept.

reversion and of short, long-memory and integration could then be proposed using this measure. A process x_t could be said to be:

- **mean-reverting** in $i(.)$, if $\forall t \lim_{\tau\to\infty} i_x(\tau, t) = 0\ \forall t$.
- **short-memory** in $i(.)$, if $\forall t\ \sum_{\tau>0} i_x(\tau, t) < \infty$.
- **long-memory** in $i(.)$, if $\forall t\ \sum_{\tau>0} i_x(\tau, t) = \infty$.
- **integrated of order d** in $i(.)$, say $x_t \sim II(d)$, if $\sum_{\tau>0} i_x(\tau, t) = \infty$, $\forall t$, and d is the smallest positive real number such that $\sum_{\tau>0} i_z(\tau, t) < \infty$, $\forall t$, with $z_t = (1-B)^d x_t$.

Remarks:

1. In principle, the function $i_x(\tau, t)$ could be any serial dependence measure capable of capturing nonlinear dependencies between the variables in the series. Remark that $\sum_{\tau=1}^{\infty} i_x(\tau, t)$ rather than on $\sum_{\tau=1}^{\infty} \rho_x(\tau)$, with $\rho_x(\tau)$ representing the ACF of x_t, is used as a *persistence measure* for non-Gaussian time series.
2. Note that the rates of convergence of $i_x(\tau, t)$ towards 0 as $\tau \to \infty$ will be different for long- and for short-memory processes. Also remark that a short-memory process is also mean-reverting, according to these definitions.

2.2 A general characterization of cointegration

The standard definition of cointegration goes as follows:

Definition 6 *(Granger, 1981 [7]) Two long-memory time series x_t, y_t, with long-memory parameter d, are said to be (linearly [2]) **cointegrated** if $\exists a \in \Re - 0$ such that the series $z_t = y_t - ax_t$ is $I(d_z)$ with $d_z < d$.*

Figure 1 illustrates a simulation example of linear cointegration with a pair of correlated random walks ($d = 1$) and for $a = 0.72$. The scatter plot clearly shows the linearity of the relationship between x_t and y_t.

[2] In Granger (1983) [8], there is no explicit mention to the term *linear*, although it is implicit.

An important shortcoming in this definition of cointegration is that it requires the cointegrating relationship between the series to be linear. As as consequence, classical cointegration testing techniques relying on these definitions yield misleading results when nonlinearity enters the true equilibrium relationship. Evidence of this problem with definition 6 was first reported by Hallman (1990) [13], who proposed applying standard cointegration tests to the ranks rather than the levels of the series. However, even though this trick succeeds in robustizing the test against monotonic nonlinearities, it fails when confronted to general forms of nonlinearity.

In general, it should be possible to find time series that are cointegrated only after applying certain nonlinear transformations on them. Indeed, an extension of the (linear) cointegration concept follows by noticing that the common low-frequency component may "live" in a higher-order moment than the mean, that is, in nonlinear transformations of the series. For example, x_t and y_t could be cointegrated when squared, while being more or less uncorrelated in their levels. To explain, suppose $y_t = x_t \epsilon_t$, with x_t an $I(1)$ series, and ϵ_t a zero mean *i.i.d.* sequence, and thus $x_t \sim I(0)$. It follows that $(y_t)^2 = \sigma_\epsilon^2 x_t^2 + (\epsilon_t^2 - \sigma_\epsilon^2) x_t^2$, where the rightmost term must be short-memory since it is the product of an $I(0)$ process $(\epsilon_t^2 - \sigma_\epsilon^2)$ and an $I(1)$ process (x_t^2). Thus $(y_t)^2$ is linearly cointegrated with $(x_t)^2$, although y_t is not cointegrated with x_t.

Example 1:

Consider the following nonlinear factor model

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} a \\ 1 \end{pmatrix} w_t + \begin{pmatrix} -b \\ 0 \end{pmatrix} w_t^2 + \begin{pmatrix} v_t \\ \xi_t \end{pmatrix} \tag{2}$$

where $a \neq 0$, $w_t = w_{t-1} + \epsilon_t$ with $w_0 = 0$, and (v_t, ξ_t, ϵ_t) are independent sequences of independent and identically Normally distributed *r.v.*'s with zero mean and joint covariance matrix equal to the identity matrix. Let $\beta'_{l,\perp} = (a, 1)$, and let $\beta'_{n,\perp} = (-b, 0)$. Thus the orthogonal complements of $\beta'_{l,\perp}$ and $\beta'_{l,\perp}$ are respectively $\beta'_l = (1, -a)$ and $\beta'_n = (0, b)$. The nonlinear cointegrating relationship can be obtained as

$$\begin{aligned} z_t &= \beta'_l \begin{pmatrix} y_t \\ x_t \end{pmatrix} + \beta'_n \begin{pmatrix} y_t \\ x_t^2 \end{pmatrix} \qquad (3) \\ &= y_t - a x_t + b x_t^2. \end{aligned}$$

Thus the cointegration errors are given by $z_t = 2bw_t\xi_t + b\xi_t^2 + v_t - a\xi_t$, and it can be easily shown that they are short-memory according to our definition.

Figure 2 illustrates a simulation experiment of nonlinear cointegration with series having a common factor, and obtained with the model (2), with $a = 2.0$ and $b = 0.05$. Figure 3 shows a real example of an apparently nonlinearly cointegrating relationship. In both cases, the scatter plots below clearly show that the dependence between the variables is not linear.

Some previous concepts of nonlinear cointegration are the following:

Definition 7 *(Granger and Hallman,1991 [11]) A pair of series x_t, y_t, are said to have a cointegrating* **nonlinear attractor** *if there are nonlinear measurable functions $f(.), g(.)$ such that $f(x_t)$ and $g(y_t)$ are both $I(d)$, $d > 0$, and $z_t = g(y_t) - f(x_t)$ is $\sim$ $I(d_z)$, with $d_z < d$.*

Remark:
Assuming that f and g can be expanded as Taylor series up to some order $p \geq 2$ around the origin, we may write $z_t = c_0 + c_1 u_t + HOT(x_t, y_t)$, where $u_t = y_t - ax_t$, and $HOT(.,.)$ denotes higher-order terms. It follows that the linear approximation, u_t, to the true cointegration residuals differs from the latter by some higher-order terms. These terms express that the *strengh of attraction* onto the cointegration line $y_t = ax_t$ may vary with the levels of the series, x_t, y_t, when nonlinearities exist in their relationship.

As stated in the introduction, a difficulty with the application of this definition is the need to find proper estimates of the cointegrating functions $f(.)$ and $g(.)$ in order to test for cointegration.

Escribano and Mira (1996) [5] propose the following alternative definition of nonlinear cointegration based on the concepts of *α-mixing* (Rosenblatt, 1974 [18]) and *near-epoch dependence* (NED) (Wooldridge, 1986 [20]).

Definition 8 *(Escribano and Mira, 1996 [5]) A pair of series x_t, y_t, are* **nonlinear cointegrated** *with cointegration function $g(\ldots, \gamma)$ (where γ is a parameter), if $g(y_t, x_t, \gamma^*)$ is NED (α-mixing) on some α-mixing series, but $g(y_t, x_t, \gamma)$ is not NED (α-mixing) for any $\gamma \neq \gamma^*$.*

Unfortunately, this definition relies on concepts of dependence that are generally difficult to check in practice.

We propose now a most general characterization of cointegration which circumvents some of the difficulties encountered with the previous ones.

Let x_t, y_t be time series from processes that are long-memory in $i(.)$, and let $i_{x,y}(\tau, t)$ represent a general measure of serial cross-dependence between x_t, y_t.

Definition 9 *A pair of time series x_t, y_t, that are long-memory in $i(.)$, are said to be* **cointegrated** *in $i(.)$ (in short, CII) if*

$$\lim_{\tau \to \infty} \frac{i_{x,y}(\tau, t)}{i_x(\tau, t)} = b, \ \forall t \tag{4}$$

where b is a nonzero and finite real number.

Remarks:

1. Intuitively, the definition states that, under cointegration, the remote past of y_t should be as useful as the remote past of x_t in long-term forecasting x_t. A particular feature of this characterization is that it focusses on the relative behaviour of measures of serial autodependence and of cross-dependence at long lags.

2. This more general characterization of cointegration relies on the different limit behaviour of $i_x(\tau, t)$ and $i_{x,y}(\tau, t)$, under non-cointegration. If cointegration holds, we cannot have different convergence rates for $i_x(\tau, t)$ and for $i_{x,y}(\tau, t)$. The possibly different rates of convergence could be used to construct a measure of the degree of non-cointegration. Suppose that $i_x(\tau, t) \sim \tau^{-\alpha}$, and that $i_{x,y}(\tau, t) \sim \tau^{-\beta}$ for τ large enough. In numerical applications we may find that neither $i_{x,y}(\tau, t)$ nor $i_x(\tau, t)$ is either infinite or zero for any finite τ. So we may safely take the logarithm of the ratio $i_x(\tau, t)/i_{x,y}(\tau, t)$ and plot it as a function of $log\tau$. This function will tend towards an asymptote as τ grows to infinity. The slope of this asymptote is just $\alpha - \beta$, and it is always non-negative, since we expect that $\alpha \leq \beta$. Thus the larger its value the farther the hypothesis of information-cointegrateness between the series is from being realized.

3. If we replace $i_x(\tau, t)$ by the ACF of x_t, and $i_{x,y}(\tau, t)$ by the cross-correlation function between x_t and y_t, say $\rho_{x,y}(\tau, t)$, then our definition becomes a re-statement of the standard definition of linear cointegration proposed by Granger (1981) [7], and amounts at comparing the behaviour at the origin of the spectral densities of the series.

An alternative condition for cointegration is the following one. Let $S_n^{(x,y)} = \sum_{\tau=1}^{n} i_{x,y}(\tau,t)$.

Proposition 1 *If the series y_t, x_t are* **cointegrated** *in $i(.)$ then the sequence of partial sums $S_n^{(x,y)}$ diverges as $n \rightarrow \infty$.*

PROOF:

Suppose the series are cointegrated in $i(.)$. Then from our definition, it follows that there exists a nonzero real number $b = \sup_t(b_t)$ and a finite real number C such that $\lim_{n\rightarrow\infty} S_n^{(x,y)} = b \lim_{n\rightarrow\infty} S_n^{(x,x)} + C$. And the divergence of $S_n^{(x,y)}$ follows from the divergence of $S_n^{(x,x)}$, since x_t has long memory in $i(.)$.

3 Some information-theoretic measures of data variability and dependence

In this section we present the information-theoretic concepts which will form the basis of the new characterization that we proposed for the relationship between integrated time series.

3.1 Information-theoretic measures for partitions

A most basic problem in information theory is that of assigning a measure of uncertainty to the ocurrence or nonocurrence of any event in a partition $\mathcal{P}$ of the set of outcomes of an underlying experiment. We call this measure of uncertainty the *entropy* of the partition, and denote it by $H(\mathcal{P})$. The construction of this functional stems from some postulates which must be satisfied in order to provide such measure of uncertainty. Suppose now that we have a partition of a sample space $\mathcal{S}$ with M events $\mathcal{A}_i$, $i = 1, \cdots, M$, and that the event $\mathcal{A}_i$ occurs with probability p_i. It can be shown that the convex functional

$$H(\mathcal{P}) = -\sum_{i=1}^{M} p_i log(p_i) \tag{5}$$

yields a proper measure of average uncertainty in the partition $\mathcal{P}$.

Similarly, when we know about the ocurrence of a subset $\mathcal{M}$ of events from a different partition, $\mathcal{Q}$ of $\mathcal{S}$, the remaining uncertainty in the partition $\mathcal{P}$ can be measured by the nonnegative functional

$$H(\mathcal{P}|\mathcal{M}) = -\sum_{i=1}^{M} P(\mathcal{A}_i|\mathcal{M}) log P(\mathcal{A}_i|\mathcal{M}), \tag{6}$$

which is called the *conditional entropy* of $\mathcal{P}$ given $\mathcal{M}$. Notice that if the events in $\mathcal{P}$ are independent of those in $\mathcal{M}$ then $H(\mathcal{P}|\mathcal{M}) = H(\mathcal{P})$. In general, $\mathcal{M}$ may convey information about the events in $\mathcal{P}$, and this *mutual information* can be quantified by the functional

$$I(\mathcal{P},\mathcal{M}) = H(\mathcal{P}) - H(\mathcal{P}|\mathcal{M}). \tag{7}$$

That is, the observation of $\mathcal{M}$ reduces the uncertainty about $\mathcal{P}$ from $H(\mathcal{P})$ to $H(\mathcal{P}|\mathcal{M})$, so the information that $\mathcal{M}$ conveys about $\mathcal{P}$ is just $I(\mathcal{P},\mathcal{M})$. Notice that $\mathcal{M}$ can convey at most $H(\mathcal{P})$ bits of information about the events in $\mathcal{P}$, and since $H(\mathcal{P}|\mathcal{M}) < H(\mathcal{P})$, $I(\mathcal{P},\mathcal{M})$ must also be nonnegative.

Now let us denote by $H(\mathcal{P},\mathcal{Q})$ the *joint entropy* functional for the partition whose events are the intersections of the events in $\mathcal{P}$ and $\mathcal{Q}$. The resulting partition is called a *refinement* of both $\mathcal{P}$ and $\mathcal{Q}$. Notice that to observe the joint partition we must observe both $\mathcal{P}$ and $\mathcal{Q}$. It follows that the uncertainty in the joint partition must be at least equal to that of the elementary partitions. Rigorously speaking, by convexity of the entropy functional it is easy to show that $H(\mathcal{P},\mathcal{Q}) \geq H(\mathcal{P})$ and that $H(\mathcal{P},\mathcal{Q}) \geq H(\mathcal{Q})$ (i.e. Papoulis, 1991 [16]). In fact, we have

$$\begin{aligned} H(\mathcal{P},\mathcal{Q}) &= H(\mathcal{Q}) + H(\mathcal{P}|\mathcal{Q}) \\ &= H(\mathcal{P}) + H(\mathcal{Q}|\mathcal{P}) \qquad (8) \\ &\leq H(\mathcal{P}) + H(\mathcal{Q}) \qquad (9) \end{aligned}$$

Clearly, the maximum value of $H(\mathcal{P},\mathcal{Q})$ is attained when $\mathcal{P}$ and $\mathcal{Q}$ are independent. Also, by manipulating equations (7) and (8), we obtain

$$I(\mathcal{P},\mathcal{M}) = H(\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P},\mathcal{Q}). \tag{10}$$

3.2 Information-theoretic measures for random variables

So far we have introduced the concept of entropy of a given partition of the sample space of an experiment. It is possible to define the entropy of a *r.v.* by forming a suitable partition. This is straightforward for discrete-valued *r.v.*'s. For example, if a *r.v.* X takes a countable set of values $\{x_i\}$, $i = 1, 2, \cdots$, with probabilities p_i, we can form the partition in which each event corresponds to a different value of X. Thus the definition of entropy as given in the previous paragraph also applies here, and we can define the entropy of the *r.v.* X as

$$H(X) = -\sum_i p_i log(p_i). \tag{11}$$

The definitions for the rest of the uncertainty measures discussed in the preceeding paragraph, such as conditional and joint entropies, and the mutual information, remain also valid in this case. When dealing with continuous-valued *r.v.*'s the extension of these concepts is not immediate. The difficulty here is that the events $\{X = x_i\}$ do no longer form a partition, since they are not countable. Therefore, to define the entropy we must first convert X into a discrete- valued *r.v.*. That is, we can define the entropy of a quantized version of X given by $X_\delta = m\delta$ if $X \in (m\delta - \delta, m\delta]$. If we assume that X has a probability density function (*pdf*), $f_x()$ is then easy to show that

$$\lim_{\delta \to 0}\,[H(X_\delta) + log\delta] = -\int_{-\infty}^{\infty} f_x(X) log f_x(X) dX. \tag{12}$$

We remark that $\lim_{\delta \to 0} H(X_\delta) = \infty$. However, in practice, we can only observe X with finite accuracy because of noise and quantification errors from the measurement instrument. Since the term $-log\delta$ only reflects this lack of observation accuracy (which is instrument-dependent), we may define an uncertainty measure intrinsic to the variable, by leaving this term out:

$$h(X) = -\int_{-\infty}^{\infty} f_x(X) log f_x(X). \tag{13}$$

However, contrary to the entropy of a partition, the latter measure can take negative values, and thus it does only have sense when used to measure changes in uncertainty. This is why it is often referred to as *differential entropy*. In the same way, we may define joint and conditional differential entropies for any two continuous *r.v.*'s, X, Y:

$$\begin{aligned} h(X,Y) &= -E(log f_{x,y}(X,Y)), & (14)\\ h(X|Y) &= -E(log f_{x|y}(X)), & (15) \end{aligned}$$

where $f_{x,y}(.)$ and $f_{x|y}()$ denote the joint and conditional *pdf*'s of the variables (respect.), and $E(.)$ is the expectation operator. Clearly, when X is independent of Y we have $h(X,Y) = h(X) + h(Y)$, and $h(X|Y) = 0$. The previous expressions generalize straightforwardly to more than two variables. In general, the different information-theoretic concepts discussed for partitions also apply to continuous-valued *r.v.*'s as long as they only refer to differences of entropies. Thus the mutual information for continuous *r.v.*'s, defined as

$$\begin{aligned} I(X,Y) &= h(X) + h(Y) - h(X,Y), & (16)\\ &= E\left[log \frac{f_{x,y}(X,Y)}{f_x(X) f_y(Y)}\right], & (17) \end{aligned}$$

conveys the same idea of dependence among the variables, as for partitions.

For the purpose of illustration, we give the values of these information-theoretic quantities for

Gaussian *r.v.*'s.

Let X, Y be two jointly Gaussian *r.v.*'s, such that $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, and suppose that their joint *pdf* is given by

$$f_{x,y}(X,Y) = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_x\sigma_y} e^{-((X-\mu_x)^2/\sigma_x^2 + (Y-\mu_y)^2/\sigma_y^2 - 2\rho(X-\mu_x)(Y-\mu_y)/(\sigma_x\sigma_y))} \tag{18}$$

where ρ is the correlation coefficient between the X and Y variables. Then it can be shown (i.e. Papoulis, 1991 [16]) the following:

$$\begin{aligned} h(X) &= log(\sigma_x\sqrt{2\pi e}), & (19)\\ h(Y) &= log(\sigma_y\sqrt{2\pi e}), & (20)\\ h(X,Y) &= log(2\pi e) + log(\sqrt{\Delta}), & (21)\\ h(X|Y) &= log(\sigma_x\sqrt{2\pi e}) + \frac{1}{2}log(1-\rho^2), & (22)\\ I(X,Y) &= -\frac{1}{2}log(1-\rho^2) & (23) \end{aligned}$$

where Δ is the determinant of the variance-covariance matrix of the variables, that is $\Delta = \sigma_x^2\sigma_y^2(1-\rho^2)$. In general, given n jointly Gaussian *r.v.*'s, $X_1, \cdots, X_n$, with variance-covariance matrix Σ, the joint differential entropy is given by

$$h(X_1, \cdots, X_n) = \frac{n}{2}log(2\pi e) + log(\sqrt{\Delta}) \tag{24}$$

where Δ is the determinant of Σ.

3.3 Information-theoretic measures for stochastic processes

Stochastic processes are defined in terms of the joint distributions for all subsets of their *r.v.*'s. In particular, the information gained when the m *r.v.*'s $X_{t_1}, \cdots, X_{t_m}$ of a continuous-valued stochastic process x_t are observed, is given by their *mth-order joint differential entropy*, defined as

$$h(X_{t_1}, \cdots, X_{t_m}) = -E\left(log f_{t_1,\cdots,t_m}(X_{t_1}, \cdots, X_{t_m})\right) \tag{25}$$

Obviously, the uncertainty about the values of x_t on any finite interval of t, is infinite. However, if x_t can be expressed in terms of its samples on a countable set of sampling instants $\{t_i\}_i$ (i.e. to the extent that x_t can be approximated by a narrowband process) it may be possible to define entropy measures. Henceforth we will assume that this is the case. Now, if there exists a conditional stationary *pdf*'s for x_t, we can define a measure of the uncertainty about any variable of the process, when its most recent values are known. For example, the *mth-order (differential) conditional entropy* of x_t, $h(X_n|X_{n-1},\cdots,X_{n-m})$ captures the remaining uncertainty about any *r.v.* from x_t, when information about its m-th history has been collected. This functional is, obviously, decreasing in m, and its rate of decay contains important information about the type of serial dependence in the process. For $m \rightarrow \infty$ we obtain a measure of the unknown information about any variable X_n once we know its entire past. Clearly, for a deterministic process, this measure, call it $h_r(x) = \lim_{m\rightarrow\infty} h(X_n|X_{n-1},\cdots,X_{n-m})$, equals zero. It is customary to call $h_r(x)$ the *entropy rate* of the process x_t. This name acknowledges the fact that when x_t is stationary we can write

$$h_r(x) = \lim_{m\rightarrow\infty} \frac{1}{m} h(X_1,\cdots,X_m). \tag{26}$$

Clearly, the limit on the right of the previous equality measures the speed at which the uncertainty grows as we try to guess at the values of an ever-increasing number of *r.v.*'s from the process. As a way of illustration, for a wide-sense stationary Gaussian process, x_t, we have

$$h_r(x) = log(\sqrt{2\pi e}) + \frac{1}{2} \lim_{m\rightarrow\infty} log\left(\frac{\Delta_{m+1}}{\Delta_m}\right) \tag{27}$$

where Δ_m is the determinant of the m-th order variance-covariance matrix of the process.

4 An information-theoretic characterization of memory

In the previous section, we saw that the mutual information in a pair of *r.v.*'s could be interpreted as a measure of general dependence between them, in contrast with their correlation, which only measures the adequacy of any variable for *linearly* predicting the other. Similarly, we can establish the serial dependence and cross-dependence properties of wide-sense stationary stochastic processes, in terms of a *mutual information function* (MIF), with generalizes the standard autocorrelation function (ACF). However, in order to extent the new characterization to processes having stochastic trends, we must again allow some scope for heterogeneity, and thus our measures will in general depend on time. Let the MIF of x_t as $i_x(\tau,t) = I(X_t, X_{t-\tau})$. Our information-theoretic characterization of *mean reversion*, *short* and *long memory* follows from the definitions in section 2.1. We will then say that a series is either **mean-reverting, short-memory, long-memory** or **integrated in information.**

Remarks:

1. In the Gaussian case, $i_x(\tau, t)$ is related to the ACF, and thus for a Gaussian short-memory process $i_x(\tau, t)$ must converge exponentially fast to zero, while for a Gaussian long-memory process this convergence must be slower (typically, only hyperbolically fast).

2. The information quantities can be re-written as (differential) entropy changes. That is,

$$i_x(\tau, t) = h(X_t) - h(X_t|X_{t-\tau}). \tag{28}$$

   This supports our intuition that entropy differences are most useful at characterizing the dependence properties of a process.

3. There are some connections between Granger's most general definition of mean-reversion, introduced in a previous paragraph, and the MIF. This can be seen by re-interpreting the latter as some sort of *mixing coefficients*. Given a stochastic process x_t, the standard α-mixing coefficients are given by (Rosenblatt, 1974 [18])

$$\alpha(\tau, t) = \sup_t \sup_{X \in \mathcal{F}_x^{-\infty,t}; X^* \in \mathcal{F}_x^{t+\tau,\infty}} |P(X^*, X) - P(X^*)P(X)| \tag{29}$$

   where $P(.)$ is a probability measure defined on the Borel σ-field of x_t. In contrast, the "information-mixing coefficients" $i_x(\tau. t)$ can be expressed as

$$i_x(\tau. t) = E\left(log f_{x.x}(X_t. X_{t-\tau}) - log f_x(X_t) f_x(X_{t-\tau})\right), \tag{30}$$

   where $f_{x,x}(,)$ and $f_x(.)$ denote the bivariate and univariate *pdf* for x_t. We remark that both types of mixing coefficients allow for heterogeneity in the process. However. in contrast to the α-mixing coefficients $\alpha(\tau, t)$, the quantities $i_x(\tau, t)$ can be easily estimated in many cases as statistical averages.

4. An alternative characterization could be made in terms of the conditional densities. Let $\mathcal{F}_{-\infty,t-\tau+1}^{t-\tau-1,t-1}$ denote the σ-field generated by the *r.v.*'s $X_{t-1}, \cdots, X_{t-\tau+1}; X_{t-\tau-1}, \cdots$. A generally nonstationary time series of x_t could be said to be **conditionally short-memory in information**. if the sequence of partial sums $R_n^{(x)} = \sum_{\tau>0}^n I(X_t, X_{t-\tau}|\mathcal{F}_{-\infty,t-\tau+1}^{t-\tau-1,t-1})$ converges as n grows to infinity. If, on the contrary, $R_n^{(x)}$ diverges, then x_t could be said to

be **conditionally long-memory in information.** These alternative definitions rephrase the former ones in terms of a *partial serial dependence measure*, which could be regarded as a generalization of the concept of *partial autocorrelation function* (PACF) in the linear context. However, when working with conditional densities may encounter severe difficulties in practice (i.e. need for very large data sets, curse of dimensionality, etc.), which make us prefer the former approach.

A few examples may help to illustrate the behaviour of the new unconditional dependence measures. Consider the following cases:

- Let $x_t = ax_{t-1} + \epsilon_t$ where ϵ_t is an *i.i.d.* sequence of Gaussian *r.v.*'s with zero mean and variance σ^2, in short $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, and $|a| < 1$. This model generates a stationary Gaussian Markov process, for which $cov(x_t, x_{t-\tau}) = \sigma^2 a^\tau$, which converges to zero exponentially fast as $\tau - \infty$. The information mixing coefficients, defined for $\tau > 0$, are given in this case by

$$i_x(\tau, t) = i_x(\tau) = -\frac{1}{2} log(1 - a^{2\tau}), \tag{31}$$

which clearly converges exponentially fast to zero as τ grows to ∞, thus implying that $\sum_{\tau>0} i_x(\tau, t) < \infty$. We may therefore conclude that x_t is both $I(0)$ and $II(0)$. On the contrary, if $a = 1$ we have a non-mixing process with an unit root, for which $corr(x_t, x_{t-\tau}) = 1$ and $i_x(\tau, t) = \infty$ for any τ and any t. Therefore, we may classify this $I(1)$ process as $II(1)$.

- Let x_t be a Gaussian stationary long-memory process with long-memory parameter d ($0 < d < 0.5$), that is $(1 - B)^d x_t = \epsilon_t$ with ϵ_t representing a stationary zero-mean short-memory Gaussian process. This mean-reverting process is characterized by an ACF which decays hyperbolically fast, that is, $cov(x_t, x_{t-\tau}) \sim \tau^{2d-1}$ for large τ (e.g. Hosking, 1981 [14]) and thus we write $x_t \sim I(d)$. On the other hand, we obtain the following approximation for large τ,

$$i_x(\tau, t) = i_x(\tau) \sim -\frac{1}{2} log(1 - c_d \tau^{4d-2}), \tag{32}$$

where c_d is a constant depending only on d. Clearly, $i_x(\tau)$ also converges to zero, but this time the convergence is only hyperbolically fast. Noting that $log(1 - c_d \tau^{4d-2}) \approx c_d \tau^{4d-2}$ for sufficiently large τ, the divergence of $\sum_{\tau>0} i_x(\tau, t)$ follows inmediately. Therefore, x_t is long-memory in information.

Now let us have a look at these measures from the viewpoint of the conditional (differential) entropies. Let $h_{c,\tau}(X_t) = h(X_t|X_{t-1}, \cdots, X_{t-\tau})$, or equivalently, $h^*_{c,\tau}(X_t) = h(X_t|X_{t-\tau}, \cdots, X_{t-\infty})$.

Proposition 2 If $h^*_{c,\tau}(X_t) < h(X_t)$ $\forall\tau$ and $\forall t$, then the process is neither mean-reverting nor short-memory in information.

PROOF:

Let $I(X_t; X_{t-\tau}, X_{t-\tau-1}, \cdots, X_{t-\infty})$ denote the information on X_t conveyed by the variables $X_{t-\tau}, X_{t-\tau-1}, \cdots$. We can write:

$$\begin{aligned} I(X_t; X_{t-\tau}, X_{t-\tau-1}, \cdots, X_{t-\infty}) &= h(X_t) - h(X_t|X_{t-\tau}, X_{t-\tau-1}, \cdots, X_{t-\infty}) && (33) \\ &> 0. && (34) \end{aligned}$$

Thus, we must have $\lim_{\tau \to \infty} I(X_t, X_{t-\tau}) > 0$, implying that x_t is neither mean-reverting nor short-memory in information. □

Remark:

The condition in the proposition clearly expresses when the remote past of a process does still contribute in information about its present state.

We shall assume in the following examples that our processes are Gaussian. Therefore, recalling equation (27), the *τth order conditional (differential) entropy* for a Gaussian process x_t is

$$h_{c,\tau}(X_t) = log(\sqrt{2\pi e}) + \frac{1}{2} log\left(\frac{\Delta_{\tau+1,t}}{\Delta_{\tau,t}}\right). \tag{35}$$

where $\Delta_{\tau,t}$ is the determinant of the τth-order variance-covariance matrix of x_t.

In the following, we will determine the conditional entropies and some implications for the classes of processes previously characterized in terms of the MIF.

- Let $x_t = ax_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2)$. If $|a| < 1$ then we can write $h_{c,\tau}(X_t) = h(X_t|X_{t-1}) = log\left(\sigma\sqrt{2\pi e}\right)$ for any $\tau > 0$. It follows that $I(X_t; X_{t-1}, \cdots, X_{t-\tau}) = I(X_t, X_{t-1}) = h(X_t) - h_{c,\tau}(X_t) = -\frac{1}{2}log(1-|a|^2) < \infty$, for any $\tau > 0$. On the contrary, if $a = 1$ then $I(X_t; X_{t-1}, \cdots, X_{t-\tau})$ is infinity for any $\tau > 0$.

- Let x_t be a stationary autoregressive process of order p, in short $x_t \sim AR(p)$. If x_t is Gaussian then we have the following result from Kay (1988 [19], pp. 169-178):

$$\frac{\Delta_{\tau+1,t}}{\Delta_{\tau,t}} = \frac{\Delta_{\tau+1}}{\Delta_\tau} = \sigma^2 \prod_{k=1}^{\tau}(1 - |r_k|^2) \tag{36}$$

where r_k is the partial autocorrelation at lag k. Thus, at long lags,

$$\frac{\Delta_{\tau+1}}{\Delta_\tau} = \sigma^2 \prod_{k=1}^{p}(1 - |r_k|^2) \tag{37}$$

since $r_k = 0$ for $k > p$. Now, since $|r_k| < 1$, $\forall k$, it follows from equation (27) that $h_r(x)$ is bounded, and that $I(X_t; X_{t-1}, \cdots, X_{t-\infty}) < \infty$.

- Suppose x_t is a Gaussian stationary long-memory process with long-memory parameter d $(0 < d < 0.5)$. Then since the partial autocorrelations of this process r_k satisfy $0 < r_k < 1$ for any finite k (see Hosking,1981 [14]) then

$$\begin{aligned} \lim_{\tau \to \infty} \frac{\Delta_{\tau+1}}{\Delta_\tau} &= \sigma^2 \lim_{\tau \to \infty} \prod_{k=1}^{\tau}(1 - |r_k|^2) \\ &= 0 \end{aligned} \tag{38}$$

The latter implies that $h_r(x) = -\infty$, which in turns leads to an infinite value for the mutual information between X_t and its infinite history, that is $I(X_t; X_{t-1}, \cdots, X_{t-\infty}) = \infty$.

These examples seem to support our intuition that the persistence of the shocks in a process results in that its entire past contains an infinite amount of information about its present. On the contrary, this amount of information is bounded for mixing processes.

The connection of the latter discussion with our characterization of dependence in terms of the information mixing numbers $i_x(\tau)$ comes by realizing that each variable from the past contributes a small portion of information about the present variable. X_t. In other words, we must have

$$I(X_t; X_{t-1}, \cdots, X_{t-\infty}) \leq \sum_{\tau=1}^{\infty} i_x(\tau, t). \tag{39}$$

Now, the fact that $I(X_t; X_{t-1}, \cdots, X_{t-\infty}) = \infty$ for persistent Gaussian processes implies that $i_x(\tau, t)$ cannot decrease with τ faster than $O(\tau^{-2+\delta})$ for some $\delta > 0$. Alternatively, for stationary Gaussian processes we obtained $I(X_t; X_{t-1}, \cdots, X_{t-\infty}) < \infty$, which is consistent with an exponentially fast decay of $i_x(\tau, t)$ for growing τ.

4.1 Some implementation issues

We briefly explain how the mutual information quantities were estimated in the experiments that follow. The MIF, $i_x(\tau)$, was evaluated using the following estimator, where N is the sample size,

$$\begin{aligned} \hat{i}_x^{(N)}(\tau) &= N^{-1}\sum_{t=1}^{N}\hat{i}_x(\tau,t) \\ &\approx N_\gamma^{-1}\sum_{t\in\mathcal{S}} c_t(\gamma) log\left(\frac{\hat{f}_{x,x}(X_t,X_{t-\tau})}{\hat{f}_x(X_{t-\tau})^2}\right), \end{aligned} \tag{40}$$

with

$$c_t(\gamma) = \begin{cases} 1+\gamma, & \text{for } t \text{ odd} \\ 1-\gamma, & \text{for } t \text{ even} \end{cases}$$

where $\gamma \geq 0$, $N_\gamma = N$ for N even, and $N_\gamma = N + \gamma$, for N odd. Here X_t represents a generic vector variable, $\hat{f}_{x,x}(.,.)$ and $\hat{f}_x(.)$ are estimators of the bivariate and univariate *pdf*'s (which may be time-varying), and the set $\mathcal{S}$ is introduced to make explicit the exclusion of certain inocuous summands, which can occur, for example, when $\hat{f}_{x,x}(.,.) \leq 0$ or $\hat{f}_x(.) \leq 0$, or when logarithms cannot be taken. The densities can be estimated using *kernel smoothers* (Breiman et al., 1977 [3]). In general, given a set of $N - n$ n-dimensional vectors X_t, $t = 1, N - n$, a kernel density estimator with kernel K and bandwidth α, of their unconditional *pdf*, say $f(.)$, has the form

$$\hat{f}(X) = (N-n)^{-1}\alpha^{-1}\sum_{t=1}^{N-n} K[\alpha^{-1}(X - X_t)] \tag{41}$$

where the kernel K is a function verifying $\int_{\Re^n} K(Y)dY = 1$. Robinson (1991) [17] proved the consistency of a similar estimator under the assumption of stationarity in the series and for $n = 1$. For the experiments, we choose Gaussian kernels:

$$K(X) = (2\pi)^{-n/2}\, exp(-X'X/2) \tag{42}$$

Even though the form of the kernel is not critical to the results, the bandwidth is. We can deal with this problem by means of adaptive bandwidths. This technique consists in allowing the kernels to shrink in rather densily populated regions of the n-dimensional embedding space, and to widen in regions with few data points. The likelihood of introducing important biases is greatly reduced in this way, since the smoothing becomes only important at those regions of the embedding space containing a large number of points. Initially, we took a fixed bandwidth for the kernels, α, and the initial density estimates were subsequently used to obtain locally adapted bandwidths, say $\beta(X)$, according to

$$\beta(X) \propto 1/\hat{f}_\alpha(X) \tag{43}$$

where $\hat{f}_\alpha(X)$ denotes a rough estimate of the *pdf* at X using a kernel estimator with the fixed bandwidth, α.

4.2 Information-theoretic characterization of cointegration

Let x_t, y_t be long-memory in information. The concept of **cointegration in information** arises when letting $i_{x,y}(\tau, t) = I(X_t, Y_{t-\tau})$ in the characterization of cointegration proposed in section 2.2 (see definition 9).

Remarks :

1. The information-cointegrateness concept states that for any long-run predictor of X_t based on $X_{t-\tau}$, we can find a predictor based on $Y_{t-\tau}$ which conveys exactly the same information about X_t.
2. Our characterization applies to both integer and fractionally integrated processes. Besides, the processes involved are not required to have the same integration order. For instance, consider the case in which $x_t \sim II(d_x)$, $y_t \sim II(d_y)$, with $d_x \neq d_y$, and $\phi(.)$ is a nonlinear one-to-one transformation such that $z_t = \phi(y_t) \sim II(d_z)$ with $d_z = d_x$. This situation can be understood noting that both the entropy and the mutual information of the variables in a process are invariant to one-to-one transformations of the latter (see, for instance, Papoulis, 1991 [16], p. 565).
3. The information-cointegration definition can equally handle multivariate processes, which enter naturally as arguments of the information measures.

In figure 4 we compare the behaviour of a normalized version of the *generalized sample correlations*, $\hat{i}^{(N)}_{x,y}(\tau)/\hat{i}^{(N)}_{x}(1)$ and $\hat{i}^{(N)}_{x}(\tau)/\hat{i}^{(N)}_{x}(1)$ as functions of τ, by means of Montecarlo simulations. Here $\hat{i}^{(N)}_{x,y}(\tau)$ is given by:

$$\begin{aligned} \hat{i}^{(N)}_{x,y}(\tau) &= N^{-1}\sum_{t=1}^{N}\hat{i}_{x,y}(\tau,t) \\ &\approx N_\gamma^{-1}\sum_{t\in S} c_t(\gamma) log\left(\frac{\hat{f}_{x,y}(X_t, Y_{t-\tau})}{\hat{f}_x(Y_{t-\tau})^2}\right), \end{aligned} \qquad (44)$$

where the coefficients $c_t(\gamma)$ and γ are as in the previous section. The curves shown in the figure represent statistical averages computed from 20 simulated pairs of series. Plots (a), (b) and (c) correspond to linearly cointegrated, nonlinearly cointegrated, and non-cointegrated series, respectively. The horizontal scale shows $\tau + 1$. The linear cointegrated series were generated as those in figures 1, while the nonlinearly cointegrated ones were obtained by applying third-order polynomial transformations to a common random-walk component.

Example 2:

Consider the following linear *common factor model*:

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} a \\ 1 \end{pmatrix} w_t + \begin{pmatrix} v_t \\ \xi_t \end{pmatrix} \tag{45}$$

where $a \neq 0$, $w_t = w_{t-1} + \epsilon_t$ with $w_0 = 0$, and (v_t, ξ_t, ϵ_t) are independent sequences of independent and identically Normally distributed *r.v.*'s with zero mean and joint covariance matrix equal to the identity matrix. If we now define $z_t = y_t - ax_t$, and

$$\rho_x(\tau, t) = \frac{cov(x_t x_{t-\tau})}{\sigma_{x_t}\sigma_{x_{t-\tau}}}, \tag{46}$$

$$\rho_{x,y}(\tau, t) = \frac{cov(y_t x_{t-\tau})}{\sigma_{x_t}\sigma_{x_{t-\tau}}} \tag{47}$$

we obtain after some algebra

$$\rho_x(\tau, t) = \frac{(t-\tau)\sigma_\epsilon^2}{\sqrt{(}t\sigma_\epsilon^2 + \sigma_\xi^2)\sqrt{(}(t-\tau)\sigma_\epsilon^2 + \sigma_\xi^2)}, \tag{48}$$

$$\rho_{x,y}(\tau, t) = \frac{a(t-\tau)\sigma_\epsilon^2}{\sqrt{(}a^2(t\sigma_\epsilon^2 + \sigma_\xi^2) + \sigma_z^2)\sqrt{(}(t-\tau)\sigma_\epsilon^2 + \sigma_\xi^2)}. \tag{49}$$

It follows that for sufficiently large t, $\rho_{x,y}(\tau, t) \approx \rho_{x,y}(\tau, t)$.

Now since $i_{x,y}(\tau, t) = -\frac{1}{2}log(1 - \rho_{x,y}^2(\tau, t))$, and $i_x(\tau, t) = -\frac{1}{2}log(1 - \rho_x^2(\tau, t))$, it follows that $i_{x,y}(\tau, t)/i_x(\tau, t) \approx 1$ for any τ.

An alternative condition for the information-cointegrateness of (x_t, y_t) can be given using conditional entropies:

$$\lim_{\tau\to\infty} \frac{h(Y_t|\mathcal{F}_x^{-\infty, t-\tau})}{h(Y_t|\mathcal{F}_y^{-\infty, t-\tau})} \neq 0, \ \forall t \tag{50}$$

$$\lim_{\tau\to\infty} \frac{h(X_t|\mathcal{F}_y^{-\infty, t-\tau})}{h(X_t|\mathcal{F}_x^{-\infty, t-\tau})} \neq 0, \ \forall t. \tag{51}$$

At this point, it is also interesting to analyze the links between the concepts of cointegration in information and causality. To do so, we first propose a definition of *non-causality in information*, which merely express the non-causality idea of Granger (1969) [9] in terms of information statistics.

Definition 10 *A series x_t* **non-causes in information** *a series y_t if $h(Y_t|\mathcal{F}_x^{-\infty,t-1};\mathcal{F}_y^{-\infty,t-1}) = h(Y_t|\mathcal{F}_y^{-\infty,t-1})$.*

Accordingly, there is no causality among the variables if the remaining uncertainty in either variable after conditioning on its own past is not reduced by knowledge of the other's past.

5 Testing for long-memory and cointegration in information

Testing a cointegrating relationship involves two major steps: (1) a test for long-memory in the series; and (2) a test of cointegration.

5.1 Long-memory testing

It may be possible to test for long-memory in information for any of the variables, say x_t, by working out the consequences of our characterization of short and long memory in information. Recall that for x_t to be short-memory in information we must have $\sum_{\tau>0} i_x(\tau,t) < \infty$, which implies that for any $\delta > 0$ and any t, $i_x(\tau,t) = o(\tau^{-2+\delta})$. That is, there exists positive real numbers τ_0, b such that $i_x(\tau,t) < b\tau^{-2}$ $\forall\tau > \tau_0$ and $\forall t$. On the contrary, if x_t is long-memory in information then there exists positive real numbers τ_1, c_t and $2 > r > 0$ such that $i_x(\tau,t) \approx c_t\tau^{-r}$ $\forall\tau > \tau_1$. Or taking logs,

$$logi_x(\tau,t) \approx logc_t - rlog\tau + \xi_{\tau,t}, \quad \forall\tau \gg \tau_1, \tag{52}$$

where $\xi_{\tau,t}$ is an error sequence. Therefore we could check the property of short memory in information by testing the null hypothesis H_0 : $r \geq 2$.

For most empirical series, a finite sample size prevents the possibility of adjusting the previous regression line at large lags. However, a frequency-domain version of this testing device allows us to do the analysis at low frequencies ($\lambda - 0$) instead of at very long lags ($\tau - \infty$). In this way, we can take advantage of the full information contained in the sample. For this, let us first define a *generalized periodogram* as

$$G_x^{(N)}(\lambda,t) = \sum_{\tau=1}^{N} w_\tau i_x(\tau,t)exp(-j2\pi\lambda\tau), \tag{53}$$

where $j^2 = -1$, w_τ is a spectral window, and N is the sample size. Now, if x_t is long-memory in information we should have

$$G_x^{(N)}(\lambda,t) \approx u_x(\lambda,t)\lambda^{-2d}, \tag{54}$$

for small λ's. Here $d > 0$, and $u_x(\tau, t)$ is a slowly-varying function of τ, that is $\lim_{\lambda \to a} u_x(c\lambda, t)/u_x(\lambda, t) = 1$ $\forall c$ and for $a = 0$ and $a = \infty$.

Again, taking logs we obtain

$$logG_x^{(N)}(\lambda, t) = logu_x(\lambda, t) - 2dlog\lambda + v_{\lambda,t}, \tag{55}$$

for small λ's, and with $v_{\lambda,t}$ representing an error sequence. Now we can test the null-hypothesis of short-memory in information $H_0 : d = 0$ once we have an estimate of the slope of the previous regression line.

Remark:

Notice that when $i_x(\tau, t) = \rho_x(\tau, t)$ and x_t is supposed to be stationary we obtain the device proposed by Geweke and Porter-Hudak (1983) [6].

5.2 Cointegration testing

Based on definition 9, a candidate test statistic that provides a measure of cointegration in a pair of series x_t, y_t, could be

$$\tau_{m,q}(x, y) = N^{-1} \sum_{t=1}^{N} \sum_{\tau=m}^{m+q} \left(1 - \hat{i}_{x,y}(\tau, t)/\hat{i}_x(\tau, t)\right) \tag{56}$$

where m must be sufficiently large (i.e. larger than the short-memory span of the series) in order to capture only the long-wave discrepancies, q should be such that $m + q < N$, where N is the sample size.

As we said in the preceeding section, under cointegration $i_{x,y}(\tau, t)$ will be of the same order of magnitude as $i_x(\tau, t)$ for sufficiently large τ and $\forall t$. On the contrary, under non-cointegration, $i_{x,y}(\tau, t) \ll i_x(\tau, t) > 0$ for sufficiently large τ and $\forall t$. This implies a tendency for the values of $\tau_{m,q}(x, y)$ to cluster around 1 under non-cointegration.

The limiting distribution of our statistic may be difficult to find using standard asymptotic theory, since we are dealing with non-mixing processes. Yet we can test the null hypothesis of cointegration by constructing an empirical confidence interval for the test statistic. That is, for fixed values of m, q, we estimate the empirical critical value b_α such that $P(\tau_{m,q}(x, y) > b_\alpha) = \alpha$ under the assumption of information-cointegrateness, for the given significance level, α. Therefore this hypothesis will be rejected at this level when $\tau_{m,q}(x, y) > b_\alpha$.

6 Experiments on simulated series

To assess the potentialities of a cointegration test based on the statistic $\tau_{m,q}(x,y)$ in equation (56), we generated 100 pairs of linearly, nonlinearly and non-cointegrated series. The linearly cointegrated series were obtained as in figures 1. The nonlinearly cointegrated ones were computed applying third-order polynomial transformations to a common random walk component. The coefficients of these polynomials were chosen at random. Finally, the non-cointegrated series were either pairs of independent random walks ($H_{2,1}$) or mutually dependent short-memory series ($H_{2,2}$). In the latter case, the series were generated according to the model $y_t = x_t + \epsilon_t$, $z_t = a_0 + a_1 x_t + a_2 x_t^2 + a_3 x_t^3 + \epsilon_t'$, where $x_t = a_4 e_{t-2} e_{t-1} + e_t$, $\epsilon_t, \epsilon_t', e_t$ are mutually independent *i.i.d.* sequences, and the a_i were chosen at random. For the experiment, we selected $q = 0, m = 10$, and a sample size of $N = 1000$. In all the replications the value of $\tau_{10,0}(x,y)$ was comparatively large and positive under non-cointegration, but small and with varying sign under cointegration, both in the linear and the nonlinear cases. Table 1 shows the mean, standard deviation and mean absolute value of $\tau_{10,0}(x,y)$ obtained in the experiment.

The histogram plots of $\tau_{10,0}(x,y)$ for the different cases are given in figure 5. Using the 5% empirical critical values of this statistic under $H_{2,1}$, estimated from 1000 Monte Carlo replicas, the percentage rejection approached 85% of the simulated cointegrated pairs.

7 Experiment on financial data

The statistic $\tau_{m,q}(x,y)$ proposed in the previous section for testing cointegration and linearity in cointegration, respectively, is here evaluated on two pairs of exchange rate series (figure 7), and on a pair of stock return series (STR1,STR2) from a Japanese food company (figure 6). The former group of series were the rates of exchange of the US Dollar (EXRPD), the Deutsch Mark (EXRPM) and the Japanese Yen (EXRPY) against the Spanish Peseta. We took the first $N = 1000$ daily observations from series starting at January the first 1987. For the exchange-rate data, EXRPD was taken as reference.

The hypothesis of a unit root could not be rejected by a standard Dickey-Fuller test for any of the series, for the given sample size. To test for cointegration, we first run an *Augmented Dickey-Fuller* (ADF) test (the conventional DF test was augmented with one lag in the first differences of the series) on the regression residuals of the three pairs considered above. The values taken by the test statistic τ_{df} are reported in table 2.

Using the critical values computed by Mackinnon (1990) [15] (−2.57, −1.94 and −1.62 at the 1%, 5% and 10% levels, respectively), the hypothesis of cointegration (i.e. that τ takes values smaller than the tabulated critical values) was only accepted for the pair of stock return series (STR1,STR2). In contrast, the values of the test statistic $\tau_{10,0}(x, y)$ of equation (56), shown in table 3. suggests evidence of cointegration in both (EXRPD,EXRPY) and (STR1,STR2), when using a one-standard-deviation empirical confidence interval.

8 Conclusion

Long-memory and cointegration are two important features of many economic time series. Standard methods to characterize these features do not take into account possible nonlinearities in the data generating processes or in their relationship. This calls for a more general characterization of memory in time series, and of cointegration between pairs of time series, where nonlinearity is allowed in the long-run relationship between the variables. In this paper, we proposed one such alternative characterization based on the mutual information in pairs of variables, but which could be used in connection with any measure of serial dependence. Our methodology does not contraint the integration orders of the individual series to be equal. and could be generalized to the analysis of vector cointegrating relationships. Finally, we suggest new devices for exploratory data analysis and for testing the hypotheses of short-memory and cointegration. The performances of our cointegration testing device was shown on both simulated and some real-world financial series. Our results point to a gain in robustness of the proposed schemes over standard ones when the integrated variables are nonlinearly related.

References

[1] F.M. Aparicio Acosta and C.W.J. Granger. Information-theoretic schemes for linearity testing under long-range dependence and under cointegration. Working Paper (Depart. of Economics of the University of California at San Diego), March 1995.

[2] L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80(391):580–619, 1985.

[3] L. Breiman, W.S. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities and their calibration. *Technometrics*, 19:135–144, 1977.

[4] R.F. Engle and C.W.J. Granger. Co-integration and error correction: representation, estimation and testing. *Econometrica*, 55:251–276, 1987.

[5] A. Escribano and S. Mira. Nonlinear cointegration and nonlinear error correction. Working paper 96-54 of the Depart. of Statistics of Universidad Carlos III de Madrid, 1996.

[6] J. Geweke and S. Porter-Hudak. The estimation and application of long-memory time series models. *Journal of time series analysis*, 4(4):221–238, 1983.

[7] C.W.J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, 16:121–130, 1981.

[8] C.W.J. Granger. Co-integrated variables and error-correcting models. Discussion paper 83-13 (Depart. of Economics of the University of California at San Diego), 1983.

[9] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1990.

[10] C.W.J. Granger. Modelling nonlinear relationships between extended-memory variables. *Econometrica*, 63(2):265–279, 1995.

[11] C.W.J. Granger and Hallman J.J. Long-memory series with attractors. *Oxford Bulletin of Economics and Statistics*, 53(1):11–26, 1991.

[12] C.W.J. Granger and T. Terasvirta. *Modeling nonlinear economic relationships*. Oxford University Press, Oxford, 1993.

[13] J.J. Hallman. *Nonlinear integrated series, cointegration, and an application*. PhD thesis, Department of Economics of the University of California at San Diego, La Jolla, USA, 1990.

[14] J.R.M. Hosking. Fractional differencing. *Biometrika*, 68:165–176, 1981.

[15] J. G. MacKinnon. Critical values for cointegration. Working Paper (Depart. of Economics of the University of California at San Diego), January 1990.

[16] A. Papoulis. *Probability, random variables and stochastic processes.* McGraw-Hill, N.Y., 1991.

[17] P.M. Robinson. Consistent nonparametric entropy-based testing. *Review of Economic Studies*, 58:437–453, 1991.

[18] M. Rosenblatt. *Random processes.* Springer-Verlag, N.Y., 1974.

[19] S.M.Kay. *Modern spectral estimation.* Prentice Hall, N.Y., 1988.

[20] J.M. Wooldridge. *Asymptotic properties of econometric estimators.* PhD thesis, Dpt. of Economics of the University of California at San Diego, La Jolla, USA, 1986.

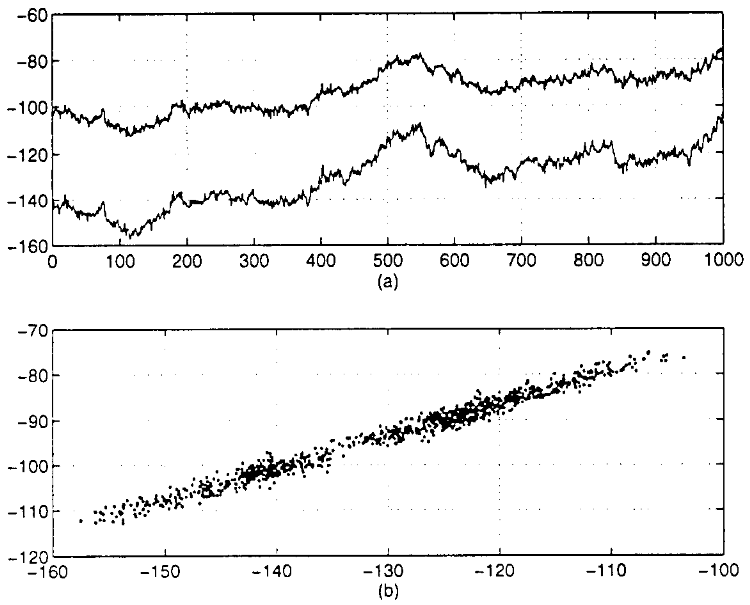

Figure 1: Two simulated linearly cointegrated random walks (a) and their scatter plot (b). The series, x_t, x'_t were generated with the model: $x_t = aw_t + \epsilon_t$, $x'_t = w_t + \epsilon'_t$, $w_t = w_{t-1} + \xi_t$, with $w_0 = 0$ and where $\epsilon_t, \epsilon'_t, \xi_t$ are independent sequences of *i.i.d.* Gaussian *r.v.*'s.

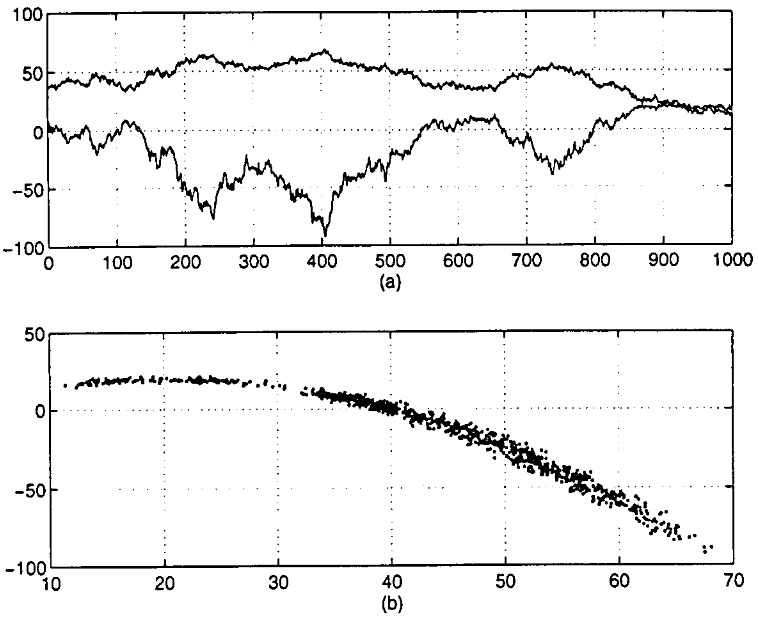

Figure 2: Two simulated nonlinearly cointegrated series (a) and their scatter plot (b). The upper series was obtained as $x_t = w_t+\xi_t$, where $w_t = w_{t-1}+\varepsilon_t$ with $w_0 = 0$, and the lower one corresponds to $y_t = 2w_t - 0.05w_t^2 + \epsilon_t$. The errors $\epsilon_t, \varepsilon_t, \xi_t$ are independent sequences of *i.i.d.* Gaussian *r.v.*'s.

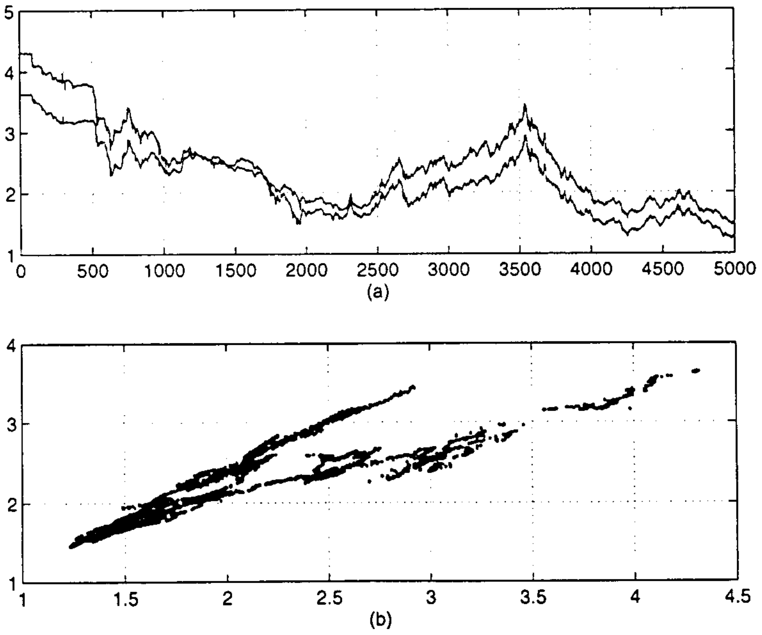

Figure 3: Two apparently nonlinearly cointegrated time series of stock returns from a Japanese food company *Ajinomoto* (a). Clearly, the strength of attraction varies accross time, as shown in the scatter plot (b).

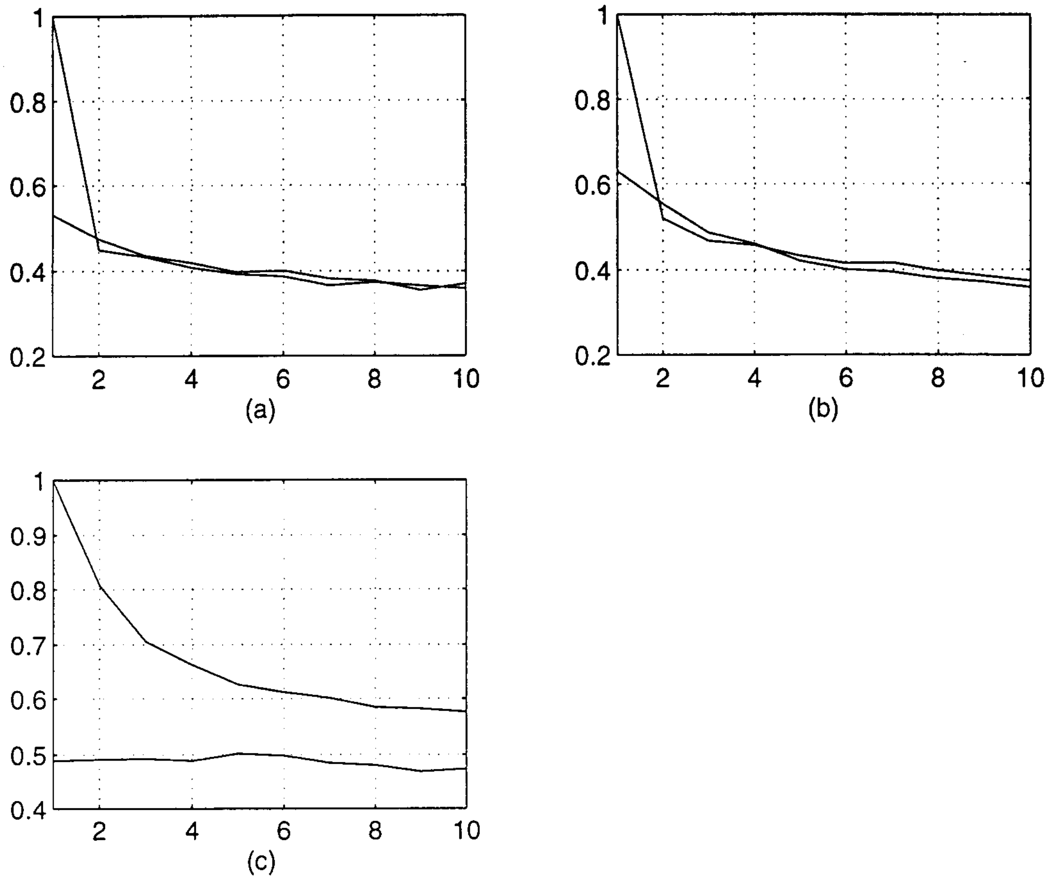

Figure 4: Plots of the generalized correlations $\hat{\imath}_{x,y}^{(N)}(\tau)/\hat{\imath}_{x}^{(N)}(1)$ and $\hat{\imath}_{x}^{(N)}(\tau)/\hat{\imath}_{x}^{(N)}(1)$ versus $\tau + 1$ for linearly (a). nonlinearly (b). and non-cointegrated (c) series. The plots show the average curves obtained from 20 Monte Carlo simulated pairs of series. The nonlinearly cointegrated series were generated by applying third-order polynomial transformations to a common random walk component. The non-cointegrated series were independent random walks.

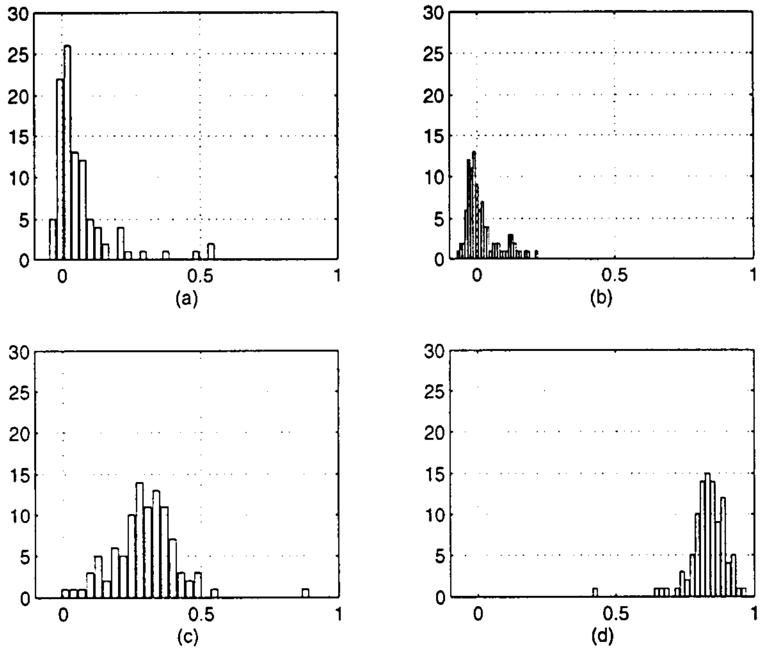

Figure 5: Histogram plots of $r_{10,0}(x, y)$, where (x, y) represents linearly (a), nonlinearly (b), and non-cointegrated (c)-(d) pairs of series. Plots (c)-(d) corresponds to non-cointegrated series from the alternative hypotheses $H_{2,1}$ and $H_{2,2}$, respectively. The nonlinearly cointegrated series were obtained as in figure 2.

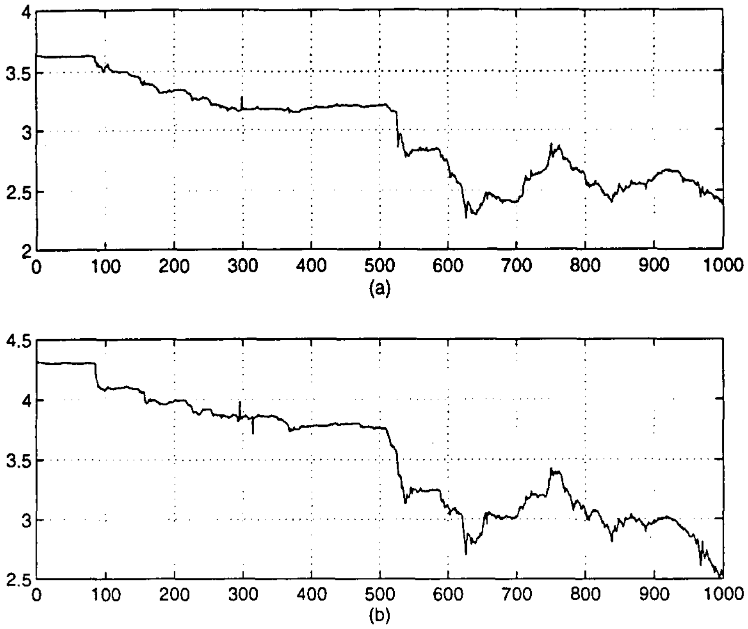

Figure 6: Two stock return series from the Japanese food company *Ajinomoto.*

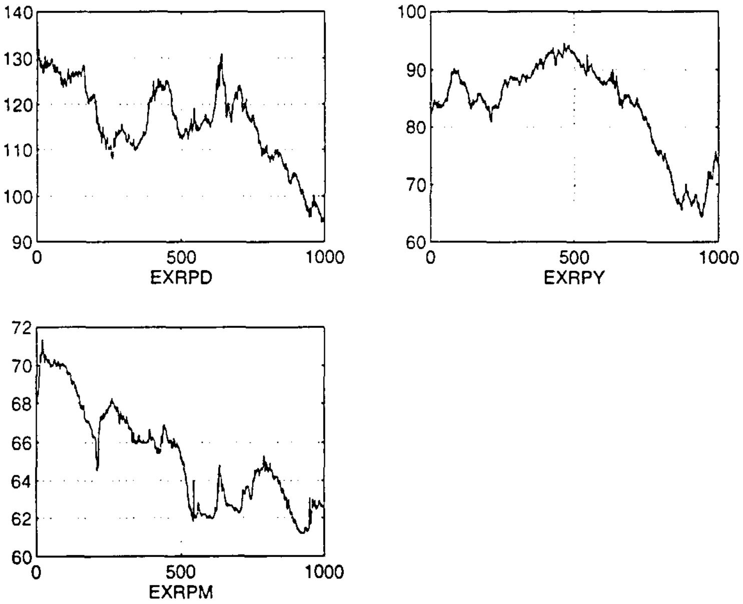

Figure 7: Daily foreign exchange-rate series from January 1987: EXRPD (Peseta/US Dollar), EXRPY (Peseta/100 Yens), EXRPM (Peseta/Deutsch Mark).

Test statistic	*linear cointeg.*	*nonlin. cointeg.*	*non-cointeg.* $(H_{2,1})$	*non-cointeg.* $(H_{2,2})$
$E(\tau_{10,0}(x,y))$	0.0619	0.0189	0.2953	0.8307
$std(\tau_{10,0}(x,y))$	0.117	0.061	0.12	0.07
$E(\|\tau_{10,0}(x,y)\|)$	0.0718	0.0434	0.2953	0.8307

Table 1: Mean, standard deviation and absolute mean values of $\tau_{10,0}(x,y)$ for linearly, nonlinearly and non-cointegrated series.

Series	*EXRPY/EXRPD*	*EXRPM/EXRPD*	*STR1/STR2*
$\tau(x,y)$	-0.328	-0.686	-21.28

Table 2: Values taken by the Dickey-Fuller test statistic $\tau_{df}(x,y) = N(\hat{a}-1)$ on the two pairs of foreign exchange rate series and the pair of stock return series. Here $\hat{a}$ is the OLS estimator of the parameter in the regression of y_t on x_t.

Series	*EXRPY/EXRPD*	*EXRPM/EXRPD*	*STR1/STR2*
$\tau_{10,0}(x,y)$	-0.0113	0.257	0.1169

Table 3: Values taken by the cointegration test statistic $\tau_{10,0}(x,y)$ on two pairs of foreign exchange rate and a pair of stock return series.