Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Macroeconomics 28 (2006) 195–224
Journal of MACROECONOMICS
www.elsevier.com/locate/jmacro

Non-parametric determination of real-time lag structure between two time series: The "optimal thermal causal path" method with applications to economic data

Wei-Xing Zhou [a,b], Didier Sornette [b,c,*]

[a] School of Business and State Key Laboratory of Chemical Reaction Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] Laboratoire de Physique de la Matiére Condensée, CNRS UMR 6622 and Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France
[c] Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, 1693 Geology Building, Los Angeles, CA 90095-1567, USA

Received 15 October 2005; accepted 30 October 2005
Available online 5 January 2006

Abstract

We introduce a novel non-parametric methodology to test for the dynamical time evolution of the lag–lead structure between two arbitrary time series. The method consists in constructing a distance matrix based on the matching of all sample data pairs between the two time series. Then, the lag–lead structure is searched as the optimal path in the distance matrix landscape that minimizes the total mismatch between the two time series, and that obeys a one-to-one causal matching condition. We apply our method to the question of the causality between the US stock market and the treasury bond yields and confirm earlier results on a causal arrow of the stock markets preceding the Federal Reserve Funds adjustments as well as the yield rates at short maturities in the period 2000–2003. The application to inflation, inflation change, GDP growth rate and unemployment rate unearths non-trivial causal relationships: the GDP changes lead inflation especially since the 1980s, inflation changes lead GDP only in the 1980 decade, and inflation leads unemployment rates since the 1970s.

* Corresponding author. Address: Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, 1693 Geology Building, Los Angeles, CA 90095-1567, USA. Tel.: +1 310 825 2863; fax: +1 310 206 3051.
E-mail address: sornette@moho.ess.ucla.edu (D. Sornette).
URL: http://www.ess.ucla.edu/faculty/sornette/ (D. Sornette).

0164-0704/$ - see front matter
doi:10.1016/j.jmacro.2005.10.015

In addition, we detect multiple competing causality paths in which one can have inflation leading GDP with a certain lag time and GDP feeding back/leading inflation with another lag time.

JEL classification: C14; E31; E58; G10

Keywords: Causality; Time-dependent correlation; Distance matrix; Lead–lag; Time series; Stock markets; Bond yields; Inflation; GDP growth; Unemployment

1. Introduction

Determining the arrow of causality between two time series $X(t)$ and $Y(t)$ has a long history, especially in economics, econometrics and finance, as it is often asked which economic variable might influence other economic phenomena (see e.g. Chamberlain, 1982; Geweke, 1984). This question is raised in particular for the relationships between respectively inflation and GDP, inflation and growth rate, interest rate and stock market returns, exchange rate and stock prices, bond yields and stock prices, returns and volatility (Chan et al., 2001), advertising and consumption and so on. One simple naive measure is the lagged cross-correlation function $C_{X,Y}(\tau) = \langle X(t)Y(t+\tau)\rangle / \sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}$, where the brackets $\langle x\rangle$ denotes the statistical expectation of the random variable x. Then, a maximum of $C_{X,Y}(\tau)$ at some non-zero positive time lag τ implies that the knowledge of X at time t gives some information on the future realization of Y at the later time $t+\tau$. However, such correlations do not imply necessarily causality in a strict sense as a correlation may be mediated by a common source influencing the two time series at different times. The concept of Granger causality bypasses this problem by taking a pragmatic approach based on predictability: if the knowledge of $X(t)$ and of its past values improves the prediction of $Y(t+\tau)$ for some $\tau > 0$, then it is said that X Granger causes Y (Ashley et al., 1980; Geweke, 1984, see also Chen et al. (2004) for a recent extension to nonlinear time series). Such a definition does not address the fundamental philosophical and epistemological question of the real causality links between X and Y but has been found useful in practice. Our approach is similar in that it does not address the question of the existence of a genuine causality but attempts to detect a dependence structure between two time series at non-zero lags. We thus use the term "causality" in a loose sense embodying the notion of a dependence between two time series with a non-zero lag time.

However, most economic and financial time series are not strictly stationary and the lagged correlation and/or causality between two time series may be changing as a function time, for instance reflecting regime switches and/or changing agent expectations. It is thus important to define tests of causality or of lagged dependence which are sufficiently reactive to such regime switches, allowing to follow almost in real time the evolving structure of the causality. Cross-correlation methods and Granger causality tests require rather substantial amount of data in order to obtain reliable conclusions. In addition, cross-correlation techniques are fundamentally linear measures of dependence and may miss important nonlinear dependence properties. Granger causality tests are most often formulated using linear parametric auto-regressive models. The new technique introduced in this paper, called the "optimal thermal causal path," is both non-parametric and sufficiently general so as to detect a priori arbitrary nonlinear dependence structures. Moreover, it is specif-

ically conceived so as to adapt to the time evolution of the causality structure. The "optimal thermal causal path" can be viewed as an extension of the "time distance" measure which amounts to comparing trend lines in horizontal differences of two time series (Granger and Jeon, 1997).

The organization of the paper is as follows. Section 2 introduces the concept of a time-dependent time lag and sets the general framework that we propose to determine it. Section 3 presents a first implementation for the determination of optimal paths and local time lags, that we refer to as the multi-layer algorithms. Section 4 presents our second implementation, that we refer to as the "optimal thermal causal path" method. While the first implementation of Section 3 seemed a priori more general, it turns out to be too sensitive and leads to over-fitting. As a consequence, the second method of Section 4 shows much better performance as demonstrated in synthetic tests presented in Section 5 on auto-regressive models. Section 6 then presents an application of the optimal thermal causal path method on two important economic problems: the causal relationship between the US treasury bond yields and the stock market in the aftermath of the Internet bubble collapse and between inflation, inflation change, gross domestic product rate and unemployment rate in the United States. Section 7 concludes.

2. The concept of an optimal lag function

2.1. Distance matrix

We consider time series updated in discrete time, in units of some elementary discretization step, taken unity without loss of generality. Let us denote $\{X(t_1): t_1 = 1, \ldots, N_1\}$ and $\{Y(t_2): t_2 = 1, \ldots, N_2\}$ the two time series that we would like to test for causality. Note that the lengths N_1 and N_2 of the two series can in principle be different as our method generalizes straightforwardly to this case, but for the sake of pedagogy, we restrict here to the case $N_1 = N_2 = N$. These time series $\{X(t_1)\}$ and $\{Y(t_2)\}$ can be very different in nature with largely different units and meanings. To make them comparable, we normalize them by their respective standard deviations, so that both normalized time series have comparable typical values. From now on, the two time series $\{X(t_1)\}$ and $\{Y(t_2)\}$ denote these normalized time series.

We introduce a distance matrix $E_{X,Y}$ between X to Y with elements defined as

$$\epsilon(t_1, t_2) = |X(t_1) - Y(t_2)|. \tag{1}$$

The value $|X(t_1) - Y(t_2)|$ defines the distance between the realization of the first time series at time t_1 and the realization of the second time series at time t_2. Other distances can be considered and our method described below applies without modifications for any possible choice of distances. Depending on the nature of the time series, it may be interesting to use other distances, which for instance put more weight on large discrepancies $|X(t_1) - Y(t_2)|$ such as by using distances of the form $|X(t_1) - Y(t_2)|^q$ with $q > 1$.

When $Y(t)$ is the same time series as $X(t)$, a matrix deduced from (1) by introducing a threshold so that entries of the matrix (1) smaller (respectively larger) than the threshold are set to 0 (respectively 1) has been introduced by Eckmann et al. (1987) under the name "recurrence plot" to analyze complex chaotic time series. In the physical literature, the binary matrix deduced from (1) with the use of a threshold for two different time series is called a cross-recurrence plot. This matrix and several of its statistical properties have

been used to characterize the cross-correlation structure between pairs of time series by Strozzia et al. (2002), Quiroga et al. (2002), Marwan and Kurths (2002) and Marwan et al. (2002).

Consider the simple example in which $Y(t) = X(t-k)$ with $k > 0$ fixed. Then, $\epsilon(t_1, t_2) = 0$ for $t_2 = t_1 + k$ and is typically non-zero otherwise. The detection of this causal relationship then amounts in this case to find the line with zero values which is parallel to the main diagonal of the distance matrix. This line defines the affine mapping $t_2 = \phi(t_1) = t_1 + k$, corresponding to a constant translation. More generally, we would like to determine a sequence of elements of this distance matrix along which the elements are the smallest, as we describe next.

As this last example makes clear, the distance matrix (1) tracks the co-monotonic relationship between X and Y. But, two time series can be more anti-monotonic than monotonic, i.e., they tend to take opposite signs. Then, to diagnose such occurrence, we need to consider the "anti-monotonic" distance

$$\epsilon_{\text{anti}}(t_1, t_2) = |X(t_1) + Y(t_2)|. \tag{2}$$

The + sign makes the distance minimum when X and Y have opposite signs. In the following, we do not explore this and other possibilities and only use (1). This implies that we bias our search for lagged dependence between two time series towards lagged co-monotonic behaviors.

2.2. Optimal path at zero temperature

When the relationship between $X(t_1)$ and $Y(t_2)$ is more complex than a simple constant lead–lag of the form $Y(t) = X(t-k)$, the determination of the correspondence between the two time series is less obvious. A first approach would correspond to associate to each entry $X(t_1)$ of the first time series the value $Y(t_2)$ of the second time series which makes the distance (1) minimum over all possible t_2 for a fixed t_1. This defines the mapping $t_1 \to t_2 = \phi(t_1)$ from the t_1-variable to the t_2-variable as

$$\phi(t_1) = \text{Min}_{t_2} |X(t_1) - Y(t_2)|. \tag{3}$$

Note that this procedure analyzes each time t_1 independently of the others. The problem with this approach is that it produces mappings $t_2 = \phi(t_1)$ with two undesirable properties: (i) numerous large jumps and (ii) absence of one-to-one matching (ϕ is no more a function since the curve can have overhangs and "cliffs") which can also be viewed as a backward (non-causal) time propagation. Property (i) means that, in the presence of noise in two time series, with large probability, there will be quite a few values of t_1 such that $\phi(t_1 + 1) - \phi(t_1)$ is large and of the order of the total duration N of the time series. Most of the time, we can expect lags to be slowly varying function of time and large jumps in the function ϕ are not reasonable. The second property means that, with large probability, a given t_1 could be associated with several t_2, and therefore there will be pairs of times $t_1 < t_1'$ such that $\phi(t_1) > \phi(t_1')$: an occurrence in the future in the first time series is associated with an event in the past in the second time series. This is not excluded as lags between two time series can shift from positive to negative as a function of time, as in our example (17) below. But such occurrences should be relatively rare in real-time series which are not dominated by noise. Obviously, these two properties disqualify the method (3) as a suitable construction of a time correspondence between the two time series. This

reflects the fact that the obtained description of the lag structure between the two time series is erratic, noisy and unreliable.

To address these two problems, a first natural idea is to search for a smooth mapping $t_1 \to t_2 = \phi(t_1)$:

$$0 \leqslant \phi(t_1 + 1) - \phi(t_1) \leqslant 1. \tag{4}$$

In the continuous time limit, this amounts to imposing that the mapping ϕ should be continuous. Then, the correspondence $t_1 \to t_2 = \phi(t_1)$ can be interpreted as a reasonable time-lag or time-lead structure of the two time series. For some applications, it may be desirable to constraint even further by ensuring the differentiability (and not only the continuity) of the mapping (in the continuous limit). This can be done by a generalization of the global optimization problem (5) defined below by adding a path "curvature" energy term. Here, we do not pursue this idea further. Then, the causal relationship between the two time series is searched in the form of a mapping $t_2 = \phi(t_1)$ between the times $\{t_1\}$ of the first time series and the times $\{t_2\}$ of the second time series such that the two times series are the closest in a certain sense, i.e., $X(t_1)$ and $Y(\phi(t_1))$ match best, in the presence of these two constraints.

To implement these ideas, our first proposal is to replace the mapping (3) determined by a local minimization by a mapping obtained by the following global minimization:

$$\mathrm{Min}_{\{\phi(t_1), t_1 = 1,2,\ldots,N\}} \sum_{t_1=1}^{N} |X(t_1) - Y(\phi(t_1))|, \tag{5}$$

under the constraint (4). Note that, without the constraint (4), the solution for the mapping of the minimization (5) would recover the mapping obtained from the local minimization (3), as the minimum of the unconstrained sum is equal to the sum of the minima. In contrast, the presence of the continuity constraint changes the problem into a global optimization problem.

This problem has actually a long history and has been extensively studied, in particular in statistical physics (see Halpin-Healy and Zhang (1995) for a review and references therein), under the name of the "random directed polymer at zero temperature." Indeed, the distance matrix $E_{X,Y}$ given by (1) can be interpreted as an energy landscape in the plane (t_1, t_2) in which the local distance $\epsilon(t_1, t_2)$ is the energy associated with the node (t_1, t_2). The continuity constraint means that the mapping defines a path or line or "polymer" of equation $(t_1, t_2 = \phi(t_1))$ with a "surface tension" preventing discontinuities. The condition that $\phi(t_1)$ is non-decreasing translates in the fact that the polymer should be directed (it does not turn backward and there are no overhangs). The global minimization problem (5) translates into searching for the polymer configuration with minimum energy. In the case where the two time series are random, the distance matrix (and thus energy landscape) is random, and the optimal path is then called a random directed polymer at zero temperature (this last adjective "at zero temperature" will become clear in Sections 3.3 and 4.2; it suffices to say here that it corresponds to searching for the absolute minimum). Of course, we are interested in non-random time series, or at least in time series with some non-random components: this amounts to having the distance matrix and the energy landscape to have hopefully coherent structures (i.e., non-white noise) that we can detect. Intuitively, the lag–lead structure of the two time series will reveal itself through the organization and structure of the optimal path.

It is important to stress the non-local nature of the optimization problem (5), as the best path from an origin to an end point requires the knowledge of the distance matrix (energy landscape) $E_{X,Y}$ both to the left as well as to the right of any point in the plane (t_1, t_2). There is a general and powerful method invented by Derrida et al. (1978) and Derrida and Vannimenus (1983) to solve this problem in polynomial time using the transfer matrix method. Fig. 1 shows a realization of the distance (or energy) landscape $E_{X,Y}$ given by (1) and the corresponding optimal path.

The two next sections describe two different implementations of the search for the optimal lag path between two time series. In the next Section 3, we relax the condition (4) by allowing larger jumps on no more that one time series. This natural approach, as we explain, turns out to be sub-performing when compared with the approach of the subsequent Section 4 implementing the condition (4).

3. Optimal paths with vertical and horizontal moves forbidden: Multi-layer algorithms

3.1. The matching of two time series with vertical and horizontal moves forbidden

For the sake of illustration, let us consider two time series of stock prices $P_1(t_1)$ and $P_2(t_2)$. We are interested in testing their possible lagged dependence. For this, we need to compare their returns because they are more stationary than the prices themselves. We first need to construct a two-dimensional landscape in the coordinates (t_1, t_2). Consider a pair (t_1, t_2) and its price pair $(P_1(t_1), P_2(t_2))$. This pair of prices can be viewed from the view point of many previous price pairs $(P_1(t_1 - i), P_2(t_2 - j))$, leading to the formation of the return pairs $\left(\ln\left[\frac{P_1(t_1)}{P_1(t_1-i)}\right], \ln\left[\frac{P_2(t_2)}{P_2(t_2-j)}\right]\right)$. It is the return pairs that we would like to match along some lag trajectory to define the optimal lagged-dependence structure between to the price time series.

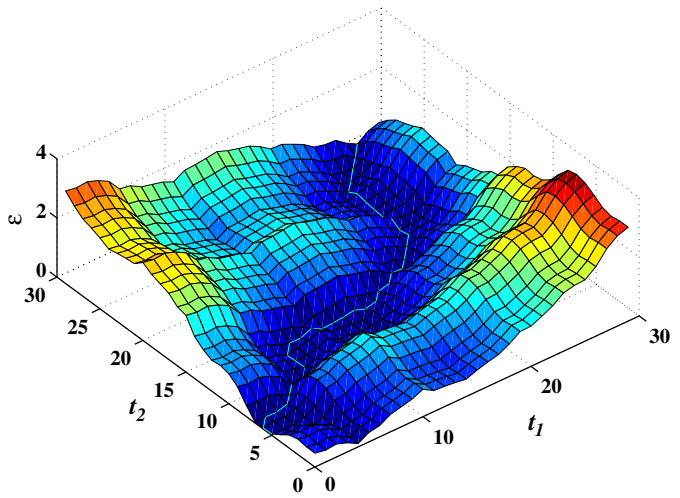

Fig. 1. An example of energy landscape $E_{X,Y}$ given by (1) for two noisy time series and the corresponding optimal path wandering at the bottom of the valley similarly to a river. This optimal path defines the mapping $t_1 \to t_2 = \phi(t_1)$.

It is easy to realize that, in order to break down the path into its most elementary constituents, we need to impose that at least one of the time increments i,j is unity. For $i=1$ and $j \geqslant 1$, this means that the first time series has its price changing from $P(t_1-1)$ to $P(t_1)$ with a return $\ln[P_1(t_1)/P_1(t_1-1)]$. Meanwhile, the second time series has its price changing from $P(t_2-j)$ to $P(t_2)$ with a return $\ln[P_2(t_2)/P_2(t_2-j)]$. Comparing the return $\ln[P_1(t_1)/P_1(t_1-1)]$ with the return $\ln[P_2(t_2)/P_2(t_2-j)]$ means that the time t_2 of P_2 is accelerating or flowing faster (or contracting) with respect to the time t_1 of P_1 and thus the full total return from t_2-j to t_2 must be compared with the return of the first time series from t_1-1 to t_1. This leads to defining a distance (or "energy")

$$\epsilon[(t_1-1,t_2-j)\to(t_1,t_2)] = \left|\ln\left[\frac{P_1(t_1)}{P_1(t_1-1)}\right] - \ln\left[\frac{P_2(t_2)}{P_2(t_2-j)}\right]\right|, \tag{6a}$$

associated with the transition (or link) from (t_1-1,t_2-j) to (t_1,t_2). Symmetrically, each link from (t_1-i,t_2-1) to (t_1,t_2) is assigned the distance

$$\epsilon[(t_1-i,t_2-1)\to(t_1,t_2)] = \left|\ln\left[\frac{P_1(t_1)}{P_1(t_1-i)}\right] - \ln\left[\frac{P_2(t_2)}{P_2(t_2-1)}\right]\right|. \tag{6b}$$

For each of the remaining bands, the energy is infinite. In other words, paths along these links are forbidden. In the present implementation, only these links are allowed and all others are forbidden, corresponding to an infinite distance (or energy).

In this algorithm, we forbid paths that have segments which are vertical or horizontal, because a vertical segment in the (t_1,t_2) plane means that the time t_1 of the first time series stops while the time t_2 of the second time series continues to flow. This situation seems implausible from an intuitive point of view. The formulation which takes into account the *discreteness* of the time steps is the following (all this discussion pertains to issues raised by the discreteness of time in the recorded the time series) : either (a) t_1 increases by one time step and t_2 increases by at least one time step (This situation corresponds to an acceleration of the time flow for the second time series compared with the first time series); or (b) t_1 increases by at least one time step and t_2 increases by one time step (this situation corresponds to an acceleration of the time flow of the first time series compared with the second time series).

This specification ensures that both time series have their time flowing at each increment along the optimal path of lags that matches the two time series. Recall that, along the optimal path, the lag between the two time series may vary, which implies a transient time contraction or dilation of one time series with respect to the other. But, of course, if one time series has its time slowing down (as in Einstein's special relativity) compared with the other time series, it does not stop completely. This implies in the discrete time formulation that, from one node of the path to the next, both time series must have their time increase by at least one time step at least. This thus automatically disqualifies optimal paths which would contain vertical and horizontal links. Such vertical and horizontal links are forbidden in the present implementation.

For each node (t_1,t_2) on the optimal path, there must exist an ancestor node at (t_1-1,t_2-j) where $j=1,2,\ldots,J$ or at (t_1-i,t_2-1) where $i=1,2,\ldots,I$. The values I and J are the maximum time contraction factors between the two time series. For instance, values $I=J=3$ mean that one time series cannot exhibit a transient time flow more than three times faster that the other one. Fig. 2 shows the topology of all the elementary links ending at a given node for $I=J=3$, $I=J=5$, $I=J=7$, and $I=J=10$, from

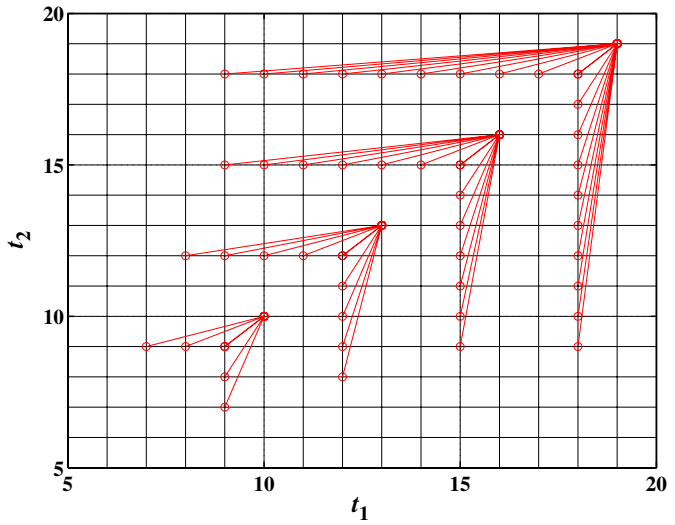

Fig. 2. Topology of all the elementary links ending at a given node for $I = J = 3$, $I = J = 5$, $I = J = 7$, and $I = J = 10$, from bottom-left to top-right.

bottom-left to top-right. All nodes that are possible ancestors to a given node (t_1, t_2) are shown in Fig. 2. The best path reaching (t_1, t_2) is thus carried by one of these segments.

3.2. Optimal path at zero temperature

At zero temperature, we can construct the optimal path with the help of transfer matrix method (Derrida et al., 1978; Derrida and Vannimenus, 1983). To each node (t_1, t_2), we associate the energy (or total distance) $E(t_1, t_2)$ defined as the minimal energy over all possible paths from the origin to (t_1, t_2). The energy of a given path from the origin to (t_1, t_2) is simply the sum of the energies of all links of this path. The path corresponding to the minimum energy (minimum matching distance) is called the optimal path. The key remark is that the energy value $E(t_1, t_2)$ can be obtained by the following recursion relationship

$$E(t_1,t_2) = \mathrm{Min}[\{E(t_1 - i, t_2 - 1) + \epsilon[(t_1 - i, t_2 - 1) \to (t_1, t_2)],\ i = 1, 2, \ldots, I\} \cup \{E(t_1 - 1, t_2 - j) + \epsilon[(t_1 - 1, t_2 - j) \to (t_1, t_2)],\ j = 1, 2, \ldots, J\}] \quad (7)$$

In words, the minimum energy of the path reaching (t_1, t_2) is obtained from the energy of one of the optimal paths reaching all the possible ancestors of (t_1, t_2) by adding the energy of the link from that ancestor to (t_1, t_2). Taking the minimum over all possible ancestors of this sum gives $E(t_1, t_2)$ and determines the corresponding optimal path from the origin to (t_1, t_2).

To see what is the outcome of the aforementioned procedure, let us consider two price series P_1 and P_2 constructed as the cumulative sums of two stationary time series $X(t_1)$ and $Y(t_2)$ respectively (which are the one-time-step returns). Let us construct $Y(t_2)$ from $X(t_1)$ as follows:

$$Y(t_2) = aX(t_2 - \tau) + \eta, \quad (8a)$$

where a is a constant, τ is the time lag, and the noise $\eta \sim N(0, \sigma_\eta)$ is serially uncorrelated. The time series $X(t_1)$ itself is generated from an AR process:

$$X(t_1) = bX(t_1 - 1) + \xi, \tag{8b}$$

where $b < 1$ and the noise $\xi \sim N(0, \sigma_\xi)$ is serially uncorrelated. The factor $f = \sigma_\eta/\sigma_\xi$ quantifies the amount of noise degrading the causal relationship between $X(t_1)$ and $Y(t_2)$. A small f corresponds to a strong causal relationship. A large f implies that $Y(t_2)$ is mostly noise and becomes unrelated to $X(t_1)$ in the limit $f \to \infty$. Specifically, $\mathrm{Var}[X] = \sigma_\xi^2/(1 - b^2)$ and

$$\mathrm{Var}[Y] = a^2\mathrm{Var}[X] + \sigma_\eta^2 = \sigma_\xi^2\left(\frac{a^2}{1 - b^2} + f^2\right) = \sigma_\xi^2\left(\frac{a^2\mathrm{Var}[X]}{\sigma_\xi^2} + f^2\right). \tag{9}$$

In our simulations, we generate X and Y of size 50 with parameters $a = 1$, $b = 0.7$, $\tau = -10$, and $f = 0$. From the definition of this model, the true path expressing the lag between X and Y should be characterized by a fixed time lag $\tau = 10$. One thus expects that the optimal path reconstructed by our algorithm should be close to this true path at time lag $\tau = 10$. We have performed extensive simulations of the model (8) using different values of $I = J$. Fig. 3 finds that the obtained optimal paths tend to track the true solution (which is parallel to the diagonal but translated by 10 time steps). Some departures at the beginning and at the end of the path are expected since the search of the optimal path imposes that it starts at the origin and ends on the main diagonal. However, we can observe in Fig. 3 the existence of deviations from the true path in the last third of the path, which are larger than expected.

This comes from the fact that our search algorithm has too much choice, by allowing paths to jump along so many possible links. Because we allow paths to compare the return of one day in one time series with the return of up to 10 days in the other time series (for

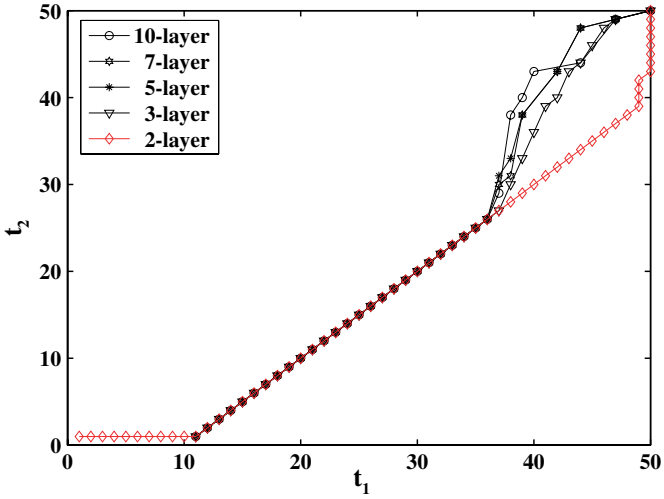

Fig. 3. Optimal paths of a typical realization of model (8) obtained with the algorithm described in Section 3.2 with $I = J = 3$, $I = J = 5$, $I = J = 7$, and $I = J = 10$. In addition, the optimal path obtained by the different algorithm described in Section 4 is also shown for comparison.

$I = J = 10$), we find spuriously very good matching due to this very large choice offered for $I = J = 10$. This suggests that we should reduce the number of available links, that is, decrease $I = J$.

We have thus tried the different values $I = J = 3$, $I = J = 5$, $I = J = 7$, and $I = J = 10$, associated with the links shown in Fig. 2. For each realization of a pair (X, Y), we obtain the four optimal paths associated with the four values $I = J = 3, 5, 7, 10$. In the example of Fig. 3, decreasing $I = J$ tends to bring the optimal path towards the true solution, albeit not much. In comparison, the optimal path recovered with the algorithm developed in Section 4 behaves significantly better by approaching much more the true solution. This result is general: we find systematically in all our simulations that the smaller values $I = J$ give optimal paths closer to the true one. Moreover, the algorithm of the next Section 4 is found systematically superior to the present one. Before turning to it, we investigate the effect of an improvement which will also be used in Section 4.

3.3. Optimal path at finite temperature

While appealing, the optimization program (5) has an important potential drawback: it assumes that the distance matrix $E_{X,Y}$ between the time series X to Y defined by (1) is made only of useful information. But, in reality, the time series $X(t_1)$ and $Y(t_2)$ can be expected to contain significant amount of noise or more generally of irrelevant structures stemming from random realizations. Then, the distance matrix $E_{X,Y}$ contains a possibly significant amount of noise, or in other words of irrelevant patterns. Therefore, the global optimal path obtained from the procedure of the previous Section 3.2 is bound to be delicately sensitive in its conformation to the specific realizations of the noises of the two time series. Other realizations of the noises decorating the two time series would lead to different distance matrices and thus different optimal paths. In the case where the noises dominates, this question amounts to investigating the sensitivity of the optimal path with respect to changes in the distance matrix. This problem has actually be studied extensively in the statistical physics literature (see Halpin-Healy and Zhang (1995) and references therein). It has been shown that small changes in the distance matrix may lead to very large jumps in the optimal path, when the distance matrix is dominated by noise. Clearly, these statistical properties would led to spurious interpretation of any causal relationship between the two time series. We thus need a method which is able to distinguish between truly informative structure and spurious patterns due to noise.

In a realistic situation, we can hope for the existence of coherent patterns in addition to noise, so that the optimal path can be "trapped" by these coherent structures in the energy landscape. Nevertheless, the sensitivity to specific realizations of the noise of the two time series may lead to spurious wandering of the optimal path, that do not reflect any genuine lag–lead structure. We thus propose a modification of the previous global optimization problem to address this question and make the determination of the mapping more robust and less sensitive to the existence of noise decorating the two time series. Of course, it is in general very difficult to separate the noise from the genuine signal, in absence of a parametric model. The advantage of the method that we now propose is that it does not require any a priori knowledge of the underlying dynamics.

The idea of the "optimal thermal causal path" method is the following. Building on the picture of the optimal path as being the conformation of a polymer or of a line minimizing its energy E in a frozen energy landscape determined by the distance matrix, we now pro-

pose to allow from "thermal" excitations or fluctuations around this path, so that path configurations with slightly larger global energies are allowed with probabilities decreasing with their energy. We specify the probability of a given path configuration with energy ΔE above the absolute minimum energy path by a multivariate logit model or equivalently by a so-called Boltzmann weight proportional to $\exp[-\Delta E/T]$, where the "temperature" T quantifies how much deviations from the minimum energy are allowed. For $T \to 0$, the probability for selecting a path configuration of incremental energy ΔE above the absolute minimum energy path goes to zero, so that we recover the previous optimization problem "at zero temperature." Increasing T allows to sample more and more paths around the minimum energy path. Increasing T thus allows us to wash out possible idiosyncratic dependencies of the path conformation on the specific realizations of the noises decorating the two time series. Of course, for too large temperatures, the energy landscape or distance matrix becomes irrelevant and one looses all information in the lag–lead relationship between the two time series. There is thus a compromise as usual between not extracting too much from the spurious noise (not too small T) and washing out too much the relevant signal (too high T). Increasing T allows one to obtain an average "optimal thermal path" over a larger and larger number of path conformations, leading to more robust estimates of the lag–lead structure between the two time series. The optimal thermal path for a given T is determined by a compromise between low energy (associated with paths with high Boltzmann probability weight) and large density (large number of contributing paths of similar energies as larger energies are sampled). This density of paths contributing to the definition of the optimal thermal path can be interpreted as an entropic contribution added to the pure energy contribution of the optimization problem of the previous Section 3.2. In a sense, the averaging over the thermally selected path configurations provides an effective way of averaging over the noise realizations of the two time series, without actually having to resampling the two times series. This intuition is confirmed by our tests below which show that the signal-over-noise ratio is indeed increased significantly by this "thermal" procedure.

Let us now describe how to implement this idea. It is convenient to use the rotated coordinate system (x,t) where t is in the main diagonal direction of the (t_1,t_2) system and x is perpendicular to t. The transformation from the coordinates (t_1,t_2) to (x,t) is given by

$$\begin{cases} t_1 = 1 + (t-x)/2, \\ t_2 = 1 + (t+x)/2. \end{cases} \tag{10}$$

We stress that the origin is ($t_1 = 1$, $t_2 = 1$) and ($x = 0$, $t = 0$).

The present framework aims at calculating different thermal averages over all possible path fluctuations at a given temperature. Of interest are the average path $\langle x(t)\rangle$ and its standard deviation. Actually, any such average can be calculated once the so-called partition functions $G(x,t)$ at time t is known.

The local time lag $\langle x(t)\rangle$ at time t is given by

$$\langle x(t)\rangle = \sum_x xG(x,t)/G(t), \tag{11}$$

where $G(t) = \sum_x G(x,t)$. In statistical physics, $G(x,t)$ is called the partition function constrained to x while $G(t)$ is the total partition function at t. Then, $G(x,t)/G(t)$ is nothing but the probability for a path be at x from the diagonal for a distance t along the diagonal. Expression (11) indeed defines $\langle x\rangle$ as the (thermal) average of the local time lag X at t.

It is standard to call it "thermal average" because G is made of the Boltzmann factors that weight each path configuration.

The partition function can be determined recursively according to the following formula, which uses the specific topology of the links described in Fig. 2:

$$G(x,t) = G(x,t-2)\mathrm{e}^{-\epsilon[(x,t-2)\to(x,t)]/T} + \sum_{j=1}^{J} G(x-j,t-2-j)\mathrm{e}^{-\epsilon[(x-j,t-2-j)\to(x,t)]/T} + \sum_{i=1}^{I} G(x+i,t-2-i)\mathrm{e}^{-\epsilon[(x+i,t-2-i)\to(x,t)]/T}, \tag{12}$$

where $\epsilon[(t_0,x_0)\to(t_1,x_1)]$ is determined by Eq. (6) and T is the temperature.

To calculate $G(x,t)$ at the tth layer, we need to know and bookkeep the previous $\max(I,J)+1$ layers from $G(\cdot,t-11)$ to $G(\cdot,t-1)$. Of course, the layer $G(\cdot,t-1)$ is not used for the calculation of $G(\cdot,t)$ but for $G(\cdot,t+1)$. After $G(\cdot,t)$ is determined, the G's at the $\max(I,J)+1$ layers are normalized by $G(t)$ so that $G(x,t)$ does not diverge at large t. We stress that the boundary condition of $G(x,t)$ plays an crucial role. For $t=0$ and $t=1$, $G(x,t)=1$ so that Eq. (11) holds. For $t>1$, the boundary condition is taken to be $G(x=\pm t,t)=0$, in order to prevent paths to stay on the boundaries.

Let us test this algorithm on the autoregressive model (8). For this, we generate two time series X and Y of size 50 with parameters $a=1$, $b=0.7$, $\tau=-10$, and $f=0$ as in Section 3.2. We have performed extensive simulations of the model (8) using the multi-layer algorithms for different $I=J$ values and for different noise levels. For each realization (a pair of X and Y), we calculate the local time lag $\langle x\rangle$ using Eq. (11) at temperature $T=1$ (since the time series are normalized by their own standard deviations, this value $T=1$ corresponds approximately to an equal weight given to the energy and to the entropy contributions in the determination of the optimal path). The resulting time lags from different

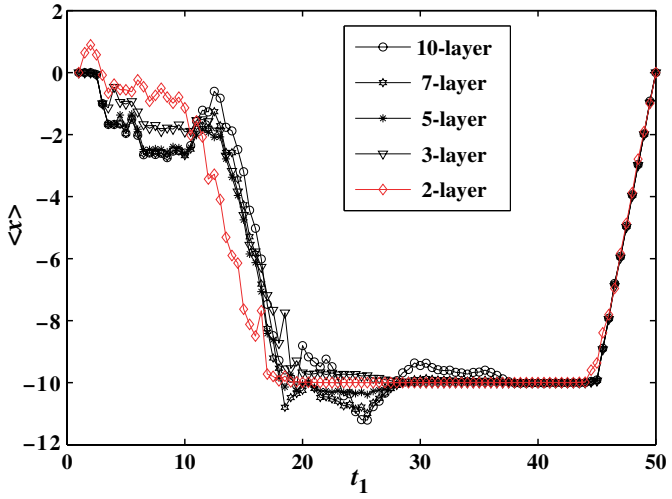

Fig. 4. Average thermal paths at temperature $T=1$ for a realization of model (8) obtained with the algorithm described in Section 3.3 with $I=J=3$, $I=J=5$, $I=J=7$, and $I=J=10$. In addition, the optimal path obtained by the different algorithm described in Section 4.2 is also shown for comparison.

multi-layer algorithms are compared with those obtained from the two-layer algorithm described below in detail in Section 4.2. In general, the local time lag $\langle x \rangle$ of the two-layer algorithm is closer to the theoretical value $\tau = -10$ than those of the multi-layer algorithms. At high temperatures ($T > 5$), all the algorithms fail to extract the correct $\langle x \rangle$. For lower temperatures, the results are similar to the case $T = 1$. A typical example is illustrated in Fig. 4. Since the two time series are lagged and have finite size and the optimal path starts and ends at $x = 0$, there is a transient at the two extremities of the paths characterized by the function $\langle x(t) \rangle$.

We have also tested these algorithms using two time series independently generated from a Gaussian distribution, which corresponds to model (8) with $a = 1$, $b = 0$, $\tau = -10$. We find that the multi-layer algorithms fail to find the correct $\langle x \rangle$ in the majority of the realizations, while the two-layer method that we now describe does a very good job.

4. Optimal paths with vertical and horizontal moves allowed: The "optimal thermal causal path" or two-layer algorithm

The previous tests have been very disappointing. We have shown only an example with a fixed time lag for a simple auto-regressive model. There, the algorithm has not been very convincing. We explain these results, as already mentioned, by the too large choices given in the search of the optimal lag path. By allowing paths to jump over up to I or J time series in one time series while the other is updated by a single time steps gives so many configurations that the optimal path does not bring much information on the real lag structure but only tries by combinatorial force to minimize its mismatch (energy).

The algorithm we present now is much more restrictive in the configurations that the path is allowed to sample. It turns out to provide very good results, as we will show by testing the method on different synthetic time series before applying it to concrete economic questions.

4.1. Optimal path at zero temperature

The algorithm we consider comes back to the condition (4) discussed at the beginning. This condition is shown in Fig. 5, which also defines the notations. The topology of the connecting bonds is very different from that of Fig. 2. The optimal path (and thus mapping) is constructed such that it can either go horizontally by one step from (t_1, t_2) to $(t_1 + 1, t_2)$, vertically by one step from (t_1, t_2) to $(t_1, t_2 + 1)$ or along the diagonal from (t_1, t_2) to $(t_1 + 1, t_2 + 1)$, as shown by the three arrows. The restriction to these three possibilities embodies the continuity condition (4) and the one-to-one mapping (for vertical segments the one-to-one correspondence is ensured by the convention to map t_1 to the largest value t_2 of the segment) between to the two time series. However, this comes at the cost that one time series may appear to "freeze" along the vertical or horizontal bonds while the other has its time still flowing.

As before, the optimal path for two identical time series is the main diagonal, so deviations from the diagonal quantify lag or lead times between the two time series. It is thus convenient to use a rotated frame (t, x) as shown in Fig. 5 such that the second coordinate x quantifies the deviation from the main diagonal, hence the lead or lag time between the two time series. In general, the optimal path is expected to wander around, above or below the main diagonal of equation $x(t) = 0$.

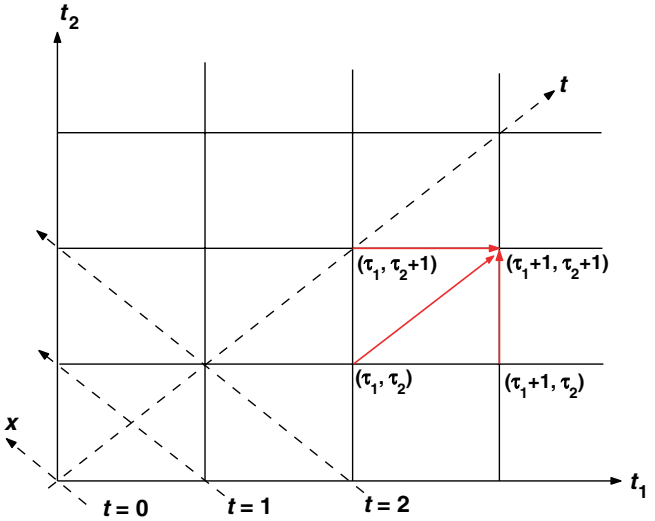

Fig. 5. Representation of the two-layer approach in the lattice (t_1, t_2) and of the rotated frame (t, x) as defined in the text. The three arrows depict the three moves that are allowed to reach any node in one step, in accordance with the continuity and monotonicity conditions (4).

A given node (t_1, t_2) in the two-dimensional lattice carries the "potential energy" or distance $\epsilon(t_1, t_2)$. Let us now denote $E(t_1, t_2)$ as the energy (cumulative distance (5)) of the optimal path starting from some origin $(t_{1,0}, t_{2,0})$ and ending at (t_1, t_2). Using the present topology of Fig. 5, the transfer matrix method providing the determination of the optimal lag path is based on the following fundamental relation:

$$E(t_1, t_2) = \epsilon(t_1, t_2) + \mathrm{Min}[E(t_1 - 1, t_2), E(t_1, t_2 - 1), E(t_1 - 1, t_2 - 1)]. \tag{13}$$

The key insight captured by this equation is that the minimum energy path that reaches point (t_1, t_2) can only come from one of the three points $(t_1 - 1, t_2)$, $(t_1, t_2 - 1)$ and $(t_1 - 1, t_2 - 1)$ preceding it. Then, the minimum energy path reaching (t_1, t_2) is nothing but an extension of the minimum energy path reaching one of these three preceding points, determined from the minimization condition (13). Then, the global optimal path is determined as follows. One needs to consider only the sub-lattice $(t_{1,0}, t_{2,0}) \times (t_1, t_2)$ as the path is directed. The determination of the optimal path now amounts to determining the forenode of each node in the sub-lattice $(t_{1,0}, t_{2,0}) \times (t_1, t_2)$. Without loss of generality, assume that $(t_{1,0}, t_{2,0})$ is the origin $(1, 1)$. Firstly, one performs a left-to-right and bottom-to-up scanning. The forenode of the bottom nodes $(\tau_1, 1)$ is $(\tau_1 - 1, 1)$, where $\tau_1 = 2, \ldots, t_1$. Then, one determines the forenodes of the nodes in the second-layer at $t_2 = 2$, based on the results of the first (or bottom) layer. This procedure is performed for $t_2 = 3$, then for $t_2 = 4, \ldots$, and so on.

The global minimization procedure is fully determined once the starting and ending points of the paths are defined. In some of our tests, it is simpler to impose the start and end points to be on the diagonal, i.e., $x = 0$. But more generally, since the lag–leads between two time series can be anything at any time, we should allow in general the starting point to lie anywhere on the horizontal axis $t_2 = 1$ or on the vertical axis $t_1 = 1$.

Similarly, we should allow the ending point to lie anywhere on the horizontal axis $t_2 = N$ or on the vertical axis $t_1 = N$. This allows for the fact that one of the two time series may precede the other one. For each given pair of starting and ending points, we obtain a minimum path (the "optimal directed polymer" with fixed end-points). The minimum energy path over all possible starting and ending points is then the solution of our global optimization problem (5) under the constraint (4). This equation of this global optimal path defines the mapping $t_1 \to t_2 = \phi(t_1)$ defining the causal relationship between the two time series.

As an example, we construct the distance matrix of the normalized returns of IBM and MSFT stocks from 2001/05/16 to 2001/06/20, as shown in Fig. 6. The normalized returns are in the first row (bottom) and first column (left). The corresponding distance matrix is with the optimal path in bold. The straight line characterizes the diagonal (no time lag).

4.2. Optimal path at finite temperature

Similarly to our discussion in Section 3.3, it is natural to define thermally average paths in the hope of obtaining more reliable lag recoveries. Here, we describe how we implement the idea of thermal path fluctuations in the context of the present algorithm.

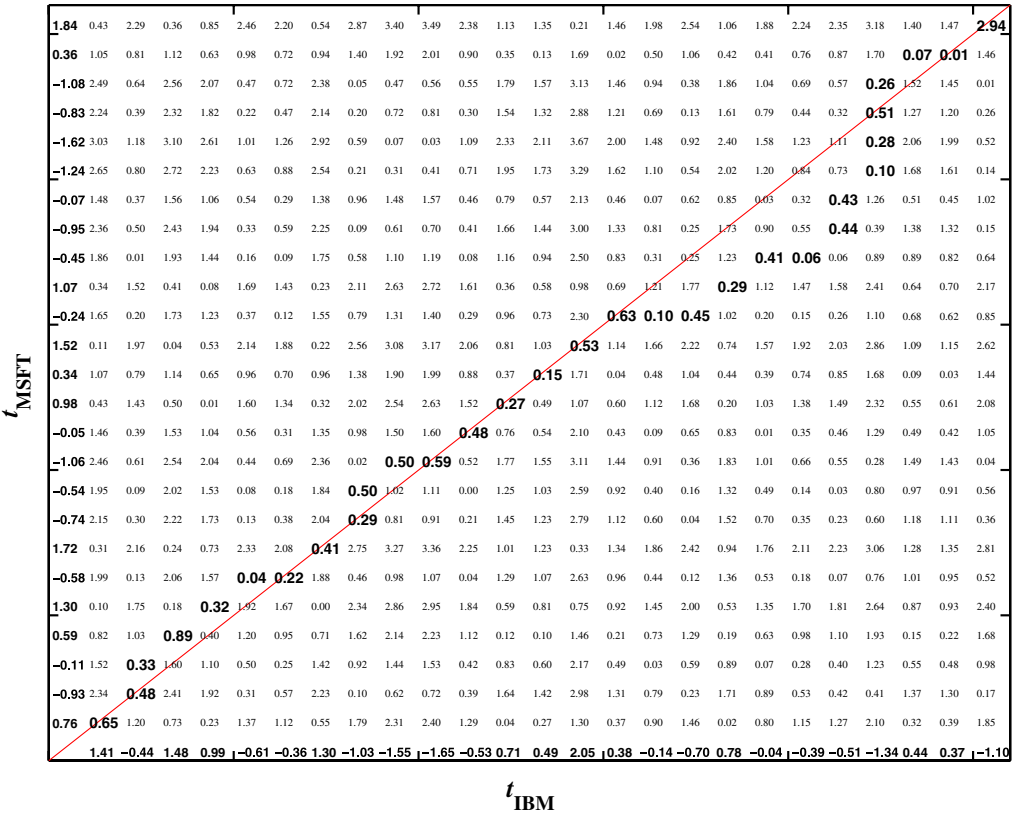

Fig. 6. Distance matrix of the normalized returns of IBM and MSFT stocks from 2001/05/16 to 2001/06/20. The normalized returns are in the first row (bottom) and first column (left). The corresponding distance matrix is with the optimal path in bold. The straight line characterizes the diagonal (no time lag).

It is again convenient to use the rotated frame (t, x) as defined in Fig. 5, in which t gives the coordinate along the main diagonal of the (t_1, t_2) lattice and x gives the coordinate in the transverse direction from the main diagonal. The transformation between the two coordinates systems was given in Eq. (10). Note that the constraint that the path is directed allows us to interpret t as an effective time and x as the position of a path at that "time" t. The local time lag $\langle x(t)\rangle$ is given by Eq. (11). Similarly to expression (11), the variance of the trajectory of the optimal thermal path reads

$$\sigma_x^2 = \sum_x (x - \langle x\rangle)^2 G(x,t)/G(t). \tag{14}$$

The variance σ_x^2 gives a measure of the uncertainty in the determination of the thermal optimal path and thus an estimate of the error in the lag–lead structure of the two time series as seen from this method.

If two time series are perfectly causally related (they are the same up to a factor), then the optimal path is the diagonal, that is, made of the diagonal bonds of the square lattice, or alternatively the nodes on the diagonals. Since the "energy" (i.e., local mismatch defined by expression (1)) is defined only on the nodes, a path has a Boltzmann weight contributed only by the nodes and there is no contribution from bonds. We should thus allow path not only along the horizontal and vertical segments of each square of the lattice but also along the main diagonal of each square. The directedness means that a given path is not allowed to go backward on any of the three allowed moves. As illustrated in Fig. 5, in order to arrive at $(t_1 + 1, t_2 + 1)$, the path can come from $(t_1 + 1, t_2)$ vertically, $(t_1, t_2 + 1)$ horizontally, or (t_1, t_2) diagonally. The recursive equation on the Boltzmann weight factor is thus

$$G(x, t+1) = [G(x-1,t) + G(x+1,t) + G(x,t-2)]\mathrm{e}^{-\epsilon(x,t)/T}. \tag{15}$$

The intuition is to imagine the polymer/path as fluctuating randomly due to random "thermal kicks" in the quenched random energy landscape. In the limit where the temperature T goes to zero, $G(x,t)/G(t)$ becomes the Dirac function $\delta[x - x_{DP}(t)]$ where $x_{DP}(t)$ is the position of the global optimal path determined previously in Section 2.2. Thus, for $T \to 0$, expression (11) leads to $\langle x\rangle = x_{DP}(t)$, showing that this thermal procedure generalizes the previous global optimization method. For non-vanishing T, the optimal thermal average $\langle x(t)\rangle$ given by (11) takes into account the set of the neighboring (in energy) paths which allows one to average out the noise contribution to the distance matrix. This recursion relation uses the same principle and has thus the same structure as expression (13) (Wang et al., 2000).

As we have learned in Section 4.2, this two-layer approach outperforms the multi-layer algorithms. In the next section, we shall determine the impact of temperature and show the potential power of the method for multi-lag time series.

5. Comprehensive numerical tests of the "optimal thermal causal path" approach

5.1. First synthetic test

Consider two stationary time series $X(t_1)$ and $Y(t_2)$ given by model (8). In our simulations, we take $\tau = 5$, $a = 0.8$, $b = 0.7$, and $\sigma_\xi = 1$ and consider time series of duration $N = 100$. For a given f, we obtain the optimal zero-temperature path by using the trans-

fer-matrix method (13) explained in Section 2.2 for 19 different starting positions around the origin and similarly 19 different ending positions around the upper-right corner at coordinate (100, 100). This corresponds to solve 19×19 transfer matrix optimization problems. The absolute optimal path is then determined as the path which has the smallest energy over all these possible starting and ending points. We also determine the optimal local time lag $\langle x(t) \rangle$, for different temperatures, typically from $T = 1/5$ to 10, using the relation (15) for the partition function.

Fig. 7(a) shows that transverse trajectory $x(t)$ as a function of the coordinate t along the main diagonal for $f = 1/10$ and for temperatures $T = 0$, 1/5, 1, and 10. This graph corresponds to the case where we restrict our attention to paths with fixed imposed starting (origin) and ending (coordinates (100, 100) on the main diagonal) points. This restriction is relaxed as we explain above and apply below to prevent the boundary effects clearly visible in Fig. 7(a). Fig. 7(b) shows the corresponding standard deviation defined by (14) of the thermal average paths.

The impact of the temperature is nicely illustrated by plotting how the energy of an optimal thermal path depends on its initial starting point $x(t = 0) = x_0$ (and ending point taken with the same value $x(t = 198) = x(0)$). For a given x_0 and temperature T, we determine the thermal optimal path and then calculate its energy $e_T(x_0)$ by the formula

$$e_T(x_0) = \frac{1}{2(N - |x_0|) - 1} \sum_{t=|x_0|}^{2N-1-|x_0|} \sum_{x} \epsilon(x, t) G(x, t)/G(t). \tag{16}$$

By construction, the time lag between the two time series is $\tau = 5$ so that we should expect $e_T(x_0)$ to be minimum for $x_0 = \tau = 5$. Fig. 8 plots $e_T(x_0)$ as a function of the average of the path $\langle x(x_0) \rangle$ with different starting points x_0 for different temperatures T respectively equal to 1/50, 1/5, 1/2, 1, 2, 5, and 10 and for $f = 1/2$. One can observe a large quasi-degeneracy for small temperatures, so that it is difficult to identify what is the value of the lag between the two time series. The narrow trough at $\langle x(x_0) \rangle = 5$ for the smallest temperatures, while at the correct value, is not clearly better than negative values of $\langle x(x_0) \rangle$. In contrast,

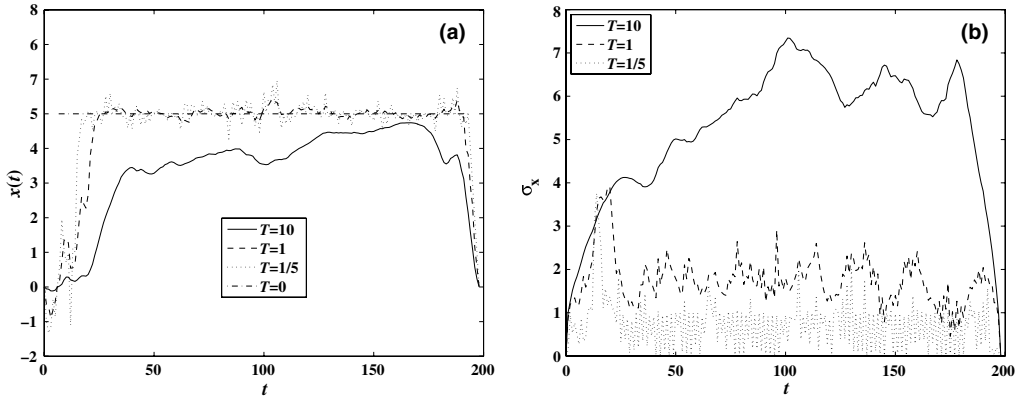

Fig. 7. (a) Thermal average $\langle x(t) \rangle$ of the transverse fluctuations with respect to t for $T = 10$, 1, and 1/5 and the directed polymer. (b) The uncertainty σ_x of the thermal average paths for different temperatures. All the paths are constrained to start from the diagonal $(x = 0, t = 0)$ and to return to it at $(x = 0, t = 198)$.

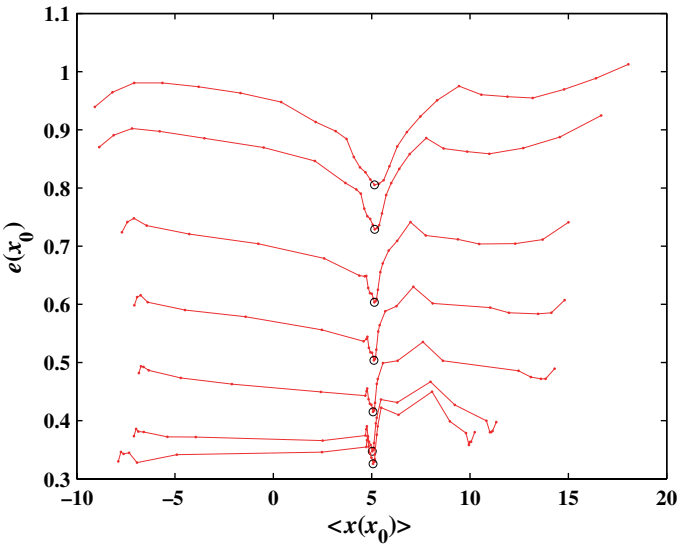

Fig. 8. Dependence of the thermal average energy $e_T(x_0)$ of the optimal thermal path as a function of the average $\langle x(x_0)\rangle$ defined in turn by the coordinate of its starting point ($t = |x_0|, x = x_0$) for different temperatures given by $T = 1/50$, $1/5$, $1/2$, 1, 2, 5 and 10 from bottom to top and for $f = 1/2$.

increasing the temperature produces a well-defined quadratic minimum bottoming at the correct value $\langle x(x_0)\rangle = \tau = 5$ and removes the degeneracies observed for the smallest temperatures. This numerical experiment illustrates the key idea underlying the introduction of the thermal averaging in Section 4.2: too small temperatures lead to optimal paths which are exceedingly sensitive to details of the distance matrix, these details being controlled by the specific irrelevant realizations of the noise η in expression (8). The theoretical underpinning of the transformation from many small competing minima to well-defined large scale minima as the temperature increases, as observed in Fig. 8, is well understood from studies using renormalization group methods (Bouchaud et al., 1991).

Fig. 9 further demonstrates the role of the temperature for different amplitudes of the noise η. It shows the position $\overline{\langle x\rangle}$ as a function of T for different relative noise level f. Recall that $\langle x(t)\rangle$ is the optimal thermal position of the path for a fixed coordinate t along the main diagonal, as defined in (11). The symbol $\overline{\langle x\rangle}$ expresses an additional average of $\langle x\rangle$ over all the possible values of the coordinate t: in other words, $\overline{\langle x\rangle}$ is the average elevation (or translation) of the optimal thermal path above (or below) the diagonal. This average position is an average measure (along the time series) of the lag/lead time between the two time series, assuming that this lag–lead time is the same for all times. In our numerical example, we should obtain $\overline{\langle x\rangle}$ close to or equal to $\tau = 5$. Fig. 9 shows the dependence of $\overline{\langle x\rangle}$ as a function of T for different values of f.

Obviously, with the increase of the signal-to-noise ratio of the realizations which is proportional to $1/f$, the accuracy of the determination of τ improves. For a noise level f, $\overline{\langle x\rangle}$ approaches the correct value $\tau = 5$ with increasing T. The beneficial impact of the temperature is clearer for more noisy signals (larger f). It is interesting to notice that an "optimal range" of temperature appears for large noise level.

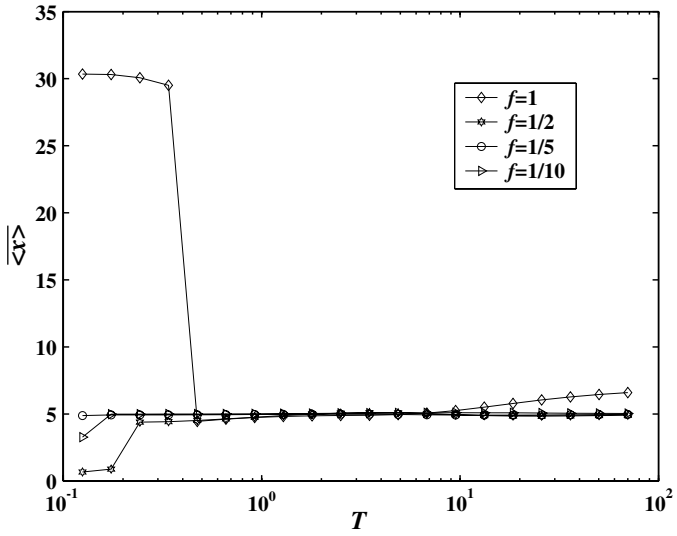

Fig. 9. Dependence of $\overline{\langle x \rangle}$ upon noise level f and temperature T.

5.2. Test on the detection of jumps or change-of-regime in time lag

We now present synthetic tests of the efficiency of the optimal thermal causal path method to detect multiple changes of regime and compare the results with a standard correlation analysis performed in moving windows of different sizes. Consider the following model:

$$Y(i) = \begin{cases} 0.8X(i) + \eta, & 1 \leqslant i \leqslant 50, \\ 0.8X(i-10) + \eta, & 51 \leqslant i \leqslant 100, \\ 0.8X(i-5) + \eta, & 101 \leqslant i \leqslant 150, \\ 0.8X(i+5) + \eta, & 151 \leqslant i \leqslant 200, \\ 0.8X(i) + \eta, & 201 \leqslant i \leqslant 250. \end{cases} \tag{17}$$

In the sense of definition (8), the time series Y is lagging behind X with $\tau = 0$, 10, 5, -5 (this negative lag time corresponds to $X(t)$ lagging behind $Y(t)$), and 0 in five successive time periods of 50 time steps each. The time series X is assumed to be the first-order AR process (8b) and η is a Gaussian white noise. Our results are essentially the same when X is itself a white Gaussian random variable. We use $f = 1/5$ in the simulations presented below.

Fig. 10 shows the standard cross-correlation function calculated over the whole time interval $1 \leqslant i \leqslant 250$ of the two time series X and Y given by (17), so as to compare with our method. Without further information, it would be difficult to conclude more than to say that the two time series are rather strongly correlated at zero time lag. It would be farfetched to associate the tiny secondary peaks of the correlation function at $\tau = \pm 5$ and 10 to genuine lags or lead times between the two time series. And since, the correlation function is estimated over the whole time interval, the time localization of possible shifts of lag/leads is impossible.

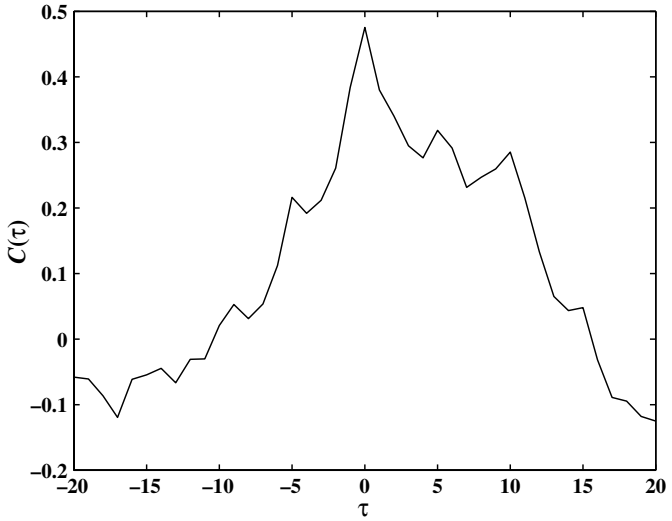

Fig. 10. Standard cross-correlation function of two time series with varying time lags $\tau = 0$, 10, 5, −5, and 0 as defined in Eq. (17).

Before presenting the results of our method, it is instructive to consider a natural extension of the correlation analysis, which consists in estimating the correlation function in a moving window $[i + 1 - D, i]$ of length D, where i runs from D to 250. We then estimate the lag–lead time $\tau_D(i)$ as the value that maximizes the correlation function in each window $[i + 1 - D, i]$. We have used $D = 10$, 20, 50, and 100 to investigate different compromises ($D = 10$ is reactive but does not give statistically robust estimates while $D = 100$ gives statistically more robust estimates but is less reactive to abrupt changes of lag). The local lags $\tau_D(i)$ thus obtained are shown in Fig. 11 as a function of the running time i. For $D = 10$, this method identifies successfully the correct time lags in the first, third, fourth, and fifth time periods, while $\tau_D(i)$ in the second time period is very noisy and fails to unveil the correct value $\tau = 10$. For $D = 20$, the correct time lags in the five time periods are identified with large fluctuations at the boundaries between two successive time periods. For $D = 50$, five successive time lags are detected but with significant delays compared to their actual inception times, with in addition high interspersed fluctuations. For $D = 100$, the delays of the detected inception times of each period reach about 50 time units, that is, comparable to the width of each period, and the method fails completely for this case.

Let us now turn to our optimal thermal causal path method. We determine the average thermal path (transverse trajectory $x(i)$ as a function of the coordinate i along the main diagonal) starting at the origin, for four different temperatures $T = 2$, 1, 1/2, and 1/5. Fig. 12 plots $x(i)$ as a function of i. The time lags in the five time periods are recovered clearly. At the joint points between the successive time periods, there are short transient crossovers from one time lag to the next. Our new method clearly outperforms the above cross-correlation analysis.

The advantage of our new method compared with the moving cross-correlation method for two time series with varying time lags can be further illustrated by a test of predictability. It is convenient to use an example with unidirectional causal lags (only positive lags)

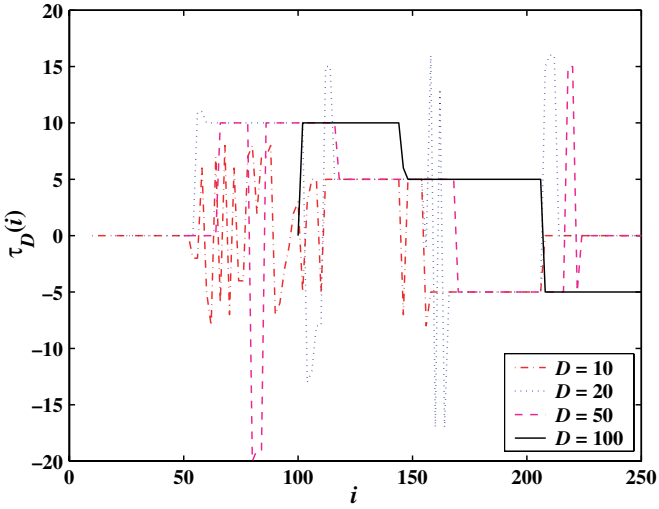

Fig. 11. Local cross-correlation analysis of the two time series defined by (17) with (8b) using moving windows of sizes $D = 10$, 20, 50, and 100. The value $\tau_D(i)$ of the lag that makes maximum the local cross-correlation function in each window $[i + 1 - D, i]$ is plotted as a function of the right-end time i. The true time lags as defined in (17) are respectively $\tau = 0$, 10, 5, -5 and 0 in five successive time periods of 50 time steps each.

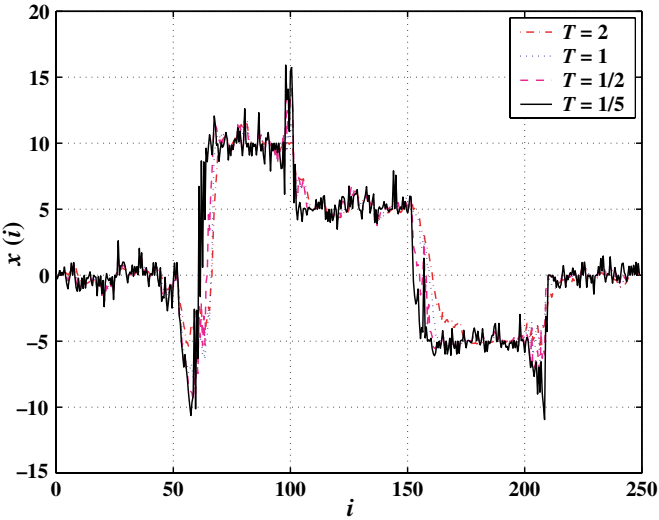

Fig. 12. Average thermal path (transverse trajectory $x(i)$ as a function of the coordinate i along the main diagonal) starting at the origin, for four different temperatures ($T = 2$ (dotted-dash), $T = 1$ (dotted), $T = 0.5$ (dashed), and 0.2 (continuous)) obtained by applying the optimal thermal causal path method to the synthetic time series (17) with (8b).

and not with bidirectional jumps as exemplified by (17). We thus consider a case in which X leads Y in general and use the following model:

$$Y(i) = \begin{cases} 0.8X(i) + \eta, & 1 \leqslant i \leqslant 50, \\ 0.8X(i-10) + \eta, & 51 \leqslant i \leqslant 100, \\ 0.8X(i-5) + \eta, & 101 \leqslant i \leqslant 150, \\ 0.8X(i-8) + \eta, & 151 \leqslant i \leqslant 200. \end{cases} \tag{18}$$

At each instant i considered to be the "present," we perform a prediction of $Y(i+1)$ for "tomorrow" at $i+1$ as follows. We first estimate the instantaneous lag–lead time $\tau(i)$. The first estimation uses the running-time cross-correlation method which delivers $\tau(i) = \tau_D(i)$. The second estimation is the average thermal position $\tau(i) = \max\{[x(i)], 0\}$ using the optimal thermal causal path method where the operator $[\cdot]$ takes the integral part of a number. We construct the prediction for $Y(i+1)$ as

$$Y(i+1) = 0.8X(i+1-\tau(i)). \tag{19}$$

In this prediction set-up, we assume that we have full knowledge of the model and the challenge is only to calibrate the lag. The standard deviations of the prediction errors are found for the cross-correlation method respectively equal to 2.04 for $D = 10$, 0.41 for $D = 20$, and 1.00 for $D = 50$. Using the optimal thermal path, we find a standard deviation of the prediction errors of 0.45 for $T = 2$, 0.39 for $T = 1$, 0.33 for $T = 1/2$, and 0.49 for $T = 1/5$. Our optimal causal thermal path method thus outperforms and is much more stable than the classic cross-correlation approach.

6. Applications to economics

6.1. Revisiting the causality between the US treasury bond yield and the stock market antibubble since August 2000

In a recent paper (Zhou and Sornette, 2004), we have found evidence for the following causality in the time period from October 2000 to September 2003: stock market → Fed Reserve (Federal funds rate) → short-term yields → long-term yields (as well as a direct and instantaneous influence of the stock market on the long-term yields). These conclusions were based on (1) lagged cross-correlation analysis in running windows and (2) the dependence of the parameters of a "log-periodic power law" calibration to the yield time series at different maturities (see Sornette and Johansen (2001), Sornette and Zhou (2002) and Sornette (2003) for recent exposition of the method and synthesis of the main results on a variety of financial markets).

Let us now revisit this question by using the optimal thermal causal path method. The data consist in the S&P 500 index, the Federal funds rate (FFR), and 10 treasury bond yields spanning three years from 2000/09/09 to 2003/09/09. The optimal thermal paths $x(i)$'s of the distance matrix between the monthly returns of the S&P 500 index with each of the monthly relative variations of the eleven yields are determined for a given temperature T, giving the corresponding lag–lead times $\tau(i) = x(i)$'s as a function of present time i. Fig. 13 shows these $\tau(i)$'s for $T = 1$, where positive values correspond to the yields lagging behind or being caused by the S&P 500 index returns. The same analysis was performed also for $T = 10$, 5, 2, 1, 1/2 and 1/5, yielding a very consistent picture, confirming indeed that τ is positive for short-term yields and not significantly different from zero for long-term yields, as shown in Fig. 13. One can also note that the lag $\tau(i)$ seems to have increased with time from September 2000 to peak in the last quarter of 2003.

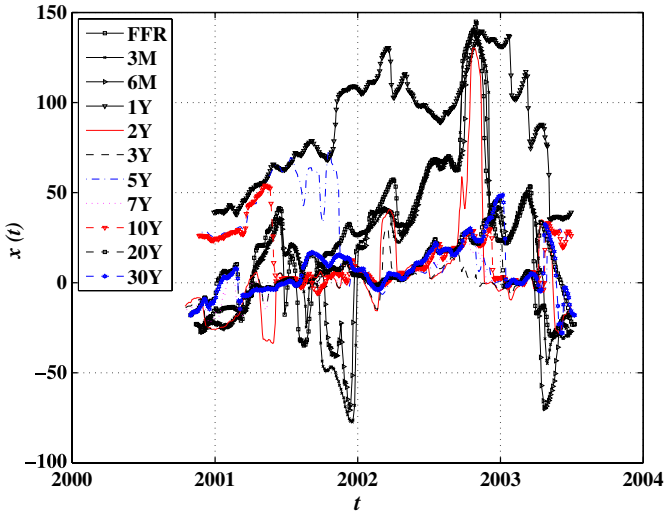

Fig. 13. Instantaneous lags between the S&P 500 index and the Federal funds rate (FFR), and between the S&P 500 index and each of 10 treasury bond yields, calculated using the optimal thermal causal path method at temperature $T = 1$ using monthly returns for the S&P 500 index and monthly relative variations for the yields. Positive lags corresponds to the yields lagging behind the S&P 500 index.

We also performed the same analysis with weakly and quarterly data of the returns and yield changes. The results (not shown) confirm the results obtained at the monthly time scale. This analysis seems to confirm the existence of a change of regime in the arrow of causality between the S&P 500 index and the Federal Funds rate: it looks as if the Fed (as well as the short term yields) started to be influenced by the stock market after a delay following the crash in 2000, waiting until mid-2001 for the causality to be revealed. The positivity of the time lag shows the causal "slaving" of the yields to the stock index. This phenomenon is consistent with the evidence previously presented by Zhou and Sornette (2004) and thus provides further evidence on the causal arrow flowing from the stock market to the treasury yields. The instantaneous lag–lead functions $\tau(t)$ provide actually much clearer signatures of the causality than our previous analysis: compare for instance with the cross-correlation coefficient shown in figure 10 of a paper by Zhou and Sornette (2004). From an economic view point, we interpret these evidences, that the FRB is causally influenced by the stock market (at least for the studied period), as an indication that the stock markets are considered as proxies of the present and are conditioning the future health of the economy, according to the FRB model of the US economy. In a related study, causality tests performed by Lamdin (2004) also confirm that stock market movements precede changes in yield spread between corporate bonds and government bonds. Abdulnasser and Manuchehr (2002) have also found that Granger causality is unidirectionally running from stock prices to effective exchange rates in Sweden.

6.2. Are there any causal relationship between inflation and gross domestic product (GDP) and inflation and unemployment in the USA?

The relationship between inflation and real economic output quantified by GDP has been discussed many times in the last several decades. Different theories have suggested

that the impact of inflation on the real economy activity could be either neutral, negative, or positive. Based on Mundell's (1963) story that higher inflation would lower real interest rates, Tobin (1965) argued that higher inflation causes a shift from money to capital investment and raise output per capita. On the contrary, Fischer (1974) suggested a negative effect, stating that higher inflation resulted in a shift from money to other assets and reduced the efficiency of transactions in the economy due to higher search costs and lower productivity. In the middle, Sidrauski (1967) proposed a neutral effect where exogenous time preference fixed the long-run real interest rate and capital intensity. These arguments are based on the rather restrictive assumption that the Phillips curve (inverse relationship between inflation and unemployment), taken in addition to be linear, is valid.

To evaluate which model characterizes better real economic systems, numerous empirical efforts have been performed. Fama (1982) applied the money demand theory and the rational expectation quality theory of money to the study of inflation in the USA and observed a negative relation during the post-1953 period. Barro (1995) used data for around 100 countries from 1960 to 1990 to assess the effects of inflation on economic output and found that an increase in average inflation led to a reduction of the growth rate of real per capita GDP, conditioned on the fact that the inflation was high. Fountas et al. (2002) used a bivariate GARCH model of inflation and output growth and found evidence that higher inflation and more inflation uncertainty lead to lower output growth in the Japanese economy. Apergis (2004) found that inflation affected causally output growth using a univariate GARCH models to a panel set for the G7 countries.

Although cross-country regressions explain that output growth often obtains a negative effect from inflation, Ericsson et al. (2001) argued that these results are not robust and demonstrated that annual time series of inflation and the log-level of output for most G7 countries are cointegrated, thus rejecting the existence of a long-run relation between output growth and inflation. A causality analysis using annual data from 1944 to 1991 in Mexico performed by Shelley and Wallace (2004) showed that it is important to separate the changes in inflation into predictable and unpredictable components whose differences respectively had a significant negative and positive effect on real GDP growth. Huh (2002) and Huh and Lee (2002) utilized a vector autoregression (VAR) model to accommodate the potentially important departure from linearity of the Phillips curve motivated by a strand of theoretical and empirical evidence in the literature suggesting nonlinearity in the output–inflation relationship. The empirical results indicated that their model captured the nonlinear features present in the data in Australia and Canada. This study implies that there might exists a nonlinear causality from inflation to economic output. It is therefore natural to use our novel method to detect possible local nonlinear causality relationship.

Our optimal thermal causal path method is applied to the GDP quarterly growth rates paired with the inflation rate updated every quarter on the one hand and with the quarterly changes of the inflation rates on the other hand, for the period from 1947 to 2003 in the USA. The GDP growth rate, the inflation rate and the inflation rate changes have been normalized by their respective standard deviations. The inflation and inflation changes are calculated from the monthly customer price index (CPI) obtained from the Fed II database (federal reserve bank). Eight different temperatures $T = 50$, 20, 10, 5, 2, 1, 1/2, and 1/5 have been investigated.

Fig. 14 shows the data used for the analysis, that is, the normalized inflation rate, its normalized quarterly change and the normalized GDP growth rate from 1947 to 2003.

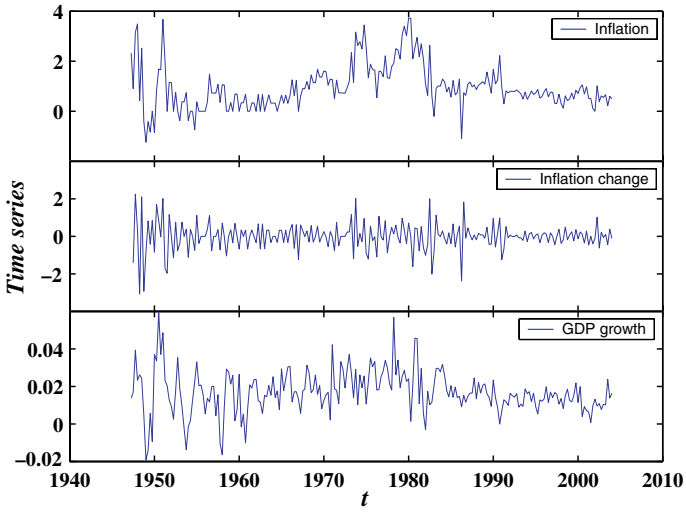

Fig. 14. Data used in our analysis, that is, the normalized inflation rate, its normalized quaterly change, and the normalized GDP growth rate from 1947 to 2003.

Fig. 15 shows the lag–lead times $\tau(t) = x(t)$'s (units in year) for the pair (inflation, GDP growth) as a function of present time t for $T = 2$ and for 19 different starting positions (and their ending counterparts) in the (t_1, t_2) plane, where positive values of $\tau(t) = x(t)$ correspond to the GDP lagging behind or being caused by inflation. This figure is representative of the information at all the investigated temperatures. Overall, we find that τ is negative in the range -2 years $\leqslant \tau \leqslant 0$ year, indicating that it is more the GPD which leads inflation than the reverse. However, this broad-brush conclusion must be toned down somewhat at a finer time resolution as two time periods can be identified in Fig. 15:

- From 1947 (and possibly earlier) to early 1980s, one can observe two clusters, one with negative -2 years $\leqslant \tau = x(t) \leqslant 0$ years implying that the GDP has a positive causal effect on future inflation, and another with positive 0 years $\leqslant \tau = x(t) \leqslant 4$ years implying that inflation has a causal effect on GDP with a longer lag.
- From the mid-1980s to the present, there is not doubt that it is GDP which has had the dominating causal impact on future inflation lagged by about 1–2 years.

In summary, our analysis suggests that the interaction between GDP and inflation is more subtle than previously discussed. Perhaps past controversies on which one causes the other one may be due to the fact that, to a certain degree, each causes the other with different time lags. Any measure of a causal relationship allowing for only one lag is bound to miss such subtle interplay. It is interesting to find that GDP impacts on future inflation with a relatively small delay of about one year while inflation has in the past influenced future GDP with a longer delay of several years.

Fig. 16 shows the lag–lead times $\tau(t) = x(t)$'s (units in year) for the pair (inflation change, GDP) as a function of present time t for $T = 2$ and for 19 different starting positions (and their ending counterparts) in the (t_1, t_2) plane, where positive values of $\tau(t) = x(t)$ correspond to the GDP lagging behind or being caused by inflation change.

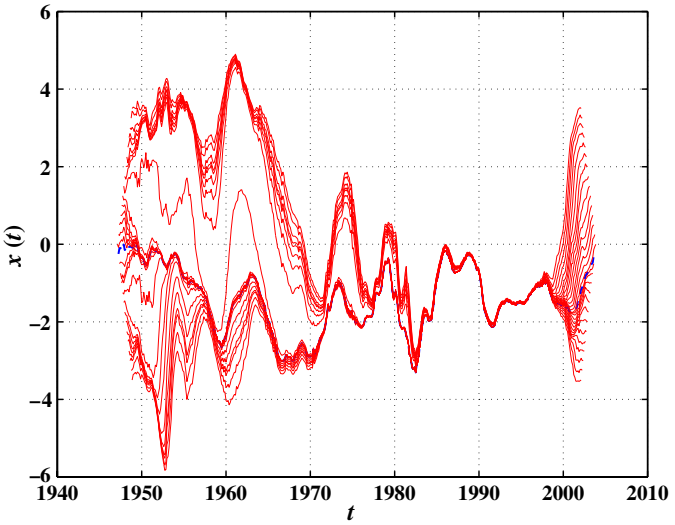

Fig. 15. (Colour online). Lag–lead times $\tau(t) = x(t)$'s (units in year) for the pair (inflation, GDP) as a function of present time t for $T = 2$ and for 19 different starting positions (and their ending counterparts) in the (t_1, t_2) plane, where positive values of $\tau(t) = x(t)$ correspond to the GDP lagging behind or being caused by inflation. The dashed blue line is the optimal path with the minimal "energy." (For interpretation of the references in color in this figure legend, the reader is referred to the Web version of this article.)

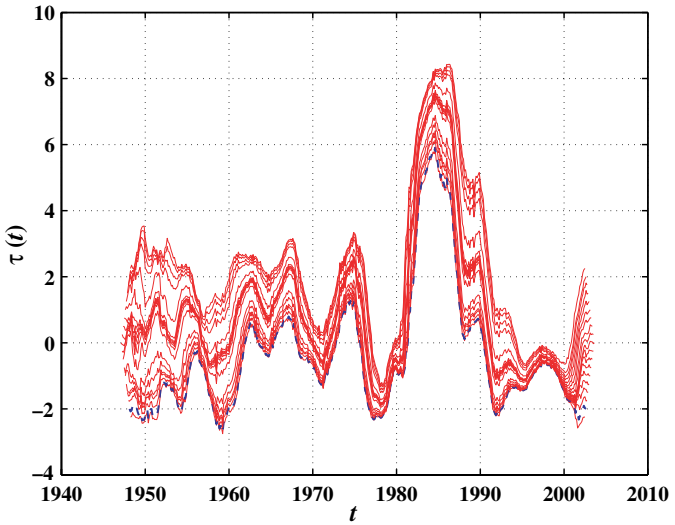

Fig. 16. (Colour online). Same as Fig. 15 for the pair (inflation change, GDP). Positive values of $\tau(t) = x(t)$ correspond to the GDP lagging behind or being caused by inflation change. The dashed blue line is the optimal path with the minimal "energy." (For interpretation of the references in color in this figure legend, the reader is referred to the Web version of this article.)

Due to the statistical fluctuations, we cannot conclude on the existence of a significant causal relationship between inflation change and GDP, except in the decade of the 1980s for

which there is strong causal effect of a change of inflation on GDP. The beginning of this decade was characterized by a strong decrease of the inflation rate from a two-digit value in 1980, following a vigorous monetary policy implemented under the Fed's chairman Paul Volker. The end of the 1970s and the better half of the 1980s were characterized by an almost stagnant GDP. In the mid-1980s, the GDP started to grow again at a strong pace. It is probably this lag between the significant reduction of inflation in the first half of the 1980s and the raise of the GDP growth that we detect here. Our analysis may help in improving our understanding in the intricate relationship between different economic variables and their impact on growth and on stability and in addressing the difficult problem of model errors, that Cogley and Sargent (2005) have argued to be the cause for the lack of significant action from the Fed in the 1970s.

Fig. 17 shows the lag–lead times $\tau(t) = x(t)$'s (units in year) for the pair (inflation, unemployment rate) as a function of present time t for $T = 2$ and for 19 different starting positions (and their ending counterparts) in the (t_1, t_2) plane, where positive values of $\tau(t) = x(t)$ correspond to the unemployment rate lagging behind or being caused by inflation. We use quarterly data from 1948 to 2004 obtained from the Fed II database (federal reserve bank). This figure is representative of the information at all the investigated temperatures.

- From 1947 (and possibly earlier) to 1970, one can observe large fluctuations with two clusters, suggesting a complex causal relationship between the two time series, similarly to the situation discussed above for the (inflation, GDP) pair.
- From 1970 to the present, there is not doubt that inflation has predated and "caused" unemployment in the sense of the optimal thermal causal path method. It is also noteworthy that the lag between unemployment and inflation has disappeared in recent

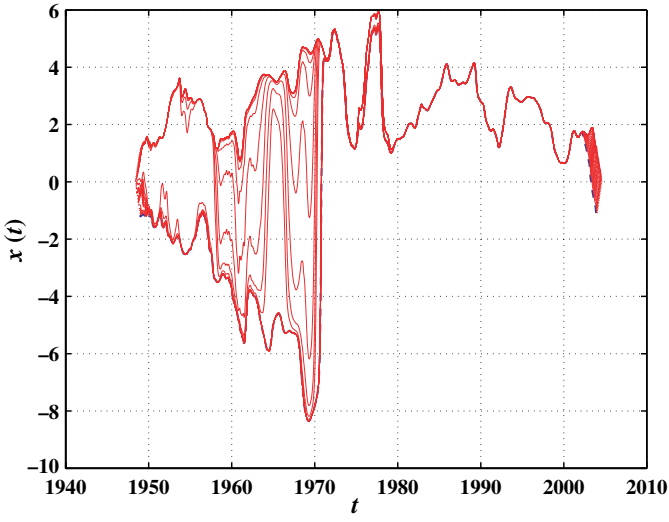

Fig. 17. (Colour online). Same as Fig. 15 for the pair (inflation, unemployment rate). Positive values of $\tau(t) = x(t)$ correspond to the unemployment lagging behind or being caused by inflation. The dashed blue line is the optimal path with the minimal "energy." (For interpretation of the references in color in this figure legend, the reader is referred to the Web version of this article.)

years. From a visual examination of Fig. 17, we surmise that what is detected is probably related to the systematic lags between inflation and employment in the four large peak pairs: (1970 for inflation; 1972 for employment), (1975 for inflation; 1976 for unemployment), (1980 for inflation; 1983 for unemployment) and (1991 for inflation; 1993 for unemployment).

One standard explanation for a causal impact of inflation on unemployment is through real wage: if inflation goes faster than the adjustment of salaries, this implies that real wages are decreasing, which favors employment according to standard economic theory, thus decreasing unemployment. Here, we find that surges of inflation "cause" increases and not decreases of unemployment. Rather than an inverse relationship between synchronous inflation and unemployment (Phillips curve), it seems that a better description of the data is a direct lagged relationship, at least in the last thirty years. The combination of increased inflation and unemployment has been known as "stagflation" and caused policymakers to abandon the notion of an exploitable Phillips curve trade-off (see for instance Lansing (2000)). Our analysis suggests a more complex multivariate description which requires taking into account inflation, inflation change, GDP, unemployment and their expectations by the agents, coupled all together through a rather complex network of lagged relationships. We leave this for a future work.

7. Concluding remarks

In summary, we have developed a novel method for the detection of causality between two time series, based on the search for a robust optimal path in a distance matrix. We have shown that the two-layer approach outperforms the multi-layer methods. Our optimal thermal causal path method determines the thermal average paths emanating from different starting lag–lead times in the distance matrix constructed from the two original time series and choose the one with minimal average mismatch ("energy"). The main advantage of our method is that it enables us to detect causality locally and is thus particularly useful when the causal relation is nonlinear and changes intermittently. An advantage of the method is that it is robust with respect to noise, i.e., it does not attribute causal relationships between two time series from patterns in the distance matrix that may arise randomly. This robustness is acquired by using the "thermal" averaging procedure which provides a compromise between optimizing the matching between the two time series and maximizing the local density of optimal paths to ensure a strong relationship.

We have applied this method to the stock market and treasury bond yields and confirmed our earlier results in Zhou and Sornette (2004) on a causal arrow of the stock markets preceding the Federal Reserve Funds adjustments as well as the yield rates at short maturities. Another application to the inflation and GDP growth rate and to unemployment have unearthed non-trivial "causal" relationships: the GDP changes lead inflation especially since the 1980s, inflation changes leads GDP only in the 1980 decade, and inflation leads unemployment rates since the 1970s.

Our approach seems to detect multiple competing causality paths with intertwined arrows of causality in which one can have inflation leading GDP with a certain lag time and GDP feeding back/leading inflation with another lag time. This suggests that the predictive skills of models with one-way causality are fundamentally limited and more elab-

orate measurements as proposed here and models with complex feedbacks are necessary to account for the multiple lagged feedback mechanisms present in the economy.

Acknowledgements

We are grateful to D. Darcet and X.-H. Wang for fruitful discussion and to N. Marwan for the permission of the use of his MATLAB programs (available at http://www.agnld.uni-potsdam.de) at the early stage of this work. This work was partly supported by NSFC (Grant No. 70501011).

References

Abdulnasser, H.-J., Manuchehr, I., 2002. On the causality between exchange rates and stock prices: A note. Bulletin of Economic Research 54, 197–203.

Apergis, N., 2004. Inflation, output growth, volatility and causality: Evidence from panel data and the G7 countries. Economics Letters 83, 185–191.

Ashley, R., Granger, C.W.J., Schmalensee, R., 1980. Advertising and aggregate consumption: An analysis of causality. Econometrica 48, 1149–1167.

Barro, R.J., 1995. Inflation and economic growth. Bank of England Quarterly Bulletin 35, 166–175.

Bouchaud, J.-P., Mézard, M., Yedidia, J., 1991. Variational theory for disordered vortex lattices. Physical Review Letters 67, 3840–3843.

Chamberlain, G., 1982. The general equivalence of Granger and Sims causality. Econometrica 50, 569–582.

Chan, K.C., Cheng, L.T.W., Lung, P.P., 2001. Implied volatility and equity returns: Impact of market microstructure and cause-effect relation. Working paper.

Chen, Y., Rangarajan, G., Feng, J., Ding, M., 2004. Analyzing multiple nonlinear time series with extended Granger causality. Physics Letters A 324, 26–35.

Cogley, T., Sargent, T.J., 2005. The conquest of US inflation: Learning and robustness to model uncertainty. Review of Economic Dynamics 8, 528–563.

Derrida, B., Vannimenus, J., 1983. Interface energy in random systems. Physical Review B 27, 4401–4411.

Derrida, B., Vannimenus, J., Pomeau, Y., 1978. Simple frustrated systems: Chains, strips and squares. Journal of Physics C 11, 4749–4765.

Eckmann, J.-P., Kamphorst, S.O., Ruelle, D., 1987. Recurrence plots of dynamical systems. Europhysics Letters 4, 973–977.

Ericsson, N.R., Irons, J.S., Tryon, R.W., 2001. Output and inflation in the long run. Journal of Applied Econometrics 16, 241–253.

Fama, E.F., 1982. Inflation, output, and money. Journal of Business 55, 201–231.

Fischer, S., 1974. Money and production function. Economic Inquiry 12, 517–533.

Fountas, S., Karanasos, M., Kim, J., 2002. Inflation and output growth uncertainty and their relationship with inflation and output growth. Economics Letters 75, 293–301.

Geweke, J., 1984. Inference and causality in economic time series models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Economics, vol. II. Elsevier Science Publisher BV, Amsterdam.

Granger, C.W.J., Jeon, Y., 1997. Measuring lag structure in forecasting models—the introduction of time distance. Discussion Paper 97-24, University of California, San Diego, October 1997.

Halpin-Healy, T., Zhang, Y.-C., 1995. Kinetic roughening phenomena, stochastic growth directed polymers and all that. Physics Reports 254, 215–415.

Huh, H.-S., 2002. Estimating asymmetric output cost of lowering inflation for Australia. Southern Economic Journal 68, 600–616.

Huh, H.-S., Lee, H.-H., 2002. Asymmetric output cost of lowering inflation: Empirical evidence for Canada. Canadian Journal of Economics 35, 218–238.

Lamdin, D.J., 2004. Corporate bond yield spreads in recent decades: an examination of trends, changes, and stock market linkages. Business Economics 39, 28–35.

Lansing, K.J., 2000. Exploring the causes of the great inflation. Federal Reserve Bank of San Francisco, 2000–21, 7 July 2000.

Marwan, N., Kurths, J., 2002. Nonlinear analysis of bivariate data with cross recurrence plots. Physics Letters A 302, 299–307.

Marwan, N., Thiel, M., Nowaczyk, N.R., 2002. Cross recurrence plot based synchronization of time series. Nonlinear Processes in Geophysics 9, 325–331.

Mundell, R., 1963. Inflation and real interest. Journal of Political Economy 71, 280–283.

Quiroga, R.Q., Kreuz, T., Grassberger, P., 2002. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. Physical Review E 66, 041904.

Shelley, G.L., Wallace, F.H., 2004. Inflation, money, and real GDP in Mexico: A causality analysis. Applied Economics Letters 11, 223–225.

Sidrauski, M., 1967. Rational choice and patterns of growth in a monetary economy. American Economic Review 57, 534–544.

Sornette, D., 2003. Why Stock Markets Crash (Critical Events in Complex Financial Systems). Princeton University Press, Princeton.

Sornette, D., Johansen, A., 2001. Significance of log-periodic precursors to financial crashes. Quantitative Finance 1, 452–471.

Sornette, D., Zhou, W.-X., 2002. The US 2000–2002 market descent: How much longer and deeper? Quantitative Finance 2, 468–481.

Strozzia, F., Zaldvarb, J.-M., Zbilutc, J.P., 2002. Application of nonlinear time series analysis techniques to high-frequency currency exchange data. Physica A 312, 520–538.

Tobin, J., 1965. Money and economic growth. Econometrica 33, 671–684.

Wang, X.H., Havlin, S., Schwartz, M., 2000. Directed polymers at finite temperatures in $1+1$ and $2+1$ dimensions. Journal of Physical Chemistry B 104, 3875–3880.

Zhou, W.-X., Sornette, D., 2004. Causal slaving of the US treasury bond yield antibubble by the stock market antibubble of August 2000. Physica A 337, 586–608.