Thick-pen transformation for time series

P. Fryzlewicz H.-S. Oh

January 1, 2011

Abstract

Traditional visualisation of time series data often consists of plotting the time series values against time and "connecting the dots". We propose an alternative, multiscale visualisation technique, motivated by the scale-space approach in computer vision. In brief, our method also "connects the dots", but uses a range of pens of varying thicknesses for this purpose. The resulting multiscale map, termed the Thick-Pen Transform (TPT) corresponds to viewing the time series from a range of distances. We formally prove that the TPT is a discriminatory statistic for two Gaussian time series with distinct correlation structures. Further, we show interesting possible applications of the TPT to measuring cross-dependence in multivariate time series, classifying time series, and testing for stationarity. In particular, we derive the asymptotic distribution of our test statistic, and argue that the test is applicable to both linear and nonlinear processes under low moment assumptions. Various other aspects of the methodology, including other possible applications, are also discussed.

1 Introduction

Traditional objectives of time series analysis are at least twofold: to obtain an understanding of certain aspects of the data, and to forecast future values, although other objectives, such as for example process control, have also been extensively studied. Naturally enough, not all of these aims are relevant to or present in every case study: for example, in time series classification problems, the focus will typically be on understanding and summarising the data rather than forecasting.

Statistical time series analysis typically tackles these aims by firstly assuming a statistical model, which can be either parametric or nonparametric, and then using suitable tools to estimate its parameters. Classical modelling, estimation and forecasting techniques for processes which are linear in their innovations and either stationary or easily transformed into such, have been extensively covered in many excellent monographs, including Brillinger (1975), Priestley (1981), Brockwell and Davis (1987) and Shumway and Stoffer (2006), while stationary but nonlinear processes, including some of the models widely used in finance, are described, for example, in Fan and Yao (2003).

Comparatively less literature exists on statistically rigorous modelling and estimation ideas for nonstationary time series, where, necessarily, some modelling effort is needed to control the degree of nonstationarity in the process before consistent statistical inference is possible. Not attempting to be exhaustive, we mention the seminal work of Priestley (1965), who introduced models based on a time-dependent transfer function in the spectral representation of nonstationary processes, and Dahlhaus (1997), who set up a framework for asymptotic considerations in nonstationary models by employing the rescaled-time principle.

In the following discussion, let X_t denote a generic univariate real-valued time series, or a real-valued univariate component of a multivariate time series (we note that the methodology proposed in this work is applicable in both univariate and multivariate time series analysis). Regardless of the nature of the problem to be solved, of the time series model to be used, and of the statistical techniques employed in the chosen model, the first step in the exploratory analysis of X_t is often the *plotting and visual inspection* of its values. While subsequent steps of the analysis, starting with the model choice, have understandably received enormous attention in the statistical literature over the years, it appears to us that the initial visualisation has been overwhelmingly skewed towards plotting the values of (t, X_t) and "connecting the dots". Useful as it undoubtedly is, the relative lack of variation in this initial step across time series literature prompts us to ask whether more can be achieved at this stage, perhaps by employing a more informative visualisation technique.

At the core of our alternative proposal for visualising time series data, which we later term the "thick-pen transform" for time series, lies the idea of looking at time series data at multiple scales, or equivalently from multiple distances. To clarify and motivate our proposal, we consider the following visual experiment. The left-hand plot of Figure 1 shows conventional "connect-the-dots" visualisation of a piecewise-stationary time series, consisting of white noise followed by a low-frequency sine wave. Moving away from the image, or alternatively looking at the image with eyes half-closed, we are likely to observe the illusory "disappearance" of the sine wave. If we believe that visibility of the time series from a distance is linked to the "volume" created by the line used to connect the dots, we can prevent this phenomenon simply by using a thicker pen to plot the second half of the data, which is done in the right-hand plot of Figure 1.

One possible lesson from this experiment is that the degree of visibility of time series data from a distance can be a helpful indicator of local structural properties of the data, as it is clearly able to discriminate between fast- and slowly-oscillating signals. Since, as argued above, it is possible to associate "visibility" from a certain distance with the "volume" created by a pen of a certain thickness used to connect the points (t, X_t), we propose to visualise a time series by plotting it using pens of various thicknesses, hoping that the resulting set of plots will provide interesting and useful information about the structure of the time series, not only in a heuristic, but also in a formal probabilistic sense.

The above discussion leads us to the first preliminary definition of our "thick-pen transform" for time series, which, without making it precise at this stage, we mean to denote a set of plots of the time series values, each performed using a pen of a different thickness. The transform is clearly multiscale, with larger thickness values bringing out coarser-scale features of the data. As a taster, we show in Figure 2 the same time series plotted with pens of thickness 5 and 30 (units are arbitrary but their ratio correctly reflects relative thickness). As the rest of this article will argue, the thick-pen transform can be useful in tasks such as nonstationarity detection, time series classification, or measuring dependence between time series, amongst others.

Time series literature is no stranger to the concept of "looking at data at multiple scales",

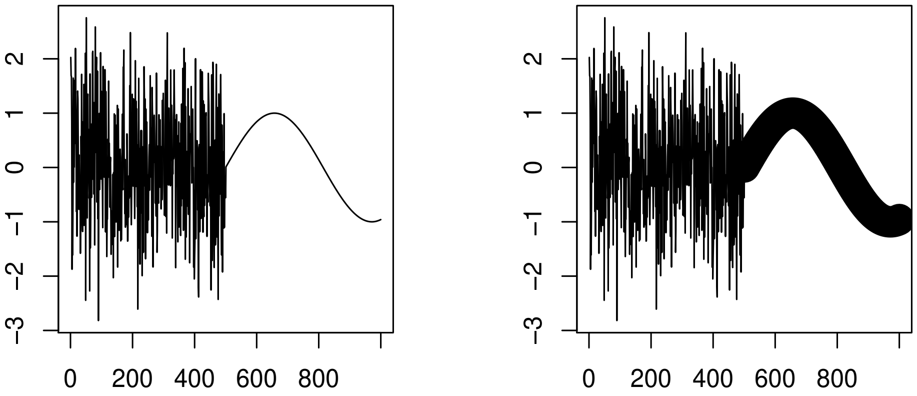


Figure 1: White noise followed by slow wave, plotted with pen of thickness 1 (left plot) and 1 followed by 20 (right plot).

and references are diverse and rather loosely connected. Wavelets, which provide linear, multiscale and local decomposition of data, have been used extensively in time series analysis; the reader is referred to the monographs of Vidakovic (1999), Percival and Walden (2000) and Nason (2008) and the references therein. Self-similarity and (multi-)fractality are often-recurring concepts in time series analysis, aiming to study parametric relationships between distributions of the process at different scales, particularly in the context of long-range dependent processes, see e.g. Doukhan et al. (2003). Other uses of multiscale methods in time series are rarer, but include, for example, Van Bellegem and von Sachs (2008), who use a multiscale technique based on iterative testing for intervals of homogeneity for the purpose of second-order structure estimation in locally stationary time series. Besides using different methodology, our approach differs from the above in that in its philosophy, it is primarily a visualisation technique (although its ultimate aim is to aid in solving some well-established time series tasks such as those listed above). Unlike wavelets, it is not linear, and is not concerned with estimation: in particular, we do not define or attempt to estimate parameters such as the long-memory parameter, fractal dimensionality or the Hurst exponent. Indeed, the thick-pen transform can also be meaningfully applied to *non*parametric time series models.

Although different in terms of its aims and methodology, our approach shares some of its spirit with SiZer (Chaudhuri and Marron (1999)), a data visualisation technique for displaying features of kernel-smoothed data as a function of location and bandwidth, simultaneously over a range of bandwidths. The similarity with thick-pen is that the latter displays the time series simultaneously over a range of pen thickness values, without attempting to choose a “best” thickness. Secondly, like SiZer, the thick-pen transform is also motivated by the “scale-space” theory in computer vision (Lindeberg (1994)) in that as argued above, large thickness values (or: larger bandwidths in SiZer) bring out features of the data best seen from a large distance. Similarly, small thickness values (smaller bandwidths

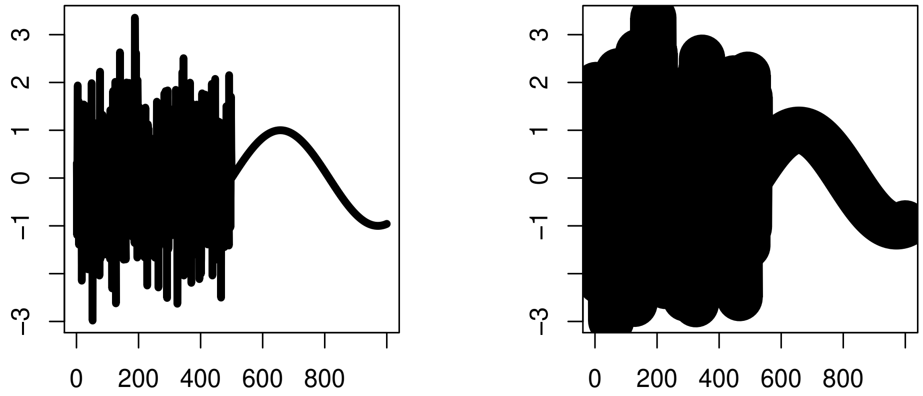


Figure 2: White noise followed by slow wave, plotted with pen of thickness 5 (left plot) and 30 (right plot).

in SiZer) permit visualisation of features observable when zooming in closer. The fundamental difference with SiZer is that the thick-pen transform (a) is designed for exploring the dependence structure of time series rather than the shape of curves, and (b) is based on nonlinear operations (as described in Section 2), as opposed to the linear smoothing used in SiZer. We note that SiZer for time series, also based on kernel smoothing, has been studied in Rondonotti et al. (2007) and Park et al. (2009).

The paper is organised as follows. Section 2 outlines the basic methodology of the thick-pen transform, and demonstrates its variation-diminishing and discrimination properties. Section 3 focuses on three possible applications of the thick-pen transform: testing for stationarity (where we propose the test, derive its asymptotic distribution under the null hypothesis and illustrate with simulated and real-data examples), classifying time series (where we propose the algorithm and apply it to a well-known geophysical dataset) and measuring dependence between time series (where we propose the measure and investigate via simulated and real-data examples). Finally, Section 4 lists possible avenues for further research.

2 Basic methodology and theory of the thick-pen transform

For the purpose of this and subsequent sections, we qualitatively define the thick-pen transform of a real-valued univariate process $(X_t)_{t=1}^n$ as follows. Let $\mathcal{T}$ denote the set of thickness parameters. For each $\tau_i \in \mathcal{T}$, $i = 1, \ldots, |\mathcal{T}|$, let $U_t^{\tau_i}$ denote the upper boundary of the area covered by a pen of thickness τ_i while connecting the points $(t, X_t)_{t=1}^n$. Similarly, let $L_t^{\tau_i}$ denote its lower boundary. The thick-pen transform $TP_{\mathcal{T}}(X_t)$ is the sequence of all pairs of boundaries, i.e.

$$TP_{\mathcal{T}}(X_t) = \{(L_t^{\tau_i}, U_t^{\tau_i})_{t=1}^n\}_{i=1,\ldots,|\mathcal{T}|}.$$

The precise mathematical form of $TP_{\mathcal{T}}(X_t)$ depends on the shape of the pen used. The examples below describe two possibly the most natural pen shapes, as well as briefly discussing other possible shapes. In all of the examples below, we let the set $\mathcal{T}$ of thickness parameters be the set of positive integers, i.e. $\mathcal{T} = \{1, 2, \ldots\}$.

Square pen. In this example, the pen is a closed square of side length $\tau \in \mathcal{T}$, positioned so that two of its sides are parallel to the time axis. For each point along the straight line connecting (t, X_t) with $(t+1, X_{t+1})$, we place the pen so that the given point is at the centre of the right-hand side of the pen. In this set-up, we have

$$U_t^\tau = \max(X_t, \ldots, X_{t+\tau}) + \frac{\tau}{2} \quad (1)$$

$$L_t^\tau = \min(X_t, \ldots, X_{t+\tau}) - \frac{\tau}{2}. \quad (2)$$

Importantly, we note the following recursive formula for computing U_t^τ thickness-by-thickness: $U_t^\tau = \max(U_t^{\tau-1}, U_{t+1}^{\tau-1}) + \frac{1}{2}$. Obviously, an analogous formula holds for L_t^τ. For extra flexibility, we admit the possibility of replacing the constants $\pm\frac{\tau}{2}$ with $\pm\gamma\frac{\tau}{2}$, where the constant γ is termed the "scaling factor" and is to be chosen by the analyst. We do not dwell on the choice of γ in this work, and set it by default to one in the examples below, unless mentioned otherwise.

Round pen. Let the pen be a closed circle with diameter $\tau \in \mathcal{T}$, positioned so that for each point along the straight line connecting (t, X_t) with $(t+1, X_{t+1})$, the pen is centred at the given point. Denoting by $\mathbb{Z}$ the set of integers, we have

$$U_t^\tau = \max_{k \in [-|\tau|/2, |\tau|/2] \cap \mathbb{Z}} \{X_{t+k} + \sqrt{\tau^2/4 - k^2}\} \quad (3)$$

$$L_t^\tau = \min_{k \in [-|\tau|/2, |\tau|/2] \cap \mathbb{Z}} \{X_{t+k} - \sqrt{\tau^2/4 - k^2}\}. \quad (4)$$

As with the square pen, for extra flexibility, the additive terms $\pm\sqrt{\tau^2/4 - k^2}$ could be replaced by $\pm\gamma\sqrt{\tau^2/4 - k^2}$. In what follows, the scaling factor γ is always set to one unless mentioned otherwise.

Other possible pen shapes. We note the following interesting connection between the above formulae (1) – (4) and kernel smoothing. Replacing the addition and subtraction by multiplication, and the "max" and "min" operators by the summation operator, we obtain, up to a multiplicative factor proportional to τ^2, kernel-smoothed versions of X_t using the one-sided uniform kernel (in the square-pen case) and the two-sided circular kernel (in the round-pen case). Obvious generalisations of the square and round pens could be obtained by employing other kernel shapes.

More on the duality between kernel smoothing and the thick pen. Despite the duality between kernel smoothing and the thick pen (as described above), these two are entirely different operations and they serve different purposes. Computed using a single bandwidth/thickness, kernel smoothing of X_t produces one linear output sequence (weighted local means of X_t), whereas the thick-pen transform produces two nonlinear output sequences L_t^τ and U_t^τ. Unlike the kernel-smoothed version of X_t (which can serve to estimate the trend but not the local dependence structure of X_t), the thick-pen transform of X_t provides useful information on both the trend and the local dependence structure, which we explain later in this section, and in particular in Theorem 2.1. By contrast, in order to use kernel smoothing to estimate or make inference on the local dependence structure of X_t, one would have to kernel-smooth not X_t, but other local statistics of X_t (e.g. the sequence $X_t X_{t+1}$ to estimate the local covariance at lag one; or local periodograms to estimate the local spectral density).

We note that for a fixed value of τ (i.e. focusing on a "single scale" of our thick-pen transform), and disregarding the additive constant $\tau/2$, formulae (1), (2) are reminiscent of the running maximum and minimum filters, used in signal and image processing for tasks such as edge detection (Lee et al. (1987), Douglas (1996), Vemis et al. (1995)), texture description (Werman and Peleg (1985)), character extraction (Ye et al. (2001)), indexing of dynamic time warping (Keogh and Ratanamahatana (2005)), or the suppression of over- and under-shoot (Cho and Bae (2006)). However, the fundamental difference with our approach is that the thick-pen transform puts these filters in a multiscale context (which, incidentally, is precisely what SiZer does to kernel smoothing) by considering a range of thickness parameters τ simultaneously. This is done with the initial aim of visualising time series data, and the eventual aim of solving some classical problems in statistical time series analysis mentioned earlier. Also, in contrast to the above heuristic approaches, ours is more rigorous in that we formally prove, later in this section, that the thick-pen transform discriminates between two time series with different correlation structures (see Theorem 2.1). One essential ingredient of this result is the multiscale context in which the thick-pen transform operates.

We now consider a toy example which shows one possible way of visualising the thick-pen transform. The top plot in Figure 3 shows the conventional visualisation of a time series consisting of Gaussian white noise in the first half, and the Gaussian AR(1) process with parameter 0.9 in the second half. Both processes have the same variance. The middle plot shows the thick-pen transform using the round pen, with pen thicknesses ranging from 10 to 100 in multiples of 10, where each trajectory is superimposed on the next thicker one. The bottom plot shows the same object plotted with a different colour pattern. The reader is invited to think of this object as the "thick-pen map" of the given process. While visual inspection of the top plot reveals obvious visual difference between the structure of the process in the first and second half, a look at the thick-pen map reveals the multiscale nature of this difference. For instance, considering the bottom plot at the "thinnest" scale, we observe the frequent brief but deep incursions of non-white colours into the white area in the second, more structured half. On the other hand, at the "thickest" scales, by looking at the darkest colours, the incursions in the second half are still present but are much longer-lasting and more shallow: consider, for example, the incursions around time $t = 350$ (lower boundary) or time $t = 400$ (upper boundary). We note that none of these features are obvious from the visual inspection of the top plot, but could serve as potentially interesting "markers" for discriminating between the two halves or detecting the change-point.

However, it is important to bear in mind that the analyst does need to rely on her or his vision to take advantage of the thick-pen transform of the data. In what follows, we describe certain summary statistics, involving the sequences U_t^τ and L_t^τ, which can be viewed as automatic "scanners" for reading off certain properties of the thick-pen map.

We first note that in the thick-pen transform as described above, one thickness value τ generates two sequences: U_t^τ and L_t^τ. In some time series problems, for example nonstationarity detection (as described later), it might be more convenient to use a single summary

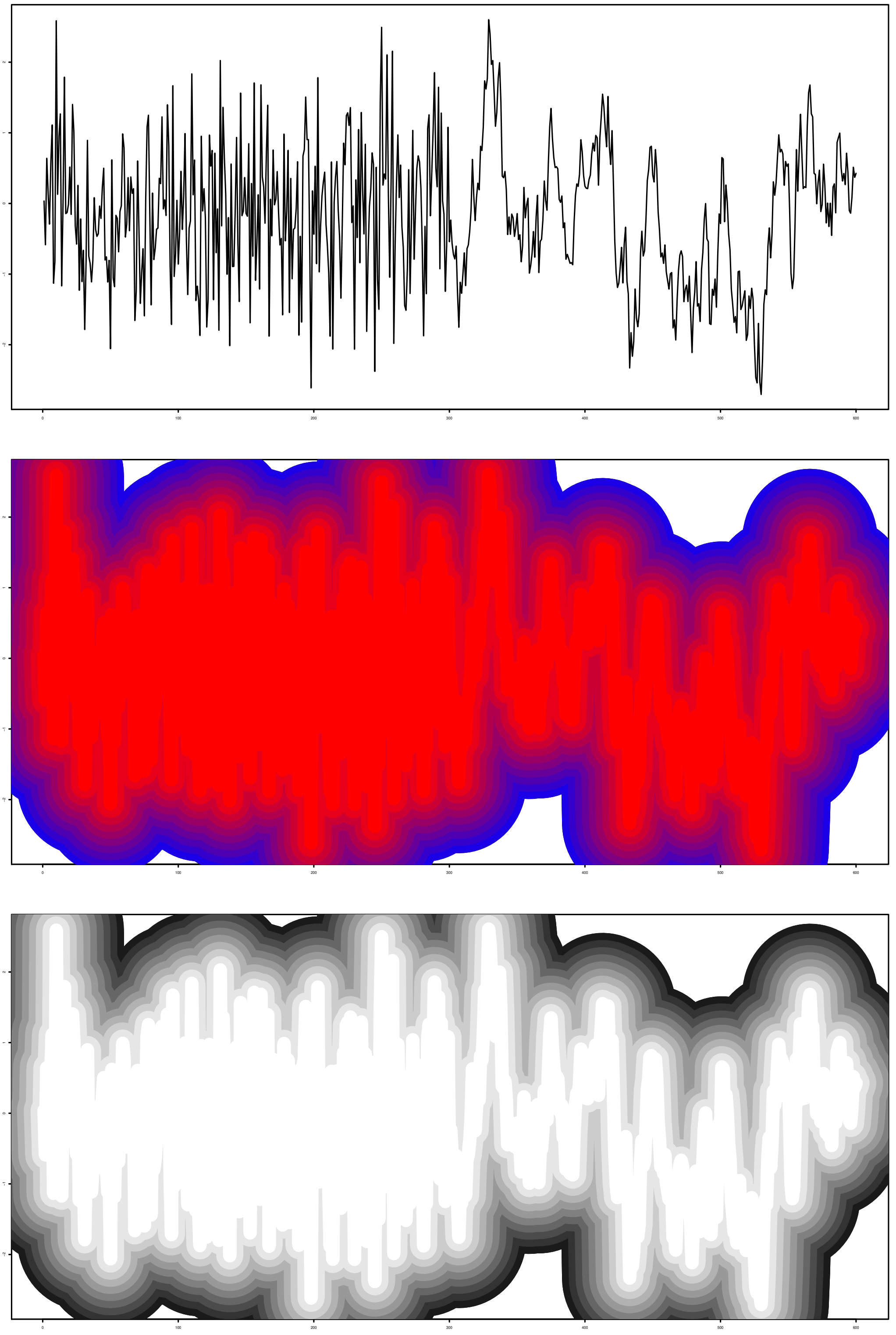


Figure 3: White noise followed by AR(1) process with parameter 0.9, top: conventional visualisation; middle and bottom plots: thick-pen visualisation.

sequence, instead of a pair. (By way of analogy, we mention that, for instance, a single scale in a wavelet transform is typically also represented by a single sequence: that of the wavelet coefficient values at a given scale.) Below, we identify and describe some particular summary statistics involving the sequences U_t^τ and L_t^τ which will be used later in the paper for the purposes of nonstationarity detection as well as measuring dependence between time series. We divide them into two classes: "basic" and "derivative" summary statistics.

Basic summary statistics. We define the following basic summary statistics involving U_t^τ and L_t^τ.

- Volume of the pen, defined as $V_t^\tau = U_t^\tau - L_t^\tau$.
- Mean of the pen, defined as $M_t^\tau = \frac{1}{2}\{U_t^\tau + L_t^\tau\}$.

Derivative summary statistics. We define the following derivative summary statistics involving U_t^τ and L_t^τ.

- Rate of change of volume with respect to time t, defined as $\frac{\Delta V_t^\tau}{\Delta t} = V_t^\tau - V_{t-1}^\tau$.
- Rate of change of volume with respect to thickness τ, defined as $\frac{\Delta V_t^\tau}{\Delta \tau} = V_t^\tau - V_t^{\tau-1}$.
- Rate of change of mean with respect to time t, defined as $\frac{\Delta M_t^\tau}{\Delta t} = M_t^\tau - M_{t-1}^\tau$.
- Rate of change of mean with respect to thickness τ, defined as $\frac{\Delta M_t^\tau}{\Delta \tau} = M_t^\tau - M_t^{\tau-1}$.

V_t^τ measures the local width of the area created by the pen of thickness τ, and M_t^τ measures its local mean level. The derivative summary statistics measure how those two quantities change with respect to t or τ. Depending on the nature of the time series and problem at hand, more complex summary statistics are also possible. One such example is described in Section 4. A few remarks are in order.

Complementarity of V_t^τ and M_t^τ. We note that V_t^τ and M_t^τ are constructed by applying the complementary operations of subtraction and addition (respectively) to U_t^τ and L_t^τ, and thus can be viewed as "symmetric" quantities. To recover U_t^τ and L_t^τ from V_t^τ and M_t^τ we use the very similar operations

$$\begin{aligned} U_t^\tau &= M_t^\tau + \frac{1}{2}V_t^\tau \\ L_t^\tau &= M_t^\tau - \frac{1}{2}V_t^\tau. \end{aligned}$$

Link between V_t^τ and tube formula. Out of the basic and derivative summary statistics described above, V_t^τ deserves special attention because of the fact that statistical literature has previously explored the concept of "the volume of a covering of data", albeit in other contexts. Weyl (1939) derived the famous "tube formula" for calculating the volume of a tube surrounding a smooth manifold embedded in a k-dimensional unit sphere, for a finite k, extending a previous result by Hotelling (1939). These results were more recently discussed, extended and applied in various statistical contexts involving smooth (but not always deterministic) curves or surfaces by, amongst others, Knowles and Siegmund (1988), Johansen and Johnstone (1990) and Sun (1993). We are unaware of any applications of tube formulae in classical time series, where sample paths are often intrinsically non-smooth.

Link between V_t^τ and estimation of fractal properties. On the other hand, in estimating the Hurst exponent or the fractal dimension of stochastic processes, two techniques involving statistics related to our V_t^τ are the Rescaled Range Analysis (Hurst (1951)) and the "box-counting" method, whose statistical properties in estimating the fractal dimension of a stationary continuous-time Gaussian process were studied in Hall and Wood (1993). In contrast to the above methodologies, we emphasise here again that our statistic V_t^τ is not an estimator of any quantity; it is merely one possible summary statistic for the multiscale thick-pen transform of any, not necessarily stationary time series, at location t and thickness ("scale") τ. If taking the thick-pen transform corresponds to viewing the time series from a range of distances, V_t^τ (or any other summary statistic derived from the thick-pen transform) can be thought of as a particular form of "eye" or "scanner" used to record certain properties of the given time series, at time t, viewed from the given range of distances. We also emphasise that the above-mentioned methodologies, Rescaled Range and box counting, have traditionally been used for stationary, possibly long-memory processes. In constract, our statistic V_t^τ is being put to work mainly in the context of nonstationary processes. For example, the following Section 3.1 demonstrates its usefulness in detecting nonstationarities in time series. Finally, we mention that while estimators of fractal properties of time series are typically considered in the limit as their "scale" parameter approaches either zero or infinity, our statistic V_t^τ is also meaningful and informative (both theoretically and empirically) for a finite number of thicknesses τ, for example when the task at hand is to detect nonstationarities in time series.

V_t^τ as a measure of self-overlap. We also mention an interesting interpretation of V_t^τ as measure of the extent to which the pen "overlaps with itself" whilst plotting a given time series. For each fixed time t, the quantity $\sum_{i=1}^t V_i^\tau$ is a total measure of the area created by the pen at thickness τ up to time t. Thus, $V_t^\tau = \sum_{i=1}^t V_i^\tau - \sum_{i=1}^{t-1} V_i^\tau$ measures how much new area appeared at time t. Subtracting V_t^τ from the area of the "tip" of the pen (for example, τ^2 for the square pen), we obtain a measure of how much the pen overlaps with itself whilst plotting the data.

Smoothness of U_t^τ, L_t^τ. Recalling the analogy between pens and kernels described earlier, we mention that pens corresponding to differentiable kernels (e.g. the round pen) will lead to smoother sequences U_t^τ, L_t^τ than pens corresponding to non-differentiable kernels, such as the square pen.

We are now ready to prove two theoretical properties of the thick-pen transform, both of which focus on the square pen (which we have found the least challenging to analyse theoretically). The first one, in analogy to SiZer, establishes a "variation diminishing property" of the thick-pen transform. The second one establishes the fundamental property that the thick-pen transform is *discriminative* for Gaussian time series, i.e. that the thick-pen transforms of two different Gaussian time series are distributed differently.

Proposition 2.1 Variation diminishing property. *Let $(X_t)_{t=1}^n$ be a time series, let $\mathcal{T} = \{1, 2, \ldots\}$ and let $TP_{\mathcal{T}}(X_t)$ be the thick-pen transform of X_t. For any sequence $(f_t)_{t=1}^m$, we define its total variation functional by*

$$\|f\|_{TV} = \sum_{t=2}^{m} |f_t - f_{t-1}|.$$

Both $\|U^\tau\|_{TV}$ and $\|L^\tau\|_{TV}$ are non-increasing functions of τ.

At this point, we mention an interesting link between the thick-pen transform and SiZer. Lindeberg (1994), Section 3.5.1 describes the *variation dimishing property* of kernels in linear smoothing (as used in SiZer), in the sense that, roughly speaking, the number of sign changes in the estimated function decreases as a function of the bandwidth if and only if the kernel used is either Gaussian or one-sided exponential. In the case of the thick-pen transform, note first that if X_t were a sequence of -1's and 1's, then $\|U^\tau\|_{TV}$ and $\|L^\tau\|_{TV}$ would simply count, up to a multiplicative factor, the number of sign changes in U_t^τ and L_t^τ. The above proposition implies that in this special case, the thick-pen transform enjoys a similar variation diminishing property, or, in other words, it demonstrates a "smoothing" property of the (nonlinear) max/min filter.

The discrimination property follows. Before we formulate the result, we introduce the following mild technical assumption, and explain it underneath.

Assumption 2.1 *For a given fixed lag $\tau > 0$, a process X_t satisfies*

$$\exists\, \lambda_0, \delta \in [0,1) \quad \forall \lambda > \lambda_0 \quad \forall\, t$$

$$P\left(\bigcup_{t \le i,j \le t+\tau; \{i,j\} \ne \{t, t+\tau\}} |X_i - X_j| > |X_t - X_{t+\tau}| \;\middle|\; |X_t - X_{t+\tau}| > \lambda \right) \le \delta.$$

More descriptively, the above assumption means that uniformly over all time locations t, conditioning on the fact that the absolute difference $|X_t - X_{t+\tau}|$ is "large", it is *not entirely unlikely* that it achieves the maximum absolute difference amongst all $|X_i - X_j|$ for i, j lying between $t, t+\tau$. Since there is no requirement on δ other than that it should be less than one, the above assumption should be viewed as a mild one. An extra discussion of this assumption appears underneath the proof of Theorem 2.1 in the Appendix.

We are now ready to state the discrimination result.

Theorem 2.1 *Let X_t, Y_t be two zero-mean Gaussian time series such that for some $s < t$, the distribution of $X_s - X_t$ is not the same as the distribution of $Y_s - Y_t$, and let both X_t and Y_t satisfy Assumption 2.1 with $\tau = t - s$. Let $TP_{\mathcal{T}}(X_t)$, $TP_{\mathcal{T}}(Y_t)$ be the thick-pen transforms of X_t, Y_t respectively, both with the square pen where the set $\mathcal{T}$ of thickness parameters is $\mathcal{T} = \{1, 2, \ldots\}$, and let $V_t^\tau(X), V_t^\tau(Y)$ be the corresponding volumes. Then, $TP_{\mathcal{T}}(X_t)$ and $TP_{\mathcal{T}}(Y_t)$ follow different probability distributions in the sense that the trivariate random vectors $(V_s^{\tau-1}(X), V_{s+1}^{\tau-1}(X), V_s^\tau(X))$ and $(V_s^{\tau-1}(Y), V_{s+1}^{\tau-1}(Y), V_s^\tau(Y))$ are distributed differently.*

Theorem 2.1, although of a purely "existential" nature, gives us hope that the thick-pen transform can act as a successful discriminator for time series, as it uniquely determines their distribution (clearly, identically distributed time series yield identically distributed thick-pen transforms, and by the above theorem, differently distributed time series lead to differently distributed thick-pen transforms). Furthermore, the above theorem also gives us a hint as to the range of thickness parameters in which to look for the distributional differences: roughly and approximately speaking, if the autocorrelation structures of the two time series differ at lag τ, differences in the distributions of the thick-pen transforms can also be expected "around" thickness τ. To the best of our knowledge, the proof technique for Theorem 2.1 is new. With some effort, it can be extended to certain non-Gaussian distributions; we leave this for future work. Note that since $V_t^0 = 0$, in the case $\tau = 1$, the statement of the theorem reduces to "$V_s^1(X)$ and $V_s^1(Y)$ are distributed differently".

3 Possible uses of the thick-pen transform

3.1 Testing for nonstationarity via the thick-pen transform

In this section, we demonstrate, via both theoretical and empirical arguments, that the thick-pen transform can serve as an efficient tool for detecting nonstationarities in time series. The logic we use for this purpose is as follows. Let K_t^τ denote any generic basic, derivative or other summary statistic involving the sequences U_t^τ and L_t^τ which the analyst believes is likely to capture the nature of possible nonstationarity in the underlying process X_t, if there is any. As a general guidance, taking $K_t^\tau = V_t^\tau$ (volume) appears to be a good choice when analysing possible changes in the dependence structure of the process, whereas taking $K_t^\tau = M_t^\tau$ is a good idea when analysing changes in the trend. Our simulation study in the latter part of this section offers more specific practical advice on the choice of K_t^τ.

Let the values of τ be positive integers. If X_t is stationary, then the vector-valued time series $(K_t^{\tau_1}, \dots, K_t^{\tau_2})_t$, where the choice of τ_1, τ_2 will be discussed later, is also stationary. For the time being, fix $\tau \in \{\tau_1, \tau_1 + 1, \dots, \tau_2\}$. The following result will form the basis of our test for stationarity.

Theorem 3.1 *Let $\{X_t\}_{t=1}^n$ be a stationary process satisfying $\mathbb{E}|X_t|^r < \infty$ for some $r > 2$. In addition let X_t be α-mixing with the mixing coefficients α_m satisfying $\alpha_m = O(m^{-s})$ for some $s > \frac{r}{r-2}$. Let $TP_\mathcal{T}(X_t)$ be the thick-pen transform of X_t using an arbitrary pen but such that both U_t^τ and L_t^τ are functions of $X_{t-C\tau}, \dots, X_{t+C\tau}$ only, for some $C > 0$. Further let the summary sequence K_t^τ be such that for each fixed τ, we have $n^{-1}\mathrm{Var}(\sum_{t=1}^n K_t^\tau) \to \sigma_\tau^2 < \infty$, and $|K_t^\tau| \le A + B|\max(X_{t-C\tau}, \dots, X_{t+C\tau})|$ for some constants $A, B > 0$, possibly depending on τ. Under these conditions, the following functional central limit results hold for each fixed τ.*

(i) Let $u \in [0,1]$. We have

$$Y_n^\tau(u) := \frac{1}{\sigma_\tau\sqrt{n}} \sum_{t=1}^{\lceil nu \rceil} K_t^\tau - \mathbb{E}(K_t^\tau) \xrightarrow{d} B_u, \qquad (5)$$

where B_u is the standard Wiener process on $[0,1]$.

(ii) Further, we have

$$Z_n^\tau(u) := Y_n^\tau(u) - \frac{\lceil nu \rceil}{n} Y_n^\tau(1) \xrightarrow{d} B_u^0,$$

where B_u^0 is the standard Brownian bridge process on $[0,1]$.

The cumulative distribution function of the range of a Brownian bridge is well-known (Kennedy (1976)) and is given by

$$F_{B^0}(x) = 1 + 2\sum_{k=1}^{\infty}(1 - 4k^2x^2)\exp(-2k^2x^2).$$

The above result naturally suggests the following procedure for testing stationarity:

1. Fix the thinnest scale τ_1 and the thickest scale τ_2. The simulation study in the latter part of this section offers some insight into suitable choices of these parameters.

2. Set the desired significance level α. With the Bonferroni correction, this becomes $\alpha_B = \alpha/(\tau_2 - \tau_1 + 1)$.

3. For each $\tau \in \{\tau_1, \tau_1 + 1, \dots, \tau_2\}$, estimate $\mathbb{E}(K_t^\tau)$ (which is independent of t under the null hypothesis of stationarity) by $\frac{1}{n}\sum_{t=1}^{n} K_t^\tau$.

4. Estimate σ_τ^2 as $\hat{s}_0^\tau + 2\sum_{k=1}^{M} \hat{s}_k^\tau$, where $\{\hat{s}_k^\tau\}_k$ is the sample autocovariance sequence of K_t^τ. The simulation study below discusses the choice of M.

5. Using the estimated versions of $\mathbb{E}(K_t^\tau)$ and σ_τ, form the Brownian bridge processes $Z_n^\tau(u)$, and calculate their ranges $R^\tau = \max_u Z_n^\tau(u) - \min_u Z_n^\tau(u)$.

6. If $F_{B^0}(\max_\tau R^\tau) > 1 - \alpha_B$, then reject the hypothesis of stationarity; otherwise accept.

A few remarks are in order.

Low moment requirements. We note the low moment requirements of the proposed test. Indeed, we only require that $\mathbb{E}|X_t|^r < \infty$ for some $r > 2$. This is because, obviously but interestingly, moments of maxima of random variables exist if an only if the corresponding moments of the variables themselves exist. By contrast, a variety of nonstationarity tests proposed in literature, see e.g. Neumann and von Sachs (2000) and the references therein, are based on local second-order statistics of the process, e.g. local periodograms. If K_t^τ were to be such a local quadratic form of X_t, we would automatically need to require the existence of $\mathbb{E}|X_t|^{2r}$ for some $r > 2$.

Difference with Rescaled Range Analysis. In the earlier part of the paper, we mention differences between our methodology in the case where our summary statistic of interest is V_t^τ, and the Rescaled Range Analysis. Another difference emerges now. Taking $K_t^\tau = V_t^\tau$, the operations taken in (5) correspond to taking the local volume (where the 'locality' is determined by the thickness parameter τ) and rescaling by the *global* quantity σ_τ. This is in contrast to the Rescaled Range Analysis, which, in its local version, would take the local range and rescale it by the *local* standard deviation. However, this would not be of use in detecting nonstationarities: we invite the reader to think of a stationary process mutliplied by a time-varying standard deviation function, for which such a local Rescaled Range Analysis would annihilate the effect of the time-varying standard deviation (due to the fact that for such a process, the local range is roughly proportional to the local standard deviation) and thus make it impossible to detect the nonstationarity of the process.

Suitability for both linear and nonlinear time series. A unique feature of our test is that it is equally valid for linear and nonlinear processes, provided they satisfy the requirements of Theorem 3.1. This is in contrast to, for example, the variety of tests based on local second-order statistics of the process, which, by construction, are not applicable to certain well-known nonlinear time series models such as (G)ARCH, which are simply "white noise" as far as their second-order properties are concerned.

Exact p-value available when $\tau_1 = \tau_2$. When $\tau_1 = \tau_2$, the test is performed using a single thickness τ only, and there is no need to perform the Bonferroni correction. In this case, it is possible to specify the exact p-value, which equals $1 - F_{B^0}(R^\tau)$.

In the simulation study that follows, we test using a single thickness τ only. This is done to "separate out" the performance of the test from the effect of the Bonferroni correction, which has a chance of spoiling things when the dependence across τ amongst $K_t^{\tau_1}, \ldots, K_t^{\tau_2}$ is too high. However, later in this section, we propose a way of alleviating this issue by constructing K_t^τ in such a way that this dependence hopefully remains low. Also, by using a single thickness only, we are able to report the exact p-value.

Our simulation study is in five parts. In part one, our test process is white noise with standard deviation changing abruptly halfway through. In parts two and three, respectively, the test processes are AR(1) and ARCH(1), for which the autoregressive parameters change halfway through, but the variance remains constant. In part four, we evaluate the performance of our test for some challenging nonlinear processes from Davis et al. (2008). In part five, we use an interesting example of an "on-off" process to exhibit the multiscale aspect of our test and explain why its performance naturally depends on the thickness used in the test. Since we test for structure rather than trend, our summary statistic of choice is $K_t^\tau = V_t^\tau$. We use the square pen, and the maximum autocovariance lag M from step 4 of the testing procedure above is set equal to $\max(\tau, \log n)$ (the τ in this expression "takes care" of the dependence arising from the construction of K_t^τ, whereas the $\log n$ is responsible for picking up the most significant autocovariances arising from the autocovariance structure of the original process X_t). We use thickness $\tau = 1$ in parts one to four (the reason for this is given in part five). In part five, we also test using a single thickness at a time, but we investigate how the performance of the test varies with the thickness used. For testing using multiple thicknesses, our recommendations for the choice of τ_1, τ_2 are given later in this section.

(a) The model is $X_t = \sigma_t \varepsilon_t$, where $\{\varepsilon_t\}_{t=1}^{500}$ are i.i.d. $N(0,1)$, and σ_t changes from 1 for $t = 1, \ldots, 250$ to σ for $t = 251, \ldots, 500$. Average p-values based on 100 simulations are shown in Table 1. We note that even in the case $\sigma = 1.35$, for which the p-value is below 10%, it is still extremely difficult to detect the nonstationarity by simple visual inspection, so our test genuinely appears to help in this case.

(b) The model is $X_t = \{1 - a_t^2\}^{1/2} Y_t$, where $Y_t = a_t Y_{t-1} + \varepsilon_t$ with $\{\varepsilon_t\}_{t=1}^{500}$ as above, and a_t changes at $t = 251$ from $a^{(1)}$ to $a^{(2)}$. Note that X_t has a constant variance throughout, so the only change is in the autoregressive parameter. We report the results in Table 2. It is remarkable that for a pair of parameters $(a^{(1)}, a^{(2)}) = (a, a + \delta)$ for a fixed δ, it is becoming easier for our test to detect the nonstationarity as a increases. This, we believe, has an appealing physical interpretation: indeed, the quantity V_t^τ can be regarded as an "eye" or "scanner" used to view the time series from a distance. Thus, it should not suprise us that the test is more sensitive for higher values of a: after all, the human eye also tends to be better at detecting the change for larger values of a, since lower values of a imply noisier appearance of the data.

(c) The model is $X_t = \{1 - a_t\}^{1/2} Y_t$, where $Y_t = \sigma_t \varepsilon_t$ with $\{\varepsilon_t\}_{t=1}^{1000}$ as above, and $\sigma_t^2 = 1 + a_t Y_{t-1}^2$. (The sample paths now have length 1000 in preparation for the next example.) The parameter a_t changes at $t = 501$ from $a^{(1)}$ to $a^{(2)}$. Again, X_t has a constant variance throughout, and the only change is in the autoregressive parameter. Obviously, the p-values are now not as impressive as in model (b), which is not surprising, as nonstationarity detection for ARCH is known to be a harder problem than for Gaussian AR processes. However, the next simulation example reassures us that this is still not a bad result and in fact outperforms what we have been able to establish is the state of the art, which is all the more interesting given the fact that our method is in no way specifically designed for ARCH or even nonlinear processes. The example follows.

(d) Davis et al. (2008) consider, amongst others, three challenging examples (from the point of view of breakpoint detection) of the GARCH(1,1) model $X_t = \sigma_t \varepsilon_t$, with $\{\varepsilon_t\}_{t=1}^{1000}$ as above, and $\sigma_t^2 = a_t^{(0)} + a_t^{(1)} X_{t-1}^2 + b_t^{(1)} \sigma_{t-1}^2$, where the triple of parameters $(a_t^{(0)}, a_t^{(1)}, b_t^{(1)})$ changes, at time $t = 501$, as follows:

(i) $(0.4, 0.1, 0.5) \rightarrow (0.4, 0.1, 0.6)$

(ii) $(0.1, 0.1, 0.8) \rightarrow (0.1, 0.1, 0.7)$

(iii) $(0.4, 0.1, 0.5) \rightarrow (0.5, 0.1, 0.5)$

We ran our nonstationarity test on these three models, and Table 4 shows the percentage of times (over 1000 simulations) that our test failed to detect their nonstationarity at 5% significance level. Comparing this to the percentage of times Davis et al. (2008) [Auto-Seg] and the competing method from Andreou and Ghysels (2002) [AG] failed to detect any breakpoints (i.e. classified the series as stationary), we can see that our method (which has not been particularly fine-tuned for GARCH or even nonlinear processes) outperforms these two state of the art techniques on these challenging examples. To check calibration, we also ran our test on the stationary examples considered in Davis et al. (2008), in which the triples of parameters were

(iv) $(0.4, 0.1, 0.5)$

(v) $(0.1, 0.1, 0.8)$

The results are in Table 5. The empirical results for our test are not far off the theoretical value of 5%.

(e) In this example, we demonstrate that results of the thick-pen test for stationarity can, understandably and interpretably, depend on the thickness at which the given process is being considered. This should not be viewed as a weakness of the thick-pen methodology, but rather as a natural implication of the fact that certain processes do not “appear” stationary when inspected at certain scales (= thicknesses) but do so at others.

For example, we saw in parts (a)–(d) that the thinnest pen (with $\tau = 1$) was “sufficient” in testing for stationarity of classical time series models such as AR(1), ARCH(1) and GARCH(1,1). This is because when the values of their parameters change, it follows that the stochastic relationship between *two consecutive* variables comprising the series also changes, which is why it suffices to consider these processes at the smallest thickness value to detect the parameter change.

However, in this example, we consider an altogether different stochastic process, which switches, in a Markovian way, between being constant and equal to zero, and being

Gaussian white noise with mean zero, variance one. (This could serve e.g. as a *very* rudimentary "cartoon" of a speech signal.) Consider the following cases:

Case A. The switching probability, at any particular time t, is $\frac{1}{20}$.

Case B. The switching probability, at any particular time t, is $\frac{1}{50}$.

Case C. The switching probability, at any particular time t, is $\frac{1}{20}$, but overall the time series is non-stationary in the sense that is it always equal to zero for the last 30% of the time (so it is as in Case A times zero for the last 30% of the time).

Simulated sample paths for each of the cases: A, B and C, are in Figure 4.

We perform exactly the same test on cases A, B, C as in parts (a)–(d), the only difference being that we now vary thickness values from 1 to 14. Figure 5 shows the average p-value of the test as a function of thickness, averaged over 100 simulated sample paths from models A (red), B (blue), C (green), each for a sample path of length 1000.

As expected, the p-value increases with thickness, for models A and B. This is to be interpreted as saying that the series begins to "look" stationary for higher thickness values, as it is for those values that we are beginning to "bridge the gaps" between the white noise parts and therefore obtain a more complete, broad scale picture of the process. For lower thickness values, the scale at which we view the process is "too fine" given the nature of the process: at those thicknesses, the transitions between the white noise and zero parts are mistakenly interpreted as breakpoints.

In particular, we note that p-values for model A begin to exceed 10% for thicknesses 3 and above. A similar threshold thickness for model B is 8. It is unsurprising that the blue curve lies underneath the red curve: model B displays fewer changes, which are in addition further apart, and therefore are more likely to be misinterpreted as breakpoints (or in other words in takes a higher thickness value, or a further zoom-out, to appreciate the stationarity of the process). On the other hand, process C is, on average, correctly interpreted as non-stationary for all thicknesses in the considered range.

Finally, we run our stationarity test (with exactly the same parameters as above and with $\tau = 1$) on the time series of logged and differenced daily closing values of the S&P 500 index, for the 500 trading days (approximately 2 years) ending 25 November 2009. Thus, the considered period overlaps with the recent financial crisis. The data are plotted in Figure 6. Our choice of the thickness parameter is motivated by the good performance of our test with $\tau = 1$ for (G)ARCH models, as illustrated above. Since our test is equally suitable for linear and nonlinear processes, the test hypothesis is "whether any stationary model, possibly nonlinear, which satisfies the assumptions of Theorem 3.1, can explain the changing volatility of the S&P 500 index over this period". However, the p-value of our test is less than 4×10^{-4}. This provides extremely strong evidence against the null hypothesis of stationarity, and in particular against nonlinear stationary models such as GARCH or its many variants. To place this result in a broader context, we mention that recently some authors have advocated the use of nonstationary models for the evolution of financial log-returns, see e.g. Starica and Granger (2005) and Fryzlewicz et al. (2008).

We conclude the section on testing for nonstationarity with a few important remarks.

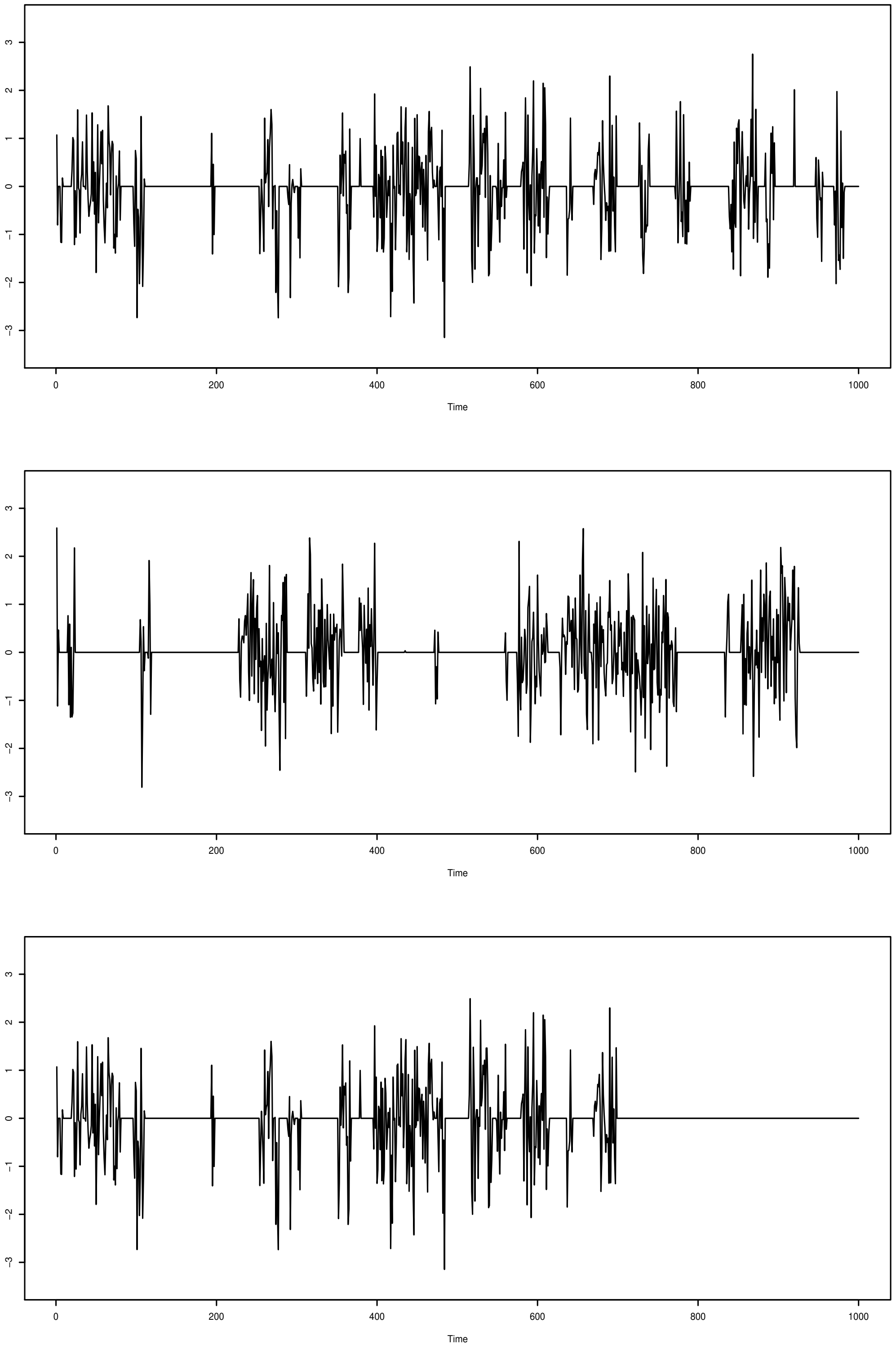


Figure 4: Simulated sample paths for cases A (top), B (middle), C (bottom) in simulation study (e), Section 3.1.

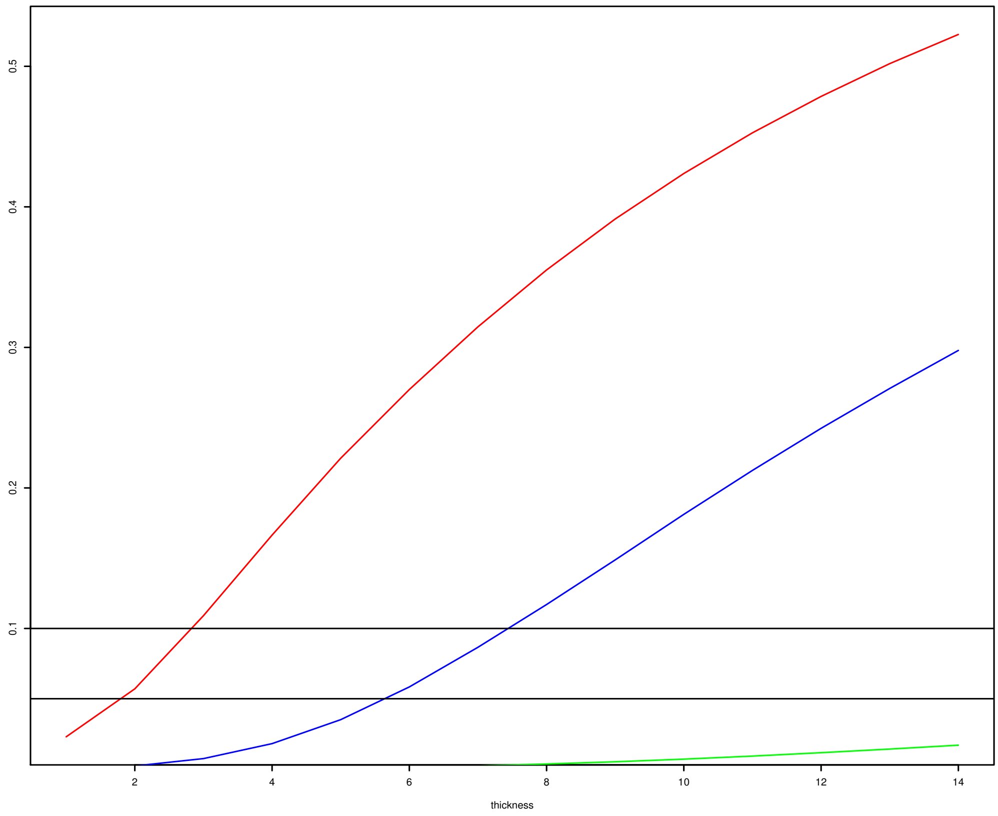


Figure 5: p-value as a function of thickness, averaged over 100 simulated sample paths from models A (red), B (blue), C (green), each for a sample path of length 1000, in simulation study (e), Section 3.1. Horizontal lines at 0.05 and 0.1 for reference.

Table 1: Average p-values for simulation model (a).

σ	1	1.1	1.2	1.3	1.35	1.4
p-value	0.59	0.47	0.29	0.1	0.07	0.04

Choice of τ_1 and τ_2. Our recommendations for the choice of τ_1 and τ_2 are as follows. Since the majority of commonly encountered nonstationary processes, such as piecewise ARMA and (G)ARCH processes, exhibit changes in their dependence structure at small lags, we suggest setting $\tau_1 = 1$ when dealing with data which could be modelled in such frameworks. In light of Theorem 2.1, it may be helpful to perform some preliminary data analysis before selecting τ_2 as, for example, the largest significant lag for which the autocovariance of the process is likely to change over time. It is much more challenging to advise on a suitable choice of τ_1, τ_2 for processes resembling that of our simulation study (e) above, and we leave this interesting question for future work.

Choice of K_t^τ for $\tau > \tau_1$. When testing using multiple thicknesses τ, it must be borne in mind that for the most common choices of the summary statistic K_t^τ (e.g. volume, mean), the degree of dependence between K_t^τ for different consecutive values of τ will

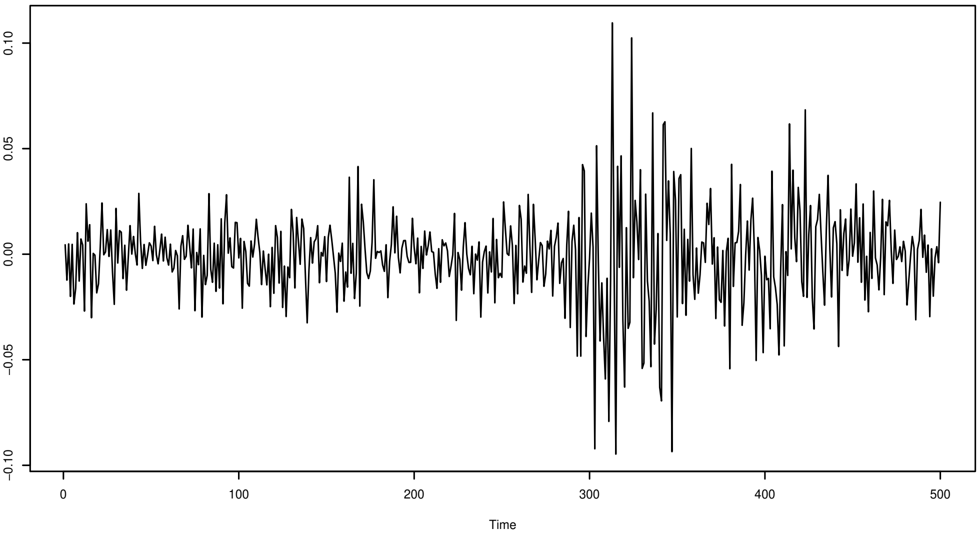


Figure 6: Logged and differenced daily closing values of the S&P 500 index, for the 500 trading days ending 25 November 2009.

typically be non-negligible, and therefore the Bonferroni correction may not be very effective. In order to “break” this dependence, instead of taking, say, $K_t^{\tau_1} = V_t^{\tau_1}$, $K_t^{\tau_1+1} = V_t^{\tau_1+1}$, $K_t^{\tau_1+2} = V_t^{\tau_1+2}$ (etc.), we suggest taking $K_t^{\tau_1} = V_t^{\tau_1}$, $K_t^{\tau_1+1} = V_t^{\tau_1+1} - V_t^{\tau_1}(= \frac{\Delta V_t^{\tau_1+1}}{\Delta\tau})$, $K_t^{\tau_1+2} = V_t^{\tau_1+2} - V_t^{\tau_1+1}(= \frac{\Delta V_t^{\tau_1+2}}{\Delta\tau})$, etc. Empirically, we have observed that this adjustment reduces the degree of dependence and thus makes the Bonferroni correction more effective.

We illustrate the above choice of K_t^{τ} with a challenging example of a non-stationary process X_t, which is Gaussian throughout, with variance one and lag-one autocorrelation equal to 1/2; however, it is AR(1) in the first half and MA(1) in the second half (we refer to this model as AR(1)-MA(1) below). Since the change in the autocorrelation structure only occurs at lag two, it is clear that in this case it will be insufficient to test using $K_t^1 = V_t^1$ as V_t^1 only uses lag-one information. Thus, we also test using $K_t^2 = V_t^2 - V_t^1$. Average p-values for our thick-pen test for stationarity as a function of the sample size, based on 100 simulated sample paths, for both K_t^1 and K_t^2, are shown in Figure 7.

It is unsurprising but interesting to observe that K_t^1 performs extremely poorly and that K_t^2 performs well at detecting the non-stationarity in this model, since the largest lags examined by K_t^1 and K_t^2 are, respectively, 1 and 2. To assess how well K_t^2 performs, we compared the performance of our test to an analogous test based directly on the local autocorrelation at lag two; in other words, we built the Brownian bridge statistic exactly like in our test but based on the sequence X_tX_{t+2} instead of K_t^{τ}. Since X_t is Gaussian and its autocorrelation varies the most prominently at lag 2, we would expect this autocorrelation-based test to be close to optimal; with some abuse of terminology, we henceforth refer to it as the “oracle” test. It is unsurprising to note that the oracle test does better than our thick-pen-based test since the former was fine-tuned to this particular Gaussian example; however, it is also interesting to note

Table 2: Average p-values for simulation model (b).

$a^{(1)} \quad \backslash \quad a^{(2)}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.59	0.51	0.39	0.29	0.1	0.03	0.01	0	0	0
0.1		0.53	0.47	0.36	0.19	0.05	0.01	0	0	0
0.2			0.58	0.5	0.28	0.15	0.02	0	0	0
0.3				0.59	0.50	0.22	0.04	0	0	0
0.4					0.57	0.39	0.15	0.01	0	0
0.5						0.58	0.45	0.09	0	0
0.6							0.52	0.35	0.01	0
0.7								0.56	0.14	0
0.8									0.61	0.01
0.9										0.57

Table 3: Average p-values for simulation model (c).

$a^{(1)} \quad \backslash \quad a^{(2)}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.57	0.53	0.51	0.51	0.49	0.39	0.3	0.17	0.13	0.06
0.1		0.56	0.54	0.53	0.5	0.4	0.33	0.2	0.15	0.02
0.2			0.59	0.57	0.5	0.46	0.38	0.21	0.1	0.05
0.3				0.55	0.6	0.48	0.39	0.33	0.17	0.05
0.4					0.54	0.55	0.48	0.30	0.22	0.05
0.5						0.55	0.54	0.45	0.29	0.08
0.6							0.53	0.54	0.34	0.12
0.7								0.59	0.46	0.2
0.8									0.6	0.33
0.9										0.61

that it does not do dramatically better. On the other hand, it is worth emphasising that a similar test based on local autocorrelations would by definition be of no use when investigating changes in the dependence structure beyond the second moment. In particular, it would have to fail in the (G)ARCH framework (a set-up where our thick-pen test does well, as demonstrated earlier in this section) as (G)ARCH processes are simply white noise as far as their second-order structure is concerned.

Testing using multiple summary statistics. If there are good reasons to believe that multiple characteristics of the time series under consideration (e.g. mean level and dependence structure) change over time, it may make sense to test using multiple summary statistics simultaneously (e.g. M_t^τ and V_t^τ) instead of, or in addition to, looking at multiple thicknesses τ simultaneously. The same procedure applies, including obviously the Bonferroni correction.

Thick-pen as the last stage in a cascade of tests. Since, as demonstrated above, the thick-pen-based test is also applicable to situations where changes in the dependence structure occur beyond the second moment, it is an option to use it as the last stage in a cascade of tests, where a test for changes in the mean is performed in the first stage, a test for changes in the second-order dependence structure in the second stage (for example that proposed in Paparoditis (2009)) and the thick-pen-based test in the final, third stage.

Table 4: Percentage of times the (nonstationary) models (i)–(iii) were considered stationary in the Auto-Seg, AG and thick-pen (TP) procedures, in simulation (d).

	(i)	(ii)	(iii)
Auto-Seg	80.4	37	87.8
AG	72	21	85
TP	71.5	8.9	81.8

Table 5: Percentage of times the (stationary) models (iv)–(v) were considered nonstationary in the Auto-Seg, AG and thick-pen (TP) procedures, in simulation (d).

	(iv)	(v)
Auto-Seg	4	4
AG	4	12
TP	3	6

3.2 Measuring time series dependence via the thick-pen transform

The purpose of this section is to argue the potential usefulness of the thick-pen transform as a measure of association/dependence between time series, especially in cases where traditional measures (e.g. cross-covariance, cross-spectrum, coherence) fail due to insufficient moment conditions, or where a more "visual" measure is required. We illustrate and expand on this below. We aim at conveying the main idea, rather than at a detailed analysis.

Let X_t and Y_t be two univariate zero-mean (or zero-median if mean does not exist) time series, roughly on the same scale (by which we mean variance if it exists or another robust measure of scale if it does not). Our main basic idea is to measure the overlap between the areas created by the thick-pen transforms of X and Y. Denoting by $L_t^\tau(Z)$ ($U_t^\tau(Z)$) the lower (upper) thick-pen boundary for a generic process Z at time t, thickness τ, we define the localised Thick-Pen Measure of Association (TPMA) between X and Y at time t, thickness τ as

$$\rho_t^\tau(X,Y) = \frac{\min\{U_t^\tau(X), U_t^\tau(Y)\} - \max\{L_t^\tau(X), L_t^\tau(Y)\}}{\max\{U_t^\tau(X), U_t^\tau(Y)\} - \min\{L_t^\tau(X), L_t^\tau(Y)\}}.$$

Note that if the intersection between $[L_t^\tau(X), U_t^\tau(X)]$ and $[L_t^\tau(Y), U_t^\tau(Y)]$ is non-empty, then $\rho_t^\tau(X,Y)$ simply measures the length of this intersection as a proportion of the length of the union of these two intervals. Indeed, if $X = Y$, then $\rho_t^\tau(X,Y) \equiv 1$. If $[L_t^\tau(X), U_t^\tau(X)]$ and $[L_t^\tau(Y), U_t^\tau(Y)]$ do not intersect, then $\rho_t^\tau(X,Y)$ measures the length of the "gap" between them as a proportion of the length of the shortest interval containing their union, times minus one. If X and Y are stationary between times t_1 and t_2, a natural averaged version of $\rho_t^\tau(X,Y)$ is

$$\bar{\rho}_{t_1,t_2}^\tau(X,Y) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \rho_t^\tau(X,Y)$$

Note that the range of $\rho_t^\tau(X,Y)$ is $(-1,1]$ and so it is a bounded random variable, which in particular possesses all finite moments, irrespective of the degree of heavy-tailedness of X or Y. This, in particular, implies that under appropriate mixing conditions (see e.g.

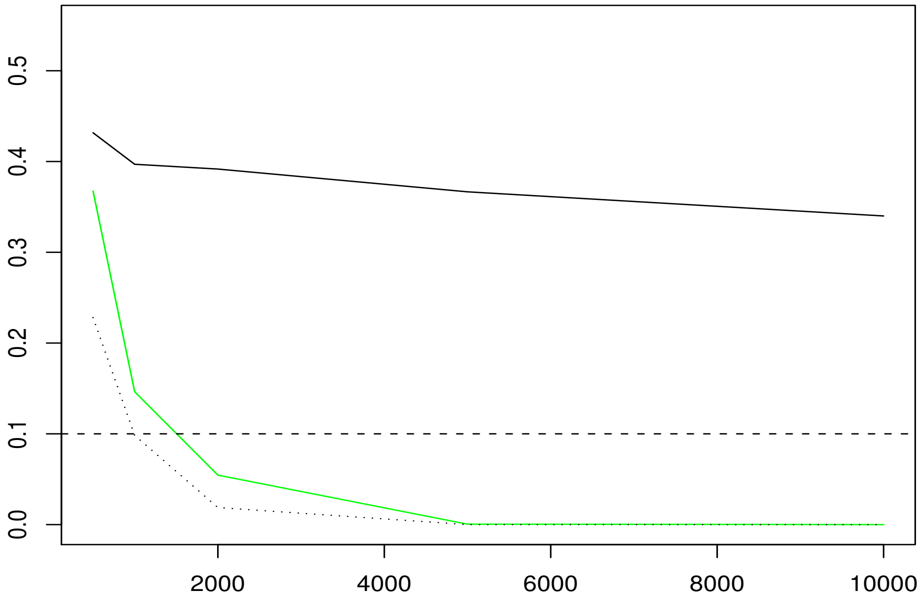


Figure 7: Average p-values (over 100 simulated sample paths) in detecting the non-stationarity of the AR(1)-MA(1) model, as a function of the sample size. Solid black: thick-pen test based on K_t^1. Green: thick-pen test based on K_t^2. Dotted black: "oracle" test. Dashed horizontal line at 0.1 for reference.

Davidson (1994), Chapter 19, for a review), we have

$$\lim_{t_2 - t_1 \to \infty} \bar{\rho}^{\tau}_{t_1,t_2}(X,Y) = \mathbb{E}(\rho^{\tau}_t(X,Y))$$

in the almost-sure sense, if X and Y are stationary (but not necessarily light-tailed). However, a similar convergence result does not hold for the sample correlation between X and Y if their second moments do not exist.

Although the range of $\rho^{\tau}_t(X,Y)$ is $(-1,1]$, the reader should not fall into the trap of identifying values of $\rho^{\tau}_t(X,Y)$ close to -1 (0, 1) with "perfect negative" ("lack of", "perfect positive") correlation between X and Y. The TPMA $\rho^{\tau}_t(X,Y)$ describes how the two time series *appear to co-vary when seen from the distance corresponding to thickness* τ. For example, if $\{X_t\}_{t=1}^{100}$ and $\{Y_t\}_{t=1}^{100}$ are two independent Gaussian white noise sequences, they will invariably appear very similar when viewed from a sufficiently large distance. The numerical analysis of the quantity $\bar{\rho}^{99}_{1,100}(X,Y)$ (note the extremely high value of the thickness parameter) applied to this example confirms this observation: indeed, in 100 simulated realisations, the value of this quantity, in this particular example, never fell below 0.97. However, when computed at a range of (lower) thickness values, TPMA is well able to discriminate between different degrees and types of dependence in time series, also in cases where covariance fails. This is demonstrated next. (As an aside, we mention that TPMA is a "visual" measure in the sense that it is often possible to deduce an approximate value of $\rho_t^\tau(X, Y)$ simply by the visual inspection of the graphs of X and Y. However, the same cannot be said of covariance as it involves multiplication, which is not an obvious operation to perform graphically.)

We demonstrate the advantages of the TPMA using three examples. The first one involves TPMA analysis of the DJIA and FTSE 100 stock indices, with an interesting possible interpretation of the results. The second one involves variable selection where the covariates (and the response) are extremely heavy-tailed. The final example demonstrates the sensitivity of TPMA to the phase between the two input time series. We use the square pen throughout.

Example 1. In this example, we perform the Thick-Pen Measure of Association (TMPA) analysis for a pair of time series: daily log-returns on the Dow Jones Industrial Average (X_t) and FTSE 100 (Y_t) indices, on 2048 trading days ending 10 March 2010; both series are scaled so that their variance is 1. We note that the initial part of both series corresponds to the final part of the burst of the "dotcom bubble" in the early 2000s, while the final part of both series covers the period of the recent financial crisis. Both of those periods are characterised by a dramatic increase in volatility. The series are displayed in Figure 8.

To analyse the series, we use the TPMA with the square pen, the thickness parameter τ ranging from 1 to 199, and the scaling factor γ equal to zero: this is done to ensure that the TPMA is invariant to changes in the marginal volatility of each series (indeed, with the scaling factor equal to zero, the value of the TPMA remains the same if each series is multiplied by the same constant σ). For each thickness parameter τ, for ease of visual interpretation, we additionally smooth the TPMA sequence $\rho_t^\tau(X, Y)$ by means of a Gaussian kernel smoother with bandwidth 200. For comparative purposes, we also compute the localised cross-correlation sequence $\gamma_t(X, Y)$ between X and Y, where the localisation is also achieved by means of the Gaussian kernel with bandwidth 200. For ease of comparison, we further normalise each sequence $\rho_t^\tau(X, Y)$ so that its overall (global) sample mean and variance match those of $\gamma_t(X, Y)$.

Results are visualised in Figure 9. The top plot shows $\gamma_t(X, Y)$ as well as $\rho_t^\tau(X, Y)$ for $\tau = 1, 4, 19$. The peak present in all the sequences around time $t = 300$ indicates increased co-dependence between the two time series during the dotcom crisis. Similarly, the peak around time $t = 1800$ indicates increased co-dependence during the recent financial crisis. It is interesting to see that the two peaks are apparent in all four indicators.

However, it is fascinating to observe that the relationship between the levels of the two peaks differs depending on the value of the thickness parameter. For $\tau = 1$, which roughly corresponds to the 'daily' scale of the data, the value of $\rho_t^\tau(X, Y)$ at the later peak is higher than that at the earlier peak. However, the opposite is true for $\tau = 4$ ('weekly' scale) and $\tau = 19$ ('monthly' scale). This is illustrated further in the middle plot, which shows the time-thickness "map" of $\rho_t^\tau(X, Y)$ (the darker the colour, the larger the value) and in the bottom plot, which shows (in red colour) the time-thickness regions where $\rho_t^\tau(X, Y)$ exceeds the threshold of 0.566.

Although this finding appears challenging to interpret, we note that the dotcom crisis was a single-sector crisis, affecting mainly companies from the highly-globalised IT sector. On the other hand, the recent financial crisis was a more complex phenomenon, involving both global-scale events (e.g. spectacular shocks in the global banking sector) and country-specific events (e.g. bailout packages, announcements of macroeconomic indicators). We

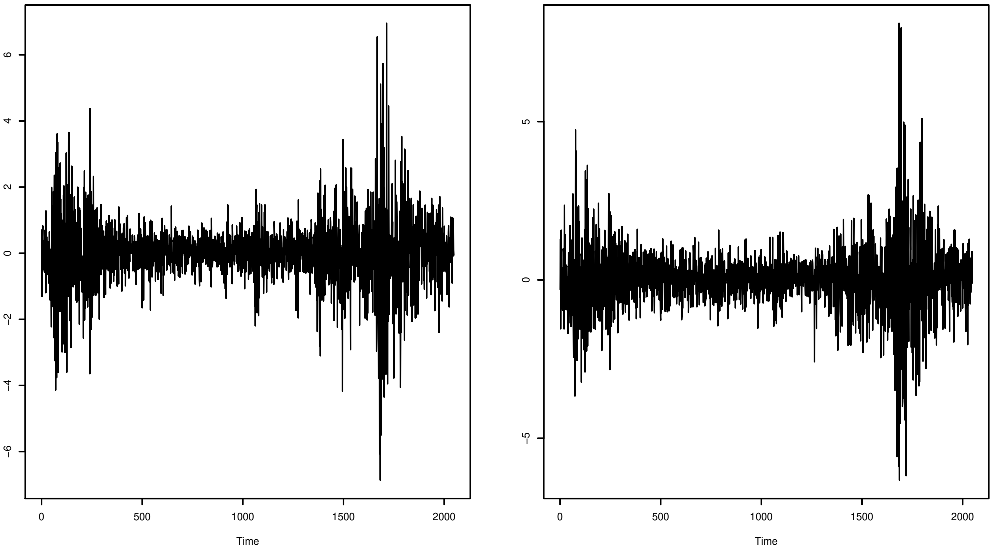


Figure 8: Left: FTSE 100 index; right: DJIA index, both on 2048 trading days (roughly 8 years) ending 10 March 2010.

can only hypothesise that it might have been the presence of those country-specific effects that led to the decreased dependence between DJIA and FTSE 100 as measured by the TPMA for larger thickness values. In other words, during the recent financial crisis, DJIA and FTSE 100 might have been responding in unison to abrupt global shocks (hence the extremely high co-dependence of the two series for smaller thickness values) but might have been less co-ordinated over longer time-scales (corresponding to higher thickness values) due to the possibly longer-term effects of country-specific factors affecting the value of these two stock indices.

Finally, we mention that a similar analysis would not have been possible only based on the local cross-correlation sequence $\gamma_t(X, Y)$, as the latter quantity does not reflect the concept of "viewing" the data at multiple time-scales.

Example 2. Unlike the other two examples, this one exceptionally departs from the domain of time series and considers a possible application of the TPMA to the problem of variable selection for heavy-tailed data. Such data arise naturally in finance (returns on financial instruments) and biostatistics (gene expressions), amongst others. The Cauchy distribution displays an extreme degree of heavy-tailedness in the sense that even its first moment does not exist. In the first part of the example, we attempt to measure the degree of linear association between X and $Z = a\,X + Y$, where X and Y are independent Cauchy variates. Having observed $\{X_t\}_{t=1}^n$ and $\{Z_t\}_{t=1}^n$, we apply (a) the sample (Pearson) correlation between X and Z, (b) Kendall's *tau* rank correlation, (c) Spearman's *rho* rank correlation, and (d) the TPMA $\bar{\rho}^1_{1,n}(X, Z)$, in an attempt to quantify the degree of dependence between X and Z, for a range of values of a. While we should not hope for either of these estimators to return a value close to a itself, it would be desirable for the sample distributions of the estimators to concentrate around values whose magnitude increases with a, to reflect the increasing degree of dependence of X and Z as a increases.

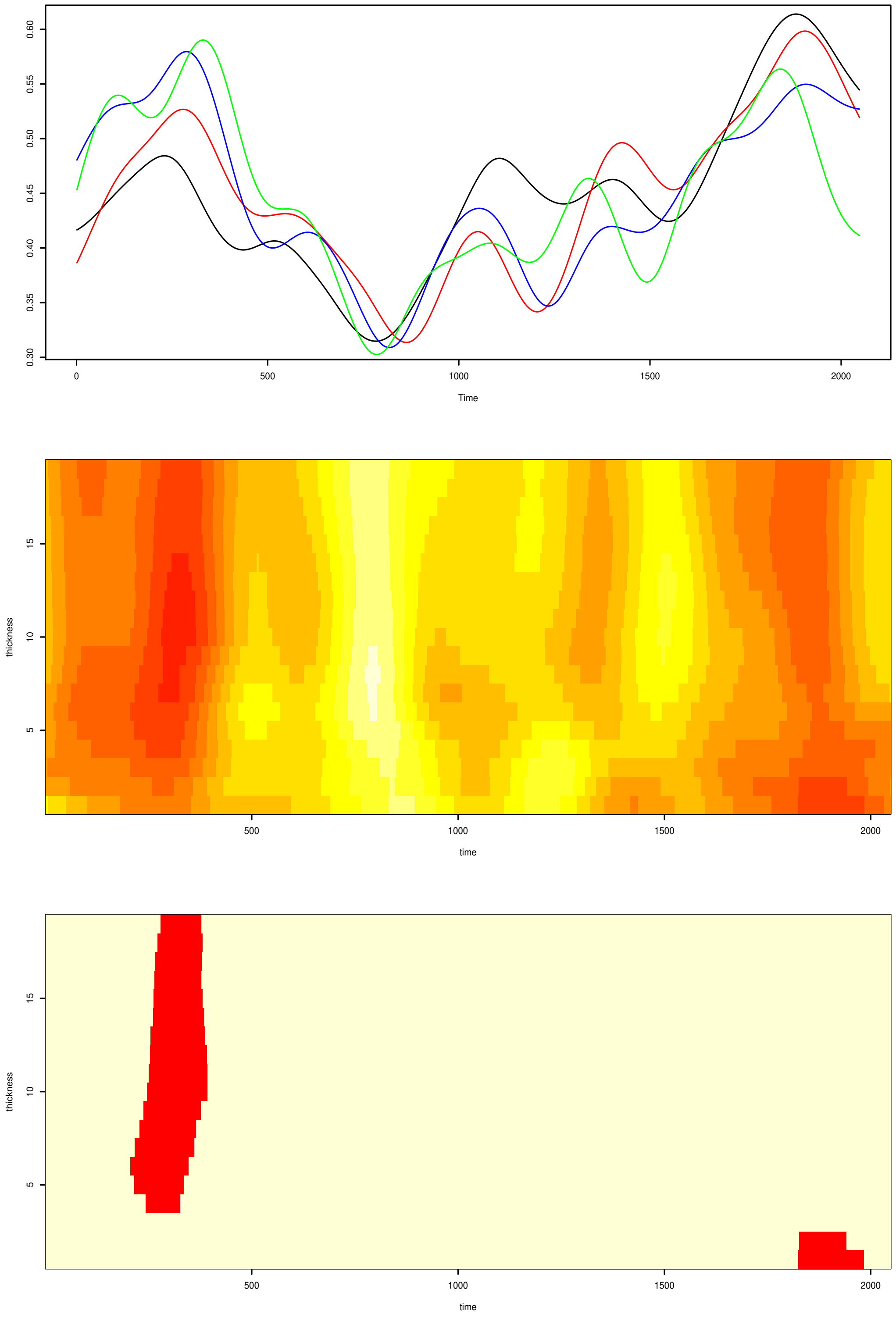


Figure 9: Top: $\gamma_t(X, Y)$ (black), $\rho_t^\tau(X, Y)$ for $\tau = 1, 4, 19$ (red, blue, green, respectively). Middle: time-thickness plot of $\rho_t^\tau(X, Y)$ (darker colour means larger value). Bottom: regions where $\rho_t^\tau(X, Y)$ exceeds 0.566 (red).

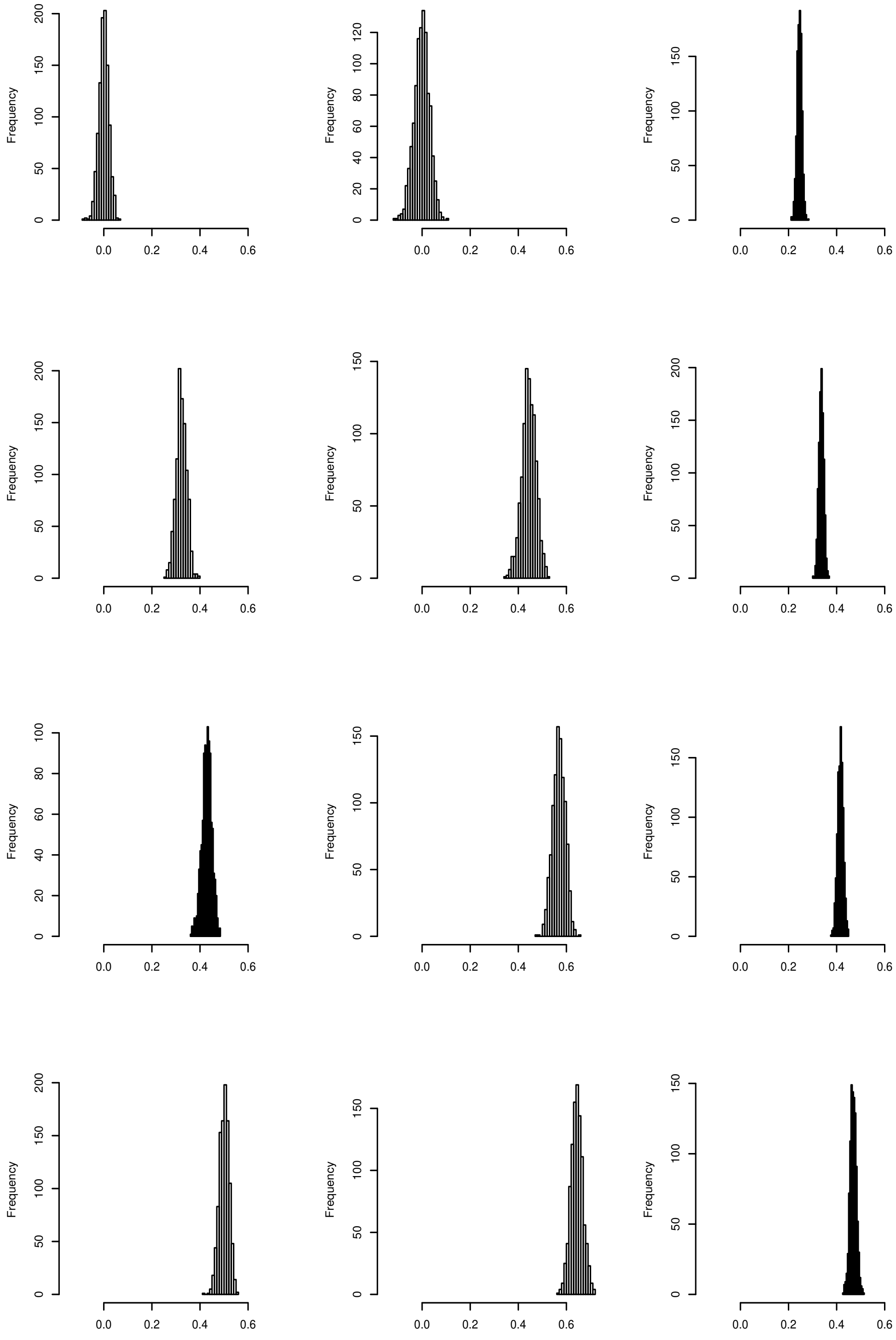


Figure 10: Left column: histograms, over 1000 simulations, of Kendall's *tau* correlations between $\{X_t\}_{t=1}^{1000}$ and $\{Z_t\}_{t=1}^{1000}$, $Z_t = aX_t + Y_t$; X, Y independent Cauchy, for $a = 0, 0.4, 0.7, 1$ (from top to bottom). Middle column: Spearman's *rho*; right column: $\bar{\rho}^1_{1,1000}(X, Z)$, for the same values of a.

Figure 10 shows the sample distributions of Kendall's *tau* correlations between $\{X_t\}_{t=1}^n$ and $\{Z_t\}_{t=1}^n$, for $n = 1000$, for a range of values of a, as well as the sample distributions of Spearman's *rho* and $\bar{\rho}_{1,n}^1(X, Z)$. It is clear that all three statistics provide concentrated and peaked distributions, and that their average values appear to increase with a, as required. However, the advantage of $\bar{\rho}_{1,n}^1(X, Z)$ is that it is computationally fast (unlike Kendall's *tau* and Spearman's *rho*, it does not involve a full sorting of the observations), a property which would be of particular significance in tasks such as high-dimensional variable selection. We note that the computational complexity of $\bar{\rho}_{1,n}^\tau(X, Z)$ is $O(n)$. The sample (Pearson) correlation is also rapid to compute, but it exhibits extremely poor performance in this example, which is not surprising as the Cauchy distribution does not possess a finite variance or mean.

To investigate this issue further, the second part of Example 1 continues this theme but in the context of variable selection. Sample correlation is the basic ingredient of most modern variable selection techniques, including, amongst others, LARS (Efron et al. (2004)) and its various special cases such as Lasso (Tibshirani (1996)) or forward stagewise selection. It could be replaced by Kendall's *tau* or Spearman's *rho* if the data were extremely heavy-tailed. We consider a simple linear regression model

$$Y = \beta_1 X^1 + \beta_2 X^2 + \varepsilon,$$

where X^1, X^2 and ε are independent Cauchy-distributed variables, and suppose that we collect $n = 1000$ independent observations. We fix $\beta_1 = 1$, and vary β_2 from 0.7 to just short of 1. Since $\beta_1 > \beta_2$, and thus the degree of dependence between Y and X^1 is greater than that between Y and X^2, it would be desirable for any measure of vector dependence to return a larger value for the pair (Y, X^1) than for (Y, X^2), i.e. to rank X^1 ahead of X^2 in importance. Table 6 shows the percentage of times (over 1000 simulations) that this correct ranking was achieved by simple sample correlation (Marginal Correlation Ranking; MCR), Kendall's *tau*, Spearman's *rho* and the TPMA. The latter three are extremely competitive, with the TPMA being marginally superior in all cases. We also emphasise again its lower computational complexity. This provides evidence of its potential usefulness in variable selection contexts where the number of covariates is large. We emphasise that TPMA performed well in this setting even though no rescaling of Y was performed, i.e. Y and X^i were not exactly on the same "scale". Also, we note that, unlike the three other measures, TPMA is not invariant with respect to permutations of the data (as it is designed as a "time series" measure). Thus, further performance improvement could be expected if we were to average over some permutations of the data, at the expense of computational efficiency.

Table 6: Percentage of cases the covariates were correctly ordered (over 1000 simulations) by Marginal Correlation Ranking (MCR), Kendall's *tau* (tau), Spearman's *rho* (rho) and the Thick-Pen Measure of Association with thickness $\tau = 1$ (TPMA), as a function of β_2.

β_2	0.7	0.8	0.9	0.95	0.99
MCR	62	59	55	52	51
tau	99	96	83	65	51
rho	99	95	81	66	53
TPMA	100	97	83	66	53

Example 3. The final example illustrating the potential applicability of the TPMA in multivariate time series analysis concerns the detection of phase between time series. In

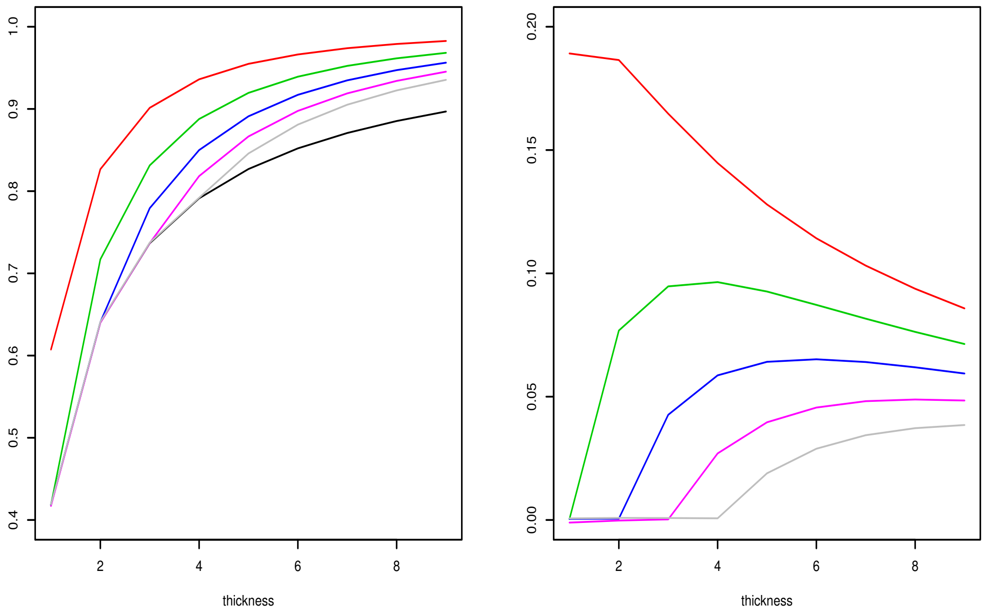


Figure 11: Thick-pen cross-spectra from Example 2. Top: cross-spectrum for $q = 1$ (red), $q = 2$ (green), $q = 3$ (blue), $q = 4$ (magenta), $q = 5$ (grey) and the independent case (black). Bottom: differences between the coloured curves and the black curve; colours correspond.

this example, both $\{X_t\}_{t=1}^{200}$ and $\{Y_t\}_{t=1}^{200}$ are i.i.d. Gaussian sequences (the reader is invited to think of them as, for example, representing residuals from two univariate model fits), which are however dependent on each other in the sense that Y is a shifted version of X:

$$Y_t = B^q X_t,$$

where B is the shift operator, and q ranges from 1 to 5. We also test the case where X is independent of Y. Figure 11 shows what we term the Thick-Pen Cross-Spectrum of X and Y, i.e. the sequence $\{\bar{\rho}^{\tau}_{1,200}(X, Y)\}_{\tau}$, here for τ ranging from 1 to 9, plotted for various values of q as well as for the case of X and Y being independent, averaged over 1000 simulated sample paths.

If $Y_t = B^q X_t$, then $X_t, \dots, X_{t+q-1}$ is independent of $Y_t, \dots, Y_{t+q-1}$ and thus, for $\tau = 1, \dots, q-1$, we have that $\mathbb{E}(\bar{\rho}^{\tau}_{1,200}(X, Y)) = \mathbb{E}(\bar{\rho}^{\tau}_{1,200}(X, Z))$, where Z is independent of X. This is illustrated in Figure 11, which shows that the averaged Thick-Pen Cross-Spectrum only diverges from the spectrum for independent series for thickness values starting from $\tau = q$, thus providing a natural way of estimating the phase q between X and Y. Exactly the same phenomenon was observed in a similar example using the Cauchy distribution.

To conclude this section, we note that unlike covariance and related measures which only make sense for a pair of time series, the TPMA naturally generalises to more than two time series. For a collection of time series $X^1, \dots, X^M$, we define

$$\rho^{\tau}_t(X^1, \dots, X^M) = \frac{\min_i\{U^{\tau}_t(X^i)\} - \max_i\{L^{\tau}_t(X^i)\}}{\max_i\{U^{\tau}_t(X^i)\} - \min_i\{L^{\tau}_t(X^i)\}}.$$

It is our belief that this extension can potentially pave the way for the application of the TPMA in tasks such as time series clustering.

3.3 Non-stationary time series classification via the thick-pen transform

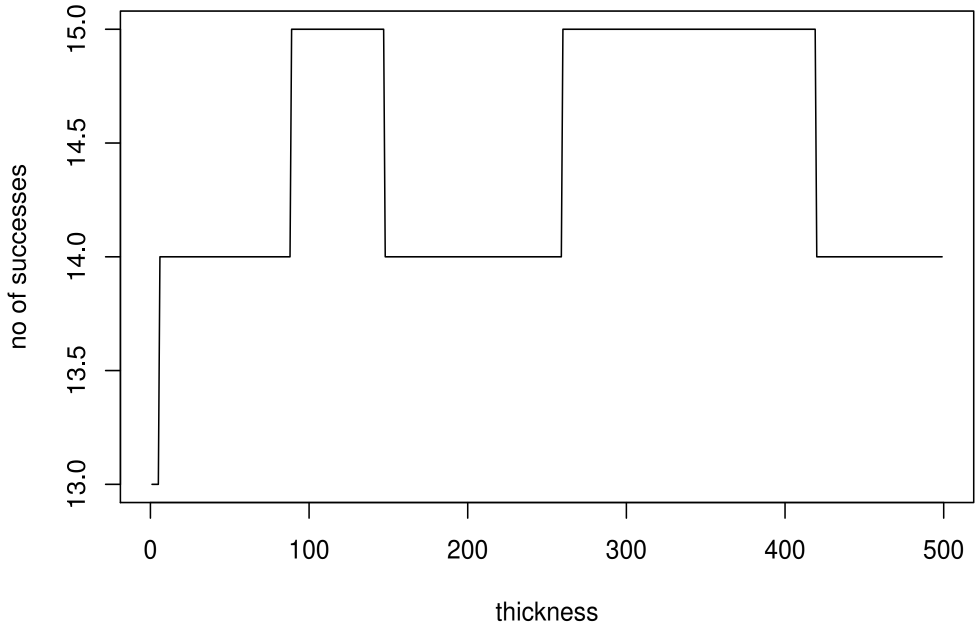
In this section, we show how the TPT can be used to classify not-necessarily-stationary time series. The setting is as follows: we have $C \geq 2$ groups of time series, where time series within each group follow the same distribution, but the distributions differ across the groups. Within each group c, we have already observed N_c time series. Given a "new arrival" X_t, we wish to classify it as belonging to one of the C groups. The generic TPT-based classification algorithm proceeds as follows:

1. Decide on a suitable summary sequence K_t^τ.
2. Compute K_t^τ for each series in each group, and average this summary sequence contemporaneously over the series within each group, to produce $\bar{K}_t^{\tau,c}$ for each of the groups $c = 1, \dots, C$.
3. Compute $K_t^\tau(X)$ for the new arrival X_t.
4. Using a pre-selected distance function $d(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, compute the distances between $K_t^\tau(X)$ and $\bar{K}_t^{\tau,c}$ for $c = 1, \dots, C$.
5. Classify X_t to the group which corresponds to the smallest distance.

By comparing $K_t^\tau(X)$ and $\bar{K}_t^{\tau,c}$, we essentially compare the (suitably understood) "shape" of the new arrival against the average "shapes" of the time series within each group, and classify the new arrival to the group where the average shape of the time series (viewed at thickness τ) resembles the most closely that of the new arrival.

We illustrate the use of the above algorithm on a well-known geophysics dataset. In the monitoring of a comprehensive test ban treaty, it is critical to develop methods for discriminating between nuclear explosions and earthquakes. We applied the proposed methodology for classifying a time series as either an explosion or an earthquake. The proliferation of nuclear explosions is monitored in regional distances of $100 - 2000$ km and the recordings of mining explosions can serve as a reasonable proxy. A data set of regional (100 – 2000 km) recordings of several typical Scandinavian earthquakes and mining explosions measured by stations in Scandinavia are used in this study. The data set, consisting of 8 earthquakes and 8 explosions, is given in Kakizawa et al. (1998). The problem is discussed in detail in Shumway and Stoffer (2006), and the data are available online from `http://lib.stat.cmu.edu/general/tsa2`.

Prior to the analysis, we center and scale each time series so that its sample mean is zero and its sample variance is one. We use the following selection of parameters: the square pen, $K_t^\tau = V_t^\tau$ (volume), $d^2(f,g) = \frac{1}{n}\sum_{t=1}^{n}(f_t - g_t)^2$. We remove each of the $2 \times 8 = 16$ series one by one, and classify it using the above algorithm. As a result, each time series is classified either to the group consisting of the 7 remaining series from the same category (successful classification) or to the group consisting of the 8 series from the other category (failed classification).



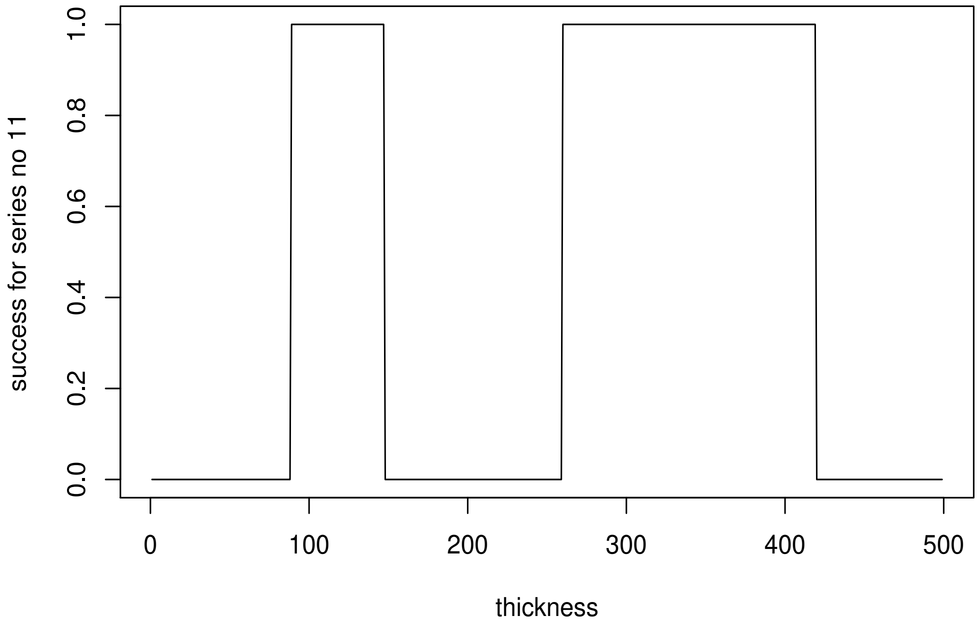


Figure 12: Top: number of successful classifications (out of 16) as a function of thickness in the data analysis of Section 3.3. Bottom: classification success for time series no. 11 as a function of thickness.

The top plot in Figure 12 shows the number of successful classifications (out of 16) as a function of thickness. The number of successes peaks at 15 for a certain range of thicknesses, demonstrating the effectiveness of the methodology. The "double bump" in the plot is caused by the series number 11, for which the classification success as a function of thickness is plotted in the bottom plot of Figure 12 and also exhibits the double bump. Albeit an unusual feature, it is very much in line with the multiscale philosophy of "viewing the data at multiple thicknesses / scales". The result can be interpreted as saying that series number 11 resembles other series from the same group when "viewed" at thickness 100 and around, as well as 300-400 and around, but not at thickness 200 and around.

The leave-one-out cross-validation as described above is a practical way of choosing the "right" thickness value(s) in classification problems.

4 Discussion

In this section, we mention a few further aspects of the thick-pen methodology, which in our view would merit further study.

Local extrema of U_t^τ, L_t^τ as a marker. Depending on the nature of the time series and problem at hand, more complex "markers" involving U_t^τ and L_t^τ are possible. Keeping track of local maxima of U_t^τ and local minima of L_t^τ is one such example. As illustrated in Figure 3, for certain thicknesses, U_t^τ appears to attain local maxima more often for time series with less dependence structure. If this were indeed the case, the same would obviously hold for local minima of L_t^τ, for symmetrically distributed time series. Thus the following binary markers might be of use in nonstationarity detection and classification:

$$\begin{aligned}\overline{U}_t^\tau &= \mathbb{I}(U_t^\tau > U_{t-1}^\tau \wedge U_t^\tau > U_{t+1}^\tau)\\ \underline{L}_t^\tau &= \mathbb{I}(L_t^\tau < L_{t-1}^\tau \wedge L_t^\tau < L_{t+1}^\tau),\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Adaptive-thickness pens. The analogy between pens and kernels, mentioned in Section 2, raises the question of whether it would be meaningful and valuable to consider pens whose shape or thickness varies over time t according to local properties of the time series X_t, e.g. its visibility from a distance.

Unequispaced time series. The thick-pen transform, unlike many linear transforms such as the Fourier or wavelet transforms, extends naturally and easily to non-equispaced time series.

Acknowledgements

PF would like to thank HSO for his hospitality during a visit to Seoul National University, Korea, where this work was initiated. We wish to thank Haeran Cho, Guy Nason, Hernando Ombao, Rainer von Sachs, Johan Segers, David Stoffer and Qiwei Yao for interesting and fruitful discussions.

A Proofs

Proof of Proposition 2.1. Consider the total variation of $\Upsilon_t^\tau := U_t^\tau - \frac{\tau}{2}$ (clearly the additive constant $\frac{\tau}{2}$ has no impact on the total variation). For any sequence f, define the shift operator B by $(Bf)_t = f_{t+1}$. Note the recursive relationship

$$\Upsilon_t^\tau = \max(\Upsilon_t^{\tau-1}, B\Upsilon_t^{\tau-1}).$$

It now suffices to observe that since $B\Upsilon_t^{\tau-1}$ is the shift of $\Upsilon_t^{\tau-1}$ in the *horizontal* direction, taking this shift and taking the maximum of the two resulting functions cannot possibly increase their total variation. Thus we have $\|\Upsilon_t^{\tau-1}\|_{TV} = \|B\Upsilon_t^{\tau-1}\|_{TV} \geq \|\Upsilon_t^\tau\|_{TV}$, and the proof is complete. □

Proof of Theorem 2.1. We first define $T_X^{s,s+\tau} = V_s^\tau(X) - \tau$ and note that

$$T_X^{s,s+\tau} = \max_{s \leq i,j \leq s+\tau} (X_i - X_j).$$

Without loss of generality, we consider $s = 1$; note that $t = \tau + 1$. In the new notation, our task is to show that the joint distributions of $\{T_X^{1,t-1}, T_X^{2,t}, T_X^{1,t}\}$ and $\{T_Y^{1,t-1}, T_Y^{2,t}, T_Y^{1,t}\}$ are different. Proceeding by contradiction, let us suppose that they are the same. If this were true, then in particular, we would have the equality

$$\begin{gathered} P(T_X^{1,t} - T_X^{1,t-1} > 0 \quad \wedge \quad T_X^{1,t} - T_X^{2,t} > 0 \quad \wedge \quad T_X^{1,t} > \lambda) = \\ P(T_Y^{1,t} - T_Y^{1,t-1} > 0 \quad \wedge \quad T_Y^{1,t} - T_Y^{2,t} > 0 \quad \wedge \quad T_Y^{1,t} > \lambda) =: a \end{gathered}$$

We now note that if $T_X^{1,t} - T_X^{1,t-1} > 0$, this implies that $T_X^{1,t}$ necessarily achieves its maximum at $|X_t - X_q|$ for some q. Also (symmetrically), if $T_X^{1,t} - T_X^{2,t} > 0$, then $T_X^{1,t}$ achieves its maximum at $|X_1 - X_q|$ for some q. Therefore,

$$T_X^{1,t} - T_X^{1,t-1} > 0 \quad \wedge \quad T_X^{1,t} - T_X^{2,t} > 0$$

means that $T_X^{1,t}$ achieves its maximum at $|X_t - X_1|$. Thus we have the event equality

$$\begin{gathered} T_X^{1,t} - T_X^{1,t-1} > 0 \quad \wedge \quad T_X^{1,t} - T_X^{2,t} > 0 \quad \wedge \quad T_X^{1,t} > \lambda = \\ T_X^{1,t} - T_X^{1,t-1} > 0 \quad \wedge \quad T_X^{1,t} - T_X^{2,t} > 0 \quad \wedge \quad |X_t - X_1| > \lambda, \end{gathered}$$

and similarly for Y.

We now introduce σ_1, σ_2, such that $X_t - X_1 \sim N(0, \sigma_1^2)$ and $Y_t - Y_1 \sim N(0, \sigma_2^2)$. By the assumptions of the Theorem they are different, and without loss of generality, we assume $\sigma_1 > \sigma_2$.

Denoting $A_{X,t} = \{T_X^{1,t} - T_X^{1,t-1} > 0 \quad \wedge \quad T_X^{1,t} - T_X^{2,t} > 0\}$, we decompose

$$\begin{aligned} P(|X_t - X_1| > \lambda) &= P(|X_t - X_1| > \lambda \quad \wedge \quad A_{X,t}) + P(|X_t - X_1| > \lambda \quad \wedge \quad A_{X,t}^c) \\ &= a + P(|X_t - X_1| > \lambda \quad \wedge \quad A_{X,t}^c). \end{aligned}$$

Similarly,

$$P(|Y_t - Y_1| > \lambda) = a + P(|Y_t - Y_1| > \lambda \quad \wedge \quad A_{Y,t}^c).$$

Observe now that Assumption 2.1 simply means that if λ is large enough, we have

$$P(A^c_{X,t} \quad \big| \quad |X_t - X_1| > \lambda) \leq \delta < 1,$$

and similarly for Y. This implies that

$$\begin{aligned}(1-\delta)P(|X_t - X_1| > \lambda) &\leq P(|X_t - X_1| > \lambda) - P(|X_t - X_1| > \lambda \wedge A^c_{X,t}) \\ &= a \leq P(|X_t - X_1| > \lambda).\end{aligned}$$

Similarly,

$$(1-\delta)P(|Y_t - Y_1| > \lambda) \leq a \leq P(|Y_t - Y_1| > \lambda).$$

In particular, the above inequalities hold in the region $\lambda \geq \sigma_2\sqrt{2\log n}$ where $n \to \infty$. However, setting $\lambda = \sigma_2\sqrt{2\log n}$ and using simple properties of the tails of univariate normal distributions, we then have that $a = a_n$ (ignoring irrelevant logarithmic terms) is simultaneously of the order n^{-1} (based on the inequality for Y) and of the order $n^{-\sigma_2^2/\sigma_1^2}$ (based on the inequality for X), which is a contradiction. This completes the proof of the theorem. □

Further clarification of Assumption 2.1. In the following, we identify an easy-to-verify, mixing-type assumpion which implies Assumption 2.1. We have, using the Bonferroni inequality on the way

$$\begin{aligned}&P(A^c_{X,t} \quad \big| \quad |X_t - X_1| > \lambda) \leq \\ &\quad P(T_X^{1,t} - T_X^{1,t-1} = 0 \quad \vee \quad T_X^{1,t} - T_X^{2,t} = 0 \quad \big| \quad |X_t - X_1| > \lambda) \leq \\ &\quad P(\max_{(i,j)\neq(1,t)} |X_i - X_j| \geq |X_t - X_1| \quad \big| \quad |X_t - X_1| > \lambda) \leq \\ &\quad P(\max_{(i,j)\neq(1,t)} |X_i - X_j| \geq \lambda \quad \big| \quad |X_t - X_1| > \lambda) = \\ &\quad \frac{P\left(\bigcup_{(i,j)\neq(1,t)}\{|X_i - X_j| \geq \lambda \wedge |X_t - X_1| > \lambda\}\right)}{P(|X_t - X_1| > \lambda)} \leq \\ &\quad \frac{\sum_{(i,j)\neq(1,t)} P(|X_i - X_j| \geq \lambda \cap |X_t - X_1| > \lambda)}{P(|X_t - X_1| > \lambda)}\end{aligned}$$

At this point, we assume that for all $(i,j) \neq (1,t)$,

$$P(|X_i - X_j| \geq \lambda \cap |X_t - X_1| > \lambda) \leq \alpha_\lambda P(|X_i - X_j| \geq \lambda)P(|X_t - X_1| > \lambda),$$

where the sequence of constants α_λ is uniform over all i, j, t, and its permitted rate of increase with λ is specified below. Note that this assumption involves no maxima or minima and can be verified via simple Gaussian integration for a particular process. Applying this assumption and continuing the above chain of inequalities, we get

$$\ldots \leq \alpha_\lambda \sum_{(i,j)\neq(1,t)} P(|X_i - X_j| \geq \lambda).$$

Denote $\sigma_{i,j} = \mathrm{Var}^{1/2}(X_i - X_j)$. Continuing,

$$\ldots \leq \alpha_\lambda t^2 \exp(-\lambda^2/\{2\max_{i,j}(\sigma^2_{i,j})\}),$$

which, provided that α_λ does not go to infinity too fast with λ, can be made arbitrarily small if λ is large enough.

Proof of Theorem 3.1. We first note that since the existence of moments of a sequence of random variables implies the existence of the corresponding moments of local maxima of these variables, we have that $\mathbb{E}|K_t^\tau - \mathbb{E}(K_t^\tau)|^r < \infty$. Further, by Theorem 14.1 in Davidson (1994), the sequence $K_t^\tau - \mathbb{E}(K_t^\tau)$, for any fixed τ, inherits the mixing properties of X_t, that is, is also α-mixing with the mixing coefficients α_m satisfying $\alpha_m = O(m^{-s})$ for some $s > \frac{r}{r-2}$. Thus, for any fixed τ, the sequence $K_t^\tau - \mathbb{E}(K_t^\tau)$ satisfies the conditions of Corollary 29.7 in Davidson (1994), and the proof of (i) is complete. For (ii), since the sequence of functions h_n that map $Y_n^\tau(u)$ to $Z_n^\tau(u)$, with the limiting mapping h taking $Y_n^\tau(u)$ to $Y_n^\tau(u) - uY_n^\tau(1)$, satisfies the Extended Continuous Mapping Theorem (see e.g. Billingsley (1968)), we have that

$$Z_n^\tau(u) = h_n(Y_n^\tau(u)) \xrightarrow{d} h(B_u) = B_u^0,$$

which completes the proof. □

References

E. Andreou and E. Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17:579–600, 2002.

P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

D. R. Brillinger. *Time Series: Data Analysis and Theory.* Holt, Rinehart & Winston, Inc., New York, 1975.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, 1987.

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823, 1999.

J. Cho and J. Bae. Edge-adaptive local min/max nonlinear filter-based shoot suppression. *IEEE Transactions on Consumer Electronics*, 52:1107–1111, 2006.

R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. of Stat.*, 25:1–37, 1997.

J. Davidson. *Stochastic Limit Theory.* Oxford University Press, 1994.

R. Davis, T. Lee, and G. Rodriguez-Yam. Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29:834–867, 2008.

S. Douglas. Running max/min calculation using a pruned ordered list. *IEEE Transactions on Signal Processing*, 44:2872–2877, 1996.

P. Doukhan, G. Oppenheim, and M. S. Taqqu, editors. *Theory and Applications of Long-Range Dependence.* Birkhäuser, 2003.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

J. Fan and Q. Yao. *Nonlinear Time Series.* Springer-Verlag, New York, 2003.

P. Fryzlewicz, T. Sapatinas, and S. Subba Rao. Normalised least-squares estimation in time-varying ARCH models. *Annals of Statistics*, 36:742–786, 2008.

P. Hall and A. Wood. On the performance of box-counting estimators of fractal dimension. *Biometrika*, 80:246–252, 1993.

H. Hotelling. Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics*, 61:440–460, 1939.

H.E. Hurst. Long term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engineers*, 116: 770–799, 1951.

S. Johansen and I.M. Johnstone. Hotelling's theorem of the volume of tubes: some illustrations in simultaneous inference and data analysis. *Annals of Statistics*, 18:652–684, 1990.

Y. Kakizawa, R. Shumway, and M. Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93:328–340, 1998.

D. Kennedy. The distribution of the maximum Brownian excursion. *J. Appl. Prob.*, 13: 371–376, 1976.

E. Keogh and C.A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7:358–386, 2005.

M. Knowles and D. Siegmund. On Hotelling's geometric approach to testing for a nonlinear parameter in regression. *International Statistical Review*, 57:205–220, 1988.

S. Lee, R.M. Haralick, and L.G. Shapiro. Morphologic edge detection. *IEEE Transactions on Robotics and Automation*, 3:142–156, 1987.

T. Lindeberg. *Scale-Space Theory in Computer Vision.* Kluwer, Boston, 1994.

G. P. Nason. *Wavelet Methods in Statistics with R.* Springer, New York, 2008.

M. Neumann and R. von Sachs. A wavelet-based test for stationarity. *J. Time Ser. Anal.*, 21:597–613, 2000.

E. Paparoditis. Testing temporal constancy of the spectral structure of a time series. *Bernoulli*, 15:1190–1221, 2009.

C. Park, J. Hannig, and K.-H. Kang. Improved SiZer for time series. *Statistica Sinica*, 19: 1511–1530, 2009.

D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis.* Cambridge University Press, 2000.

M. Priestley. Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B*, 27:204–237, 1965.

M. B. Priestley. *Spectral Analysis and Time Series.* Academic Press, 1981.

V. Rondonotti, J.S. Marron, and C. Park. SiZer for time series: A new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289, 2007.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples, 2nd Edition.* Springer, New York, 2006.

C. Starica and C. Granger. Non-stationarities in stock returns. *Review of Economics and Statistics*, 87:503–522, 2005.

J. Sun. Tail probabilities of the maxima of Gaussian random fields. *Annals of Probability*, 21:34–71, 1993.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B*, 58:267–288, 1996.

S. Van Bellegem and R. von Sachs. Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, 36:1879–1924, 2008.

M. Vemis, G. Economou, S. Fotopoulos, and A. Khodyrev. The use of Boolean functions and logical operations for edge detection in images. *Signal Processing*, 45:161–172, 1995.

B. Vidakovic. *Statistical Modeling by Wavelets.* Wiley, New York, 1999.

M. Werman and S. Peleg. Min max operators in texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:730–733, 1985.

H. Weyl. On the volume of tubes. *American Journal of Mathematics*, 61:461–472, 1939.

X. Ye, M. Cheriet, and C.Y. Suen. Stroke-model-based character extraction from gray-level document images. *IEEE Transactions on Image Processing*, 8:1152–1161, 2001.