Application of the singular-spectrum analysis to change-point detection in time series

V. Moskvina
School of Medicine, Cardiff University, UK

A.A. Zhigljavsky[1]
School of Mathematics, Cardiff University, UK

Abstract: A methodology of change-point detection in time series based on sequential application of the singular-spectrum analysis is proposed and studied. The underlying idea is that if at a certain time τ the mechanism generating the time series x_t has changed, then an increase in the distance between the l-dimensional hyperplane spanned by the eigenvectors of the so-called lag-covariance matrix, and the M-lagged vectors $(x_{\tau+1}, \ldots, x_{\tau+M})$ is to be expected. Under certain conditions, the proposed algorithm can be considered as a proper statistical procedure with the moving sum of weighted squares of random variables being the detection statistic. The correlation structure of the moving sums is studied. Several asymptotic expressions for the significance level of the algorithm are compared.

Key Words: change-point detection; singular-spectrum analysis; singular value decomposition; moving sum; boundary crossing probability.

Subject Classifications: 37M10, 62M10, 62.85.

1 INTRODUCTION

Singular-spectrum analysis (SSA) is a powerful technique of time series analysis. The main idea of SSA is in applying the principal component analysis to the 'trajectory matrix' obtained from the original time series with subsequent reconstruction of the series. The methodology has been known since the mid-eighties, see Broomhead and King (1986), Broomhead et al. (1987) and Vautard et al. (1992). See also the recent monographs of Elsner and Tsonis (1996) and Golyandina et al. (2001) and the references therein. SSA is still a relatively unknown methodology in statistical circles. On the other hand, it has already become a standard tool in the analysis of

[1]Address correspondence to Prof. Anatoly Zhigljavsky, School of Mathematics, Cardiff University, Senghennydd Rd., Cardiff CF24 4AG, UK; Fax: +44(0)2920874199; E-mail: ZhigljavskyAA@cardiff.ac.uk

climatic and meteorological time series; see, for example, Fraedrich (1986), Vautard and Ghil (1989).

In the present paper we continue the SSA-related research and develop a methodology of change-point detection in time series based on the use of SSA. The software can be downloaded from our web-site
http://www.cf.ac.uk/maths/stats/changepoint/

Let us briefly describe the main idea of the method. Let $x_1, x_2, \ldots$ be a time series, M and N be two integers ($M \leq N/2$), and set $K = N - M + 1$. Define the vectors $X_j = (x_j, \ldots, x_{j+M-1})^T$ ($j = 1, 2, \ldots$) and the matrix

$$\mathbf{X} = (x_{i+j-1})_{i,j=1}^{M,K} = (X_1, \ldots, X_K),$$

which is called the trajectory matrix.

We consider $\mathbf{X}$ as multivariate data with M characteristics and K observations. The columns X_j of $\mathbf{X}$, considered as vectors, lie in the M-dimensional space $\mathbb{R}^M$. The singular value decomposition (SVD) of the so-called lag-covariance matrix $\mathbf{R} = \mathbf{X}\mathbf{X}^T$ (and of the trajectory matrix $\mathbf{X}$ itself) provides us with a collection of M eigenvalues and eigenvectors. A particular combination of a certain number $l < M$ of these eigenvectors determines an l-dimensional hyperplane in $\mathbb{R}^M$. According to the SSA algorithm, the M-dimensional data is projected onto this l-dimensional subspace and the subsequent averaging over the diagonals gives us an approximation to the original series; see the above cited literature for details.

One of the features of the SSA algorithm is that the distance between the vectors X_j ($j = 1, \ldots, K$) and the l-dimensional hyperplane is controlled by the choice of l and can be reduced to a rather small value. If the time series $\{x_t\}_{t=1}^N$ is continued for $t > N$ and there is no change in the mechanism which generates the values x_t, then this distance should stay reasonably small for X_j, $j \geq K$ (for testing, we take Q such vectors). However, if at a certain time $N + \tau$ the mechanism generating x_t ($t \geq N + \tau$) has changed, then an increase in the distance between the l-dimensional hyperplane and the vectors X_j for $j \geq K + \tau$ is to be expected.

SSA expansion tends to pick up the main structure of the time series, if there is one. (This happens when the l-dimensional subspace approximates well the M-dimensional vectors $X_1, \ldots, X_K$.) If this structure is being found and there are no structural changes, then the SSA continuation of the time series should agree with the continued series. (That is, the Q vectors X_j for $j \geq K$ should stay close to the l-dimensional subspace.) A change in structure of the time series should force the corresponding vectors X_j out of the subspace. This is the central idea of the method we propose.

SSA performs the analysis of the time series structure in a nonsequential (off-line) manner. However, change-point detection is typically a sequential (on-line) problem, and we aim to develop an algorithm that can be used in the on-line regime. This can be achieved by sequentially applying the SVD to the lag-covariance matrices computed in a sequence of time intervals, either $[n+1, n+N]$ or $[1, n+N]$. Here $n = 0, 1, \ldots$ is the iteration number and N is the length of the time interval where the trajectory matrix is computed. The latter version produces a CUSUM-type algorithm. We, however, prefer the former version, with the sequence of time intervals $[n+1, n+N]$: this version is better accommodated to the presence of slow changes in the time series structure, to outliers and to the case of multiple changes. (The price for that is a smaller size of the sample used to construct the trajectory matrices, and therefore some loss in efficiency in the ideal situation.)

SSA and the proposed change-point detection algorithm are model–free tools and generally are not intended for precise statistical inferences; they are essentially model-building procedures. However, under certain conditions, the proposed algorithm can be considered as a proper statistical procedure, see Section 3. Studying properties of this procedure is the main purpose of the paper.

The paper is organized as follows.

Section 2 is devoted to description of the main algorithm. In Section 1.1 we provide an informal description while the the description in Section 1.2 is formal. In Section 2.3 recommendations concerning the choice of parameters are given. In Section 2.4 we discuss three numerical examples illustrating some features of the method.

Section 3 is devoted to the formulation of the statistical model and stating the main statistical questions. In Section 2.1 we consider the rationale of SSA. In Section 2.2 we formulate the null hypothesis model; this allows us to express the main detection statistic as a moving weighted sum of squares of random variables (Section 2.3) and to formulate the problem of selecting the threshold in the main algorithm as the problem of boundary crossing probability for this moving sum of squares (Section 2.4).

In Section 3 the correlation structure of the sequence of the moving weighted sums of squares is studied. In particular, it is shown that this structure very much depends on the ratio of M and Q.

In Section 4 three approximations for the significance level of the algorithm are investigated. These approximations can be applied in three cases: large M and large Q (Section 4.1), large M and small Q (Section 4.2), and large M and $Q = 1$ (Section 4.3). The approximations of Section 4.3 are

more precise than the approximations of Sections 4.1 and 4.2.

2 ALGORITHM

2.1 Informal description of the algorithm

Let $x_1, x_2, \ldots, x_{\mathbb{T}}$ be a time series with $\mathbb{T} \leq \infty$. Let us choose two integers: the window width N ($N \leq \mathbb{T}$), and the lag parameter M ($M \leq N/2$). Also, set $K = N - M + 1$.

For each suitable $n \geq 0$ we consider the time interval $[n+1, n+N]$ and construct the trajectory matrix (which will be called *base matrix*)

$$\mathbf{X}^{(n)} = (x_{n+i+j-1})_{i,j=1}^{M,K} = \begin{pmatrix} x_{n+1} & x_{n+2} & \ldots & x_{n+K} \\ x_{n+2} & x_{n+3} & \ldots & x_{n+K+1} \\ \vdots & \vdots & \vdots & \ddots \\ x_{n+M} & x_{n+M+1} & \ldots & x_{n+N} \end{pmatrix}. \qquad (1)$$

The columns of $\mathbf{X}^{(n)}$ are the vectors $X_j^{(n)}$ $(j = 1, \ldots, K)$, where

$$X_j^{(n)} = (x_{n+j}, \ldots, x_{n+M+j-1})^T, \quad j \geq -n+1\,.$$

For each $n = 0, 1, \ldots$ we define the lag-covariance matrix $\mathbf{R}_n = \mathbf{X}^{(n)}\big(\mathbf{X}^{(n)}\big)^T$. The SVD of $\mathbf{R}_n$ gives us a collection of M eigenvectors, and a particular group I of $l < M$ of them determines an l-dimensional subspace $\mathcal{L}_{n,I}$ of the M-dimensional space $\mathbb{R}^M$ of vectors $X_j^{(n)}$.

We denote the l eigenvectors that form the basis of the subspace $\mathcal{L}_{n,I}$ by $U_{i_1}, ..., U_{i_l}$ and the sum of squares of the (Euclidean) distances between the vectors $X_j^{(n)} (j = p+1, \ldots, q)$ and this l-dimensional subspace by $\mathcal{D}_{n,I,p,q}$ (the choice of p and $q = p + Q$ is discussed in Section 2.3.2). The matrix with columns $X_j^{(n)}$ $(j = p+1, \ldots, q)$ is called *test matrix*; the location of the base and test matrices is depicted in Figure 1.

Since the eigenvectors of $\mathbf{R}_n$ are orthonormal, the squared Euclidean distance between any vector $Z \in \mathbb{R}^M$ and the subspace $\mathcal{L}_{n,I}$ spanned by the l eigenvectors $U_{i_1}, \ldots, U_{i_l}$, is just

$$||Z||^2 - ||U^T Z||^2 = Z^T Z - Z^T U U^T Z\,,$$

where $||\cdot||$ is the usual Euclidean norm and U is the $(M \times l)$-matrix with columns $U_{i_1}, \ldots, U_{i_l}$. It is also the difference between the squared norms of the vector Z and the projection of Z to the space $\mathcal{L}_{n,I}$. The squared distance

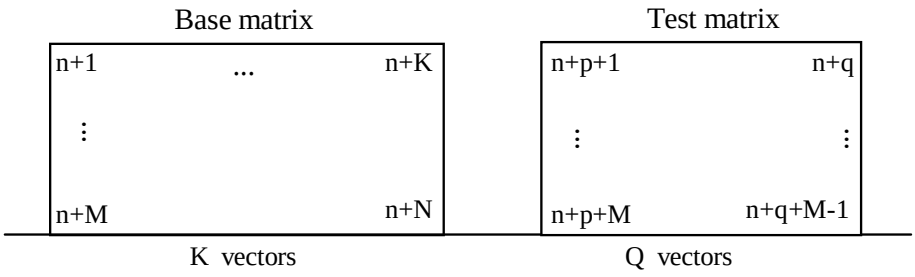

Figure 1: Construction of the base and test matrices.

$\mathcal{D}_{n,I,p,q}$ is the sum of these differences for the vectors $X_j^{(n)}$ constituting the test matrix. That is,

$$\mathcal{D}_{n,I,p,q} = \sum_{j=p+1}^{q} \left((X_j^{(n)})^T X_j^{(n)} - (X_j^{(n)})^T U U^T X_j^{(n)} \right). \tag{2}$$

If a change in the mechanism generating x_t occurs at a certain point τ, then we expect that the vectors $X_j = X_{j-n}^{(n)}$ with $j > \tau$ lie further away from the l-dimensional subspace $\mathcal{L}_{n,I}$ than the vectors X_j with $j \le \tau$. This means that we expect that as n changes, the sequence $\mathcal{D}_{n,I,p,q}$ starts growing somewhere around $\hat{n}$ such that $\hat{n}+q+M-1=\tau$. (This value $\hat{n} = \tau–q–M+1$ is the first value of n such that the test sample $x_{n+p+1}, \ldots, x_{n+q+M-1}$ contains a point with a change.) This growth continues for some time; the expected time of the growth depends on the duration of change and the relations between p, q and N. In a particular case when $p = N$ and $Q = q - p \le M$ and for an abrupt single change, the sequence $\mathcal{D}_{n,I,p,q}$ stops growing after Q iterations, around the point $n = \tau - p - M$. Then during the following $M - Q$ iterations one would expect reasonably high values of this sequence, which must be followed by its decrease to, perhaps, a new level. (This relates to the fact that the SSA decomposition should incorporate the new signal at the intervals $[n + 1, n + N]$ with $n \ge \tau - M$.) See Section 2.4 for more discussions.

The detection statistics are:

- $\mathcal{D}_{n,I,p,q}$, the sum of squared Euclidean distances between the vectors $X_j^{(n)}$ $(j=p+1, \ldots, q)$ and the $l-$dimensional subspace $\mathcal{L}_{n,I}$ of $\mathbb{R}^M$;
- the normalized sum of squared distances (the normalization is made

with respect to the number of elements in the test matrix);

$$\tilde{\mathcal{D}}_{n,I,p,q} = \frac{1}{M(q-p)}\mathcal{D}_{n,I,p,q}\,;$$

- $S_n = \tilde{\mathcal{D}}_{n,I,p,q}/\upsilon_n$.

Here υ_j is an estimate of the normalized sum of squared distances $\tilde{\mathcal{D}}_{j,I,p,q}$ at the time intervals $[j+1, j+N]$ where the hypothesis of no change can be accepted. We suggest to use $\upsilon_n = \tilde{\mathcal{D}}_{\bar{n},I,0,K}$, where $\bar{n}$ is the largest value of $j < n$ such that the null hypothesis of no change in the interval $[j+1, j+N]$ has been accepted. S_n is the squared distance normalized to the number of elements in the test and base matrices and to the variance of the residuals (which are associated with noise); this statistic is shown in graphs.

The decision rule in the algorithm we propose is to announce a change if for some n

$$S_n \geq H, \tag{3}$$

where H is a fixed threshold.

2.2 Formal description of the algorithm

Let $x_1, x_2, \ldots$ be a time series and N, M, l, p and q be fixed integers so that $0 \leq l < M \leq N/2$ and $0 \leq p < q$. The proposed change-point detection algorithm is as follows.

For each $n = 0, 1, \ldots$ we compute:

- *the base matrix $\mathbf{X}^{(n)}$, see* (1),
- *the lag-covariance matrix $\mathbf{R}_n = \mathbf{X}^{(n)}(\mathbf{X}^{(n)})^T$,*
- *the SVD of $\mathbf{R}_n$,*
- *$\mathcal{D}_{n,I,p,q}$, see* (2)*, the sum of the squared Euclidean distances between the vectors $X_j^{(n)}$ $(j = p+1, \ldots, q)$ and the l-dimensional subspace $\mathcal{L}_{n,I}$, and*
- *S_n, the normalized squared distance.*

Large values of $\mathcal{D}_{n,I,p,q}$ and S_n indicate that there is a change in the structure of the time series. If for some $n > N/2$ the inequality (3) holds, then a change in the structure of the time series is announced to have happened at about the point $\hat{\tau} = \hat{n} + q + M - 1$. Here $\hat{n}$ is the iteration number such that the statistic S_n (or $\mathcal{D}_{n,I,p,q}$) has started to grow the last time before reaching the threshold.

2.3 Choice of parameters

Significant changes in the time series structure will be detected for any reasonable choice of parameters. To detect small changes in noisy series some careful tuning of parameters may be required. Let us make some recommendations concerning such a tuning.

2.3.1 Window width N

The choice of N depends on the kind of structural changes we are looking for. A general rule is to choose N reasonably large. However, if we allow small gradual changes in the time series then we could not take N very large. Also, structural changes should not happen too often; ideally, at most one change may occur in any subseries of length N. If N is too large, then we can either miss or smooth out the effects of changes in our time series.

Alternatively, for small N precision of the SSA expansion can be poor; this would cause a haphazard behaviour of the moving squared distances $\mathcal{D}_{n,I,p,q}$. As a consequence, we may have high frequency of false alarms; also, an outlier can be recognized as a structural change.

2.3.2 Length and location of the test sample: p, q

A general recommendation is to choose $p \geq K$; this makes the columns of the base and test matrices different. If $p \geq N = M+K-1$, then the base and test matrices consist of different elements. This choice of p is reasonable if the delay time between the change-point and the moment of its detection permits such a choice.

Numerical simulations show that the choice $Q = q-p = 1$ is often very reasonable and even optimal, see Moskvina and Zhigljavsky (2003). In this case the squared distance $\mathcal{D}_{n,I,p,q}$, which is a weighted sum of squares of residuals, becomes an ordinary (unweighted) sum of squares of residuals.

To get a smoother behaviour of the test statistics $\mathcal{D}_{n,I,p,q}$ one may select $Q > 1$. If Q becomes too large, then the behaviour of $\mathcal{D}_{n,I,p,q}$ becomes too smooth (which makes it difficult to detect changes).

Below, we always assume that $Q \leq M$. However, most results of Sections 3 and 4 can be easily reformulated for the case $Q > M$; for doing this we only need to make the substitution $M \leftrightarrow Q$ in the related formulae.

2.3.3 Parameters of SSA algorithm: lag M and group I

To choose values of the lag M and the group I of indices of the eigenvectors, we have to follow standard SSA recommendations. For an extensive discussion of this problem we refer to Golyandina et al. (2001), pp. 44–78.

If N is not very large, which should be regarded as the most interesting case in practice, by default we choose $M = \lfloor N/2 \rfloor$ and $I = \{1, \ldots, l\}$, where l is such that the first l components describe well the signal and the lower $M - l$ components correspond to noise.

To choose l, visual inspection of the SSA decomposition of the whole series and some large parts of the series before applying the change-point detection algorithms is advised. If l is too small (underfitting), then we miss a part of the signal and therefore we can miss a change (the change may occur in the underestimated components). Alternatively, if l is too large (overfitting), then we approximate a part of noise together with the signal and therefore finding a change in the signal becomes more difficult.

There is also an automatic way of choosing l (such a recommendation is popular in SSA literature): largest l eigenvalues are supposed to be separated from the smallest $M - l$ ones by the largest (in a suitable sense) gap in the ordered set of eigenvalues of the lag-covariance matrix.

2.3.4 Normalization constant υ_n

The suggested normalization constant υ_n is a consistent estimate of σ^2, the variance of the noise under the null hypothesis model, see Section 3.2. Theoretically, any other consistent estimate of σ^2 may be used as well, see Section 3.5.

2.4 Numerical examples

To illustrate applications of the algorithm, let us consider three numerical examples. In the first two the data was simulated so that $N = 400$, $x_t = z_t + e_t$ $(t = 1, \ldots, 400)$, where z_t is the signal and the e_t are i.i.d.r.v., $e_t \sim N(0, 1)$ for $t = 1, \ldots, 400$ (white noise); the change-point was at $\tau = 200$.

Example 1, see Figure 2.1(a,b):

$$z_t = \begin{cases} 1.5 \sin(0.2t) & \text{for} \quad 1 \le t \le 200 \\ 1.5 \sin(0.3t) & \text{for} \quad 201 \le t \le 400\,. \end{cases}$$

Example 2, see Figure 3.2(a,b):

$$z_t = \begin{cases} -0.96z_{t-1} + z_{t-2} - 0.5z_{t-3} + 0.97z_{t-4} & \text{for} \quad 5 \le t \le 200 \\ -0.96z_{t-1} + z_{t-2} - 0.7z_{t-3} + 0.97z_{t-4} & \text{for} \ \ 201 \le t \le 400 \end{cases}$$

with $z_1 = 0, z_2 = 8, z_3 = 6, z_4 = 4$.

In these two examples the change-point is not obviously seen in the graphs. Example 2 is particularly difficult and the success of the proposed change-point detection algorithm can only be explained by the fact that the model (6) is very suitable for the corresponding time series. In Example 1 the model (6) is also suitable but the signals are simpler. In both examples we have applied the following three versions of the algorithm:
(A1) $N = 80,\ M = 40,\ p = 41,\ q = 81$,
(A2) $N = 80,\ M = 40,\ p = 80,\ q = 120$, and
(A3) $N = 80,\ M = 40,\ p = 80,\ q = 81$.
The values of l are: $l = 2$ in Example 1 and $l = 4$ in Example 2. (In both cases these values have been automatically chosen by the software.)

In Figures 2 (c), 3 (c) we plot the test statistics $S_{n'} = S_{n+q+M-1} = \tilde{\mathcal{D}}_{n,I,p,q}/\tilde{\mathcal{D}}_{\bar{n},I,0,K}$, see Section 2.1. In plotting the detection statistics we use the index $n' = n+q+M-1$ to align the expected point of increase of the statistics and the change-point, if there is one. For $n' < \tau = 200$ (while no change occurred) the values of $S_{n'}$ should be close to 1. The corresponding values of n' are in the range $[N + M, \tau] = [120, 200]$ for (A1) and (A3) and $[N+2M-1, \tau] = [159, 200]$ for (A2). Then the values of $S_{n'}$ are expected to grow and reach their highest values for n' around $\tau + M = 240$. After this, the values of $S_{n'}$ are stabilizing at perhaps another level for $n' > \tau + q + M$ ($n' > 320$, $n' > 360$ and $n' > 320$ for (A1), (A2) and (A3), respectively). This is what we roughly see in Figures 2(c) and 3(c).

Example 3, see Figure 4.
The series is a two-dimensional series with no signal and independent in time errors $\{e_t^{(1)}, e_t^{(2)}\}$ such that

$$\begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \end{pmatrix} \sim \begin{cases} N\left(0,\ \sigma^2 I_2\right) & \text{for} \ \ t = 1, \ldots, 200 \\ N\left(0,\ \sigma^2 \Sigma\right) & \text{for} \ \ t = 201, \ldots, 400, \end{cases} \tag{4}$$

where $\sigma^2 = 1$,

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \ \text{ and } \ \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

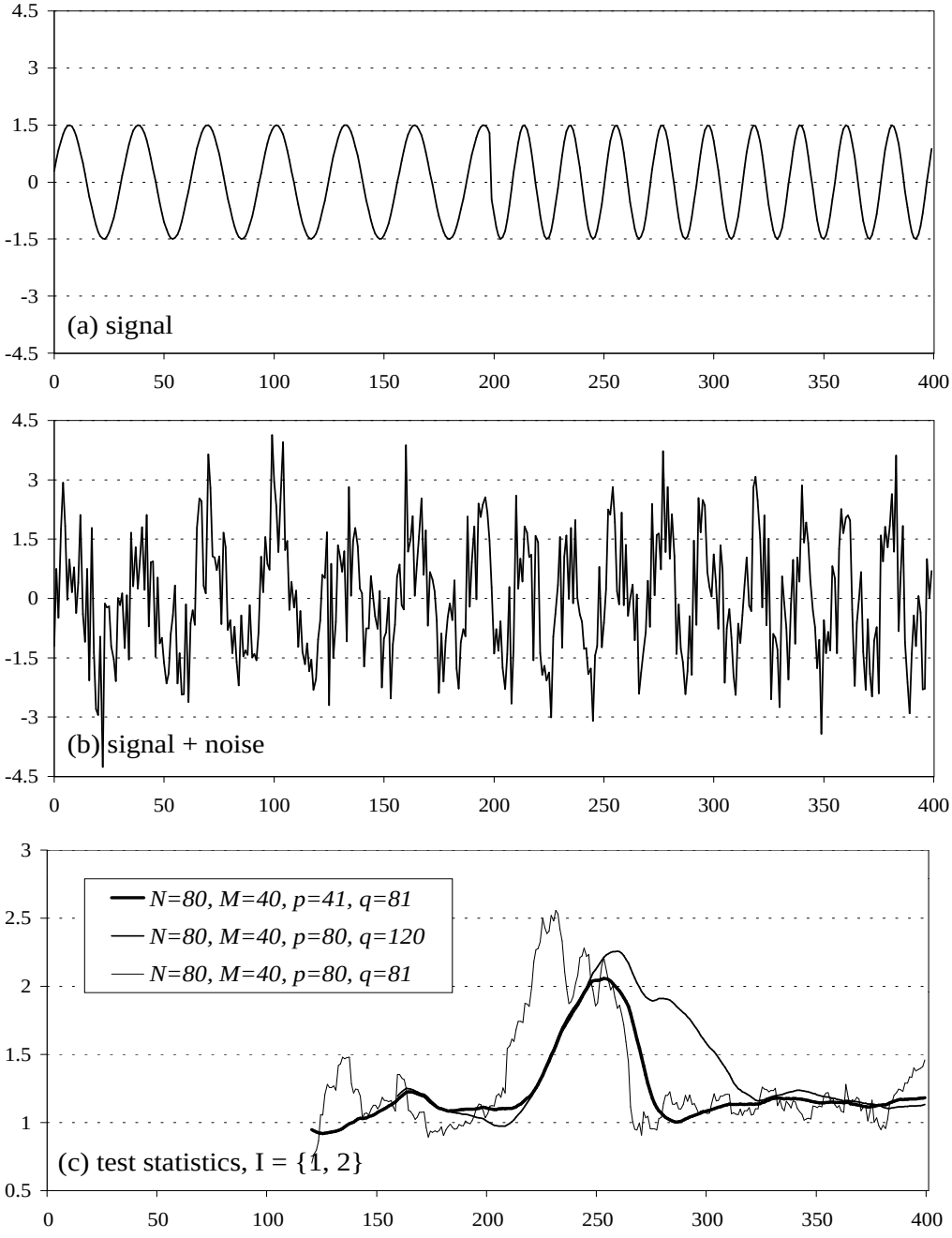

Figure 2: Model of Example 1.

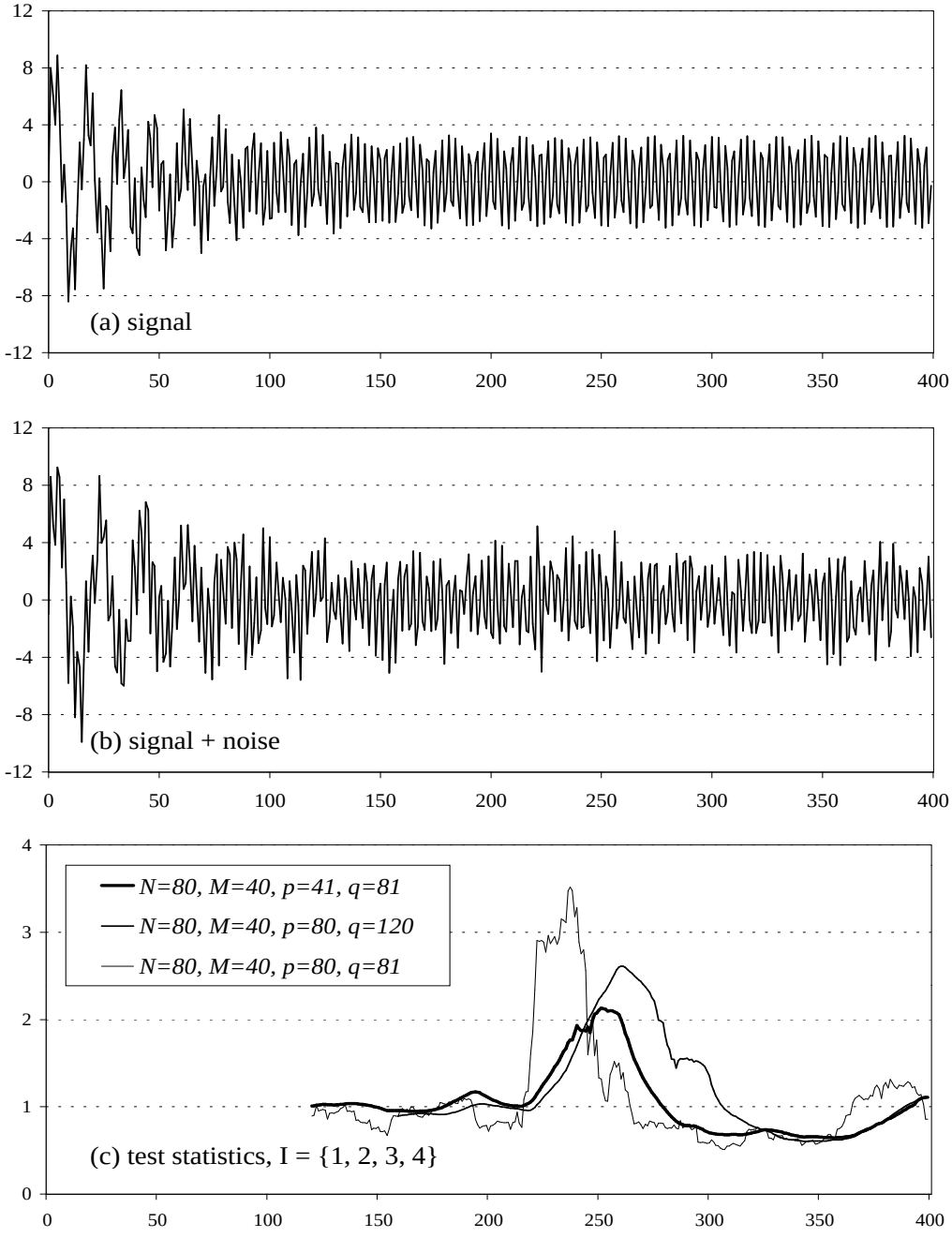

Figure 3: Model of Example 2.

The individual series $e_t^{(j)}$ $(j = 1, 2)$ do not have changes in their probabilistic structure, see Figure 4 (a,b); the change occurs in the correlation structure of the series. To detect this change we consider the sum $e'_t = e_t^{(1)} + e_t^{(2)}$, see Figure 4 (c), and apply three change-point detection algorithms to this series.

The CUSUM test with the detection statistic

$$g'_k = \sum_{i=1}^{k} \left(e_i^{(1)} + e_i^{(2)} \right)$$

and normalized moving sum test with lag $m = 100$ and the statistic

$$\tilde{g}'_k = \frac{1}{m} \sum_{i=k+1}^{k+m} \left(e_i^{(1)} + e_i^{(2)} \right)$$

do not reflect the change, see Figure 4 (d,e). However, the moving sum of squares

$$g_k = \frac{1}{2m\sigma^2} \sum_{i=k+1}^{k+m} \left(e_i^{(1)} + e_i^{(2)} \right)^2 \tag{5}$$

$(m = 100)$ does react to the change, see Figure 4 (f). Note that this algorithm corresponds to the version of the algorithm of Section 2.2 with $M = 100,\ p = 100,\ q = 101$.

The result of this example can be explained as follows. The change-point model (4) is reduced to a change in variance for the series $e_i^{(1)} + e_i^{(2)}$ $(i = 1, \ldots, 400)$ which is a sequence of independent normal r.v. with zero mean and variances $2\sigma^2$ for $i \le 200$ and $3\sigma^2$ for $i > 200$. It is, however, well-known (see, for example, Basseville and Nikiforov, 1993), that the likelihood ratio statistic for this problem is the sum of squares of $e_i^{(1)} + e_i^{(2)}$; therefore, the moving sum of squares (5) is a very natural (and close to the best possible) change-point detection statistic in this case.

3 STATISTICAL CONSIDERATIONS

3.1 SSA rationale

The proposed algorithm can hardly be considered as an automatic tool for detecting changes, it is rather a tool providing bricks for model building and helping to see heterogeneities in the original series. However, under certain

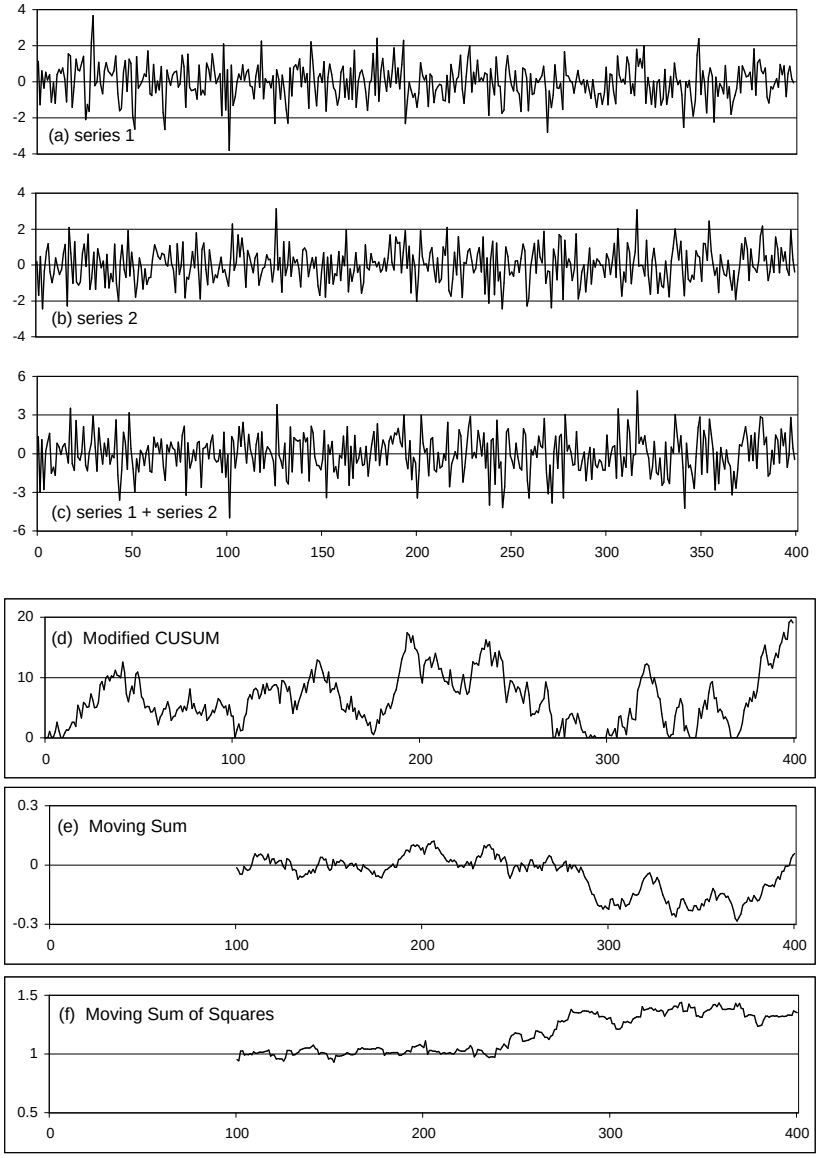

Figure 4: Model of Example 3.

conditions, which asymptotically hold under fairy general assumptions concerning the underlying time series, the algorithm may also be considered as a proper statistical procedure; this can be used for justifying the choice of the threshold H.

The underlying assumption of the SSA technique in general and the proposed change-point detection algorithm in particular is the assumption that the initial time series x_t is well approximated by a series z_t satisfying a finite-difference equation of reasonably small order or, which is equivalent, by a process of the form

$$z_t = \sum_k \alpha_k(t) e^{\mu_k t} \sin(2\pi\omega_k t + \varphi_k),$$

(where $\alpha_k(t)$ are polynomials in t, μ_k, ω_k and φ_k are arbitrary parameters) with small number of terms. That is, we assume that

$$x_t = z_t + e_t\,, \tag{6}$$

where e_t is a noise process and z_t satisfies a finite-difference equation

$$z_t = a_1 z_{t-1} + \ldots + a_d z_{t-d} \tag{7}$$

with $d < M$, some coefficients $a_1, \ldots, a_d$ and some initial conditions. The noise is any aperiodic series; it can be either random or deterministic, but it must have the property that its approximation by solutions of finite-difference equations is poor. (White noise certainly satisfies this assumption.)

Application of SSA with lag M at time intervals $[n+1, n+N]$ approximately recovers the model (6). As the SSA decomposition we obtain

$$x_t = z_t^{(n)} + e_t^{(n)}, \tag{8}$$

where $z_t^{(n)}$ is the SSA approximation for z_t, the solution of (7).

For a properly made SSA decomposition, the component $z_t^{(n)}$ in (8) can be identified as a trend of the original series plus a sum of a few oscillatory components (reflecting, for example, seasonality); the residuals $e_t^{(n)}$ can often be associated with noise. An oscillatory series is a periodic or quasi-periodic series which can be either pure or amplitude-modulated. The trend of a series is, roughly speaking, a slowly varying additive component of the series with all oscillations removed.

Note that no parametric model for the components in (8) is needed and these components are produced by the series itself. Thus, when analyzing

real-life series with the help of SSA, one can hardly hope to obtain $z_t^{(n)}$ as exact periodical series or linear trend, for example, even if this periodical components or linear trend are indeed present in the series. This is an influence of noise and a consequence of the non-parametric nature of the method. Often, however, we can get a very good approximation to these series, see Golyandina et all (2001).

In the ideal situation the components in (8) must be 'independent'. Achieving 'independence' (or 'separability') of the components $z_t^{(n)}$ and $e_t^{(n)}$ in the SSA decomposition (8) is of prime importance in SSA. One of the characteristics of separabilty is the so-called **w**-correlation between series, which for series z_t and e_t is defined as

$$\mathrm{Corr}_w(z_t, e_t) = \frac{\sum_t w_t z_t e_t}{\left(\sum_t w_t z_t^2 \sum_t w_t e_t^2\right)^{1/2}}$$

where the weight function $w_t = w_{M,p,q}(t)$ is defined below in (10).

If z_t satisfies a finite-difference equation (7) and the noise e_t is an ergodic random process with finite variance, then asymptotically, as N and $M \to \infty$, z_t is weakly asymptotically separable from e_t on the intervals $[n+1, n+N]$ implying, for example, that $\mathrm{Corr}_w(z_t^{(n)}, e_t^{(n)}) \to 0$, see Corollary 6.1 in Golyandina et al. (2001). There are also other conditions guaranteeing the asymptotic separability of z_t from e_t, see Chapter 6 in the above reference.

3.2 Null hypothesis

In studying statistical properties of the proposed change-point detection algorithm we assume the following null hypothesis $\mathbb{H}_0$:

(i) the model (6) is valid and there is no change in parameters of the finite difference equation (7),

(ii) $z_t^{(n)} = z_t$ for all n and t,

(iii) M and $\mathbb{T}$ tend to infinity in such a way that there exists the limit $\lim \mathbb{T}/M < \infty$,

(iv) $e_t = e_t^{(n)}$ is a sequence of i.i.d.r.v. with finite forth moment.

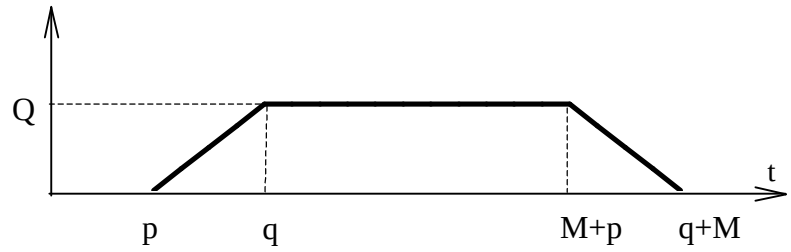

Figure 5: Function $w_{M,p,q}(t)$

3.3 The detection statistic as a moving quadratic form

Under the null hypothesis, in the change-point detection algorithm we have at iteration n

$$\mathcal{D}_{n,I,p,q} = \sum_t w_{M,n+p,n+q}(t) e_t^2 \,, \tag{9}$$

where, see Figure 5,

$$w_{M,p,q}(t) = \begin{cases} t-p & \text{for } p < t \le p+Q, \\ Q & \text{for } p+Q < t \le p+M, \\ p+M+Q-t & \text{for } p+M < t < p+M+Q, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The form of the weight function $w_{M,p,q}(t)$ is related to the structure of the trajectory matrix (1), where x_{n+1} appears once, x_{n+2} – twice, and so on.

Obviously, (9) is a quadratic form $e^T B e$, where $e = (e_1, e_2, \ldots, e_N)^T$ and B is a diagonal matrix with diagonal elements $B_{tt} = w_{M,n+p,n+q}(t)$. The first two moments of this quadratic form can easily be calculated:

$$E\mathcal{D}_{n,I,p,q} = \sigma^2 MQ, \;\; \mathrm{var}(\mathcal{D}_{n,I,p,q}) = \frac{1}{3} Q(\mu_4 - \sigma^4)(3MQ - Q^2 + 1) \,, \tag{11}$$

where $\sigma^2 = Ee_i^2$ and $\mu_4 = Ee_i^4$, the second and the forth moments of the error distribution. In the case when the errors e_i are normal $N(0, \sigma^2)$ we have $\mu_4 = 3\sigma^4$. In this case the distribution of the quadratic form $\mathcal{D}_{n,I,p,q} = e^T B e$ can be thought of as a modification of the χ^2-distribution for the weight function (10); this distribution is studied in Moskvina (2000); it can also be considered as a particular case of the distribution (3.3.1.3) in Richter (1992).

Using the Central Limit Theorem we obtain asymptotically, as $M \to \infty$,

$$\xi_n = \frac{\mathcal{D}_{n,I,p,q} - E\mathcal{D}_{n,I,p,q}}{\sqrt{\mathrm{var}(\mathcal{D}_{n,I,p,q})}} \sim N(0,1)\,. \tag{12}$$

We could have ignored the dependence structure of the sequence of squared distances $\mathcal{D}_{n,I,p,q}$ and use either the asymptotic normality (12) alone or the limiting extreme value distribution to choose the threshold H. We, however, adopt another approach, see Section 3.6, which is based on approximating the sequence $\mathcal{D}_{n,I,p,q}$ by a continuous time random process. To do that, we first need to study the correlations between $\mathcal{D}_{n,I,p,q}$ and $\mathcal{D}_{n+\nu,I,p,q}$ for $\nu > 0$. This is done in Section 4.

3.4 Significance level

As a quality characteristic of change-point detection algorithms we consider the maximum probability of false alarm on time intervals of given length rather than the expected run length, which is more standard in sequential change-point detection theory. As discussed in Lai (1995), the former criterion is more natural for the detection statistics like a moving sum (our detection statistic is the moving weighted sum of squares). According to Bakache and Nikiforov (2000) the corresponding approach can be called 'reliable detection'.

More specifically, assume that we have a change-point detection algorithm, which announces a change if at some iteration n we have $g_n \geq H$, where g_n is a detection statistic and H is a threshold. The maximum probability of false alarm is then defined as

$$P(\mathcal{T}, H, g_n) = \sup_{k \geq 0} \Pr\{g_n \geq H \text{ for at least one } n = k+1, \ldots, k+\mathcal{T} \,|\, \mathbb{H}_0\}, \tag{13}$$

where $\mathbb{H}_0$ is the null-hypothesis of no change and $\mathcal{T}$ is the length of the time interval where we monitor the false alarm. Supremum over k in (13) disappears (that is, all the probabilities inside the supremum become equal) if the statistics g_n form a stationary sequence under the null hypothesis; this is the case in our study. Therefore, without loss of generality, we can assume that $k = 0$ and there is no supremum in (13).

The value of $\mathcal{T}$ can be an arbitrary integer between 0 and the maximum possible value which in our notation is $\mathbb{T}-M-q+1$ (see Figure 1). If $\mathcal{T}$ has its maximum possible value (that is, $\mathcal{T} = \mathbb{T}-M-q+1$), then the probability (13) is the significance level of the change-point detection algorithm. We

shall assume (without loss of generality) that this is true and refer to (13) as the significance level of the change-point algorithm.

We do not consider the power function of the algorithm in this study. The problem of approximating the power function is more difficult; it also depends on the kind of alternative hypotheses we consider. Note that a classification of single changes in the model determined by (6) and (7) can be found in Section 3.2 of Golyandina et al. (2001).

3.5 Standardization

Rather than using the direct expression (13) for the probability $P_g(\mathcal{T}, H)$ it is usually more convenient to standardize the detection statistic g_n first; that is, to pass from the sequence of g_n to

$$\xi_n = (g_n - Eg_n)/\sqrt{\mathrm{var}(g_n)}\,. \tag{14}$$

If g_n forms a stationary series, then $\mu = Eg_n$ and $\delta^2 = \mathrm{var}(g_n)$ do not depend on n and we then have

$$P(\mathcal{T}, H, g_n) = P(\mathcal{T}, h, \xi_n) = \Pr\{\max_{1\le n\le \mathcal{T}} \xi_n \ge h \mid \mathbb{H}_0\} \tag{15}$$

with $h = (H - \mu)/\delta$ (alternatively, $H = \mu + \delta h$).

The two most important special cases of g_n are as follows, see Section 2.1.

(i) For $g_n = \mathcal{D}_{n,I,p,q}$ we have the expressions (11) for μ and δ^2. Thus, for $g_n = \mathcal{D}_{n,I,p,q}$ the thresholds H and h in (15) are related through

$$H = \sigma^2 MQ + h\sqrt{\frac{Q(\mu_4 - \sigma^4)(3MQ + 1 - Q^2)}{3}} \tag{16}$$

(ii) Let g_n be $S_n = \tilde{\mathcal{D}}_{n,I,p,q}/\upsilon_n$, where υ_n is a consistent (as $M \to \infty$) estimate of $\sigma^2 = Ee_n^2$. The expressions (11) imply

$$E\tilde{\mathcal{D}}_{n,I,p,q} = \sigma^2 \ \text{ and } \ \mathrm{var}(\tilde{\mathcal{D}}_{n,I,p,q}) = \frac{(\mu_4 - \sigma^4)(3MQ + 1 - Q^2)}{3\sigma^4 M^2 Q^3}\,.$$

Assume that $M \to \infty$ and e_n are Gaussian r.v. implying $\mu_4 = 3\sigma^4$ (similar formulae hold for other error distributions). In view of the asymptotic normality (12) and the celebrated Slutsky's theorem (see, for example, property (b), page 122 in Rao, 1973), which allows us to substitute σ^2 by υ_n preserving the asymptotic distribution, we obtain that as $M \to \infty$ the statistics $g_n = \tilde{\mathcal{D}}_{n,I,p,q}/\upsilon_n$ are asymptotically

normal. We also obtain 1 and $2(3MQ+1-Q^2)/3M^2Q^3$ as the limiting (as $M \to \infty$) values in (15) for μ and δ^2, respectively. Thus, for $g_n = \tilde{\mathcal{D}}_{n,I,p,q}/\upsilon_n$ in case of large M and normal e_n we can use the relationship

$$H = 1 + \sqrt{2}\frac{\sqrt{M - Q/3}}{MQ}h \tag{17}$$

between H and h in (15). If the distribution of e_n is not normal then $\sqrt{2} = \sqrt{\mu_4/\sigma^4 - 1}$ in the right-hand side of (17) must be replaced by the corresponding value (for example, by $2/\sqrt{5}$ in case of uniform distribution).

Note that asymptotically (as $M \to \infty$) the cases (i) and (ii) above lead to the same standardized sequence (14).

3.6 Continuous time approximations

The probabilities $P(\mathcal{T}, h, \xi_n)$ of (15) are $\mathcal{T}$-dimensional integrals and are difficult to compute. In Section 5 we shall use several continuous-time approximations to these probabilities. We shall assume that $M \to \infty$ and pass from the time series ξ_n ($n = 1, \ldots, \mathcal{T}$) to a continuous-time process ξ_t, $t \in [0, T]$, for some T depending on $\mathcal{T}$, M and perhaps Q. Like the series ξ_n, the process ξ_t will be standardized so that $E\xi_t = 0$ and $E\xi_t^2 = 1$ for all t. Also, the process ξ_t will be Gaussian and stationary with some autocorrelation function $R(s) = E\xi_t\xi_{t+s}$.

The probability $P(\mathcal{T}, h, \xi_n)$ will then be approximated by $P(T, h, \xi_t)$, the probability of reaching the threshold h by the process ξ_t on the interval $[0, T]$:

$$\begin{aligned} P(\mathcal{T}, h, \xi_n) &\simeq P(T, h, \xi_t) = \Pr\left\{\max_{0\le t\le T} \xi_t \ge h\right\} \\ &= \Pr\left\{\xi_t \ge h \text{ for at least one } t \in [0,T]\right\}. \end{aligned} \tag{18}$$

Two related characteristics can also be of interest: the probability density function of reaching the threshold h by the process ξ_t for the first time

$$q(t, h, \xi_t) = \frac{d}{dt}P(t, h, \xi_t), \quad 0 < t < \infty, \tag{19}$$

and the average time $\varrho(h, \xi_t)$ until the process ξ_t reaches the threshold h:

$$E(\varrho(h, \xi_t)) = \int_0^\infty tq(t, h, \xi_t)dt = \int_0^\infty tdP(t, h, \xi_t).$$

Note that it is often reasonable to assume that $\mathcal{T}$ is not very large, relative to M. Some approximations of Section 4 for the significance level, however, are reasonable only when $\mathcal{T}$ is much larger than M.

4 CORRELATIONS BETWEEN $\mathcal{D}_{n,I,p,q}$ AND $\mathcal{D}_{n+\nu,I,p,q}$

For fixed p and q the squared distances $\mathcal{D}_{n,I,p,q}$ are functions of n. The index n can be treated as time and thus the sequence $\mathcal{D}_{1,I,p,q}$, $\mathcal{D}_{2,I,p,q}, \dots$ defined in (9) can be considered as a time series. In order to understand the behaviour of this time series (in particular, to obtain approximations for the significance level of the change-point algorithm) we need to understand the behaviour of the correlations $\mathrm{Corr}(\mathcal{D}_{n,I,p,q}, \mathcal{D}_{n+\nu,I,p,q})$ with $\nu \geq 0$. Computation of these correlations is the purpose of the present section.

Without loss of generality we can assume that $n = 0$, $p = 0$ and $Q = q > 0$. Thus, in the rest of this section we shall denote $\mathcal{D}_{n+\nu,I,p,q} = \mathcal{D}_\nu$ to underline the dependence of $\mathcal{D}_{n+\nu,I,p,q}$ on the shift ν.

Let us first consider the case $\nu = 1$. Consider the quadratic forms $\mathcal{D}_0$ and $\mathcal{D}_1$. We can represent them as

$$\mathcal{D}_0 = \sum_{i=1}^{Q-1} i e_i^2 + Q \sum_{i=Q}^{M} e_i^2 + \sum_{i=M+1}^{Q+M-1} (Q+M-i) e_i^2 \quad \text{and} \quad \mathcal{D}_1 = \mathcal{D}_0 - \sum_{i=1}^{Q} e_i^2 + \sum_{i=1}^{Q} e_{M+i}^2 .$$

Using these representations we can easily compute the expectation $E(\mathcal{D}_0 \mathcal{D}_1)$:

$$E(\mathcal{D}_0 \mathcal{D}_1) = E\mathcal{D}_0^2 - Q(\mu_4 - \sigma^4) .$$

This and the formulae (11) for $E\mathcal{D}_0$ and $\mathrm{var}(\mathcal{D}_0)$ give

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_1) = \frac{E(\mathcal{D}_0 \mathcal{D}_1) - (E\mathcal{D}_0)^2}{\mathrm{var}(\mathcal{D}_0)} = 1 - \frac{3}{3MQ - Q^2 + 1} . \qquad (20)$$

Let us now consider the correlations between $\mathcal{D}_0$ and $\mathcal{D}_\nu$ for general $\nu \geq 0$. Note that these correlations (unlike covariances) do not depend on the distribution of errors e_n; this follows from the fact (see, for example, Priestley, 1981) that the spectral density of the moving average process depends only on the weights, which in our case are $w_{M,n+p,n+q}(t)$, see (9).

Thus, without loss of generality we assume that the errors e_t are normally distributed. It allows us to apply the results of Theorem 3.2d.4 in Mathai and Provost (1992). This theorem yields that if $\mathbf{X} \sim N_p(0, \Sigma)$ is p-variate

normal distribution with zero mean and covariance matrix $\Sigma > 0$, and $Q_1 = \mathbf{X}^T A_1 \mathbf{X}$, $Q_2 = \mathbf{X}^T A_2 \mathbf{X}$ are two quadratic forms then

$$\mathrm{Cov}(Q_1, Q_2) = 2\mathrm{tr}(A_1 \Sigma A_2 \Sigma). \tag{21}$$

Of course, this formula and the expression (11) for $\mathrm{var}(\mathcal{D}_0)$ yield the result (20) for $\nu = 1$.

For general $\nu \geq 1$ and $1 \leq Q \leq M$ there are five possible cases of locations of the weight functions w_t that define $\mathcal{D}_0$ and $\mathcal{D}_\nu$. These cases are illustrated in Figure 6.

Using the formula (21) we only need to carefully combine the corresponding terms to derive the correlations $\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu)$ for these five cases, see Moskvina (2001) for details. The result is as follows.

Define the function

$$f(a, b) = a(3ab - a^2 + 1)\,.$$

and note that $3\mathrm{var}(\mathcal{D}_0) = 2\sigma^4 f\,(Q, M)$, see (11). Then

Case 1. $\nu \leq Q, \quad Q{+}\nu \leq M$:

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = 1 - \frac{f\,(\nu, M)}{f\,(Q, M)}\,.$$

Case 2. $\nu \leq Q, \;\; Q{+}\nu \geq M$**:**

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = \frac{f(Q, M{+}\nu) + f(M{-}\nu, Q) - 2f(\nu, Q)}{2f\,(Q, M)}\,.$$

Case 3. $Q \leq \nu \leq M, \;\; Q{+}\nu \leq M$:

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = 1 - \frac{f\,(Q, \nu)}{f\,(Q, M)}\,. \tag{22}$$

Case 4. $Q \leq \nu \leq M, \quad \nu{+}Q \geq M$:

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = \frac{f(M{-}\nu, Q) - f(Q, \nu{-}M)}{2f\,(Q, M)}\,.$$

Case 5. $M \leq \nu < Q{+}M{-}1$**:**

$$\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = \frac{f(\nu, M{+}Q) - f(M{+}Q, \nu)}{2f\,(Q, M)}\,.$$

Clearly, if $\nu \geq Q{+}M{-}1$ then $\mathcal{D}_0$ and $\mathcal{D}_\nu$ are independent implying $\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu) = 0$.

Figure 7 illustrates the behaviour of the autocorrelation function $\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu)$ as a function of ν.

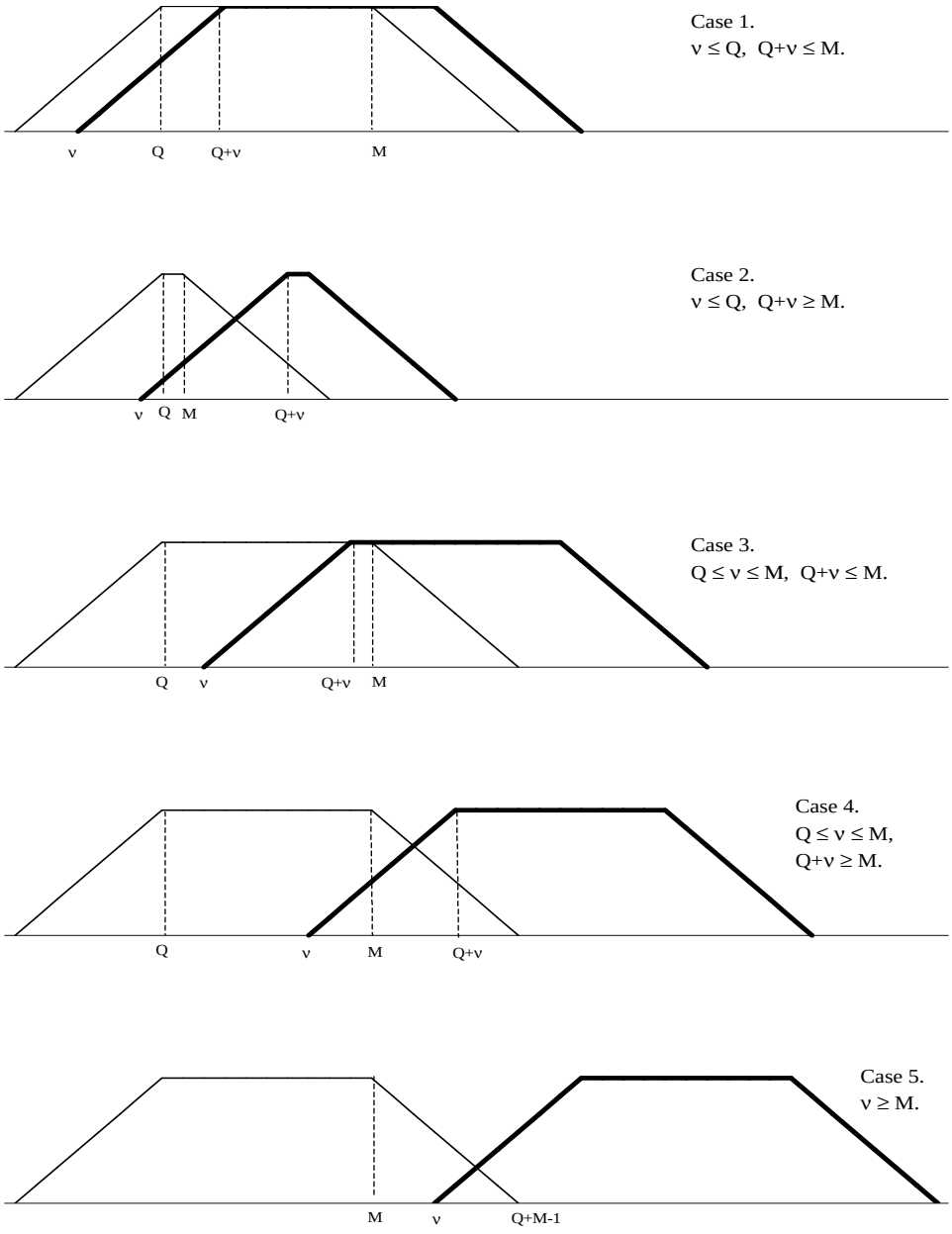

Figure 6: Weight functions for $\mathcal{D}_n$ and $\mathcal{D}_{n+\nu}$ with $\nu \geq 1$. Cases 1-5.

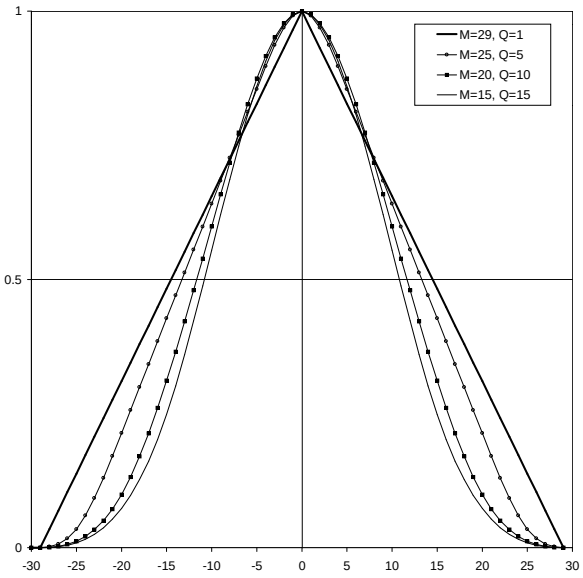

Figure 7: Examples of the autocorrelation function $\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu)$.

5 APPROXIMATIONS FOR THE SIGNIFICANCE LEVEL

For small ν, the behaviour of the autocorrelation function $\mathrm{Corr}(\mathcal{D}_0, \mathcal{D}_\nu)$ as $M \to \infty$ depends on Q. In this section we consider three different approximations to the significance level of the proposed algorithm. These three approximations are valid depending on whether Q is large, small or just equal to 1.

5.1 Smooth covariance functions: large M and Q

Consider the sequence of random variables $\xi_0, \xi_1, \ldots, \xi_{\mathcal{T}}$ defined as

$$\xi_n = \frac{\mathcal{D}_{n,I,p,q} - E\mathcal{D}_{n,I,p,q}}{\sqrt{\mathrm{var}(\mathcal{D}_{n,I,p,q})}} \quad (n = 0, \ldots, \mathcal{T}), \tag{23}$$

see case (i) in Section 3.5.

In view of (20), the correlation between ξ_n and ξ_{n+1} is

$$\mathrm{Corr}(\xi_n, \xi_{n+1}) = 1 - \frac{3}{3MQ - Q^2 + 1}. \tag{24}$$

Assume that both M and Q are large; that is, $M, Q \to \infty$ in such a way that the limit $\lambda = \lim Q/M$ exists and $0 < \lambda \leq 1$. Set $\Delta = 1/\sqrt{MQ}$ and

$$t_n = n\Delta \quad (n = 0, 1, \ldots, \mathcal{T}) \quad \text{so that} \quad t_n \in [0, T] \quad \text{with} \quad T = \mathcal{T}\Delta\,. \tag{25}$$

Define a piece-wise linear continuous-time process $\xi_t^{(M)}$, $t \in [0, T]$, as follows:

$$\xi_t^{(M)} = \frac{1}{\Delta}[(t - t_n)\xi_{n-1} + (t - t_{n-1})\xi_n] \quad \text{for} \quad t \in [t_{n-1}, t_n],\ n = 1, \ldots, \mathcal{T}. \tag{26}$$

The process $\xi_t^{(M)}$ is such that $\xi_{t_n}^{(M)} = \xi_n$ for $n = 0, \ldots, \mathcal{T}$; it is a second-order stationary process in the sense that $E\xi_t^{(M)}$, $\mathrm{var}(\xi_t^{(M)})$ and the autocorrelation function $R_\xi^{(M)}(t, t + k\Delta) = \mathrm{Corr}(\xi_t^{(M)}, \xi_{t+k\Delta}^{(M)})$ do not depend on t. The limiting process ξ_t is stationary Gaussian with some autocorrelation function $R_\xi(t, t + s) = R(s)$ which is illustrated in Figure 7. In the case $\lambda = \lim Q/M > 0$ we have for the autocorrelation function $R(\cdot)$:

$$R'(0-) = R'(0+) = \lim_{M,Q\to\infty} \frac{R(\Delta) - 1}{\Delta} = \lim_{M,Q\to\infty} \frac{-3\sqrt{MQ}}{3MQ - Q^2 + 1} = 0;$$

here we have used the facts that $R(0) = 1$, $\Delta = 1/\sqrt{MQ}$ and $R(\Delta) = 1 - 3/(3MQ - Q^2 + 1)$. We therefore have $R'(0) = 0$.

We similarly obtain

$$R''(0) = \lim_{M,Q\to\infty} \frac{R(\Delta) + R(-\Delta) - 2R(0)}{\Delta^2} = \lim_{M,Q\to\infty} \frac{-6MQ}{3MQ - Q^2 + 1} = -\frac{6}{3 - \lambda}\,. \tag{27}$$

For a Gaussian stationary process ξ_t with $E\xi_t = 0$ and $E\xi_t^2 = 1$ and autocorrelation function $R(\cdot)$ such that $R'(0) = 0$ and $R''(0) < 0$ we can use the following two well-known approximations.

Approximation 1, see Theorem 8.2.7 in Leadbetter, Lindgren and Rootzen (1983):

$$\lim_{T\to\infty} P\left\{ \max_{0\leq t\leq T} \xi_t \leq \underbrace{\frac{x + \log\frac{\sqrt{-R''(0)}}{2\pi}}{\sqrt{2\log T}} + \sqrt{2\log T}}_{h} \right\} = \exp(-e^{-x})\,.$$

Expressing x in terms of h, we obtain

$$\lim_{T\to\infty} P\left\{ \max_{0\leq t\leq T} \xi_t \geq h \right\} = 1 - \exp(-e^{-x}), \tag{28}$$

where $x = x_1 = \gamma(h-\gamma) + c$ with

$$\gamma = \gamma(T) = \sqrt{2\log T} \quad \text{and} \quad c = -\log \frac{\sqrt{-R''(0)}}{2\pi} = -\log \frac{1}{2\pi}\sqrt{\frac{6}{3-\lambda}}\,. \quad (29)$$

Approximation 2, see Cramer (1965):

$$\lim_{T\to\infty} P\left\{ \max_{0\le t\le T} \xi_t \le \underbrace{\sqrt{2\log\mu(T)} + \frac{x}{\sqrt{2\log\mu(T)}}}_{h} \right\} = \exp(-e^{-x}),$$

where

$$\mu(T) = \frac{T\sqrt{-R''(0)}}{2\pi} = \frac{T}{2\pi}\sqrt{\frac{6}{3-\lambda}}\,.$$

We thus obtain (28) with

$$x = x_2 = \sqrt{2\log\mu(T)}\left(h - \sqrt{2\log\mu(T)}\right).$$

We have

$$\sqrt{2\log\mu(T)} = \sqrt{\gamma^2 - 2c} = \gamma - \frac{c}{\gamma} + O\left(\frac{1}{\gamma^3}\right), \quad \gamma\to\infty$$

where γ and c are defined in (29). Therefore, for large T (and, therefore, large γ) we have

$$x_2 \simeq \left(h - \gamma + \frac{c}{\gamma}\right)\left(\gamma - \frac{c}{\gamma}\right) = \underbrace{(h-\gamma)\gamma + c}_{x_1} - \frac{(h-\gamma)c}{\gamma} - \frac{c^2}{\gamma^2}\,.$$

We use this fact to construct another approximation.

Combined approximation: the formula (28) with

$$x = \begin{cases} x_1 - \frac{(h-\gamma)c}{\gamma} - \frac{c^2}{\gamma^2} & \text{for} \quad h \le \gamma - \frac{c}{\gamma}\,, \\ x_1 & \text{for} \quad h \ge \gamma - \frac{c}{\gamma}\,. \end{cases}$$

Of course, asymptotically (as $T\to\infty$) all three approximations give similar results (note that the approximations are guaranteed to work only for large T and h). In practice, however, approximations have to work for values of T and h that are not too large.

A large number of simulations have been performed, see Moskvina (2001) and Moskvina and Zhigljavsky (2003) for details, to assess the quality of

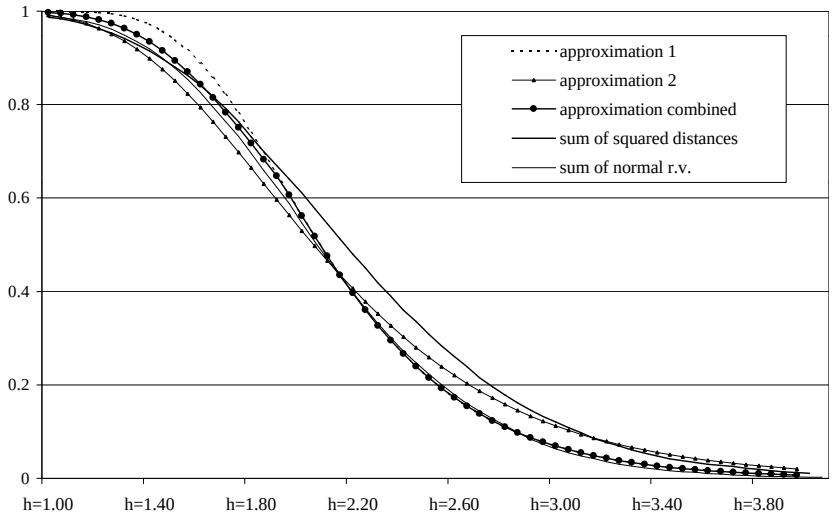

Figure 8: Approximations for the significance level for the weighted sum of normal r.v. and of their squares; smooth covariance functions: $M = 100$, $Q = 100$, $\mathcal{T} = 2000$.

these three approximations and influence of the values of T, M and Q and the distribution of errors e_t on the behaviour of these approximations. Along with the squared distances $\mathcal{D}_{n,I,p,q} = \sum_t w_t e_t^2$, where e_t are i.i.d. normal $N(0,1)$ random variables, we also considered the case when the squares of normal random variables e_t^2 are substituted by the e_t giving the moving sum $\mathcal{D}'_{n,I,p,q} = \sum_t w_t e_t$. In this case the distribution of the sum is exactly normal and we approximate the probability of reaching the threshold for the moving weighted sum of normal r.v. Non-normal distributions for e_t have also been studied.

Figure 8 shows the quality of the three approximations for $\mathcal{D}_{n,I,p,q}$ and $\mathcal{D}'_{n,I,p,q}$ with $M = Q = 100$ and $\mathcal{T} = 2000$ (so that T=20). In each case we performed 100 000 simulations of the standardized moving sums $\mathcal{D}_{n,I,p,q}$ and $\mathcal{D}'_{n,I,p,q}$; the results, presented in Figure 8, are the values of proportions of the cases when the threshold h has been reached.

The simulation results show that the combined approximation is typically the best of the three. For small M and Q and the distributions with long tails the approximations are poor; for large M, Q and T and for finite-support error distribution the approximations are good. For small T (say, $T \leq 10$) all the approximations are poor (this is related to the method of

deriving these approximations which ignores the dependence between high excursions of the process ξ_t).

5.2 Durbin's tangent approximation: large M and small Q

Consider again the sequence of r.v. defined by (23). Unlike in Section 5.1, consider now the asymptotics when $M \to \infty$ but Q is fixed. Set $\Delta = 1/M$, $T = \mathcal{T}\Delta$. Define t_n $(n = 0, 1, \ldots, \mathcal{T})$ as in (25) and consider the piece-wise linear continuous-time process $\xi_t^{(M)}$ defined by (26). The limiting process (as $M \to \infty$) is again some Gaussian second-order stationary process ξ_t with a covariance function $R_\xi(t, t+s) = R(s)$. To apply the approximation below, we shall need the value of

$$\left.\frac{\partial R_\xi(t,s)}{\partial s}\right|_{s=t+} = R(0+).$$

Using (24) and the fact that $\Delta = 1/M$, we have

$$R'(0+) = \lim_{M\to\infty} \frac{R(\Delta) - R(0)}{\Delta} = -\lim_{M\to\infty} \frac{3M}{3MQ - Q^2 + 1} = -\frac{1}{Q}.$$

Describe now the approximation we are going to use in the case $R'(0+) \neq 0$.

Let $\xi(t)$ be a Gaussian random process on $[0, T]$ with $E\xi(t)$=0 and some covariance function $R_\xi(t, s)$; let $h(t)$ be some threshold. One of the most known approximations for (19), the density function $q(t, h, \xi_t)$ of the first passage time, and therefore for (18), the first passage probability $P(T, h, \xi_t)$, is the tangent approximation suggested in Durbin (1985). For $q(t, h, \xi_t)$ this approximation can be written as

$$q(t, h, \xi_t) \simeq b_0(t, h) f(t, h)\,, \tag{30}$$

with

$$f(t,h) = \frac{1}{\sqrt{2\pi R_\xi(t,t)}}\, e^{-\frac{h^2(t)}{2R_\xi(t,t)}}\,, \quad b_0(t,h) = -\frac{h(t)}{R_\xi(t,t)} \left.\frac{\partial R_\xi(s,t)}{\partial s}\right|_{s=t+} - \frac{dh(t)}{dt}\,.$$

In view of (19) the related approximation for the first passage probability $P(T, h, \xi_t)$ is

$$P(T, h, \xi_t) \simeq \int_0^T b_0(t, h) f(t, h) dt\,.$$

In the case when the process is defined by the moving sum of squared distances and $h(t) = h$ is constant, we have

$$b_0(t, h) = -hR'(0+) = \frac{h}{Q}, \quad q(t, h, \xi_t) \simeq \frac{h}{\sqrt{2\pi}Q}\, e^{-h^2/2}$$

and therefore

$$P(T, h, \xi_t) \simeq \frac{hT}{\sqrt{2\pi}Q} e^{-h^2/2}. \tag{31}$$

The quality of the approximation (31) is poor unless h is very large, see Figure 9.

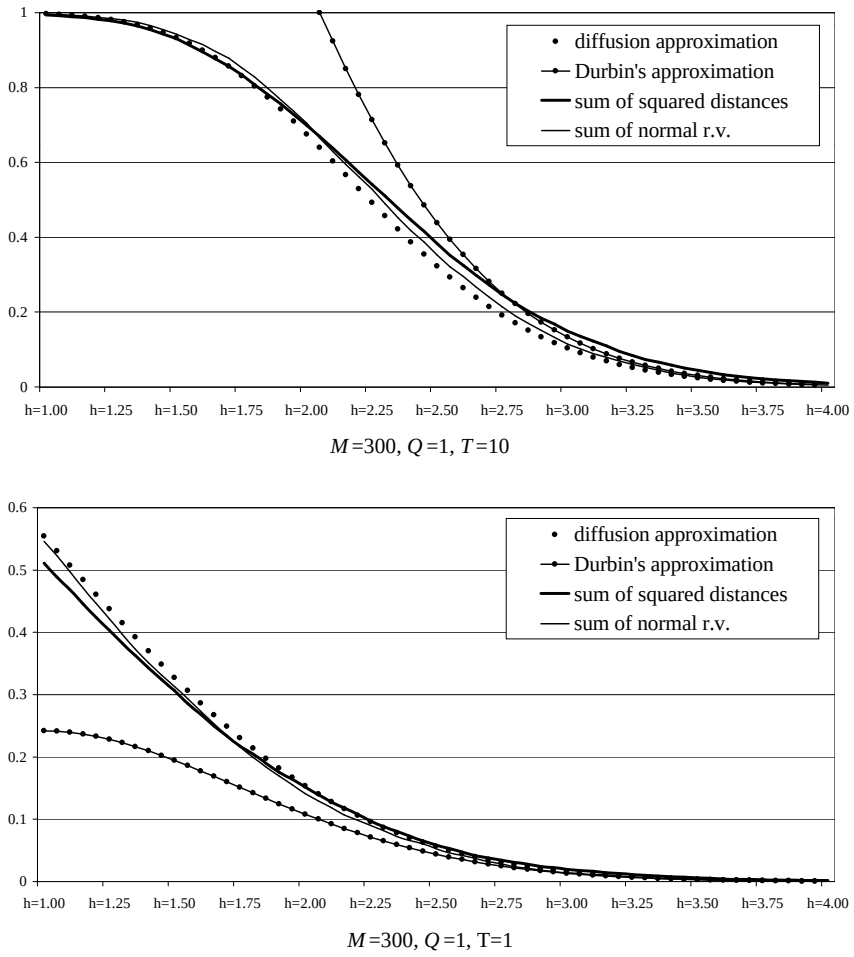

Figure 9: Diffusion and Durbin's approximations for the sum of normal r.v. and their squares.

5.3 Diffusion approximation: $Q = 1$ and large M

In the present section we shall assume that $Q = 1$ meaning that the squared distances $\mathcal{D}_{n,I,p,q}$ are simple (unweighted) sums of squares of errors e_j. For approximating the boundary crossing probabilities $P(\mathcal{T}, h, \xi_n)$ we shall apply the approach used in Zhigljavsky and Kraskovsky (1988) for a different change-point detection problem. The resulting approximation will be called *diffusion approximation.*

Consider a sequence of random variables ξ_n defined in (12), see also case (i) in Section 3.5. Since ξ_n are standardized, we have $E\xi_n = 0$ and $\mathrm{var}(\xi_n) = 1$. To compute correlations $\mathrm{Corr}(\xi_n, \xi_{n+\nu})$ for $\nu \geq 0$, we refer to the case 3 in Section 4. The formula (22) with $Q = 1$ gives

$$\mathrm{Corr}(\xi_n, \xi_{n+\nu}) = 1 - \frac{\nu}{M}\,, \quad 1 \leq \nu \leq M\,, \tag{32}$$

see Figure 7, case $Q = 1$ (of course, (32) could have been easily derived directly without referring to Section 4).

As in Section 5.2 we set $\Delta = 1/M$, $T = \mathcal{T}\Delta$, define t_n $(n = 0, 1, \ldots, \mathcal{T})$ as in (25) and consider the piece-wise linear continuous-time process $\xi_t^{(M)}$ defined by (26). As $M \to \infty$, the sequence of processes $\xi_t^{(M)}$ converges (in the sense of convergence in metric of the space $C[0, T]$) to a limiting process ζ_t, $t \in [0, T]$. The process ζ_t is a stationary Gaussian process with zero mean and the triangular covariance function

$$R(u) = E\zeta_t\zeta_{t+u} = \max\{0,\ 1-|u|\}\,; \tag{33}$$

this is a consequence of (32) and the fact that $E\zeta_t^2 = E\xi_n^2 = 1$ for all t and n.

Therefore, the boundary crossing probabilities $P(\mathcal{T}, h, \xi_n)$ defined in (15) can be approximated by the corresponding probabilities for the process ζ_t with covariance function (33):

$$P(\mathcal{T}, h, \xi_n) \simeq P(T, h, \zeta_t) = \Pr\left\{\sup_{0\leq t\leq T} \zeta_t \geq h\right\}. \tag{34}$$

The problem of exact computation of the boundary crossing probabilities $P(T, h, \zeta_t)$ has been studied in several papers including Mehr and McFadden (1965), Shepp (1966), Shepp (1971), Shepp and Slepian (1976). The results can be summarized as follows:
for $T = 1$

$$P_h = P(1, h, \zeta_t) = 1 - \Phi^2(h) + \frac{1}{\sqrt{2\pi}} h e^{-h^2/2}\Phi(h) + \frac{1}{2\pi} e^{-h^2}\,; \tag{35}$$

more generally, for $T \leq 1$

$$P(T,h,\zeta_t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{h} \Phi\left(\frac{h(T+1) - x(-T+1)}{2\sqrt{T}}\right) e^{-x^2/2} dx + \quad (36)$$

$$+\sqrt{\frac{2}{\pi}} \frac{hT\, e^{-h^2/2}}{(T+1)} \Phi\left(h\sqrt{T}\right) + \frac{\sqrt{T} e^{-h^2/2(T+1)}}{\pi(T+1)}.$$

Here

$$\Phi(h) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{h} e^{-t^2/2} dt \quad \text{and} \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,.$$

For $T > 1$ the expressions for $P(T,h,\zeta_t)$ are complicated and difficult to implement.

An approximation to $P(T,h,\zeta_t)$ for $T > 1$ has been developed in Zhigljavsky and Kraskovsky (1988). The main formula has the form

$$P(T,h,\zeta_t) \simeq P_h + (1 - P_h)(1 - \lambda_h^{T-1})\,, \quad (37)$$

where $P_h = P(1,h,\zeta_t)$ is defined in (35) and

$$\lambda_h = \frac{\Phi(h) + \sqrt{16 - 7\Phi^2(h) - 16P_h}}{4}\,.$$

The quality of this approximation seems to be rather high, even for not very large values of h, see Figure 9. If h is large ($h \to \infty$), we can derive from (36) and (37) a simple approximation

$$P(T,h,\zeta_t) \simeq \frac{hT}{\sqrt{2\pi}} e^{-h^2/2}\,, \quad h \to \infty\,, \quad (38)$$

which is valid for all $T > 0$ (for details, see the derivation of formula (2.51), p.80, in Zhigljavsky and Kraskovsky, 1988).

It is important to note that the large threshold approximation (38) is exactly the Durbin's tangent approximation (31) with $Q = 1$.

The quality of the diffusion and Durbin's approximations is shown in Figure 9 for $Q = 1$, $M = 300$, and $\mathcal{T} = 300$ and 3000 so that $T = 1$ and 10. The plots in this figure demonstrate that the approximations (37) and especially (35) are much more precise for typical h than the approximation (31), which only works when h is very large.

Acknowledgements

The authors are grateful to Prof. Igor Nikiforov (Universite Technologie de Troyes, France) for many useful comments.

References

[1] Bakhache, B. and Nikiforov, I. (2000) Reliable Detection of Faults in Measurement Systems, *International Journal of Adaptive Control and Signal Processing*, **14**, 683-700.

[2] Basseville, M. And Nikiforov, I.V. (1993) *Detection of Abrupt Changes: Theory and Applications*, Prentice Hall, Englewood Cliffs, New Jersey.

[3] Broomhead, D.S., Jones, R. and King, G.P. (1987) Topological Dimension and Local Coordinates From Time Series Data. *Physica* **A**, **20**, L563-L569.

[4] Broomhead, D.S. A and King, G.P. (1986) Extracting Qualitative Dynamics from Experimental Data. *Physica* **D**, **20**, 217-236

[5] Cramer H., (1965) A Limit Theorem for fhe Maximum Values of Certain Stochastic Processes. Theory of Probability and Applications **10**, 137-139.

[6] Durbin, J. (1985) The First–Passage Density of a Continuous Gaussian Process to a General Boundary. *Applied Probability* **22**, 99-122.

[7] Elsner, J. and Tsonis, A. (1996) Singular Spectrum Analysis: a New Tool in Time Series Analysis, *Plenum Press, New York*.

[8] Fraedrich, K. (1986) Estimating the Dimension of Weather and Climate Attractors, *J. Atmos. Sci.*, **43**, 419-432.

[9] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2000) *Analysis of Time Series Structure: SSA and Related Techniques*, London: Chapman And Hall.

[10] Lai T.L. (1995) Sequential Changepoint Detection in Quality Control and Dynamical Systems, *Journal of Royal Statistical Society*, **B**, 613-658.

[11] Leadbetter, M.R., Lindgren, G., And Rootzen, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*, Springer Series in Statistics, Springer-Verlag.

[12] Mathai, A.M., and Provost, S.B. (1992) *Quadratic Forms in Random Variables: Theory and Applications*, Marcel Dekker.

[13] Mehr, C.B., Mcfadden, J.A., (1965) Certain Properties Of Gaussian Processes and Their First Passage Times, *Journal of Royal Statistical Society*, **B**, 505-522.

[14] Moskvina, V. (2000) Distribution of Random Quadratic Forms Arising in Singular-Spectrum Analysis, *Mathematical Communications*, **5**, 161-171.

[15] Moskvina, V. (2001) Application of the Singular Spectrum Analysis for Change-Point Detection in Time Series, *PhD thesis*, Cardiff University, 2001.

[16] Moskvina, V. and Zhigljavsky A.A. (2003) An Algorithm Based on Singular-Spectrum Analysis for Change-Point Detection, *Communication in Statistics. Statistics and Simulations*, **32**, 319-352.

[17] Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Vol. 1, London, Academic Press.

[18] Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, 2Nd Ed., N.Y., Wiley.

[19] Richter, M. (1992), *Approximation of Gaussian Random Elements and Statistics*, B.G. Teubner Verlagsgesellscaft, Leipzig.

[20] Shepp, L.A. (1966) Radon-Nicodym Derivatives of Gaussian Measures, *Annual Math. Statistics*, Vol. 37, 321-354.

[21] Shepp, L.A. (1971) First Passage Time for a Particular Gaussian Process, *Annual Math. Statistics*, Vol. 42, 946-951.

[22] Shepp, L.A. And Slepian D. (1976) First Passage Time for a Particular Stationary Periodic Gaussian Process, *Journal of Applied Probability*, Vol. 13, 27-38.

[23] Vautard, R. And Ghil, M. (1989) Singular-Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time Series, *Physica* **D**, **35**, 395-424.

[24] Vautard, R., Yiou, P. and Ghil, M. (1992) Singular-Spectrum Analysis: a Toolkit for Short, Noisy Chaotic Signals, *Physica* **D**, **58**, 95-126.

[25] Zhigljavsky, A.A. and Kraskovsky, A.E. (1988) *Change-Point Detection in Random Processes for Radio-Engineering*, St.Petersburg University Press (in Russian).