Change–Point Detection in Time Series by means of the Singular Spectrum Analysis

V.G. Moskvina and A.A. Zhigljavsky [*]

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 of the main algorithm is that if at a certain time moment 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 M-lagged vectors $(x_{\tau+1}, \ldots, x_{\tau+M})$ is to be expected. The algorithm is more a model–building procedure rather than a precise statistical tool. However under certain conditions the algorithms could be considered as proper statistical procedures. Asymptotic expressions for the probability of false alarm in these algorithms are derived. Results of applications of proposed algorithms to several sets of data are displayed. Among the examples, we consider the famous airline data where presence of a change in trend is apparent.

Key Words: Principal components; Singular Spectrum Analysis; Sequential algorithm; Singular Value Decomposition.

1 Introduction

SSA, the Singular Spectrum Analysis, is a powerful technique of time series analysis. The main idea of SSA is the application of the Principal Components Analysis to the "trajectory matrix" obtained from the original time series with a subsequent reconstruction of the series. The methodology has become known since mid-eighties due to essential works of Broomhead and

[*] Anatoly Zhigljavsky is Professor, Chair in Statistics, Department of Mathematics Cardiff University, Cardiff CF2 4YH, UK, e-mail ZhigljavskyAA@cf.ac.uk. Valentina Moskvina is a Research Assistant at the same department, e-mail MoskvinaV@cf.ac.uk

King (1986), Broomhead, Jones, and King (1987) and also Vautard, Yiou, and Ghil (1992), Elsner and Tsonis (1996) and references therein. SSA is still a relatively unknown methodology in statistical circles. On the contrary, it recently became a standard tool in analysis of climatic time series, see for example Fraedrich (1986), Ghil and Vautard (1991), 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. 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 $K = N - M + 1$. Define the vectors $X_j = (x_j, \ldots, x_{j+M-1})^T$ $(j = 1, 2, \ldots)$ and the matrix

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

which is called the trajectory matrix.

Consider $\mathbf{X}$ as a multivariate data with M characteristics and K observations. The columns X_j of $\mathbf{X}$, considered as vectors, lie in an M-dimensional space. Define the matrix $R = \frac{1}{K}\mathbf{X}\mathbf{X}^T$. ($R$ is called the lag–covariance matrix.) Singular value decomposition (SVD) of R provides us with the collections of M eigen–values, eigen–vectors and principal components. A particular combination of a certain number $l < M$ of eigen-vectors determines an l-dimensional hyperplane in the M-dimensional space. According to the SSA algorithm, the M-dimensional data is projected onto this l-dimensional subspace and the subsequent averaging over the diagonals allows to get an approximation to the original series, see [....] 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 could 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 generating x_t then this distance should stay reasonably small for $X_j, j \geq K$. However, if at a certain time moment $N + \tau$ the mechanism generating x_t, $t \geq N + \tau$, has changed then an increase in the distance between the l-dimensional hyperplane and vectors X_j for $j \geq K + \tau$ is to be expected.

SSA expansions tend to pick up the main structure of the time series, if there is one. (In our discussion this corresponds to that the l-dimensional subspace approximates well the M-dimensional vectors.) If this structure is being picked up then the SSA continuation of the time series should agree with the continuation of the time series. (That is, the vectors X_j for $j \geq K$ should lie close to the l-dimensional subspace.) A change in the structure of the time series should move the corresponding vectors X_j out of the subspace. This is the central idea of the methods we propose and study.

SSA analyses time series structure in a nonsequential (off-line) manner. However, change–point detection problems are typically sequential (on–line) problems, and we aim at developing algorithms that could be used in an on–line regime in addition to the standard for time series analysis off–line manner. These algorithms would also be better accommodated to a presence of a slow change in time series structure, to outliers and to the case of multiply changes. All this will be achieved by sequential application of the SVD to the lag–covariance matrix computed in the time interval $[n+1, n+m]$ of fixed length m rather than to the trajectory matrix (1). Here $n = 0, 1, \ldots$ is the iteration number.

SSA as well as the associated change–point detection algorithms below are nonparametric and are not intended for precise statistical inferences, they are essentially model–building procedures. However under certain conditions the proposed algorithms could be considered as proper statistical procedures.

SSA analyses time series structure in a nonsequential (off-line) manner. However, change–point detection problems are typically sequential (on–line) problems, and we aim at developing algorithms that could be used in an on–line regime in addition to the standard for time series analysis off–line manner. Also these algorithms will be accommodated to a presence of a slow change in time series structure, to outliers and to the case of multiply changes. All this will be achieved by applying SVD to the trajectory matrix computed in a sequence of moving time intervals of a given length m rather than to the whole trajectory matrix (1). Payment for that is a decrease in sample size and therefore certain loss in efficiency in the ideal situation.

The paper is organised as follows.

In Section 2 we describe the basic scheme of SSA in the style of Danilov and Zhigljavsky (1997). In Section 3 we formulate change-point detection algorithms and give recommendations on the choice of parameters. In Section 4 we discuss an underlying statistical model, derive asymptotic expression for the first type error probability and provide an expression for the threshold conditionally this probability is fixed. The main versions of the change–point detection algorithms correspond to the case when $R = \mathbf{X}\mathbf{X}^T$. In Section 4 we extend the analysis of the algorithms to the case when R is the covariance matrix of the multivariate data $\mathbf{X}$. In Section 5 we demonstrate applications of the algorithms to several sets of data.

2 Description of the Algorithm

2.1 Rationale

If one accepts that SSA is a reliable tool in picking up the structure of time series and their continuation (that is the point of view in the literature on SSA, see for example Elsner and Tsonis (1996)) then he/she could be quite confident that SSA expansions could be a base of efficient change–point detection algorithms as well.

2.2 Informal description of the algorithms

Let $x_1, x_2, \dots, x_N$ be a time series where N is possibly ∞ and N is large enough. Let us choose two integers: an even integer m, $(m \leq N)$, the window width, and M $(M \leq m/2)$, the lag parameter. Define also $K = m - M + 1$.

For each $n = 0, 1, \dots, N - m$ we take time intervals $[n + 1, n + m]$ and construct the trajectory matrices

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

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

$$X_j^{(n)} = (x_{n+j}, \dots, x_{n+M+j-1})^T, \quad j = -n+1, -n+2, \dots, N-n-M+1 .$$

In line with (**??**), for every n define the lag–covariance matrix $\mathbf{R}_n = \frac{1}{K}\mathbf{X}^{(n)}(\mathbf{X}^{(n)})^T$. SVD of $\mathbf{R}_n$ gives us a collections of M eigen–vectors, and a particular combination of $l < M$ of them determines an l-dimensional subspace $\mathcal{S}_{n,l}$ in the M-dimensional space of vectors $X_j^{(n)}$. Denote the l eigenvectors that determine the subspace $\mathcal{S}_{n,l}$ by $P_1, \dots, P_l$ and the sum of squares of the (Euclidean) distances between vectors $X_j^{(n)} (j = p + 1, \dots, q)$ and this l-dimensional subspace by $\mathcal{D}_{n,l,p,q}$.

Since the eigen–vectors are orthogonal, the square of the Euclidean distance between an $M-$vector $ZY = X_j^{(n)}$ and the subspace $\mathcal{S}_{n,l}$ spanned by l eigen–vectors $P_1, \dots, P_l$, is just

$$||Z||^2 - ||P^T Z||^2 = Z^T Z - Z^T P P^T Z$$

where $||\cdot||$ is the Euclidean norm and P is the $M \times l$-matrix with columns $P_1, \dots, P_l$. Therefore

$$\mathcal{D}_{n,l,p,q} = \sum_{j=p+1}^{q} (X_j^{(n)})^T X_j^{(n)} - (X_j^{(n)})^T P P^T X_j^{(n)} \tag{3}$$

For fixed n, the part of sample $x_{n+1}, \dots, x_{n+m}$ that is used to construct the trajectory matrix $\mathbf{X}^{(n)}$ will be called 'training sample', and another part, $x_{n+p+1}, \dots, x_{n+q+M-1}$, which is used to construct the vectors $X_j^{(n)}(j = p+1, \dots, q)$ and thus to compute the sum of squared distances $\mathcal{D}_{n,l,p,q}$, will be called 'validation sample'. Of course, the training and validation samples may intersect.

If a change in the mechanism generating x_t occurred somewhere inside the time interval $[n+1, n+m]$ then the l-dimensional subspace $\mathcal{S}_{n,l}$ would provide a poorer approximation to the vectors $X_j^{(n)}$ and the values of $\mathcal{D}_{n,l}$ are going to be bigger.

Largest values of $\mathcal{D}_{n,l}$ are to be expected for n such that the change–point is in the middle of the interval $[n+1, n+m]$. The decision rule in the algorithm which we denote A(M,m,l,p,q,h) is to announce a change if for a certain n

$$\mathcal{D}_{n,l,p,q}/\mu_{n,l,p,q} \geq h \tag{4}$$

where h is a fixed threshold (see Section 3.2 concerning the choice of h) and $\mu_{n,l,p,q}$ is any estimator of the sum of squared distances $\mathcal{D}_{j,l,p,q}$ at the time intervals $[j+1, j+m]$ where the hypothesis of no change points has been accepted. We can assume for example

$$\mu_{n,l} = \frac{1}{n - m/2} \sum_{i=0}^{n-m/2-1} \mathcal{D}_{i,l} \tag{5}$$

In the cases where we allow either a slow change in time series structure or multiply change–points or even when n is very large, averaging for i from 0 to $n - m/2 - 1$ should be replaced by the averaging in a shorter interval, say, from $n - 3m/2$ to $n - m/2 - 1$.

2.3 Formal Description of the Algorithm

Let m, M, l, p and q be some integers such that m is even, $M \leq m/2$, $0 \leq p < q$. Define also $K = m - M + 1$. The change–point detection algorithm, denoted $\mathcal{A} = \mathcal{A}(m, M, l, p, q)$ is as follows.

For every $n = 0, 1, \ldots, N - m$ compute the trajectory matrix $\mathbf{X}_n$, see (2), the lag-covariance matrix $R_n = \frac{1}{K}\mathbf{X}^{(n)}(\mathbf{X}^{(n)})^T$, its SVD and $\mathcal{D}_{n,l,p,q}$, see (3), the sum of squared Euclidean distances between the vectors $X_j^{(n)} (j = p + 1, \ldots, q)$ and the $l-$dimensional subspace $\mathcal{S}_{n,l}$ spanned on the first l eigen-vectors of R_n. If for some $n > m/2$ the inequality (4) holds then a change in the structure of time series is announced to have happened in the interval $[n + 1, n + m]$.

An increase in $\mathcal{D}_{i,l}(m_0, m_1)$ considered as a function of n is to be expected starting at $n = \tau + m_0$ where τ is the point where a change has happen. This increase is expected to continue until about the point $n = \tau + m/2 + m_0$, then average squared distances $\mathcal{D}_{n,l}(m_0, m_1)$ are to be decreasing and stabilising (if there is no other change in time series) at perhaps another level for $n > \tau + m + m_0$.

2.4 Choice of parameters

Significant changes in 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 this tuning.

Length and location of the validation sample: m_0, m_1**.**

Three important special cases for the pair (m_0, m_1) in Algorithm 3 are:

(i) $(m_0, m_1) = (0, K)$ where $K = m - M + 1$, in this case Algorithm 3 is exactly Algorithm 2;

(ii) $(m_0, m_1) = (m - M, m)$, in this case we use $2M$ observations $x_{n+m-M+1}$, $\ldots$, x_{n+m+M} including M new points, to construct M test vectors $X_j^{(n)} (j = m - M + 1, \ldots, m)$;

(iii) $(m_0, m_1) = (m, m + M)$, in this case we use $2M$ new observations $x_{n+m+1}, \ldots, x_{n+m+2M}$ to construct M test vectors $X_j^{(n)} (j = m + 1, \ldots, m + M)$.

In numerical studies we mostly use Algorithm 3 with (m_0, m_1) selected according to (i) and (ii). If there are enough observations and slow variations in the trend are not allowed then (iii) is slightly preferable to both (i) and (ii) but the difference between (ii) and (iii) is almost insignificant. Algorithm 3 in cases (ii) and (iii), where 'training' samples are different from 'validation' samples, is more sensitive to changes than the more economical Algorithm 2, that is Algorithm 3 in case (i).

To get a smooth behaviour of the test statistics $\mathcal{D}_{n,l}(m_0, m_1)$ we need to select m_1 slightly bigger than m_0. If the difference $m_1 - m_0$ is too big then the behaviour of $\mathcal{D}_{n,l}(m_0, m_1)$ becomes too smooth, that's what is happening in Algorithm 2. There are no particular reasons why $m_1 - m_0$ should equal M.

Length of the training sample, window width: m.

The choice of m depends on what kind of structural changes we are looking for. If we allow small gradual changes in the time series then we could not take m very large. On the contrary, if we take m small then an outlier could be recognised as a structural change. A general rule is that value of m has to be reasonably large. Of course, if m is too large then we could either miss or smooth out all changes in our time series.

Parameters of the SSA algorithm: M, l.

To choose values of the lag M and the number of eigen–vectors chosen to approximate the trajectory matrices R_n, we have to follow standard SSA recommendations discussed in Section 2.3. We thus choose $M = m/2$ (recall that m is assumed even) and l such that the first l components describe well the signal and the lower $M - l$ components correspond to noise. To choose l, SSA decomposition of the whole series and some large parts of the series before applying the change-point detection algorithms is advised. Alternatively, if the problem is really sequential and preliminary study of the time series is impossible, the recommendation is to use all visual SSA tools in the first part of the series to choose l. (See Section **??**.)

If l is too small (underfitting) then we miss a part of the signal and therefore we could miss a change (that may happen in the underestimated components). If l is too large (overfitting) then we approximate a part of noise together with signal and finding a change in signal becomes more difficult.

3 Choice of the Threshold

3.1 Zero Hypothesis Model

The proposed algorithms are by no means the automatic tools to detect change–points, they are rather bricks for model building and visualization tools helping to see time non-homogenuities. However under certain conditions, that asymptotically hold under fairy general assumptions concerning the underlying time series, the algorithms could be considered as proper statistical procedures. It is the purpose of this section to demonstrate that.

The underlying assumption of the SSA technique in general and the proposed change–point detection algorithms in particular (Algorithms 2 and 3) is the assumption that the initial time series is well approximated by the series z_t, the solution of a finite–difference equation (**??**), that is, by a process of the form (**??**) with a 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 the finite–difference equation

$$z_t = a_1 z_{t-1} + \ldots + a_M z_{t-M} \tag{7}$$

with some coefficients $a_1, \ldots, a_M$ and certain initial conditions. The noise could be either random or deterministic but it has to have the property that its approximation by the solutions of the finite–difference equations is poor. (I.i.d.r.v. e_t certainly satisfy this assumption.)

Application of SSA with lag M at time intervals $[n+1, n+m]$ approximately recovers the model (6). That is we get

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

where $z_t^{(n)}$ is the SSA approximation for z_t, the solution of (7). Asymptotically, when $m \to \infty$, $M \to \infty$, and the noise e_t is an ergodic random process with finite variance, we get $z_t^{(n)} \to z_t$ for all n, see [1, p. 221]

In computing the threshold h we assume the following zero hypothesis:

(A1) the model (6) is valid and there is no change in parameters of the equation (7),

(A2) $z_t^{(n)} = z_t$ for all n and t,

(A3) either M or $m_1 - m_0$ tend to infinity, (A4) $e_t = e_t^{(n)}$ is a sequence of i.i.d.r.v., $e_t \sim N(0, \sigma^2)$ where the variance σ^2 is unknown.

The above assumptions imply that at iteration n in Algorithm 3

$$\mathcal{D}_{n,l}(m_0, m_1) = \sum_t w_{M,n+m_0,n+m_1}(t) e_t^2 \tag{9}$$

where (see Fig. 3.1), if $M \leq q - p$,

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

and, if $0 < q - p \leq M$,

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

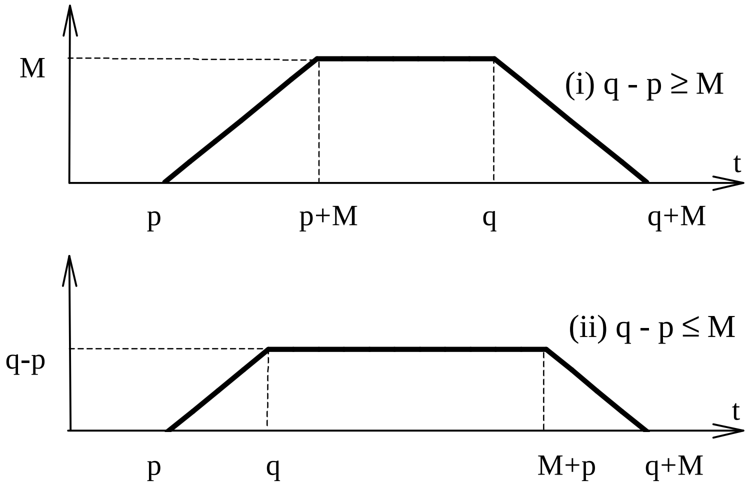

Figure 3.1: Function $w_{M,p,q}(t)$

We have

$$\sum_t w_{M,p,q}(t) = \sum_t w_{M,0,q-p}(t) = M(q - p) \tag{10}$$

$$\sum_t w^2_{M,p,q}(t) = \begin{cases} \frac{1}{3}M\left(3M(q-p)+1-M^2\right) & \text{if } q-p \geq M \\ \frac{1}{3}(q-p)\left(3M(q-p)+1-(q-p)^2\right) & \text{if } q-p \leq M \end{cases} \tag{11}$$

Obviously (9) is a quadratic form $e^T Be$ where $e = (e_1, e_2, \ldots, e_N)^T$ and $B = B(M, n, m_0, m_1)$ is a diagonal matrix with diagonal elements $B_{tt} = w_{M,n+m_0,n+m_1}(t)$.

Using the results on distributions of quadratic forms of random variables, see e.g. Searle (1971), we have the following moments of $e^T Be$:

$$E(e^T Be) = \sigma^2 \mathrm{tr} B\,, \tag{12}$$

$$\mathrm{var}(e^T Be) = 2\sigma^4 \mathrm{tr} B^2 \tag{13}$$

Note that the proof of (12) does not require normality of e_t and that both properties (12) and (13) are easily generalisable to the case when the components of e are dependent (e.g. to the case when e_t is an autoregressive process), see Searle (1971).

The representation (9) and properties (12), (13), (10), (11) imply

$$E\mathcal{D}_{n,l}(m_0, m_1) = \sigma^2 \sum_t w_{M,n+m_0,n+m_1}(t) = \sigma^2 M(m_1 - m_0)\,,$$

$$\mathrm{var}\mathcal{D}_{n,l}(m_0, m_1) = 2\sigma^4 \sum_t w^2_{M,n+m_0,n+m_1}(t)$$

$$= \frac{2\sigma^4}{3} \times \begin{cases} M\left(3M(m_1-m_0)+1-M^2\right) & \text{if } m_1 - m_0 \geq M \\ (m_1-m_0)\left(3M(m_1-m_0)+1-(m_1-m_0)^2\right) & \text{if } m_1 - m_0 \leq M \end{cases}$$

Standardising the random variable $\mathcal{D}_{n,l}(m_0, m_1) = e^T Be$ and taking into account its asymptotic normality, which is a consequence of (A3), we get asymptotically

$$\frac{\mathcal{D}_{n,l}(m_0, m_1) - E\mathcal{D}_{n,l}(m_0, m_1)}{\sqrt{\mathrm{var}\mathcal{D}_{n,l}(m_0, m_1)}} \sim N(0,1) \tag{14}$$

3.2 Choice of the Threshold h

Let α be a fixed significance level (say $\alpha = 0.05$) and t_α be such that $\Phi(t_\alpha) = 1 - \alpha$ where $\Phi(\cdot)$ is the c.d.f. of the standard normal $N(0,1)$ distribution. (For example, $t_{0.05} \simeq 1.645$.)

The asymptotic relation (14) implies that we can claim that asymptotically the probability of the event

$$\frac{\mathcal{D}_{n,l}(m_0,m_1)-E\mathcal{D}_{n,l}(m_0,m_1)}{\sqrt{\mathrm{var}\mathcal{D}_{n,l}(m_0,m_1)}} \geq t_\alpha \tag{15}$$

is α. We then rewrite this inequality in the form

$$\frac{\mathcal{D}_{n,l}(m_0,m_1)}{E\mathcal{D}_{n,l}(m_0,m_1)} \geq 1+t_\alpha \frac{\sqrt{\mathrm{var}\mathcal{D}_{n,l}(m_0,m_1)}}{E\mathcal{D}_{n,l}(m_0,m_1)} = 1+t_\alpha C_{M,m_1-m_0} \tag{16}$$

where

$$C_{u,v}=\frac{\sqrt{6}}{3uv}\times\begin{cases}\sqrt{u\,(3uv+1-u^2)} & \text{if } v\geq u\\ \sqrt{v\,(3uv+1-v^2)} & \text{if } v\leq u\end{cases}$$

The decision rule (16) takes the required form (**??**) when we set

$$h=1+t_\alpha C_{M,m_1-m_0}$$

and replace $E\mathcal{D}_{n,l}(m_0,m_1)$ by its consistent estimate $\mu_{n,l}(m_0,m_1)$ in the denominator of the test statistics in (16). This replacement does not violate the asymptotic normality of this statistics, see e.g. property (b) on page 122 in Rao (1973).

Assuming that $M=m/2$ in Algorithm 2, in three important particular cases considered in Section 3.4, we have $m_1-m_0=M$ and therefore

$$h=1+t_\alpha\frac{\sqrt{6M(2M^2+1)}}{3M^2}\sim 1+t_\alpha\sqrt{\frac{4}{3M}}+o(M^{-2}),\; M\to\infty\,. \tag{17}$$

For $\alpha=0.05$ we thus have $h\simeq 1+1.9/\sqrt{M}$, a pretty good approximation to (17), even for relatively small M.

3.3 Application of SSA with averaging

If we know the signal under zero hypothesis exactly then we substract it from the observations x_k and we do not need to make an SSA decomposition, choice $l=0$ would work. That happens when we have enough observations to estimate the signal (trend) which does not vary in time. This is an assumption which often occurs in the change–point detection literature but it is not very interesting case for us.

Assume now that under zero hypothesis the mean Ex_k is approximately constant. Of course we can estimate this constant and substract it from x_k.

If the number of observations is not large and the constant is not actually constant but a slowly varying function, this approach may create enough bias to make the related change–point problem difficult. An alternative approach, in line with the discussions above, is to to apply versions of Algorithms 2 and 3 with $l = 0$ and where row averages are substracted from the elements of the trajectory matrix. (That is, SSA algorithm with $l = 0$ and $\mu_i = \bar{x}_i$ and $\sigma_i = 1$ is applied for every n.)

Substraction of moving (row) averages from x_k makes the decision rule invariant to the mean of the process x_k but changes the expressions for the distances.

We could still apply Algorithms 2 and 3 (modified so that the row averages are substracted from the elements of the trajectory matrices) to test for changes but we have to modify the selection rule for the threshold h.

For both algorithms $\mathcal{D}_{n,l}(m_0, m_1)$ can still be represented as a quadratic form $\mathcal{D}_{n,l}(m_0, m_1) = e^T \tilde{B} e$ but the matrix $\tilde{B}$ is no longer diagonal. Let us compute the expression for $\tilde{B}$.

Since elements e_i with $i \le n$ and $i \ge n + m_1 + M$ are not present in the quadratic form $\mathcal{D}_{n,l}(m_0, m_1)$, $\tilde{B}_{ij} = 0$ if either $\min i, j \le n$ or $\max i, j \ge n + m_1 + M$. That is,

$$\tilde{B} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & B & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where B is a certain $(n+m_1+M-1) \times (n+m_1+M-1)$-matrix and $\mathbf{0}$ denote matrices of zeroes of appropriate dimensions.

This implies that

$$\mathcal{D}_{n,l}(m_0, m_1) = e^T \tilde{B} e = (e^{(n)})^T B e^{(n)}$$

where $e^{(n)} = (e_{n+1}, \ldots, e_{n+m_1+M-1})^T$.

Using the notation $e_{ij}^{(n)} = e_{n+i+j-1}$ we then have

$$\mathcal{D}_{n,l}(m_0, m_1) = \sum_{i=1}^{M} \sum_{j=m_0+1}^{m_1} (e_{ij}^{(n)} - e_{i.}^{(n)})^2$$

$$= \sum_{i=1}^{M} \sum_{j=m_0+1}^{m_1} (e_{ij}^{(n)})^2 + r \sum_{i=1}^{M} (e_{i.}^{(n)})^2 - 2r \sum_{i=1}^{M} \tilde{e}_{i.}^{(n)} e_{i.}^{(n)}$$

where $r = m_1 - m_0$, $K = m - M + 1$,

$$e_{i.}^{(n)} = \frac{1}{K} \sum_{j=1}^{K} e_{ij}^{(n)} \quad \tilde{e}_{i.}^{(n)} = \frac{1}{r} \sum_{j=m_0+1}^{m_1} e_{ij}^{(n)}$$

This yields that the matrix $B = B(M, K, r)$ is the matrix of the form

$$B = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & B_1 \end{pmatrix} + \frac{r}{K^2} \begin{pmatrix} B_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \frac{1}{K} \begin{pmatrix} \mathbf{0} & B_3 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \frac{1}{K} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ B_3^T & \mathbf{0} \end{pmatrix}$$

where B_1 is $(M{+}r{-}1) \times (M{+}r{-}1)$-matrix with elements

$$B_1 = b_{ij}^{(1)} = \begin{cases} w_{M,0,K}(i) & \text{if } \ i = j, \ 1 \le i \le K + M - 1 \\ 0 & \text{otherwise} \end{cases}$$

B_2 is $(M{+}K{-}1) \times (M{+}K{-}1)$-matrix with elements

$$B_2 = b_{ij}^{(2)} = \begin{cases} w_{i,\,0,\,K}(j), & \text{if } \ 1 \le i \le M \\ w_{M,\,0,\,K}(j), & \text{if } \ M \le i \le K \\ w_{M+K-i,\,i-K,\,i}(j), & \text{if } \ K \le i \le M + K - 1 \end{cases}$$

B_3 is $(M{+}K{-}1) \times (M{+}r{-}1)$-matrix with elements

$$B_3 = b_{ij}^{(3)} = \begin{cases} w_{i,\,0,\,r}(j), & \text{if } \ 1 \le i \le M \\ w_{M,\,0,\,r}(j), & \text{if } \ M \le i \le K \\ w_{M+K-i,\,i-K,\,r+i-K}(j), & \text{if } \ K \le i \le M + K - 1 \end{cases}$$

Matrix B_1 is a diagonal matrix with diagonal elements $w_{M,0,K}(i), \ i = 1, \dots, m{=}M{+}K{-}1$ and matrices B_2 and B_3 have the form

$$B_* = \begin{pmatrix}
1 & 1 & & \dots & & 1 & 0 & & & \dots & & & 0 \\
1 & 2 & & \dots & & 2 & 1 & & & \dots & & & 0 \\
\vdots & \vdots & \ddots & & & & & \ddots & \ddots & & & & \vdots \\
1 & 2 & \dots & M & \dots & M & \dots & 2 & 1 & 0 & & \dots & 0 \\
0 & 1 & 2 & \dots & M & \dots & M & \dots & 2 & 1 & 0 & \dots & 0 \\
\vdots & \ddots & \ddots & & & \ddots & & \ddots & & & \ddots & \ddots & \vdots \\
0 & \dots & 0 & 1 & 2 & \dots & M & \dots & M & \dots & 2 & 1 & 0 \\
0 & \dots & & 0 & 1 & 2 & \dots & M & \dots & M & \dots & 2 & 1 \\
\vdots & & & & \ddots & \ddots & & & & & \ddots & \vdots & \vdots \\
0 & & & \dots & & 0 & 1 & 2 & & \dots & & 2 & 1 \\
0 & & & \dots & & & 0 & 1 & & \dots & & 1 & 1
\end{pmatrix}$$

In case $m_0 = 0, m_1 = K$, that is Algorithm 2 and case (i) in Section 2.4, the formulas could be very much simplified:

$$e^T B e = \mathcal{D}_{n,l}(m_0, m_1) = \sum_{i=1}^{M} \sum_{j=1}^{K} (e_{ij}^{(n)} - e_{i.}^{(n)})^2 = \sum_{i=1}^{M} \sum_{j=1}^{K} (e_{ij}^{(n)})^2 - K \sum_{i=1}^{M} (e_{i.}^{(n)})^2$$

implying $r = K$, $B_2 = B_3 = B_3^T$. Therefore matrix B has the form

$$B = B_1 - \frac{1}{K} B_2$$

and elements

$$b_{ij} = \begin{cases} \frac{K-1}{K} w_{M,O,K}(j) & \text{if } \ i = j \\ -\frac{1}{K} w_{i,\,0,\,K}(j), & \text{if } \ i \neq j, \ \ 1 \leq i \leq M \\ -\frac{1}{K} w_{M,\,0,\,K}(j), & \text{if } \ i \neq j, \ \ M \leq i \leq K \\ -\frac{1}{K} w_{M+K-i,\,i-K,\,i}(j), & \text{if } \ i \neq j, \ \ K \leq i \leq M+K-1 \end{cases}$$

Hence $tr(B) = M(K-1)$,

$$tr(B^2) = \sum_{i,j} b_{ij}^2 = M^2 K - \frac{M(M^2-1)}{3} - M^2 + \frac{M^2(M^2-1)}{6K^2}$$

Even more interesting case is when $m_0 \geq K$ in Algorithm 3. (This includes versions (ii) and (iii) in Section 2.4.) In this case we have

$$tr(B) = \frac{Mr(K+1)}{K}, \quad tr(B^2) = M^2 r - \frac{M(M^2-1)}{3}$$

$$+\frac{r^2}{K^4}\left[(K-M+1)^2 M^2 - \frac{M(M-1)}{6}(5M^2 - 8KM - 7M + 4K)\right]$$

$$+\frac{2}{K^2}\left[(K-M+1)(r-M+1)M^2 - \frac{M(M-1)}{6}(5M^2 - 4M(K+r) - 7M + 2(K+r))\right]$$

Using (16) and properties (12), (13), we can now apply the decision rule (**??**) where the threshold is now

$$h = 1 + t_\alpha \frac{\sqrt{\text{var}\mathcal{D}_{n,l}(m_0,m_1)}}{E\mathcal{D}_{n,l}(m_0,m_1)} = 1 + t_\alpha \frac{\sqrt{2tr(B^2)}}{tr(B)}$$

and $tr(B)$, $tr(B^2)$ are computed above.

4 Numerical examples

To illustrate applications of Algorithms 2 and 3, let us consider five numerical examples. In the first four examples 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 e_t is noise and the change–point was always $\tau = 200$.

Example 1. (see Fig. 4.1(a,b))

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

and e_t are i.i.d.r.v. $e_t \sim N(0,1)$ for $t = 1, \ldots, 400$ (white noise).

Example 2. (see Fig. 4.2(a,b)) $z_1 = 0, z_2 = 8, z_3 = 6, z_4 = 4$,

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

and e_t are i.i.d.r.v. $e_t \sim N(0,1)$ for $t = 1, \ldots, 400$.

Example 3. (see Fig. 4.3(a))

$$z_t = \begin{cases} 0 & \text{for } 1 \le t \le 200 \\ 1 & \text{for } 201 \le t \le 400 \end{cases}$$

and e_t are i.i.d.r.v. $e_t \sim N(0,1)$ for $t = 1, \ldots, 400$.

Example 4. (see Fig. 4.4(a)) $z_t = 0$ for $t = 1, \ldots, 400$ and e_t are i.i.d.r.v. $e_t \sim N(0, \sigma_t^2)$

$$\sigma_t^2 = \begin{cases} 1 & \text{for } 1 \le t \le 200 \\ 4 & \text{for } 201 \le t \le 400 \end{cases}$$

In the first three examples the change point is not obviously seen in the graph. The most difficult is Example 2 where success of the proposed change–point detection algorithms could only be explained by the fact that the model (6) is perfect for the time series. In Examples 1 and 3 (6) is also a suitable model but the signals are simpler. Signal in Example 1 is standard in signal processing and the change–point problem of Example 3 is the most celebrated problem in the field. Example 4, where the change happens in the noise parameters, is also quite standard. Of course, when exact parametric models for signal and noise are known, CUSUM type algorithms are superior to our algorithms for the problems of Example 3 and 4. (However comparative study of different algorithms in beyond the scope of this paper.)

In all four examples we have applied Algorithm 2 with $m = 100, M = 50$ as well as Algorithm 3 with $m = 100, M = 50, m_0 = n, m_1 = n + M$ (case (ii) in Section 3.4) Since we were interested only in the main either periodics or trend components, the number of eigen-vectors, l, was small: $l = 2$ in Examples 1 and 3 $l = 4$ in Example 2 and $l = 1$ in Example 4. For Examples 3 and 4 we have also applied the Algorithms 2 and 3 with $l = 0$ and averaging, as described in Section 3.3.

In plots at Fig. 4.1(c), 4.2(c), 4.3(b,c) and 4.4(b,c) we plot test statistics $d_n = \frac{1}{tr(B)}\mathcal{D}_{n,l}(m_0, m_1)$, the normalised values of the sum of the squared distances between the vectors $X_j^{(n)} (j = m_0+1, \ldots, m_1)$ and the $l-$dimensional subspace $\mathcal{S}_{n,l}$. As soon as $n + m_1 < \tau = 200$, that is no change happened, values of d_n should be close to 1. (Corresponding values of n are in the range $[m, m+\tau] = [100, 200]$ for Algorithm 2 and $[m, m+\tau-M] = [100, 150]$ for Algorithm 3.) Then the values of d_n are expected to grow and reach highest values at n around $\tau + M = 250$ for Algorithm 2 and $\tau = 200$ for Algorithm 3. Then the time series is slowly becoming undisturbed again and therefore values of d_n are stabilising at perhaps another level (for $n > 300$ and $n > 250$, correspondingly). This is what we roughly see at Fig. 4.1(c), 4.2(c), 4.3(b,c) and 4.4(b,c) with much clearer picture for Algorithm 3 than for Algorithm 2.

Example 5. *Airlines data.* (see Fig. 4.5(a))

This celebrated data, see e.g. Box and Jenkins (1970), give logarithms of monthly totals (in thousands) of international airline passengers for January 1949 - December 1960. There are only 144 data points so we have selected a rather small m, $m = 36$. (For the sake of precision of SSA decompositions m has to be proportional to the main period which is 12.) We have also taken $M = m/2 = 18$, according to the recommendations in Section 2.4, and $m_0 = 18$, as in the case (ii) of Section 2.4, and $m_1 = 30$ (value $m_1 = 36$ would be a little too large: there is not enough data). To choose l, we have made SSA decomposition of the whole series (Algorithm 1 with $N = 144$ and $M = 18$), see also Danilov (1997). The results of the decomposition are displayed on Fig. 4.5(a): the main trend is described by the 1-st and 6-th principal components, the main period (12–months) is described by the 2-nd and 3-rd components and the second main period (6–months) is represented by the 4-th and 5-th principal components (The series reconstructed from these two components is shifted down which resulted in the second zero in the plot on Fig. 4.5(a).) Fig. 4.5(b) shows that Algorithm 3 clearly indicates on two time intervals where certain changes in trend have probably occurred.

Acknowledgements

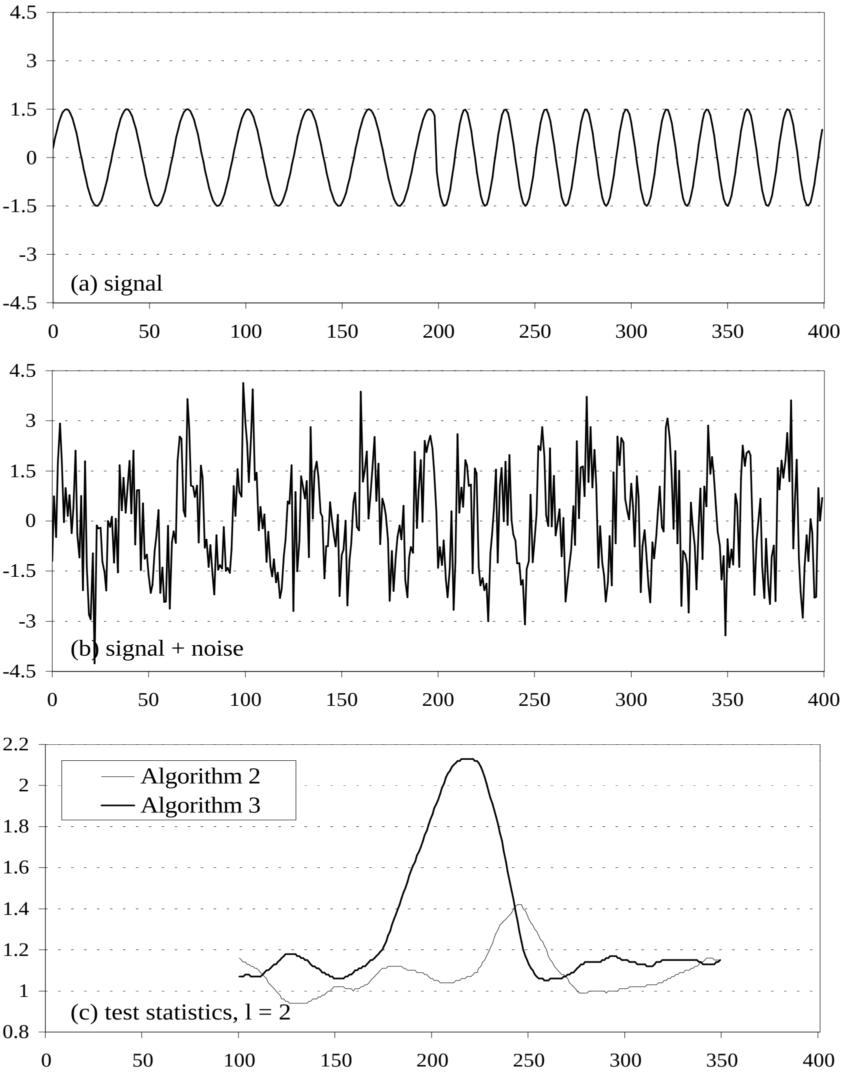

Figure 4.1: Model of Example 1.

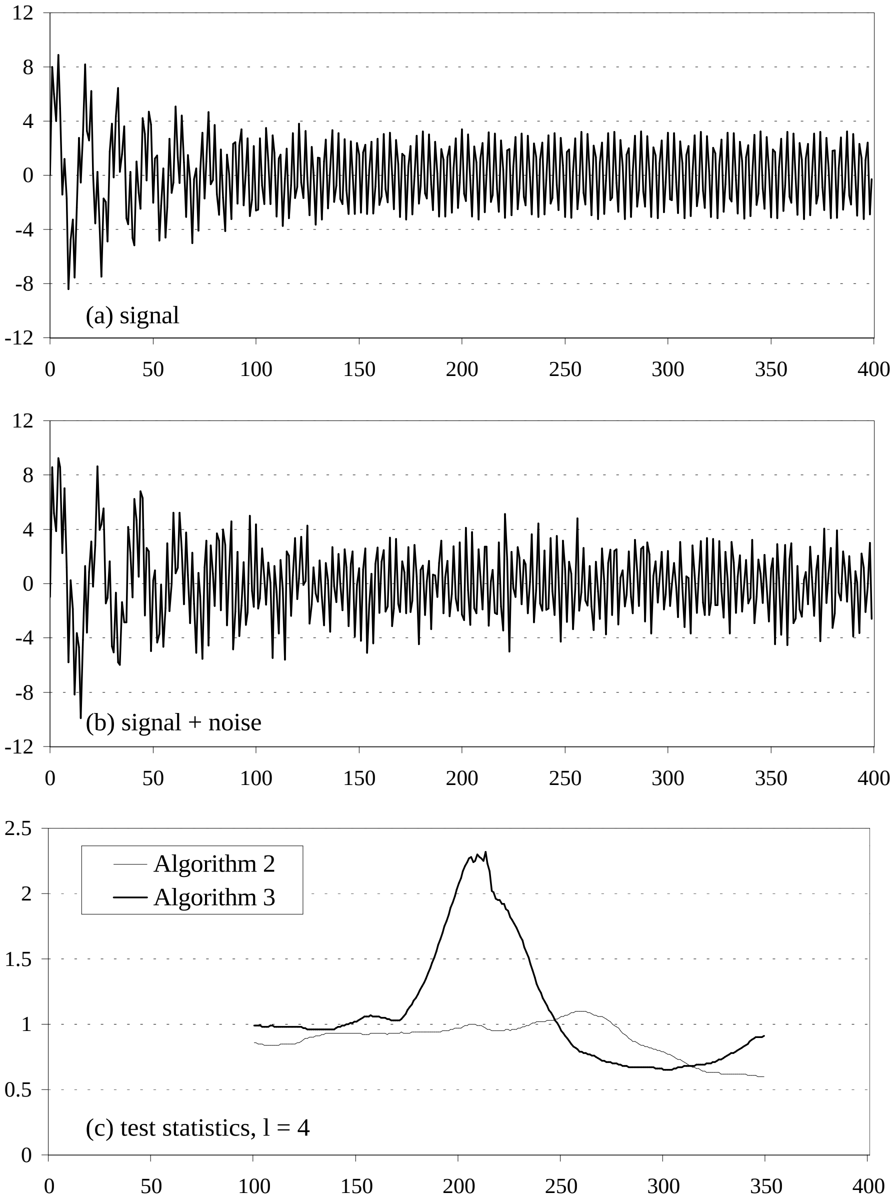

Figure 4.2: Model of Example 2.

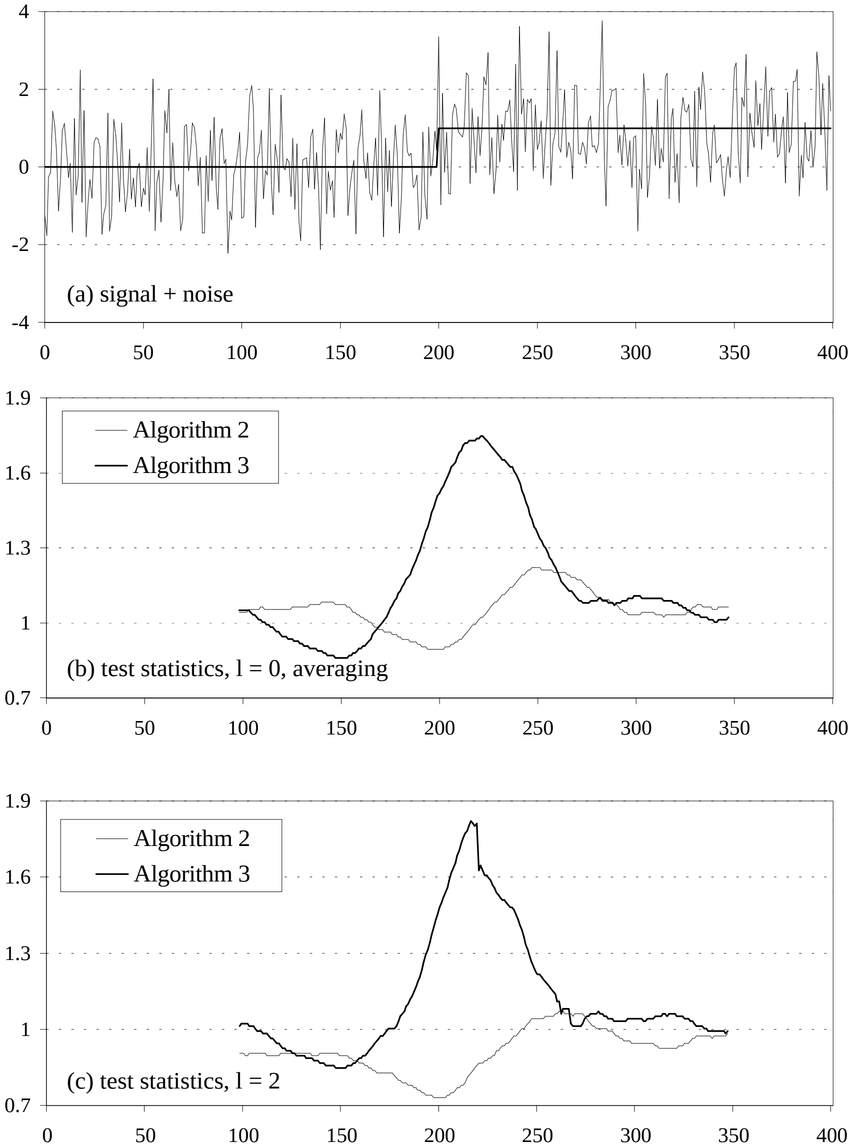

Figure 4.3: Model of Example 3.

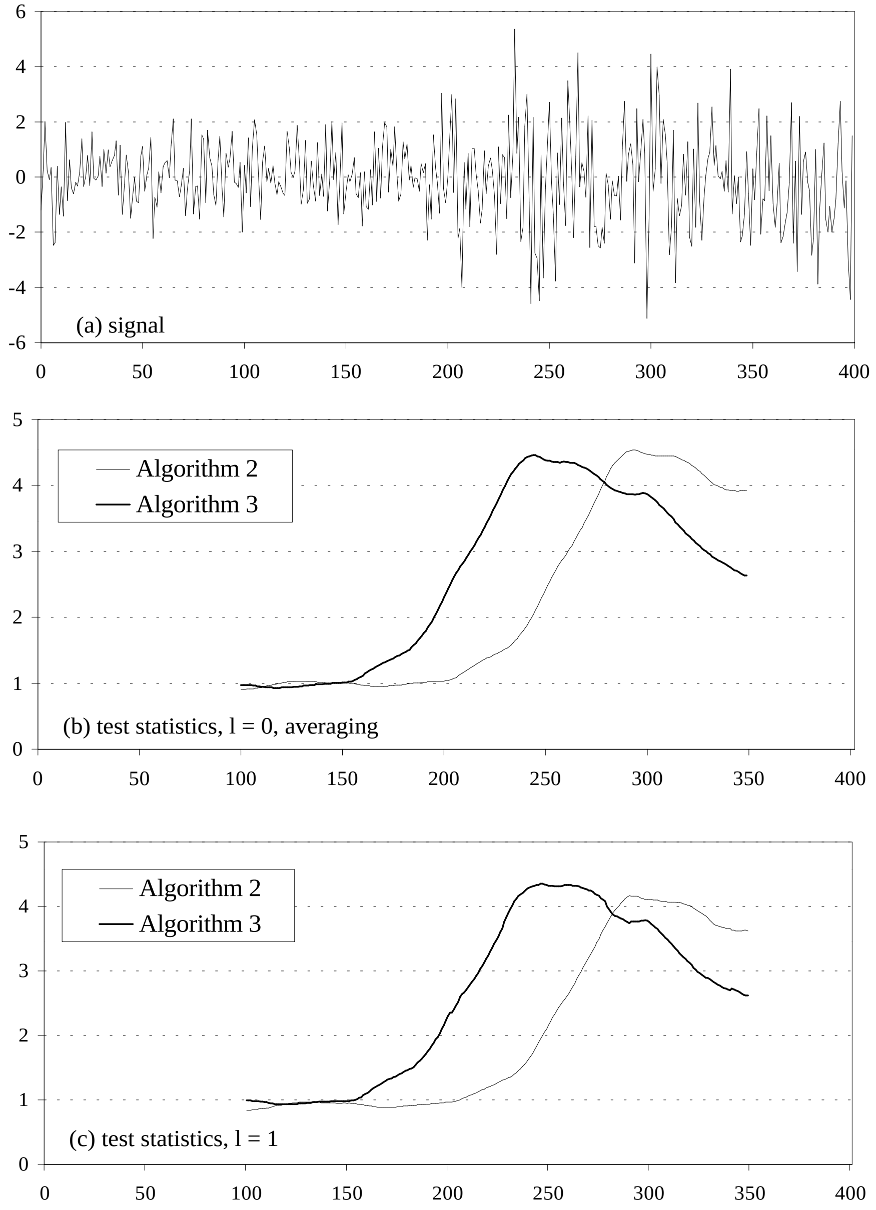

Figure 4.4: Model of Example 4.

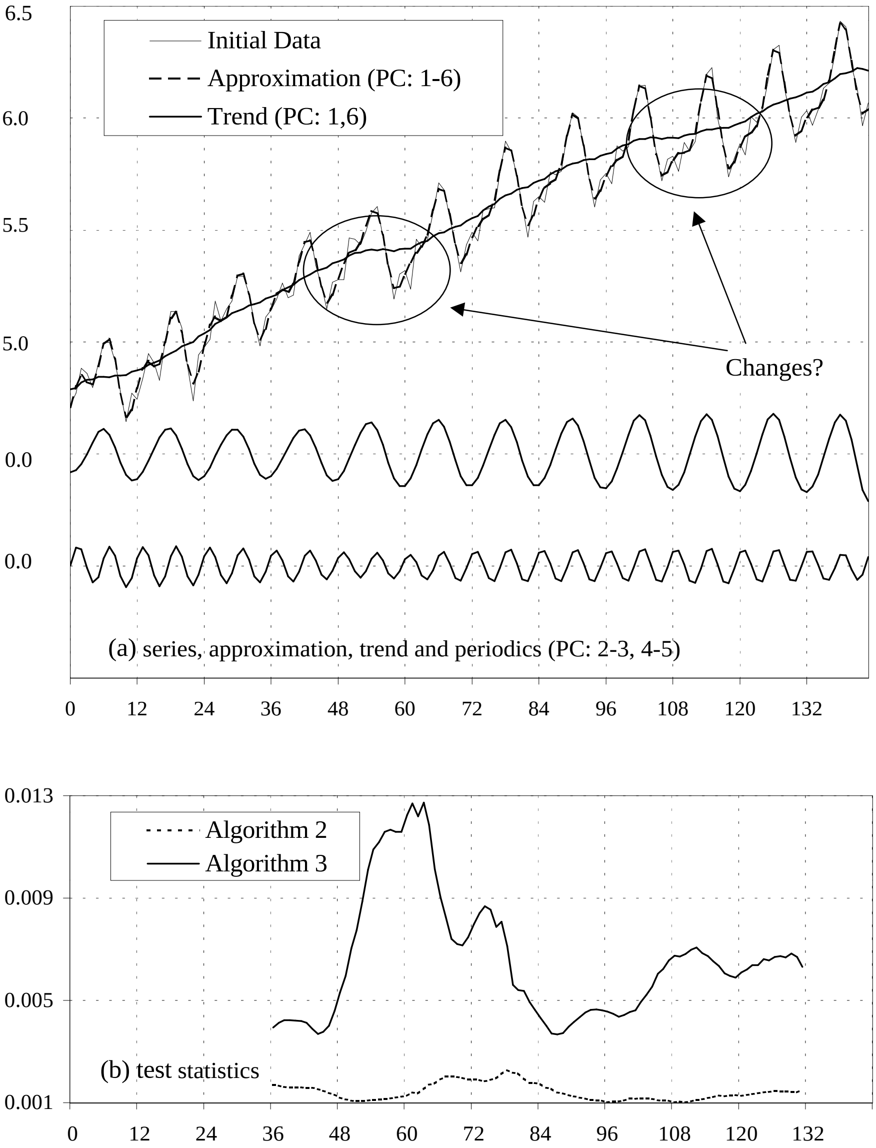

Figure 4.5: Logarithm of airline passenger numbers, Example 5. Parameters: $m = 36$, $m_0 = M = 18$, $m_1 = 30$, $l = 6$.

The work has been supported by Engineering and Physical Science Research Council, UK, grant GR/M221713 of 01/07/98.

References

[1] Box G.E.P and Jenkins G.M., (1970), *Time series analysis, Forecasting and Control*, San-Francisco: Holden-Day.

[2] Broomhead, D.S., Jones, R. and King, G.P., (1987) Topological dimension and local coordinates from time series data. *Physica* **A**, **20**, L563-L569.

[3] Broomhead, D.S. and King, G.P., (1986), Extracting qualitative dynamics from experimental data. *Physica* **D**, **20**, 217-236

[4] Danilov, D., (1997), Principal components in time series forecast, *Journal of Computational and Graphical Statistics*, 6, 112-121.

[5] Danilov, D., and Zhigljavsky, A., eds (1997), *Principal Components of Time Series: "Caterpillar" Method,* St.Petersburg University, (in Russian)

[6] Elsner, J. and Tsonis, A., (1996), Singular Spectrum Analysis: a new tool in time series analysis, *Plenum Press, New York.*

[7] Fraedrich, K., (1986), Estimating the dimension of weather and climate attractors, *J. Atmos. Sci.*, **43**, 419-432.

[8] Ghil, M., and Vautard, R., (1991), Interdecadal oscillations and the warming trend in global temperature time series, *Nature*, **350**, 324-327.

[9] Rao, C. R., (1973), *Linear statistical inference and its applications*, 2nd ed., N.Y., Wiley.

[10] Searle S.R. (1971) *Linear models*, N.Y., Wiley.

[11] Vautard, R., and Ghil, M., (1989), Singular-spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica* **D**, **35**, 395-424.

[12] Vautard, R., Yiou, P., and Ghil, M., (1992), Singular-spectrum analysis: A toolkit for short, noisy chaotic signals, *Physica* **D**, **58**, 95-126.