ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Description Length Based Signal Detection in Singular Spectrum Analysis

Md Atikur Rahman Khan and D. S. Poskitt

May 2010

Working Paper 13/10

Description Length Based Signal Detection in Singular Spectrum Analysis

Md Atikur Rahman Khan
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: Atikur.Khan@buseco.monash.edu.au

D. S. Poskitt
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: Donald.Poskitt@buseco.monash.edu.au

24 May 2010

JEL classification: C14, C22, C52

Description Length Based Signal Detection in Singular Spectrum Analysis

Abstract: This paper provides an information theoretic analysis of the signal-noise separation problem in Singular Spectrum Analysis. We present a signal–plus–noise model based on the Karhunen-Loève expansion and use this model to motivate the construction of a minimum description length criterion that can be employed to select both the window length and the signal. We show that under very general regularity conditions the criterion will identify the true signal dimension with probability one as the sample size increases, and will choose the smallest window length consistent with the Whitney embedding theorem. Empirical results obtained using simulated and real world data sets indicate that the asymptotic theory is reflected in observed behaviour, even in relatively small samples.

Keywords: Karhunen-Loève expansion, minimum description length, signal–plus–noise model, Singular Spectrum Analysis, embedding.

1 Introduction

Singular spectrum analysis (SSA) is a non-parametric technique that has gained popularity in the analysis of meteorological, bio-mechanical and hydrological time series (Ghil et al., 2002; Alonso et al., 2005; Marques et al., 2006), to name but a few. SSA is designed to look for nonlinear, non–stationary, and intermittent or transient behaviour in an observed time series, and following its successful application in the physical sciences, applications in economics and finance are now also finding favour (See for example Thomakos et al., 2002; Hassani and Zhigljavsky, 2009; Hassani et al., 2009). Indeed, although the introduction of SSA is often attributed to researchers working in the physical sciences, namely Broomhead and King (1986) and Vautard and Ghil (1989); Vautard et al. (1992), the basic building blocks of SSA where outlined by Basilevsky and Hum (1979), who argued that a discrete Karhunen-Loève analysis was more suited to applications in the social sciences where a frequency domain decomposition based on standard Fourier methods may lack appeal “because social systems rarely exhibit regular periodic behavior".

Although considerable energy has been expended on investigating and explaining the algebraic structure and interpretation of SSA, and various extensions of SSA have been considered – to forecasting, and missing data and change point problems for example (Golyandina et al., 2001) – little of the current literature analyzes the statistical properties of SSA from an abstract theoretical perspective. Our purpose in this paper is to build upon the foundations laid in Basilevsky and Hum (1979) and present what we believe to be the first detailed theoretical analysis of an automatic identification procedure designed to be implemented in the context of SSA.

By way of introduction to SSA, and in order to set the scene, suppose that $x(t)$ is a stochastic process of interest that is observed at a sequence of points $t_1 < t_2 < \cdots < t_N$ in the interval $T = (t_{\min}, t_{\max})$, giving rise to an observed time series $\{x(t_1), x(t_2), \ldots, x(t_N)\}$ of length N. The aim of SSA is to decompose an observed series into the sum of independent and interpretable components, akin to the classical decomposition of a time series into the sum of a trend plus cyclical plus seasonal plus noise, and SSA looks for such structure in an observed series via an eigen–decomposition of the so-called *trajectory* matrix. The general structure of the algorithm underlying SSA can be described in four basic steps:

1. Embedding: For a given window size m the $m \times n$ trajectory matrix is given by

$$\mathbf{X} = [\mathbf{x}_1 : \ldots : \mathbf{x}_n] \tag{1}$$

   where $n = N - m + 1$ and $\mathbf{x}_i = (x(t_i), x(t_{i+1}), \ldots, x(t_{i+m-1}))'$ for $(i = 1, 2, \ldots, n)$ are known as the m–lagged vectors of $\mathbf{X}$. The parameter m is variously referred to as the trajectory matrix window size or length, the lag length, or the dimension of the trajectory space. Following standard practice we will suppose that m is assigned by the practitioner such that $2 < m \leq N/2 \leq n$.

2. Singular Value Decomposition: Let $\ell_1 \geq \ell_2 \geq \ldots, \geq \ell_m \geq 0$ denote the eigenvalues of $\mathbf{XX}'$, arranged in descending order of magnitude, and denote by $\mathbf{U}_1, \ldots, \mathbf{U}_m$ the associated orthonormal system of eigenvectors. Then the row space of the trajectory matrix, $\mathscr{H}^{(d)}$, has dimension d where $d = max\{i : \ell_i > 0\}$ and $\mathbf{X}$ can be expressed exactly as

the sum of $d \leq m$ rank one projections

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d \,, \tag{2}$$

wherein $\mathbf{X}_i = \sqrt{\ell_i}\mathbf{U}_i\mathbf{V}_i'$ and $\sqrt{\ell_i}$ is the ith singular value, and $\mathbf{U}_i$ and $\mathbf{V}_i = \mathbf{X}'\mathbf{U}_i/\sqrt{\ell_i}$ are the ith left and right eigenvectors of $\mathbf{X}$. In SSA the vectors $\mathbf{U}_i$ and $\mathbf{V}_i$ are known as the ith *empirical orthogonal function* and the ith *principle component*, respectively.

3. Noise Reduction: It is well known that $||\mathbf{X}||^2 = \text{trace}\{\mathbf{X}\mathbf{X}'\} = \sum_{i=1}^m \ell_i$ and $||\mathbf{X}_i||^2 = \ell_i$ for $i = 1,\ldots,m$, and $\ell_i/\sum_{i=1}^m \ell_i$ can be interpreted as the proportion of the total variation in $\mathbf{X}$ attributable to $\mathbf{X}_i$. Since every *eigentriple*, $\{\ell_i, \mathbf{U}_i, \mathbf{V}_i\}$, need not contribute to the overall variation, the next step is to determine a subset of eigentriples that encompass the variation in $\mathbf{X}$. In practice, this amounts to selecting a lower dimensional space $\mathscr{H}^{(k)} \subseteq \mathscr{H}^{(d)}$ where $k \leq d$ on which to project the trajectory matrix, with the associated residual being interpreted as noise. The resulting "noise–free" representation of $\mathbf{X}$ is now given by

$$\mathbf{X} \approx \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_k} \,, \tag{3}$$

where the jth component is $\mathbf{X}_{i_j} = \sqrt{\ell_{i_j}}\mathbf{U}_{i_j}\mathbf{V}_{i_j}'$ for $j = 1,\ldots,k$ and $\{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, d\}$ denotes the designated or chosen subset.

4. Time Series Reconstruction: The purpose of this step is to transform the "noise–free" representation of $\mathbf{X}$ into a "noise–free" reconstruction of $\{x(t_1), x(t_2), \ldots, x(t_N)\}$. Noting that $\mathbf{X}$ is a Hankel matrix, this is achieved by a process of diagonal averaging or *Hankelization*. For a given $\mathbf{X}_i$, the Hankelized version is obtained by replacing the r,cth element of $\mathbf{X}_i$, $r = 1,\ldots,m$, $c = 1,\ldots,n$, by the average over all r and c such that $r + c = t + 1$ where $t = 1,\ldots,N$. The resulting Hankel matrix, $\widetilde{\mathbf{X}}_i$ say, implicitly defines an embedded series $\{\widetilde{x}^i(t_1), \widetilde{x}^i(t_2), \ldots, \widetilde{x}^i(t_N)\}$, and by applying the Hankelization procedure to each of the components in (3) we obtain the "noise–free" expansion $x(t) \approx \sum_{j=1}^k \widetilde{x}^{i_j}(t)$, $t = t_1, \ldots, t_N$, of the original time series.

For more detailed particulars on SSA we refer to Golyandina et al. (2001), where a description of the technique and its practical application (with several examples) can be found.

Here we note that in SSA it is more conventional to refer to the third step as *Grouping*. This corresponds to dividing the elementary matrices $\mathbf{X}_i$ into disjoint subsets. If $I = \{i_1, \ldots, i_g\} \subseteq$

$\{1,\ldots,d\}$ then the resultant matrix corresponding to group I is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_g}$, and given G disjoint groups $I_1,\ldots,I_G$ the grouped decomposition of the trajectory matrix becomes $\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_G}$. Much of the literature on SSA is concerned with recognizing and interpreting the structure of the series inherent in such groupings, but it is clear that a preliminary and fundamental issue that must be overcome is how to determine the binary decomposition $\mathbf{X} = \mathbf{X}_{I_\mathscr{S}} + \mathbf{X}_{I_\mathscr{N}}$ where the group $I_\mathscr{S}$ corresponds to the signal and the group $I_\mathscr{N}$ corresponds to the noise components of the observed series. It is the automatic identification of this signal–plus–noise decomposition that our paper is designed to address.

From the preceding discussion it is apparent that SSA depends upon two basic parameters that must be assigned or chosen by the practitioner, namely, the window length of the embedding and the index set that defines the signal component to be used in the reconstruction of the "noise–free" series. Golyandina et al. (2001) recommend the selection of a value for m large enough to ensure that the signal and noise components are easily (in the terminology of SSA) *separated*. To achieve this goal they propose computing a weighted correlation between the signal and noise parts of the reconstructed series once the signal and noise groupings have been determined, a small correlation indicating that strong separability has been obtained. The signal–noise groupings are determined via a singular spectrum evaluation procedure that employs pattern recognition techniques (Golyandina et al., 2001, sec. 1.6) and methods similar to those used in standard principal component analysis (the use of the scree–plot and various correlation methods as described in Jolliffe, 2002, chap. 6). The difficulty with this approach is that in the absence of clear cut statistical decision rules and with few guidelines on how to set appropriate thresholds, the modeling involves substantial subjective assessment. Software available at `http://www.gistatgroup.com` implements the methods outlined in Golyandina et al. (2001) via interactive manual inspection.

In this paper we develop an objective data driven model specification technique in which m is determined as part of the model selection process and is identified from the data along with the dimension of the signal. The identification procedure is based upon the use of a model selection criterion structured in terms of a maximized pseudo log–likelihood with a penalty term for the model complexity. The rationale underlying the criterion represents an adaptation to SSA of the Minimum Description Length (MDL) principle for signal extraction, see Rissanen (2007), Hansen and Yu (2001) and Grünwald (2007).

The plan of the remainder of the paper is as follows. In the following section we outline the signal–plus–noise model that underlies our analysis and in Section 3 we use this model to motivate the construction of our MDL criterion function. In Section 4 we show that under very general regularity conditions the criterion will identify the true signal with probability one as the sample size N increases, and will choose the smallest value of m consistent with the true signal dimension. In Section 5 we indicate the modifications necessary to allow for mean corrections. Section 6 demonstrates the use of our identification procedure via (i) some simulation experiments that illustrate the practical impact of our theoretical results, and (ii) some empirical examples.

2 Signal-Noise Model

Suppose that $\{x(t) : t \in T\}$ is a zero mean stochastic process defined on a probability space $\mathscr{P} = \{\Omega, \mathscr{B}, P\}$, continuous in mean square, with the continuous covariance kernel $K(t,s) = \mathbb{E}[x(t)x(s)]$ on $T \times T$. By Mercer's theorem

$$K(t,s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t)\phi_j(s)$$

where the $\{\phi_j\}$ are continuous orthonormal eigenfunctions of K corresponding to the eigenvalues $\{\lambda_j\}$, namely

$$\int_T K(t,s)\phi_j(s)ds = \lambda_j \phi_j(t),$$

and the series converges uniformly and absolutely on $T \times T$. Moreover, since

$$K(t,t) = \sum_{j=1}^{\infty} \lambda_j |\phi_j(t)|^2$$

converges, the stochastic series $\sum_{j=1}^{\infty} z_j \phi_j(t)$ constructed using the random coefficients

$$z_j = \int_T x(t)\phi_j(t)dt \qquad j = 1, 2, \ldots \tag{4}$$

converges in mean square, by the Cauchy criterion, because

$$\mathbb{E}[z_j z_k] = \int_T \int_T K(t,s)\phi_j(t)\phi_k(s)dtds = \lambda_j \int_T \phi_j(t)\phi_k(t)dt = \begin{cases} 0, & j \neq k; \\ \lambda_j, & j = k, \end{cases}$$

implying that $\lim_{n,m\to\infty} \mathbb{E}[|\sum_{j=n+1}^{m} z_j\phi_j(t)|^2] = \lim_{n,m\to\infty} \sum_{j=n+1}^{m} \lambda_j|\phi_j(t)|^2 = 0$.

Now,

$$\mathbb{E}[z_j x(t)] = \mathbb{E}[\int_T \phi_j(s)x(s)x(t)ds] = \int_T \phi_j(s)K(s,t)ds = \lambda_j\phi_j(t),$$

and using Fatou's lemma we have

$$\begin{aligned}
\mathbb{E}[|\sum_{j=1}^{\infty} z_j\phi_j(t) - x(t)|^2] &= \mathbb{E}[\liminf_{k\to\infty} |\sum_{j=1}^{k} z_j\phi_j(t) - x(t)|^2] \\
&\leq \liminf_{k\to\infty} \mathbb{E}[|\sum_{j=1}^{k} z_j\phi_j(t) - x(t)|^2] \\
&= K(t,t) - \lim_{k\to\infty} \sum_{j=1}^{k} \lambda_j|\phi_j(t)|^2 \\
&= \lim_{k\to\infty} \sum_{j=k+1}^{\infty} \lambda_j|\phi_j(t)|^2 \\
&= 0,
\end{aligned}$$

uniformly in t. Thus $\sum_{j=1}^{k} z_j\phi_j(t)$ converges uniformly in mean square to $x(t)$ as $k \to \infty$ and the limiting expression

$$x(t) = \sum_{j=1}^{\infty} z_j\phi_j(t)$$

is known as the Karhunen-Loève expansion. Such processes are members of the, so called, Karhunen class (Rao, 1985).

If $x(t)$ fluctuates around a non-zero mean μ then we can repeat the above argument with $x(t)$ replaced by $x(t) - \mu$ to show that $x(t) - \mu$ belongs to the Karhunen class and $x(t)$ can be expanded as $x(t) = \mu + \sum_{j=1}^{\infty} z_j\phi_j(t)$. Now let us suppose that in passage to the limit given by the Karhunen-Loève expansion there exists a value of k such that the difference $x(t) - \mu - \sum_{j=1}^{k} z_j\phi_j(t)$ behaves as a weakly stationary white–noise process, so that we may write the observed process as

$$x(t) = \mu + \sum_{j=1}^{k} z_j \phi_j(t) + \nu(t) \tag{5}$$

where $\mathbb{E}[\nu(t)] = 0$ and the covariance kernel of $\nu(t)$ is

$$E[\nu(t)\nu(s)] = \begin{cases} 0, & t \neq s; \\ \sigma^2, & t = s. \end{cases} .$$

This yields a signal–plus–noise model for $x(t)$ in which the signal $s(t) = \mu + \sum_{j=1}^{k} z_j \phi_j(t)$ and the noise $\nu(t)$ are orthogonal by construction. Given that there exists a zero mean Gaussian process defined on $\mathscr{P}$ with covariance kernel $K(t,s)$ we will also assume that $x(t)$ is Gaussian, implying that the coefficients z_j are independently distributed as $N(0,\lambda_j)$ random variables, $z_j \sim N(0,\lambda_j)$, $j = 1,\ldots,k$, and are independent of $\nu(t) \sim N(0,\sigma^2)$ for all t.

To relate this model to SSA, note from (5) that if the model obtains the m-lagged vectors of the trajectory matrix can be written as

$$\mathbf{x}_i = \mu \mathbf{1}_m + \sum_{j=1}^{k} z_j \boldsymbol{\phi}_j(\mathbf{t}_i) + \boldsymbol{\nu}_i\,, \tag{6}$$

where

$$\mathbf{1}_m = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{\phi}_j(\mathbf{t}_i) = \begin{bmatrix} \phi_j(t_i) \\ \vdots \\ \phi_j(t_{i+m-1}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\nu}_i = \begin{bmatrix} \nu(t_i) \\ \vdots \\ \nu(t_{i+m-1}) \end{bmatrix}.$$

If the ϕ_j, $j = 1,\ldots,k$, are sufficiently smooth (smoothness of the dominant eigenfunctions is commonly supposed in SSA) then they will satisfy the Lipschitz condition $|\phi_j(t_i) - \phi_j(t_{i-1})| \leq M(t_i - t_{i-1})$ and $\|\boldsymbol{\phi}_j(\mathbf{t}_i) - \boldsymbol{\phi}_j(\mathbf{t}_{i-1})\| \leq \sqrt{m}M\triangle t$ where $\triangle t \leq (t_{\max} - t_{\min})/N$. Let $\boldsymbol{\varphi}_j$ be a point on the line segment joining $\boldsymbol{\phi}_j(\mathbf{t}_i)$ to $\boldsymbol{\phi}_j(\mathbf{t}_{i-1})$ and set $\zeta_j = (\boldsymbol{\varphi}_j'\boldsymbol{\varphi}_j)^{-1}\boldsymbol{\varphi}_j'\boldsymbol{\phi}_j(\mathbf{t}_i)z_j$. Then $\zeta_j\boldsymbol{\varphi}_j = z_j\boldsymbol{\phi}_j(\mathbf{t}_i)$ and (6) can be reexpressed in matrix–vector form as

$$\mathbf{x}_i = \mu \mathbf{1}_m + \boldsymbol{\Phi}\mathbf{z}_i + \boldsymbol{\nu}_i \tag{7}$$

where $\mathbf{z}_i = (\zeta_1, \ldots, \zeta_k)'$ and $\mathbf{\Phi} = [\varphi_1 : \cdots : \varphi_k]$ is an $m \times k$ matrix of functional values. Note from (7) that the orthogonality between $\mathbf{z}_i$ and $\boldsymbol{\nu}_i$ ensures that the signal and noise subspaces of $\mathbf{X}$ are theoretically separable.

Furthermore, $|1 - (\varphi_j'\varphi_j)^{-1}\varphi_j'\boldsymbol{\phi}_j(\mathbf{t}_i)| \leq \|\varphi_j - \boldsymbol{\phi}_j(\mathbf{t}_i)\|/\|\varphi_j\|^2$ and $\|\varphi_j - \boldsymbol{\phi}_j(\mathbf{t}_i)\| \leq \sqrt{m}M\triangle t$. Hence, if $\triangle t \to 0$ as $N \to \infty$ then when N is large (6) will be equivalent to (7) where, with a slight duplication of notation, $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{\Lambda})$ with $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\}$ independently of $\boldsymbol{\nu}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This specification generates a combined functional–structural relationship (Kendal and Stuart, 1979, chap. 29) for $\mathbf{x}_i$ with an $mk + k + 2$ element parameter vector $\boldsymbol{\theta}_{mk} = (\lambda_1, \ldots, \lambda_k, \sigma^2, \mu, \varphi_1', \ldots, \varphi_k')'$.

Now let $(x_{1i}, \ldots, x_{mi})'$, $(z_{1i}, \ldots, z_{ki})'$ and $(\nu_{1i}, \ldots, \nu_{mi})'$ denote realized values of $\mathbf{x}_i$, $\mathbf{z}_i$ and $\boldsymbol{\nu}_i$ respectively. Then the model in (7) implies that the likelihood of $\boldsymbol{\theta}_{mk}$ given $\mathbf{x}_i$ is

$$L(\boldsymbol{\theta}_{mk}|x_{1i}, \ldots, x_{mi}) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\lambda_j}} \exp\left\{-\frac{z_{ji}^2}{2\lambda_j}\right\} \prod_{l=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\nu_{li}^2}{2\sigma^2}\right\}, \tag{8}$$

where $\nu_{li} = x_{li} - \mu - \sum_{j=1}^{k} \varphi_{lj} z_{ji}$, giving an expression for the likelihood in terms of the parameters, the observed value of $\mathbf{x}_i$ and the unobserved values of the underlying random variables.

Recognition that both the columns and the rows of the trajectory matrix are sub–series of $\{x(t_1), x(t_2), \ldots, x(t_N)\}$ indicates that the exact likelihood for a given $\mathbf{X}$ must incorporate complex across component covariance constraints that would make manipulations involving the exact likelihood extremely difficult and cumbersome, if not intractable, both algebraically and numerically. To overcome such problems we will replace the exact likelihood by the product of the marginal likelihoods for each of the m–lagged vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, and consider what we will call the *rolling–window* likelihood

$$L(\boldsymbol{\theta}_{mk}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} L(\boldsymbol{\theta}_{mk}|x_{1i}, \ldots, x_{mi}). \tag{9}$$

The *rolling–window* likelihood implicitly replaces a joint density by a product of marginal densities, an adaptation to SSA of a technique previously employed by Poskitt and Zhang

(2005) to circumvent the complexities associated with evaluating the exact likelihood in hidden Markov models. Substituting (8) into (9) yields

$$\log L(\boldsymbol{\theta}_{mk}|\mathbf{x}_1,\ldots,\mathbf{x}_n) = -\frac{n}{2}\sum_{j=1}^{k}\left(\log(2\pi\lambda_j) + \frac{\sum_{i=1}^{n} z_{ji}^2}{2\lambda_j}\right) - \frac{nm}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}\sum_{l=1}^{m} v_{li}^2}{2\sigma^2} \qquad (10)$$

for the rolling–window log likelihood.

3 Minimum Description Length

For a given class models with their associated likelihoods Grünwald (2007, sec. 14.2) defines the *simple refined MDL model selection* procedure as the process of selecting the model that yields the greatest normalized maximum likelihood; or equivalently, in the notation of this paper, the process of selecting the model that minimizes the description length

$$DL(m,k) = -\log L(\widehat{\boldsymbol{\theta}}_{mk}|\mathbf{x}_1,\ldots,\mathbf{x}_n) + g(\kappa)h(N), \qquad (11)$$

where $\widehat{\boldsymbol{\theta}}_{mk}$ denotes the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_{mk}$, and $g(\cdot)$ and $h(\cdot)$ are non-decreasing functions of κ, the number of freely varying parameters in the model, and N, the total number of data points, respectively. The criterion $DL(k)$ embodies what is conventionally thought of as the tradeoff between the model fit and the model, or parameter, complexity, as measured by the negative of the log likelihood and the penalty term $g(\kappa)h(N)$ respectively.

In order to determine $\widehat{\boldsymbol{\theta}}_{mk}$ and the maximum of the rolling–window likelihood we proceed as follows. Concentrating $\log L(\boldsymbol{\theta}_{mk}|\mathbf{x}_1,\ldots,\mathbf{x}_n)$ with respect to σ^2, noting that

$$\sum_{i=1}^{n}\sum_{l=1}^{m} v_{li}^2 = \sum_{i=1}^{n}\|\mathbf{x}_i - \mu\mathbf{1}_m - \boldsymbol{\Phi}\mathbf{z}_i\|^2,$$

it is readily verified that

$$\widehat{\sigma}^2 = \frac{\min\sum_{i=1}^{n}\|\mathbf{x}_i - \mu\mathbf{1}_m - \boldsymbol{\Phi}\mathbf{z}_i\|^2}{nm}$$

where the minimum is taken over all possible values of μ, $\boldsymbol{\Phi}$ and $\mathbf{z}_i$, $i = 1,\ldots,n$. Set $\bar{\mathbf{x}} =$

$n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ and denote the $m \times n$ matrix whose ith column is $\mathbf{x}_i - \bar{\mathbf{x}}$ by $\mathbb{X}$. Let $\{\bar{\ell}_i, \bar{\mathbf{U}}_i, \bar{\mathbf{V}}_i\}$, $i = 1, \ldots, m$, be the eigentriples of $\mathbb{X}$. From Rao (1965, sec. 8g.2, Complements and Problems 1.1) it follows that

$$\begin{aligned}\sum_{i=1}^{n}\|\mathbf{x}_i - \mu\mathbf{1}_m - \boldsymbol{\Phi}\mathbf{z}_i\|^2 \;&\geq\; \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}} - \sum_{j=1}^{k}\bar{\mathbf{U}}_j(\mathbf{x}_i - \bar{\mathbf{x}})'\bar{\mathbf{U}}_j\|^2 + \|\bar{\mathbf{x}} - \widehat{\mu}\mathbf{1}_m\|^2 \\ &= \sum_{i=1}^{n}\|\mathbf{x}_i - \widehat{\mu}\mathbf{1}_m - \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{z}}_i\|^2 \qquad (12)\end{aligned}$$

where the MLE of the mean $\widehat{\mu} = m^{-1}\bar{\mathbf{x}}'\mathbf{1}_m = (nm)^{-1}\sum_{r=1}^{m}\sum_{c=1}^{n} x(t_{r+c-1})$. The natural estimator to take for $\boldsymbol{\Phi}$ is the matrix of empirical orthogonal functions, and setting $\widehat{\boldsymbol{\Phi}} = [\bar{\mathbf{U}}_1 : \ldots : \bar{\mathbf{U}}_k]$ with $\widehat{\mathbf{z}}_i = (\bar{\mathbf{U}}_1'(\mathbf{x}_i - \bar{\mathbf{x}}), \ldots, \bar{\mathbf{U}}_k'(\mathbf{x}_i - \bar{\mathbf{x}}))'$, the ith centered principle component, mimics the generation of the Karhunen-Loève coefficients in (4). Let $\bar{\ell}_0 = n\|\bar{\mathbf{x}} - \widehat{\mu}\mathbf{1}_m\|^2$. Then the minimum in (12) equals $\bar{\ell}_0 + \bar{\ell}_{k+1} + \cdots + \bar{\ell}_m$ and thus we can conclude that

$$\widehat{\sigma}^2 = \frac{1}{nm}\sum_{i=1}^{n}\|\mathbf{x}_i - \widehat{\mu}\mathbf{1}_m - \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{z}}_i\|^2 = \frac{1}{mn}(\bar{\ell}_0 + \sum_{j=k+1}^{m}\bar{\ell}_j).$$

Substituting $\widehat{\mathbf{z}}_i$ back into the score equations we then find that

$$\widehat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n}\widehat{z}_{ji}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{U}}_j'(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\bar{\mathbf{U}}_j = \frac{\bar{\ell}_j}{n}, \quad j = 1, \ldots, k.$$

Evaluating the rolling–window log likelihood at the point $\boldsymbol{\theta}_{mk} = \widehat{\boldsymbol{\theta}}_{mk}$ and substituting into (11) we get

$$DL(m,k) = \frac{n}{2}\left[(k+m)(1+\log 2\pi) + \sum_{j=1}^{k}\log\left(\frac{\bar{\ell}_j}{n}\right) + m\log\left(\frac{\bar{\ell}_0 + \sum_{j=k+1}^{m}\bar{\ell}_j}{nm}\right)\right] + g(\kappa)h(N) \qquad (13)$$

for the description length criterion where $\kappa = 2 + (2mk - k^2 + k)/2$. The value of kappa reflects the fact that the values of $\boldsymbol{\Phi}$ and $\mathbf{z}_i$, $i = 1, \ldots, n$, that maximize $L(\boldsymbol{\theta}_{mk}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are only uniquely defined up to multiplication by a $k \times k$ nonsingular matrix and we have selected a particular member from within the observational equivalence class by imposing $\frac{1}{2}k(k+1)$ parameter constraints.

REMARK: Under the assumption that $\mathbf{x}_i$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) observations from an m component Gaussian random vector, the likelihood ratio statistic for testing the null hypothesis that the smallest $m-k$ eigenvalues of the variance-covariance matrix are equal against the alternative that at least two are distinct is given by

$$-2\log LR(m,k) = n\left[(m-k)\log\left(\frac{\sum_{j=k+1}^{m}\bar{\ell}_j}{n(m-k)}\right) - \sum_{j=k+1}^{m}\log\left(\frac{\bar{\ell}_j}{n}\right)\right],$$

Anderson (1984, sec. 11.7.3). In conventional principle components analysis the use of the statistic $-2\log LR(m,k)$ to test for significant components – and determine the number of components that contribute substantial amounts to the overall variation – is sometimes referred to as Bartlett's test procedure, see Jolliffe (2002, sec. 3.7.3, 6.1.4). The decomposition

$$DL(m,k) = \frac{n}{2}\log\det\left(n^{-1}\mathbb{X}\mathbb{X}'\right) - \log LR(m,k) + CR(m,k),$$

where

$$\begin{aligned} CR(m,k) \quad = \quad & \frac{n}{2}\left[(k+m)\left(1+\log 2\pi\right) + k\log\left(\frac{\sum_{j=k+1}^{m}\bar{\ell}_j}{n(m-k)}\right)\right. \\ & \left. + m\log\left(\frac{m-k}{m}\right) + m\log\left(1+\frac{\bar{\ell}_0}{\sum_{j=k+1}^{m}\bar{\ell}_j}\right)\right] + g(\kappa)h(N), \end{aligned}$$

suggests, in light of Bartlett's test procedure, that the use of the MDL model selection process in the context of SSA can be likened to the application of a generalized likelihood ratio test with a data dependent critical value chosen as a function of the penalty term $g(\kappa)h(N)$. Indeed, working with i.i.d. observations on a Gaussian array with a fixed and known dimension, Wax and Kailath (1985) construct information theoretic signal extraction criteria that are precisely adjustments to the likelihood ratio statistic. ∎

The observations made in the previous remark motivate the following development. Let us divide $DL(m,k)$ into the sum of three parts, thus,

$$DL(m,k) = SDL(k|m) + WDL(m|k) + \left[m\log\left(\frac{m-k}{m}\right) + m\log\left(1+\frac{\bar{\ell}_0}{\sum_{j=k+1}^{m}\bar{\ell}_j}\right)\right] \quad (14)$$

where $SDL(k|m) = SLL(k|m) + \frac{1}{2}g(\kappa)h(N)$,

$$SLL(k|m) = \frac{n}{2}\left[m\left(1+\log 2\pi\right) + \sum_{j=1}^{k}\log\left(\frac{\bar{\ell}_j}{n}\right) + (m-k)\log\left(\frac{\sum_{j=k+1}^{m}\bar{\ell}_j}{n(m-k)}\right)\right],$$

and $WDL(m|k) = WLL(m|k) + \frac{1}{2}g(\kappa)h(N)$,

$$WLL(m|k) = \frac{n}{2}\left[k\left(1+\log 2\pi\right) + k\log\left(\frac{\sum_{j=k+1}^{m}\bar{\ell}_j}{n(m-k)}\right)\right].$$

Now consider selecting appropriate values for the window–length and the signal dimension. First, for each $k \in \{0,\ldots,M-1\}$, where M is preassigned, a window–length may be chosen as

$$\widehat{m}_k = \arg\min_{m=k+1,\ldots,M} WDL(m|k).$$

Second, for each $m \in \{2,\ldots,M\}$, a signal dimension can be chosen as

$$\widehat{k}_m = \arg\min_{k=0,\ldots,m-1} SDL(k|m).$$

To ensure compatibility between the two choices the specification where the window–length $\widehat{m}$ and dimension of the signal $\widehat{k}$ are given by the pair

$$(\widehat{m},\widehat{k}) = (\widehat{r}+1,\widehat{r}) \quad \text{where} \quad \widehat{r} = \arg_{r\in\{1,\ldots,M-1\}}\min(\widehat{m}_r - \widehat{k}_{r+1})$$

is selected. In what follows we will show that under appropriate regularity the statistic $\widehat{k}$ will equal a constant for all N sufficiently large, that constant being the true dimension of the signal, and the window length estimate $\widehat{m}$ will converge to the minimum embedding length of the signal, the minimum trajectory dimension consistent with the reproduction of the true signal.

4 Strong Consistency

The assumption that $x(t)$ is a Gaussian process was introduced above for convenience in the derivation of $DL(m,k)$ and so forth, but now we wish to examine the behaviour of the decision criterion under more general conditions. At this point, therefore, we will dispense with the normality assumption. In order for the results presented in this section to have broad applicability we state our basic assumption in generic form.

Assumption 1: The probability generating mechanism underlying the stochastic process $x(t)$ satisfies sufficient conditions to ensure that for any trajectory matrix window length $m \leq (\log N)^c$, $c < \infty$: First, $\bar{\mathbf{x}}$ converges to $\mu\mathbf{1}_m$ almost surely and $n^{-1}\mathbb{X}\mathbb{X}'$ converges to a positive definite limit, denoted by $\mathbf{\Gamma}$; Second, $\|\bar{\mathbf{x}} - \mu\mathbf{1}_m\| = O(Q_n)$ and $\|n^{-1}\mathbb{X}\mathbb{X}' - \mathbf{\Gamma}\| = O(Q_n)$ where $Q_n = \sqrt{\log\log n/n}$, $n = N - m + 1$, as $N \to \infty$.

Various results that facilitate the validation of Assumption 1 under different scenarios are currently available. If $x(t)$ is a weakly stationary and ergodic process, for example, then the first part of Assumption 1 follows directly. Moreover, if $x(t)$ has a rational spectrum and is driven by a martingale innovation process with finite fourth moment then the second part follows from Theorem 5.3.2 of Hannan and Deistler (1988). For the more general Karhunen class of processes, suppose that $x(t)$ gives rise to a trajectory matrix whose m–lagged vectors can be modeled as in (7). Then we have

$$\bar{\mathbf{x}} = \mu\mathbf{1}_m + \mathbf{\Phi}\bar{\mathbf{z}} + \bar{\boldsymbol{\nu}}$$

where $\bar{\mathbf{z}} = n^{-1}\sum_{i=1}^n \mathbf{z}_i$ and $\bar{\boldsymbol{\nu}} = n^{-1}\sum_{i=1}^n \boldsymbol{\nu}_i$, and

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{\Phi}\mathbf{z}_i + \boldsymbol{\nu}_i)(\mathbf{\Phi}\mathbf{z}_i + \boldsymbol{\nu}_i)' - (\mathbf{\Phi}\bar{\mathbf{z}} + \bar{\boldsymbol{\nu}})(\mathbf{\Phi}\bar{\mathbf{z}} + \bar{\boldsymbol{\nu}})'.$$

Now assume that $\mathbf{z}_i$ and $\boldsymbol{\nu}_i$ are near epoch dependent (mixing) processes that are such that $\bar{\mathbf{z}}$ and $\bar{\boldsymbol{\nu}}$ converge to zero at a rate governed by the law of the iterated logarithm, and similarly, $n^{-1}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i'$ and $n^{-1}\sum_{i=1}^n \boldsymbol{\nu}_i\boldsymbol{\nu}_i'$ converge to $\mathbf{\Lambda}$ and $\sigma^2\mathbf{I}$ respectively (Davidson, 1994, chap. 24). (Given the nature of the approximation inherent in (7), supposing that the stochastic structure of $\mathbf{z}_i$ and $\boldsymbol{\nu}_i$ follows that of near epoch dependent (mixing) processes seems natural.) Since $\mathbf{z}_i$ and $\boldsymbol{\nu}_i$ are orthogonal it follows that $x(t)$ will satisfy Assumption 1 with $\mathbf{\Gamma} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}' + \sigma^2\mathbf{I}$. Other examples of processes that satisfy Assumption 1 are presented below.

Lemma 1 : *Suppose that $x(t)$ satisfies Assumption 1 and let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m$ denote the ordered eigenvalues of $\mathbf{\Gamma}$. Then $\max_{j=1,\ldots,m}|\gamma_j - \bar{\ell}_j/n| = O(Q_n)$.*

Proof of Lemma 1: From the inequality $\sum_{j=1}^m \gamma_j\bar{\ell}_j \geq \mathrm{tr}(\mathbf{\Gamma}\mathbb{X}\mathbb{X}')$ (Anderson and Gupta, 1963) we have

$$\begin{aligned}\sum_{j=1}^{m}(\gamma_j - \frac{\bar{\ell}_j}{n})^2 &= \sum_{j=1}^{m}\gamma_j^2 - 2\sum_{j=1}^{m}\gamma_j\left(\frac{\bar{\ell}_j}{n}\right) + \sum_{j=1}^{T}\left(\frac{\bar{\ell}_j}{n}\right)^2 \\ &\leq \mathrm{tr}(\mathbf{\Gamma}^2) - 2n^{-1}\mathrm{tr}(\mathbf{\Gamma}\mathbb{X}\mathbb{X}') + n^{-2}\mathrm{tr}((\mathbb{X}\mathbb{X}')^2) \\ &= \|\mathbf{\Gamma} - n^{-1}\mathbb{X}\mathbb{X}'\|^2 .\end{aligned}$$

By Assumption 1 $\|\mathbf{\Gamma} - n^{-1}\mathbb{X}\mathbb{X}'\|^2$ is $O(Q_n^2)$, and from the inequality $\max_{1\leq j\leq m}(\gamma_j - \bar{\ell}_j/n)^2 \leq \sum_{j=1}^{m}(\gamma_j - \bar{\ell}_j/n)^2$ it follows that $\max_{1\leq j\leq m}|\gamma_j - \bar{\ell}_j/n| = O(Q_n)$, as required. □

Lemma 2 *: Suppose that $x(t)$ satisfies Assumption 1. Set*

$$\overline{SLL}(k|m) = \frac{n}{2}\left[m\left(1+\log 2\pi\right) + \sum_{j=1}^{k}\log\left(\gamma_j\right) + (m-k)\log\left(\frac{\sum_{j=k+1}^{m}\gamma_j}{(m-k)}\right)\right],$$

and let

$$\overline{WLL}(m|k) = \frac{n}{2}\left[k\left(1+\log 2\pi\right) + k\log\left(\frac{\sum_{j=k+1}^{m}\gamma_j}{(m-k)}\right)\right]$$

where $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m$ denote the ordered eigenvalues of $\mathbf{\Gamma}$. Then $|SLL(k|m) - \overline{SLL}(k|m)| = O(mnQ_n)$ uniformly in k and $|WLL(m|k) - \overline{WLL}(m|k)| = O(knQ_n)$ uniformly in m.

Proof of Lemma 2: The distance between $SLL(k|m)$ and $\overline{SLL}(k|m)$ is

$$\begin{aligned}|SLL(k|m) - \overline{SLL}(k|m)| &= \frac{n}{2}\left|\sum_{i=1}^{k}\log\left(\frac{\bar{\ell}_i}{n\gamma_i}\right) + (m-k)\log\left(\frac{\sum_{i=k+1}^{m}\bar{\ell}_i}{n\sum_{i=k+1}^{m}\gamma_i}\right)\right| \\ &\leq \frac{n}{2}\sum_{i=1}^{k}\left|\log\left(\frac{\bar{\ell}_i}{n\gamma_i}\right)\right| + \frac{n(m-k)}{2}\left|\log\left(\frac{\sum_{i=k+1}^{m}\bar{\ell}_i}{n\sum_{i=k+1}^{m}\gamma_i}\right)\right| . \quad (15)\end{aligned}$$

Now, from Lemma 1 we can deduce that

$$\max_{1\leq j\leq m}\left|\frac{\bar{\ell}_j}{n\gamma_j} - 1\right| = O(Q_n),$$

and similarly, that

$$\left|\frac{\sum_{i=k+1}^{m}\bar{\ell}_i}{n\sum_{i=k+1}^{m}\gamma_i} - 1\right| = O(Q_n)$$

uniformly in k. Since $\log(1+y) = O(y)$ as $y \to 0$ we can conclude that the upper bound on the right hand side of (15) is $O(mnQ_n)$ uniformly in k. The proof that $|WLL(m|k) - \overline{WLL}(m|k)| =$

$O(knQ_n)$ uniformly in m follows in like manner, the details are omitted. □

Proposition 1 *Assume that the data generating process is such that there exists a* $k = k_x$ *for which (i)* $x(t)$ *produces a trajectory matrix whose* m*–lagged vectors can be modeled as in (7), and (ii)* $x(t)$ *satisfies Assumption 1 with* $\mathbf{\Gamma} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}' + \sigma^2\mathbf{I}$*. Then for any* $m \in \{2, \ldots, M\}$ *where* $M \le (\log N)^c$*,* $c < \infty$*, the following statements hold with probability one:*

(1.1) If $\lim_{N\to\infty} h(N)/n = 0$ *then*

(1.1.1) $\lim_{N\to\infty} \widehat{k}_m = m - 1$ *when* $m \le k_x$*, and*

(1.1.2) $\liminf_{N\to\infty} \widehat{k}_m \ge k_x$ *when* $m > k_x$*.*

(1.2) If the components of the penalty term satisfy

$$\begin{cases} h(N') - h(N) & \ge 0,\ N' > N \text{ and} \\ g(\kappa') - g(\kappa) & \ge m,\ \kappa' > \kappa\,, \end{cases} \tag{16}$$

then $\limsup_{N\to\infty} \widehat{k}_m \le k_x$ *when* $m > k_x$*.*

Proof of Proposition 1: Set $\overline{SDL}(k|m) = \overline{SLL}(k|m) + \frac{1}{2}g(\kappa)h(N)$, where $m \in \{2, \ldots, M\}$, and let $\bar{k}_m = \arg\min_{k=0,\ldots,m-1} \overline{SDL}(k|m)$. By Lemma 2 we have

$$|SDL(k|m) - \overline{SDL}(k|m)| = |SLL(k|m) - \overline{SLL}(k|m)| = O(mnQ_n) \quad \text{a.s.}$$

uniformly in k. Since by assumption $m \le M \le (\log N)^c$, $c < \infty$, we can therefore conclude that $2|SDL(k|m) - \overline{SDL}(k|m)|/n = o(1)$ a.s. uniformly in k, which implies in turn that $Prob\left(\lim_{N\to\infty} |\bar{k}_m - \widehat{k}_m| > 0\right) = 0$.

If $\lim_{N\to\infty} h(N)/n = 0$, then for any $k' < k$

$$\begin{aligned} \frac{2}{n}\left[\overline{SDL}(k|m) - \overline{SDL}(k'|m)\right] &= \frac{2}{n}\left[\overline{SLL}(k|m) - \overline{SLL}(k'|m)\right] + \frac{(g(\kappa) - g(\kappa'))h(N)}{n} \\ &= \log\left[\frac{\gamma_1\gamma_2\cdots\gamma_k\left(\frac{1}{m-k}\sum_{i=k+1}^m \gamma_i\right)^{m-k}}{\gamma_1\gamma_2\cdots\gamma_{k'}\left(\frac{1}{m-k'}\sum_{i=k'+1}^m \gamma_i\right)^{m-k'}}\right] + o(1)\,. \end{aligned}$$

Using the inequality between the geometric mean and the arithmetic mean we can deduce that

$$\gamma_1\gamma_2\cdots\gamma_{k'}\cdots\gamma_k\left(\frac{1}{m-k}\sum_{i=k+1}^{m}\gamma_i\right)^{m-k} \leq \gamma_1\gamma_2\cdots\gamma_{k'}\left(\frac{1}{m-k'}\sum_{i=k'+1}^{m}\gamma_i\right)^{m-k'}, \quad (17)$$

with equality if and only if $\gamma_{k'+1} = \cdots = \gamma_m$. It is straightforward to verify that the ordered eigenvalues of $\mathbf{\Gamma} = \mathbf{\Phi\Lambda\Phi}' + \sigma^2\mathbf{I}$ are $\gamma_i = \upsilon_i + \sigma^2$, $i = 1, \ldots, m$, when $m \leq k_x$, and $\gamma_i = \upsilon_i + \sigma^2$, $i = 1, \ldots, k_x$, and $\gamma_i = \sigma^2$, $i = k_x+1, \ldots, m$, when $m > k_x$, where $\upsilon_1 > \upsilon_2 > \cdots > \upsilon_{\min(m,k_x)} > 0$ denote the ordered, nonzero eigenvalues of $\mathbf{\Phi\Lambda\Phi}'$.

It therefore follows that $\limsup_{N\to\infty} 2\left[\overline{SDL}(k|m) - \overline{SDL}(k'|m)\right]/n < 0$ for $k' < k \leq \min(m, k_x)$, and $\lim_{N\to\infty} 2\left[\overline{SDL}(k|m) - \overline{SDL}(k'|m)\right]/n = 0$ when $k > k' \geq k_x$. Thus we can infer that $\bar{k}_m = m-1$ when $m \leq k_x$, $\bar{k}_m = k_x$ when $m = k_x + 1$, and $\bar{k}_m \geq k_x$ when $m > k_x + 1$, and the same is true of $\widehat{k}_m$ since $\lim_{N\to\infty} |\bar{k}_m - \widehat{k}_m| = 0$ with probability one.

To prove the second part of the proposition it is now only necessary to consider the case where $m > k > k_x$. From (17) it follows that

$$\begin{aligned}\frac{2}{m}\left[\overline{SDL}(k|m) - \overline{SDL}(k_x|m)\right] &= \frac{2}{m}\left[\overline{SLL}(k|m) - \overline{SLL}(k_x|m)\right] + \frac{(g(\kappa) - g(\kappa_x))h(N)}{m} \\ &= \frac{(g(\kappa) - g(\kappa_x))h(N)}{m},\end{aligned}$$

which implies via the conditions imposed on $g(\kappa)$ that

$$\frac{2}{m}\left[\overline{SDL}(k|m) - \overline{SDL}(k_x|m)\right] > h(N) > 0\,.$$

Hence we can conclude that $\bar{k}_m \leq k_x$ for all N and thus that $\limsup_{N\to\infty} \widehat{k}_m \leq k_x$. $\square$

Proposition 2 *Assume that the data generating process is such that there exists a $k = k_x$ for which (i) $x(t)$ produces a trajectory matrix whose m–lagged vectors can be modeled as in (7), and (ii) $x(t)$ satisfies Assumption 1 with $\mathbf{\Gamma} = \mathbf{\Phi\Lambda\Phi}' + \sigma^2\mathbf{I}$. Then for any $k \in \{0, \ldots, M-1\}$ where $M \leq (\log N)^c$, $c < \infty$, the following statements hold with probability one:*

(2.1) If $\lim_{N\to\infty} h(N)/n = 0$ then $\liminf_{N\to\infty} \widehat{m}_k \geq k_x + 1$.

(2.2) If the components of the penalty term satisfy

$$\begin{cases} h(N') - h(N) & \geq 0,\ N' > N \text{ and} \\ g(\kappa') - g(\kappa) & \geq k,\ \kappa' > \kappa\,, \end{cases} \tag{18}$$

then $\lim_{N\to\infty} \widehat{m}_k = k+1$ *when* $k \geq k_x$.

Proof of Proposition 2: Let $\bar{m}_k = \arg\min_{m=k+1,\ldots,M} \overline{WDL}(m|k)$ where $\overline{WDL}(m|k) = \overline{WLL}(m|k) + \frac{1}{2}g(\kappa)h(N)$. Clearly $\bar{m}_k \geq k+1$ for all k, so obviously $\bar{m}_k \geq k_x + 1$ when $k \geq k_x$. Consider then, the case where $k < k_x$. If $\lim_{N\to\infty} h(N)/n = 0$ then

$$\begin{aligned} \frac{2}{n}\left[\overline{WDL}(m|k) - \overline{WDL}(m'|k)\right] &= \frac{2}{n}\left[\overline{WLL}(m|k) - \overline{WLL}(m'|k)\right] + \frac{(g(\kappa) - g(\kappa'))h(N)}{n} \\ &= k\log\left[\frac{\frac{1}{m-k}\sum_{i=k+1}^{m}\gamma_i}{\frac{1}{m'-k}\sum_{i=k+1}^{m'}\gamma_i}\right] + o(1)\,, \end{aligned}$$

and for any $k < k_x$ we have

$$\frac{1}{m-k}\sum_{i=k+1}^{m}\gamma_i = \sigma^2 + \frac{1}{m-k}\sum_{i=k+1}^{\min\{m,k_x\}} \upsilon_i < \sigma^2 + \frac{1}{m'-k}\sum_{i=k+1}^{\min\{m',k_x\}} \upsilon_i = \frac{1}{m'-k}\sum_{i=k+1}^{m'}\gamma_i\,.$$

It therefore follows that $\limsup_{N\to\infty} 2\left[\overline{WDL}(m|k) - \overline{WDL}(m'|k)\right]/n < 0$ for any $m > m' > k$. Thus we can conclude that $\bar{m}_k \geq k_x + 1$ when $k < k_x$.

Now consider the case where $m > m' > k \geq k_x$. When $k \geq k_x$ the average of the $m-k$ smallest eigenvalues $(m-k)^{-1}\sum_{i=k+1}^{m}\gamma_i = \sigma^2$ for all $m > k$ and it follows that

$$\begin{aligned} \frac{2}{k}\left[\overline{WDL}(m|k) - \overline{WDL}(m'|k)\right] &= \frac{2}{k}\left[\overline{WLL}(m|k) - \overline{WLL}(m'|k)\right] + \frac{(g(\kappa) - g(\kappa'))h(N)}{k} \\ &= \frac{(g(\kappa) - g(\kappa'))h(N)}{k}\,. \end{aligned}$$

The conditions imposed on $g(\kappa)$ now imply that

$$\frac{2}{k}\left[\overline{WDL}(m|k) - \overline{WDL}(m'|k)\right] > h(N) > 0$$

and hence that $\bar{m}_k \leq k+1$ for all N. We can therefore conclude that $\bar{m}_k = k+1$.

By Lemma 2, however,

$$|WDL(m|k) - \overline{WDL}(m|k)| = |WLL(m|k) - \overline{WLL}(m|k)| = O(knQ_n) \quad \text{a.s.}$$

uniformly in m, and by assumption $k < M \leq (\log N)^c$, $c < \infty$. It follows that $2|WDL(m|k) - \overline{WDL}(m|k)|/n = o(1)$ a.s. uniformly in m, which implies that $\lim_{N\to\infty} |\bar{m}_k - \widehat{m}_k| = 0$ with probability one. This completes the proof. □

Suppose the conditions of Proposition 2 hold and let $\mathbf{s}_i = (s(t_i), s(t_{i+1}), \ldots, s(t_{i+m-1}))'$ for $i = 1, 2, \ldots, n$ be the m–lagged vectors of $\mathbf{S} = [s(t_{r+c-1})]$, $r = 1, \ldots, m$, $c = 1, \ldots, n$, the trajectory matrix of the signal, where $\mathbf{s}_i = \mathbf{\Phi}\mathbf{z}_i$ and $n > m \geq k_x + 1$. Now let $\mathbf{P}$ be any $m \times (m - k_x)$ matrix whose columns span the null space of $\mathbf{\Phi}$. Then $\mathbf{P}'\mathbf{s}_i = \mathbf{0}$, and via a sequence of elementary row transformations we can express each of $s(t_{i+k_x}), \ldots, s(t_{i+m-1})$ as a linear combination of $s(t_i), \ldots, s(t_{i+k_x-1})$, implying that the signal satisfies a linear recurrence relation of order k_x. The $(k_x + 1)k_x$ elements in the sub–matrix $\mathbf{S}_{11} = [s(t_{r+c-1})]$, $r = 1, \ldots, k_x + 1$, $c = 1, \ldots, k_x$, can be used to determine the coefficients of the linear recurrence formula, from which all other elements of $\mathbf{S}$ can be generated. But the Hankel structure of $\mathbf{S}_{11}$ means that all the elements of $\mathbf{S}_{11}$, and hence $\mathbf{S}$, are uniquely defined by the $2k_x$ values $s(t_1), \ldots, s(t_{2k_x})$. Now, the Whitney embedding theorem states that any smooth k-dimensional manifold with $k > 0$ (that is also Hausdorff and second-countable) can be smoothly embedded in Euclidean $2k$-space. Thus the criterion $WDL(m|k)$ leads to the selection of the minimum window length consistent with the Whitney embedding theorem.

Lets us now examine the asymptotic behaviour of $\widehat{m}_k$ and $\widehat{k}_m$ for pairs $(m, k) \in \{2, \ldots, M\} \times \{0, \ldots, m-1\}$ assuming that the conditions in Propositions 1 and 2 hold. The properties leading to Theorem 1 below are illustrated in Figure 1, which depicts the hypothetical large sample values of $\widehat{m}_k$ and $\widehat{k}_m$ for (m, k) on the triangular grid $\{2, \ldots, 10\} \times \{0, \ldots, m-1\}$ supposing that $k_x = 4$. Consider, in general, the limiting value of the pair $(\widehat{m}_k, \widehat{k}_m)$. When $m \leq k_x$ it follows from (1.1.1) that $\lim_{N\to\infty} \widehat{k}_m = m - 1$ and when $m \geq k_x + 1$ it follows from (1.1.2) and (1.2) that $\lim_{N\to\infty} \widehat{k}_m = k_x$. Now, from (2.1) it follows that $\lim_{N\to\infty} \widehat{m}_k \geq k_x + 1$ for $k = 0, \ldots, k_x - 1$, and from (2.1) and (2.2) we have $\lim_{N\to\infty} \widehat{m}_k = k + 1$ when $k \geq k_x$. These asymptotic values for $\widehat{m}_k$ and $\widehat{k}_m$ imply that $\lim_{N\to\infty}(\widehat{m}_k, \widehat{k}_m) = (m, k)$ if, and only if,

						$\widehat{k}_m$					
	(m,k)	0	1	2	3	4	5	6	7	8	9
	2	·	$\widehat{k}_2$								
	3	·	·	$\widehat{k}_3$							
	4	·	·	·	$\widehat{k}_4$						
	5	$\widehat{m}_0$	$\widehat{m}_1$	$\widehat{m}_2$	$\widehat{m}_3$	$(\widehat{m},\widehat{k})$					
$\widehat{m}_k$	6	·	·	·	·	$\widehat{k}_6$	$\widehat{m}_5$				
	7	·	·	·	·	$\widehat{k}_7$	·	$\widehat{m}_6$			
	8	·	·	·	·	$\widehat{k}_8$	·	·	$\widehat{m}_7$		
	9	·	·	·	·	$\widehat{k}_9$	·	·	·	$\widehat{m}_8$	
	10	·	·	·	·	$\widehat{k}_{10}$	·	·	·	·	$\widehat{m}_9$

Figure 1: *Large sample values of $\widehat{m}_k$ and $\widehat{k}_m$. Hypothetical values for $M = 10$ and $k_x = 4$. At $(\widehat{m},\widehat{k}) = (5,4)$, $\widehat{m}_4 = 5$ and $\widehat{k}_5 = 4$.*

$(m,k) = (k_x + 1, k_x)$. We can also conclude that

$$\lim_{N\to\infty} (\widehat{m}_r - \widehat{k}_{r+1}) \begin{cases} \geq k_x + 1 - r > 1, & \text{when } r < k_x; \\ = k_x + 1 - k_x = 1, & \text{when } r = k_x; \\ = r + 1 - k_x > 1, & \text{when } r > k_x. \end{cases}$$

We have therefore established the following theorem.

Theorem 1 *Assume that the data generating process $x(t)$ satisfies the conditions of Propositions 1 and 2, and let $(\widehat{m},\widehat{k}) = (\widehat{r}+1,\widehat{r})$ where $\widehat{r} = \arg_{r\in\{1,\ldots,M-1\}} \min(\widehat{m}_r - \widehat{k}_{r+1})$ and $M \leq (\log N)^c$, $c < \infty$. Then if the components of the penalty term satisfy*

$$\begin{cases} h(N') - h(N) & \geq 0,\ N' > N\,, \\ \lim_{N\to\infty} \frac{h(N)}{n} & = 0\,, \qquad \textit{and} \\ g(\kappa') - g(\kappa) & \begin{cases} \geq m, & \textit{when } k' > k; \\ \geq k, & \textit{when } m' > m; \end{cases} \end{cases} \tag{19}$$

then $\lim_{N\to\infty}(\widehat{m},\widehat{k}) = (k_x + 1, k_x)$ with probability one.

REMARK: Note that under the conditions of Assumption 1 the third term in (14) will collapse to $-(r+1)\log(r+1)$ when $(m,k) = (r+1,r)$ since $\bar{\ell}_0 = o(1)$ almost surely as $N \to \infty$. Had this term been added to $SDL(k|m)$ its effect would have therefore disappeared asymptotically. Thus the consequences of neglecting this term will be negligible for all N sufficiently large. ∎

5 Mean Correction: Centered–SSA

In the introduction we indicated how basic SSA is implemented via an orthogonal decomposition of the trajectory matrix $\mathbf{X}$ using the eigentriples $\{\ell_i, \mathbf{U}_i, \mathbf{V}_i\}$, $i = 1, \ldots, m$, of $\mathbf{X}$ itself. The MDL selection process works in terms of the eigentriples $\{\bar{\ell}_i, \bar{\mathbf{U}}_i, \bar{\mathbf{V}}_i\}$, $i = 1, \ldots, m$, of the re–centered matrix $\mathbb{X}$, however, where the centering process subtracts the m-dimensional mean vector from each of the m–lagged vectors of $\mathbf{X}$ to give $\mathbb{X} = [\mathbf{x}_1 - \bar{\mathbf{x}} : \ldots : \mathbf{x}_n - \bar{\mathbf{x}}] = \mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n'$.

To reconcile the two, let $\mathbf{V}_0 = \mathbf{1}_n/\sqrt{n}$ and set $\ell_0 = n\|\bar{\mathbf{x}}\|^2$. Then direct calculation shows that $\mathbf{U}_0 = \mathbf{X}\mathbf{V}_0/\sqrt{\ell_0} = \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|$ and $\mathbf{V}_0'\mathbf{X}'\mathbf{X}\mathbf{V}_0 = \ell_0$; $\{\ell_0, \mathbf{U}_0, \mathbf{V}_0\}$ is often referred to as the *first average triple* of $\mathbf{X}$. Now, $\mathbb{X}\mathbf{V}_0 = \mathbf{X}\mathbf{V}_0 - \bar{\mathbf{x}}\mathbf{1}_n'\mathbf{V}_0 = \mathbf{0}$ and $\mathbf{V}_0'\mathbb{X}'\mathbb{X}\bar{\mathbf{V}}_i = \bar{\ell}_i\mathbf{V}_0'\bar{\mathbf{V}}_i = 0$, implying that $\mathbf{V}_0$ is orthogonal to $\bar{\mathbf{V}}_i$ for all i such that $\bar{\ell}_i \neq 0$. Thus $\bar{\mathbf{V}}_i$ for $i = 1, 2, \ldots, m$ and $\mathbf{V}_0$ form an orthonormal system and

$$\mathbf{X} = \mathbb{X} + \bar{\mathbf{x}}\mathbf{1}_n' = \sum_{i=1}^{m} \sqrt{\bar{\ell}_i}\bar{\mathbf{U}}_i\bar{\mathbf{V}}_i' + \sqrt{\ell_0}\mathbf{U}_0\mathbf{V}_0'$$

yields an alternative orthogonal decomposition of $\mathbf{X}$ such that $\|\mathbf{X}\|^2 = \ell_0 + \sum_{i=1}^{m} \bar{\ell}_i$. Centered–SSA now proceeds as for basic SSA by replacing $\{\ell_i, \mathbf{U}_i, \mathbf{V}_i\}$, $i = 1, \ldots, m$, by $\{\ell_0, \mathbf{U}_0, \mathbf{V}_0\}$ and $\{\bar{\ell}_i, \bar{\mathbf{U}}_i, \bar{\mathbf{V}}_i\}$, $i = 1, \ldots, m$. For further details see Golyandina et al. (2001, sec. 6.3).

The simple refined MDL signal–plus–noise Centered–SSA model is given by the specification $\mathbf{X} = \mathbf{X}_{I_{\mathscr{S}}} + \mathbf{X}_{I_{\mathscr{N}}}$ where the dimension of the trajectory matrix is $\widehat{m}$ and the group $I_{\mathscr{S}}$ corresponds to the first average triple plus the eigentriples $\{\bar{\ell}_i, \bar{\mathbf{U}}_i, \bar{\mathbf{V}}_i\}$, $i = 1, \ldots, \widehat{k}$, so that

$$\mathbf{X}_{I_{\mathscr{S}}} = \sqrt{\ell_0}\mathbf{U}_0\mathbf{V}_0' + \sum_{i=1}^{\widehat{k}} \sqrt{\bar{\ell}_i}\bar{\mathbf{U}}_i\bar{\mathbf{V}}_i'$$

and

$$\mathbf{X}_{I_{\mathscr{N}}} = \sum_{i=\widehat{k}+1}^{m} \sqrt{\bar{\ell}_i}\bar{\mathbf{U}}_i\bar{\mathbf{V}}_i' .$$

Thus the orthonormal system given by $\{\ell_0, \mathbf{U}_0, \mathbf{V}_0\}$ and $\{\bar{\ell}_i, \bar{\mathbf{U}}_i, \bar{\mathbf{V}}_i\}$, $i = 1, \ldots, m$, leads to a signal component that is deemed to be a constant plus a process of dimension $\widehat{k}$, mimicking the representation $s(t) = \mu + \sum_{j=1}^{k} z_j\phi_j(t)$ of the underlying signal derived from the Karhunen-Loève expansion.

Let us now consider constructing the description length criteria using the eigenvalues ℓ_j, $j =$

$1, \ldots, m$, of the basic trajectory matrix $\mathbf{X}$. Namely,

$$SDL(k|m) = \frac{n}{2}\left[m\left(1+\log 2\pi\right)+\sum_{j=1}^{k}\log\left(\frac{\ell_j}{n}\right)+(m-k)\log\left(\frac{\sum_{j=k+1}^{m}\ell_j}{n(m-k)}\right)\right]+\frac{1}{2}g(\kappa)h(N), \tag{20}$$

and

$$WDL(m|k) = \frac{n}{2}\left[k\left(1+\log 2\pi\right)+k\log\left(\frac{\sum_{j=k+1}^{m}\ell_j}{n(m-k)}\right)\right]+\frac{1}{2}g(\kappa)h(N). \tag{21}$$

Strictly speaking we should modify our notation so as to identify the different eigenvalues used to calculate the criteria, but before doing so let us continue to use generic notation and recall that $(\widehat{m}, \widehat{k}) = (\widehat{r}+1, \widehat{r})$ where $\widehat{r} = \arg_{r\in\{1,\ldots,M-1\}}\min(\widehat{m}_r - \widehat{k}_{r+1})$, $\widehat{m}_k = \arg\min_{m=k+1,\ldots,M} WDL(m|k)$ and $\widehat{k}_m = \arg\min_{k=0,\ldots,m-1} SDL(k|m)$.

Assumption 1 implies that $n^{-1}\mathbf{X}\mathbf{X}'$ converges to $\mu^2\mathbf{1}_m\mathbf{1}_m' + \boldsymbol{\Gamma}$. Let ρ_j, $j = 1, \ldots, m$, denote the eigenvalues of $\mu^2\mathbf{1}_m\mathbf{1}_m' + \boldsymbol{\Gamma}$. Using standard eigenvalue inequalities (Lütkepohl, 1996, sec. 9.13.3–(11), sec. 9.13.3–(12)) we can deduce that $\rho_j \geq \gamma_j$ for $j = 1, \ldots, m$ when $m \leq k_x + 1$, and $\rho_j \geq \gamma_j$ for $j = 1, \ldots, k_x + 1$ and $\rho_j = \sigma^2$ for $j = k_x + 2, \ldots, m$ when $m > k_x + 1$. We can now establish non-centered versions of Lemmas 1 and 2. Replacing γ_j by ρ_j, and $\bar{\ell}_j$ by ℓ_j, $j = 1, \ldots, m$, the statement of the lemmas and the steps in their proofs exactly parallel those of Lemmas 1 and 2. Proofs of the following propositions are now virtually identical to those of their centered counterparts, Propositions 1 and 2, once allowance is made for the two possibilities $\rho_{k_x+1} = \gamma_{k_x+1} = \sigma^2$ and $\rho_{k_x+1} > \gamma_{k_x+1} = \sigma^2$. We omit the details.

Proposition 3 *Assume that the conditions of Proposition 1 hold. Then the following statements hold with probability one:*

(1.1) If $\lim_{N\to\infty} h(N)/n = 0$ *then*

- *(1.1.1)* $\lim_{N\to\infty}\widehat{k}_m = m - 1$ *when (i)* $\rho_{k_x+1} = \sigma^2$ *and* $m \leq k_x$, *or when (ii)* $\rho_{k_x+1} > \sigma^2$ *and* $m \leq k_x + 1$.

- *(1.1.2)* $\liminf_{N\to\infty}\widehat{k}_m \geq k_x$ *when* $\rho_{k_x+1} = \sigma^2$ *and* $m > k_x$, *and* $\liminf_{N\to\infty}\widehat{k}_m \geq k_x + 1$ *when* $\rho_{k_x+1} > \sigma^2$ *and* $m > k_x + 1$

(1.2) If the components of the penalty term satisfy (16) *then* $\limsup_{N\to\infty}\widehat{k}_m \leq k_x$ *when* $\rho_{k_x+1} = \sigma^2$ *and* $m > k_x$, *and* $\limsup_{N\to\infty}\widehat{k}_m \leq k_x + 1$ *when* $\rho_{k_x+1} > \sigma^2$ *and* $m > k_x + 1$.

Proposition 4 *Assume that the conditions of Proposition 2 hold. Then the following statements hold with probability one:*

(2.1) If $\lim_{N\to\infty} h(N)/n = 0$ *then* $\liminf_{N\to\infty} \widehat{m}_k \geq k_x+1$ *when* $\rho_{k_x+1} = \sigma^2$*, and when* $\rho_{k_x+1} > \sigma^2$ $\liminf_{N\to\infty} \widehat{m}_k \geq k_x+2$.

(2.2) If the components of the penalty term satisfy (18) *then* $\lim_{N\to\infty} \widehat{m}_k = k+1$ *when (i)* $\rho_{k_x+1} = \sigma^2$ *and* $k \geq k_x$*, or when (ii)* $\rho_{k_x+1} > \sigma^2$ *and* $k \geq k_x+1$.

Henceforth, to distinguish between statistics calculated using the basic trajectory matrix $\mathbf{X}$ and those calculated using the centered version $\mathbb{X}$, we will add the superscript (B) for the former and (C) for the latter. When $\mu = 0$ we can immediately conclude from Propositions 1 and 2, and Propositions 3 and 4, that the centered and basic estimates are asymptotically equivalent and $Prob\left(\lim_{N\to\infty} |\widehat{m}_k^{(B)} - \widehat{m}_k^{(C)}| > 0\right)$ and $Prob\left(\lim_{N\to\infty} |\widehat{k}_m^{(B)} - \widehat{k}_m^{(C)}| > 0\right)$ equal zero because, obviously, $\rho_j = \gamma_j$ for all $j = 1, \ldots, m$, when $\mu = 0$. This gives us the first part of the following theorem.

Theorem 2 *Assume that the data generating process* $x(t)$ *satisfies the conditions of Propositions 1, 2, 3 and 4. For* $a = B, C$*, let* $(\widehat{m}^{(a)}, \widehat{k}^{(a)}) = (\widehat{r}^{(a)}+1, \widehat{r}^{(a)})$ *where* $\widehat{r}^{(a)} = \arg_{r\in\{1,\ldots,M-1\}} \min(\widehat{m}_r^{(a)} - \widehat{k}_{r+1}^{(a)})$ *where* $M \leq (\log N)^c$*,* $c < \infty$*, and suppose the components of the penalty term satisfy the conditions in* (19)*. Then with probability one;*

(i) $\lim_{N\to\infty} (\widehat{m}^{(B)}, \widehat{k}^{(B)}) - (\widehat{m}^{(C)}, \widehat{k}^{(C)}) = (0,0)$ *when* $\mu = 0$*, and*

(ii) $\lim_{N\to\infty} (\widehat{m}^{(B)}, \widehat{k}^{(B)}) - (\widehat{m}^{(C)}, \widehat{k}^{(C)})$ *equals either* $(0,0)$ *or* $(1,1)$ *when* $\mu \neq 0$.

Proof of Theorem 2: It is sufficient to consider the $\mu \neq 0$ case. When $\mu \neq 0$ the difference between the basic and the centered estimates depends on whether or not the k_x+1'th eigenvector of $\mathbf{\Gamma}$, ξ_{k_x+1} say, is orthogonal to $\mathbf{1}_m$. If $\mathbf{1}_m'\xi_{k_x+1} = 0$ then $\rho_{k_x+1} = \gamma_{k_x+1} = \sigma^2$, otherwise $\rho_{k_x+1} > \gamma_{k_x+1} = \sigma^2$, and we can use Propositions 3 and 4 to show that $\lim_{N\to\infty} (\widehat{m}^{(B)}, \widehat{k}^{(B)}) = (k_x+1, k_x)$ if $\rho_{k_x+1} = \sigma^2$, and $\lim_{N\to\infty} (\widehat{m}^{(B)}, \widehat{k}^{(B)}) = (k_x+2, k_x+1)$ if $\rho_{k_x+1} > \sigma^2$; in the same manner that Propositions 1 and 2 were used to prove Theorem 1. Theorem 2 follows since Theorem 1 indicates that $\lim_{N\to\infty} (\widehat{m}^{(C)}, \widehat{k}^{(C)}) = (k_x+1, k_x)$. □

For basic SSA, the simple refined MDL signal–plus–noise model is given by the specification $\mathbf{X} = \mathbf{X}_{I_{\mathscr{S}}} + \mathbf{X}_{I_{\mathscr{N}}}$ where the dimension of $\mathbf{X}$ is $\widehat{m}$, and $\mathbf{X}_{I_{\mathscr{S}}} = \sum_{i=1}^{\widehat{k}} \sqrt{\ell_i}\mathbf{U}_i\mathbf{V}_i'$ and $\mathbf{X}_{I_{\mathscr{N}}} = \sum_{i=\widehat{k}+1}^{\widehat{m}} \sqrt{\ell_i}\mathbf{U}_i\mathbf{V}_i'$, with $\widehat{m}$ and $\widehat{k}$ determined using the criteria $SDL^{(B)}(k|m)$ and $WDL^{(B)}(m|k)$

computed as in (20) and (21).

6 Numerical Illustrations

Our purpose in this section of the paper is to examine the empirical performance of the criteria using simulated and real world data series. Before doing so, however, we will briefly outline some alternative criteria that have been advanced in the literature that the practitioner might contemplate employing in SSA to identify the dimension of the signal. These alternative criteria will provide us with natural bench marks when investigating and illustrating the operational characteristics of $SDL^{(a)}(k|m)$, $a = B, C$.

6.1 Some Related Criteria

It is clear from the work of Grünwald (2007) that description length criteria are not uniquely defined, and a pioneering version of such criteria is that due to Rissanen. In basic SSA, as described above, the value of k used in the signal–plus–noise representation of the observed process is determined via the decomposition $\mathbf{X} = \sum_{i=1}^{k} \sqrt{\ell_i}\mathbf{U}_i\mathbf{V}_i' + \sum_{i=k+1}^{m} \sqrt{\ell_i}\mathbf{U}_i\mathbf{V}_i'$ of the trajectory matrix in terms of the eigentriples $\{\lambda_j, \mathbf{U}_j, \mathbf{V}_j\}$. If we think of this decomposition of $\mathbf{X}$ in terms of a multivariate regression, with m regressands, k orthonormal regressors, residual sum of squares $\sum_{i=k+1}^{m} \ell_i$ and explained sum of squares $\sum_{i=1}^{k} \ell_i$, then the description length criterion of Rissanen becomes

$$DL_R(k|m) = (n-k)m \log\left(\frac{\sum_{i=k+1}^{m} \ell_i}{(n-k)m}\right) + mk \log\left(\frac{\sum_{i=1}^{k} \ell_i}{mk}\right) - \log\left(\frac{mk}{(n-k)m}\right).$$

Similarly, upon multiplication by 2, the criterion of Hansen and Yu (2001) can be written as $DL_{HY}(k|m) = DL_R(k|m) - 2\log mk$. See Rissanen (2007, sec. 9.4) and Grünwald (2007, chap. 14) and the references contained therein.

More recently, Poskitt and Sengarapillai (2009) derived a description length criterion designed to maximize the signal–to–noise ratio for a given choice of k. Although they derive their criterion in the context of Functional Data Analysis, their arguments are couched in terms of the Karhunen-Loève expansion and can therefore be adapted to SSA. In the SSA context the

Poskitt and Sengarapillai (2009) criterion becomes

$$DL_{PS}(k|m) = n\log\left(\frac{\sum_{i=k+1}^{m}\ell_i}{\sum_{i=1}^{m}\ell_i}\right) + k\log\left(\frac{\sum_{i=1}^{k}\ell_i}{\sum_{i=k+1}^{m}\ell_i}\frac{[nm-\nu(k)]}{\nu(k)}\right) + \log(\nu(k)\,[mn-\nu(k)])$$

where $\nu(k) = k(m+1) - \frac{1}{2}k(k+1)$.

For Centered–SSA the criteria $DL_R(k|m)$, $DL_{HY}(k|m)$ and $DL_{PS}(k|m)$ can be calculated using $\bar{\ell}_i$, $i = 1,\ldots,m$, in place of ℓ_i, $i = 1,\ldots,m$. Note that all three are evaluated on the presumption that m is known. Their performance can therefore be compared with that of $SDL^{(a)}(k|m)$, $a = B, C$, for m preassigned, but they do not allow the window length to be simultaneously determined, as with the use of $SDL^{(a)}(k|m)$ in conjunction with $WDL^{(a)}(k|m)$.

6.2 Simulated Example

Let us consider an observed process $x(t)$ such that

$$x(t) = \mu + \sum_{r=1}^{p} A_r cos(\lambda_r t + \theta_r) + \epsilon(t)$$

where μ is the mean level of the signal, A_r the amplitude, λ_r the frequency (cycles per unit time), and θ_r the phase shift of the rth cosinusoid, and $\epsilon(t)$ is a white noise process with variance σ^2.

If the θ_r are independent and uniformly distributed over the interval $(-\pi,\pi)$ it is straightforward to show that $x(t)$ is a stationary process with mean μ and covariance kernel

$$E[x(t)x(s)] = \begin{cases} \mu^2 + \frac{1}{2}\sum_{r=1}^{p} A_r^2 cos(\lambda_r(t-s)) & \text{if } t \neq s; \\ \mu^2 + \frac{1}{2}\sum_{r=1}^{p} A_r^2 + \sigma^2 & \text{if } t = s. \end{cases}$$

Moreover, $x(t)$ will satisfy Assumption 1 with

$$\mathbf{\Gamma} = \frac{1}{2}\sum_{r=1}^{p} A_r^2\left[\mathbf{c}_r\mathbf{c}_r' + \mathbf{s}_r\mathbf{s}_r'\right] + \sigma^2\mathbf{I}$$

where $\mathbf{c}_r = [1, \cos(\lambda_r), \cdots, \cos((m-1)\lambda_r)]'$ and $\mathbf{s}_r = [1, \sin(\lambda_r), \cdots, \sin((m-1)\lambda_r)]'$. The rank of $\mathbf{\Gamma} - \sigma^2\mathbf{I}$ is $2p$, the dimension of the cosinusoidal signal component.

Now, the signal–to–noise ratio is

$$SNR = \frac{\mu^2 + \frac{1}{2}\sum_{r=1}^{p} A_r^2}{\sigma^2},$$

and for known mean μ and amplitudes $A_1, \ldots, A_p$ simulated realizations from processes with different pre–assigned signal–to–noise ratios can be generated by setting the noise variance

$$\sigma^2 = \frac{\mu^2 + \frac{1}{2}\sum_{r=1}^{p} A_r^2}{SNR}.$$

In our experiments we employed $\mu = 0.25$, $p = 2$ with $A_1 = 1.0$, $A_2 = 0.5$ and $\lambda_1 = 2\pi/7$, $\lambda_2 = 2\pi/10$. The noise process was i.i.d. Gaussian with variance $\sigma^2 = 0.6875/SNR$, $SNR = 3.0, 2.0, 1.0, 0.6, 0.4$.

6.2.1 Window Length and Proportion of Correct Selection

In order to implement the selection process outlined above the practitioner must assign values to $g(\kappa)$, $h(N)$ and M. Following what is now common practice in other fields we will set $g(\kappa) = \kappa$ and $h(N) = \log n$, this corresponds to using an MDL, SSA version of Schwarz's Bayesian information criterion (Grünwald, 2007, sec. 17.3). The choice for the maximum window length is more problematic.

In general the window length is assumed to satisfy $2 \le m \le \frac{N}{2}$, and from the previous theoretical development we require that $M \le (\log N)^c$, $c < \infty$. Our aim here is to determine an interval (c_L, c_U) such that for $m = (\log N)^c$ with $c \in (c_L, c_U)$ the proportion of times $SDL^{(a)}(k|m)$, $a = B, C$, select the correct dimension is maximized. Thus we consider a sequence of possible window lengths given by the Fibonacci sequence $10, 15, 25, \ldots$ and compute the simulated proportion of correct selection. For our cosinusoidal signal $k_x = 4$ and the proportion of correct selection is defined by

$$Pr(\tilde{k} = k) = \frac{\sum_{r=1}^{R} \chi_r\{\tilde{k} = k\}}{R}$$

where R is the number of simulated replications, 25000, and $\chi_r\{\tilde{k} = k\}$ indicates that $\tilde{k}$ selects dimension k on replication r, $k = k_x = 4$ for $\tilde{k} = \widehat{k}_m^{(C)}$ and $k = k_x + 1 = 5$ for $\tilde{k} = \widehat{k}_m^{(B)}$ (See Propositions 1 and 3, and Theorem 2).

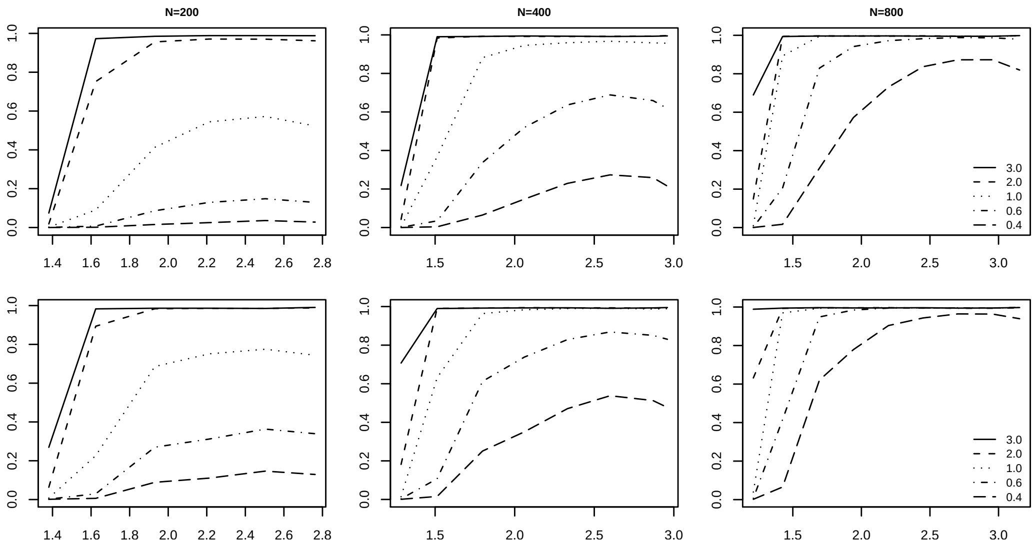

Figure 2: *Simulated proportion of correct selection for* $m = (\log N)^c$*: The top panels show* $SDL^{(B)}(k|m)$*, the bottom panels* $SDL^{(C)}(k|m)$*, the left hand panels for sample size* $N = 200$*, the center panels for* $N = 400$ *and the right hand panels for* $N = 800$*. Results are processed from 25000 replications for Gaussian noise with SNR 3.0, 2.0, 1.0, 0.6, and 0.4.*

Figure 2 displays the proportion of correct specification as a function of c. For any given c the proportion of correct selection is clearly smaller for the smaller signal-to-noise ratio. However, it is also apparent that for window lengths $m = (\log N)^c$ with $c \in (1.5, 2.5)$ these proportions are more or less stable, and they all exceed 90% by the time $N = 800$ for all but the very noisy processes ($SNR = 0.6, 0.4$). This suggests that a simple practical rule might be to set the maximum window length $M = (\log N)^c$ with $c \in (1.5, 2.5)$. When $c = 1.5$ and $N = 200$, this gives $M = 12 \approx 0.06N$, and when $c = 2.5$ and $N = 800$, $M = 115 \approx 0.14N$, giving bounds on window length that are noticeably smaller than are conventionally recommended?

6.2.2 Comparative Study

Figure 2 indicates that $SDL^{(B)}(k|m)$ performs well provided that a sensible choice of window length is used. To compare this criterion with the alternative criteria $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ we examined the case $N = 200$ with $m = 40$. This value of m equals $(\log N)^c$ with $c = 2.2124$. The average value of $SDL^{(B)}(k|m)$, $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$, computed from 25000 replications, with SNR set equal to 3.0, 2.0, 1.0, 0.6 and 0.4, are graphed in Figure 3.

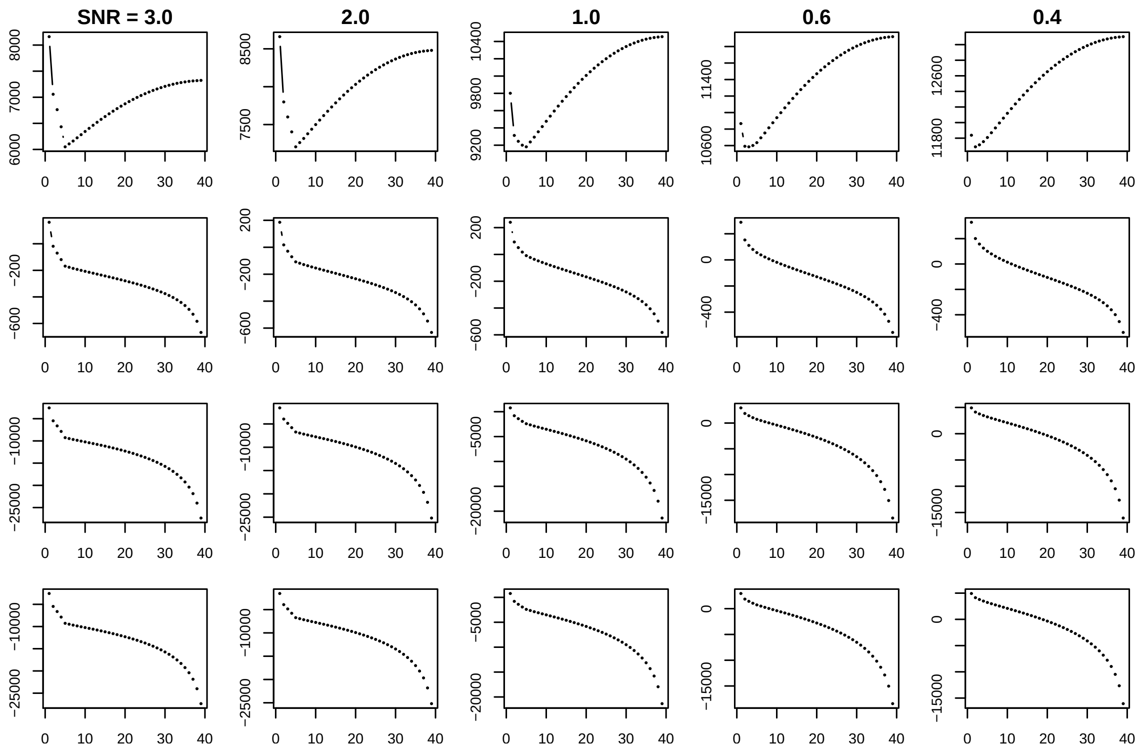

Figure 3: *Expected value of criterion function: The top row shows $SDL^{(B)}(k|m)$, the second row $DL_{PS}(k|m)$, the third row $DL_R(k|m)$ and the bottom row $DL_{HY}(k|m)$. Average values computed from 25000 replications for $N = 200$, $m = 40$ and SNR 3.0, 2.0, 1.0, 0.6 and 0.4.*

Not unexpectedly, we find that the characteristics of $DL_R(k|m)$ and $DL_{HY}(k|m)$ match each other, and, perhaps rather more surprisingly, we also find that $DL_{PS}(k|m)$ behaves similarly. Although these criteria hint at the true dimension by locating a kink or point of inflexion at $k = k_x + 1 = 5$, all three will be misleading if applied automatically due to their downward trend with increasing numbers of components. Each of these criteria will tend to select the saturated model with $k = m - 1$, irrespective of the SNR value.

On the other hand, $SDL^{(B)}(k|m)$ reaches a well defined global minimum with $k \ll m - 1$ in all cases. When SNR is small and the process is very noisy it is clear that at this sample size $SDL^{(B)}(k|m)$ will tend to underestimate the true dimension, but the criterion minimum is at the true dimension of the data generating process whenever SNR is sufficiently large. Even for relatively noisy data ($SNR = 1.0$) the $SDL^{(B)}(k|m)$ curve has a sharp minimum at $k = k_x + 1 = 5$.

Very similar conclusions concerning the characteristics of the different criteria are also obtained in the centered case. We do not present all the details here, but in order to give some idea of the similarity, and to illustrate the workings of Propositions 1 and 3, and Theorem 2, we present in Figure 4 the average values of $SDL^{(a)}(k|m)$, $a = B, C$. The overall appearance

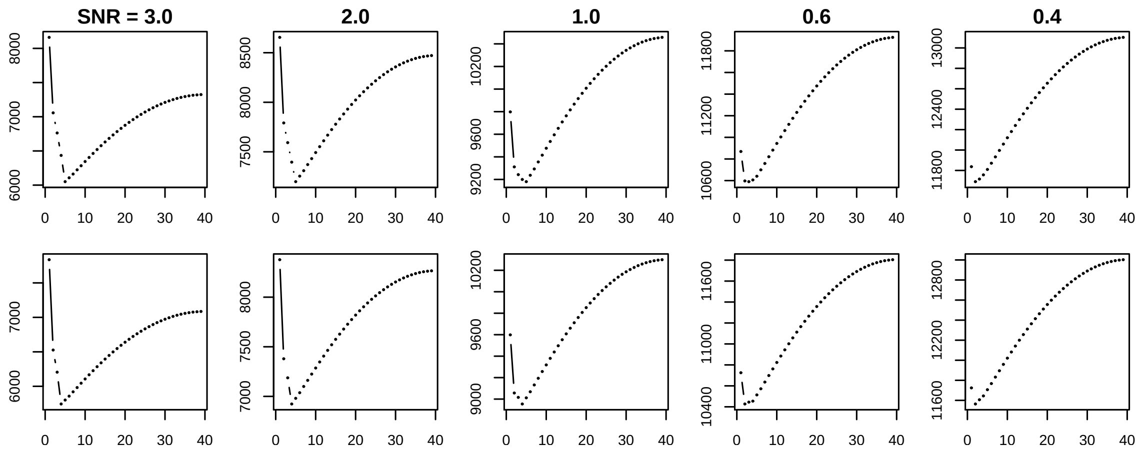

Figure 4: *Expected value of criterion function: Basic and centered SSA, the top panel shows $SDL^{(B)}(k|m)$ and the bottom panel $SDL^{(C)}(k|m)$. Average values computed from 25000 replications for $N = 200$, $m = 40$ and SNR 3.0, 2.0, 1.0, 0.6 and 0.4.*

of the graphs of $SDL^{(B)}(k|m)$ and $SDL^{(C)}(k|m)$ is such as to make them virtually indistinguishable on the basis of a perfunctory glance, but the curves differ in that the minimum is at $k = k_x + 1 = 5$ for $SDL^{(B)}(k|m)$ and at $k = k_x = 4$ for $SDL^{(C)}(k|m)$ when SNR equals 3.0, 2.0, 1.0, at $k = 3$ and $k = 2$ when SNR is 0.6, and both are minimized at $k = 2$ when SNR is 0.4.

Increasing the sample size to $N = 400$ we find that both $SDL^{(B)}(k|m)$ and $SDL^{(C)}(k|m)$ are minimized at the correct dimension when SNR equals 3.0, 2.0, 1.0 and 0.6, and at $k = 4$ and $k = 3$, respectively, when SNR is 0.4. When $N = 800$ both the basic and centered versions of the MDL, SSA criterion are minimized at the correct dimension, even in the noisiest case (SNR=0.4). The characteristics of $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ remain little changed, however, whatever the sample size.

6.2.3 Determination of Window Length

In order to illustrate the workings of Proposition 2 we present in Figure 5 the average value of $WDL^{(B)}(m|k)$ for $m \in \{k+1,\ldots,M\}$, $k = 1,3,5,7$, with $M = 40$, in the case $N = 200$. The properties indicated in Proposition 2 are reflected in the observed behaviour. However, closer examination of the value of $WDL^{(B)}(m|k)$ when $k < k_x$ reveals that although $WDL^{(B)}(m|k)$ does have a local minimum at $m = k_x + 1 = 5$ the criterion will have a tendency to select

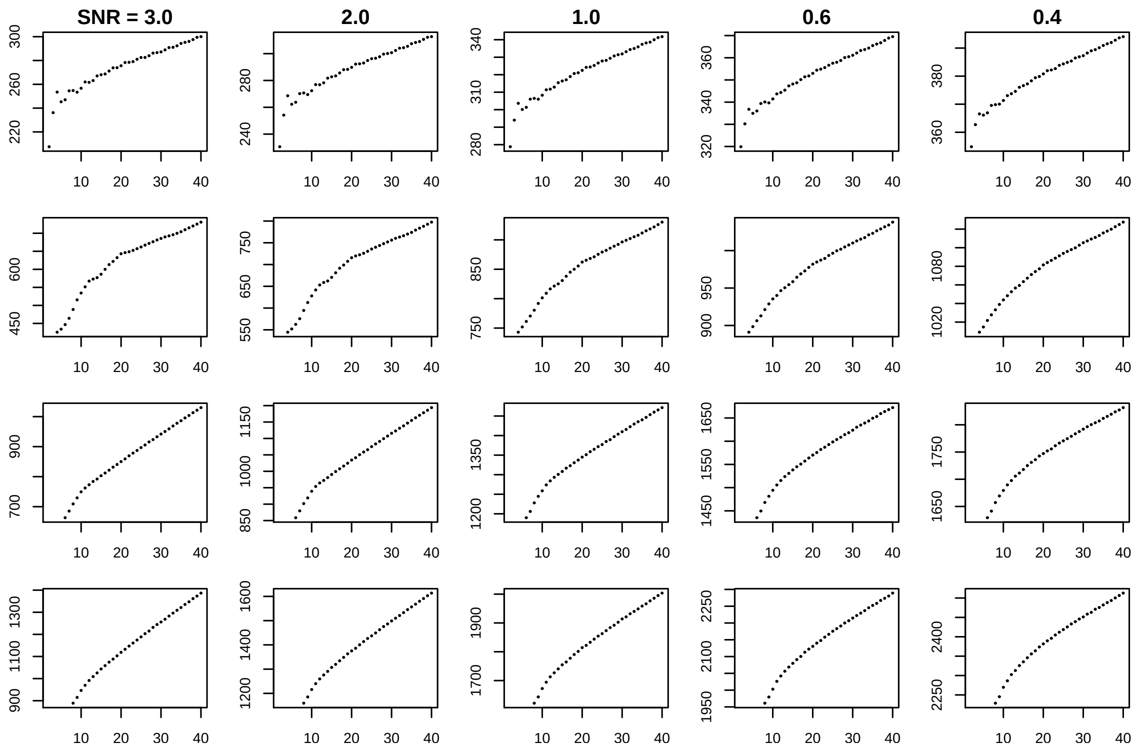

Figure 5: *Expected value of criterion function:* $WDL^{(B)}(m|k)$*, top to bottom panels give cases* $k = 1, 3, 5, 7$*. Average values computed from 25000 replications for* $N = 200$*,* $M = 40$ *and SNR 3.0, 2.0, 1.0, 0.6 and 0.4.*

$m = k+1$, in line with Whitney embedding theorem. Additional results not presented here also indicate that when $k < k_x$ and $N \leq 800$ the predictions of asymptotic theory are unlikely to manifest themselves with a high degree of regularity. Very similar characteristics are observed with $WDL^{(C)}(m|k)$. These results suggest that $WDL^{(a)}(m|k)$, $a = B, C$, will be of less practical use than is its counterpart $SDL^{(a)}(k|m)$, $a = B, C$, unless the sample size is very large, $N \gg 800$ say.

6.3 Real Data Sets

In this section we examine three data sets: (a) Accidental deaths in USA (monthly) from January 1973 to December 1978; (b) Monthly rose wine sales (thousand of litres) in Australia from July 1980 to June 1994; and (c) Number of daily births in Quebec, Canada from January 01, 1977 to December 31, 1990. (We have employed R routines to implement our algorithms. The first data set is readily available in R, and the others can be found at www.gistatgroup.com/cat/book2/bookdata.html.) Since we are still interested in comparing the behaviour of our criterion with $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ we have assigned fixed window lengths of 24, 36 and 112 for the accidental deaths, rose wine sales and daily births data respectively – in each case this corresponds to setting $m = (\log N)^c$ with $c = 2.2$.

Figure 6 graphs the values obtained for $SDL^{(a)}(k|m)$, $a = B, C$. Comparing these with those presented in Figure 4 we see that, as in the simulations, the values given by $SDL^{(B)}(k|m)$ and $SDL^{(C)}(k|m)$ are very nearly the same, apart from a possible difference in the location of the global minimum.

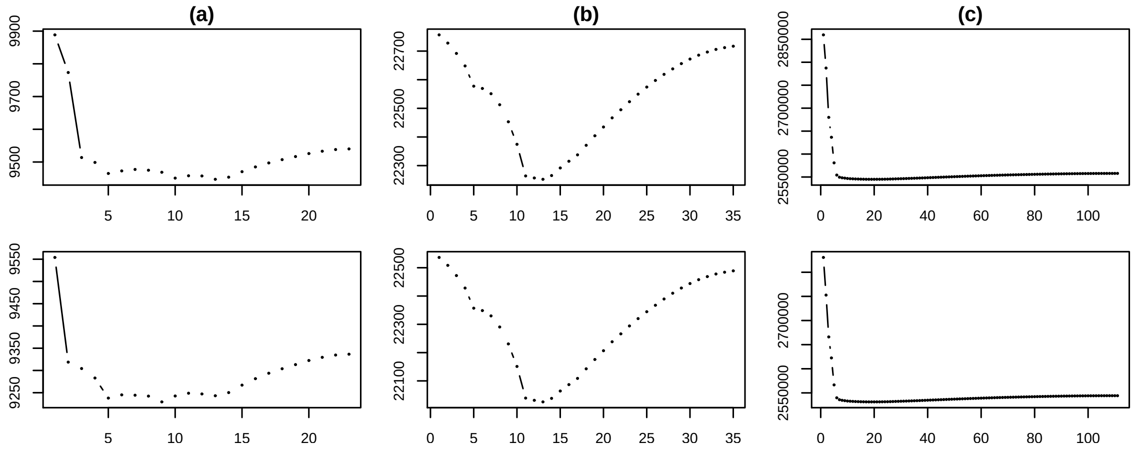

Figure 6: *Criterion functions for (a) accidental deaths, (b) rose wine sales and (c) daily births data. Upper panels give $SDL^{(B)}(k|m)$ and the bottom panels $SDL^{(C)}(k|m)$.*

It is of interest to note that using a window length of $m = 24$ in conjunction with standard SSA techniques – the use of auxiliary information, a scree plot, phase plots, periodogram analysis and separability evaluations – Hassani (2007) identifies a 12 eigentriple model for the USA accidental deaths data. The criterion $SDL^{(B)}(k|m)$ selects $k = 13$ directly. Both the basic and centered versions of the MDL criterion have three local minima with similar values however, and the global minimum of $SDL^{(C)}(k|m)$ occurs at $k = 9$, suggesting that the choice between $k = 13$ and $k = 9$ is not clear cut. Golyandina et al. (2001, sec. 1.4.1) applied conventional SSA with $m = 84$ to the rose wine sales data and determined a 14 component model; using a window length of $m = 36$ we find that $\widehat{k}_m^{(B)} = \widehat{k}_m^{(C)} = 13$. For the daily births data Golyandina et al. (2001, sec. 1.3.4)) set $m = 365$ and using standard SSA techniques chose $k = 19$. With $m = 112 = (\log N)^{2.2}$ both $SDL^{(B)}(k|m)$ and $SDL^{(C)}(k|m)$ also select $k = 19$ for the daily births data.

Figure 7 graphs the values of $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ for the three data sets. Once again properties previously seen in the simulations are repeated with the real world data sets, in particular, all three criteria behave in a very similar fashion, and if applied automatically they would lead to the selection of the saturated model (*cf*. Figure 3). As with $SDL^{(a)}(k|m)$, $a = B, C$, the three criteria indicate a more uncertain structure for the accidental deaths data, but unlike the former criteria, $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ offer very ambiguous guides to the appropriate signal–noise separation for the rose wine sales and daily births data.

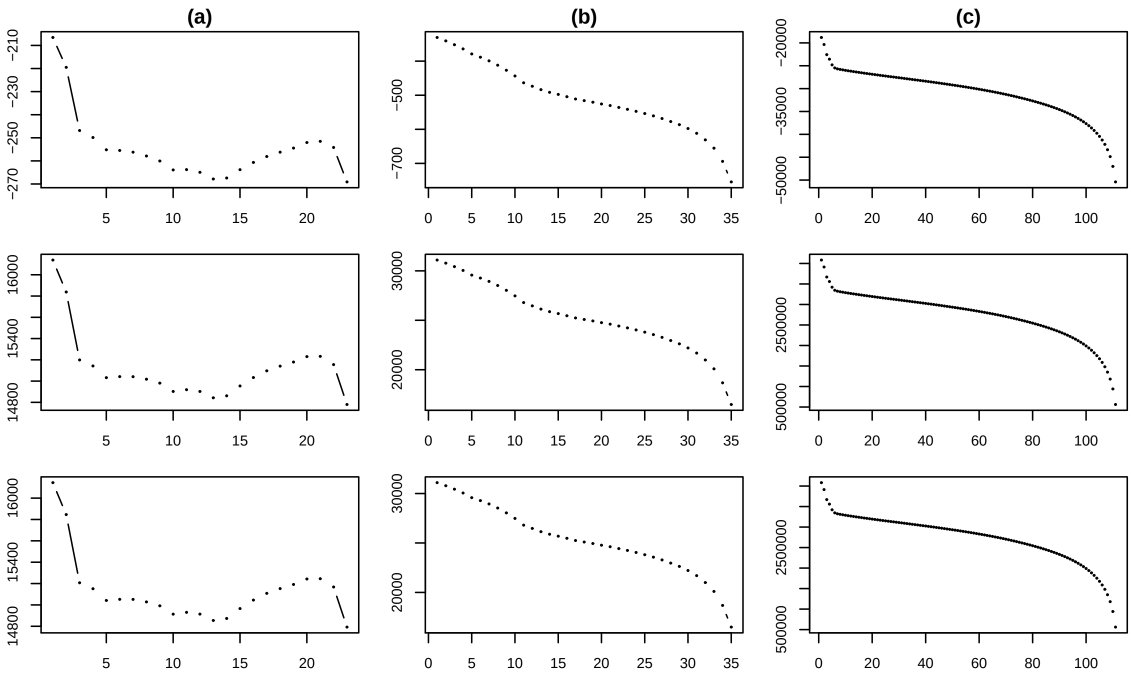

Figure 7: *Criterion functions for (a) accidental deaths, (b) rose wine sales and (c) daily births data: The first, second and third rows are $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ respectively for basic SSA.*

REMARK: That $DL_{PS}(k|m)$, $DL_R(k|m)$ and $DL_{HY}(k|m)$ appear to be ill conditioned with respect to signal–noise separation in SSA merits further comment. Starting at the origin, as k increases these criteria generally exhibit local turning points – points of inflection or local maxima or minima – before finally reaching a global minimum at the saturation boundary. This behaviour presents a problem when searching automatically across k. The phenomenon is due to the fact that all three criteria can be expressed as functions of the residual mean square, represented in the form of the arithmetic mean of the smallest eigenvalues. As $k \to m$

the residual mean square approaches zero and ultimately the increase in the penalty term is not large enough to counteract the decrease in the dominant term of the criteria brought about by the approach of the residual mean square to zero. The problem does not manifest itself with $SDL^{(B)}(k|m)$ and $SDL^{(C)}(k|m)$ since for these criteria the dominant term depends upon a balance between the arithmetic mean of the smallest eigenvalues and the geometric mean of the largest. ∎

7 Conclusion

Signal-noise separation is a critical issue in SSA, the quality of which depends upon two basic parameters that must be chosen by the practitioner, the window length of the embedding and the index set that defines the signal component. In this paper we have presented a minimum description length criterion that can be employed to automatically select both the window length and the signal. We showed that under very general regularity conditions the criterion will identify the true signal dimension with probability one as the sample size increases, and will choose the smallest window length consistent with the Whitney embedding theorem. Empirical results obtained using simulated and real world data sets indicate that the asymptotic theory is reflected in observed behaviour. Overall our results suggest that, other things being equal, the MDL, SSA criteria will favour the minimal null model as the signal–to–noise ratio decreases, but we can expect the true signal dimension to be obtained even when SNR is small provided the sample size is reasonably large.

References

Alonso, F. J., Castillo, J., and Pintado, P. (2005), "Application of Singular Spectrum Analysis to the Smoothing of Raw Kinematic Signals," *Journal of Biomechanics*, 38(5), 1085–1092.

Anderson, T. (1984), *Introduction to Multivariate Statistical Analysis,* New York: John Wiley.

Anderson, T. W., and Gupta, S. D. (1963), "Some Inequalities on Characteristic Roots of Matrices," *Biometrika*, 50(3-4), 522.

Basilevsky, A., and Hum, D. P. J. (1979), "Karhunen-Loève Analysis of Historical Time Series With an Application to Plantation Births in Jamaica," *Journal of the American Statistical*

Association, 74(366), 284–290.

Broomhead, D. S., and King, G. P. (1986), "Extracting Qualitative Dynamics from Experimental Data," *Physica D: Nonlinear Phenomena*, 20, 217–236.

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

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F. et al. (2002), "Advanced Spectral Methods for Climatic Time Series," *Rev. Geophys*, 40(1), 1003.

Golyandina, N., Nekrutkin, V. V., and Zhigljavski, A. A. (2001), *Analysis of Time Series Structure: SSA and Related Techniques,* Boca Raton: CRC Press.

Grünwald, P. D. (2007), *The Minimum Description Length Principle,* Cambridge: The MIT Press.

Hannan, E. J., and Deistler, M. (1988), *The Statistical Theory of Linear Systems,* New York: John Wiley.

Hansen, M. H., and Yu, B. (2001), "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, 96(454), 746–774.

Hassani, H. (2007), "Singular Spectrum Analysis: Methodology and Comparison," *Journal of Data Science*, 5(2), 239–257.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2009), "Forecasting European Industrial Production with Singular Spectrum Analysis," *International Journal of Forecasting*, 25, 103–118.

Hassani, H., and Zhigljavsky, A. (2009), "Singular Spectrum Analysis: Methodology and Application to Economics Data," *Journal of Systems Science and Complexity*, 22(3), 372–394.

Jolliffe, I. T. (2002), *Principal Component Analysis,* New York: Springer-Verlag.

Kendal, M., and Stuart, A. S. (1979), *The Advanced Theory of Statistics* (Vol. 2, 4th ed.), London: Griffin.

Lütkepohl, H. (1996), *Handbook of Matrices,* Chichester: John Wiley.

Marques, C. A. F., Ferreira, J. A., Rocha, A., Castanheira, J. M., Melo-Gonçalves, P., Vaz, N., and Dias, J. M. (2006), "Singular Spectrum Analysis and Forecasting of Hydrological Time

Series," *Physics and Chemistry of the Earth*, 31(18), 1172–1179.

Poskitt, D., and Sengarapillai, A. (2009), "Description Length and Dimensionality Reduction in Functional Data Analysis," *Monash Econometrics and Business Statistics Working Papers*, .

Poskitt, D., and Zhang, J. (2005), "Estimating Components in Finite Mixtures and Hidden Markov Models," *Australian & New Zealand Journal of Statistics*, 47(3), 269–286.

Rao, C. R. (1965), *Linear Statistical Inference and its Applications,* New York: John Wiley.

Rao, M. M. (1985), "Time Series in the Time Domain," *Handbook of Statistics*, Vol. 5, chap. 10, Harmonizable, Cramer and Kurhunen Classes of Processes, Amsterdam: North–Holland.

Rissanen, J. (2007), *Information and Complexity in Statistical Modeling,* New York: Springer-Verlag.

Thomakos, D. D., Wang, T., and Wille, L. T. (2002), "Modeling Daily Realized Futures Volatility with Singular Spectrum Analysis," *Physica A: Statistical Mechanics and its Applications*, 312(3-4), 505–519.

Vautard, R., and Ghil, M. (1989), "Singular Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time Series," *Physica D: Nonlinear Phenomena*, 35(3), 395–424.

Vautard, R., Yiou, P., and Ghil, M. (1992), "Singular-spectrum Analysis: A Toolkit for Short, Noisy Chaotic Signals," *Physica D: Nonlinear Phenomena*, 58(1-4), 95–126.

Wax, M., and Kailath, T. (1985), "Detection of Signals by Information Theoretic Criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2), 387–392.