ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

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

On The Theory and Practice of Singular Spectrum Analysis Forecasting

M. Atikur Rahman Khan and D.S. Poskitt

February 2014

Working Paper 03/14

On The Theory and Practice of Singular Spectrum Analysis Forecasting

M. Atikur Rahman Khan and D. S. Poskitt[*]

Abstract

Theoretical results on the properties of forecasts obtained using singular spectrum analysis are presented in this paper. The mean squared forecast error is derived under broad regularity conditions, and it is shown that the forecasts obtained in practice will converge to their population ensemble counterparts. The theoretical results are illustrated by examining the performance of singular spectrum analysis forecasts when applied to autoregressive processes and a random walk process. Simulation experiments suggest that the asymptotic properties developed are reflected in observed finite sample behaviour. Empirical applications using real world data sets indicate that forecasts based on singular spectrum analysis are competitive with other methods currently in vogue.

Keywords: Linear recurrent formula, Mean squared forecast error, Signal dimension, Window length.

JEL Classification: C51, C52, C53

1 Introduction

Singular spectrum analysis (SSA) is a nonparametric technique designed to be used for signal extraction and the prediction of irregular time series that may exhibit non–stationary and nonlinear properties, as well as intermittent or transient behaviour. The development of SSA is often attributed to researchers working in the physical sciences, namely Broomhead & King (1986), Vautard & Ghil (1989) and Vautard et al. (1992), although many of the basic building blocks were outlined by Basilevsky & Hum (1979) in a socioeconomic setting and an early formulation of some of the key ideas can be found in the work of Prony (1795). An introduction to SSA is presented in Elsner & Tsonis (1996) and a more detailed examination of the methodology with an emphasis on the algebraic structure and algorithms is available in Golyandina et al. (2001).

The application of SSA to forecasting has gained popularity over recent years, see for example Thomakos et al. (2002), Hassani et al. (2009), Hassani & Zhigljavsky (2009) and Hassani et al. (2010) for applications in business and economics, and the general finding appears to be that SSA performs well. In these studies SSA forecasts have been examined by investigating real world applications and comparing the performance of SSA to other benchmarks like $ARIMA$ models and Holt-Winters procedures. However, with real world data the true data generating mechanism is not known and making a comparison with such benchmarks does

[*] Corresponding author: D. S. Poskitt, Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia. Telephone:+61-3-99059378, Email: Donald.Poskitt@monash.edu.

not convey the full picture – to know that SSA outperforms a benchmark serves only to show that the benchmark is suboptimal and therefore that the benchmark does not provide an appropriate base line.

Our purpose in this paper is to provide what we believe to be the first theoretical analysis of the forecasting performance of SSA under appropriate regularity conditions concerning the true data generating mechanism. We present a formulation of the SSA mean squared forecast error ($MSFE$) for a general class of processes. The usefulness of such formulae lies not only in the fact that they provide a neat mathematical characterization of the SSA forecast error, but also in the fact that they allow a comparison to be made between SSA and the optimal mean squared error solution for a known random processes. The minimal mean squared error ($MMSE$) predictor obviously provides a (*gold*) standard against which all other procedures can be measured.

Irrespective of the actual structure of the observed process SSA forecasts are obtained by calculating a linear recurrence formula (LRF) that is used to construct a prediction of the future value(s) of the realized time series. Given a univariate time series of length N, the coefficients of the LRF are computed from a spectral decomposition of an $m \times n$ dimensional *Hankel* matrix known as the trajectory matrix. The dimension m is called the window length, and $n = N - m + 1$ is referred to as the window width. For a known window length the Gramian of the trajectory matrix is constructed and the eigenvalue decomposition of the Gramian evaluated. This is then used to decompose the observed series into a signal component, constructed from k eigentriples of the Hankel matrix (the first k left and right hand eigenvalues and their associated singular values), and a residual. The resulting signal plus noise decomposition is then employed to produce a forecast via the LRF coefficients. Details are presented in the following section where we outline the basic structure of the calculations underlying the construction of a $SSA(m, k)$ model and the associated forecasts.

Section 3 presents the theoretical $MSFE$ of a $SSA(m, k)$ model under very broad assumptions. The formulae that we derive indicate how the use of different values of m, a tuning parameter, and k, a modeling parameter, will interact to influence the $MSFE$ obtained from a given $SSA(m, k)$ model. In Section 4 it is shown that when appropriate regularity conditions are satisfied SSA forecasts constructed in practice, and their associated $MSFE$ estimates, will converge to their theoretical population ensemble counterparts.

Section 5 illustrates the theoretical results obtained in Sections 3 and 4 using simulation experiments based on autoregressive processes and a random walk. The examination of SSA forecasting presented in Section 5 indicates that for some processes (a simple autoregression) different $SSA(m, k)$ models will not achieve the same $MSFE$ performance as the MMSE predictor for any combination of window length and signal dimension, whereas for other processes (a random walk) the simplest $SSA(2, 1)$ model closely approximates the forecasting performance of the optimal predictor, to which it will converge as the effective sample size increases.

Section 6 demonstrates the application of SSA forecasting to different real world time series. It shows that SSA forecasts can exhibit considerable improvements in empirical $MSFE$ performance over conventional benchmark models that have been previously used to characterize these series. Section 7 presents a brief conclusion.

2 The Mechanics of SSA forecasting

Singular spectrum analysis (SSA) is based on the basic idea that there is an isomorphism between an observed time series $\{x(t) : t = 1, \ldots, N\}$ and the vector space of $m \times n$ Hankel matrices defined by the mapping

$$\{x(t) : t = 1, \ldots, N\} \mapsto \mathbf{X} = \begin{bmatrix} x(1) & x(2) & \ldots & x(n) \\ x(2) & x(3) & \ldots & x(n+1) \\ \vdots & \vdots & & \vdots \\ x(m) & x(m+1) & \ldots & x(N) \end{bmatrix} = [\mathbf{x}_1 : \ldots : \mathbf{x}_n], \qquad (1)$$

where m is a preassigned window length, $n = N - m + 1$, $\mathbf{x}_t = (x(t), x(t+1), \ldots, x(t+m-1))'$ and the so called trajectory matrix $\mathbf{X} = [x(i + t - 1)]$ for $i = 1, \ldots, m$ and $t = 1, \ldots, n$. Let $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_m > 0$ denote the eigenvalues of $\mathbf{XX}'$ arranged in descending order of magnitude and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ the corresponding orthonormal system of eigenvectors. The trajectory matrix can be expressed as $\mathbf{X} = \sum_{i=1}^{m} \mathbf{X}_i$, the sum of m rank one projections $\mathbf{X}_i = \sqrt{\ell_i}\mathbf{u}_i\mathbf{v}_i' = \mathbf{u}_i\mathbf{u}_i'\mathbf{X}$ where $\mathbf{u}_i$ and $\mathbf{v}_i = \mathbf{X}'\mathbf{u}_i/\sqrt{\ell_i}$, $i = 1, \ldots, m$, are the left and right eigenvectors of $\mathbf{X}$. Now suppose that a large proportion of the total variation in $\mathbf{XX}'$ can be associated with a subset of dominant eigentriples $\{\ell_i, \mathbf{u}_i, \mathbf{v}_i\}$, $i = 1, \ldots, k$. The projection of $\mathbf{X}$ onto the space spanned by $\mathbf{u}_i$, $i = 1, \ldots, k$, $\mathbf{S}_k = \sum_{i=1}^{k} \mathbf{X}_i$, can then be viewed as the component of $\mathbf{X}$ due to the presence of a signal in the original series, with k being the designated dimension of the signal, and the remainder $\mathbf{E}_k = \sum_{i=k+1}^{m} \mathbf{X}_i$ taken as the component due to noise. Henceforth we will refer to this as an $SSA(m, k)$ model.

Suppose that a $SSA(m, k)$ model has been fitted to time series data $x(1), x(2), \ldots, x(N)$. Since $\mathbf{S}_k$ has rank $k < m$ there exists an $m \times (m - k)$ matrix $\mathbf{P}$ whose columns span the null space of $\mathbf{S}_k$, implying that $\mathbf{P}'\mathbf{S}_k = \mathbf{0}$, and hence that the last row of $\mathbf{S}_k$ can be expressed as a linear combination of the first $m-1$ rows. This in turn implies that the signal satisfies, in the terminology of SSA, a linear recurrent formula (LRF), namely, $s(t) = \sum_{j=1}^{m-1} a_j s(t - m + j)$ where the coefficients $a_1, \ldots, a_{m-1}$ in the LRF are calculated by forming the projection of $\mathbf{S}_k^l$, the last row of the signal component $\mathbf{S}_k$, onto $\mathbf{S}_k^u$, its first $m - 1$ rows. The forecasts of $x(N + \tau)$ for $\tau = 1, \ldots, h$ are then generated sequentially from the recursions

$$\hat{x}(N + \tau|N) = \begin{cases} \sum_{i=1}^{m-1} a_i s_k(N + \tau - m + i)\,, \quad \tau = 1\,, \\ \sum_{i=1}^{\tau-1} a_{m-i}\hat{x}(N + \tau - i|N) + \sum_{i=1}^{m-\tau} a_i s_k(N + \tau - m + i)\,, \tau \leq m - 1\,, \\ \sum_{i=1}^{m-1} a_{m-i}\hat{x}(N + \tau - i|N)\,, \quad \tau > m - 1\,. \end{cases} \qquad (2)$$

Lemma 1 *Let* $\mathbf{U}_k = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ *denote the matrix containing the first* k *eigenvectors of* $\mathbf{XX}'$ *and let* $a_1, \ldots, a_{m-1}$ *denote the coefficients formed by projecting* $\mathbf{S}_k^l$*, the last row of*

$\mathbf{S}_k = \mathbf{U}_k\mathbf{U}_k'\mathbf{X}$, onto $\mathbf{S}_k^u$, its first $m-1$ rows. Then $(a_1, \ldots, a_{m-1}) = (1 - \mathbf{U}_k^l\mathbf{U}_k^{l'})^{-1}\mathbf{U}_k^l\mathbf{U}_k^{u'}$ where $\mathbf{U}_k^l$ is the last row $\mathbf{U}_k$ and $\mathbf{U}_k^u$ is the matrix containing the first $m-1$ rows.

For a more detailed exposition of the algebra underlying SSA modelling and forecasting we refer the interested reader to Golyandina et al. (2001). For current purposes it is sufficient to note that many of the ideas and techniques in SSA have been developed in a physical science/engineering context where signal extraction is the overriding objective, consequently SSA forecasting is often predicated on the assumption that $x(t) = s(t) + \varepsilon(t)$ where $\varepsilon(t)$ is a white noise process that is orthogonal to the signal $s(t)$. From the theory of stochastic processes we know that in this case forecasting $x(t)$ is equivalent to filtering combined with extrapolation, and that the minimum mean squared error (MMSE) approximation to $x(t+h)$ given $x(\tau)$, $\tau \leq t$, coincides with the MMSE approximation to $s(t+h)$ given $x(\tau)$, $\tau \leq t$. The formulation in (2) can thus be viewed as an application of the recursion $\widehat{s}_k(t) = \sum_{i=1}^{m-1} a_i\widehat{s}_k(t-m+i)$, wherein the a_i, $i = 1, \ldots, m-1$, are estimates of the coefficients in the MMSE linear projection of $s(t)$ onto the space spanned by $s(t-m), \ldots, s(t-1)$, and $\widehat{s}_k(\tau)$ is replaced by $\widehat{x}(\tau) = \widehat{s}_k(\tau)$ whenever $\tau > N$, and by $s_k(\tau)$ whenever $\tau \leq N$.

3 Theoretical Properties of SSA Forecasts

Following common practice in SSA, let us assume that the data-generating mechanism of the underlying stochastic process $x(t)$ is such that there exists a $k < m$ for which the $m-$lagged vectors of the trajectory matrix $\mathbf{X}$ can be modeled as

$$\mathbf{x}_t = \mathbf{\Phi}\mathbf{z}_t + \boldsymbol{\varepsilon}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t\,, \tag{3}$$

where $\mathbf{z}_t = (\zeta_{1t}, \ldots, \zeta_{kt})'$ and $\mathbf{\Phi} = [\boldsymbol{\varphi}_1 : \cdots : \boldsymbol{\varphi}_k]$ is an $m \times k$ coefficient matrix, $\mathbf{z}_t \sim (\mathbf{0}, \mathbf{\Lambda})$ with $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\}$ and is orthogonal to $\boldsymbol{\varepsilon}_t \sim (\mathbf{0}, \sigma^2\mathbf{I})$. The signal-plus-noise decomposition in (3) gives rise to the characterization $\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') = \mathbf{\Gamma} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}' + \sigma^2\mathbf{I}$. For any orthogonal transformation matrix $\mathbf{T}$ we can re-express (3) as

$$\mathbf{x}_t = \mathbf{\Psi}\mathbf{w}_t + \boldsymbol{\varepsilon}_t\,, \tag{4}$$

where $\mathbf{\Psi} = \mathbf{\Phi}\mathbf{T}$ and $\mathbf{w}_t = \mathbf{T}'\mathbf{z}_t$, implying that $\mathbf{\Gamma} = \mathbf{\Psi}\mathbf{T}'\mathbf{\Lambda}\mathbf{T}\mathbf{\Psi}' + \sigma^2\mathbf{I}$ and (3) and (4) are observationally equivalent.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ denote the eigenvectors of $\mathbf{\Gamma}$ and set $\mathbf{\Upsilon} = [\mathbf{\Upsilon}_k, \mathbf{\Upsilon}_{m-k}]$ where $\mathbf{\Upsilon}_k = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]$ and $\mathbf{\Upsilon}_{m-k} = [\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_m]$. It is straightforward to verify that the ordered eigenvalues of $\mathbf{\Gamma} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}' + \sigma^2\mathbf{I}$ are $\gamma_i = v_i + \sigma^2$ for $i = 1, \ldots, k$ and $\gamma_i = \sigma^2$ for $i = k+1, \ldots, m$, where $v_1 \geq \cdots \geq v_k$ are the ordered eigenvalues of $\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}'$. Set $\mathbf{V} = \mathrm{diag}(v_1, \ldots, v_k)$. Then

$$\mathbf{\Upsilon}'\mathbf{\Gamma}\mathbf{\Upsilon} - \sigma^2\mathbf{I} = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{\Upsilon}_k'(\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}')\mathbf{\Upsilon}_k & \mathbf{\Upsilon}_k'(\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}')\mathbf{\Upsilon}_{m-k} \\ \mathbf{\Upsilon}_{m-k}'(\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}')\mathbf{\Upsilon}_k & \mathbf{\Upsilon}_{m-k}'(\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}')\mathbf{\Upsilon}_{m-k} \end{bmatrix}. \tag{5}$$

Let $\mathbf{z}_t^* = \boldsymbol{\Upsilon}_k'\boldsymbol{\Phi}\mathbf{z}_t$. Then from (5) it follows that $\mathbb{E}(\mathbf{z}_t^*\mathbf{z}_t^{*'}) = \mathbf{V}$ and an equivalent representation of (3) is given by

$$\mathbf{x}_t = \boldsymbol{\Upsilon}_k\mathbf{z}_t^* + \boldsymbol{\varepsilon}_t^* = \mathbf{s}_t^* + \boldsymbol{\varepsilon}_t^* \,, \tag{6}$$

where $\mathbf{s}_t^* = \boldsymbol{\Upsilon}_k\mathbf{z}_t^* = \boldsymbol{\Upsilon}_k\boldsymbol{\Upsilon}_k'\boldsymbol{\Phi}\mathbf{z}_t = \boldsymbol{\Upsilon}_k\boldsymbol{\Upsilon}_k'\mathbf{s}_t$ and

$$\mathbb{E}(\mathbf{s}_t^*\mathbf{s}_t^{*'}) = \boldsymbol{\Upsilon}_k[\boldsymbol{\Upsilon}_k'\mathbb{E}(\mathbf{s}_t\mathbf{s}_t')\boldsymbol{\Upsilon}_k]\boldsymbol{\Upsilon}_k' = \boldsymbol{\Upsilon}_k[\boldsymbol{\Upsilon}_k'(\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}')\boldsymbol{\Upsilon}_k]\boldsymbol{\Upsilon}_k' = \boldsymbol{\Upsilon}_k\mathbf{V}\boldsymbol{\Upsilon}_k' = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}' \,.$$

Furthermore,

$$\begin{aligned}\boldsymbol{\varepsilon}_t^* = \mathbf{x}_t - \mathbf{s}_t^* \quad &= \mathbf{s}_t - \mathbf{s}_t^* + \boldsymbol{\varepsilon}_t \\ &= (\boldsymbol{\Upsilon}_k\boldsymbol{\Upsilon}_k' + \boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}')\mathbf{s}_t - \mathbf{s}_t^* + \boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}'\mathbf{s}_t + \boldsymbol{\varepsilon}_t \,,\end{aligned}$$

and since $\mathbb{E}(\mathbf{s}_t^*\mathbf{s}_t^{*'}) = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}' = \mathbb{E}(\mathbf{s}_t\mathbf{s}_t')$ we deduce from (5) that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\varepsilon}_t^*\boldsymbol{\varepsilon}_t^{*'}) \quad &= \boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}'\mathbb{E}(\mathbf{s}_t\mathbf{s}_t')\boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}' + \mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') \\ &= \boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}'(\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}')\boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}' + \sigma^2\mathbf{I} \\ &= \sigma^2\mathbf{I} \,,\end{aligned}$$

and $\mathbb{E}(\boldsymbol{\varepsilon}_t^*\mathbf{s}_t^{*'}) = \mathbb{E}[(\boldsymbol{\Upsilon}_{m-k}\boldsymbol{\Upsilon}_{m-k}'\mathbf{s}_t + \boldsymbol{\varepsilon}_t)(\boldsymbol{\Upsilon}_k\boldsymbol{\Upsilon}_k'\mathbf{s}_t)'] = \mathbf{0}$. Thus $\mathbb{E}(\mathbf{s}_t^*\mathbf{s}_t^{*'}) = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}' = \mathbb{E}(\mathbf{s}_t\mathbf{s}_t')$ and $\mathbb{E}(\boldsymbol{\varepsilon}_t^*\boldsymbol{\varepsilon}_t^{*'}) = \sigma^2\mathbf{I} = \mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')$. The representation of the $m-$lagged vectors in (3), and likewise in (4), is therefore observationally equivalent to

$$\mathbf{x}_t = \mathbf{s}_t^* + \boldsymbol{\varepsilon}_t^* = \boldsymbol{\Upsilon}_k\mathbf{z}_t^* + \boldsymbol{\varepsilon}_t^*, \tag{7}$$

where $\mathbf{s}_t^*$ and $\boldsymbol{\varepsilon}_t^*$ are orthogonal. Thus (7) provides a canonical representation for the equivalence class of (3) and (4). To avoid over complex notation we will henceforth suppress the asterisk in (7) and write $\mathbf{x}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t = \boldsymbol{\Upsilon}_k\mathbf{z}_t + \boldsymbol{\varepsilon}_t$ and so on, on the understanding that the model is represented in canonical form.

For the canonical representation we have $\boldsymbol{\Upsilon}_k\mathbf{z}_t = \mathbf{s}_t$ and multiplying both sides by $\boldsymbol{\Upsilon}_k'$ we obtain $\mathbf{z}_t = \boldsymbol{\Upsilon}_k'\mathbf{s}_t$, from which we deduce that

$$\mathbf{s}_t = \boldsymbol{\Upsilon}_k\boldsymbol{\Upsilon}_k'\mathbf{s}_t \,, \tag{8}$$

and recognising that the $m \times (m-k)$ matrix $\boldsymbol{\Upsilon}_{m-k}$ spans the null space of $\boldsymbol{\Upsilon}_k$ we have $\boldsymbol{\Upsilon}_{m-k}'\mathbf{s}_t = \boldsymbol{\Upsilon}_{m-k}'\boldsymbol{\Upsilon}_k\mathbf{z}_t = \mathbf{0}$, implying that $\mathbf{s}_t$ satisfies a LRF. If we now partition $\boldsymbol{\Upsilon}_k$ such that $\boldsymbol{\Upsilon}_k = \begin{pmatrix}\boldsymbol{\Upsilon}_k'^u \; \boldsymbol{\Upsilon}_k'^l\end{pmatrix}'$ where $\boldsymbol{\Upsilon}_k^l$ is the row vector of elements in the last row of $\boldsymbol{\Upsilon}_k$ and $\boldsymbol{\Upsilon}_k^u$ is a $(m-1) \times k$ matrix of the first $m-1$ rows of $\boldsymbol{\Upsilon}_k$, and we partition $\mathbf{s}_t$ conformable with the partition of $\boldsymbol{\Upsilon}_k$, we can re-express (8) as

$$\begin{pmatrix}\mathbf{s}_t^u \\ \mathbf{s}_t^l\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Upsilon}_k^u \\ \boldsymbol{\Upsilon}_k^l\end{pmatrix}\begin{pmatrix}\boldsymbol{\Upsilon}_k^{u'} & \boldsymbol{\Upsilon}_k^{l'}\end{pmatrix}\begin{pmatrix}\mathbf{s}_t^u \\ \mathbf{s}_t^l\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Upsilon}_k^u\boldsymbol{\Upsilon}_k^{u'}\mathbf{s}_t^u + \boldsymbol{\Upsilon}_k^u\boldsymbol{\Upsilon}_k^{l'}\mathbf{s}_t^l \\ \boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{u'}\mathbf{s}_t^u + \boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{l'}\mathbf{s}_t^l\end{pmatrix} . \tag{9}$$

The projection of the last element of the vector $\mathbf{s}_t$ on to its first $m-1$ elements is thus given by

$$\mathbf{s}_t^l \quad = \boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{u'}\mathbf{s}_t^u + \boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{l'}\mathbf{s}_t^l = (1 - \boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{l'})^{-1}\boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{u'}\mathbf{s}_t^u = \boldsymbol{\alpha}'\mathbf{s}_t^u, \tag{10}$$

where $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_{m-1}) = (1 - \boldsymbol{\Upsilon}_k^l \boldsymbol{\Upsilon}_k^{l\prime})^{-1} \boldsymbol{\Upsilon}_k^l \boldsymbol{\Upsilon}_k^{u\prime}$.

These coefficients yield the LRF

$$s(t) = \sum_{j=1}^{m-1} \alpha_j s(t - m + j), \tag{11}$$

which can be used to forecast $s(t+j)$, and hence $x(t+j)$, for $j = 1, 2, \ldots, h$. The one-step-ahead forecast is

$$s(t+1|t) = \sum_{j=1}^{m-1} \alpha_j s(t + 1 - m + j), \tag{12}$$

and for $j = 2, \ldots, h$ the forecasts are obtained recursively by using the equation

$$s(t+j|t) = \begin{cases} \sum_{i=1}^{j-1} \alpha_{m-i} s(t+j-i|t) + \sum_{i=1}^{m-j} \alpha_i s(t+j-m+i) & \text{for } j \leq m-1; \\ \sum_{i=1}^{m-1} \alpha_{m-i} s(t+j-i|t) & \text{for } j > m-1. \end{cases} \tag{13}$$

The development from (8) through (13) provides an obvious theoretical stochastic process parallel to the previous derivation of the empirical LRF forecasting formula. Extending the current theoretical development yields the following proposition wherein $\boldsymbol{\Gamma}_{m+h-1} = \mathbb{E}[\boldsymbol{\xi}_{m+h-1}\boldsymbol{\xi}_{m+h-1}']$ where $\boldsymbol{\xi}_{m+h-1} = [x(t-m+2), \ldots, x(t+h)]'$, and $\boldsymbol{\Sigma}_{m+h-1} = \mathbb{E}[\boldsymbol{\eta}_{m+h-1}\boldsymbol{\eta}_{m+h-1}'] = \text{diag}(\sigma^2 \mathbf{1}_{m-1}', \mathbf{0}_h')$ where $\boldsymbol{\eta}_{m+h-1} = [\varepsilon(t-m+2), \ldots, \varepsilon(t), \mathbf{0}_h']'$.

Proposition 1 *Suppose that the m-lagged vectors can be represented in canonical form* $\mathbf{x}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t = \boldsymbol{\Upsilon}_k \mathbf{z}_t + \boldsymbol{\varepsilon}_t$*, and that* $s(t+j|t)$ *is used to forecast* $x(t+j)$ *where* $s(t+j|t)$, $j = 1, 2, \ldots, h$*, are generated as in equations* (12) *and* (13)*. Let* $\varepsilon(t+j|t) = x(t+j) - s(t+j|t)$, $j = 1, 2, \ldots, h$*. Then the mean squared forecast error (MSFE)*

$$\mathbb{E}[\varepsilon(t+j|t)^2] = A_{t+j}(\boldsymbol{\Gamma}_{m+h-1} - \boldsymbol{\Sigma}_{m+h-1})A_{t+j}'$$

where $A_{t+1} = (-\boldsymbol{\alpha}', 1, \mathbf{0}_{h-1}')$ *and* A_{t+j} *are obtained recursively as follows:*

$$A_{t+j} = \begin{cases} (\mathbf{0}_{j-1}', -\boldsymbol{\alpha}', 1, \mathbf{0}_{h-j}') + \sum_{i=1}^{j-1} \alpha_{m-i} A_{t+j-i} & \textit{for } j \leq m-1, \\ (\mathbf{0}_{j-1}', -\boldsymbol{\alpha}', 1, \mathbf{0}_{h-j}') + \sum_{i=1}^{m-1} \alpha_{m-i} A_{t+j-i} & \textit{for } j > m-1. \end{cases}$$

The specification in (3) resembles a classical common factor model and, as shown in Watanabe (1965), it is directly related to the discrete Karhunen-Loéve expansion. In order to generalize the model and expand the previous results suppose now that $\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') = \boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} > 0$ but is otherwise unconstrained. As previously, let $\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_m$ denote the eigenvectors of $\boldsymbol{\Gamma}$, so $\boldsymbol{\Upsilon}'\boldsymbol{\Gamma}\boldsymbol{\Upsilon} = \text{diag}(\gamma_1, \ldots, \gamma_m)$ where $\gamma_1 \geq \ldots \geq \gamma_m > 0$ are the corresponding eigenvalues. Then the mean squared error (MSE) $\mathbb{E}[\|\mathbf{x}_t - \mathbf{s}_t\|^2]$ is minimized across all possible choices of $\mathbf{s}_t$ of dimension k by setting $\mathbf{s}_t = \sum_{i=1}^k \boldsymbol{\upsilon}_i \zeta_{it}$ where $\mathbf{z}_t = (\zeta_{1t}, \ldots, \zeta_{kt})' \sim (\mathbf{0}, \text{diag}\{\gamma_1, \ldots, \gamma_k\})$, with a MMSE of $\sum_{i=k+1}^m \gamma_i$ (See, for example, Rao 1965, §8.g Complements and Problems 1.1). The resulting decomposition $\mathbf{x}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t$ corresponds to the previous canonical form, only now $\boldsymbol{\varepsilon}_t$ is a colored noise process with covariance $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \sum_{i=k+1}^m \gamma_i \boldsymbol{\upsilon}_i \boldsymbol{\upsilon}_i'$.

Repeating the argument leading from (8) through to (13) provides an immediate generalization of the previous derivation of the theoretical LRF forecasting formula. This in turn leads to the following generalization of Proposition 1.

Proposition 2 *Suppose that the m-lagged vectors satisfy $\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') = \sum_{j=1}^{m} \gamma_i \boldsymbol{\upsilon}_j \boldsymbol{\upsilon}_j'$ and are decomposed as $\mathbf{x}_t = \boldsymbol{\Upsilon}_k \mathbf{z}_t + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{\Upsilon}_k = [\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_k]$ and $\mathbf{z}_t \sim (\mathbf{0}, diag\{\gamma_1, \ldots, \gamma_k\})$. Let $\boldsymbol{\Upsilon}_{m-k}^u$ denote the $(m-1) \times k$ sub-matrix in the first $m-1$ rows of $\boldsymbol{\Upsilon}_{m-k} = [\boldsymbol{\upsilon}_{k+1}, \ldots, \boldsymbol{\upsilon}_m]$ and set $\mathbf{G} = diag(\gamma_{k+1}, \ldots, \gamma_m)$. Let $\varepsilon(t+j|t) = x(t+j) - s(t+j|t)$ where $s(t+j|t)$, $j = 1, 2, \ldots, h$, are generated as in equations (12) and (13). Then the MSFE*

$$\mathbb{E}[\varepsilon(t+j|t)^2] = A_{t+j}(\boldsymbol{\Gamma}_{m+h-1} - \boldsymbol{\Sigma}_{m+h-1})A_{t+j}'$$

where the MSFE coefficients A_{t+j}, $j = 1, \ldots, h$, are as in Proposition 1 and

$$\boldsymbol{\Sigma}_{m+h-1} = \begin{pmatrix} \boldsymbol{\Upsilon}_{m-k}^u \mathbf{G} \boldsymbol{\Upsilon}_{m-k}^{u'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Example: Consider the first order moving average process $x(t) = \varepsilon(t) + \theta\varepsilon(t-1)$ where $\varepsilon(t)$ is a zero mean white noise process with variance σ^2, $\varepsilon(t) \sim WN(0, \sigma^2)$. Then the m-lagged vectors $\mathbf{x}_t = (x(t), x(t+1), \ldots, x(t+m-1))'$ have a tridiagonal Toeplitz covariance matrix

$$\boldsymbol{\Gamma} = \sigma^2 \begin{pmatrix} (1+\theta^2) & \theta & & & \\ \theta & (1+\theta^2) & \theta & & \\ & \ddots & \ddots & \ddots & \\ & & \theta & (1+\theta^2) & \theta \\ & & & \theta & (1+\theta^2) \end{pmatrix}.$$

The eigenvalues are $\gamma_j = \sigma^2(1 + \theta^2 + 2\theta\cos(\omega_j))$ where $\omega_j = \pi j/(m+1)$ $j = 1, \ldots, m$, with corresponding eigenvectors $\boldsymbol{\upsilon}_j = \sqrt{\frac{2}{(m+1)}}(\sin(\omega_j), \ldots, \sin(m\omega_j)'$, $j = 1, \ldots, m$. Employing the SSA signal plus noise decomposition and the associated LRF forecasting formula with any given $k < m$ implies that the $m-k$ eigenvalue-eigenvector pairs that do not correspond to dominant parts of the power spectrum of the process will be relegated to the noise and neglected. □

4 Forecast Consistency

In order to relate the empirical SSA forecasting procedure to its population counterpart we must introduce some basic assumptions concerning the observed process. Rather than specifying primitive regularity conditions we will suppose that $x(t)$ is a zero mean stochastic process that satisfies the following assumption.

Assumption 1 *The data generating mechanism underlying the stochastic process $x(t)$ satisfies sufficient conditions to ensure that for any trajectory matrix window length $m = (\log N)^c$,*

$c < \infty$, there exists a positive definite matrix $\mathbf{\Gamma}$ such that $\|n^{-1}\mathbf{X}\mathbf{X}' - \mathbf{\Gamma}\| = O(Q_n)$ a.s. where $Q_n \to 0$ as $n = N - m + 1 \to \infty$ as $N \to \infty$.

Conditions under which statistical ergodic theorems of the type implicit in Assumption 1 are valid are well documented in the time series literature, see Brockwell & Davis (1991) for example. The generality of Assumption 1 implies our results have broad applicability and processes that satisfy Assumption 1 are examined in the following section. To establish our consistency results we will appeal to the following lemma relating the spectral decomposition of $\mathbf{X}\mathbf{X}'/n$ to the spectral decomposition of $\mathbf{\Gamma}$.

Lemma 2 *Suppose that the stochastic process $x(t)$ satisfies Assumption 1. If the eigenvalue-eigenvector pairs of $\mathbf{X}\mathbf{X}'/n$ are denoted $\{\ell_j/n, \mathbf{u}_j\}$, $j = 1, \ldots, m$, and those of $\mathbf{\Gamma}$ are denoted by $\{\gamma_j, \boldsymbol{\upsilon}_j\}$, $j = 1, \ldots, m$, then $|\ell_j/n - \gamma_j| = O(Q_n)$ and $\|\varsigma_j \mathbf{u}_j - \boldsymbol{\upsilon}_j\| = O(Q_n)$, $j = 1, \ldots, m$, where $\varsigma_j = sign(\boldsymbol{\upsilon}_j' \mathbf{u}_j)$.*

Before proceeding let us note that the $MSFE$ values presented in Propositions 1 and 2 will not be available to the practitioner. They can be estimated from the data, however, using obvious "plug in" estimates. Thus A_{t+j}, $j = 1, \ldots, h$, can be estimated by substituting $\mathbf{a}$ for $\boldsymbol{\alpha}$, $\mathbf{\Gamma}_{m+h-1}$ can be estimated using $\sum_{t=1}^{n'} \boldsymbol{\xi}(t)\boldsymbol{\xi}(t)'/n'$ where $\boldsymbol{\xi}(t) = (x(t), \ldots, x(t+m+h-1))'$ and $n' = N - m - h + 1$, and $\mathbf{\Sigma}_{m+h-1}$ can be estimated by replacing $\mathbf{\Upsilon}^u_{m-k}\mathbf{G}\mathbf{\Upsilon}^{u'}_{m-k}$ with, in an obvious notation, $n^{-1}\mathbf{U}^u_{m-k}\text{diag}(\ell_{k+1}, \ldots, \ell_m)\mathbf{U}^{u'}_{m-k}$. Denote these estimates by $\widehat{A}_{t+j}$, $j = 1, \ldots, h$, $\widehat{\mathbf{\Gamma}}_{m+h-1}$ and $\widehat{\mathbf{\Sigma}}_{m+h-1}$.

Theorem 1 *Suppose that the stochastic process $x(t)$ satisfies Assumption 1. Then the SSA forecasting coefficients satisfy $\|\mathbf{a} - \boldsymbol{\alpha}\| = O(Q_n)$. Moreover, the $MSFE$ estimation errors $\|\widehat{A}_{t+j} - A_{t+j}\|$, $j = 1, \ldots, h$, $\|\widehat{\mathbf{\Gamma}}_{m+h-1} - \mathbf{\Gamma}_{m+h-1}\|$ and $\|\widehat{\mathbf{\Sigma}}_{m+h-1} - \mathbf{\Sigma}_{m+h-1}\|$ are all $O(Q_n)$.*

Theorem 1 shows that when a $SSA(m,k)$ model is fitted to data the values obtained will be consistent for the corresponding forecasting formulae derived from the stochastic process giving rise to the observed realization. In the next section we present illustrative examples that demonstrate the $MSFE$ performance of different $SSA(m,k)$ models and compares them to the optimal MSE forecast for known processes.

5 Theoretical Illustrations

5.1 Forecasting an $AR(1)$ process

Consider a zero mean $AR(1)$ process $x(t) = \phi x(t-1) + \varepsilon(t)$ where $\varepsilon(t) \sim WN(0, \sigma^2)$ and $|\phi| < 1$. The autocovariance generating function of this process is $\gamma(z) = \gamma(0)\sum \phi^i z^i$, where $\gamma(0) = \sigma^2/(1-\phi^2)$, and Assumption 1 is satisfied with Γ equal to the Toeplitz matrix with first row $\gamma(0)(1, \phi, \ldots, \phi^{m-1})$, $\mathbf{\Gamma} = \gamma(0)T\{1, \phi, \ldots, \phi^{m-1}\}$. For an $AR(1)$ process the optimal MSE forecast of $x(t+j)$ given $x(\tau)$, $\tau \le t$, is $x(t+j|t) = \phi^j x(t)$, $j = 1, 2, \ldots, h$, with a $MSFE$ of $MSFE_{AR(1)}(j) = \sigma^2(1-\phi^{2j})/(1-\phi^2)$ for the jth forecast horizon.

5.1.1 The $SSA(2,1)$ model

The simplest possible SSA specification for any observed time series is an $SSA(2,1)$ model. If such a model is applied to an $AR(1)$ process the eigenvalue-eigenvector pairs of $\mathbf{\Gamma} = \gamma(0)T\{1,\phi\}$ are $\{\gamma_1 = \sigma^2(1+|\phi|)/(1-\phi^2), \boldsymbol{\psi}_1 = (1,1)'/\sqrt{2}\}$ and $\{\gamma_2 = \sigma^2(1-|\phi|)/(1-\phi^2), \boldsymbol{\psi}_2 = (1,-1)'/\sqrt{2}\}$. Projecting the second element of $\boldsymbol{\psi}_1$ onto the first gives $\alpha = 1$ for the LRF coefficient. This leads to the $MSFE$ coefficients

$$A_{t+j} = \sum_{i=1}^{j}(\mathbf{0}'_{j-i}, -1, 1, \mathbf{0}'_{h-j+i-1}) = (-1, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j}), j = 1,\ldots,h$$

and the $MSFE$ of the $SSA(2,1)$ model for each j across the forecast horizon is

$$\mathbb{E}[\varepsilon(t+j|t)^2] = (-1, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j})(\mathbf{\Gamma}_{h+1} - \mathbf{\Sigma}_{h+1})(-1, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j})' \tag{14}$$

where $\mathbf{\Gamma}_{h+1} = \gamma(0)T\{1,\phi,\ldots,\phi^h\}$ and $\mathbf{\Sigma}_{h+1} = \operatorname{diag}(\gamma_2/2, \mathbf{0}'_h)$. Evaluating (14) for $j = 1,\ldots,h$ gives the $MSFE$ of a $SSA(2,1)$ model when applied to an $AR(1)$ process as

$$MSFE_{SSA(2,1)}(j) = \frac{2(1-\phi^j)}{1-\phi^2}\sigma^2 - \frac{(1-|\phi|)}{2(1-\phi^2)}\sigma^2 = \frac{4(1-\phi^j)-(1-|\phi|)}{2(1-\phi^2)}\sigma^2.$$

Figure 1 depicts the theoretical $MSFE$ of the $SSA(2,1)$ model and the optimal $AR(1)$ MSE forecast over the horizon $h = 20$ when $\phi = 0.5, 0.7, 0.9$ and σ^2 is set so that $\gamma(0) = 5$. To ascertain the practical relevance of the theoretical formulas we carried out a simulation experiment using an $AR(1)$ data generating process with the same parameter values. We generated $R = 10,000$ time series $\{x^{(r)}(t) : t = 1,\ldots,N+h\}$ with $N = 300$ and $h = 20$, $r = 1,\ldots,R$. For each replication r we fitted an $SSA(2,1)$ model and an $AR(1)$ model to the first N data values and constructed the empirical forecasts $\hat{x}^{(r)}(N+j|N)$, $j = 1,\ldots,h$. For each model we then compute the simulated $MSFE$ by averaging the squared forecast error across the R replications, $\overline{MSFE}(j) = \sum_{r=1}^{R}\left(x^{(r)}(N+j) - \hat{x}^{(r)}(N+j|N)\right)^2/R$, $j = 1,\ldots,h$. Figure 1 also graphs $\overline{MSFE}(j)$, $j = 1,\ldots,h$. Two obvious conclusions can be drawn from the $MSFE$ curves plotted in Figure 1: First, the simulated $MSFE$ closely

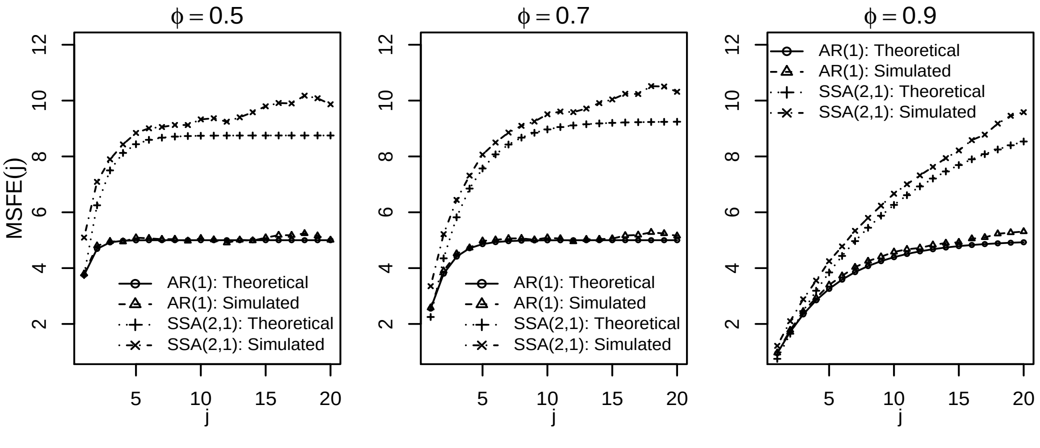

Figure 1: Theoretical $MSFE(j)$ and simulated $\overline{MSFE}(j)$ for $SSA(2,1)$ model and optimal $AR(1)$ MSE forecast, $x(t) = \phi x(t-1) + \varepsilon(t)$, $\varepsilon(t) \sim WN(0,1)$, $\phi = 0.5, 0.7, 0.9$ and $j = 1,\ldots,20$.

resembles the theoretical $MSFE$. Second, the $MSFE$ of the $SSA(2,1)$ model exceeds that of the optimal $AR(1)$ MSE forecast by a considerable margin. A third conclusion seems to be that the $SSA(2,1)$ model exhibits greater experimental sampling variation in $\overline{MSFE}(j)$ than the $AR(1)$ model, particularly when $j > 12$ – a forecast horizon that exceeds six times the window length.

A natural measure of the signal-to-noise ratio (SNR) of an $AR(1)$ process is $\phi^2/(1-\phi^2)$, and for $\phi = 0.5, 0.7, 0.9$, $SNR = 0.33', 0.96, 4.26$. For the $SSA(2,1)$ model SNR is given by $\gamma_1/\gamma_2 = (1+|\phi|)/(1-|\phi|)$. This gives $SNR = 3.0, 5.66', 19$ when $\phi = 0.5, 0.7, 0.9$, respectively. The relative magnitudes of SNR for the $SSA(2,1)$ model and the $AR(1)$ process are not too dissimilar, a feature that is partly reflected in the comparative curvatures seen in the different panels in Figure 1. The increasingly poor performance of the $SSA(2,1)$ model as the forecast horizon increases reflects that the ratio $MSFE_{SSA(2,1)}(j)/MSFE_{AR(1)}(j) \rightarrow 2 - \frac{1}{2}(1-|\phi|)$ as $j \rightarrow \infty$. It is clear that the $SSA(2,1)$ model does not capture the structure of the $AR(1)$ process well.

5.1.2 The $SSA(m,k)$ model

For the general $SSA(m,k)$ model values must be allocated to the window length m and to the signal dimension k in order to decompose the time series and compute the LRF forecasts. Here we select $m \in \{2, M\}$ and $k \in \{1, m-1\}$ such that the absolute difference between the SNR of the $SSA(m,k)$ model and the $AR(1)$ process is a minimum. For an $SSA(m,k)$ model $SNR = \sum_{j=1}^{k} \gamma_j / \sum_{j=k+1}^{m} \gamma_j$ where, for the $AR(1)$ process, γ_j is the jth eigenvalue of $\mathbf{\Gamma} = \gamma(0)T\{1, \phi, \ldots, \phi^{m-1}\}$. The eigenvalues are $\gamma_j = \sigma^2/(1+\phi^2 - 2\phi\cos(\omega_j))$ where ω_j, $j = 1, \ldots, m$, are the positive roots of $\omega = \tan^{-1}\{(1-\phi^2)\sin\omega/(1+\phi^2)\cos\omega - 2\phi)\}/m$, the eigenvectors are non-harmonic sinusoids. The eigenvalue-eigenvector pairs are transcendental functions that can be readily computed once ϕ and σ^2 are known, and the SNR of the $SSA(m,k)$ model evaluated accordingly. For the preassigned value $M = 20$ this leads to $SSA(11,1)$, $SSA(8,1)$ and $SSA(7,1)$ models being selected when $\phi = 0.5, 0.7, 0.9$, respectively. Note in passing that although the value selected for m exceeds two in all three cases, we also have $m < M$ and $k = 1$. We will indicate how to choose m and k in practice in Section 6.

Figure 2 shows the theoretical and simulated $MSFE$ of the optimal MSE $AR(1)$ predictor and the selected $SSA(m,k)$ models when evaluated using the same simulation experiments that gave rise to Figure 1. As in Figure 1, the simulated $MSFE$ closely resembles the theoretical $MSFE$. The $MSFE$ of the $SSA(m,1)$ models still exceeds that of the optimal $AR(1)$ MSE forecast, but the increase in window length has obviously improved performance, the margin between $MSFE_{SSA(m,1)}(j)$ and $MSFE_{AR(1)}(j)$ in Figure 2 is at worst no more than one half of that seen in Figure 1. That the use of an $SSA(2,1)$ model produces inferior results relative to $SSA(m,1)$ models with $m > 2$ is clear, but as will be seen in the following example, increasing window length does not always improve performance.

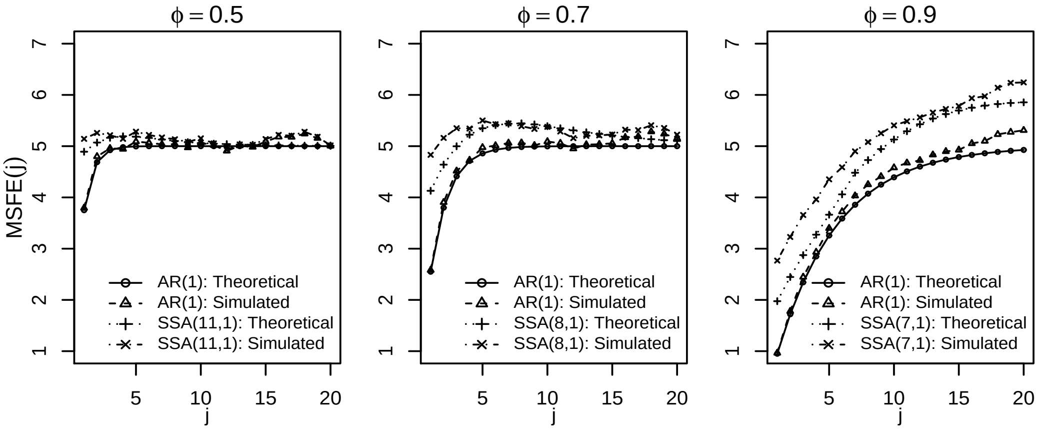

Figure 2: Theoretical $MSFE(j)$ and simulated $\overline{MSFE}(j)$ for $SSA(m,k)$ model and optimal $AR(1)$ MSE forecast, $x(t) = \phi x(t-1) + \varepsilon(t)$, $\varepsilon(t) \sim WN(0,1)$, $\phi = 0.5, 0.7, 0.9$ and $j = 1, \ldots, 20$.

Before we proceed it is of interest to observe that the use of plug in values yields very reasonable estimates of the $MSFE$. Table 1, for example, lists the average value and the variance of the plug in estimate $\widehat{MSFE}(j)$, $j = 1, \ldots, h = 20$, observed across the replications that gave rise to the far right hand panel in Figure 2. A comparison of these values with the theoretical $MSFE$ suggests that appropriately constructed SSA forecast confidence bands obtained using $\widehat{MSFE}(j)$ will possess the correct coverage probability.

Table 1 $MSFE$ for $SSA(7,1)$ model of $AR(1)$ process with $\gamma(0) = 5$, $\phi = 0.9$.

Horizon	Theoretical	Average	Variance
j	$MSFE(j)$	$\widehat{MSFE}(j)$	
1	1.975	1.972	0.044
2	2.447	2.441	0.073
3	2.874	2.865	0.115
4	3.273	3.261	0.171
5	3.662	3.647	0.245
6	4.059	4.041	0.348
7	4.481	4.459	0.491
8	4.729	4.706	0.600
9	4.944	4.919	0.717
10	5.130	5.105	0.840
11	5.291	5.267	0.967
12	5.427	5.405	1.095
13	5.540	5.519	1.218
14	5.626	5.608	1.331
15	5.695	5.681	1.439
16	5.750	5.739	1.541
17	5.791	5.786	1.636
18	5.822	5.821	1.722
19	5.843	5.846	1.799
20	5.855	5.863	1.867

5.2 Forecasting a random walk series

Consider an observed process $x(t)$ such that for $t = 1, \dots$

$$x(t) = \sum_{\tau=0}^{t-1} \varepsilon(t-\tau)$$

where $\varepsilon(t)$ is a white noise processes with unit variance, $\varepsilon(t) \sim WN(0,1)$. Exploiting the strong Markov property of the random walk we can express the m–lagged vector as

$$\mathbf{x}_t = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \sum_{\tau=0}^{t-2} \varepsilon(t-1-\tau) + \begin{bmatrix} \varepsilon(t) \\ \varepsilon(t)+\varepsilon(t+1) \\ \varepsilon(t)+\varepsilon(t+1)+\varepsilon(t+2) \\ \vdots \\ \varepsilon(t+1)+\ldots+\varepsilon(t+m-1) \end{bmatrix}, \tag{15}$$

where the two components on the right hand side of (15) are orthogonal and give the decomposition $\mathbf{x}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t$ directly. Since $\mathbb{E}[\left(\sum_{\tau=1}^{t-1}\varepsilon(t-\tau)\right)^2] = t-1$ and $\mathbb{E}[\sum_{\tau=0}^{r}\varepsilon(t-\tau)\sum_{\tau=0}^{s}\varepsilon(t-\tau)] = \min(r+1, s+1)$, for the $m-$lagged vector $\mathbf{x}_t$ we obtain $\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') = (t-1)\mathbf{1}_m\mathbf{1}_m' + \boldsymbol{\Psi}$ where $\boldsymbol{\Psi} = [\min(r,c)]_{r,c=1,\dots,m}$ and $\mathbf{1}_m' = (1,\dots,1)$, and for the trajectory matrix $\mathbf{X}$ we therefore have

$$\mathbb{E}[n^{-1}\mathbf{X}\mathbf{X}'] = n^{-1}\sum_{t=1}^{n}\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'] = \frac{n-1}{2}\mathbf{1}_m\mathbf{1}_m' + \boldsymbol{\Psi}\,.$$

Applying Donsker's theorem and the fact that $n^{-3/2}\sum_{t=1}^{n} x(t-1)\varepsilon(t) = O(\sqrt{\log\log n})$ (Poskitt 2000, Lemma A.1.(ii)) we can also deduce that $\|n^{-1}\mathbf{X}\mathbf{X}' - \boldsymbol{\Gamma}\| = O(Q_n)$ where $\boldsymbol{\Gamma} = n\beta_n^2\mathbf{1}_m\mathbf{1}_m' + \boldsymbol{\Psi}$ and

$$\beta_n^2 = \frac{1}{n^2}\sum_{t=1}^{n} x(t-1)^2 + O(Q_n) \overset{D}{\to} \int_0^1 \mathbb{B}^2(\omega)d\omega,$$

where $Q_n = \sqrt{\log\log n/n}$ and $\mathbb{B}(\omega)$ denotes standard Brownian motion.

Let $\mathbf{H} = n\beta_n^2\mathbf{1}_m\mathbf{1}_m'$. Then the eigenvalues of $\mathbf{H}$ are $\lambda_1 = mn\beta_n^2$ with eigenvector $\boldsymbol{\varphi}_1 = \mathbf{1}_m/\sqrt{m}$, and $\lambda_m = 0$ with multiplicity $m-1$ and eigenvectors

$$\begin{aligned} \boldsymbol{\varphi}_2 &= (-1, 1, \mathbf{0}_{m-2}')'/\sqrt{2} \\ \boldsymbol{\varphi}_3 &= (1, 1, -2, \mathbf{0}_{m-3}')'/\sqrt{6} \\ \boldsymbol{\varphi}_4 &= (-1, -1, -1, 3, \mathbf{0}_{m-4}')'/\sqrt{12} \\ &\vdots \\ \boldsymbol{\varphi}_m &= (-1)^{m-1}(1, 1, \dots, 1, -(m-1))'/\sqrt{m(m-1)}\,. \end{aligned}$$

Moreover, when $m << n$ the spectral decomposition of $\boldsymbol{\Gamma} = \mathbf{H} + \boldsymbol{\Psi}$ is dominated by that of $\mathbf{H}$ to the point where the effect of the matrix $\boldsymbol{\Psi}$ vanishes as $n \to \infty$. To verify this note that $\|\boldsymbol{\Gamma} - \mathbf{H}\|/mn = \|\boldsymbol{\Psi}\|/mn = O(m/n)$ and the eigenvalues of $(mn)^{-1}\boldsymbol{\Gamma}$ and $(mn)^{-1}\mathbf{H}$

are γ_j/mn and λ_j/mn, $j = 1, \ldots, m$, respectively. By a repetition of the argument that leads from Assumption 1 to Lemma 2 it therefore follows that $|\gamma_1/mn - \beta_n^2| = O(m/n)$ and $|\gamma_j/mn| = O(m/n)$, $j = 2, \ldots, m$, and that $\|\varsigma_j \boldsymbol{v}_j - \boldsymbol{\varphi}_j\| = O(m/n)$ where $\varsigma_j = \boldsymbol{\varphi}_j' \boldsymbol{v}_j$. The spectral decomposition of $\mathbf{\Gamma}$ is thus dominated by its largest eigenvalue, as can be seen by observing that

$$\frac{\gamma_1}{\sum_{j=1}^{m} \gamma_j} = \frac{\gamma_1/mn}{\sum_{j=1}^{m} \gamma_j/mn} = \frac{\beta_n^2 + O(m/n)}{\beta_n^2 + O(m^2/n)} = 1 + O(m^2/n)$$

approaches unity as $n \to \infty$ provided $m^2/n \to 0$ as $n \to \infty$. This clearly shows that the signal component eventually dominates the behaviour of the m-lagged vectors and the contribution of the noise component all but disappears.

Let $\mathbf{\Phi}_k = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k]$ where $1 \leq k \leq m-1$. Then it is straightforward to verify that the linear recurrent coefficient evaluated from the eigenvalue decomposition of $\mathbf{H}$, namely $(1 - \mathbf{\Phi}_k^l \mathbf{\Phi}_k^{l'})^{-1} \mathbf{\Phi}_k^u \mathbf{\Phi}_k^{l'}$, equals $(m-1)^{-1} \mathbf{1}_{m-1}$. From the development in the previous paragraph it follows that $\|(m-1)\boldsymbol{\alpha} - \mathbf{1}_{m-1}\| = O(m^2/n)$, where we recall that for an $SSA(m,k)$ model $\boldsymbol{\alpha} = (1 - \mathbf{\Upsilon}_k^l \mathbf{\Upsilon}_k^{l'})^{-1} \mathbf{\Upsilon}_k^u \mathbf{\Upsilon}_k^{l'}$. The sum of the coefficients, $\sum_{i=1}^{m-1} \alpha_i$, will therefore approach one as the effective sample size grows, that is as $n \to \infty$. This phenomenon is illustrated in Figure 3, which plots $\sum_{i=1}^{m-1} \alpha_i$ against m for an $SSA(m,1)$ model evaluated from the spectral decomposition of $\mathbb{E}[n^{-1}\mathbf{X}\mathbf{X}'] = \frac{1}{2}(n-1)\mathbf{1}_m \mathbf{1}_m' + \mathbf{\Psi}$. Figure 3 indicates that $\sum_{i=1}^{m-1} \alpha_i \geq 1$ and that the proximity of $\sum_{i=1}^{m-1} \alpha_i$ to unity varies directly with the magnitude of m/n. When m/n is not sufficiently small the impact of $\mathbf{\Psi}$ on the spectral decomposition is not negligible and has the effect of increasing the magnitude of $\sum_{i=1}^{m-1} \alpha_i$ beyond that achieved by the dominant eigenvalue. This is seen in $\sum_{i=1}^{m-1} \alpha_i - 1$ increasing as m increases for a given N, and decreasing as the effective sample size $n = N - m + 1$ increases for a given m.

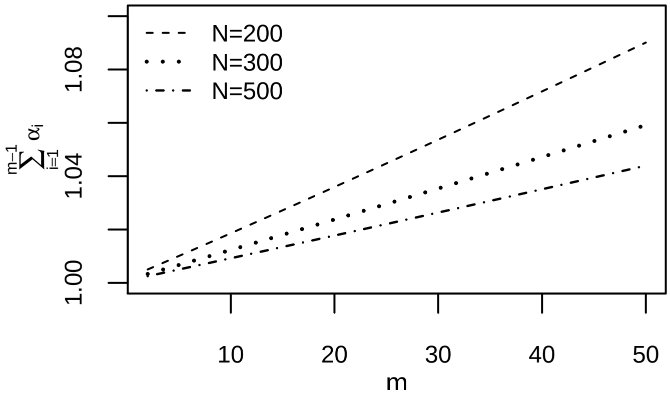

Figure 3: Sum of linear recurrent coefficients for $SSA(m,1)$ model of random walk process.

Remark: Let $\mathbf{z}_t = (x(t-n+1), x(t-n+2), \ldots, x(t))'$ and $\mathbf{Z}_{t-1} = [\mathbf{z}_{t-m+1} : \ldots : \mathbf{z}_{t-1}]$. Then the regression coefficient from regressing $x(t)$ on $x(t-m+1), \ldots, x(t-1)$ for $t = m, \ldots, N$ is obtained by solving the normal equations $\mathbf{Z}_{t-1}'\mathbf{Z}_{t-1}\hat{\boldsymbol{\alpha}} = \mathbf{Z}_{t-1}'\mathbf{z}_t$. For the random walk process it can be shown by following the results of Poskitt (2000, Lemma A.1.) that

$n^{-1}\mathbf{Z}'_{t-1}\mathbf{z}_t = n\beta_n^2\mathbf{1}_{m-1} + O(Q_n)$, and that $n^{-1}\mathbf{Z}'_{t-1}\mathbf{Z}_{t-1} = n\beta_n^2\mathbf{1}_{m-1}\mathbf{1}'_{m-1} + O(Q_n)$. It follows that $\mathbf{1}_{m-1}\mathbf{1}'_{m-1}\hat{\boldsymbol{\alpha}} = \mathbf{1}_{m-1} + O(Q_n)$, that is, $\hat{\boldsymbol{\alpha}} = (m-1)^{-1}\mathbf{1}_{m-1} + O(Q_n)$. Thus, in the case of a random walk process, the linear recurrent coefficient estimated via the spectral decomposition of the trajectory matrix is asymptotically equivalent to a least squares estimate. □

5.2.1 The $SSA(2,1)$ model

To ascertain the practical relevance of the theoretical formulas we have employed a simulation procedure similar to the one used previously and generated data from a random walk process $x(t) = \sum_{\tau=0}^{t-1}\varepsilon(t-\tau)$ where $\varepsilon(\tau) \sim WN(0,1)$. We generated $R = 10,000$ time series and for each replication we fitted an $SSA(2,1)$ model and constructed empirical forecasts which were then used to compute the simulated $MSFE$. To construct the theoretical $MSFE$ of the $SSA(2,1)$ model we have evaluated the spectral decomposition of $\mathbb{E}[n^{-1}\mathbf{X}\mathbf{X}'] = \frac{1}{2}(n-1)\mathbf{1}_2\mathbf{1}_2' + \boldsymbol{\Psi}$. The eigenvalue-eigenvector pairs are $\{\gamma_1 = \frac{1}{2}(n+2+\sqrt{(n+1)^2+1}), \boldsymbol{\psi}_1 = (1, r_n)'/\sqrt{1+r_n^2}\}$ and $\{\gamma_2 = \frac{1}{2}(n+2-\sqrt{(n+1)^2+1}), \boldsymbol{\psi}_2 = (-r_n, 1)'/\sqrt{1+r_n^2}\}$ where $r_n = (1+\sqrt{(n+1)^2+1})/(n+1)$. Projecting the second element of $\boldsymbol{\psi}_1$ onto the first gives $\alpha = r_n$ for the LRF coefficient. This leads to the MSFE coefficients $A_{t+j} = (-r_n^j, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j})$, $j = 1,\ldots,h$. Setting $\boldsymbol{\Gamma}_{h+1} = (n-1)\mathbf{1}_{h+1}\mathbf{1}'_{h+1} + \boldsymbol{\Psi} = \mathbb{E}[\boldsymbol{\xi}_{h+1}\boldsymbol{\xi}'_{h+1}]$, $\boldsymbol{\xi}_{h+1} = [x(n),\ldots,x(n+h)]'$, and $\boldsymbol{\Sigma}_{h+1} = \mathrm{diag}(\gamma_2 r_n^2/(1+r_n^2), \mathbf{0}'_h)$ in Proposition 2 yields

$$\begin{aligned}\mathbb{E}[\varepsilon(N+j|N)^2] &= (-r_n^j, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j})(\boldsymbol{\Gamma}_{h+1} - \boldsymbol{\Sigma}_{h+1})(-r_n^j, \mathbf{0}'_{j-1}, 1, \mathbf{0}'_{h-j})' \\ &= j + n\left((1-r_n^j)^2 - \tfrac{r_n^{2j}}{4}\left\{\tfrac{(n+1)+2r_n}{(n+1)-r_n}\right\}\left\{\tfrac{(n+1)(1-r_n)+2}{n}\right\}\right)\end{aligned} \qquad (16)$$

for $MSFE(j)$ of a $SSA(2,1)$ model when applied to a random walk process. The $MSFE$ for the optimal MSE predictor of the random walk process equals the conditional variance, namely $Var[x(t+j)|\{x(\tau) : \tau \leq t\}] = j$.

Figure 4 plots the theoretical $MSFE$ of the $SSA(2,1)$ model and the $MMSE$ predictor of the random walk process along with their simulated versions $\overline{MSFE}$ across the forecast horizon $j = 1,\ldots,20$ when $N = 200, 300, 500$. It can be seen that the $MSFE$ of the

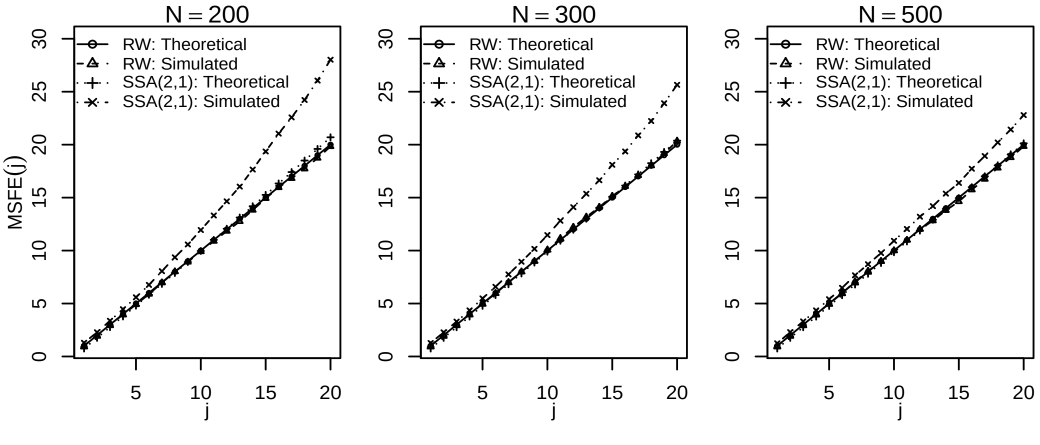

Figure 4: Theoretical and simulated $MSFE$ of $SSA(2,1)$ model and random walk process computed from 10,000 replications of $x(t) = \sum_{\tau=0}^{t-1}\varepsilon(t-\tau)$, $\varepsilon(t) \sim WN(0,1)$.

$SSA(2,1)$ model exceeds that of the optimal MSE predictor by a tiny margin for $j \leq 15$ but the deviation increases slightly in a manor consistent with expression (16) as the forecast horizon increases and $j \to 20$. A comparison of the three panels indicates how the discrepancy between the $MSFE$ of the $SSA(2,1)$ model and the $MMSE$ predictor diminishes as the effective sample size increases and the LRF coefficient $\alpha = r_n \to 1$ as $n \to \infty$, in line with the previous development. The difference between $\overline{MSFE}(j)$ and $MSFE(j)$ for the $SSA(2,1)$ model is much larger than that observed with the optimal MSE predictor, for which this difference is virtually zero. This latter feature reflects that for the $SSA(2,1)$ model $MSFE$ has been calculated on the basis of $\mathbb{E}[n^{-1}\mathbf{X}\mathbf{X}']$ whereas the magnitude of $\overline{MSFE}$ reflects the additional variation present in $n^{-1}\mathbf{X}\mathbf{X}' = n\beta_n^2\mathbf{1}_m\mathbf{1}_m' + \mathbf{\Psi} + O(\sqrt{\log\log n/n})$ that does not subside as $n \to \infty$.

5.2.2 $SSA(m,1)$ Models

Given that the spectral decomposition associated with a random walk process is dominated by the first eigenvalue we consider here the performance of different $SSA(m,1)$ models. Figure 5 plots the outcomes resulting from the application of such models to a random walk process when $N = 300$ using window lengths $m = 4, 6, 10$. Figure 5 shows the theoretical and simulated $MSFE$ of the optimal MSE predictor and the selected $SSA(m,1)$ models when evaluated using the same experimental data that gave rise to the centre panel in Figure 4. The panels in the Figure 5 demonstrate that the $MSFE$ of the $SSA(m,1)$ models increases with increasing window length, and use of an $SSA(m,1)$ model with $m > 2$ does not improve on the performance of the simple $SSA(2,1)$ model.

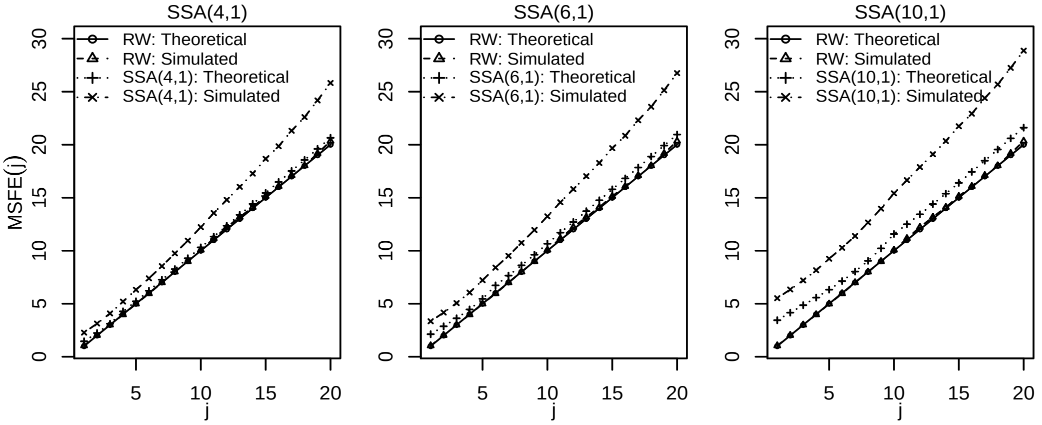

Figure 5: Theoretical and simulated $MSFE$ of $SSA(m,1)$ models, $m = 4, 6, 10$, and random walk process computed from 10,000 replications of $x(t) = \sum_{\tau=0}^{t-1} \varepsilon(t-\tau)$, $\varepsilon(t) \sim WN(0,1)$.

6 Empirical Applications

When examining real world data sets the predictive performance of $SSA(m,k)$ models can only be evaluated by comparing it to other competing models – in practice the true data generating mechanism is unknown and the optimal MSE predictor is not available for analysis as was the case with the theoretical processes examined in the previous section. We have therefore selected three different time series that have been examined elsewhere in the literature – (i) Airline passenger data (Box & Jenkins 1976); (ii) Nile river data (Hipel & McLeod 1994); (iii) USA accidental death data (Brockwell & Davis 2002) – and used models previously fitted to these data sets as benchmarks.

The airline passenger data records monthly totals of international airline passengers in the USA from January 1949 to December 1960. The period January 1949 to December 1958 (the sample data) was used for modeling and the period January 1958 to December 1960 (the test data) was reserved for checking forecast accuracy. Since this series exhibits a strong seasonal pattern and marked heteroscedasticity $ARIMA(p,d,q)(P,D,Q)_{12}$ models were fitted to both the original sample data and it's logarithmic transformation. Employing the automatic selection algorithm of Hyndman & Khandakar (2008) leads to an $ARIMA(0,1,1)(0,1,1)_{12}$ model being selected via AIC in both cases (*cf.* Box & Jenkins 1976, Chatfield 2004). In similar vein, $SSA(m,k)$ models were selected automatically using the minimum description length criterion of Khan & Poskitt (2010). The latter criterion determines the window length m and the signal dimension k simultaneously and for the airline passenger sample data gives a $SSA(12,11)$ model for both the original and the log–transformed series.

For both the $ARIMA(0,1,1)(0,1,1)_{12}$ model and the $SSA(12,11)$ model forecasts of the out of sample test data were constructed after the models had been fitted using the sample data. The performance of the two models was then compared by calculating the empirical root mean squared forecast error

$$RMSFE_h = \left\{ \frac{1}{h}\sum_{j=1}^{h}(x(N+j)-\hat{x}(N+j|N))^2 \right\}^{\frac{1}{2}} ,$$

wherein N is the sample size, $\hat{x}(N+j|N)$, $j=1,\ldots,h$, denote the forecasts, and h is the total forecast horizon. The $RMSFE_h$ figures derived from forecasts of airline passenger numbers for the two years January 1958 to December 1960 are presented in Table 2, where $N=120$ and $h=6,12,18$ and 24. In all cases considered the $SSA(12,11)$ model improves on the $RMSFE$ performance of the $ARIMA(0,1,1)(0,1,1)_{12}$ model by at least 18% and as much as 56.4%. It is noteworthy that whereas for the $ARIMA(0,1,1)(0,1,1)_{12}$ model the $RMSFE$ obtained when the modeling is conducted in terms of the log–transformed data is about one half of that obtained when the modeling is conducted in terms of the original data, the $RMSFE$ of the $SSA(12,11)$ model does not change anywhere near as dramatically. This

Table 2 $RMSFE_h$ of $SSA(12,11)$ and $ARIMA(0,1,1)(0,1,1)_{12}$ models for airline data.

Airline data	Model	$RMSFE_h$			
		$h=6$	$h=12$	$h=18$	$h=24$
Original	$SSA(12,11)$	19.8197	22.2352	26.7928	33.1892
	$ARIMA(0,1,1)(0,1,1)_{12}$	37.4336	46.2710	61.4668	71.9598
Log-transformed	$SSA(12,11)$	17.0925	20.8780	23.4169	27.2589
	$ARIMA(0,1,1)(0,1,1)_{12}$	20.8866	28.1668	29.3763	34.2055

latter feature reflects that SSA automatically adapts to nonlinear and localized features of the series far more readily than do conventional seasonal $ARIMA$ models.

The Nile river data comprises mean monthly river flows (m^3/s) from January 1870 to December 1934. The values from January 1870 to December 1932 were employed as sample data and the period from January 1933 to December 1934 as test data. Montanari et al. (2000) have suggested that the short-memory $ARMA(1, 0, 1)(1, 0, 1)_{12}$ model fits the series well (Montanari et al. 2000, eq.25–26). This model gave a value of $AIC = 8209.08$ when fitted to the sample data, but the minimization of AIC leads to an $ARIMA(1, 0, 1)(1, 1, 1)_{12}$ model being selected for the sample data with $AIC = 8043.97$. This latter model was therefore used to compute forecasts and $RMSFE_h$ values, computed as described above, are presented in Table 3. This table also presents the $RMSFE_h$ values derived from the $SSA(18, 17)$ model determined using the Khan & Poskitt (2010) criterion. The entries in Table 3 show that $SSA(m, k)$ models do not uniformly dominate conventional seasonal $ARIMA$ models.

Table 3 $RMSFE_h$ of $SSA(18, 17)$ and $ARIMA(1, 0, 1)(1, 1, 1)_{12}$ models for Nile data.

Model	$RMSFE_h$			
	$h = 6$	$h = 12$	$h = 18$	$h = 24$
$SSA(18, 17)$	66.1722	49.3324	51.7531	47.6137
$ARIMA(1, 0, 1)(1, 1, 1)_{12}$	68.8752	52.3328	48.4641	47.3151

The USA accident data gives the number of accidental deaths per month from January 1973 to June 1979. The numbers from January 1973 to December 1978 were used as sample data and those from January 1979 to June 1979 were used as test data. Brockwell & Davis (2002) have analyzed this data set and computed forecasts from three models, an $ARIMA(0, 1, 1)(0, 1, 1)_{12}$, a coefficient constrained subset $ARIMA(0, 1, 13)(0, 1, 0)_{12}$ (Brockwell & Davis 2002, eq.6.5.8 and eq.6.5.9, p.208), and a Holt-Winters seasonal model (Brockwell & Davis 2002, §9.3). These models were used to forecast the number of accidental deaths from January 1979 to June 1979 and the $RMSFE$ values where $N = 72$ and $h = 6$ are presented in Table 4 (*cf.* Brockwell & Davis 2002, Table 6.1, p.210 and Table 9.3, p.327). The automatic model selection algorithm of Hyndman & Khandakar (2008) yields an $ARIMA(2, 0, 0)(2, 1, 0)_{12}$ model for this sample data when used in conjunction with AIC, and the minimum description length criterion of Khan & Poskitt (2010) gives an $SSA(24, 13)$ model. The relative $RMSFE_h$ values in Table 4 show once again that an $SSA(m, k)$ model can improve on the best performing of conventional benchmark models by anything from 10.5% to more than 55%.

Table 4 $RMSFE_h$ of different models for USA accidental death data.

Model	$RMSFE_h$	$Relative - RMSFE_h$
$ARIMA(0, 1, 1)(0, 1, 1)_{12}$	582.6261	2.2705
Subset $ARIMA(0, 1, 13)(0, 1, 0)_{12}$	500.5004	1.9504
Holt-Winters	401.2626	1.5637
$ARIMA(2, 0, 0)(2, 1, 0)_{12}$	286.9519	1.1182
$SSA(24, 13)$	256.6120	1

7 Concluding remarks

The theoretical examination of SSA forecasting presented above indicates that for some processes (a simple autoregression) different $SSA(m, k)$ models will not achieve the same $MSFE$ performance as the $MMSE$ predictor for any combination of window length and signal dimension, whereas for other processes (a random walk) the simplest $SSA(2, 1)$ model closely approximates the forecasting performance of the optimal predictor, to which it will converge as the effective sample size increases. These theoretical results are clearly reflected in behaviour observed in simulation experiments. When applied to different real world time series, however, SSA can exhibit considerable improvements in empirical $MSFE$ performance over conventional benchmark models that have been previously used to characterize the series.

The contrast between the relative performance of SSA when it is compared to the $MMSE$ predictor as apposed to it's superior empirical performance when compared to benchmark models might be viewed as a paradox. It is however part of folklore that model specification is of paramount importance in forecasting and that the use of a class of flexible but parsimonious models can be critical in determining performance. This suggests that the solution to the apparent paradox lies in thinking of SSA as a nonparametric modeling methodology that produces accurate approximations of minimal dimension, and that further insight may be achieved by examining SSA from this perspective. It is hoped to investigate this issue further elsewhere.

References

Basilevsky, A. & 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.

Box, G. E. P. & Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

Brockwell, P. & Davis, R. A. (1991), *Time Series: Theory and Methods*, 2nd Ed, New York: Springer-Verlag.

Brockwell, P. J. & Davis, R. A. (2002), *Introduction to Time Series and Forecasting*, 2nd Ed, New York: Springer.

Broomhead, D. & King, G. (1986), 'Extracting qualitative dynamics from experimental data', *Physica D: Nonlinear Phenomena* **20**(2-3), 217–236.

Chatfield, C. (2004), *The Analysis of Time Series: An Introduction*, 6th Ed, Boca Raton: CRC Press.

Elsner, J. B. & Tsonis, A. A. (1996), *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, New York: Plenum Press.

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

Hassani, H., Heravi, S. & Zhigljavsky, A. (2009), 'Forecasting European industrial production with singular spectrum analysis', *International Journal of Forecasting* **25**, 103–118.

Hassani, H., Soofi, A. & Zhigljavsky, A. (2010), 'Predicting daily exchange rate with singular spectrum analysis', *Nonlinear Analysis: Real World Applications* **11**(3), 2023–2034.

Hassani, H. & Zhigljavsky, A. (2009), 'Singular spectrum analysis: Methodology and application to economics data', *Journal of Systems Science and Complexity* **22**(3), 372–394.

Hipel, K. W. & McLeod, A. I. (1994), *Time Series Modelling of Water Resources and Environmental Systems*, Vol. 45, Amsterdam: Elsevier.

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: The forecast package for R', *Journal of Statistical Software* **27**(3), 2023–2034.

Khan, M. A. R. & Poskitt, D. S. (2010), 'Description length based signal detection in singular spectrum analysis', *Monash Econometrics and Business Statistics Working Papers* **13/10**.

Khan, M. A. R. & Poskitt, D. S. (2013), 'A note on window length selection in singular spectrum analysis', *Australian & New Zealand Journal of Statistics* **55**(2), 87–108.

Montanari, A., Rosso, R. & Taqqu, M. (2000), 'A seasonal fractional ARIMA model applied to the Nile river monthly flows at Aswan', *Water Resources Research* **36**(5), 1249–1259.

Poskitt, D. (2000), 'Strongly consistent determination of cointegrating rank via canonical correlations', *Journal of Business & Economic Statistics* **18**(1), 77–90.

Prony, G. S. (1795), 'Essai experimental et analytique: Sur les lois de la dilatabilitie de fluides elasttique et sur celles de la force expansive de la vapeur de l'alkool, a differentes temperatures', *Journal de l'Ecole Polytechnique* **1**, 24–76.

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

Thomakos, D., Wang, T. & Wille, L. (2002), 'Modeling daily realized futures volatility with singular spectrum analysis', *Physica A: Statistical Mechanics and its Applications* **312**(3-4), 505–519.

Vautard, R. & 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. & Ghil, M. (1992), 'Singular-spectrum analysis: A toolkit for short, noisy chaotic signals', *Physica D: Nonlinear Phenomena* **58**(1-4), 95–126.

Watanabe, S. (1965), Karhunen-Loéve expansion and factor analysis: Theoretical remarks and applications, *in* ' *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*', Prague: Czechoslovak Academy of Sciences, pp. 635–660.

APPENDIX **Proofs**

Proof of Lemma 1: Partition the matrix $\mathbf{U}_k$ such that $\mathbf{U}_k = \left(\mathbf{U}_k^{\prime u}\ \mathbf{U}_k^{\prime l}\right)'$ where $\mathbf{U}_k^l$ is the row vector of elements in last row of $\mathbf{U}_k$ and $\mathbf{U}_k^u$ is the $(m-1)\times k$ matrix containing the first $m-1$ rows. Partition $\mathbf{S}_k = \mathbf{U}_k\mathbf{U}_k'\mathbf{X}$ and $\mathbf{X}$ conformable with the partition of $\mathbf{U}_k$. Then

$$\begin{pmatrix}\mathbf{S}_k^u\\ \mathbf{S}_k^l\end{pmatrix} = \begin{pmatrix}\mathbf{U}_k^u\\ \mathbf{U}_k^l\end{pmatrix}\begin{pmatrix}\mathbf{U}_k^{u\prime} & \mathbf{U}_k^{l\prime}\end{pmatrix}\begin{pmatrix}\mathbf{X}^u\\ \mathbf{X}^l\end{pmatrix} = \begin{pmatrix}\mathbf{U}_k^u\mathbf{U}_k^{u\prime}\mathbf{X}^u + \mathbf{U}_k^u\mathbf{U}_k^{l\prime}\mathbf{X}^l\\ \mathbf{U}_k^l\mathbf{U}_k^{u\prime}\mathbf{X}^u + \mathbf{U}_k^l\mathbf{U}_k^{l\prime}\mathbf{X}^l\end{pmatrix}$$

and the projection of the last row of $\mathbf{S}_k$ onto its first $m-1$ rows is given by

$$\begin{aligned}\mathbf{S}_k^l &= \mathbf{U}_k^l\mathbf{U}_k^{u\prime}\mathbf{X}^u + \mathbf{U}_k^l\mathbf{U}_k^{l\prime}\mathbf{X}^l\\ &= \mathbf{U}_k^l\mathbf{U}_k^{u\prime}(\mathbf{S}_k^u + \mathbf{E}_k^u) + \mathbf{U}_k^l\mathbf{U}_k^{l\prime}(\mathbf{S}_k^l + \mathbf{E}_k^l)\\ &= (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}[\mathbf{U}_k^l\mathbf{U}_k^{u\prime}(\mathbf{S}_k^u + \mathbf{E}_k^u) + \mathbf{U}_k^l\mathbf{U}_k^{l\prime}\mathbf{E}_k^l]\\ &= (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}\mathbf{U}_k^l\mathbf{U}_k^{u\prime}\mathbf{S}_k^u + (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}\mathbf{U}_k^l[\mathbf{U}_k^{u\prime}\mathbf{E}_k^u + \mathbf{U}_k^{l\prime}\mathbf{E}_k^l]\\ &= (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}\mathbf{U}_k^l\mathbf{U}_k^{u\prime}\mathbf{S}_k^u + (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}\mathbf{U}_k^l\mathbf{U}_k'\mathbf{E}_k\,.\end{aligned}$$

Let $\mathbf{U}_{m-k} = [\mathbf{u}_{k+1},\ldots,\mathbf{u}_m]$ denote the matrix containing the last $m-k$ eigenvectors of $\mathbf{X}\mathbf{X}'$. Then $\mathbf{E}_k = \mathbf{U}_{m-k}\mathbf{U}_{m-k}'\mathbf{X}$, and since $\mathbf{U}_k$ and $\mathbf{U}_{m-k}$ are orthogonal it follows that $\mathbf{S}_k^l = \mathbf{a}'\mathbf{S}_k^u$ where the vector $\mathbf{a}' = (a_1,\ldots,a_{m-1}) = (1-\mathbf{U}_k^l\mathbf{U}_k^{l\prime})^{-1}\mathbf{U}_k^l\mathbf{U}_k^{u\prime}$. ∎

Proof of Proposition 1: First decompose the forecast error as

$$\begin{aligned}\varepsilon(t+j|t) &= x(t+j) - s(t+j|t)\\ &= [x(t+j) - x(t+j|t)] + [x(t+j|t) - s(t+j|t)],\end{aligned}\qquad\text{(A.1)}$$

where for $j=1,\ldots,h$, $x(t+j|t) = \sum_{i=1}^{m-1}\alpha_i x(t+j-m+i)$. The first part of $\varepsilon(t+j|t)$ is

$$\begin{aligned}x(t+1) - x(t+1|t) &= x(t+1) - \sum_{i=1}^{m-1}\alpha_i x(t+1-m+i)\\ &= (-\boldsymbol{\alpha}', 1, \mathbf{0}_{h-1}')\boldsymbol{\xi}_{m+h-1},\\ x(t+2) - x(t+2|t) &= x(t+2) - \sum_{i=1}^{m-1}\alpha_i x(t+2-m+i)\\ &= (0, -\boldsymbol{\alpha}', 1, \mathbf{0}_{h-2}')\boldsymbol{\xi}_{m+h-1},\\ &\ \ \vdots\\ x(t+j) - x(t+j|t) &= x(t+j) - \sum_{i=1}^{m-1}\alpha_i x(t+j-m+i)\\ &= (\mathbf{0}_{j-1}', -\boldsymbol{\alpha}', 1, \mathbf{0}_{h-j}')\boldsymbol{\xi}_{m+h-1}.\end{aligned}$$

The second part of (A.1) can be expressed as

$$\begin{aligned} x(t+1|t) - s(t+1|t) &= \sum_{i=1}^{m-1} \alpha_{m-i}[x(t+1-m+i) - s(t+1-m+i)] \\ &= \sum_{i=1}^{m-1} \alpha_{m-i}\varepsilon(t+1-i) = (\boldsymbol{\alpha}', \mathbf{0}'_h)\boldsymbol{\eta}_{m+h-1}, \end{aligned}$$

and for $j = 2, \dots, m-1$,

$$\begin{aligned} x(t+j|t) - s(t+j|t) &= \textstyle\sum_{i=1}^{m-1} \alpha_i x(t+j-m+i) - \sum_{i=1}^{j-1} \alpha_{m-i} s(t+j-i|t) \\ &\qquad - \textstyle\sum_{i=1}^{m-j} \alpha_i s(t+j-m+i) \\ &= \textstyle\sum_{i=1}^{j-1} \alpha_{m-i}\varepsilon(t+j-i|t) + \sum_{i=1}^{m-j} \alpha_i \varepsilon(t+j-m+i) \\ &= \textstyle\sum_{i=1}^{j-1} \alpha_{m-i}\varepsilon(t+j-i|t) + (\mathbf{0}'_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1})\boldsymbol{\eta}_{m+h-1} \,. \end{aligned}$$

For $j > m-1$,

$$\begin{aligned} x(t+j|t) - s(t+j|t) &= \sum_{i=1}^{m-1} \alpha_i x(t+j-m+i) - \sum_{i=1}^{m-1} \alpha_{m-i} s(t+j-i|t) \\ &= \sum_{i=1}^{m-1} \alpha_{m-i}\varepsilon(t+j-i|t)\,, \end{aligned}$$

which for notational consistency and convenience we can reexpress as

$$x(t+j|t) - s(t+j|t) = \sum_{i=1}^{m-1} \alpha_{m-i}\varepsilon(t+j-i|t) + (\mathbf{0}'_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1})\boldsymbol{\eta}_{m+h-1}$$

since for $j > m-1$ we have $(\mathbf{0}'_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1})\boldsymbol{\eta}_{m+h-1} = 0$.

Collecting these terms together gives us

$$\begin{aligned} \varepsilon(t+1|t) &= (-\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-1})\boldsymbol{\xi}_{m+h-1} + (\boldsymbol{\alpha}', \mathbf{0}'_h)\boldsymbol{\eta}_{m+h-1} \\ &= A_{t+1}\boldsymbol{\xi}_{m+h-1} + B_{t+1}\boldsymbol{\eta}_{m+h-1}\,, \end{aligned}$$

where $A_{t+1} = (-\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-1})$ and $B_{t+1} = (\boldsymbol{\alpha}', \mathbf{0}'_h)$. For $2 \leq j \leq m-1$,

$$\begin{aligned}
\varepsilon(t+j|t) &= (\mathbf{0}'_{j-1}, -\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-j})\boldsymbol{\xi}_{m+h-1} + (\mathbf{0}_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1})\boldsymbol{\eta}_{m+h-1} \\
&\quad + \sum_{i=1}^{j-1} \alpha_{m-i}\varepsilon(t+j-i|t) \\
&= (\mathbf{0}'_{j-1}, -\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-j})\boldsymbol{\xi}_{m+h-1} + (\mathbf{0}'_{j-1}, \mathbf{a}', \mathbf{0}'_{h-j+1})\boldsymbol{\eta}_{m+h-1} \\
&\quad + \sum_{i=1}^{j-1} \alpha_{m-i}[A_{t+j-i}\boldsymbol{\xi}_{m+h-1} + B_{t+j-i}\boldsymbol{\eta}_{m+h-1}] \\
&= [(\mathbf{0}'_{j-1}, -\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-j}) + \sum_{i=1}^{j-1} \alpha_{m-i}A_{t+j-i}]\boldsymbol{\xi}_{m+h-1} \\
&\qquad + [(\mathbf{0}'_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1}) + \sum_{i=1}^{j-1} \alpha_{m-i}B_{t+j-i}]\boldsymbol{\eta}_{m+h-1} \\
&= A_{t+j}\boldsymbol{\xi}_{m+h-1} + B_{t+j}\boldsymbol{\eta}_{m+h-1}\,,
\end{aligned}$$

and similarly for $j > m-1$,

$$\begin{aligned}
\varepsilon(t+j|t) &= [(\mathbf{0}'_{j-1}, -\boldsymbol{\alpha}', 1, \mathbf{0}'_{h-j}) + \sum_{i=1}^{m-1} \alpha_{m-i}A_{t+j-i}]\boldsymbol{\xi}_{m+h-1} \\
&\qquad + [(\mathbf{0}'_{j-1}, \boldsymbol{\alpha}', \mathbf{0}'_{h-j+1}) + \sum_{i=1}^{m-1} \alpha_{m-i}B_{t+j-i}]\boldsymbol{\eta}_{m+h-1} \\
&= A_{t+j}\boldsymbol{\xi}_{m+h-1} + B_{t+j}\boldsymbol{\eta}_{m+h-1}\,.
\end{aligned}$$

Thus for $j = 1, \ldots, h$ the forecast error can be expressed as

$$\varepsilon(t+j|t) = x(t+j) - s(t+j|t) = A_{t+j}\boldsymbol{\xi}_{m+h-1} + B_{t+j}\boldsymbol{\eta}_{m+h-1},$$

where A_{t+j}, $j = 1, \ldots, h$, are generated recursively as specified in the proposition, and B_{t+j}, $j = 2, \ldots, h$, satisfy the same recursions but start at $B_{t+1} = (\boldsymbol{\alpha}', \mathbf{0}'_h)$.

The mean squared forecast error is therefore equal to

$$\begin{aligned}
\mathbb{E}[\varepsilon(t+j|t)^2] &= A_{t+j}\mathbb{E}[\boldsymbol{\xi}_{m+h-1}\boldsymbol{\xi}'_{m+h-1}]A'_{t+j} + B_{t+j}\mathbb{E}[\boldsymbol{\eta}_{m+h-1}\boldsymbol{\eta}'_{m+h-1}]B'_{t+j} \\
&\qquad + 2A_{t+j}\mathbb{E}[\boldsymbol{\xi}_{m+h-1}\boldsymbol{\eta}'_{m+h-1}]B'_{t+j} \\
&= A_{t+j}\boldsymbol{\Gamma}_{m+h-1}A'_{t+j} + B_{t+j}\boldsymbol{\Sigma}_{m+h-1}B'_{t+j} + 2A_{t+j}\boldsymbol{\Sigma}_{m+h-1}B'_{t+j} \\
&= A_{t+j}(\boldsymbol{\Gamma}_{m+h-1} - \boldsymbol{\Sigma}_{m+h-1})A'_{t+j} + (A_{t+j} + B_{t+j})\boldsymbol{\Sigma}_{m+h-1}(A_{t+j} + B_{t+j})'
\end{aligned}$$

where the penultimate line follows because $\mathbb{E}[\boldsymbol{\xi}_{m+h-1}\boldsymbol{\eta}'_{m+h-1}] = \mathbb{E}[\boldsymbol{\eta}_{m+h-1}\boldsymbol{\eta}'_{m+h-1}]$ since $x(t) = s(t) + \varepsilon(t)$ where $s(t)$ is orthogonal to $\varepsilon(t)$. The required result now follows because for all $j = 1, \ldots, h$ the first $m-1$ elements of $A_{t+j} + B_{t+j}$ are zero and $\boldsymbol{\Sigma}_{m+h-1} = \text{diag}(\sigma^2\mathbf{1}'_{m-1}, \mathbf{0}'_h)$, and so $(A_{t+j} + B_{t+j})\boldsymbol{\Sigma}_{m+h-1}(A_{t+j} + B_{t+j})' = 0$, $j = 1, \ldots, h$. ∎

Proof of Proposition 2: This proposition differs from Proposition 1 only in the specification of $\boldsymbol{\Sigma}_{m+h-1} = \mathbb{E}[\boldsymbol{\eta}_{m+h-1}\boldsymbol{\eta}'_{m+h-1}]$. The proof is otherwise identical to that of Proposition 1 and is therefore omitted. ∎

Proof of Lemma 2: By Assumption 1 $\|\mathbf{XX}'/n-\boldsymbol{\Gamma}\| = O(Q_n)$ and the Hoffman-Wielandt Theorem states that $\sum_{j=1}^{m}(\ell_j/n-\gamma_j)^2 \leq \|\mathbf{XX}'/n-\boldsymbol{\Gamma}\|^2$, implying that $|\ell_j/n-\gamma_j| = O(Q_n)$ for $j = 1,\dots,m$. (See also Khan & Poskitt 2013, Lemma 2).

Since the eigenvectors are orthonormal and span $\mathbb{R}^m$ we may set $\mathbf{u}_k = \sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j$ where the coefficients $c_j = \boldsymbol{\upsilon}_j'\mathbf{u}_k$ are such that $|c_j| \leq 1$ and $\sum_{j=1}^{m} c_j^2 = 1$. It follows that

$$(\mathbf{XX}'/n-\boldsymbol{\Gamma}+\boldsymbol{\Gamma})\sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j = (\ell_k/n-\gamma_k+\gamma_k)\sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j\,,$$

which, because $\|\mathbf{XX}'/n-\boldsymbol{\Gamma}\| = O(Q_n)$ and $|\ell_k/n-\gamma_k| = O(Q_n)$, can be re-expressed as

$$\boldsymbol{\Gamma}\sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j = \gamma_k\sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j + O(Q_n)\,,$$

implying that $\sum_{j=1}^{m} c_j^2(\gamma_j-\gamma_k)^2 = O(Q_n^2)$. Thus we can conclude that $c_j = O(Q_n)$ whenever $\gamma_j \neq \gamma_k$ and hence that $|c_k| = 1+O(Q_n)$. Multiplying $\mathbf{u}_k$ by $\operatorname{sgn}(c_k)$ we obtain

$$\operatorname{sgn}(c_k)\mathbf{u}_k = \operatorname{sgn}(c_k)\sum_{j=1}^{m} c_j\boldsymbol{\upsilon}_j = |c_k|\boldsymbol{\upsilon}_k + \operatorname{sgn}(c_k)\sum_{\substack{j=1\\j\neq k}}^{m} c_j\boldsymbol{\upsilon}_j\,,$$

and subtracting $\boldsymbol{\upsilon}_k$ from either side and substituting $c_j = O(Q_n)$ $j = 1,\dots,m$, $j \neq k$, and $|c_k| = 1+Q(Q_n)$ into the resulting equation we have

$$\operatorname{sgn}(c_k)\mathbf{u}_k - \boldsymbol{\upsilon}_k = (|c_k|-1)\boldsymbol{\upsilon}_k + \operatorname{sgn}(c_k)\sum_{\substack{j=1\\j\neq k}}^{m} c_j\boldsymbol{\upsilon}_j = O(Q_n)\,.$$

Thus for the kth eigenvector we find that $\|\varsigma_k\mathbf{u}_k - \boldsymbol{\upsilon}_k\| = O(Q_n)$ where $\varsigma_k = \operatorname{sgn}(c_k)$, that is, the orthonormal eigenvectors of $\mathbf{XX}'/n$ converge to the orthonormal eigenvectors of $\boldsymbol{\Gamma}$ modulo a change in sign. ∎

Proof of Theorem 1: From the equality

$$\mathbf{U}_k\overline{\mathbf{P}}_k\overline{\mathbf{P}}_k'\mathbf{U}_k' = \mathbf{U}_k\mathbf{U}_k' = \begin{pmatrix} \mathbf{U}_k^u\mathbf{U}_k^{u'} & \mathbf{U}_k^u\mathbf{U}_k^{l'} \\ \mathbf{U}_k^l\mathbf{U}_k^{u'} & \mathbf{U}_k^l\mathbf{U}_k^{l'} \end{pmatrix}$$

where $\overline{\mathbf{P}}_k = \operatorname{diag}(\pm1,\dots,\pm1)$, it is clear that the coefficient $\mathbf{a}' = (1-\mathbf{U}_k^l\mathbf{U}_k^{l'})^{-1}\mathbf{U}_k^l\mathbf{U}_k^{u'}$ is invariant to changes in sign in the eigenvectors $\mathbf{u}_1,\dots,\mathbf{u}_k$. The coefficient $\boldsymbol{\alpha}'$ equals $(1-\boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{l'})^{-1}\boldsymbol{\Upsilon}_k^l\boldsymbol{\Upsilon}_k^{u'}$ and from Lemma 2 $\|\mathbf{U}_k\mathbf{P}_k-\boldsymbol{\Upsilon}_k\| = O(Q_n)$, where $\mathbf{P}_k = \operatorname{diag}(\varsigma_1,\dots,\varsigma_k)$, and hence we can conclude that $\|\mathbf{a}-\boldsymbol{\alpha}\| = O(Q_n)$.

That $\|\widehat{A}_{t+j} - A_{t+j}\|$, $j = 1, \ldots, h$, have the same order of magnitude as $\|\mathbf{a} - \boldsymbol{\alpha}\|$ is immediate, and $\|\widehat{\boldsymbol{\Gamma}}_{m+h-1} - \boldsymbol{\Gamma}_{m+h-1}\| = O(Q_n)$ follows directly from Assumption 1. That $\|\widehat{\boldsymbol{\Sigma}}_{m+h-1} - \boldsymbol{\Sigma}_{m+h-1}\| = O(Q_n)$ follows by noting that $\mathbf{U}^u_{m-k}\mathrm{diag}(\ell_{k+1}, \ldots, \ell_m)\mathbf{U}^{u'}_{m-k}$ is invariant to changes in sign in the eigenvectors $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_m$. Now, by Lemma 2 $\|\mathbf{U}_{m-k}\mathbf{P}_{m-k} - \boldsymbol{\Upsilon}_{m-k}\| = O(Q_n)$, where $\mathbf{P}_{m-k} = \mathrm{diag}(\varsigma_{k+1}, \ldots, \varsigma_m)$, and $|\ell_j/n - \gamma_j| = O(Q_n)$, $j = k+1, \ldots, m$, implying that

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}_{m+h-1} - \boldsymbol{\Sigma}_{m+h-1}\| &= \|\boldsymbol{\Upsilon}^u_{m-k}\mathbf{G}\boldsymbol{\Upsilon}^{u'}_{m-k} - \mathbf{U}^u_{m-k}\mathrm{diag}(\ell_{k+1}/n, \ldots, \ell_m/n)\mathbf{U}^{u'}_{m-k}\| \\ &= O(Q_n), \end{aligned}$$

thus completing the proof. ∎