# Difference-based Methods for Truncating the Singular Value Decomposition

Dafydd Evans and Jonathan Gillard

Cardiff School of Mathematics

Cardiff University

{EvansD8, GillardJW}@Cardiff.ac.uk

December 3, 2013

**Abstract**

Given a noisy time series (or signal), one may wish to remove the noise from the observed series. Assuming that the noise-free series lies in some low dimensional subspace of rank $r$, a common approach is to embed the noisy time series into a Hankel trajectory matrix. The singular value decomposition is then used to deconstruct the Hankel matrix into a sum of rank-one components. We wish to demonstrate that there may be some potential in using difference-based methods of the observed series in order to provide guidance regarding the separation of the noise from the signal, and to estimate the rank of the low dimensional subspace in which the true signal is assumed to lie.

## 1 Introduction

### 1.1 Problem motivation and aims of paper

Let $\mathbf{y} = (y_1, \ldots, y_N)$ denote an observed finite realisation of a stochastic process. We assume that $\mathbf{y}$ has been measured with noise. Thus a common requirement is to apply some denoising methodology to $\mathbf{y}$ in order to try to separate the noise from the signal. Throughout this paper we will use both the terms observed 'series' and 'signal' for $\mathbf{y}$, in order to deliberately blur the distinction between the statistical literature (on time series analysis) and signal processing.

Common approaches map the one-dimensional series $\mathbf{y}$ to a multidimensional series of lagged vectors to be

contained within some $L \times K$ matrix. This permits further analysis. For a recent review of such subspace-based techniques, the reader is referred to Golyandina [10]. Such a mapping puts $\mathbf{y}$ into a structured matrix (commonly Hankel) and singular value decomposition (SVD) methods are regularly used to deconstruct the matrix into a sum of rank-one components. Under certain conditions, components which are associated with noise, and components associated with the true signal may be separated.

The aim of this paper is to introduce a computationally efficient difference-based estimator of the noise in $\mathbf{y}$ in order to automate the separation of the noise from the signal. Common approaches concentrate on estimation in the frequency domain [5, 13] whilst there has also been much discussion in the time domain [18, 20]. The novelty of this paper is in using difference-based methods to estimate the variance of the noise, this estimate can then be utilised to estimate the rank of the Hankel matrix in which the signal $\mathbf{y}$ is embedded. This will lead to the truncated SVD which is often computed to estimate the noise-free signal. Different modifications of the SVD are also discussed. The estimation of the variance of the noise is a largely debated topic in time series analysis. In this paper we solely wish to demonstrate the potential of difference-based methods for assisting with the truncation of the SVD. We do not wish to offer a comprehensive comparison study of our method with the many alternatives available in the literature. We now describe in more detail the form of the SVD we shall use throughout the remainder of the paper.

## 1.2 Singular value decomposition (SVD)

Let the time series or signal $\mathbf{y}$ be mapped onto an $L \times K$ Hankel matrix $\mathbf{X}$ as follows:

$$\mathbf{X} = [X_1, \ldots, X_K] = \begin{pmatrix} y_1 & y_2 & \cdots & y_K \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_N \end{pmatrix}. \tag{1}$$

$L$ is a parameter known as the window length, an integer such that $2 \leq L < N$. The window length $L$ is the sole parameter in this mapping. Selection of $L$ depends on the problem in hand and on preliminarily information about $\mathbf{y}$. Namely, if we know that $\mathbf{y}$ has a periodic component with an integer period, then for better separability of this component it is advisable to take $L$ proportional to that period (see [17]). Empirical results tell us that $L$ should be large enough but not greater than $\frac{N}{2}$ (see for example [11, 16]).

Assuming that $L = \min\{L, K\}$, the SVD of $\mathbf{X}$ is given by $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{L \times L}$ , $\mathbf{V} \in \mathbb{R}^{K \times K}$ and $\mathbf{\Sigma} \in \mathbb{R}^{L \times K}$ consists of $L$ singular values with $\sigma_1 \geq \ldots \geq \sigma_L$. The Eckart-Young-Mirsky theorem (see [6] for example) states that the closest rank $r$ approximation (with respect to the Frobenius norm) is given

by setting $\sigma_{r+1} = \ldots = \sigma_L = 0$ and computing $\mathbf{X}_r = \mathbf{U}\mathbf{\Sigma}_r\mathbf{V}^T$ where $\mathbf{\Sigma}_r$ is the modified matrix of singular values. Also $||\mathbf{X} - \mathbf{X}_r||_F^2 = \sum_{j=r+1}^{L} \sigma_j^2$, where $||\cdot||_F$ is the Frobenius norm.

The singular values may be adjusted by applying some function $f$ to them (in order to obtain a modified least squares estimate [24], a minimum variance estimate [6], a time-domain constraint estimate [7] or to achieve some other estimate [23]). The reconstructed signal component may estimated from $\mathbf{X}_r$ by averaging across its anti-diagonals. This is equivalent to finding a Hankel matrix approximation of $\mathbf{X}_r$ (see [11]).

Forms of the function $f$ described in this paper include

$$f_{LS}(\mathbf{\Sigma}) = \mathbf{\Sigma}$$
$$f_{MV}(\mathbf{\Sigma}) = (\mathbf{\Sigma}^2 - K\hat{\theta}\mathbf{I})\mathbf{\Sigma}^{-1},$$

where $\hat{\theta}$ is an estimate of the variance of the noise within the signal. These are known as the least squares, and minimum variance reconstruction respectively. Derivation and motivation of the selection of these functions is given in De Moor [6]. These forms of the function $f$ have also been studied in [15].

Inherent within this method is the assumption that the pure signal $\mathbf{y}$ lies in a rank $r$ low-dimensional subspace of $\mathbb{R}^N$. A typical aim of many signal processing methods is to approximate this subspace, and estimate the signal within it. Statisticians also often wish to approximate this subspace and perhaps forecast future values of the series based on this approximation. The assumption that $\mathbf{y}$ lies in a low-dimensional subspace of $\mathbb{R}^N$ may not hold exactly in all applications, but is nevertheless a good model for many series and signals [14] and has been demonstrated to work well in speech processing [19] and for a number of statistical applications [26, 3, 2, 27, 9, 21, 22].

### 1.3 Structure of paper

We structure the paper as follows. Section 2 describes the difference-based method to estimate the noise variance in the observed time series. After discussing some properties of our proposed method, Section 3 details how our estimate of the noise variance may be used to separate the noise from the observed series. After a simulation study and analysis of a real-life data example motivated by the problem of denoising measurements of the time variation of the intensity of a white dwarf star (Section 4), the paper is concluded in Section 5. We compare various versions of the SVD popular in the signal processing literature, and show that our method appears effective in estimating the rank of the signal.

# 2 Difference-based methods

## 2.1 Noise variance estimation and assumptions

Here we introduce the decomposition of the time series into the latent signal and noise components, and thus render the contribution of the paper. Suppose that the time series $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ is a discrete realisation of a random process $Y(t)$, observed at times $t = 1, 2, \ldots, N$. Suppose also that $Y(t)$ is the sum of a mean function $s(t)$ and a stationary zero-mean random process $Z(t)$,

$$Y(t) = s(t) + Z(t), \qquad \text{where} \quad s(t) = E\big(Y(t)\big),$$

and the expectation is taken with respect to the distribution of $Z(t)$. The mean function $s(t)$ is the quantity of interest (signal), with $Z(t)$ representing all nuisance variables (noise). We assume that the error term $Z(t)$ is a zero mean stationary process. We also assume that this process is homoscedastic with its distribution being independent of $t$. This is true if the process strongly stationary with higher order moments (specifically moments of order larger than two) not depending on time, or if the process is weakly stationary and Gaussian.

If the signal $s(t)$ is deterministic, then it follows that $s(t)$ and $Z(t)$ are uncorrelated. If the signal is stochastic, we make the assumption that $s(t)$ and $Z(t)$ are uncorrelated, independent processes.

Let $Y = (Y_1, Y_2, \ldots, Y_N)$ represent the discrete-time random sample, with $Y_t = s_t + Z_t$ for $t = 1, 2, \ldots, N$, and let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be a realisation of the sample. We assume that the noise is homoscedastic, so that $Z_1, \ldots, Z_N$ are independent and identically distributed zero-mean random variables with common variance $\theta = \mathrm{var}(Z)$. It is of interest to extract the signal $s(t)$ from the observed time series $Y_t$. To achieve this, we first need an estimate of the noise variance $\theta$.

## 2.2 Difference-based methods

Difference-based methods for estimating residual variance in time series can be traced back to von Neumann [25] who, for the additive model $Y_t = s_t + Z_t$ and homoscedastic noise, proposed that the variance $\theta = \mathrm{var}(Z)$ can be estimated by the average of the squared differences between successive observations,

$$\hat{\theta} = \frac{1}{2(N-1)} \sum_{t=2}^{N} (y_t - y_{t-1})^2 \,. \tag{2}$$

By independence, $\mathrm{var}\big(Y(t)\big) = \mathrm{var}\big(s(t)\big) + \mathrm{var}\big(Z(t)\big)$. The *signal-to-noise* ratio is defined to be the ratio of

the signal variance and the noise variance,

$$SNR = \frac{\mathrm{var}\big(s_t\big)}{\mathrm{var}\big(Z(t)\big)} = \frac{\mathrm{var}\big(Y(t)\big)}{\mathrm{var}\big(Z(t)\big)} - 1$$

As the SNR decreases, it becomes more difficult to extract the signal from the observed time series.

### 2.2.1 Proposed difference-based estimator

In this paper we proceed with the difference-based estimator introduced by Gasser [8]. This estimator, which we call $\Delta$, is a function of the second central difference operator of half-width $\tau \in \mathbb{R}^+$,

$$\Delta_Y^2(t,\tau) = \big(Y(t+\tau) - 2Y(t) + Y(t-\tau)\big)^2$$

In the discrete case $k \in \mathbb{Z}^+$,

$$\Delta_Y^2(t,k) = \big(Y_{t+k} - 2Y_t + Y_{t-k}\big)^2$$

$Z(t)$ is a zero-mean process, and $s(t)$ and $Z(t)$ are independent. Thus the expected value of $\Delta_Y^2(t,\tau)$ taken with respect to the distribution of $Z(t)$ (with $t$ fixed) satisfies

$$E\big(\Delta_Y^2(t,\tau)\big) = \Delta_s^2(t,\tau) + E\big(\Delta_Z^2(t,\tau)\big) \tag{3}$$

Let $\mathcal{D}_Y^2(\tau)$ be the expected value of $\Delta_Y^2(t,\tau)$ over all $t$. If $Z(t)$ is a stationary process,

$$\mathcal{D}_Y^2(\tau) = \mathcal{D}_s^2(\tau) + \mathcal{D}_Z^2(\tau) \tag{4}$$

For $\tau > 0$, if $s(t)$ is Lipschitz continuous, then by the mean value theorem it follows that

$$s(t-\tau) = s(t) - \tau s'(t) + \frac{1}{2}\tau^2 s''(t)\big(1 + O(\tau)\big) \qquad \text{as } \tau \to 0$$
$$s(t+\tau) = s(t) + \tau s'(t) + \frac{1}{2}\tau^2 s''(t)\big(1 + O(\tau)\big) \qquad \text{as } \tau \to 0$$

Thus for sufficiently smooth functions $s(t)$ we have

$$\Delta_s^2(t,\tau) = \tau^4 s''(t)^2\big(1 + O(\tau)\big) \qquad \text{as } \tau \to 0$$

and

$$\mathcal{D}_s^2(\tau) = \tau^4 E\big(s''(t)^2\big)\big(1 + O(\tau)\big) \qquad \text{as} \qquad \tau \to 0$$

Note that

$$s''(t) = \lim_{\tau \to 0} \frac{Y(t+\tau) - 2Y(t) + Y(t-\tau)}{\tau^2} = \lim_{\tau \to 0} \frac{\Delta_s(t,\tau)}{\tau^2}$$

Under the resolution imposed by the sampling rate, we can restrict our attention to functions of the form

$$s(t) = \int_0^\infty S(\omega) e^{i\omega t}\, d\omega \qquad \big(S(\omega) \in \mathbb{C}\big) \tag{5}$$

where $\omega = 2\pi f$ is the angular frequency. The discrete nature of the time series $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ limits the frequency resolution to $f \in [1, N]$, or equivalently to periodicities $T \in [1/N, 1]$ . High frequencies $f > N$, corresponding to short periods $T < 1$, appear as noise in the time series; while low frequencies $f < 1$, corresponding to long periods $T > N$, appear as trend.

### 2.2.2 Main idea

Since $Z(t)$ is a zero-mean stationary process,

$$\begin{aligned}
\mathcal{D}_Z^2(\tau) = E\big(\Delta_Z^2(t,\tau)\big) &= E\Big[\big(Z(t-\tau) - 2Z(t) + Z(t+\tau)\big)^2\Big] \\
&= E\big(Z(t-\tau)^2 + 4Z(t)^2 + Z(t+\tau)^2\big) - 4E\big(Z(t)Z(t-\tau) + Z(t)Z(t+\tau)\big) + 2E\big(Z(t-\tau)Z(t+\tau)\big) \\
&= 6\mathcal{R}_Z(0) - 8\mathcal{R}_Z(\tau) + 2\mathcal{R}_Z(2\tau)
\end{aligned}$$

where $\mathcal{R}_Z(\tau)$ is the autocovariance function of $Z(t)$,

$$\mathcal{R}_Z(\tau) = E\big(Z(t)Z(t-\tau)\big) = \int_{-\infty}^{\infty} Z(t)Z(t-\tau)\, dt$$

and $\mathcal{R}_Z(0) = \mathrm{var}(Z)$ is the variance of the noise process $Z(t)$.

Thus we have

$$\begin{aligned}
\mathcal{D}_Y^2(\tau) &= \mathcal{D}_s^2(\tau) + \mathcal{D}_Z^2(\tau) \\
&= \mathcal{D}_s^2(\tau) + 6\mathcal{R}_Z(0) - 8\mathcal{R}_Z(\tau) + 2\mathcal{R}_Z(2\tau)
\end{aligned}$$

Note that $\mathcal{D}_s^2(\tau) \to 0$ as $\tau \to 0$.

For brevity of notation, let $\eta(\tau) = \frac{1}{6}\mathcal{D}_Y^2(\tau)$ so that

$$\eta(\tau) = \text{var}(Z) - \frac{4\mathcal{R}_Z(\tau)}{3} + \frac{\mathcal{R}_Z(2\tau)}{3} + \frac{\mathcal{D}_s^2(\tau)}{6}$$

Note that $\eta(\tau) \to 0$ as $\tau \to 0$.

Let $\tau_c$ be a value of $\tau$ for which the (negative) term $\frac{-4\mathcal{R}_Z(\tau)}{3} + \frac{\mathcal{R}_Z(2\tau)}{3}$ due to noise correlation cancels the (positive) term $\frac{\mathcal{D}_s^2(\tau)}{6}$ due to local non-linearity of $s(t)$;

$$\tau_c = \arg\min_{\tau} \left\{ \left| 8\mathcal{R}_Z(\tau) - 2\mathcal{R}_Z(2\tau) - \mathcal{D}_s^2(\tau) \right| \right\}$$

If we can estimate $\tau_c$, we then immediately have an estimate $\hat{\theta} = \eta(\hat{\tau}_c)$ for the noise variance $\text{var}(Z)$.

## 2.3 Estimation of $\tau_c$

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, be a realisation of the random process $Y_t = s_t + Z_t$, observed at times $t = 1, \ldots, N$. Based on the previous discussion, we compute a sequence $0 = \eta_0, \eta_1, \eta_2, \ldots$ defined by

$$\eta_k = \frac{1}{6(N-2k)} \sum_{t=k+1}^{N-k} (y_{t-k} - 2y_t + y_{t+k})^2 = \frac{1}{6(N-2k)} \sum_{t=k+1}^{N-k} \Delta_{\mathbf{y}}^2(t,k) \qquad k = 0, 1, 2, \ldots, k_{\max}$$

We need to estimate $\tau_c$ from the sequence $\eta_0, \eta_1, \ldots$.

Let $\tau_1, \tau_2 \in [0, \infty)$ be such that

$$\begin{aligned} \mathcal{R}_Z(\tau) &\approx 0 && \text{for all} \quad \tau > \tau_1 \\ \mathcal{R}_s(\tau) &\approx \mathcal{R}_s(0) && \text{for all} \quad \tau < \tau_2 \,. \end{aligned} \tag{6}$$

Here, $\tau_1$ is the *correlation length* of the noise process and $\tau_2$ depends on the amplitude of the minimum periodicity $T_{\min}$ present in the signal. To separate the signal $s(t)$ from the observed process $Y(t)$, the correlation length $\tau_1$ of the noise process $Z(t)$ must be of smaller order than the minimum periodicity $T_{\min}$ present in the signal. If $\tau_1 < \tau_2$, the function $\eta(\tau)$ satisfies

$$\eta(\tau) \approx \begin{cases} \text{var}(Z) - \frac{4\mathcal{R}_Z(\tau)}{3} + \frac{\mathcal{R}_Z(2\tau)}{3} & \tau < \tau_1 \\ \text{var}(Z) & \tau_1 \le \tau \le \tau_2 \\ \text{var}(Z) + \frac{\mathcal{D}_s^2(\tau)}{6} & \tau > \tau_2 \end{cases}$$

If $\tau_1 < \tau_2$, the function $\eta(\tau)$ will increase as $\tau$ increases from 0 to $\tau_1$, stabilize around $\text{var}(Z)$ as $\tau$ increases from $\tau_1$ to $\tau_2$, then start to increase again as $\tau$ moves beyond $\tau_2$. Thus we wish to estimate $\tau_1$, the correlation length of the noise process. To remain within the scale at which $s(t)$ is approximately linear, $\tau_2$ should be no greater than 1/4 of the minimum periodicity that we wish to detect. Up to this point, $\eta(\tau)$ is a non-decreasing function of $\tau$. If $\eta(\tau)$ stabilises between $\tau_1$ and $\tau_2$, the onset of the plateau at $\tau_1$ can thus be estimated by the point at which the empirical second derivative (second order difference) of the sequence $\eta_0, \eta_1, \ldots, \eta_{k_{\max}}$ reaches its minimum value, which indicates the point of maximum curvature.

Although $\eta(\tau)$ is a non-decreasing function over $[0, \tau_2]$, statistical fluctuations may produce spurious local minima in the empirical derivatives of the sequence $\eta_0, \eta_1, \ldots, \eta_{k_{\max}}$. To mitigate such effects, we construct an estimator $\hat{\eta}(\tau)$ of the function $\eta(\tau)$, based on the sequence $\eta_0, \eta_1, \ldots, \eta_{k_{\max}}$. The general form of such an estimator should have $\hat{\eta}(0) = 0$, and be able to model two inflexion points $\tau_1$ and $\tau_2$ (the minimum second derivative will lie between these inflexion points). The presence of two inflexion points in $[0, \tau_2]$ indicates that $\tau_1 < \tau_2$, and therefore that the signal can be separated from the noise.

To estimate the correlation length $\tau_1$, since $\mathcal{D}_s^2(\tau) \sim \tau^2$, it may be advantageous to consider the reduced function

$$\eta^*(\tau) = \frac{\eta(\tau)}{\tau}$$

and construct a polynomial estimate of the form

$$\hat{\eta}^*(\tau) = a_0\tau^3 + a_1\tau^2 + a_2\tau + a_3$$

Alternatively, we might consider $\eta(\tau)/\tau^2$, which makes the bias due to the local non-linearity of the signal independent of $\tau$. In either case, the 'peak', 'knee' or 'trough' we wish to detect still occurs at $\tau = \tau_c$. Of course, the estimate $\hat{\tau}_c$ for the correlation length of the noise sequence should be substituted into $\hat{\eta}$ to obtain an estimate for $\theta = \text{var}(Z)$.

# 3 Automatic truncation of the SVD

Let $\mathbf{X}$ denote the observed $L \times K$ trajectory matrix, and let $\mathbf{X}_S$ and $\mathbf{X}_Z$ denote the (unobserved) trajectory matrices, due to signal and noise respectively (as defined earlier, $\mathbf{X}_S = ||s_{i+j-1}||_{i,j=1}^{L,K}$ and $\mathbf{X}_Z = ||z_{i+j-1}||_{i,j=1}^{L,K}$). The SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ of $\mathbf{X} = \mathbf{X}_s + \mathbf{X}_Z$ yields a sequence of singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_L \geq 0$, where $\sigma_i^2$ quantifies the variance along the $i$th (rotated) coordinate.

It follows that,

$$\|\mathbf{X}\|_F^2 = \sum_{i=1}^{L}\sum_{j=1}^{K} x_{i+j-1}^2 = \text{trace}(\mathbf{X}\mathbf{X}^T) = \sum_{i=1}^{L}\sigma_i^2$$

This corresponds to the total energy/variability in the matrix entries. By independence we have $\|\mathbf{X}\|_F = \|\mathbf{X}_S\|_F + \|\mathbf{X}_Z\|_F$. The expected norm of the noise matrix $\mathbf{X}_Z$ satisfies

$$E(\|\mathbf{X}_Z\|_F^2) = \sum_{i=1}^{L}\sum_{j=1}^{K} E(z_{i+j-1}^2) = LK\theta \qquad \text{where } \theta = \text{var}(Z)$$

The rank of the signal matrix $\mathbf{X}_S$ can be estimated by the number of singular values that exceed the noise level $\theta = \text{var}(Z)$. Because the singular values are listed in decreasing order of magnitude, our estimator is obtained by truncating all singular values that are below the noise level

$$\hat{r} = \arg\max_i \{\sigma_i^2 > K\hat{\theta}\} \tag{7}$$

Let

$$\mathbf{\Sigma}_{\hat{r}} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\hat{r}}, 0, 0, \ldots, 0) \in \mathbb{R}^{L\times L}$$

where $\hat{r}$ is an estimate for the true rank. We now compute $\widehat{\mathbf{X}}_S = \mathbf{U}_1 f(\mathbf{\Sigma}_{\hat{r}})\mathbf{V}_1^T$. The signal estimate $\hat{\mathbf{s}}$ may be extracted from $\widehat{\mathbf{X}}_S$ by averaging $\widehat{\mathbf{X}}_S$ over its anti-diagonals.

Forms of the function $f$ considered in this paper are

$$f_{LS}(\mathbf{\Sigma}_{\hat{r}}) = \mathbf{\Sigma}_{\hat{r}} \tag{8}$$

$$f_{MV}(\mathbf{\Sigma}_{\hat{r}}) = (\mathbf{\Sigma}_{\hat{r}} - K\hat{\theta}\mathbf{I}_r)\mathbf{\Sigma}_r^{-1} \tag{9}$$

As described earlier, it can be shown that $f_{LS}$ yields a least-squares estimate for $\mathbf{X}_S$, while $f_{MV}$ yields a minimum variance estimate (see for example De Moor [6]).

In summary, our proposed algorithm is described below:

1. Construct the Hankel trajectory matrix $\mathbf{X}$ from the observed signal.
2. Compute the SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
3. Compute an estimate $\hat{\theta}$ of the noise variance $\theta = \text{var}(Z)$.
4. Compute an estimate $\hat{r}$ of the rank $r$ of the signal matrix $\mathbf{X}_S$.
5. Choose a reconstruction function $f(\mathbf{\Sigma}_r)$ and estimate the signal matrix $\widehat{\mathbf{X}}_S = \mathbf{U}_1 f(\mathbf{\Sigma}_r)\mathbf{V}_1^T$.

6. Extract the signal estimate $\hat{\mathbf{s}}$ from $\widehat{\mathbf{X}}_S$ (average over anti-diagonals).

The automatic extraction and forecast of time series components within the framework of a subspace-based technique known as singular spectrum analysis is described in [1] and [12].

# 4 Simulation studies and examples

In a time series of $N$ observations, slowly-varying components (of period $T > N$ or frequency $\omega < 2\pi/N$) appear in the time series as trend, while rapid oscillations (of period $T \leq 1$ or frequency $\omega \geq 2\pi$) cannot be resolved, and appear as noise.

Suppose that $s(t)$ is a sinusoidal function with periodicities $T \in \{1, \ldots, N\}$, which we write as

$$s(t) = \sum_{n=1}^{N} C_n e^{i\omega_n t} \quad \text{where } \omega_n = \frac{2\pi}{T_n} \text{ and } C_n \in \mathbb{C}$$

In each experiment, we choose a random phase $\phi \sim \text{Uniform}[0, 2\pi]$ and consider sequences of the form

$$s(t) = \sum_{n=1}^{N} A_n e^{i\phi} e^{i\omega_n t} \quad \text{where } A_n \in \mathbb{R} \text{ is an amplitude.}$$

To construct our synthetic time series, the signal process $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ and noise process $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ are realised independently, then combined to form the observed time series $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ having first been re-scaled to achieve the required signal-to-noise ratio, and such that the observed time series has zero mean and unit variance. Let the correlation length be given by $\ell$.

## 4.1 Illustrative example

As a simple illustrate example, consider a signal with $N = 1000$. We also take $L = 500$. As for $L \times K$ Hankel matrices. $N = L + K - 1$ then $K = 501$ The signal is generated with frequencies $f \in \{5, 10\}$, correlation length $\ell = 5$, and $SNR = 1$. As an additional demonstration, we will also forecast ahead $M = 100$ points. We produce the forecast using the linear recurrent formula constructed from $\hat{\mathbf{X}}_s$, for further details the reader is referred to [11]. The true signal, noise signal, and the resultant series from summing both of these is provided in Figure 1. The true rank of the signal is $r = 4$. In this example we adjust the singular values by using the minimum variance method (as given in equation (9)).

Figure 2 contains plots of the inverted autocovariance function and mean squared central differences. The

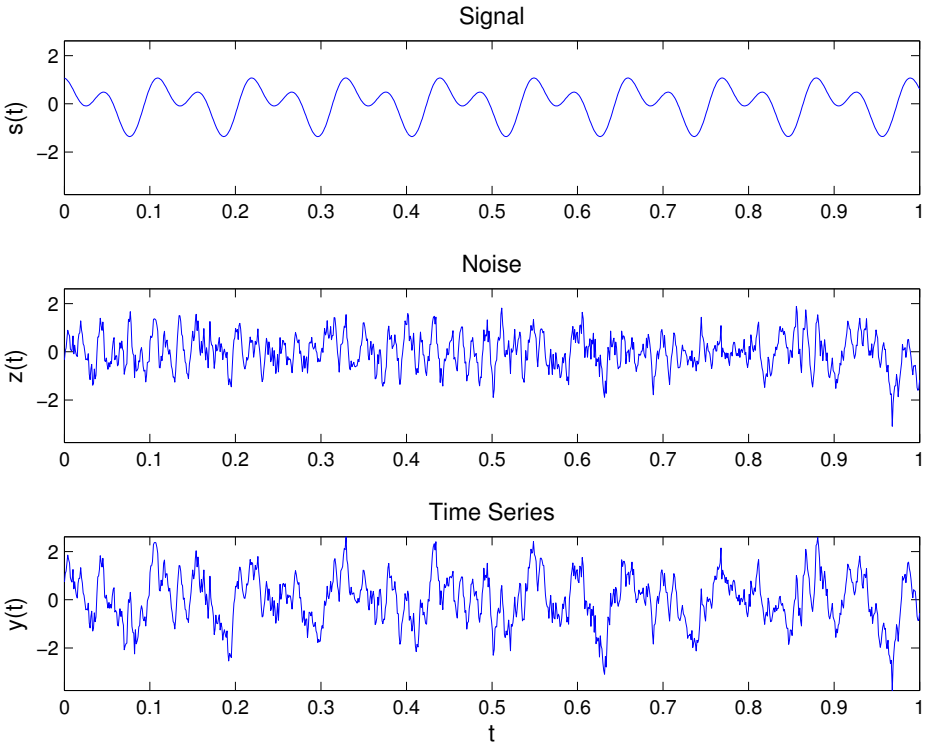
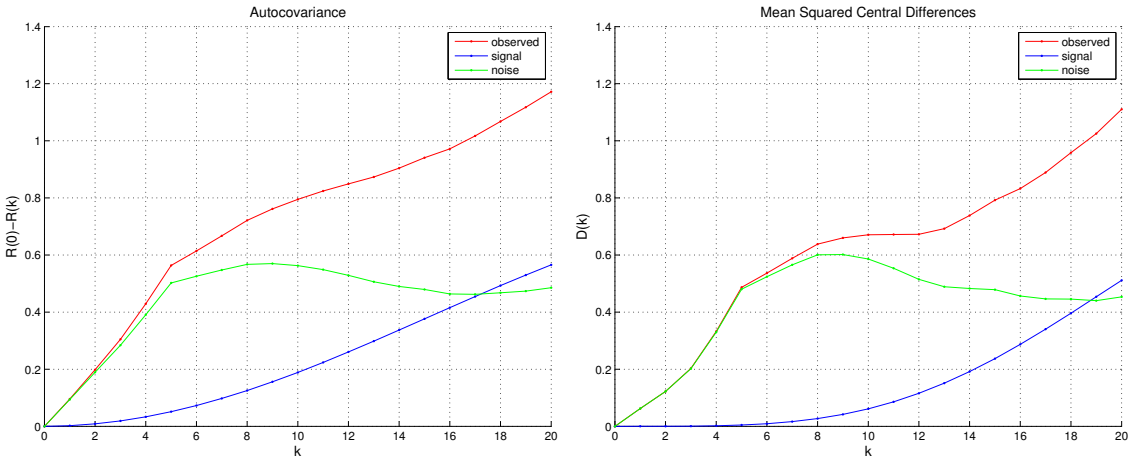



Figure 1: True signal, noise and combined sequences.

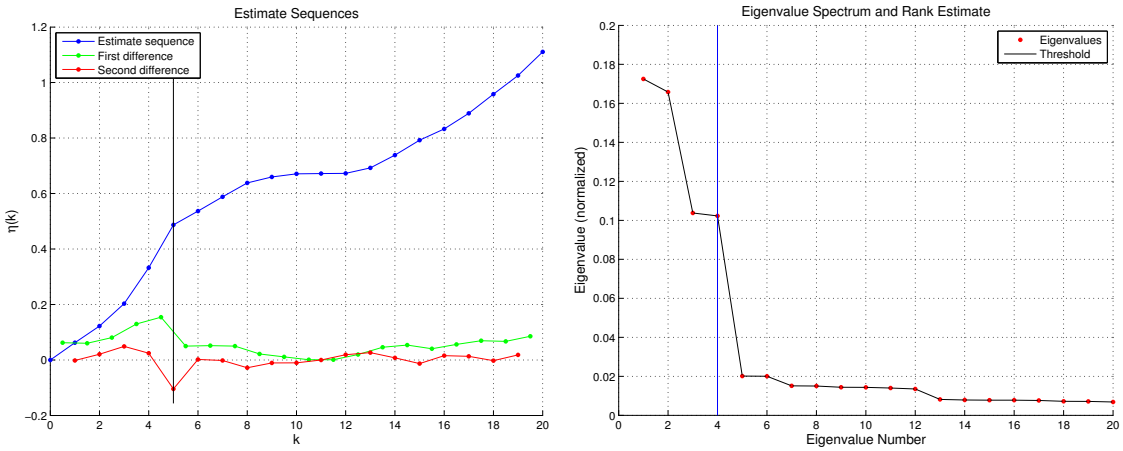
correlation length of $\ell = 5$ can be confirmed; both plots demonstrate noticeable 'kinks' at this value. Figure 3 demonstrates the core idea of our methodology. The noise variance $\theta$ is estimated from the sequence $\eta_1, \eta_2, \ldots$ as defined in Section 2.3. There is a clear dip in the second differences of sequence $\eta_1, \eta_2, \ldots$ located at 5, and thus our noise is estimated as $\hat{\theta} = \hat{\eta}(5) = 0.4867$. The true value of $\theta$ was selected to be 0.5 for this example. We estimate the rank by computing $\hat{r} = \arg\max_i \{\sigma_i^2 > K\hat{\theta}\} = 4$.




(a) $\mathcal{R}(0) - \mathcal{R}(k)$ (b) $\mathcal{D}(k)$

Figure 2: Inverted autocovariance and mean squared central differences.

Figure 4 contains the observed data, the true, noise free-signal, our reconstruction and forecast. The retro-




(a) Noise estimate sequences

(b) Eigenvalue spectrum and rank estimate

Figure 3: Noise and rank estimation.

spective MSE for our model for the signal is 0.016749, whilst the MSE for our forecast is 0.055313.

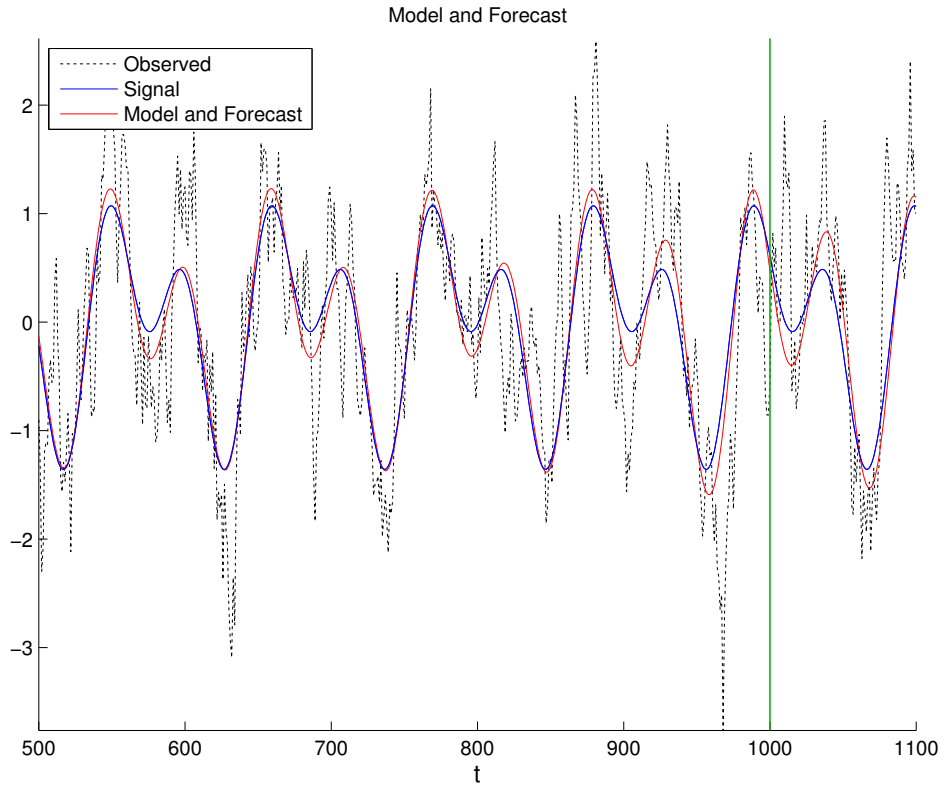



Figure 4: Model and forecast.

## 4.2 Experimental Scheme

For the two simulation studies that follow, we adopt the following parameter settings.

| | |
|---|---|
| Frequencies | 10, 20, 30, 40, 50 |
| True rank (r) | 10 |
| Repetitions | 1000 |

We also select $L = \frac{N}{2}$ with $N = L + K - 1 = 1000$. In our simulation studies below we consider the impact of increasing the correlation length $\ell$, and reducing the signal-to-noise ratio SNR.

#### 4.2.1 Simulation Study 1

In this simulation study we attempt to estimate the rank $r$ of the true signal, while increasing the correlation length $\ell$. Table 1 contains $\hat{r}$ taken over 1000 simulations, with varying $\ell$. As the correlation increases, then the performance of our rank estimator worsens. Potential reasons for this are illustrated in Figure 5.

| $\ell$ | Mean | SD |
|---|---|---|
| 1 | 11.0411 | 2.2641 |
| 3 | 8.9812 | 1.0001 |
| 5 | 6.7864 | 0.9071 |
| 7 | 3.7770 | 0.8444 |

Table 1: Table of the mean and SD of values of $\hat{r}$ taken over 1000 simulations, with varying $\ell$.

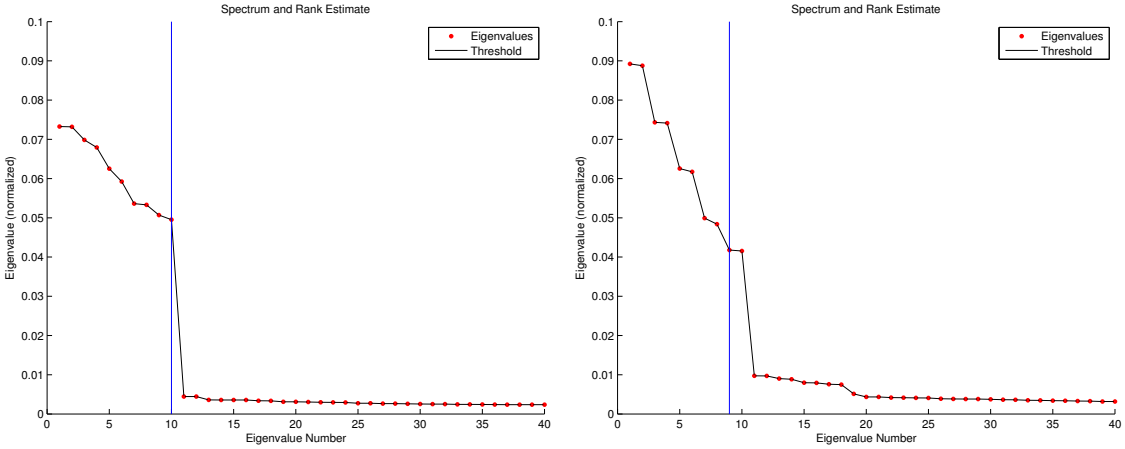
Figure 5 contains plots of typical spectra observed for different correlation lengths. The effect of increasing $\ell$ is to dampen the sharp drop in the value of the observed eigenvalues that appears over $r = 10$, and increase the number of sudden changes in the eigenvalue spectra. As a result, increasing $\ell$ gives the impression that there (at least visually) a number of possibilities of a valid choice for an estimated $r$.

#### 4.2.2 Simulation Study 2

In this simulation study we attempt to estimate the rank $r$ of the true signal, while decreasing the SNR. We also compare the MSE of our reconstructions, and the MSE of a forecast $M = 100$ points ahead using both the least squares (8) and minimum variance (9) adjustment to the singular values.

Table 2 contains the mean values (and standard deviations in brackets) of our estimated ranks, the MSE of our reconstruction, and our forecast, averaged over 1000 simulations. Our estimator is seemingly robust to decreasing the SNR, only when $SNR = 0.1$ do we underestimate the rank. The standard deviation of our estimate of the rank also increases for this value of the SNR.

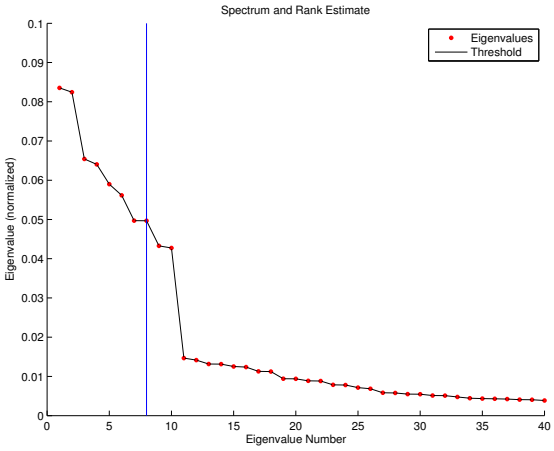
The MSE's for both the reconstruction of the data and the forecast increase as the SNR decreases, and this effect is to be expected. There is a marginal improvement in the MSE's using the minimum variance



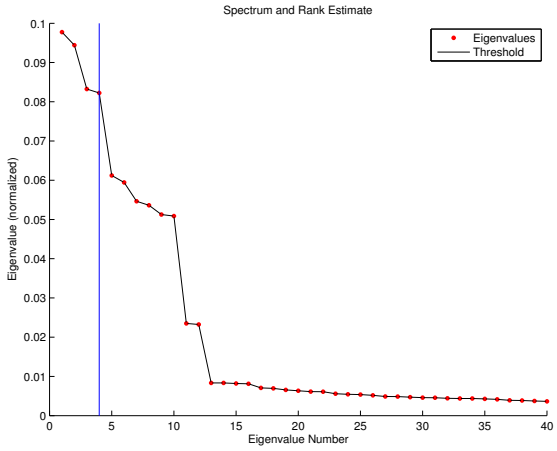

(a) $\ell = 1$ ($\hat{r} = 10$)




(b) $\ell = 3$ ($\hat{r} = 9$)




(c) $\ell = 5$ ($\hat{r} = 8$)




(d) $\ell = 7$ ($\hat{r} = 4$)

Figure 5: Spectra for various correlation lengths ($r = 10$)

adjustment to the singular values (see equation (9)), but the differences between the least squares and minimum variance adjustments are not statistically different. The point of this example is to take comfort in the observation that the automatic procedure of estimating the rank closely agrees, under moderate SNR, with the true value of the rank fixed in the study.

| | | Model MSE | | Forecast MSE | |
|---|---|---|---|---|---|
| SNR | Rank | LS | MV | LS | MV |
| 1 | 10.8123 (2.2122) | 0.0214 (0.0107) | 0.0210 (0.0089) | 0.0476 (0.0358) | 0.0453 (0.0352) |
| 0.5 | 10.1301 (2.1111) | 0.0320 (0.0171) | 0.0317 (0.0171) | 0.0669 (0.0333) | 0.0622 (0.0330) |
| 0.25 | 9.3309 (2.3735) | 0.0459 (0.0211) | 0.0356 (0.0210) | 0.0718 (0.0326) | 0.0666 (0.0317) |
| 0.1 | 7.0009 (3.8221) | 0.0655 (0.0235) | 0.0600 (0.0215) | 0.0899 (0.0379) | 0.0844 (0.0274) |

Table 2: The effect of SNR on rank estimation, model reconstruction and forecasting. Mean values of the rank, with standard deviations in brackets provided. Average MSE values (and standard deviations) for the least squares (LS) and minimum variance (MV) adjustments for the singular values for both the model reconstruction, and the forecast also provided.

Figure 6 contains plots of typical eigenvalue spectra for varying SNR. The effect of increasing the noise variance is to dampen the sharp dip observed in the spectra over the point $r = 10$. Hence decreasing the SNR makes it exceptionally difficult to visually identify the rank of the true signal.

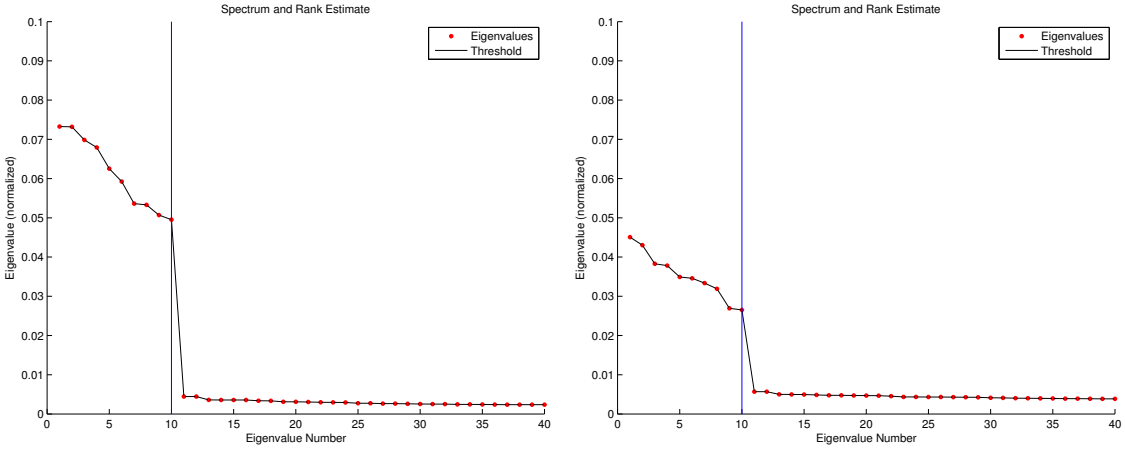
### 4.3 Application: Removing noise from 'White Dwarf' data

In this example we consider the so-called 'White Dwarf' data which contains $N = 618$ noisy measurements of the time variation of the intensity of the white dwarf star PG1159-035 during March 1989. The data were also discussed by Clemens [4]. For simplicity we took $L = 309$.

Figure 7 contains the original 'White Dwarf' data along with the $\hat{r} = 13$ reconstruction (as found by our methodology). Solely from inspecting the principal components of the data, Golyandina et al. [11] also investigated the data, and suggested that $\hat{r} = 11$. However their result was based on visual inspections of the data; they did not perform any formal analysis of the rank.

Figure 8 contains a time series plot of the residuals of the data with our $\hat{r} = 13$ fit; the residuals seem to have no evident structure. Moreover, further analysis suggests that the residuals appear as Gaussian white noise (see Figure 9).

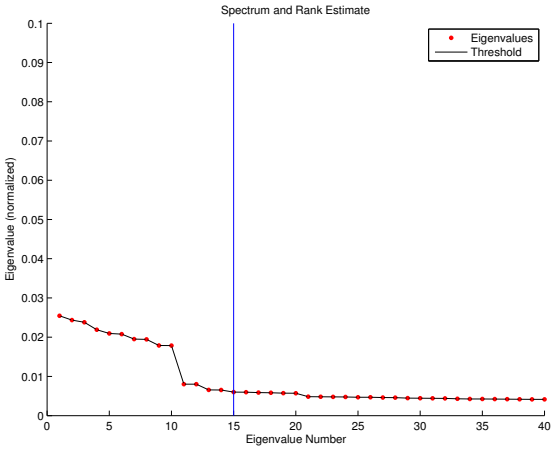
Thus we may assume in this case the smoothing procedure leads to noise reduction. This example also provides a further motivation for our paper. Data such as that considered in this example is now in abundance; it is desirable to have computationally efficient methods to denoise, and reconstruct the data prior to additional analyses. Note however that in this example there may be additional improvement to be gained upon a more refined selection of the parameter $L$. In this example we solely wish to demonstrate the potential of



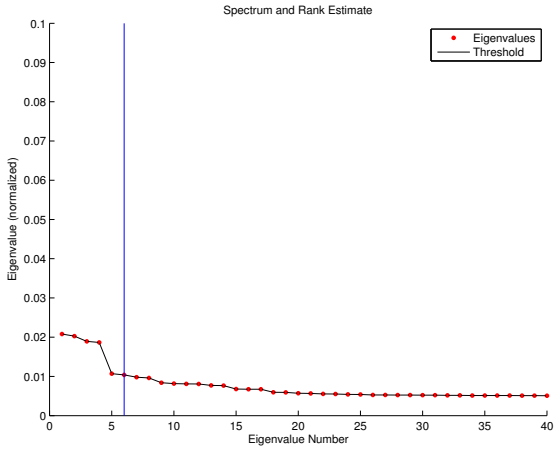

(a) SNR=1




(b) SNR=0.5




(c) SNR=0.25




(d) SNR=0.1

Figure 6: Spectra for various SNR

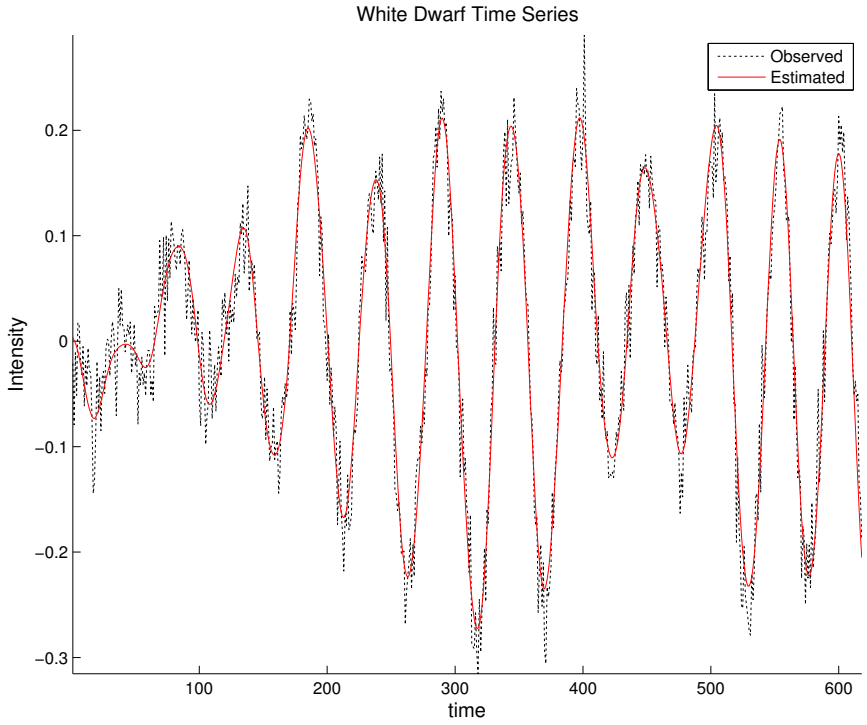



Figure 7: White Dwarf data with reconstruction with $\hat{r} = 13$.

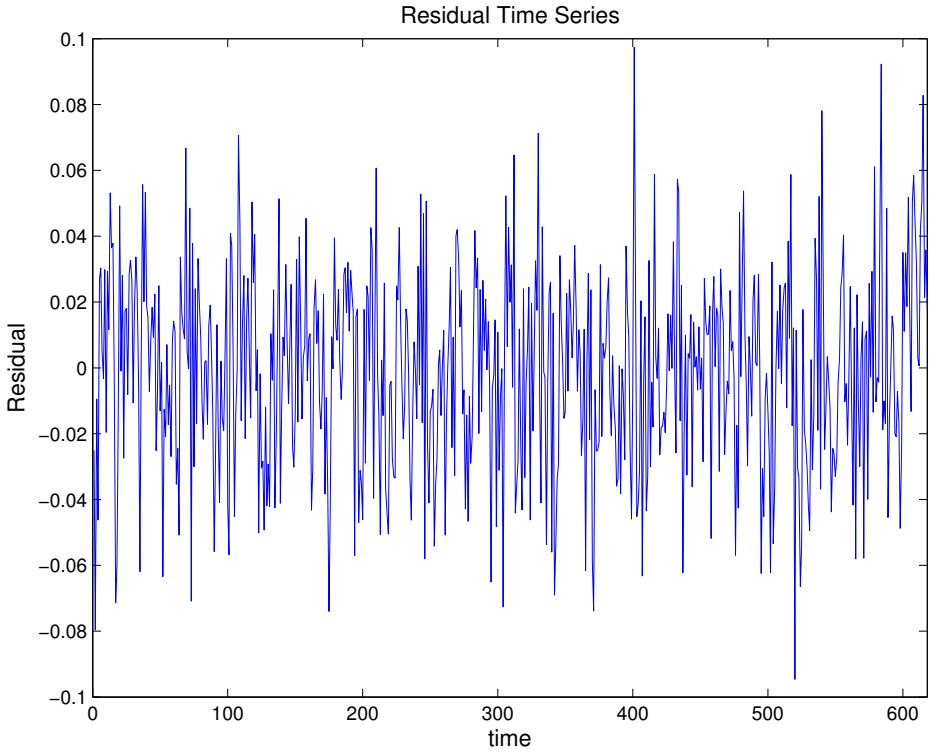
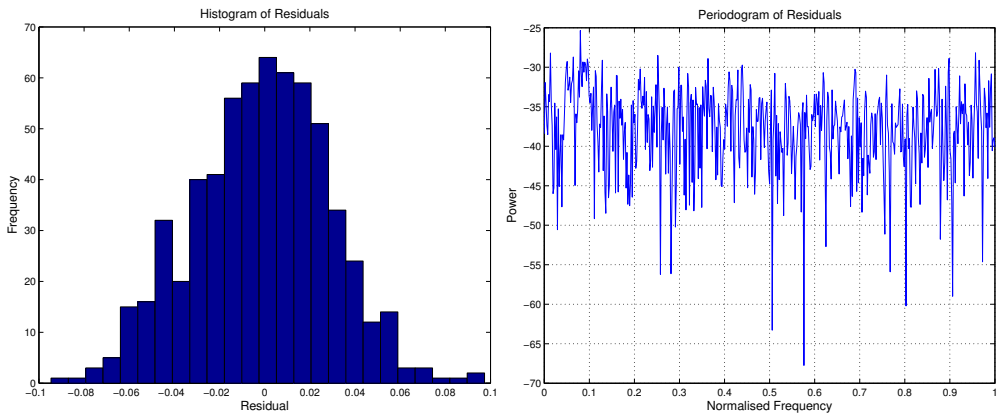



Figure 8: Residuals of reconstruction with $\hat{r} = 13$.




(a) Histogram of residuals

(b) Periodogram of residuals

Figure 9: Further analysis of residuals given in Figure 8

the difference-based method proposed. A recent discussion as to optimal selection of $L$ is available in [16].

# 5 Conclusion

In this paper we have proposed a computationally efficient difference-based estimator of the noise variance $\theta$, and consequently an estimator for the true rank of the signal $r$. As alluded to in the previous example, with the number of large data sets now becoming available across a number of occasions, a method which may reduce the model order of the observed data automatically is likely to be appealing to a number of practitioners. Further work will investigate the statistical properties of our estimators, and will concentrate on developing methods which will produce robust forecasts on the basis of our model reconstructions.

It is important to note that any automatic procedure heavily depends on the model assumptions in hand. Moreover, any automatic procedure may occasionally give very wrong results (if these assumptions fail). It would be a sensible idea to prevent serious mistakes by testing (using appropriate samples of the observations) whether the methodology proposed give adequate and stable results.

# References

[1] Th. Alexandrov and N. Golyandina. Automatic extraction and forecast of time series cyclic components within the framework of ssa. In *Proc. of the 5th St. Petersburg Workshop on Simulation*, pages 45–50, 2005.

[2] D. S. Broomhead, R. Jones, G. P. King, and E. R. Pike. *Singular system analysis with application to dynamical systems*. CRC Press, Bristol, 1987.

[3] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Physica D*, 20(2-3):217–236, 1986.

[4] J. C. Clemens. *Whole earth telescope observation of the white dwarf star PG1159-035*, volume Time Series Prediction: Forecasting the Future and Understanding the Past. Addison-Wesley, Reading, 1994.

[5] H. T. Davis and R. H. Jones. Estimation of the innovation variance of a stationary time series. *J. Amer. Statist. Assoc.*, 63:141–149, 1968.

[6] B. de Moor. Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Trans. on Signal Processing*, 42(11):3104–3113, 1994.

[7] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. on Speech and Audio Processing*, 3(4):251–266, 1995.

[8] T. Gasser, L. Sroka, and C. Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633, 1986.

[9] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):3.1–3.41, 2001.

[10] N. Golyandina. On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Stat. Interface*, 3(3):259–279, 2010.

[11] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of time series structure*, volume 90 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2001. SSA and related techniques.

[12] N. Golyandina and A. Zhigljavsky. *Singular Spectrum Analysis for Time Series*. Springer, 2013.

[13] E. J. Hannan and D. F. Nicholls. The estimation of the prediction error variance. *J. Amer. Statist. Assoc.*, 72(360, part 1):834–840, 1977.

[14] P. C. Hansen and S. H. Jensen. Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: survey and analysis. *IEEE Trans. on Signal Processing*, Volume 2007:1–24, 2007.

[15] H. Hassani. Singular spectrum analysis based on the minimum variance estimator. *Nonlinear Analysis: Real World Applications*, 11(3):2065–2077, 2010.

[16] H. Hassani, R. Mahmoudvand, and M. Zokaei. Separability and window length in singular spectrum analysis. *Comptes Rendus Mathematique*, 349(17):987–990, 2011.

[17] H. Hassani and A. Zhigljavsky. Singular spectrum analysis: methodology and application to economics data. *J. Syst. Sci. Complex.*, 22(3):372–394, 2009.

[18] C. Loader. *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York, 1999.

[19] R. J. McAulay and T. F. Quateri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

[20] M. Pourahmadi. *Foundations of time series analysis and prediction theory*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2001.

[21] X. Rodo, M. Pascual, G. Fuchs, and A. S. G. Faruque. ENSO and cholera: A nonstationary link related to climate change? *PNAS*, 99(20):12901–12906, 2002.

[22] D. D. Thomakos, T. Wang, and L. T. Wille. Modeling daily realized futures volatility with singular spectrum analysis. *Physica A: Statistical Mechanics and its Applications*, 312(3-4):505–519, 2002.

[23] A. J. Thorpe and L. L. Scharf. Data adaptive rank-shaping methods for solving least squares problems. *IEEE Trans. on Signal Processing*, 43(7):1591–1601, 1995.

[24] S. van Huffel. Enhanced resolution based on minimum variance estimation and exponential data modelling. *Signal Processing*, 33(3):333–355, 1993.

[25] J. von Neumann. Distribution of the ratio of the mean squared successive difference to the variance. *Ann. Math. Stat.*, 12:367–395, 1941.

[26] B. C. Weare and J. S. Nasstrom. Examples of extended empirical orthogonal functions. *Monthly Weather Review*, 110:481–485, 1982.

[27] P. Yiou, E. Baert, and M. F. Loutre. Spectral analysis of climate data. *Surveys in Geophysics*, 17:619–663, 1996.