Automated outlier detection in Singular Spectrum Analysis

Jacques de Klerk *

North-West University, Potchefstroom, South Africa 23239603@nwu.ac.za

Abstract

Singular Spectrum Analysis (SSA) is a powerful non-parametric time series technique which is finding wide application in time series analysis. SSA is particular powerful for time series that exhibit seasonal variation with/without trend components and find application in time series found in market research, economics, meteorology and oceanology, to name but a few. Outliers that might be present in time series can unduly influence model fitting and forecasting results. This paper compares automated outlier identification techniques in SSA by simulating time series from the broad spectrum of time series that SSA can handle. Specific attention is paid to modern robust principal component analysis techniques such as ROBPCA which employs projection pursuit combined with estimation of robust covariance matrices. The latter is employed to outlier maps, which essentially represents multivariate data in a two dimensional plot consisting of projected orthogonal distances plotted against score distances, in order to identify outliers. Promising results are obtained by robust principal component methods and also applying additional convex hull peeling methods to outliers. A well-known time series with an additive outlier present is used to illustrate the usefulness of the techniques.

Keywords: Convex hull peeling, outlier maps, projection pursuit, robust principal component analysis

1. Introduction

Singular Spectrum Analysis (SSA) is a powerful non-parametric time series technique that found its origins in the field of Physics (Takens, 1981; Broomhead and King, 1986). Golyandina et al. (2001) provide a thorough introduction to SSA and can be consulted in gaining insight into the underlying theory and applications.

According to SSA methodology a time series $\{y_t\}_{t=1}^{N}$ is unfolded into the column vectors of a Hankel structured matrix

$$X_{L\times(N-L+1)} = \left(x_{ij}\right)_{i,j=1}^{L,N-L+1} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_{N-L+1} \\ y_2 & y_3 & y_4 & \cdots & y_{N-L+2} \\ y_3 & y_4 & y_5 & \cdots & y_{N-L+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{bmatrix}. \quad (1.1)$$

Note that off-diagonal elements in the matrix are not unique. The matrix has been coined the *trajectory matrix* in SSA literature and places a univariate time series into a multivariate framework. The dimension L into which the column vectors are unfolded is termed the *window length* and restricted by the choice $2 \le L \le floor\left[(N+1)/2\right]$.

Buchstaber (1994) showed that time series sampled from the following broad class of functions with an additive property can be handled by SSA, viz.

$$y(t) = \sum_{k=1}^{K} p_k(t)\exp(\alpha_k t)\sin(2\pi\varpi t + \phi_k) \quad (1.2)$$

where $p_k(t)$ indicate polynomials.

Golyandina et al. (2001) indicated that SSA can actually handle a broader class of functions than the above in the form of finite difference equations or so-called linear recurrent formulae (LRF) of the form

$$y_{t+r} = \sum_{k=1}^{r} a_k y_{t+r-k},\ 1 \le t \le N-r \tag{1.3}$$

where $a_1,...,a_r$ are coefficients and r is the rank (structure) of the time series. It is clear that SSA can handle a wide variety of time series structure which can include trend with/without seasonality. The interested reader can refer to Golyandina et al. (2001) to further understand how singular value decomposition is used to extract signal structures from an observed time series and also how forecasting can be approached. It is clear that SSA has only two "parameters", i.e. the window length (L) and number of leading eigenvectors (r).

Given the above schema, a single additive time series outlier at position $t = t^*$ will be present in consecutive column vectors of the trajectory matrix. It is not difficult to show that, once consecutive column vectors in the trajectory matrix have been flagged as outlying, that the location of the additive time series outlier will be given by

$$t^* = \begin{cases} o_{(t_1-1)} + L - 1 & if \quad \underset{1 \le t_1 < t_1+1 < \ldots \le t_2}{\arg\max} \sum_{t=t_1}^{t_2} I_t\left(o_{(t)} - o_{(t-1)}\right) = L-1 \\ o_{(t_2)} & if \quad \underset{1 \le t_1 < t_1+1 < \ldots \le t_2}{\arg\max} \sum_{t=t_1}^{t_2} I_t\left(o_{(t)} - o_{(t-1)}\right) < L-1 \ and \ 2 \le o_{(t_2)} < L \\ o_{(t_1-1)} + L - 1 & if \quad \underset{1 \le t_1 < t_1+1 < \ldots \le t_2}{\arg\max} \sum_{t=t_1}^{t_2} I_t\left(o_{(t)} - o_{(t-1)}\right) < L-1 \ and \ o_{(t_2)} > (N - 2L + 3) \end{cases} \tag{1.4}$$

where

- $\left(o_{(1)},...,o_{(n)}\right)$ represents a column vector consisting of the ordered index values of column vectors in the trajectory matrix, which were identified as outlying by some method;
- t_1 and t_2 (where $t_1 < t_2$) are the first and last index values of consecutive column vectors identified as outliers;
- $I_t\left(o_{(t)} - o_{(t-1)}\right) = \begin{cases} 1 & if \quad o_{(t)} - o_{(t-1)} = 0 \\ 0 & otherwise \end{cases}$.

Since the column vectors of trajectory matrix in (1.1) places the time series into a multivariate setting, this paper proposes that methods which identify multivariate outliers can be combined with (1.4) to identify a single additive time series outlier.

2. Outlier maps and Robust Principal Component Analysis (ROBPCA)

Outlier maps were introduced by Hubert et al. (2005) to assist as a diagnostic plot in identifying multivariate outliers in Principal Component Analysis (PCA). The method has not been used in the SSA context to date. Three types of multivariate outliers exist according to Hubert et al. (2005), viz. good leverage points (points 5 & 6 in figure

below), orthogonal outliers (points 3 & 4 in figure below) and bad leverage points (points 1 & 2 in figure below).

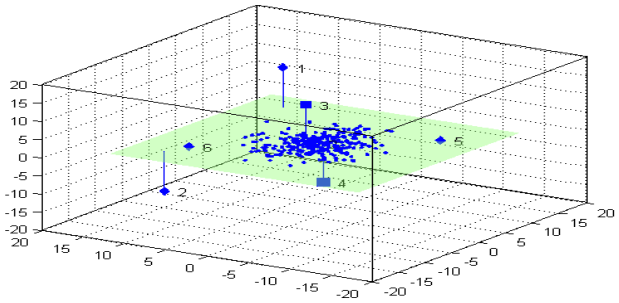

Figure 1. PCA outlier types

Outlier maps constitute a plot of individual p-dimensional multivariate observations' orthogonal distance (OD_i) against score distance (SD_i). Calculations of these measures can be consulted in Hubert et al. (2005). In classic multivariate analysis the row vectors of a data matrix are used to calculate OD_i and SD_i. In the SSA context, however, the column vectors of the trajectory matrix represent observations and are used in the calculations. In this study the classic mean vector and PCA loadings (CPCA) were used to calculate these measures as well as robust PCA (ROBPCA) versions using the LIBRA (Library for Robust Analysis) package in MATLAB that was developed by Verboven and Hubert (2005). The ROBPCA method employs projection pursuit combined with estimation of robust covariance matrices.

Multivariate outliers are identified by employing the outlier map using cut-off values for OD_i and SD_i. Interested readers can consult Hubert et al. (2005) regarding details of the cut-of limits and theoretical justifications. In this study we also used the cut-off limits as a first iteration to identify multivariate outliers. Practical applications yielded many situations where this alone did not always successfully identify outliers in the SSA context. As a second iteration we then applied a quasi convex hull peeling (CVHP) technique. In this approach we identified outlying column vectors in the trajectory matrix using the cut-off limits proposed by Hubert et al. (2005). We then removed these observations from the trajectory matrix and used MATLAB routines to identify the convex hull based on coordinates (OD_i, SD_i) of remaining column vectors. Column vectors for which multivariate observations with OD_i on the convex hull greater than the 80-th percentile of OD_i, or SD_i on the convex hull greater than the 80-th percentile of SD_i were then flagged as outliers in addition to those already identified using cut-off limits. The latter CVHP was applied twice. This alternative approach to devise cut-off limits for the outlier maps was applied to outlier maps generated using the CPCA and ROBPCA approaches. Results are reported in the ensuing section and a practical example of an outlier map with CVHP is illustrated in the final section of this paper. Other methods that were attempted included k-means cluster analysis applied to the outlier maps, instead of applying cut-off limits to OD_i and SD_i. Simulation studies, however, indicated that this approach only performed well when the additive time series outlier was very large and in many cases led to false positive outlier identification.

3. Monte Carlo simulation studies

Monte Carlo simulations were performed as part of this study. A rank $r = 6$ time series of the form $f_t = (300 + 1.98t) + 100(1 - 0.12(\sin(2\pi t/12) + 1.17(\sin(2\pi t/6)))) + \varepsilon_t$ for $t = 1,...,144$ was simulated 200 times with $\varepsilon_t \sim Uniform[-a, a]$. The latter choice of noise

made it possible to control noise-to-signal ratios better. An additive outlier $f_t = f_t + \delta_t$ where $\delta_t = 75$ was added to each of the time series observation. The percentage time that the outlier was correctly identified for $t = 1,...,144$ during the 200 Monte Carlo simulations are summarised in **Figures 2** to **4**, below. The accuracy of outlier identification was tested for choices of window length in the range $L \in [7, 28]$. Four different methods were compared for their effectiveness in identifying outliers, viz. Classic PCA (CPCA), Classic PCA combined with convex hull peeling (CPCA+CVHP), Robust PCA (ROBPCA) and Robust PCA combined with convex hull peeling (ROBPCA+CVHP).

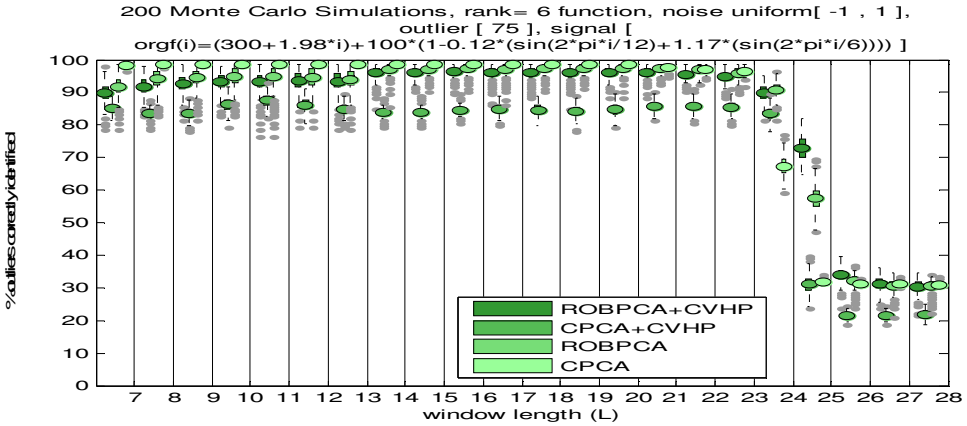

Figure 2. Monte Carlo simulation results (Uniform[-1,1], $\delta_t = 75$)

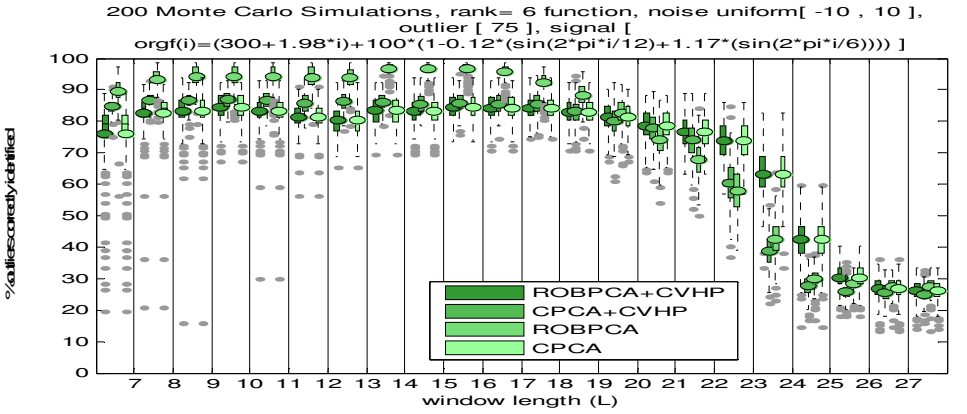

Figure 3. Monte Carlo simulation results (Uniform[-10,10], $\delta_t = 75$)

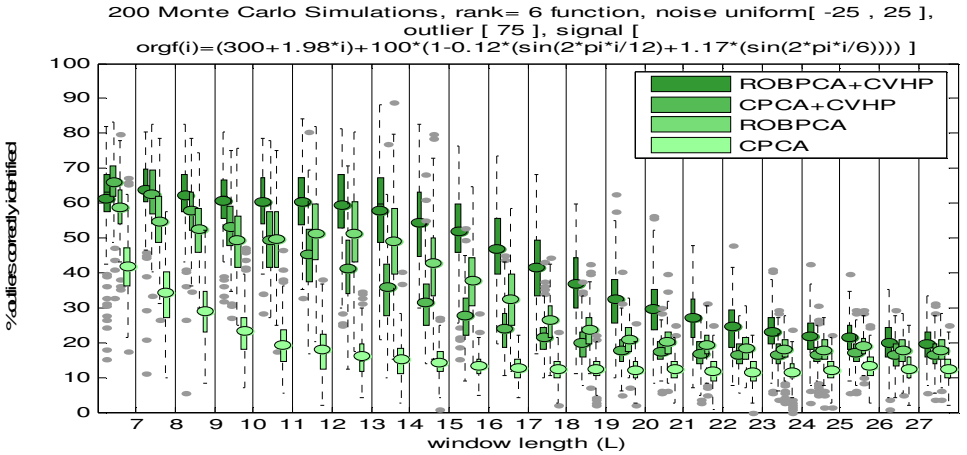

Figure 4. Monte Carlo simulation results (Uniform[-25,25], $\delta_t = 75$)

It is clear from the simulated results that outlier detection for practically noise-free signals are preferring the CPCA and ROBPCA for lower window lengths. As noise levels increase, the ROBPCA combined with convex hull peeling followed by ROBPCA methods reign supreme.

4. Practical application

Tsay (1988) identified an additive outlier at time t=14 in the log-transformed version of the well-known airline passenger time series. **Figure 5**, below, depicts the original time series and log-transformed time series. The additive outlier is evident in the log-transformed series at t=14.

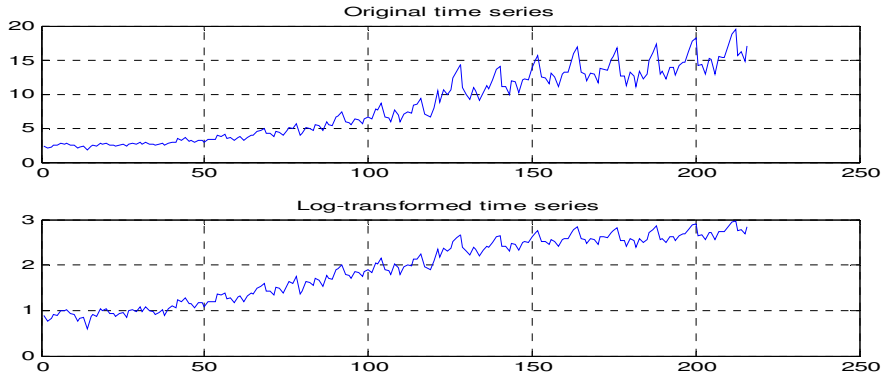

Figure 5. Airline passenger time series

Table 1, below, summarizes results from outlier position identification using (a) ROBUST PCA combined with convex hull peeling as proposed in this paper and (b) ROBPCA when using SSA. Both the latter outlier identification methods were applied to time series lengths in the range $N \in [50,...,69]$, window length (L) in the range $L \in [4,...,9]$ using the leading $r = 3$ eigenvectors. The MATLAB version of the ROBPCA routines developed by Verboven and Hubert (2005) was employed to resist 90% of outliers during multivariate outlier detection. It is clear from the results that the ROBPCA method combined with convex hull peeling outperforms the ROBPCA technique in identifying the additive outlier present in this time series for this particular time series. The time series lengths were purposefully chosen to avoid the later section in the times series that exhibits a level change and changing variation.

Table 1. Comparison of outlier detection techniques

(a) Outlier identification using ROBPCA combined with convex hull peeling

		N (time series length)																			
		50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
L	4	*	*	*	*	14	14	14	14	14	14	*	14	14	14	14	14	14	14	14	14
	5	14	14	14	14	14	14	14	14	14	14	14	14	14	14	*	*	*	14	14	14
	6	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
	7	14	14	*	14	*	14	14	14	14	14	14	14	14	14	14	14	*	14	14	14
	8	*	*	*	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
	9	14	8	*	14	14	*	14	14	14	14	*	*	14	14	14	14	14	14	14	*

(b) Outlier identification using ROBPCA

		N (time series length)																			
		50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
L	4	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
	5	14	14	14	14	*	*	14	14	14	14	14	14	14	14	14	14	14	14	14	14
	6	14	*	14	*	*	*	*	14	14	14	*	14	14	14	14	*	*	*	14	14
	7	*	14	14	*	*	14	*	14	*	14	14	14	14	*	*	*	14	14	14	*
	8	*	*	*	*	*	14	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	9	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	7	*	8	*	*

Notes: An asterisk indicates that no additive outlier was identified by the method

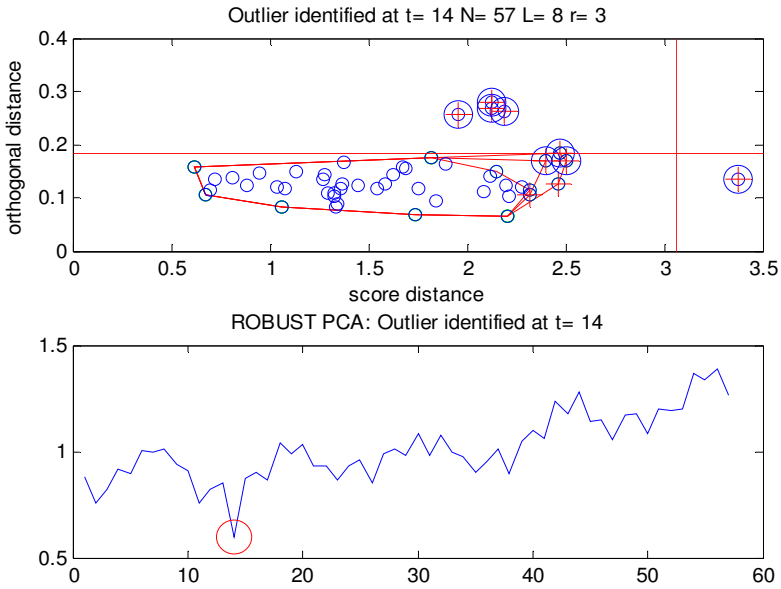

Figure 6. Example of outlier map (ROBPCA combined with convex hull peeling)

Figure 6, above, is an example of an outlier map that was constructed using the first $N = 57$ log-transformed airline passenger time series observations, window length $L = 8$ and using the leading $r = 3$ eigenvectors. Double circled crosshairs indicate consecutive column vectors in the trajectory matrix that were identified as multivariate outliers. The cut-off limits, used for outlier detection as proposed by Hubert et al. (2005), are indicated by the vertical and horizontal lines in the plot. It is clear from the above depiction how the addition of convex hull peeling identifies additional columns as outlying and correctly identifies, using (1.4), the position of the outlier at t=14.

5. Conclusions

This paper proposed the use of robust principal component analysis (ROBPCA) techniques with the aid of outlier maps to identify outliers in the SSA context. It was clear from Monte Carlo simulation results that both the ROBPCA and ROBPCA combined with a convex hull peeling approaches provided promising results.

References

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

Buchstaber, V.M. (1994) "Time Series Analysis and Grassmannians", *Amer. Math. Soc. Transl.*, 162, 1-17.

Golyandina, N., Nekrutkin, V., Zhigljavsky, A. (2001) "Analysis of Time Series Structure – SSA and Related Techniques", *Monographs on Statistics and applied Probability*, 90, Chapman and Hall/CRC.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K. (2005). "ROBPCA: A new approach to robust principal component analysis", *Technometrics*, 47, 64-79.

Takens, F. (1981). "Detecting strange attractors in turbulence", *Lecture notes in Mathematics*, 898, 366-381.

Tsay, R.S. (1988). "Outliers, Level Shifts, and Variance Changes in Time Series", *Journal of Forecasting*, 7, 1-20.

Verboven S., and Hubert, M. (2005). "LIBRA: a MATLAB library for robust analysis", *Chemometrics and Intelligent Laboratory Systems*, 75**,** 127-136.