TIME SERIES OUTLIER DETECTION USING THE TRAJECTORY MATRIX IN SINGULAR SPECTRUM ANALYSIS WITH OUTLIER MAPS AND ROBPCA

J. de Klerk [1]
North-West University, Potchefstroom, South Africa
e-mail: *jacques.deklerk@nwu.ac.za*

Key words: Convex hull peeling, Hankel matrix, outlier maps, robust principal component analysis, singular spectrum analysis.

Summary: Singular spectrum analysis is a powerful non-parametric time series method that applies singular value decomposition to a Hankel structured matrix. The method can handle complex time series structures that include combinations of polynomials, sinusoids and exponentials. Outlier maps combined with robust principal component analysis is considered and shown to compare very favourably with existing time series methods to identify an additive time series outlier. The well-known airline time series as well as a South African tourism time series are used to illustrate the usefulness of the methodology.

1. Introduction

Singular Spectrum Analysis (SSA) is a powerful non-parametric time series method which originated in the field of Physics (Takens, 1981; Broomhead and King, 1986). A thorough introduction to SSA theory and methods can be consulted in Golyandina, Nekrutkin and Zhigljavsky (2001). More recent advances in the field of SSA are considered in Golyandina and Zhigljavsky (2013).

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

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

[1] Corresponding author.

AMS: 62H99, 62M10, 62M15, 62P99

It is clear that anti-diagonal elements of secondary diagonals in this matrix are equal and is also strategic to time series outlier detection in the SSA context. 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 called the *window length* (embedding dimension in Physics literature) and restricted by the choice $2 \leq L \leq floor\left[(N+1)/2\right]$. Additive outliers can unduly influence forecasts, if they form part of the vectors employed by recurrent- or vector forecasting techniques described in Golyandina et al. (2001). Not only can outliers cause problems when forecasting, but their presence can cause bias in any form of model based bootstrap method employed in the SSA context. Employing bootstrap methods for forecast confidence intervals comes to mind in this regard. It makes sense that any outlier identification method in the SSA context can assist to guard against these possible issues and motivates methods developed as part of this research.

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

$$y(t) = \sum_{k=1}^{K} p_k(t) \exp\left(\alpha_k t\right) \sin\left(2\pi\varpi_k t + \phi_k\right) \tag{2}$$

where $p_k(t)$ indicate polynomials.

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

$$y_{t+r} = \sum_{k=1}^{r} a_k y_{t+r-k},\ 1 \leq t \leq N-r \tag{3}$$

where $a_1,...,a_r$ are coefficients and r is the rank (structure) of the time series. SSA can evidently handle a wide variety of time series structure which can include trend with/without seasonality. According to Golyandina et al. (2001) the column vectors of the trajectory matrix all lie on a single linear subspace of rank r if a noise-free signal series of rank r governed by an LRF, as described above, is observed. Noise contaminated time series causes the column vectors of the trajectory matrix not to lie on a single linear subspace and methods are proposed by Golyandina et al. (2001) to reconstruct an approximate signal series. This is where SVD (as part of Basic SSA) or PCA (as part of centred SSA) can be applied to extract singular- or eingenvectors from the noise contaminated Hankel structured trajectory matrix to produce a signal series and forecasts.

According to basic SSA methodology (Golyandina et al., 2001) the Hankel structured trajectory matrix can be reproduced by adding rank 1 elementary matrices, i.e. $\mathbf{X}_{L\times(N-L+1)} = (\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_r)$. The latter is actually the spectral decomposition of a matrix, i.e. $\mathbf{X}_{L\times(N-L+1)} = \sum_{i=1}^{r} \mathbf{X}_i = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^T$, where SSA derives the nomenclature from. Note that any given elementary matrix can be reconstructed using $\mathbf{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^T$ for $i = 1,...,r$. Here r denotes the rank of the trajectory matrix and in effect the time series. Eigenvalues are extracted from the non-centered square matrix $\mathbf{X}\mathbf{X}^T$, which results in eigentriples $(\lambda_i, \mathbf{u}_i, \mathbf{v}_i)$ for $i = 1,...,L$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L \geq 0$ are the ordered (in decreasing order of magnitude) eigenvalues and $(\mathbf{u}_i, \mathbf{v}_i)$ the orthonormal singular vectors. Note that $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$ since the singular vectors are mutually orthogonal and the same holds true that $\mathbf{v}_i^T\mathbf{v}_j = 0$ for $i \neq j$. 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 positioned at time $t = t^*(1 \leq t^* \leq N)$ will be present in a number of consecutive column vectors $(1 \leq ... \leq n^* \ll n = N - L + 1)$ of the trajectory matrix. It is not difficult to show that, once consecutive column vectors in the Hankel structured trajectory matrix have been identified as outlying by some multivariate statistical method that the position of the additive time series outlier will be given by

$$t^* = \begin{cases} o_{(i_1-1)} + L - 1 & if \quad \underset{1\leq i_1<i_2\leq n^*}{\text{argmax}} \sum_{i=i_1}^{i_2} I_i(o_{(i)} - o_{(i-1)}) = L - 1 \\ o_{(i_2)} & if \quad \underset{1\leq i_1<i_2\leq n^*}{\text{argmax}} \sum_{i=i_1}^{i_2} I_i(o_{(i)} - o_{(i-1)}) < L - 1 \; and \; 2 \leq o_{(i_2)} < L \\ o_{(i_1-1)} + L - 1 & if \quad \underset{1\leq i_1<i_2\leq n^*}{\text{argmax}} \sum_{i=i_1}^{i_2} I_i(o_{(i)} - o_{(i-1)}) < L - 1 \; and \; o_{(i_2)} > (N - 2L + 3) \end{cases} \tag{4}$$

where

- n^*, such that $(1 \leq ... \leq n^* \ll n = N - L + 1)$, is the number of column vectors identified by some statistical method as outlying;
- $\mathbf{o}_{n^*\times 1} = (o_{(1)}, ..., o_{(n^*)})^T$ is a column vector with elements consisting of the ordered index values of column vectors in the trajectory matrix, which were identified as outlying by some statistical method;
- $I_i(o_{(i)} - o_{(i-1)}) = \begin{cases} 1 & if \quad o_{(i)} - o_{(i-1)} = 1 \\ 0 & otherwise; \end{cases}$
- i_1 and i_2 (where $1 \leq i_1 < i_2 \leq n^*$) are the first and last index positions in $\mathbf{o}_{n^*\times 1}$ for which the sum over the above indicator function is maximized. The purpose is to locate the maximum number of consecutive column vectors identified as outlying.

Since the column vectors of the trajectory matrix in (1) places a time series governed by an LRF into a multivariate setting, this paper proposes that multivariate statistical methods which identify multivariate outliers can be combined with (4) to identify a single additive time series outlier when applying SSA methodology. It must be noted that methods have been devised for change-point detection in SSA (Moskvina and Zhigljavsky, 2003), but that the methodology described here is solely responsible for the detection of an additive time series outlier.

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

Outlier maps were introduced by Hubert, Rousseeuw and Vanden Branden (2005) as a diagnostic plot assisting multivariate outlier identification in Principal Component Analysis (PCA). The method has not been applied in the SSA context to date, even though PCA can be applied in SSA to extract time series structures. Three types of multivariate outliers can be identified in multivariate datasets in the PCA context according to Hubert et al. (2005), viz. good leverage points (points 5 and 6, cf. Figure 1), orthogonal outliers (points 3 and 4, cf. Figure 1) and bad leverage points (points 1 and 2, cf. Figure 1). Here we suppose that multivariate observations can be arranged in a matrix $\mathbf{X}_{nxp}$, with n denoting the number of observations and p the number of variables measured on each observation. In Figure 1, below, the plane represents a two-dimensional PCA subspace and scatter points the observed multivariate data scatter.

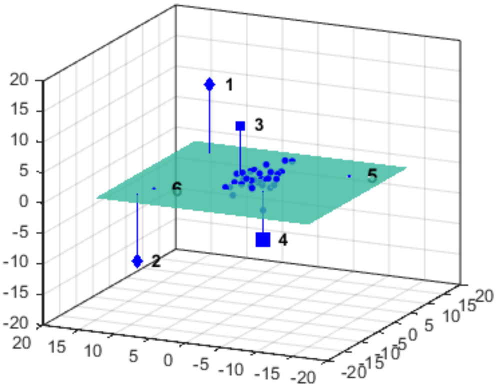

Figure 1: PCA outlier types.

Effectively, an outlier map (Hubert et al., 2005) constitutes a two-dimensional plot of the orthogonal distance (OD_i) against score distance (SD_i), which is calculated for each individual p-dimensional multivariate observation. When applying classic PCA (CPCA), these measures are calculated using

$$OD_i = \left\| \mathbf{x}_i - \underline{\hat{\mu}}_x - \mathbf{P}_{p,k}\mathbf{P}^T_{p,k}(\mathbf{x}_i - \underline{\hat{\mu}}_x) \right\| \tag{5}$$

and

$$SD_i = \sqrt{(\mathbf{x}_i - \underline{\hat{\mu}}_x)^T \mathbf{P}_{p,k}\mathbf{L}^{-1}_{k,k}\mathbf{P}^T_{p,k}(\mathbf{x}_i - \underline{\hat{\mu}}_x)} \tag{6}$$

where $\underline{\mu}_x$ represents the estimated mean vector of the multivariate dataset, $\mathbf{P}_{p,k}$ a matrix consisting of the leading k PCA loadings and $\mathbf{L}_{k\times k}$ a diagonal matrix with elements the eigenvalues based on the covariance matrix of the multivariate dataset. It is clear that the score distance is a typical statistical distance measure as described in Johnson and Wichern (2007). It is clear in (6) that $(\mathbf{x}_i - \underline{\hat{\mu}}_x)^T\mathbf{P}_{p,k}$ is the i-th principal component and that the diagonal elements of $\mathbf{L}_{k\times k}$ are the respective variances of the principal components. An example of an outlier map constructed using the multivariate data scatter with $p = 3$ and $k = 2$ in Figure 1, is illustrated in Figure2, below. The

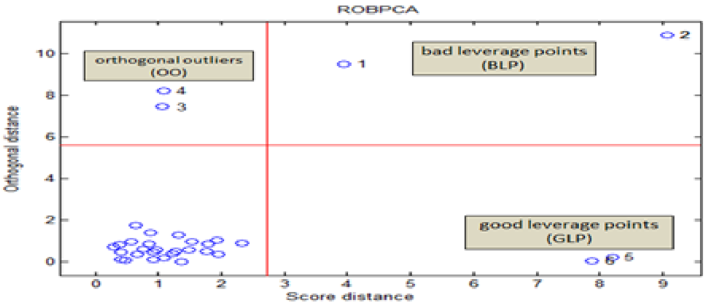

Figure 2: Example of an outlier map.

horizontal axis in Figure 2, above, indicates that the score distance was calculated using 2 leading vectors or PCs.

Hubert et al. (2005) suggest two cut-off limits, each respectively for the orthogonal- and score distances to assist in identifying multivariate outliers in the PCA context. The cut-off limit $c_{SD} = \sqrt{\chi^2_{k;0.975}}$, which is the square root of the 97.5th percentile of the chi-square distribution, is applied to calculated score distances. In case of the orthogonal distances, the cut-off limit $c_{OD} = (\hat{\mu}_{MCD} + \hat{\sigma}_{MCD} z_{0.975})^{3/2}$ is applied. In the latter case $\hat{\mu}_{MCD}$ and $\sigma_{\hat{MCD}}$ respectively represents the minimum covariance determinant (MCD) estimator of the univariate score distance location and spread, and $z_{0.975}$ denotes the 97.5th percentile of a Gaussian distribution. Figure 2, above, illustrates in which quadrant of the outlier map each of the mentioned three types of outliers are identified.

In order to apply the notion of an outlier map in the SSA context, slight changes are required in the above formulae's notation. In the SSA context the column vectors of the trajectory matrix represent multivariate observations and hence $n = N - L + 1$ and $p = L$. The number of leading PCA loadings to use will be either $k = r$ or $k = r - 1$, depending on the time series structure. Applying PCA in the SSA context (referred to as centred SSA in Golyandina et al., 2001), instead of SVD to the trajectory matrix (referred to as Basic SSA in Golyandina et al. (2001)), affects the choice of the leading eigenvectors. The interested reader is referred to Golyandina et al. (2001) to gain clarity in this regard. From this point on forward all reference to outlier maps will be in the context of the trajectory matrix, formed as part of SSA methodology.

Verboven and Hubert (2005) developed a software package in MATLAB, called LIBRA (LiBrary for Robust Analysis), which can readily be used to construct outlier maps. The package can be downloaded from the website `https://wis.kuleuven.be/stat/robust/LIBRA`. The package makes distinction in different approaches used to calculate the orthogonal- and score distances as formulated in (5) and (6). When applying Classic PCA (CPCA) the measures used in calculating the aforementioned formulae uses the usual non-robust PCA extracted from the centered trajectory matrix. The LIBRA package can also apply Robust Principal Component Analysis (ROBPCA) as described in Hubert et al. (2005) and Hubert, Rousseeuw and Van Aelst (2008). The ROBPCA method employs projection pursuit combined with estimation of robust covariance matrices. The interested reader can consult the relevant literature in this regard to gain insight into the methodology applied. In case of ROBPCA the robust center $(\underline{\hat{\mu}}_x)$ of the trajectory matrix, MCD loadings in $\mathbf{L}_{k\times k}$

and robust scatter measures are employed. The ROBPCA routine in the LIBRA package also allows the user to specify the fraction of outliers to resist as part of the robust calculations. Throughout this research the selection was set at 90 percent.

The motivation for employing outlier maps in the SSA context stems from the fact that PCA can also be used to extract time series structure of an approximating signal series governed by an LRF of rank r from an observed noise contaminate time series. Another motivation stems from the fact that the column vectors of a noise-free time series governed by an LRF of known rank lies on the same subspace, which can be found by applying centred SSA (Golyandina et al., 2001). Hence, instead of focusing on the time series itself and additive outlier at a single position, the focus may be shifted to the Hankel structured matrix. By extracting the principal components from the Hankel structured matrix, as applied by the outlier map methodology, we are in fact extracting the 'structure' of the time series in the form of principal components and we can also form the projection matrix of the PCA subspace. By projecting the column vectors of the Hankel structured trajectory matrix onto a linear subspace of chosen rank (k) and taking the necessary distance norm from this approximating subspace, we are in fact taking the signal structure into account when calculating the distance of the trajectory matrix column vectors to the subspace governed by rank k (i.e. the signal structure). This results in using the orthogonal distance to this approximating subspace and coincidentally is measured along the vertical axes of the outlier map. It is the Hankel structure of the trajectory matrix combined with SVD or PCA that makes SSA an appealing method to use. The use of the multivariate distance component of the outlier map, coined the score distance, is perhaps more difficult to justify in the SSA context. Literature in the SSA context does not exist with regards to the distributional properties of the column vectors in the trajectory matrix. It is, however, a well-known fact that PCA is a distribution-free method. Hence, the statistical distance employed in the outlier maps uses the notion of a statistical distance in the context of the principal components. It can only be noted here that empirical evidence in the form of Monte Carlo simulations, which are presented as part of this study, using outlier maps indicated that it worked very well in identifying additive time series outliers in simulated time series governed by an LRF of known rank. The outlier map also makes it possible to test whether column vectors in the trajectory matrix are possibly one of three outlier types identified, in terms of the PCA subspace, or close to the PCA subspace. In the ensuing section we consider an example of a real-life time series where SSA is combined with the notion of an outlier map and it will be clear that the method actually works very well.

2.1. Example of outlier maps combined with ROBPCA

In this section the usefulness of outlier maps in the SSA context is illustrated. The well-known time series consisting of air-passenger miles flown by passengers within the USA between the periods January 1960 and December 1977, as sourced from Cryer (1986), is entertained. The time series is illustrated in Figure 3, below.

The airline time series exhibits a number of interesting structural issues, which include increasing variance and structural changes over time. Tsay (1988) identified an additive outlier in the log-transformed time series at time $t = 14$ when a SARIMA model was fitted. We will, for the purpose of illustration, only consider the log-transformed time series over the period January 1960 up to February 1964 as illustrated in Figure 4, below.

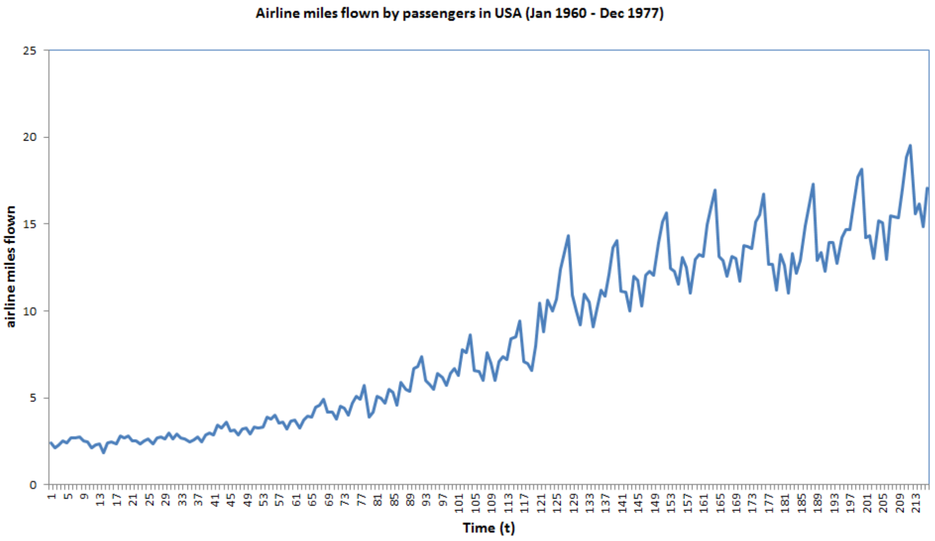

Figure 3: Airline passenger miles flown in USA (January 1960 to December 1977).

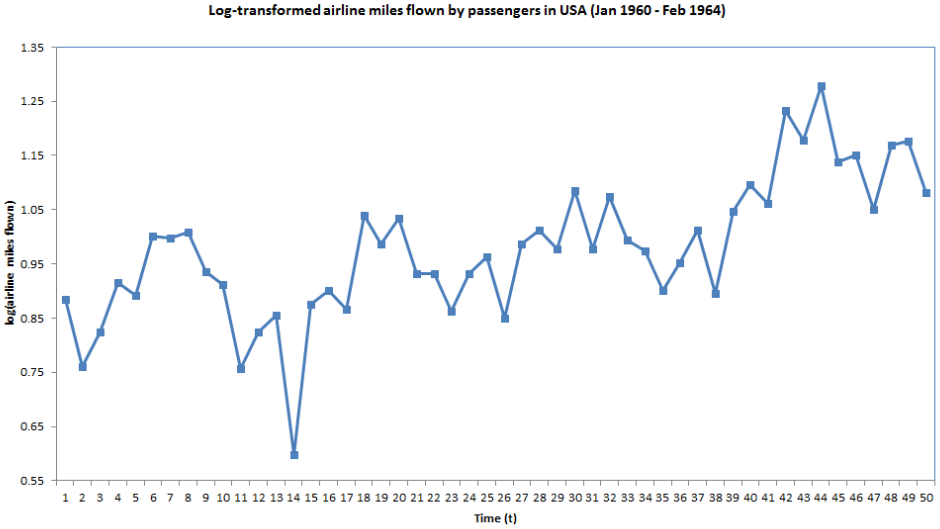

Figure 4: Log-transformed airline passenger miles flown in USA (January 1960 to February 1964).

ROBPCA was employed in constructing the outlier map in Figure 5, based on the trajectory matrix formed using the log-transformed time series. The window length was set at $L = 5$ and the leading $k = 4$ PCAs were used. An additive time series outlier was identified at position $t = 14$. Double encircled cross hairs in the outlier map (cf. Figure 5) represent consecutive column vectors in the trajectory matrix identified as outlying. Single circled cross hairs indicate other column vectors

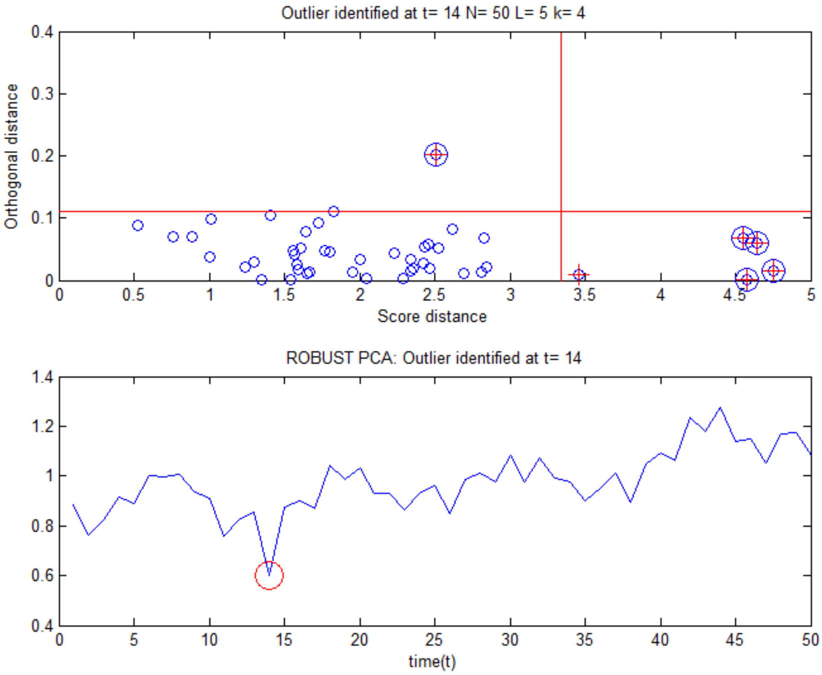

Figure 5: Outlier map (top graph using ROBPCA) and position of identified additive time series outlier (lower graph).

in the trajectory matrix which were also identified as outlying.

Solely studying orthogonal distances as outlier identification method in SSA has its drawbacks, as is clearly indicated by the outlier map in Figure 5. Only a single outlying column vector in the trajectory matrix would be identified and hence the time series outlier would not be identified. According to the outlier map there are 6 outliers of which 5 are classified as good leverage points (GLP). Additional identification of these outlying column vectors makes it possible to correctly identify the additive time series outlier using the outlier map.

Table 1, below, illustrates how the algorithm proposed in (4) identified the position of the outlier, after the outlier map identified column vectors $(10,11,12,13,14,38)$ as outlying. Note that the choice $i_1 = 2$ and $i_2 = 5$ maximized the sum over the indicator function, i.e. $\sum_{i=2}^{5} I_i(o_{(i)} - o_{(i-1)}) = 4$, and the time series outlier was identified at $t^* = o_{(i_1-1)} + L - 1 = 10 + 5 - 1 = 14$.

Table 1: Example applying outlier identification algorithm.

i	**1**	**2** (i_1)	**3**	**4**	**5** (i_2)	**6**
$\mathbf{o_{(i)}}$ **and outlier type**	10 (GLP)	11 (GLP)	12 (OO)	13 (GLP)	14 (GLP)	38 (GLP)
$\mathbf{I_i(o_{(i)} - o_{(i-1)})}$	0	1	1	1	1	0

2.2. Example of outlier map combined with ROBPCA and convex hull peeling

It was experienced that the cut-off limits applied to outlier maps (Hubert et al., 2005) not always identified all the consecutive outlying column vectors in a trajectory matrix via use of the outlier map methodology, albeit in a small fraction of cases and especially for heavy noise contaminated series. Monte Carlo simulated results in the next section of this article will clearly substantiate the latter observation.

An example of the above mentioned situation is illustrated in Figure 6, below. In this example the first $N = 50$ log-transformed airline time series observations were unfolded into a trajectory matrix using window length $L = 7$ and the leading $k = 4$ PCA factor loadings in the ROBPCA approach to identify outlying column vectors. It is clear from inspection of Figure 6 that the ROBPCA procedure and proposed cut-off limits only identified 6 consecutive outlying column vectors, i.e. double circled cross hair points outside the cut-off limits indicated by the horizontal line (c_{OD}) and vertical line (c_{SD}).

A procedure is proposed here, whereby bivariate convex hull peeling (CVHP) is applied in conjunction with the cut-off limits to identify additional potential outlying column vectors in the trajectory matrix using the outlier map. The convex hulls are clearly indicated in Figure 6, below. It is clear that the approach identified an additional 4 column vectors which had large orthogonal distances and/or large score distances. One of these candidates, indicated by a double circled cross hair on the outer convex hull, was part of consecutive outlying column vectors in the trajectory matrix that would correctly identify the time series outlier at time $t = 14$.

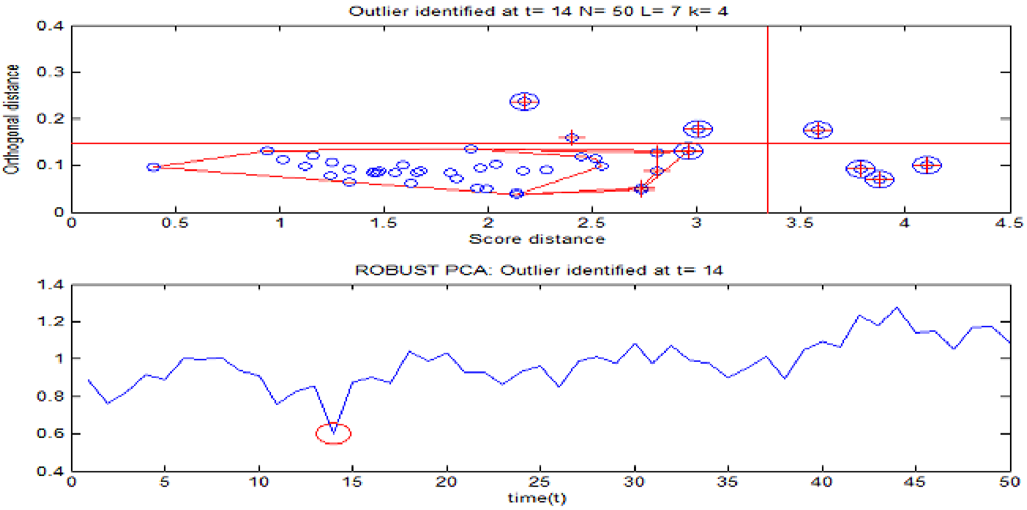

Figure 6: Outlier map (using ROBPCA + CVHP) and position of identified additive time series outlier.

Literature on the application of convex hull theory can be found in numerous published papers in the field of Statistics. Efron (1965) applied convex hulls to the study of random points. Bebbington (1978) considered the use of convex hulls in the trimming of bivariate data as robust estimation of the correlation coefficient. More recent advances that employ convex hull peeling to multivariate datasets can be found in Porzio and Ragozini (2000).

The algorithmic approach of applying bivariate convex hull peeling in conjunction with the limits in the outlier maps, as proposed in this article, can be listed as follows:

STEP 1: Identify outlying column vectors using limits proposed by Hubert et al. (2005) in the outlier map;

STEP 2: Remove the outlying column vectors identified in STEP 1 from the outlier map and construct the bivariate convex hull based on the remainder of points in the outlier map;

STEP 3: Flag column vectors which are on the convex hull where OD_i is greater than the 80-th percentile of the orthogonal distances used to construct the convex hull in this step, or where SD_i is greater than the 80-th percentile of the score distances used to construct the convex hull in this step;

STEP 4: Remove the column vectors identified in STEP 3 from the outlier map and again apply the methodology described in STEP 3. This implies that the depth of data peeling applied were two iterations of the CVHP approach.

Monte Carlo simulations comparing the above algorithm approach will be studied in the ensuing section.

On a note regarding the choice of convex hull peeling, it can be reasoned that another approach would simply involve changing the percentile used in the score distance cut-off limit, e.g. changing $\chi^2_{k;0.975}$ to $\chi^2_{k;0.90}$. This approach would clearly just shift the cut-off limit left towards the mass of points in the outlier map and possibly identify points with a high score distance, but low orthogonal distance as possible outlying column vectors. The same would hold true for changing $z_{0.975}$ in the limit applied to the orthogonal distances (c_{OD}). Even by changing both percentiles used in the limits proposed by Hubert et al. (2005) would merely result in shifting the limits closer to the mass of points in the outlier map and possibly identifying too many candidates that are clearly not outlying. The CVHP approach is proposed here for its simplicity and fact that it concentrates on the most extreme points in the outlier map, which were not also identified as outlying candidates.

Surely other approaches to the use of CVHP in the current context can be utilised and is an open research question. This paper will, however, by use of Monte Carlo simulations in the ensuing section clearly illustrate that CPCA and ROBPCA are very attractive methods for identifying outliers and that the addition of CVHP is a mere possible alternative which compares very favourably in the case of highly noise contaminated simulated time series.

3. Monte Carlo simulations

Monte Carlo simulations were conducted to compare the proposed outlier identification approaches, viz. using classic PCA and outlier maps (denoted by CPCA), using robust PCA and outlier maps (denoted by ROBPCA), using classic PCA combined with the CVHP technique and outlier maps (denoted by CPCA+CVHP) and, finally, using robust PCA and combining CVHP with the outlier maps (denoted by ROBPCA+CVHP).

Two hundred rank $r = 6$ series of the form $y_t = (300 + 1.98t) + 100(1 - 0.12(sin(2\pi t/12) + 1.17(sin(2\pi t/6)))) + \varepsilon_t$, where $\varepsilon_t \sim uniform[-a,a]$ and $a \in [1,10,25]$ were simulated. The latter choice of noise made it possible to control noise-to- signal ratios better. Let the simulated time series be noted by $\{y_t\}_{t,i=1}^{144,200}$. An additive time series outlier of magnitude $\delta_t = 75$ was then introduced to each of the time series observations, i.e. forming $y^*_{t,i} = y_{t,i} + \delta_t$ where $t = 1,...,144$ and $i = 1,...,200$. Each of the proposed outlier detection methods were then applied to the same simulated time series with an additive time series outlier added, in order to perform outlier identification. The percentage of outliers correctly identified in each simulated time series of length $N = 144$ was then calculated. Hence, there would be 200 such accuracy percentages, each per Monte Carlo simulated time series. Outlier identification at each stage was performed using $k = 5$ leading PCA loadings and window lengths in the range $L \in [7,28]$. Adjacent boxplots for the proposed outlier methods, based on the percentage accuracy for the 200 simulated series, with simulated noise choices $a = 1$ and $a = 10$ can be consulted in Figures 7 and 8. Not all the Monte Carlo results are graphically reported here to conserve space.

It is clear from the simulation results that breakdown of the outlier detection methods occur beyond a certain choice of the window length(L) . This is due to the well-known curse of dimensionality plaguing multivariate data when attempting outlier detection. Rousseeuw and Van Zomeren (1990) suggest, as rule of thumb, that there be at least five observations per dimension. Hence, a restriction $(N-L+1)/L > 5$ (or simply $L < floor[(N+1)/6]$) needs to be placed on the choice of window length when performing outlier detection in the SSA context. Since the simulated time series were all of length $N = 144$ this amounts to a restriction of $L < 24$, which is clearly beyond which breakdown of the multivariate outlier detection methods were observed. It was generally found that the accuracy of the proposed outlier detection methods decreased for larger choices of L. Accuracy of the CPCA and ROBPCA methods are fairly similar for time series with not too much noise contamination ($a \in [1,10]$). The ROBPCA method was more accurate as noise contamination in the time series increased. In the case of heavy noise contamination ($a = 25$), the ROBPCA+CVHP method performed best. Finally, even if the structure of the time series was miss-specified and $k = 6$ leading eigenvectors used, the mentioned findings still held true.

A number of valuable conclusions can be reached from the simulation studies. First and foremost, that a suitable choice of window length (L) can be made in the range $k < L < floor[(N+1)/6]$ and preferable be a small choice. Next, that the use of CPCA or ROBPCA be limited to time series with little noise contamination. In case heavy noise contamination being present in a time series, that ROBPCA+CVHP be employed. The combination of CPCA+CVHP did not produce results on par with other methods compared and is not recommended.

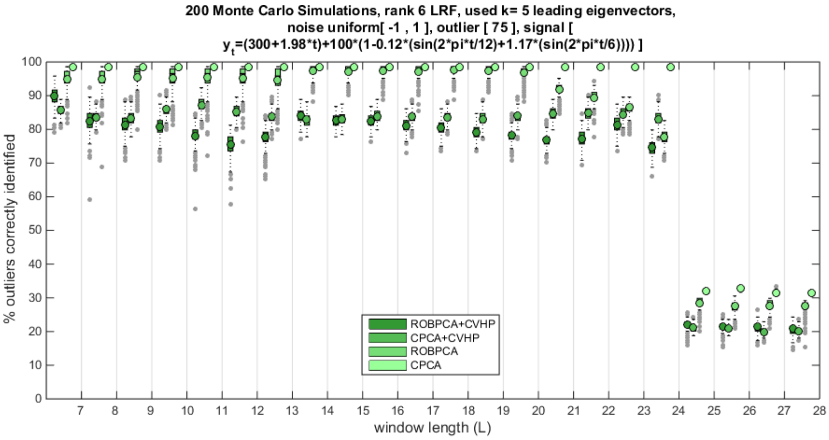

Figure 7: Monte Carlo simulation results $(uniform[-1,1], \delta_t = 75, k = 5)$.

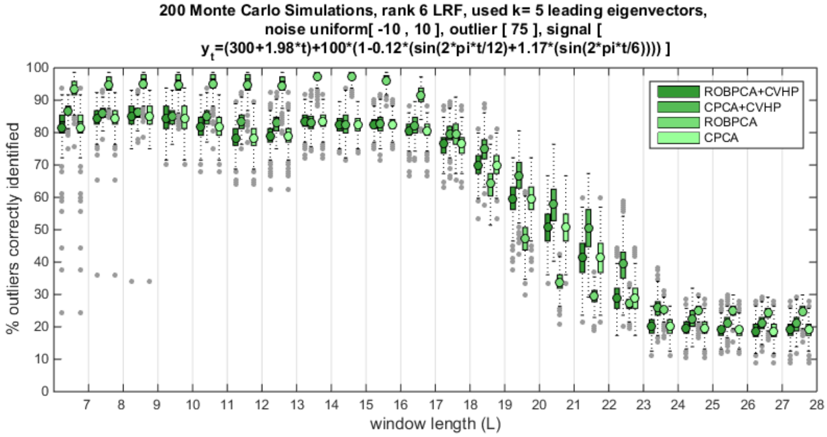

Figure 8: Monte Carlo simulation results $(uniform[-10,10], \delta_t = 75, k = 5)$.

4. Application to South African tourism time series

The proposed outlier identification methodology will now be applied to a South African tourism time series that has been studied by De Klerk (2014). The time series T110444, which was obtained from Statistics South Africa, will be studied for time series outliers prevalent therein. The series consists of the monthly foreign travellers from Switzerland to South Africa. We will only consider the period June 1998 to April 2013. It is evident from a time series plot (cf. Figure 9, below) that the time series exhibits trend, monthly seasonality and a few other harmonic oscillatory patterns. From closer visual inspection of the time series it is evident that outliers are present at positions $t = 3$ and $t = 92$.

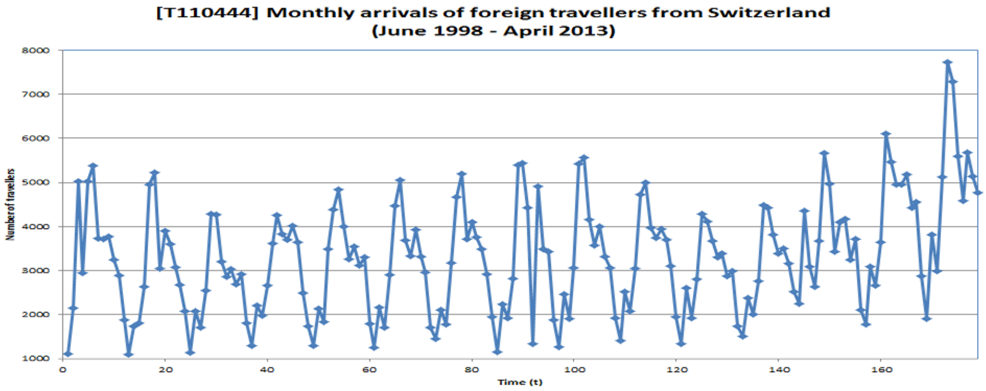

Figure 9: Monthly arrivals of foreign travellers to South Africa from Switzerland [June 1998 to April 2013].

4.1. Outlier identification using SAS/ETS software

The PROC ARIMA procedure in SAS/ETS software was used to both model and identify outliers present in the tourism time series. Only the first $N = 100$ observations were considered for the purpose of time series modeling and outlier identification.

A $SARIMA(0,1,1)_{12}$ (with no intercept) model was fitted to the log-transformed time series. The fitted model was of the form $(1-B^{12})(1-0.57132B^{12})ln(y_t) = z_t$ where B denotes the usual backshift operator. A 12th order difference was applied to the time series due to the monthly seasonal variation. The procedure was instructed to identify a maximum of 2 outliers in the time series.

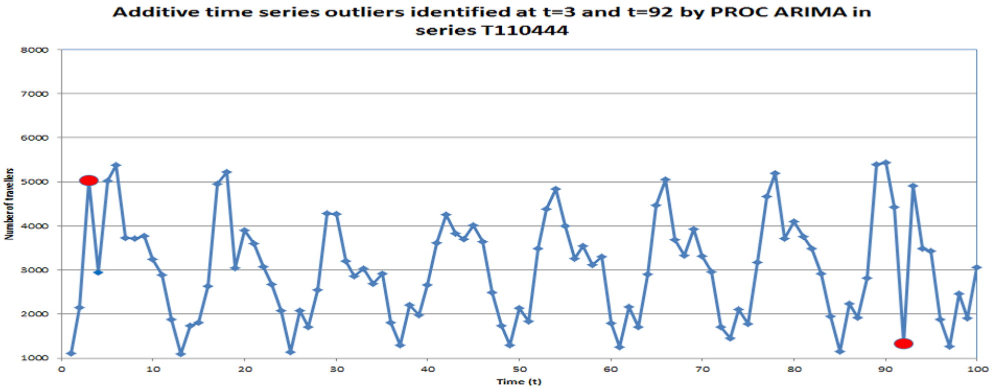

Figure 10: Additive outliers identified by PROC ARIMA in monthly arrivals of foreign travellers to South Africa from Switzerland [June 1998 to April 2013] .

It is clear from Figure10, above, that the ARIMA procedure identified additive time series outliers at positions $t^* = 3$ and $t^* = 92$.

4.2. Outlier identification using SSA combined with outlier maps

De Klerk (2014) illustrated how the leading 12 left singular vectors described the structure of the log-transformed South African tourism time series. As part of the approach followed, SVD was directly applied to the Hankel structured trajectory matrix, which was formed using the log-transformed series. In the current example the $k = 11$ leading PCA loadings were used to perform outlier identification.

In order to compare the accuracy of outlier identification using CPCA, ROBPCA and ROBPCA+CVHP time series lengths in the range $N \in [88, ..., 100]$ and window lengths in the range $L \in [12, 13, 14]$ were employed.

Table 2: Position of outliers identified using CPCA, ROBPCA and ROBPCA+CVHP using $k = 11$.

		CPCA			**ROBPCA**			**ROBPCA+ CVHP**		
		L			**L**			**L**		
		12	**13**	**14**	**12**	**13**	**14**	**12**	**13**	**14**
N	**88**		3	3		3	3	3	3	3
	89		3	3			3	3	3	3
	90		3	3	3	3	3	3	3	3
	91		3	3	3	3	3	3	3	3
	92		3	3		3	3	3	3	3
	93	92	92	3		92	3	3	3	
	94	92	92	92	92	92	92	3	3	3
	95	92	92	92	92	92	92	92	92	92
	96	92	92	92	92	92	92		92	92
	97	92	92	92	92	92	92	92	92	92
	98	92	92	92	92	92	92	92	92	92
	99	92	92	92	92	92	92	92	92	92
	100	92	92	92	92	92	92	92	92	92

Golyandina et al. (2001) actually recommends that a window length as an integer multiple of the major seasonal periodicity, which is 12 in this time series, be used to extract the signal as best possible. The choice of window length (L) is, however, restricted here to the range $k < L < [(N+1)/6]$ due to curse of dimensionality which was illustrated in the Monte Carlo simulations. Hence, the choice of window length was restricted to $12 \leq L \leq 14$. Table 2, above, summarizes the position (t^*) in the time series where the aforementioned outlier identification approaches identified an additive outlier for the particular combination of time series- and window length. Perusing through the table, it is clear that ROBPCA+CVHP identified an additive outlier at when the first $N = 88$ log-transformed time series observations were used combined with the choice $k = 11$ leading eigenvectors and window length $L = 12$. It is clear from the results that the ROBPCA+CVHP approach accurately identified the outlier at $t^* = 3$ for time series lengths in the range $N \in [88, ..., 94]$ when a window length in the range $L \in [12,13]$ was used. The time series outlier present at $t^* = 92$ is much larger in magnitude than the one at $t^* = 3$ and explains why the ROBPCA+CVHP approach identifies $t^* = 92$ for time series lengths in the range $N \in [95,...,100]$. It also seems, for this particular example, that the ROBPCA+CVHP approach was most accurate in identifying an additive time series outlier for different time series- and window lengths.

5. Concluding remarks

Singular spectrum analysis and four methods employing outlier maps were considered in this paper to identify an additive time series outlier. At present the methodology is designed to deal with a single additive time series outlier. The procedures were compared using Monte Carlo simulations and it was evident that CPCA and ROBPCA worked very well in case time series had slight noise contamination. ROBPCA and ROBPCA+CVHP seem to perform best for time series with higher levels of noise contamination. As a rule of thumb the choice of window length is restricted to $k+1 \leq L \leq floor([N+1]/6)$, in which case the upper bound results due to the inherent curse of dimensionality plaguing multivariate outlier detection methods. The method proposed in this paper was also compared to outlier identification methods available in commercial software using much more advanced time series techniques. The methods proposed in this paper compared on par with regards to accuracy in identifying an additive outlier.

Acknowledgements

The author is most grateful for constructive comments by the anonymous reviewer and associate editor that resulted in a much improved version of this paper.

References

BEBBINGTON, A. C. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient. *Journal of the Royal Statistical Society Series C*, **27**, 221–226.

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. *American Mathematical Society Translations*, **162**, 1–17.

CRYER, J. D. (1986). *Time Series Analysis*. Duxbury Press, Boston.

DE KLERK, J. (2014). Identifying South African tourism time series structure using singular spectrum analysis. *South African Statistical Journal*, **48**, 19–39.

EFRON, B. (1965). The convex hull of a random set of points. *Biometrika*, **52**, 331–343.

GOLYANDINA, N., NEKRUTKIN, V., AND ZHIGLJAVSKY, A. (2001). *Analysis of Time Series Structure - SSA and Related Techniques*, volume 90 of *Monographs on Statistics and applied Probability*. Chapman and Hall/CRC.

GOLYANDINA, N. AND ZHIGLJAVSKY, A. (2013). *Singular Spectrum Analysis for Time Series*. Springer, New York.

HUBERT, M., ROUSSEEUW, P. J., AND VAN AELST, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, **23**, 92–119.

HUBERT, M., ROUSSEEUW, P. J., AND VANDEN BRANDEN, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, **47**, 64–79.

JOHNSON, R. A. AND WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis.* Sixth edition. Pearson Prentice Hall.

MOSKVINA, V. AND ZHIGLJAVSKY, A. (2003). An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics*, **32** (2), 319–352.

PORZIO, G. AND RAGOZINI, G. (2000). Peeling multivariate data sets: a new approach. *Quanderni di Statistica*, **2**, 85–99.

ROUSSEEUW, P. J. AND VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633–651.

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.

Manuscript received, 2014-07-11, revised, 2014-10-12, accepted, 2014-12-11.