Tests for contaminated time series

Marcio Valk¹, Aluisio²

¹ Federal University of Rio Grande do Sul, Brazil
² University of Campinas, Brazil

E-mail for correspondence: apinheiro.unicamp@gmail.com

Abstract: The presence of outliers in time series may cause some problems in model specification, parameter estimation and forecasting. We propose a non-parametric algorithm with three main objectives: clustering; testing groupings; and classifying new time series. We employ a robust kernel quasi U-statistic and show that it works well even if some (or all) time series are contaminated by outliers. The set-up is based on models for which the probability of occurrence of outliers may be time-dependent. We motivate the methodology through its theoretical properties. The procedure is then illustrated in a simulation study and by its application in a real data set concerning Heart Rate Variability (HRV).

Keywords: Quasi U-statistics, Outliers, Dissimilarity Measures.

1 Introduction

An Additive Outlier (AO) affects only a specific observation while the influence of an Innovational Outlier (IO) propagates to subsequent observations (Fox, 1972). An extensive literature on outliers in time series is available (Chang et al., 1988; Ljung, 1993; Burridge et al., 2006; Huang et al., 2013). Ma and Genton (2000) address the problem of the robustness of the sample autocovariance function. Recent discussions on outliers in time series can be found in Fajardo et al. (2009), Hotta and Tsay (2011) and Reisen and Molinares (2012). A communality in these works is that the probability of occurrence of one or more outliers is constant in time. We consider model with time-dependent probabilities, which can fit in a realistically way phenomena under various external factors. As an example we analyze the Heart Rate Variability (HRV) (Spang and Dutter, 2007). We study the performance of clustering methods when data is contaminated, using tests which belong to the class of quasi U-statistics (Pinheiro et al., 2011; Valk and Pinheiro, 2012).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
2 Tests for contaminated time series

2 The Test Statistic

The model is defined by

\[ Y_t = Z_t + \sum_{j=1}^{m} \omega_j X_{jt}, \quad (1) \]

where \( X_{jt} \) take values in \( \{0, 1, -1\} \). We define \( P(X_{jt} = 1) = P(X_{jt} = -1) = p_{jt}/2 \) and \( P(X_{jt} = 0) = 1 - p_{jt} \), for all \( t = 1, \ldots, T \) and \( j = 1, \ldots, m \). Figure 1 presents six configurations for the vector of probabilities \( \mathbf{p} \) which are successful in modeling a wide range of data (Hotta and Tsay, 2011).

![Figure 1. Time dependent probabilities for outliers’s occurrence](image)

Fajardo et al. (2009) proposes the robust estimator of the periodogram:

\[ RSDE(\omega_j) = \frac{1}{2\pi} \sum_{h=-(T-1)}^{T-1} \hat{\gamma}_R(h) \cos(h\omega_j). \quad (2) \]

where \( \hat{\gamma}_R(h) = [Q_{n-h}^2(u + v) + Q_{n-h}^2(u - v)]/4 \), for vectors \( u \) and \( v \) of the first \( n - h \) and last \( n - h \) observations, respectively (Ma and Genton, 2000). \( Q_n(\cdot) \) is the \( \kappa \)th order statistic of \( \binom{n}{2} \) distances \( \{|Z_i - Z_j|, i < j\} \), i.e., \( Q_n(Z) = c \times \{|Z_i - Z_j|, i < j\}(\tau), \) for \( Z = (Z_1, Z_2, \ldots, Z_n) \) and \( c \) a constant used to guarantee consistency.

We employ two quasi-U-statistics. \( B_n \) is based on the usual periodogram, and \( RB_n \) is based on \( RSDE \). We refer the reader to Pinheiro et al. (2011) for the general properties of quasi-U-statistics, and for Valk and Pinheiro (2012) for the specific properties under a time series set-up. Basically, under a null hypothesis of homogeneity between all pairs of groups, i.e., no real groups do exist, \( B_n \) and \( RB_n \) are centered at 0, \( n\sqrt{T}B_n \) and \( n\sqrt{T}RB_n \) are asymptotically normal, where \( n \) is the sample size and \( T \), the length of the series. Under the alternative hypothesis of heterogeneous groups, \( B_n \) and \( RB_n \) have positive means and \( n\sqrt{T}(B_n - E[B_n]) \) and \( n\sqrt{T}(RB_n - E[RB_n]) \) are asymptotically normal.
3 Simulation and Application to real data sample of ECG

We use the six configurations in Figure 1. Three different amplitudes of the outliers are considered \( w = 0, 3, 10 \), where \( w = 0 \) means no outliers. The length of the time series varies as \( T = 250, 500, 1000 \). Two test statistics are shown here: \( B_n \), based on the usual periodogram; and \( RB_n \), based on the robust spectral density estimator RSDE. The simulations are performed in the software R with 1000 replications. Three underlying error structures are used: \( M_1 \) is a pure error; \( M_2 \) is an AR(1) with \( \phi = 0.5 \); and \( M_3 \) is an ARMA(1,1) with \( \phi = 0.5 \) and \( \theta = -0.8 \). Four series were generated in each group being compared. Table 1 presents the empirical test sizes. An amplitude of \( w = 0 \) means that the time series are not contaminated. The significance level of the test is \( \alpha = 0.05 \). One should note the good empirical test sizes for \( RB_n \) even for \( w = 10 \), and its overall superior performance compared to \( B_n \)’s.

<table>
<thead>
<tr>
<th>Amp. Model</th>
<th>T</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w = 0 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.03</td>
<td>0.34</td>
<td>0.05</td>
<td>0.26</td>
<td>0.06</td>
<td>0.55</td>
<td>0.02</td>
<td>0.19</td>
<td>0.04</td>
<td>0.35</td>
</tr>
<tr>
<td>1000</td>
<td>0.05</td>
<td>1.00</td>
<td>0.06</td>
<td>0.73</td>
<td>0.05</td>
<td>1.00</td>
<td>0.03</td>
<td>0.11</td>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>( w = 3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.02</td>
<td>0.30</td>
<td>0.04</td>
<td>0.50</td>
<td>0.02</td>
<td>0.75</td>
<td>0.04</td>
<td>0.19</td>
<td>0.04</td>
<td>0.75</td>
</tr>
<tr>
<td>1000</td>
<td>0.05</td>
<td>1.00</td>
<td>0.06</td>
<td>1.00</td>
<td>0.05</td>
<td>1.00</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>( w = 10 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.04</td>
<td>0.63</td>
<td>0.07</td>
<td>0.09</td>
<td>0.05</td>
<td>0.73</td>
<td>0.04</td>
<td>0.15</td>
<td>0.05</td>
<td>0.84</td>
</tr>
<tr>
<td>1000</td>
<td>0.06</td>
<td>1.00</td>
<td>0.06</td>
<td>1.00</td>
<td>0.05</td>
<td>1.00</td>
<td>0.06</td>
<td>0.16</td>
<td>0.05</td>
<td>0.98</td>
</tr>
</tbody>
</table>

The ECG data set used here is available at the MIT-BIH Arrhythmia Database (http://www.physionet.org/physiobank/database/mitdb/) and is described in Moody and Mark (2001). It consists of ECG recordings of healthy and unhealthy patients with clinically significant arrhythmias. We focus on the Heart Rate Variability (HRV) which is a continuous beat-by-beat measurement of interbeat intervals. The \{RHRV\} package from software R was used to obtain the HRV time series from the ECG records. Outliers in the HRV can appear by several factors such as activity, emotion, sex, and age. However, in this case, the group of healthy patients is medically homogeneous (Spangl and Dutter, 2007). Using the non robust test \( Bn \) one finds two spurious groups of healthy patients. The robust test provides the correct decision of not separating patients within the homogeneous group.
4 Conclusions

We propose homogeneity tests for groups of time series. The importance of a robust kernel is illustrated by simulation and in a real data time series concerning MIT-BIH Arrhythmia. In both instances, spurious grouping may result from lack of robustness of the test statistic. The test behavior is greatly improved by the robust kernel.

References


