Tests for contaminated time series

Marcio Valk[1], Aluísio[2]

[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).

---

2 The Test Statistic

The model is defined by

$$Y_t = Z_t + \sum_{j=1}^{m} \omega_j X_{jt}, \tag{1}$$

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).

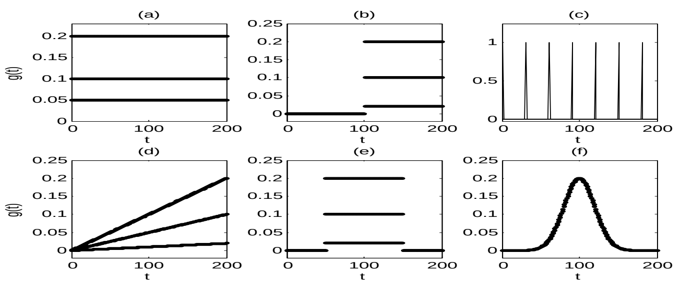

FIGURE 1. Time dependent probabilities for outliers's occurrence

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

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

where $\hat{\gamma}_R(h) = [Q^2_{n-h}(u+v) + Q^2_{n-h}(u-v)]/4$, 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 κth 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 $r$eal 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 $\sqrt{nT}(B_n - E[B_n])$ and $\sqrt{nT}(RB_n - E[RB_n])$ 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 1. Empirical Test Sizes for B_n and RB_n.

			cases for probability of occurence											
			a		b		c		d		e		f	
Amp.	Model	T	RB_n	B_n	RB_n	B_n	RB_n	B_n	RB_n	B_n	RB_n	B_n	RB_n	B_n
	M_1	250	0.04	0.91	0.05	0.26	0.08	0.75	0.02	0.19	0.04	0.35	0.05	0.04
	$\times$	500	0.05	1.00	0.06	0.60	0.06	1.00	0.02	0.58	0.05	0.55	0.05	0.14
	M_1	1000	0.05	1.00	0.05	0.85	0.06	1.00	0.08	0.92	0.04	0.81	0.04	0.24
0	M_2	250	0.07	0.30	0.07	0.09	0.07	0.19	0.06	0.07	0.04	0.05	0.04	0.05
$\times$	$\times$	500	0.04	0.63	0.07	0.09	0.05	0.73	0.04	0.15	0.05	0.11	0.05	0.06
3	M_2	1000	0.06	0.95	0.06	0.21	0.05	1.00	0.05	0.13	0.05	0.16	0.04	0.04
	M_3	250	0.07	0.71	0.04	0.19	0.04	0.43	0.04	0.20	0.03	0.20	0.05	0.09
	$\times$	500	0.04	0.98	0.06	0.41	0.06	0.98	0.02	0.30	0.05	0.23	0.02	0.08
	M_3	1000	0.04	1.00	0.06	0.61	0.04	1.00	0.02	0.68	0.05	0.68	0.08	0.21
	M_1	250	0.07	1.00	0.08	0.97	0.06	1.00	0.05	0.93	0.03	0.97	0.02	0.72
	$\times$	500	0.05	1.00	0.04	1.00	0.07	1.00	0.05	1.00	0.05	1.00	0.07	0.96
	M_1	1000	0.09	1.00	0.06	1.00	0.06	1.00	0.06	1.00	0.05	1.00	0.06	1.00
0	M_2	250	0.10	1.00	0.07	0.93	0.06	1.00	0.04	0.91	0.10	0.98	0.08	0.63
$\times$	$\times$	500	0.10	1.00	0.04	1.00	0.07	1.00	0.05	1.00	0.06	1.00	0.06	0.85
10	M_2	1000	0.07	1.00	0.06	1.00	0.06	1.00	0.06	1.00	0.05	1.00	0.05	1.00
	M_3	250	0.05	1.00	0.04	0.94	0.06	1.00	0.06	0.95	0.04	1.00	0.03	0.75
	$\times$	500	0.04	1.00	0.07	1.00	0.05	1.00	0.08	1.00	0.05	1.00	0.04	0.92
	M_3	1000	0.07	1.00	0.06	1.00	0.05	1.00	0.07	1.00	0.06	1.00	0.05	1.00

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

Burridge, P. and Taylor, A. M. R. (2006). Additive outlier detection via extreme-value theory. *J. Time Ser. Anal.*, **27**(5), 685701.

Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, **30**(2), 193-204.

Fajardo, F., Reisen, V., and Cribari-Neto, F. (2009). Robust estimation in long-memory processes under additive outliers. *J. Statist. Plann. Inference*, **139**(8), 2511-2525.

Fox, A. J. (1972). Outliers in time series. *J. Roy. Statist. Soc. Ser. B*, **34**, 350-363.

Hotta, L. K. and Tsay, R. S. (2011). Outliers in garch processes. In *Economic Time Series: Modeling and Seasonality*, pages 337-358. Chapman and Hall/CRC.

Huang, H., Mehrotra, K., and Mohan, C. K. (2013). Rank-based outlier detection. *J. Stat. Comput. Simul.*, 83(3), 518-531.

Ljung, G. M. (1993). On outlier detection in time series. *J. Roy. Statist. Soc. Ser. B*, **55**(2), 559-567.

Ma, Y. and Genton, M. G. (2000). Highly robust estimation of the autocovariance function. *J. Time Ser. Anal.*, **21**(6), 663-684.

Moody, G. and Mark, R. (2001). The impact of the mit-bih arrhythmia database. *IEEE Eng. Med. Biol.*, **20**(3), 45-50.

Pinheiro, A., Sen, P. K., and Pinheiro, H. P. (2011). A class of asymptotically normal degenerate quasi U-statistics. *Ann. Instit. Statist. Math.*, **63**(6), 1165-1182.

Reisen, V. A. and Fajardo, F. (2012). Robust estimation in time series with long and short memory properties. *Ann. Math. Inform.*, **39**, 207-224.

Spangl, B. and Dutter, R. (2007). Estimating spectral density functions robustly. *REVSTAT*, **5**(1), 41-61.

Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi U-statistics. *J. Time Ser. Anal.*, **33**(4), 608-619.