# Outliers in Time Series: an Empirical Likelihood Approach

## *Dati Anomali in Serie Storiche: un Approccio Basato sulla Verosimiglianza Empirica*

Roberto Baragona and Domenico Cucina

**Abstract** The empirical likelihood method is known to be a flexible and effective approach for testing hypotheses and constructing confidence regions in a non-parametric setting. This framework is adopted here for dealing with the outlier problem in time series where conventional distributional assumptions may be inappropriate in most cases. The procedure is illustrated by a simulation experiment.
**Abstract** *Il metodo basato sulla verosimiglianza empirica è noto essere un approccio flessibile ed efficiente per il controllo di ipotesi e la costruzione di regioni di confidenza in ambito non parametrico. Questo approccio è qui adottato per affrontare il problema dei dati anomali in serie storiche dove le ipotesi sulla distribuzione di probabilità dei dati risultano spesso inappropriate*

**Key words:** empirical likelihood, outliers, time series, confidence regions

## 1 Introduction

Outliers in time series may be defined as those observations that do not conform to the overall behavior of the data sequence. In the independent data framework usually outliers are searched for among either the largest or the smallest observations. In time series framework outliers are to be found instead among observations that show some unexpected departure from predicted value or fail to fit the correlation structure inferred by the majority of the data. Such a result may be produced by outlying observations characterized by different shapes which reflect on time series statistics in some peculiar ways. [11] distinguished outlying observations of four types that may distort linear model parameter estimates, i.e. additive (AO), innovation (IO),

R. Baragona
Lumsa University of Rome, Italy e-mail: r.baragona@lumsa.it

D. Cucina
Sapienza University of Rome, Italy e-mail: domenico.cucina@uniroma1.it

transient (TC) and permanent (LC) level change. Other outlier types which have been considered in the literature are the so called patches , i.e. a sequence of consecutive outlying observations that do not show a steady pattern ([2]), and outliers in generalized autoregressive conditional heteroscedastic (GARCH) models which may impact either levels or volatility or both ([1]). Further extensions refer to outliers in non linear and in vector time series (see, e.g., [5] for a review).

Statistical inference on outliers in time series usually relies on distributional assumptions on some appropriate data generating process. In this paper a distribution-free schema for building confidence regions for parameter estimates and conducting hypothesis testing in the context of time series data possibly affected by outlying observations is considered. The empirical likelihood (EL) methods ([9]) are adopted so that the familiar likelihood ratio statistic may be used which allows the statistical inference to be based essentially on the chi squared distribution. New developments that prove to be necessary in order to handle difficult situations are employed which came to be known as adjusted EL and balanced EL ([4, 6]). Attention is specially directed to outliers of the AO type and outliers which induce a permanent LC. A rather general framework is provided however, that allows several different other types to be handled along very similar guidelines. Some simulation experiments are presented to illustrate the effectiveness of the method in case of moderate sample size.

The plan of the paper is as follows. In Sect. 2 the EL framework is briefly considered. In Sect. 3 inference methods for outliers in time series are developed based on EL methods. In Sect. 4 the behavior of the statistics for inference in small samples is outlined by means of some simulation experiments. Conclusions and possible suggestions for further research are provided in Sect. 5.

## 2 The empirical likelihood

EL methods have been introduced by [7, 8, 9] and have been used afterwards for many applications, including time series analysis. Basically an unknown probability $p_i$ is assigned to each observation in a sample $y = (y_1, y_2, \ldots, y_n)'$ to define an empirical distribution $F$ specified by $(y_i, p_i), i = 1, \ldots, n$. This way the necessity to assume a family of probability distributions on which statistical inference may be based is avoided. The EL is defined instead as $L(F) = \prod_{i=1}^{n} p_i$ under the constraints $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$. The probability distribution $F$ may possibly depend on a parameters set $\theta$ so that one has to consider the maximum of $F(\theta)$ to obtain a well defined probability distribution. Let $\theta$ be e.g. the mean $\theta = \sum_i p_i y_i$, then the largest probability to obtain the observed sample is

$$EL(\theta) = \max_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} p_i \,:\, p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i y_i = \theta \right\}. \tag{1}$$

This is called the EL profile for $\theta$. The maximum is reached at $\theta = \bar{y} = \sum y_i/n$ and equals $n^{-n}$, corresponding to $p_i = 1/n$, $i = 1,\ldots,n$. Thus, $\bar{y}$ is the maximum EL estimator of $\theta$. The EL ratio is obtained dividing (1) by its maximum $n^{-n}$.

The addition of the so called estimating equations ([9, 10]) to the constraint set is a further step that allows complicated models to be estimated and statistical inference for building confidence regions and conducting tests of hypotheses to be based on EL ratio. Let the data $y$ be generated by a model which depends on a parameter vector $\theta$ of length $p$ and assume that $r \geq p$ equations of the type

$$\mathrm{E}\{g(y,\theta)\} = 0, \quad g = (g_1,\ldots,g_r)', \tag{2}$$

exist that uniquely describe the relationships between the data and the model parameters. The functions $g_1,\ldots,g_r$ are called the estimating functions and Equations (2) are called the estimating equations. The EL ratio may be written

$$\mathrm{ELR}(\theta) = \max_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} (np_i) | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(y_i,\theta) = 0 \right\}. \tag{3}$$

If $r = p$ Equations (2) are as many as the number of the unknown parameters. A model for which this circumstance occurs is often called a just identified model. In what follows such assumption will be held satisfied.

Let $\theta_0$ denote in Equation (3) the true parameter vector. Then it may be shown ([9, 10]) that under some regularity conditions, as $n \to \infty$, the statistic $-2\log\{\mathrm{ELR}(\theta)\}$ in $\theta = \theta_0$ is distributed as a $\chi^2$ with $p$ degrees of freedom, in close agreement with the similar property which holds for ordinary parametric likelihood. So even in the absence of any assumption on the probability distribution of the data, confidence regions and tests of hypotheses may be computed all the same.

## 3 Empirical likelihood for inference on outliers in time series

It seems convenient in the present EL context to consider the following general time series model with outliers

$$y_t = f(x_t,\theta) + \varepsilon_t, \tag{4}$$

where $x_t$ summarizes all explanatory variables possibly including one or more dummies which account for outliers occurring at known time instants, and $\varepsilon_t$ is a zero mean random error for which no distributional assumptions are made. The vector parameter $\theta$ assumed of length $p$ includes both model parameters and outlier sizes. The following procedure may be used to inscribe the inference problems related to model in Equation (4) in the EL framework. Let $e_t = y_t - f(x_t,\theta)$. The least squares estimate $\hat{\theta}$ is obtained by solving the normal equations

$$\frac{\partial}{\partial \theta_k} \sum_{t=1}^{n} e_t^2 = 2\sum_{t=1}^{n} (y_t - f(x_t,\theta)) \left\{ -\frac{\partial}{\partial \theta_k} f(x_t,\theta) \right\} = 0, \quad k = 1,\ldots,p. \tag{5}$$

The estimating functions $g_k(x_t,\theta)$ in the EL framework are each of the terms in the sum which defines Equation (5).

Two cases will be considered here in some details, i.e. the AO and the LC outlier type. In both cases an autoregressive (AR) model of order $p$ will be assumed in the presence of a single outlier of size $\omega$, which occurs at time $t=q$. For the AO a dummy variable $c_t$ may be built that takes the form $c_t=1$ if $t=q$ and zero otherwise. The dummy for the LC will be $c_t=1$ if $t\geq q$ and zero elsewhere. More details on how either an AO or an LC impacts the observed time series $y_t$ may be found in [11]. The explanatory variables are $x_t=(y_{t-1},\ldots,y_{t-p},c_t)'$ and the parameter vector is $\theta=(\phi_1,\ldots,\phi_p,\omega)'$. So model in Equation (4) becomes in more compact form $y_t=x_t'\theta+\varepsilon_t$. The estimating functions $g_k(x_t,\theta), k=1,\ldots,p+1$, may be derived from Equation (5).

$$g_k(x_t,\theta)=(y_{t-k}-c_{t-k}\omega)\,e_t,\quad k=1,\ldots,p$$
$$g_{p+1}(x_t,\theta)=\left(\sum_{j=1}^{p}c_{t-j}\phi_j-c_t\right)e_t.$$

For each $\theta$, the likelihood ratio function $\mathrm{ELR}(\theta)$ is well defined only if the convex hull of $\{g(x_t,\theta), t=1,\ldots,n\}$ contains the $(p+1)$-dimensional vector 0. For AO outliers the probability that the vector 0 is not contained in the convex hull of $g$ increases when autoregressive order decreases. This problem may well be exemplified by an AR(1) model with an AO. In this peculiar case the last constraint in Equation (3) becomes

$$p_q g_2(x_q,\theta)+p_{q+1}g_2(x_{q+1},\theta)=0.$$

If the estimating functions have the same sign the unique solution is $p_q=0, p_{q+1}=0$ and $-2\log\{\mathrm{ELR}(\theta)\}$ goes to infinity. Two kinds of EL adjustments have been suggested to address the convex hull constraint, i.e. the Adjusted EL (AEL) and the Balanced EL (BEL). An AEL has been proposed by [3] which consists of adding an artificial observation and then calculating the EL statistic based on the augmented data set. In the present example, this amounts to set $g_2(x_{n+1},\theta)=-a_n\bar{g}$. [4, 6] proposed a BEL where two balanced points are added to the data set, i.e. $g_2(x_{n+1},\theta)=\delta$ and $g_2(x_{n+2},\theta)=2\bar{g}-\delta$. The investigation on the BEL method for inference about a parameter vector $\theta$ seems very important for improving the method performance. An appropriate choice of location for the new extra points is proposed in such a way as to guarantee that correct coverage levels are obtained.

## 4 Some results from a simulation experiment

The first example is concerned with an LC in an AR(1) model. 250 standard normal random numbers have been generated and used for building an AR(1) time series with parameter $\phi=0.7$. The first 50 values have been discarded. An outlier

of size $\omega = 5$ has been added starting from time $q = 100$ on. The 90% confidence region for the ELR test compared to the likelihood test in normality hypothesis is displayed in Fig. 1 (left hand panel). The confidence regions are quite similar in spite of the fact that much less information has been employed for building the ELR test. For nominal $1-\alpha$ confidence level, the observed coverage frequencies computed on 1000 replications refer to EL $(1-\alpha_{EL})$ and normal-based $(1-\alpha_N)$ confidence regions. Nominal and observed coverage level are displayed in Table 1. Results are quite satisfying overall, and with the only exception of the 90% confidence probability level the EL coverage frequencies are slightly more accurate than their normal-based counterpart.

The second example is concerned with an AO in the same AR(1) model. An AO of size $\omega = 5$ has been added at time $q = 100$. The 90% confidence region for the ELR test compared to the parametric likelihood ratio test under normality hypothesis is displayed in Fig. 1 (right hand panel). The confidence region computed under the hypothesis of normality is narrower than that computed by the ELR statistic since the former benefits of a strong distributional assumption. However as far as the AR parameter is concerned the difference is negligible. Note that the BEL had to be employed necessarily for the EL method to work properly, in accordance with the argument developed in the preceding Sect. 3.

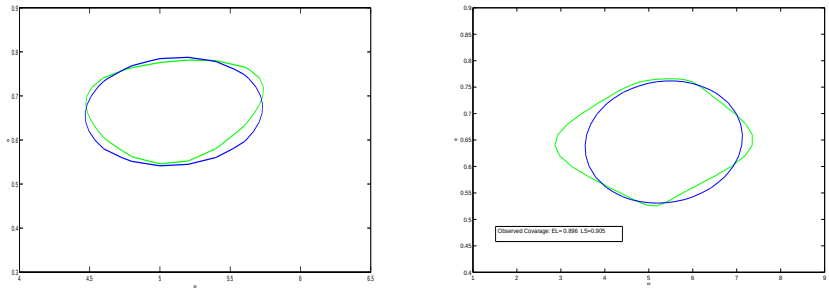


**Fig. 1** LC (left hand panel) and AO (right hand panel) simulated in an AR(1) $\phi = 0.7$ $n = 200$ $\omega = 5$ $q = 100$. Confidence regions at 90% Green=ELR Blue=normal ellipsoid

**Table 1** Coverage frequencies for a Level Change in an AR(1) model

| $1-\alpha$ | $1-\alpha_{EL}$ | $1-\alpha_N$ |
|---|---|---|
| 0.95 | 0.946 | 0.933 |
| 0.90 | 0.889 | 0.891 |
| 0.80 | 0.794 | 0.786 |
| 0.70 | 0.699 | 0.698 |
| 0.60 | 0.607 | 0.610 |
| 0.50 | 0.502 | 0.510 |

## 5 Conclusions

Empirical likelihood methods have been considered for estimating outliers size in time series models and computing confidence regions for the estimated sizes. The use of the balanced empirical likelihood has been taken into account aiming at obtaining more accurate coverage and larger power in testing of hypotheses, and allowing the computation of outlier size estimates in cases where plain empirical likelihood fails to provide feasible solutions. The empirical likelihood based methods for making inference on outliers in time series have been illustrated on two simulated examples concerned with a level change outlier and an additive outlier in a first order autoregressive model. Further interesting topics, e.g. other outlier types, and outliers identification and estimation in a wider class of time series models, such as the general autoregressive moving average and the non linear models, are left for future research.

**Acknowledgements** This research was supported by the grant C26A1145RM of the Università di Roma La Sapienza, and the national research PRIN2011 ”Forecasting economic and financial time series: understanding the complexity and modelling structural change”, funded by Ministero dell’Istruzione dell’Università e della Ricerca.

## References

1. Balke, N. S., Fomby, T. B.: Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. J. Appl. Econometrics **9**, 181–200 (1994)
2. Bruce, A.G., Martin, R.D.: Leave-k-out diagnostics for time series. J. R. Stat. Soc. Ser. B Stat. Methodol. **51**, 363–424 (1989)
3. Chen, J., Variyath, A. M., Abraham, B.: Adjusted empirical likelihood and its properties. J. Comput. Graph. Statist. **3**, 426-443 (2008)
4. Emerson, S., Owen, A. B.: Calibration of the empirical likelihood method for a vector mean. Electron. J. Stat. (2009) doi: 10.1214/09-EJS518
5. Galeano, P., Peña, D.: Finding outliers in linear and nonlinear time series. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures, pp. 243-260. Springer, Heidelberg (2013)
6. Liu, Y., Chen, J.: Adjusted empirical likelihood with high-order precision. Ann. Statist. **38** 1341–1362 (2010)
7. Owen, A. B.: Empirical likelihood ratio confidence intervals for a single functional. Biometrika **75**, 237–249 (1988)
8. Owen, A. B.: Empirical likelihood for linear models. Ann. Statist. **19**, 1725-1747 (1991)
9. Owen, A.B.: Empirical Likelihood. Chapman & Hall/CRC, Boca Raton (2001)
10. Qin, J., Lawless, J.: Empirical likelihood and general estimating equations. Ann. Statist. **22**, 300–325 (1994)
11. Tsay, R. S.: Outliers, level shifts and variance changes in time series. J. Forecasting **7**, 1–20 (1988)