NEW PROPOSALS IN MULTIVARIATE OUTLIERS IDENTIFICATION

Roberto Baragona[1], Francesco Battaglia[2]

[1] Dipartimento di Sociologia e Comunicazione, Università di Roma "La Sapienza", Via Salaria 113, 00198 Roma, Italy
(roberto.baragona@uniroma1.it)

[2] Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", Piazzale Aldo Moro 5, 00100 Roma, Italy
(francesco.battaglia@uniroma1.it)

ABSTRACT: Occurrences of outliers in multivariate time series are unpredictable events which may severely distort the analysis of the series. It may be noticed that a convenient way for representing multiple outliers consists in superimposing a deterministic disturbance to a Gaussian multivariate time series. Then outliers may be modelled as non – Gaussian time series components. The independent component analysis is a recently developed tool that is likely to be able to extract possible outlier patterns. In practice the independent component analysis may be used to analyze multivariate observable time series and separate regular and outlying unobservable components. In the factor models framework too, independent component analysis turns out to be a useful tool for outliers detection in multivariate time series.

KEYWORDS: Independent component analysis, Multivariate time series, Outliers.

1 Introduction

A problem that arises in time series analysis is outlier detection in multivariate time series. An "outlier signal" may usually be described by a simple deterministic sequence. This signal is however difficult to recognize as it is embedded in a data set whose dynamic structure may either mask the outlying observations or spread the disturbance to data otherwise unaffected. Therefore the outlier problem in time series is usually harder than in a random sample. The extent of the anomaly is to be judged with respect to its conditional distribution given the remaining observations, rather than to the marginal distribution. So outliers are not necessarily the largest or smallest data often detected by simple graphical inspection but are data which are not locally coherent with their surrounding observations.

Methods that extend well established procedures for univariate time series to the multivariate framework have been proposed e.g. by Tsay, Peña and Pankratz (2000) based on fitting multiple ARIMA models. Widespread techniques for handling outliers aim at estimating both occurrence times and sizes. Nevertheless, if times of

outliers occurrences were known, then estimating their magnitude (size) would be simplified considerably.

Galeano, Peña and Tsay (2004) used projection pursuit techniques in order to find the linear combination of a multivariate time series that maximizes kurtosis with the purpose of best reproducing the outlying signal. Then, detection of time points and estimating the magnitudes of multiple outliers may be accomplished by employing univariate searching methods.

Here we propose the independent component analysis (ICA) as a tool capable of identifying the locations of multiple outliers in multivariate time series. The reason is that outlying components have a very large kurtosis, as we explain in the next Section.

The ICA aims at identifying a set of independent unobservable variables that are supposed to generate the data set of interest. An unknown mixing matrix is postulated to linearly transform the unobservable variables to produce a set of observable mixed ones. Both unobservable variables and the mixing matrix have to be estimated from the data. The ICA has been applied successfully to a variety of fields such as biomedicine, speech, sonar and radar, signal processing and time series. A comprehensive account of ICA theory and applications may be found in Hyvärinen, Karhunen and Oja (2001).

Suppose that we observe a contaminated multivariate time series obtained by linearly mixing some independent Gaussian signals, and adding, only at some fixed time points, a constant to each observed component. When the series is decomposed by ICA, the most important non – Gaussian component is likely to represent the outlying pattern, while the remaining independent components would be essentially similar to Gaussian linear combinations of the outlier free time series.

2 The multivariate outlier model

Let us denote by $\{\Delta_t\}$ the sequence describing the outlier pattern, that is $\Delta_t = 1$ if at time $t = t_0$ there is an outlying observation, and $\Delta_t = 0$ otherwise. A level change at time $t = t_0$ may be represented by the sequence $\Delta_t = 1$ if $t \geq t_0$ and 0 otherwise. An outlier patch (a sequence of consecutive outliers) of length m beginning at time $t = t_0$ is easily modelled by defining $\Delta_t = 1$ if $t = t_0+1, \ldots, t_0+m$ and 0 otherwise. The perturbation series $\{\Delta_t\}$ is far from normality and its kurtosis coefficient is large in modulus. If n data are recorded and $\Delta_t = 1$ for q distinct times $1 \leq t_1 < \ldots < t_q \leq n$, while $\Delta_t = 0$ for the remaining times, then letting $\alpha = q/n$ it may be seen that its kurtosis coefficient is $\{\alpha(1-\alpha)\}^{-1}-6$ which is considerably large in modulus for moderate α (the proportion of outlying observations).

Suppose that the observed series $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{ht})'$, $t = 1, \ldots, n$, is obtained by adding perturbations to an outlier free time series $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{ht})'$ having a multivariate Gaussian distribution with $E(\mathbf{x}_t) = \mathbf{0}$ and $E(\mathbf{x}_t\mathbf{x}_t') = \mathbf{\Sigma}_x$ positive definite. Then we may consider the model $\mathbf{y}_t = \mathbf{x}_t+\boldsymbol{\omega}\Delta_t$ where $\boldsymbol{\omega}_t=(\omega_1, \omega_2, \ldots, \omega_h)'$ is the outlier magnitude. In order to apply ICA, we shall write $\mathbf{y}_t$ as a linear combination of independent possibly non – Gaussian series. Since the variance – covariance matrix $\mathbf{\Sigma}_x$ is positive definite, the regular symmetric matrix $\mathbf{S} = \mathbf{\Sigma}_x^{-1/2}$ exists. Let $\boldsymbol{\lambda} = \mathbf{S}\boldsymbol{\omega}$ and

denote also the standardized variables by $\mathbf{u}_t = \mathbf{S}\mathbf{x}_t$, so that $E(\mathbf{u}_t\mathbf{u}_t') = \mathbf{I}$. The observed series may be written $\mathbf{y}_t = \mathbf{S}^{-1}\mathbf{u}_t + \mathbf{S}^{-1}\boldsymbol{\lambda}\Delta_t$. Next, we define an orthonormal basis of $\Re^h$, $(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_h)$ with $\mathbf{m}_1 = \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|$, and consider the matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_h]$. Then $\mathbf{M}'\boldsymbol{\lambda} = (\|\boldsymbol{\lambda}\|, 0, \ldots, 0)'$ and $\mathbf{z}_t = \mathbf{M}'\mathbf{u}_t = \mathbf{M}'\mathbf{S}\mathbf{x}_t$ are also independent. The observed series may be written $\mathbf{y}_t = \mathbf{S}^{-1}\mathbf{M}\{\mathbf{z}_t + (\|\boldsymbol{\lambda}\|, 0, \ldots, 0)'\Delta_t\}$, where the first component in the linear transformation, $z_{1t} + \|\boldsymbol{\lambda}\|\Delta_t$, is contaminated by outliers and has large kurtosis, while the other components z_{jt}, $j = 2, \ldots, h$, are Gaussian variables. Therefore, ICA will be able to identify the de – mixing matrix and the outliers effects will be confined in just one of the extracted components. Whether such effects will be evident and clearly recoverable, it depends on the stochastic behaviour of z_{1t}, which has zero mean and variance one. Thus the outlier will be clearly evident in the component if $\|\boldsymbol{\lambda}\| = \{\boldsymbol{\omega}'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\omega}\}^{1/2}$ is large compared to a standard Gaussian variate. However, if the x_{jt}'s are highly autocorrelated series, then the occurrence of a perturbation in z_{1t} may be better recognized using a univariate outlier detection method (see e.g. Tsay, 1988, and Chen and Liu, 1993).

We turn to consider now a different model, which appears to be suitable for representing real data, specially with a large dimension, and allows for a factor structure and noise. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{ht})'$, $t = 1, \ldots, n$, denote as before the observed data, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{ht})'$ denote a sequence of independent identically distributed vector Gaussian variables with mean $\mathbf{0}$ and variance – covariance matrix $\boldsymbol{\Sigma}_\varepsilon = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_h^2\}$. We suppose that the observed time series is obtained by a linear transformation of a k – variate (unobservable) time series $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{kt})'$ with $k << h$, an outlier with vector magnitude $\boldsymbol{\omega}$ and pattern Δ_t, plus a noise represented by the $\{\boldsymbol{\varepsilon}_t\}$ series. Namely, we have $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\omega}\Delta_t + \boldsymbol{\varepsilon}_t$. If we set $\boldsymbol{\omega} = \mathbf{0}$, then the model known as dynamic factor model is obtained which has been often used in time series analysis and econometric studies (see e.g. Peña and Box, 1987, Forni and Reichlin, 1998, and Bai and Ng, 2002).

In this case also it may be shown (Baragona and Battaglia, 2006) that the independent component analysis helps outlier identification: the h sources extracted by means of ICA may be characterized as follows. (a) One component includes only the term closely related to the outlier pattern, plus a noise term. (b) Each of k components contains one of the original factor series, a possible term due to the outliers and a random error. (c) The remaining $h - k - 1$ components are purely random noise.

In the particular case that the outlier magnitude vector $\boldsymbol{\omega}$ is exactly a linear combination of the columns of the factor matrix $\mathbf{A}$, i.e. $\boldsymbol{\omega} = \mathbf{A}\boldsymbol{\alpha}$, the perturbation in the observed series may be entirely explained by outliers in the factors, and the first extracted component (a) disappears, since we only can recover the perturbed factor series.

If the observed series has two outliers at different times, and their outlier magnitude vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are proportional, they will appear in the same component, whereas if $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are linearly independent, two sources of type (a) will be extracted, each one related to a different outlier.

Formal statistical tests for the null hypothesis of absence of outliers at a given time could be easily derived, owing to Gaussianity of the components, if the factor matrix coefficients were known. This is obviously not true in general, and estimation

of such parameters may be biased by the presence of the outliers themselves. Thus, a more cautious procedure may be based on checking each value against bounds $m \pm cs$, where m and s are the mean and standard error (or a robust estimate of them) of the extracted source. An outlier is detected whenever at least one of the extracted values exceeds such bounds. To assess the value of c, we think that the Gaussian quantiles are not appropriate, since our procedure tends to enhance non – Gaussianity of the data. A more conservative choice may be obtained resorting to the Chebyshev inequality. Also we may take into account each source and treat it as a univariate time series for applying methods for univariate outlier detection.

3 Experience and conclusions

The outlined procedure has been applied to a wide set of simulated data, and on some real multivariate series concerning economic data and biomedical recordings. The proposed method performed very satisfactorily in detection of isolated outliers, and good performances were observed as well for patches and level shifts.

The main advantages in comparison to ARIMA-model based methods are that no explicit model has to be estimated, and that outliers also at the very beginning or end of the series may be easily detected. Our method assumes an instantaneous ICA mixing model, therefore it does not exploit the time dynamics of the series. This could be possibly done by choosing the objective function according to dynamic relationships, it will be a subject for further research.

References

BAI, J., & NG, S. 2002. Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191-221.

BARAGONA, R., & BATTAGLIA, F. 2006. Outlier detection in multivariate time series by independent component analysis. Submitted.

CHEN, C., & LIU, L. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, **88**, 284-297.

FORNI, M., & REICHLIN, L. 1998. Let's get real: a dynamic factor analytical approach to disaggregated business cycle. *Review of Economic Studies*, **65**, 453-474.

GALEANO, P., PEÑA, D., & TSAY, R. S. 2004. Outlier detection in multivariate time series via projection pursuit. Working paper.

HYVÄRINEN, A., KARHUNEN, J., & OJA, E. 2001. *Independent Component Analysis*. New York: Wiley.

PEÑA, D., & BOX, G. E. P. 1987. Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836-843.

TSAY, R. S. 1988. Outliers, level shifts and variance changes in time series. *Journal of Forecasting*, **7**, 1-20.

TSAY, R. S., PEÑA, D., & PANKRATZ, A. E. 2000. Outliers in multivariate time series. *Biometrika*, **87**, 789-804.