

Projection Methods for Outlier Detection in Multivariate Time Series ^(*)

Metodi di proiezione per l'individuazione di dati anomali in serie temporali multivariate

Roberto Baragona¹ and Francesco Battaglia²

¹ Dipartimento di Sociologia e Comunicazione, Università di Roma “La Sapienza”
e-mail: roberto.baragona@uniroma1.it

² Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma “La Sapienza”

Riassunto: I dati anomali in serie temporali sono dati non conformi alla struttura dinamica suggerita dalla maggior parte delle restanti osservazioni. In questo contesto i dati anomali non debbono essere necessariamente cercati tra i valori estremi. Nelle serie multivariate le relazioni fra le componenti propagano spesso le anomalie di una componente sulle altre e l'individuazione di dati anomali risulta ancora più complicata. Un promettente approccio consiste nello studio dei dati anomali in combinazioni lineari univariate delle serie componenti secondo proiezioni scelte con appropriati criteri. E' presentato un metodo per l'individuazione di dati anomali in serie temporali multivariate basato sugli algoritmi genetici. Un esperimento di simulazione viene effettuato per fornire una prima valutazione della procedura in comparazione con l'analisi in componenti indipendenti e i modelli multivariati.

Keywords: multivariate time series, outliers, projection pursuit, independent component analysis, genetic algorithms.

1. Introduction

Investigation and treatment of outlying observations in time series analysis are relevant issues, because outliers may suggest the occurrence of unexpected events and, on the other hand, may produce biased parameter estimation. Outliers in time series differ from outliers in random samples as they are not necessarily extreme values. An observation in a time series is to be classified as outlier if it does not conform to the pattern exhibited by neighboring data. Moreover, the dynamic structure of the process underlying the observed time series may originate ‘masking’ and ‘smearing’ effects. In the multivariate time series framework outliers identification is apparently more difficult as the joint dynamics of the component series has to be properly considered. As a matter of fact, outliers in a component may either produce or hidden outlying observations in other components.

Most outlier studies concerned with univariate time series assume that the data are generated by some linear structure, for instance autoregressive integrated moving-average (ARIMA) models and finite linear interpolators (FLI). In this view an observation is classified outlier if it differs considerably from its conditional mean computed according to the model. Univariate procedures may well be extended to the multivariate framework. Tsay, Peña and Pankratz (2000) study outliers in multivariate time series by generalizing

(*) Work partially supported by the University of Rome “La Sapienza” (2004)

the approach for univariate time series based on ARIMA modeling (see, for instance, Tsay, 1986, Chang, Tiao and Chen, 1988, and Tsay, 1988). A multivariate ARIMA model has to be built, however, and the presence of multiple outliers may render this issue very difficult.

An approach that has been recently repeatedly suggested for discovering outlying observations in multivariate time series is finding the projection that best highlights irregular behavior of the dynamic vector structure. This view has already been considered in robustness studies to become a well established practice (see, e. g., Maronna and Yohai, 1995). Projection pursuit (PP) and independent component analysis (ICA) were the most promising frameworks for designing effective and efficient procedures to handle outliers in multivariate time series. Hyvärinen, Karhunen and Oja (2001) provide a comprehensive review of ICA theory and applications. For PP we may refer to Huber (1985) and Friedman (1987). A PP method based on kurtosis maximization has been proposed and investigated by Galeano, Peña and Tsay (2004). An identification procedure has been developed in the ICA framework by Baragona and Battaglia (2005). Graphical inspection of the most non-normal projections or independent components usually shows the presence of outlying observations with greatest evidence. Nevertheless, establishing some rules for automatic identification is of interest specially if the number of component time series is large or outliers' detection is required to be quickly and accurately accomplished. Also identifying outliers by means of graphical inspection is a somewhat subjective procedure that may produce different results according to different researchers.

In this paper a method based on PP is illustrated that takes the dynamic structure of the multivariate time series into account and, on the other hand, uses meta-heuristic methods for finding optimal solutions. Meta-heuristic methods (see Pirlot, 1996) are called as well general heuristics or general local search methods because they are not structured algorithms but all-purpose stochastic search methods. They offer basic ideas to be implemented properly for developing procedures to cope efficiently with particular problems. For example, simulated annealing, tabu search and the genetic algorithms are well known meta-heuristic methods. Furthermore, a method based on ICA will be examined which uses the Hyvärinen's fixed-point algorithm (Hyvärinen and Oja, 1997). Usually outliers in time series are classified into several types, but we shall consider only outliers of the additive type. The notion of patch (sequence of consecutive outlying observations) is likely to be useful for unifying outliers' treatment, simplifying formulation of procedures for outliers detection and allowing more efficient methods to be designed. While we are interested only about the time points where outliers occur, this may be considered as a first step, that is a second step has to be possibly devoted to examine if each patch may be modeled in accordance to some structure. We shall examine some rules that may help to decide whether and where the projections or extracted components of a multivariate time series are giving evidence of the presence of multiple outlying multivariate observations.

The paper is organized as follows. In Section 2 models for multivariate outliers and projection methods will be described. A genetic search projection method for outliers detection will be introduced in Section 3. In Section 4 the performances of the procedures will be checked by means of a Monte Carlo simulation study. In Section 5 some final remarks are included.

2. Outlier models and projection methods

Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{ht})'$, $t = 1, 2, \dots, n$, be an observed multivariate time series that includes outlying observations at unknown time points. We assume

$$\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\omega} \Delta_t, \quad (1)$$

where $\mathbf{x}_t = (x_{1t}, x_{2t}, \dots, x_{ht})'$, $t = 1, 2, \dots, n$, is a unobservable time series normally distributed with mean zero and positive definite variance-covariance matrix Σ_x . The array $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_h)'$ defines the outlier size. The outlier pattern may be modeled as a binary sequence $\{\Delta_t\}$, where $\Delta_t = 1$ if at the time point t an outlying observation is present. Equation (1) may be used for modeling outlier patterns such as isolated outliers, outliers patches and changes in the level.

Unlike $\mathbf{x}_t$ the series $\{\Delta_t\}$ is not normal. We may calculate its kurtosis coefficient as a measure of departure from normality. Let $\Delta_t = 1$ for q time points and $\Delta_t = 0$ for the remaining $n - q$. The standardized version of $\{\Delta_t\}$ equals $\sqrt{(1 - \alpha)/\alpha}$ with frequency q and $-\sqrt{\alpha/(1 - \alpha)}$ with frequency $n - q$, where $\alpha = q/n$ is the proportion of outlying observations. The kurtosis is

$$\kappa_4 = \mu_4 - 3(\sigma^2)^2 = \frac{1}{\alpha(1 - \alpha)} - 6. \quad (2)$$

It follows from Equation (2) that the kurtosis coefficient, as a function of the outlier percentage, is symmetric around $\alpha = 0.5$. Its absolute value decreases when α increases from zero to 0.5 ($1 - 1/\sqrt{3}$) ≈ 0.211 and then increases when α is approaching 0.5 . The absolute value of the kurtosis of the series Δ_t is always larger than that of the normal components $\mathbf{x}_t$. It has to be noticed that the kurtosis does not depend on the outliers pattern but it depends only on the ratio q/n . In particular, in the presence of a level change, kurtosis may be as well useful for identification purpose, unless $(n - q)/q$ is near $(\sqrt{3} + 1)/(\sqrt{3} - 1) \approx 3.73$.

The analysis of the kurtosis has been suggested by Galeano, Peña and Tsay (2004) for finding directions useful for outlier detection. The projections in one-dimensional space are computed as linear combinations of the components of the observed multivariate time series. The coefficients that define each projection are calculated by solving a constrained optimization problem. In order to find the optimum the stationary points are computed where the gradient of the objective function is equal to zero. Such points are possibly relative maxima or relative minima. Starting points are needed to initialize the iterative algorithm for computing the stationary points and this choice may impact the final result. We suggest to use meta-heuristic methods as they are known to be less sensible than quasi-Newton-type algorithms, for instance, to the initial values. Moreover, meta-heuristic methods perform a direct search to yield the global optimum and avoid the stationary points to be exhaustively examined. Another relevant issue is concerned with the objective function that has to be optimized to find interesting projections. The methods that aim at maximizing the kurtosis do not take into account the dynamics of the multivariate time series. Objective functions built on the dynamic structure of the data may possibly yield univariate time series with potential outliers easily identifiable. Such objective functions may be based on either prediction or interpolation error, for instance. In particular, the residuals computed from the FLI are known to be strictly related to the additive outliers

(Peña, 1990). For a given univariate time series y_t the FLI is defined

$$y_t = \sum_{j=1}^m \varphi_j (y_{t-j} + y_{t+j}) + e_t, \quad t = m+1, \dots, n-m, \quad (3)$$

for some pre-specified positive integer m . The coefficients φ_j may be estimated by ordinary least squares. $\{e_t\}$ is the FLI errors series. If an additive outlier occurs at time point t , then e_t is the maximum likelihood estimate of its size. The maximum of the absolute values of the series $\{e_t\}$ shall be considered as a suitable objective function for finding interesting projections to detect outliers. Meta-heuristic methods offer valuable tools to implement this procedure as far as intensive computational effort is needed. In particular we suggest that GAs may be used as they allow several paths towards solutions to be explored in a parallel fashion. This peculiar feature is believed to increase the chance that the global optimum will be attained.

The ICA aims at identifying a set of independent unobservable variables that are supposed to generate the observed data set. It is postulated that the unobserved variables have been linearly transformed according to some unknown mixing matrix. Both unobservable variables and the mixing matrix have to be estimated from the data.

In order to apply ICA, we shall write $\mathbf{y}_t$ as a linear combination of independent possibly non-normal series. Since the variance-covariance matrix Σ_x is positive definite, the regular symmetric matrix $\mathbf{S} = \Sigma_x^{-1/2}$ exists. Let $\boldsymbol{\lambda} = \mathbf{S}\boldsymbol{\omega}$ and denote the standardized variables by $\mathbf{u}_t = \mathbf{S}\mathbf{x}_t$, so that the variance-covariance matrix of $\{\mathbf{u}_t\}$ is the identity matrix. 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 $\mathbb{R}^h$, $(\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_h)$ with $\mathbf{m}_1 = \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|$, and consider the matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_h]$. Then $\mathbf{M}'\boldsymbol{\lambda} = (\|\boldsymbol{\lambda}\|, 0, \dots, 0)'$ 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

$$\mathbf{y}_t = \mathbf{S}^{-1}\mathbf{M} \{ \mathbf{z}_t + (\|\boldsymbol{\lambda}\|, 0, \dots, 0)' \Delta_t \},$$

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 $z_{jt}, j = 2, \dots, h$, are normal variables. Therefore, ICA will be able to identify the de-mixing matrix and the outliers effects will be confined in one of the extracted components. Whether such effects will be evident and clearly recoverable, it depends on the stochastic behavior of z_{1t} , which has zero mean and variance one. The outlier will be clearly evident in the component as far as $\|\boldsymbol{\lambda}\| = \{\boldsymbol{\omega}'\Sigma_x^{-1}\boldsymbol{\omega}\}^{1/2}$ is large.

3. A genetic search projection method for multivariate time series outliers

GAs were introduced by Holland (1975) as a framework for studying the behavior of both natural and artificial adaptive systems. The evolution of a biological population through several generations to fit the environment may be modeled by means of a selection rule and some genetic operators. GAs perform optimization of an objective

function by evolving the artificial population composed of a subset of the set of potential solutions. The fittest individual found after some iterations is assumed to be the optimal solution. Mitchell (1996) presented several applications of GAs as numerical optimizers. Chatterjee, Laudato and Lynch (1996) and Chatterjee and Laudato (1997) dealt with applications in statistics, for instance variable selection and parameter estimation in linear regression modeling.

Our GA algorithm is essentially structured as a simple GA according to the guidelines provided by Goldberg (1989). Given positive integers s and ℓ , a set of s binary strings $(i_1, i_2, \dots, i_\ell)$ is uniformly randomly generated. This set of s strings is called “initial population”. Starting from the initial population, N iterations are performed in each of which the s strings are allowed to change according to the “genetic operators” selection, crossover and mutation.

- **Selection.** The objective function to be optimized is called in this context the “fitness function”. The GAs generally assume that the fitness function is positive and has to be maximized. For s times a string in the current population is chosen (with replacement) with probability proportional to its fitness function. We obtain, in this way, s strings that replace the current population. We may want the population to be only partially replaced. In this case, we choose a “generational gap” G and in the next generations only $G \times s$ strings are considered for replacement.
- **Crossover.** We may assume for simplicity that s is an even integer. Then, the strings in the current population may be paired at random to yield $s/2$ couples. Each pair may undergo the crossover with a pre-specified probability. The crossover acts as follows. (1) An integer j is chosen uniformly randomly in the interval $(1, \ell - 1)$. The number j is called the “cutting point”. (2) The bits from $j + 1$ to ℓ are exchanged between the strings that are paired.
- **Mutation.** Each bit of each string in the current population is allowed to flip with a pre-specified mutation probability. The probability for a bit to mutate is typically small and ranges usually between 0.001 and 0.01.

The “elitist strategy” (see Jennison and Sheehan, 1995, for motivation) applies, that is the best string found in each of the N iterations is always maintained in the current population unless an even better string appears due to the genetic operators. If this latter is not the case, then such string replaces the string with the worst fitness function.

We shall examine random search by GAs of univariate projections either with the aim of maximizing the absolute value of the kurtosis, or, alternatively, for maximizing the absolute value of the interpolation errors.

Let $\{y_{jt}, j = 1, \dots, h, t = 1, \dots, n\}$ be the observed multivariate time series and $(c_1, \dots, c_h)$ be real coefficients such that $\sum c_j^2 = 1$ and

$$p_t = \sum_{j=1}^h c_j y_{jt} \tag{4}$$

is the univariate projection to the direction defined by the c_j ’s coefficients. If we aim at finding by means of GAs the linear combination (4) that allows outliers to be easily discovered, then appropriate encoding and fitness function have to be defined.

As for encoding, it is well known that $h - 1$ angles suffice to completely and uniquely determine a point on the unit h -dimensional sphere. Let $0 \leq \phi \leq 2\pi$ and $0 \leq \theta_j \leq \pi$, $j = 1, \dots, h - 2$. Then

$$\begin{aligned} c_1 &= \sin(\theta_1)\sin(\theta_2) \times \dots \times \sin(\theta_{h-2})\sin(\phi) \\ c_2 &= \sin(\theta_1)\sin(\theta_2) \times \dots \times \sin(\theta_{h-2})\cos(\phi) \\ c_3 &= \sin(\theta_1)\sin(\theta_2) \times \dots \times \cos(\theta_{h-2}) \\ c_4 &= \sin(\theta_1)\sin(\theta_2) \times \dots \times \cos(\theta_{h-3}) \\ &\dots \\ c_h &= \cos(\theta_1) \end{aligned}$$

are unit norm coefficients that define a univariate projection. Every feasible solutions that GAs may take into account may be coded as a sequence $\{\theta_1, \theta_2, \dots, \theta_{h-2}, \phi\}$. A set of sequences of this type forms the population to be evolved through the pre-specified number of generations to reach the optimal projection. We consider optimality according to the following two criteria:

- kurtosis, and the fitness function is assumed

$$f(\theta_1, \theta_2, \dots, \theta_{h-2}, \phi) = f(c_1, c_2, \dots, c_h) = \left| \frac{\mu_4}{\text{var}(p_t)^2} - 3 \right|$$

- largest interpolation error, with fitness function

$$f(\theta_1, \theta_2, \dots, \theta_{h-2}, \phi) = f(c_1, c_2, \dots, c_h) = \max_s \left(\frac{|e_s|}{\sqrt{\frac{1}{n} \sum e_t^2}} \right)$$

where e_t is the linear interpolation error defined in (3) with input series p_t defined in (4), and the unknown parameters $\{\varphi_j, j = 1, \dots, m\}$ are estimated by least squares.

The actual chromosomes are binary vectors of length ℓ obtained by standard binary coding of $\{\theta_1, \theta_2, \dots, \theta_{h-2}, \phi\}$. We assume 20 bits to represent each real parameter so that the length of each string is $\ell = 20 \times (h - 1)$. The generational gap is assumed $G = 0.9$. The simple mutation and single cut point crossover operators are employed with probabilities equal to $0.7/\ell$ and 0.7 respectively.

4. Monte Carlo experiment

A simulation experiment was performed to compare different methods for detecting multiple outliers in multivariate time series. Two models with 5 components series were considered, a factorial model and a vector autoregressive (VAR) model. Computations were performed by means of the Matlab language, using specific toolboxes for GAs (Genetic Algorithm Toolbox 1.2, developed by Chipperfield, Fleming, Pohlheim and Fonseca, 1994) and for the ICA (FastICA package 2.5, developed by Gävert, Hurri, Särelä and Hyvärinen, 2005).

As far as the factorial model simulation is concerned, we considered 4 time series generated by the first order autoregressive models

$$\xi_{jt} = \pm 0.7 \xi_{j,t-1} + a_t$$

where coefficient 0.7 was assumed for $j = 1, 3$ and -0.7 for $j = 2, 4$, and $\{a_t, t = 1, \dots, n\}$ was a zero-mean unit variance white noise. Then the following series

$$\Delta_t = \begin{cases} 1 & \text{if } t = \tau, \dots, \tau + q - 1 \\ 0 & \text{otherwise} \end{cases}$$

defined the outlier pattern. Two cases were considered, that is a single outlier in the middle of the series ($\tau = 50, q = 1$) and a sequence (patch) of outlying observations ($\tau = 51, q = 3$). The observed multivariate time series was

$$\mathbf{y}_t = \mathbf{A} \begin{pmatrix} \xi_t \\ \Delta_t \end{pmatrix} + \eta_t$$

with $\mathbf{A}$ the 5×5 mixing matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 3 \\ 0.5 & 1 & 0 & 0 & 3 \\ 0 & 0.5 & 1 & 0 & 3 \\ 0 & 0 & 0.5 & 1 & 3 \\ 0 & 0 & 0 & 0.5 & 3 \end{pmatrix}$$

where the last column includes the outliers' sizes, and with $\{\eta_t, t = 1, \dots, n\}$ multivariate zero-mean uncorrelated white noise with common variance 0.04. The standard deviations of the components range approximately from 1 to 1.5 so that 3 seemed an appropriate choice for the outlier's size. Though the time series that generated the multivariate time series were independent, the mixing matrix yielded a rather strong linear structure.

As far as the simulation of the VAR(5) model is concerned, we assumed the relationship

$$\mathbf{y}_t = \Phi \mathbf{y}_{t-1} + \omega \Delta_t + \epsilon_t,$$

where ω is the column vector $(4, 4, 4, 4, 4)'$ and Φ is the matrix coefficient

$$\Phi = \begin{pmatrix} 0.7 & 0.2 & 0.1 & 0.25 & -0.1 \\ 0.2 & 0.75 & 0.1 & 0.1 & -0.1 \\ -0.1 & -0.2 & 0.7 & -0.1 & -0.1 \\ 0.2 & -0.1 & 0.25 & 0.7 & -0.1 \\ 0.2 & 0.1 & 0.1 & -0.1 & 0.7 \end{pmatrix}.$$

The eigenvalues are all less than 1 in modulus so that the VAR is stationary. The innovation $\{\epsilon_t\}$ variance-covariance matrix was the 5-dimension identity matrix. The standard deviations of the component series are larger than those of the factorial model (about 2). This circumstance suggested that 4 could be an appropriate size for the outliers. The series are linearly slightly more correlated than the components of the factorial model.

In Table 1 the quantiles of the statistics used for outlier detection are reported. The figures were assessed by simulation by generating 10000 outlier-free artificial time series according to the factorial model described so far. Two GAs-based procedures were considered, the first one with fitness function equal to the absolute value of the kurtosis,

the second one with fitness function equal to the maximum of the absolute value of the normalized interpolation error of the projected series. We assumed 30 as population size and 100 as the number of generations. The quantiles were referred to the statistics obtained by taking the absolute value of the normalized interpolation error and, for the procedure aimed at maximizing the kurtosis, the absolute value of the normalized projection. For the ICA-based procedure the FastICA algorithm was used. The search for outliers was performed on the first extracted component and the quantiles were concerned with the distribution of the normalized component data. It turns out that the quantiles in Table 1 are not far from the usual thresholds of 3.0 or 3.5 that are common practice in outlier analysis. Finally, the Tsay, Peña and Pankratz (2000) procedure (TPP) was tried and quantiles were assessed for the statistic J in the case of additive outliers. This is necessary because the crude approximation by a chi-square distribution is not satisfactory, and no extensive tables are available.

Table 1: *Quantiles of the statistics used for outlier detection assessed by simulation. 10000 replications of outlier-free artificial multivariate series of length 100 each were generated.*

Algorithm	stat. outlier detect.	95% quantile	99% quantile
GAs (kurtosis)	$ e_t /\sigma_e$	1.9276	2.7094
GAs (kurtosis)	$ p_t - m /\sigma_p$	2.85	4.64
GAs (int. err.)	$ e_t /\sigma_e$	1.9191	3.2653
ICA (FastICA)	$ p_t - m /\sigma_p$	1.7867	2.5825
TPP (AO)	J	20.9548	24.244

In Table 2 the case of a single outlier that occurs at time $t = 50$ is considered as far as the factorial model is concerned. Comparison is reported between the two GAs-based procedure, the ICA-based procedure and the TPP procedure. The GAs-based procedure uses as fitness function either the kurtosis or the largest interpolation error. In both cases the rule for outlier detection was assumed whether the absolute value of the normalized interpolation error at time t exceeded either 3.0 or 3.5. Figures for both thresholds are shown in Table 2 and it may be noted that the results are rather stable. For the ICA-based procedure also the rule for outlier detection was assumed whether the absolute value of the standardized first extracted component exceeded either 3.0 or 3.5. For the TPP procedure an outlier was assumed as soon as the statistic J exceeded 25.

The Monte Carlo simulation included 100 replications for computing the percentage of detection of the isolated outlier and the percentage of time series where outliers were incorrectly detected. Within each replication 100 iterations were performed to compute the standard errors of the detection percentages.

All methods are perfectly able to detect the isolated outlier, and differences amongst procedures are very small. We may note only slight differences as the independent component analysis-based method yields 94.5% success, the multivariate model-based method 100% success, and the genetic algorithms-based methods yield percentages of success in between.

In Table 3 the case of a sequence of consecutive outliers (patch) at $t = 51, 52, 53$ is considered. The same procedures displayed in Table 2 were taken into account but we

Table 2: Average percentages of outlier correct and incorrect identification in 100 replications of multivariate series generated by a factorial model each repeated 100 times with a single multivariate outlier of size 3 at $t = 50$. Standard errors are enclosed in parentheses.

	Threshold 3.0		Threshold 3.5	
Procedure	% correct	% false	% correct	% false
GAs (kurtosis)	98.8 (1.03)	2.70 (1.48)	98.7 (1.03)	0.67 (0.88)
GAs (int. err.)	99.9 (0.22)	5.61 (2.14)	99.9 (0.22)	0.89 (0.99)
ICA	94.5 (2.12)	3.17 (1.49)	94.5 (2.12)	2.66 (1.37)
TPP (Threshold 25)			100 (0.0)	0.66 (0.84)

considered further to refine the GAs-based procedure by performing a univariate search for outliers in the projection series.

The Monte Carlo simulation included 100 replications for computing the percentage of detection of at least one outlier in the patch and of the entire patch, and the percentage of time series where outliers were incorrectly detected. Within each replication 100 iterations were performed as before to compute the standard errors of the detection percentages.

It may be seen that the methods based on interpolation errors (GAs) and prediction errors (TPP) are not adequate, since they are generally able to identify at least one of the outliers but frequently fail in identifying the complete patch. On the contrary, the ICA-based method and the GAs-based procedure aimed at maximizing the kurtosis (using the normalized projection series directly for outliers detection) are considerably more efficient, and a complete identification is obtained in more than 90% of the series. A possible explanation is that the masking and smearing effects seriously affect both interpolation and prediction error estimates. If we use in a further step a univariate method specifically derived for identifying multiple outliers, such drawback may be partially amended. The last two rows of Table 3 show results obtained by means of a multiple outlier detection method (Baragona, Battaglia and Calzini, 2001) based on a GA which evaluates every possible outlier configuration through penalized likelihood. It turns out that such method may improve the best results obtained by applying simple detection rules. In addition, this method provides estimates of the outliers' size along with their location. These advantages are partly contrasted by the circumstance that the percentages of false detections seem rather large.

In Table 4 results are reported concerned with outlier patch detection in the VAR(5) model. We considered the GAs-based procedure for kurtosis maximization, the ICA-based procedure and the TPP procedure for the thresholds corresponding to 95% and 99% levels.

The stronger linear structure is likely to impact negatively on the searching methods performances. The percentage of success in detecting at least an outlier was nearly 90%

Table 3: Average percentages of correct identification of at least one outlier, of the complete patch, and of incorrect outlier detection in 100 replications of multivariate series generated by a factorial model each repeated 100 times with a sequence of outliers in $t = 51, 52, 53$. Standard errors are enclosed in parentheses.

Procedure	Threshold 3.0			Threshold 3.5		
	% id. ≥ 1	% id. all	% false	% id. ≥ 1	% id. all	% false
GAs (kurtosis)	97.6	7.4	10.9	85.5	1.11	2.67
(int. err.)	(1.43)	(2.73)	(2.98)	(3.35)	(1.28)	(1.57)
GAs (kurtosis)	99.9	99.6	0.01	99.9	97.4	0.0
(norm. ser.)	(0.1)	(0.67)	(0.1)	(0.1)	(1.52)	(0.0)
GAs (int. err.)	99.9	53.4	3.04	99.9	21.6	1.07
	(0.38)	(4.13)	(1.64)	(0.38)	(3.56)	(1.07)
ICA	93.7	93.3	2.91	93.7	91.2	2.62
	(2.19)	(2.33)	(1.73)	(2.19)	(2.60)	(1.59)
TPP				19.3	7.23	0.43
(Threshold 25)				(3.76)	(2.83)	(0.68)
Mult. GA search				99.2	98.7	16.9
(kurtosis)				(1.6)	(2.18)	(7.09)
Mult. GA search				99.9	63.15	20.1
(int. err.)				(0.38)	(5.02)	(3.99)

but on the average only two out of three outliers were identified. The complete patch was detected for no more than 75% of cases, and this percentage was considerably lower (about 47%) for the higher threshold. Percentages of false detections are larger than their nominal values but for the GAs-based method. This latter also yielded the best results. We may notice that the most promising projections were obtained by maximizing some non-normality measures. We found that the detection of outliers was more accurate when the absolute value of the kurtosis of the projection series was large. This circumstance may indicate that the projections with largest kurtosis may display the outlying observations with greatest evidence. The GAs-based method performed slightly better than the ICA-based method possibly because the former aimed directly at maximizing the kurtosis. For the factorial model the kurtosis of the extracted component was on the average 4.38 for the GAs-based method with fitness function the greatest interpolation error, 15.44 for the GAs-based method with fitness function the greatest kurtosis and 14.53 for the ICA-based method. For the VAR model these figures were 1.41, 5.39 and 3.70 respectively.

5. Conclusions

Finding interesting projections of a multivariate time series seems to be a promising approach for multiple outlier detection. Projection pursuit and independent component analysis techniques may be used for extracting one-dimensional time series that best displays the pattern of the outlying observations. Complexity of the computations needed to perform this analysis satisfactorily suggests that meta-heuristic methods should be

Table 4: Average percentages of correct identification of at least one outlier, of the complete patch, and of incorrect outlier detection in 100 replications of multivariate series generated by a VAR(5) model each repeated 100 times with a sequence of outliers in $t = 51, 52, 53$. Standard errors are enclosed in parentheses.

Procedure	Threshold 3.0			Threshold 3.5		
	% id. ≥ 1	% id. all	% false	% id. ≥ 1	% id. all	% false
GAs	90.4 (2.78)	74.6 (3.76)	5.46 (2.09)	89.0 (3.09)	46.7 (5.01)	3.19 (1.65)
ICA	60.8 (5.24)	52.0 (4.96)	8.9 (2.84)	60.3 (5.01)	33.2 (5.10)	7.0 (2.34)
Procedure	Threshold 21			Threshold 25		
	% id. ≥ 1	% id. all	% false	% id. ≥ 1	% id. all	% false
TPP	73.7 (4.51)	33.4 (4.25)	26.2 (3.91)	39.1 (4.88)	7.6 (2.45)	9.2 (3.13)

used. We propose a detection procedure based on genetic algorithms for both finding the most useful projection and identifying the multiple outlier pattern. Once some optional parameters and a proper objective function are provided the presence of outliers is automatically detected. Some objective functions and rules for outlier identification are considered and compared by means of simulation experiments. An outlier detection procedure based on multivariate models is also included in the simulations. The results show that outliers occurring in isolation are easily detected by all methods. We may note only slight differences as the independent component analysis-based method yields 94.5% success, the multivariate model-based method 100% success, and the genetic algorithms-based methods yield percentages of success in between. If many outliers occur close together then detection of the entire sequence is performed satisfactorily only by the independent component analysis-based method and by the genetic algorithms-based method aimed at maximizing the kurtosis. In these two cases percentages of success exceed 90%. Further univariate search for outliers in the projection series may slightly improve the results. For the VAR model the best methods are still the same though smaller percentages of success (less than 75%) are obtained. Extended simulation studies and applications to real series are possibly needed to check further the performance of methods for outlier detection in multivariate time series.

References

- Baragona R., Battaglia F. (2005) Outlier detection in multivariate time series by independent component analysis, submitted.
- Baragona R., Battaglia F., Calzini C. (2001) Genetic algorithms for the identification of additive and innovation outliers in time series, *Computational Statistics & Data Analysis*, 37, 1–12.
- Chang I., Tiao G. C., Chen C. (1988) Estimation of time series parameters in the presence

- of outliers, *Technometrics*, 30, 193–204.
- Chatterjee S., Laudato M. (1997) Genetic algorithms in statistics: procedures and applications, *Communications in Statistics - Theory and Methods*, 26(4), 1617–1630.
- Chatterjee S., Laudato M., Lynch L. A. (1996) Genetic algorithms and their statistical applications: an introduction, *Computational Statistics & Data Analysis*, 22, 633–651.
- Chipperfield, A., Fleming, P., Pohlheim, H., Fonseca, C. M. (1994) Genetic algorithm toolbox user's guide, ACSE Research Report No. 512, University of Sheffield, 1994.
- Friedman J. (1987) Exploratory projection pursuit, *Journal of the American Statistical Association*, 82, 249–266.
- Galeano P., Peña D., Tsay R. S. (2004) Outlier detection in multivariate time series via projection pursuit, Working Paper 04-42, Statistics and Econometrics Series 11, September 2004, Departamento de Estadística, Universidad Carlos III de Madrid.
- Gävert, H., Hurri, J., Särelä, J., Hyvärinen, A. (2005) The FastICA package for Matlab, Laboratory of Information and Computer Science in the Helsinki University of Technology.
- Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Massachusetts.
- Holland J. (1975) *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor.
- Huber P. J. (1985) Projection pursuit, *Annals of Statistics*, 13, 435–475.
- Hyvärinen A., Karhunen J., Oja E. (2001) *Independent Component Analysis*, Wiley, New York.
- Hyvärinen A., Oja E. (1997) A fast fixed-point algorithm for independent component analysis, *Neural Computation*, 9, 1483–1492.
- Jennison, C., Sheehan, N. (1995) Theoretical and empirical properties of the genetic algorithm as a numerical optimizer, *Journal of Computational and Graphical Statistics*, 4, 296–318.
- Maronna, R. A., Yohai, V. J. (1995) The behavior of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, 90, 330–341.
- Mitchell M. (1996) *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, Massachusetts.
- Peña D. (1990) Influential observations in time series, *Journal of Business & Economic Statistics*, 8, 235–241.
- Pirlot M. (1996) General local search methods, *European Journal of Operational Research*, 92, 493–511.
- Tsay R. S. (1986) Time series model specification in the presence of outliers, *Journal of the American Statistical Association*, 81, 132–141.
- 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.