Chemometrics and Intelligent Laboratory Systems 132 (2014) 103–110

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Outliers detection in multivariate time series using genetic algorithms

Domenico Cucina *, Antonietta di Salvatore, Mattheos K. Protopapas

Department of Statistics, Sapienza University of Rome, Piazzale Aldo Moro 5, I-00100 Roma, Italy

ARTICLE INFO

Article history:
Received 10 October 2012
Received in revised form 13 January 2014
Accepted 13 January 2014
Available online 21 January 2014

Keywords:
Genetic algorithms
Outlier detection
Outlier patches
Generalized AIC criterion
Vector ARMA model
Linear interpolator

ABSTRACT

A genetic algorithm to detect multiple additive outliers in multivariate time series is proposed. In contrast with many of the existing methods, it does not require to specify a vector ARMA model for the data and is able to detect any number of potential outliers simultaneously reducing possible masking and swamping effects. A generalized AIC-like criterion is used as objective function. The comparison and the performance of the proposed method are illustrated by simulation studies and real data analysis. Simulation results show that the proposed approach is able to handle patches of additive outliers.

1. Introduction

The treatment of outlier in a data set is important for many reasons. First, outliers can have a considerable influence on the results of an analysis (e.g., leading to model misspecification, biased parameter estimation and incorrect forecasting). It is therefore important to identify them prior to modeling and analysis of the data. For example, outlier detection is one of the most important steps for building high-quality univariate and multivariate calibration models [1,2]. Second, although outliers are often caused by measurement or recording errors, some of them can represent phenomena of interest, something significant from the viewpoint of the application domain (e.g., credit card fraud detection, clinical trials, network intrusion, severe weather prediction, geographic information systems). Third, for many applications, exceptions identified can often lead to the discovery of unexpected knowledge. Outliers are of primary interest when analysing chemical data. For example, in geochemical exploration they are indications for mineral deposits. In a manufacturing process, outliers may represent failure of the underlying mechanical system, materials of inferior quality, or unexpected experimental conditions and results. In environmental applications, outliers may represent highly contaminated areas. A series of papers on outlier detection were published in the chemometric literature. [3] proposed an outlier detection method for non-bilinear data, a situation in which more common detection techniques have difficulty. [4] proposed a procedure based on the PROP influence function that works effectively at identifying outliers in univariate as well as multivariate data sets of all sizes. [5] introduced a method for outlier detection in building linear regression for data with interval-bounded error. [1,2] showed that outliers incorporated into a multivariate calibration model can significantly reduce the performance of the model and propose robust statistical methods for the detection of outliers. Other useful references for the outlier detection are [6–8].

The chemist frequently encounters continuous processes often described as time series. Examples are continuous industrial processes where deviations from pre-set limits can result in poor quality of a product or sometimes industrial accidents. Time series occurring in geochemistry when measuring compounds down the core, in environmental chemistry when monitoring seasonal diurnal changes in composition, in clinical chemistry when monitoring biorhythm and finally when tracking reaction kinetics by methods such as stopped-flow. These time series are useful to provide a description of a dynamic system and to develop models for monitoring or controlling continuous process. The first step in building statistical process monitoring system for multivariable continuous processes is to develop an accurate process model. One example of an important chemical process control problem is wastewater treatment [9]. The empirical process model should be built from reliable data. The presence of just a few items of anomalous data can lead to model misspecification, biased parameter estimation (see [10]), and poor forecasts (see [11]). The presence of outliers in the variable of interest affects the reliability of the data which can result in erroneous interpretations concerning the variable of interest. For these reasons outlier-free time series data are essential to develop accurate models. Therefore, it is essential to identify them, estimate their magnitude and correct the time series and at the same time avoid false identifications (i.e. observations that are identified as outliers while they are not).

* Corresponding author. Tel.: +39 0649910657; fax: +39 064959241.
E-mail address: domenico.cucina@uniroma1.it (D. Cucina).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.007

Several approaches have been proposed in the literature for handling outliers in univariate time series. Among these methods we can distinguish those based on an explicit model (parametric approach) from those that use non-explicit models (nonparametric approach). For the parametric approach, [12] developed a likelihood ratio test for detecting outliers in a pure autoregressive model. [10,13–15,11] extended this test to an autoregressive integrated moving-average (ARIMA) model and proposed an iterative procedure for detecting multiple outliers.

For the non-parametric approach, [16–20] proposed specific procedures based on the relationship between the additive outliers and the linear interpolator, while [21] used a genetic algorithm.

Outliers can occur in several variables simultaneously caused by common-mode sources like instrumentation system failures (e.g., loss of a common power source for multiple instruments), communication system failures, or the simultaneous influence of a process upset on several measured variables. For multivariate time series, only three procedures have been proposed. [22] proposed a sequential detection procedure, which we will call the TPP method, based on individual and joint likelihood ratio statistics; this method requires an initial specification of a vector ARMA model. [23,24] proposed a method based on univariate outlier detection applied to some useful linear combinations of the vector time series. The optimal combinations are found by projection pursuit in the first paper and independent component analysis in the second one.

Multiple outliers, especially those occurring close in time, often have severe masking effect and smearing effect that can easily render outlier detection methods inefficient. Although there is no rigorous mathematical definition of masking and swamping effect, we report the definitions from [25]:

- **Masking effect**. An outlier masks a second one that is close by if the latter can be considered an outlier by itself, but not if it is considered along with the first one.
- **Swamping effect**. An outlier swamps another instance if the latter can be considered outlier only under the presence of the first one.

The term masking is due to [26] while the term swapping is defined in [27]. Several procedures for independent data are proposed for reducing masking and swamping effects (e.g., [1,28,2,29]). A special case of multiple outliers is a patch of additive outliers. For univariate time series this problem has been addressed firstly by [30]. They define a procedure for detecting outlier patches by examining blocks of consecutive observations. Other useful references for the patch detection are [31–33]. For multivariate time series, only [24] report simulation results for an outlier patch.

Unlike the univariate case where there are specific procedures on the identification of consecutive outliers, in multivariate time series framework, methods for identification of consecutive outliers do not exist.

In this paper we propose a genetic algorithm (GA) [34–36] for identifying multiple additive outliers in multivariate time series. The use of GAs for outlier detection seems attractive because several outliers may be processed simultaneously, in this way they are less vulnerable to the masking and smearing effects. Note that almost all available methods for outlier detection are iterative, but there is a crucial difference with GAs. In this latter case, any potential location may change through the iterations. In existing methods, once a location has been selected, it remains fixed in the subsequent iterations. So, the GAs seem able to provide more flexibility and adaptation to the outlier detection problem.

2. Theory

2.1. Genetic algorithms

Many optimization problems do not satisfy the necessary conditions to guarantee the convergence of traditional numerical methods. For instance, in order to apply standard gradient methods to maximum likelihood estimation we need a globally convex likelihood function, however there are a number of relevant cases with non convex likelihood functions or functions with several local optima. Another class of "hard" problems is when the solution space is discrete and large. These problems are known as combinatorial problems. A simple approach for solving an instance of a combinatorial problem is to list all the feasible solutions, evaluate their objective function, and pick the best. However, for a combinatorial problem of a reasonable size, the complete enumeration of its elements is too computationally expensive, and most available searching algorithms are likely to yield some local optimum as a result [37].

GAs are often used to solve such problem instances. GAs do not rely on a set of strong assumptions about the optimization problem, on the contrary, they are robust to changes in the characteristics of the problem. On the other hand, they do not produce a deterministic solution but a high quality stochastic approximation to the global optimum.

GAs, inspired by [34], imitate the evolution process of biological systems, to optimize a given function. GAs use a set of candidate solutions, called population, instead of one single current solution. In GAs terminology, any candidate solution is encoded via a numerical vector called chromosome. The GAs proceed by updating the population of active chromosomes (the sets of current candidate solutions) in rounds, called generations. In each generation, some of the active chromosomes are selected (parents-chromosomes) to form the chromosomes of the next generation (children-chromosomes). The selection process is based on an evaluation measure called fitness function, linked to the objective function, that assigns to each chromosome a positive number. This fitness is the determining factor for calculating the probability to select a chromosome as a parent. A higher fitness value leads to higher probability that the corresponding chromosome will be one of the parents used to form the children-chromosomes. Children are formed by recombining (crossover) randomly the genetic material of their two parents-chromosomes and perhaps after a random alteration of some of the genes (single digits of the chromosome) which is called mutation (see [34,35], for a detailed description).

Several papers on genetic algorithms in chemometrics have been published recently (e.g., [38–41]). Two articles deal with the problem of outliers [42,43]. [36] and [44]reported a wide review of applications of GAs to chemometric problems.

2.2. Solution encoding

Each solution ξ^c is a binary string with length N, where N is the number of observations of the time series: $\xi^c = (\xi_1^c, \xi_2^c, \ldots, \xi_N^c)$, where ξ_i^c takes the value 1 if at time i there is an outlier and 0 otherwise. Each chromosome is composed of N bits, therefore their storage space is not large also when N is much large. Obviously, the number of outliers for a given time series is unknown. We allow solutions with a maximum number of outliers equal to g. The value of g should be chosen according to the series length and every relevant a priori on its accuracy and instability. The constant g should be chosen large enough to allow detection of any reasonable number of outliers in the series.

Binary encoding implies that the solution space Ω consists of $\sum_{k=0}^{g} \binom{N}{k}$ distinct elements, since the total number of outliers is limited to a constant g. We can see that Ω is really large even if g is considerably lower than the length of the time series. For example, with $g = 5$, the solution space Ω is of order 2×10^9 when the sample size is $N = 200$, and it is of order 8×10^{10} when the sample size is $N = 400$. If we increase the value of g or N it seems reasonable to increase also the generation number of the GA because a larger solution space has to be explored.

2.3. Fitness function

Let $\mathbf{y}_t = [y_{1,t}, \ldots, y_{s,t}]'$ be a Gaussian s-dimensional jointly second order stationary real-valued vector time series, with mean zero for