Computational Statistics and Data Analysis 128 (2018) 184–199

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

ICS for multivariate outlier detection with application to quality control[☆]

Aurore Archimbaud [a], Klaus Nordhausen [b], Anne Ruiz-Gazen [a,*]

[a] Toulouse School of Economics, University of Toulouse 1 Capitole, 21 allée de Brienne, 31015 Toulouse cedex 6, France
[b] CSTAT - Computational Statistics, Institute of Statistics & Mathematical Methods in Economics Vienna University of Technology, Wiedner Hauptstr. 7, A-1040 Vienna, Austria

HIGHLIGHTS

- Detecting automatically multivariate outliers in high reliability standards fields.
- Combining the advantages of Mahalanobis distance and Principal Component Analysis.
- Simple and efficient procedure in the context of a small proportion of outliers.
- Reducing the number of false positives compared to competitors.
- An R package available: ICSOutlier.

ARTICLE INFO

Article history:
Received 25 December 2017
Received in revised form 23 June 2018
Accepted 26 June 2018
Available online 17 July 2018

Keywords:
Affine invariance
Mahalanobis distance
Principal component analysis
Scatter estimators
Unsupervised outlier identification

ABSTRACT

In high reliability standards fields such as automotive, avionics or aerospace, the detection of anomalies is crucial. An efficient methodology for automatically detecting multivariate outliers is introduced. It takes advantage of the remarkable properties of the Invariant Coordinate Selection (ICS) method which leads to an affine invariant coordinate system in which the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original coordinates. The limitations of MD are highlighted using theoretical arguments in a context where the dimension of the data is large. Owing to the resulting dimension reduction, ICS is expected to improve the power of outlier detection rules such as MD-based criteria. The paper includes practical guidelines for using ICS in the context of a small proportion of outliers. The use of the regular covariance matrix and the so called matrix of fourth moments as the scatter pair is recommended. This choice combines the simplicity of implementation together with the possibility to derive theoretical results. The selection of relevant invariant components through parallel analysis and normality tests is addressed. A simulation study confirms the good properties of the proposal and provides a comparison with Principal Component Analysis and MD. The performance of the proposal is also evaluated on two real data sets using a user-friendly R package accompanying the paper.

[☆] Supplementary material is provided. It contains some scatterplot matrices to visualize the six simulated data sets and the R code to generate these data sets. It also includes the R code to reproduce the results of Table 4 for the Reliability data and the HTP data sets.

[*] Corresponding author.

E-mail addresses: aurore.archimbaud@tse-fr.eu (A. Archimbaud), klaus.nordhausen@tuwien.ac.at (K. Nordhausen), anne.ruiz-gazen@tse-fr.eu (A. Ruiz-Gazen).

https://doi.org/10.1016/j.csda.2018.06.011
0167-9473/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Detecting outliers in multivariate data sets is of particular interest in many industrial, medical and financial applications (Aggarwal, 2017). Some classical statistical detection methods are based on the Mahalanobis distance and its robust counterparts (see e.g. Rousseeuw and Van Zomeren (1990), Cerioli et al. (2009) and Cerioli (2010)) or on robust principal component analysis (see e.g Hubert et al. (2005)). One advantage of the Mahalanobis distance is its affine invariance while Principal Component Analysis (PCA) is only invariant under orthogonal transformations. For its part, PCA allows some components selection and facilitates the interpretation of the detected outliers. All these methods are adapted to the context of casewise contamination while other methods are adapted to the case of cellwise contamination (see e.g. Agostinelli et al. (2015) and Rousseeuw and Bossche (2018)). Furthermore, several other recent references tackle the problem of outlier detection in high dimension where the number of observations may be smaller than the number of variables (see e.g. Croux et al. (2013) and Hubert et al. (2016)).

In the present paper, we propose an alternative to the Mahalanobis distance and to PCA, in a casewise contamination context and when the number of observations is larger than the number of variables. The focus is on applications with high level of quality control, such as in the automotive, avionics or aerospace fields, where only a small proportion of outliers, up to 2%, is plausible. From our experience in such application fields, a small proportion of parts potentially defective are to be detected with very limited false detection.

The method we consider is the Invariant Coordinate Selection (ICS) as proposed by Tyler et al. (2009). The principle of ICS is quite similar to Principal Component Analysis (PCA) with coordinates or components derived from an eigendecomposition followed by a projection of the data on selected eigenvectors. However, ICS differs in many respects from PCA. It relies on the simultaneous spectral decomposition of two scatter matrices instead of one for PCA. While principal components are orthogonally invariant but scale dependent, the invariant components are affine invariant for affine equivariant scatter matrices. Moreover, under some elliptical mixture models, the Fisher's linear discriminant subspace coincides with a subset of invariant components in the case where group identifications are unknown (see Theorem 4 in Tyler et al. (2009)). This remarkable property is of interest for outlier detection since outliers can be viewed as data observations that differ from the remaining data and form separate clusters.

Despite its attractive properties, ICS has not been extensively studied in the literature on outlier detection. An early version of ICS was proposed in Caussinus and Ruiz (1990) for multivariate outlier detection and studied further in e.g. Penny and Jolliffe (1999) and Caussinus et al. (2003) for two specific scatter matrices. Recent articles by Nordhausen et al. (2008) and Tyler et al. (2009) argue that ICS is useful for outlier detection. However, a thorough evaluation of ICS in this context is still missing and the present paper is a first step aimed at filling the gap.

This article is organized as follows. In Section 2 we observe the behavior of the usual and the robust Mahalanobis distances for large dimensions when the outlierness structure lies in a small dimensional subspace. This result motivates the use of some selected invariant components for outlier detection. ICS is described in a general framework in Section 3 and in the context of a small proportion of outliers in Section 4. Section 5 provides results from a simulation study and derives practical guidelines for the choice of the scatter matrices pair and the components selection method. Comparisons with the Mahalanobis distance and PCA are also provided. Two real data sets are analyzed in Section 6. The second example in particular illustrates the use of the selected invariant components to characterize the detected outliers. Following the same principle as PCA, one can calculate the correlations of the selected invariant components with the initial variables, and look at the largest correlations in order to find out and plot the variables responsible for the outlierness behavior. Finally, conclusions and perspectives are drawn in Section 7. The proof of Proposition 1 is given in Appendix A.

2. Behavior of the Mahalanobis distance in large dimension

There exist several papers using the (robust) Mahalanobis distance for outlier detection. However, the present section illustrates the fact that in high dimensions, one needs to look at alternatives. Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a p-multivariate real random vector and assume the distribution of $\mathbf{X}$ is a mixture of $(q+1)$ Gaussian distributions with $q+1 < p$, different location parameters $\boldsymbol{\mu}_h$, for $h = 0, \ldots, q$, and the same definite positive covariance matrix $\boldsymbol{\Sigma}_W$:

$$\mathbf{X} \sim (1-\epsilon)\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W) + \sum_{h=1}^{q} \epsilon_h \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W) \tag{1}$$

where $\epsilon = \sum_{h=1}^{q}\epsilon_h < 1/2$.

Such a distribution can be interpreted as a model for outliers where the majority of the data follows a given Gaussian distribution and outliers are clustered in q clusters with Gaussian distributions with different locations than the majority group. This model is a generalization of the well-known mean-shift outlier model to more than two groups.

For such a model, the mean is $\boldsymbol{\mu}_{\mathbf{X}} = (1-\epsilon)\,\boldsymbol{\mu}_0 + \sum_{h=1}^{q}\epsilon_h\boldsymbol{\mu}_h$, the within covariance matrix is $\boldsymbol{\Sigma}_W$, the between covariance is $\boldsymbol{\Sigma}_B = (1-\epsilon)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})' + \sum_{h=1}^{q}\epsilon_h(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})'$, where the prime symbol denotes the transpose vector or matrix, and the total covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$. Let us consider the following squared Mahalanobis distances:

$$d^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}), \tag{2}$$

$$d_R^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_0)'{\boldsymbol{\Sigma}_W}^{-1}(\mathbf{X} - \boldsymbol{\mu}_0). \tag{3}$$