SPARSE AND ROBUST ESTIMATORS FOR OUTLIER DETECTION IN DISTRIBUTIONAL DATA

Pedro Duarte Silva [1], Peter Filzmoser [2] and Paula Brito [3]

[1] Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Porto, Portugal, (e-mail: psilva@ucp.pt)

[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria, (e-mail: peter.filzmoser@tuwien.ac.at)

[3] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal, (e-mail: mpbrito@fep.up.pt)

ABSTRACT: The classical data representation model is too restrictive when the data to be analysed are not real numbers but comprise variability. In this talk, we are interested in numerical distributional data, where units are described by histogram or interval-valued variables. We consider parametric probabilistic models, which are based on the representation of each distribution by a location measure and interquantile ranges. A multivariate outlier detection method is proposed that makes use of restricted configurations for the covariance matrix, and is based on a sparse robust estimator of its inverse. The computations rely on an efficient adaptation of the graphical lasso algorithm. A simulation study puts in evidence the usefulness of the robust estimates for outlier detection.

KEYWORDS: outliers, robust statistics, distributional data, Mahalanobis distance, graphical lasso

1 Introduction

Multivariate datasets often include atypical data points known as *outliers*, i.e. points that deviate from the main pattern. Outlier detection is important because outlying data points may reveal nonconforming phenomena and the results of usual multivariate methods can be heavily influenced by them.

In this paper we address the problem of outlier detection in multivariate distributional data. Distributional data may result from the aggregation of large amounts of open/collected/generated data, or may be directly available in a structured or unstructured form, describing the variability of some features. In recent years, different approaches have been investigated and methods proposed for the analysis of such data. However, most existing methods rely on non-parametric descriptive approaches.

A common approach for multivariate outlier detection measures outlyingness by Mahalanobis distances. Given a sample of n observations, a point i is considered an outlier if its distance $D^2_{\hat{\mu},\hat{\Theta}}(i)$ from an appropriate mean estimate, $\hat{\mu}$, is above a relevant threshold. Here, $\hat{\Theta}$ is an estimate of the precision matrix, $\Theta = \Sigma^{-1}$, and Σ denotes the population covariance. However, if $\hat{\mu}$ and $\hat{\Theta}$ are chosen to be the classical sample mean vector and inverse covariance matrix, S^{-1}, this procedure is not reliable, as $D^2_{\hat{\mu},\hat{\Theta}}(i)$ may be strongly affected by atypical observations. Furthermore, S^{-1} has a large sample variability when its dimension, d, is close to n, and it is is not even computable when $d > n$. To address these issues Öllerer and Croux (Öllerer & Croux, 2015), proposed sparse precision matrix estimators based on the GLASSO L_1-penalized log-likelihood function (Friedman *et al.* , 2008).

In this paper we address the problem of outlier detection in distributional data, combining Öllerer and Croux estimators with a parametric modelling of distributional data, along the lines of Brito & Duarte Silva, 2012, and Duarte Silva *et al.* , 2018.

2 Distributional Variables

Let $S = \{s_1, \ldots, s_n\}$, be the set of n units under analysis. We consider that for each unit, the descriptive variables are (in general) not constant, but present variability.

We represent the "values" of a numerical distributional variable by an ordered vector of quantiles, always including the minimum and the maximum. Formally, a numerical distributional variable is defined by an application

$$\begin{gathered} Y : S \to T \\ s_i \to Y(s_i) = (Min_i, \psi_{1i}, \ldots, \psi_{qi}, Max_i) \end{gathered}$$

Let $Y_1, \ldots, Y_p$ be the p numerical distributional variables, defined on S. Here we assume that all variables are represented by the same set of $q+2$ quantiles, and that $Min_{ij} < \psi_{1ij} < \ldots < \psi_{qij} < Max_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq p$ (strict inequalities).

The model consists in representing $Y_j(s_i)$ by

- a central statistic C_{ij}, typically the Median Med_{ij} or the MidPoint $\frac{Max_{ij}+Min_{ij}}{2}$
- the $[Min, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - Min_{ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- ...

- the $[\psi_q, Max[$ range: $R_{mij} = Max_{ij} - \psi_{qij}$

Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles.

The proposed model consists in assuming that the joint distribution of the central statistic C and the logarithms of the ranges $R^*_\ell, \ell = 1, \ldots, m$, is Gaussian:

$(C, R^*_1, \ldots, R^*_m) \sim N_{(m+1)p}(\mu, \Sigma)$

In the most general formulation (configuration 1) we allow for non-zero correlations among all central statistics and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables Y_j are non-correlated, but for each variable, the central statistic and all its log-ranges may be correlated among themselves (configuration 2); central statistics (respectively, log-ranges) of different variables may be correlated, but no correlation between central statistics and log-ranges is allowed (configuration 3); central statistics (respectively, each log-range) of different variables may be correlated, but no correlation between central statistics and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all central statistics and log-ranges are non-correlated (configuration 5).

3 Outlier Detection of Distributional Data

Let $X_i = \left[C_i^t, R^{*\,t}_{1i}, \ldots, R^{*\,t}_{mi}\right]^t$ be the $d = (m+1)p$ dimensional column vector comprising all central statistics and log-ranges for s_i, $i = 1, \ldots, n$.

The identification of outliers is based on robust Mahalanobis distances, $D^2_{\hat{\mu},\hat{\Theta}}(i) = (x_i - \hat{\mu})^t \hat{\Theta}(x_i - \hat{\mu})$ from each data point to a robust location vector, $\hat{\mu}$, which are then compared with the 97.5% quantile of a chi-squared distribution with d-degrees of freedom. In our approach we choose as location vector, the L_1 median (Fritz *et al.* , 2012), which has a break-down point of 0.5 and, given our Gaussian assumption, is a robust estimator of μ.

Following Öllerer and Croux (2015) we estimate $\Theta = \Sigma^{-1}$ by

$$\hat{\Theta} = argmax_{\Theta \in \vartheta} \log \det(\Theta) - tr(\hat{\Sigma}\Theta) - \rho \sum_{j,k=1}^{d} |(\Theta)_{jk}| \qquad (1)$$

where $\vartheta := \{\Theta \in \mathbb{R}^{d \times d} : \Theta \succ 0\}$ is the space of d-dimensional positive-definite matrices, $\hat{\Sigma}$ is a robust covariance estimate, and ρ a regularization parameter.

For each covariance configuration, we set the null elements of Σ to zero in its initial $\hat{\Sigma}$ estimate, and for the remaining elements we use the formula

$$\hat{\Sigma}_{j,k} = scale(X^j)\, scale(X^k)\, r(X^j, X^k) \tag{2}$$

where X^j, X^k are the j^{th} and k^{th} columns of X, $scale(X^j), scale(X^k)$ are robust scale estimators (see Rousseeuw & Croux, 1993), and $r(X^j, X^k)$ is the Gaussian rank correlation (Boudt *et al.* , 2012) between X^j and X^k.

The above procedure was evaluated in a controlled simulation experiment that showed promising results for the proposed approach.

References

BOUDT, KRIS, CORNELISSEN, JONATHAN, & CROUX, CHRISTOPHE. 2012. The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, **22**, 471–483.

BRITO, PAULA, & DUARTE SILVA, A. PEDRO. 2012. Modelling Interval Data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, **39**(1), 3–20.

DUARTE SILVA, A. PEDRO, FILZMOSER, P., & BRITO, PAULA. 2018. Outlier detection in Interval Data. *Advances in Data Analysis and Classification*, **12**(3), 785–822.

FRIEDMAN, JEROME, HASTIE, TREVOR, & TIBSHIRANI, ROBERT. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

FRITZ, HEINRICH, FILZMOSER, PETER, & CROUX, CHRISTOPHE. 2012. A comparison of algorithms for the multivariate L1-median. *Computational Statistics*, **27**, 393–410.

ÖLLERER, VIKTORIA, & CROUX, CHRISTOPHE. 2015. Robust high-dimensional precision matrix estimation. *Pages 325–350 of: Modern Nonparametric, Robust and Multivariate Methods*. Springer.

ROUSSEEUW, PETER J, & CROUX, CHRISTOPHE. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, **88**(424), 1273–1283.