CLUSTERING THREE-WAY DATA WITH OUTLIERS

Katharine M. Clark [1] and Paul D. McNicholas[1]

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada (e-mail: clarkkm2@mcmaster.ca, paul@math.mcmaster.ca)

ABSTRACT: An approach for clustering three-way data is discussed. The approach, which is based on mixtures of matrix-variate distributions, uses an iterative subset log-likelihood approach to detect and trim outliers.

KEYWORDS: clustering, matrix-variate, mixture models, outliers, three-way data.

1 Introduction

Grubbs (1969) describes an outlier as an observation "that appears to deviate markedly from other members of the sample in which it occurs." Outliers, and their treatment, is a long-studied topic in the field of applied statistics. The problem of handling outliers in multivariate clustering has been studied in several contexts including work by García-Escudero *et al.* (2008), Punzo & McNicholas (2016), Punzo *et al.* (2020), and Clark & McNicholas (2023). The approach of Clark & McNicholas (2023) is extended to the matrix-variate paradigm, i.e., to account for three-way data such as multivariate longitudinal data. The OCLUST algorithm introduced in Clark & McNicholas (2023), and supported by the R package `oclust` (Clark & McNicholas, 2022), is based on the mixture model-based clustering framework (see, e.g., McNicholas, 2016) and uses an iterative subset log-likelihood approach to detect and trim outliers. An analogue of the OCLUST algorithm is developed for three-way data.

2 Background

The density of a finite mixture model is $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$, where $\boldsymbol{\vartheta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_G\}$, $\pi_g > 0$ is the gth mixing proportion with $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density with parameters $\boldsymbol{\theta}_g$. Most (mixture) model-based clustering methods assume, either explicitly or implicitly, that the data are free of outliers. Outlier algorithms in (multivariate) model-based clustering usually fall into either one of two paradigms: outlier-inclusion and outlier trimming. Focusing on the latter, Cuesta-Albertos *et al.* (1997)

developed an impartial trimming approach for k-means clustering; however, this method maintains the drawbacks of k-means clustering, where the clusters are spherical with equal — or, in practice, similar — radii. García-Escudero *et al.* (2008) improved upon trimmed k-means with the TCLUST algorithm. TCLUST places a restriction on the eigenvalue ratio of the covariance matrix, as well as implementing a weight on the clusters, allowing for clusters of various elliptical shapes and sizes. An obvious challenge with these methods is that the eigenvalue ratio must also be known *a priori*. There exists an estimation scheme for the proportion of outliers but it is heavily influenced by the choices for number of clusters and eigenvalue ratio.

The OCLUST algorithm (Clark & McNicholas, 2023) uses the fact that the Mahalanobis distance is χ^2_p for p-dimensional multivariate normal data (Mardia *et al.*, 1979) to derive the distribution of subset log-likelihoods for clustering multivariate normal data. A subset log-likelihood is considered to be the log-likelihood of a model fitted with $n-1$ of the data points. There are n such subsets. The OCLUST algorithm uses the subset log-likelihoods and their distribution to identify and trim outliers.

Two-way data can be regarded as the observation of n vectors, whereas three-way data can be considered the observation of n matrices. Mixtures of matrix-variate distributions have been used to cluster three-way data (e.g., Viroli, 2011; Anderlucci & Viroli, 2015; Gallaugher & McNicholas, 2018). An $r \times c$ random matrix $\mathcal{X}$ comes from a matrix-variate normal distribution if its density is of the form

$$\phi_{r\times c}(\boldsymbol{X} \mid \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{U}) = \frac{1}{(2\pi)^{\frac{rc}{2}} |\boldsymbol{V}|^{\frac{r}{2}} |\boldsymbol{U}|^{\frac{c}{2}}} \exp\left\{ -\frac{1}{2} \text{tr}\big(\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M})^\top \boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{M})\big) \right\}, \tag{1}$$

where $\boldsymbol{M}$ is the $r \times c$ mean matrix, $\boldsymbol{U}$ is the $r \times r$ row covariance matrix, and $\boldsymbol{V}$ is the $c \times c$ column covariance matrix. Note that there is an identifiability issue with regard to the parameters $\boldsymbol{U}$ and $\boldsymbol{V}$, i.e., if k is a strictly positive constant, then replacing $\boldsymbol{U}$ and $\boldsymbol{V}$ by $(1/k)\boldsymbol{U}$ and $k\boldsymbol{V}$, respectively, leaves (1) unchanged. Various different solutions have been proposed to resolve this issue, including setting $\text{tr}(\boldsymbol{U}) = r$ or $\boldsymbol{U}_{11} = 1$.

For multivariate normal data, the Mahalanobis distance can be expressed as $\mathcal{D}(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$. Pocuca *et al.* (2023) derive a similar expression for matrix-variate normal data:

$$\mathcal{D}_M(\boldsymbol{X}_i, \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{U}) = \text{tr}\left\{ \boldsymbol{U}^{-1}(\boldsymbol{X}_i - \boldsymbol{M})\boldsymbol{V}^{-1}(\boldsymbol{X}_i - \boldsymbol{M})^\top \right\}, \tag{2}$$

and prove that if a Kronecker product structure exists for $\mathbf{\Sigma}$, then

$$\mathcal{D}_M(\boldsymbol{X}_i, \hat{\boldsymbol{M}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}) \xrightarrow{P} \mathcal{D}_M(\boldsymbol{X}_i, \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}), \tag{3}$$

where $\xrightarrow{P}$ denotes convergence in probability.

3 Methodology

As in the multivariate case, consider a subset log-likelihood in the matrix-variate case to be the log-likelihood of a model fitted with $n-1$ of the data points. Formally, if we denote our complete dataset as $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, then the jth subset is defined as the complete dataset with the jth point removed, i.e., $\mathcal{X} \setminus \boldsymbol{X}_j = \{\boldsymbol{X}, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_n\}$. Analogous to the multivariate case, treat point $\boldsymbol{X}_k$, whose absence produced the largest subset log-likelihood, as our candidate outlier, ie.

Definition 1 (Candidate Outlier)**.** *We define our candidate outlier as $\boldsymbol{X}_k$, where*

$$k = \arg\max_{j \in [1,n]} \ell_{\mathcal{X} \setminus \boldsymbol{X}_j},$$

and $\ell_{\mathcal{X} \setminus \boldsymbol{X}_j}$ is the log-likelihood of the subset model with the jth point removed.

Remove candidate outliers one-by-one until we obtain our best model, which is determined by the distribution of our subset log-likelihoods, stated in Proposition 1.

Proposition 1. *For a point $\boldsymbol{X}_j$ belonging to the hth cluster, if $Q_{\mathcal{X}}$ is a simplified log-likelihood and $Y_j = Q_{\mathcal{X} \setminus \boldsymbol{X}_j} - Q_{\mathcal{X}}$, then Y_j has an approximate shifted gamma density*

$$Y_j \sim f_{gamma}\left(y_j - k \;\middle|\; \alpha = \frac{p}{2}, 1\right), \tag{4}$$

for $y_j - k \geq 0, \alpha > 0$, where

$$k = -\log \pi_h + \frac{rc}{2}\log(2\pi) + \frac{c}{2}\log|\boldsymbol{U}_h| + \frac{r}{2}\log|\boldsymbol{V}_h|,$$

n_h is the number of points in cluster h, and $\pi_h = n_h/n$.

The mathematical results for this proposition will be given in the full paper, along with other technical details as well as illustrations via real and simulated data.

References

ANDERLUCCI, L., & VIROLI, C.. 2015. Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, **9**(2), 777–800.

CLARK, K. M., & MCNICHOLAS, P. D. 2022. *oclust: Gaussian model-based clustering with outliers*. R package version 0.2.0.

CLARK, K. M., & MCNICHOLAS, P. D. 2023. *Using subset log-likelihoods to trim outliers in Gaussian mixture models*. arXiv preprint arXiv:1907.01136v4.

CUESTA-ALBERTOS, J. A., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed *k*-means: an attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

GALLAUGHER, M. P. B., & MCNICHOLAS, P. D. 2018. Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83 – 93.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A General Trimming Approach to Robust Cluster Analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

GRUBBS, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, **11**(1), 1–21.

MARDIA, K. V., KENT, J. T., & BIBBY, J. M. 1979. *Multivariate Analysis*. London: Academic Press.

MCNICHOLAS, P. D. 2016. Model-Based Clustering. *Journal of Classification*, **33**(3), 331–373.

POCUCA, N., GALLAUGHER, M. P. B., CLARK, K. M., & MCNICHOLAS, P. D. 2023. Assessing and visualizing matrix-variate normality. *Australian and New Zealand Journal of Statistics*. In press

PUNZO, A., BLOSTEIN, M., & MCNICHOLAS, P. D. 2020. High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition*, **98**, 107031.

PUNZO, A., & MCNICHOLAS, P. D. 2016. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

VIROLI, C. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and computing*, **21**(4), 511–522.