Robust clustering and outlier rejection using the Mahalanobis distance distribution

Violeta Roizman
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
violeta.roizman@centralesupelec.fr

Matthieu Jonckheere
Instituto de Cálculo, CONICET, Universidad de Buenos Aires, Buenos Aires, Argentina.
mjonckhe@dm.uba.ar

Frédéric Pascal
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
frederic.pascal@centralesupelec.fr

***Abstract*—Both clustering and outlier detection tasks have a wide range of applications in signal processing. We focus here on the case where the data is corrupted with outliers and samples are relatively small. We study approximations of the distribution of the Mahalanobis distance when using robust estimators for the mean and the scatter matrix. We develop clustering and outlier rejection methods in the context of robust mixture modelling. We leverage on robust clustering and parameter estimations on a portion of the data, and we perform outlier detection on the rest of the data. We illustrate the importance of our method with synthetic simulations where we compare the theoretical asymptotic distribution and an approximated distribution to the empirical distribution. We conclude with an application using the well-known data set MNIST contaminated with noise.**

***Index Terms*—clustering, outlier rejection, Mahalanobis distance, robust estimation**

I. Introduction

The clustering task consists in finding coherent groups in a data set. Given a notion of distance or similarity, we expect similar points to be in the same cluster. This task has became of great importance in the recent years due to the huge increase of unlabeled data available. Clustering applications are numberless and include brain signal processing to recognize motor task [1], land use detection in radar images [2], community segmentation in social networks and image color quantization [3], just to name a few examples. In this paper, we address the clustering problem by possibly dividing the data into two parts, building an initial clustering on the first chunk and classifying the remaining part of the data. Of course, this is also applicable to the case where new independent data is available at a later stage and has to to be labelled once a main clustering is available. We refer to the latter task as a classification problem.

We follow a model-based approach and we build the clustering task on recent adaptations of the well-known Expectation-Maximization (EM) algorithm to the case of diverse and non-Gaussian data, as explained below. Besides, we use precise distributional properties of the Mahalanobis distance adapted to small samples in order to detect outliers.

The EM algorithm in the case of a Gaussian Mixture Model (GMM) models the data as drawn from a mixture of Gaussian distributions and each cluster is associated to one of the distributions of the mixture. The EM is an iterative algorithm that alternates until convergence between two steps. In the E-step, the membership of the data points are computed based on the *a posteriori* probability to belong to each cluster. Subsequently, the M-step is where the estimators of the model parameters are updated. To estimate the parameters of the GMM, closed-form equations are derived from the expectation of the likelihood of the model. Once the original data is clustered, the fitted model can be used to easily classify additional data points by computing the probability of the new data to belong to each of the found clusters.

The EM clustering algorithm (in the GMM case) suffers from noise and outliers [4]. In order to solve these drawbacks, several modifications of the algorithm have been proposed. Mainly two approaches appeared, either modelling the noise or using usual robust estimation techniques. The first category includes among others the mixture of Student's t-distributions where outliers are allowed in the model as part of the heavy tails of the distributions [5]. The second one includes methods that replace the classic sample estimators by robust versions. For example, the trimming TCLUST algorithm leaves out of the estimation a percentage of the data that lie far from the estimated centers of the clusters [6]. We use here the clustering algorithm presented in [7], that uses a semi-parametric paradigm to estimate an unknown scale parameter per data point and allowing the algorithm to accommodate for heavier tails distributions.

The Mahalanobis distance is a well-known measure in multivariate statistics and signal processing. For example, it appears in the density function of the multivariate Gaussian distribution and it has been used in the context of clustering and classification [8]. It measures the distance of a set of data points to a center by taking into account the dispersion of the data around that center. If the data follow a Gaussian distribution, then the distribution of the distance is known to follow a chi-square distribution. The knowledge of the distribution of this distance allows to design statistical tests to recognize outliers. However, in the clustering application, the real center and the true covariance matrix of the distribution of each cluster are unknown. A crucial observation is that as we replace the real values of the parameters by their estimators, the exact distribution of the Mahalanobis distance

changes. In this case, we can usually prove that the asymptotic distribution is chi-square. However, if the number of observations is small the true distribution of the distance can be significantly different and, in consequence, the outlier rejection based on the asymptotic distribution can be misleading. For example, in [9] the authors study the distribution of a robust variant of the Mahalanobis distance, when using the Minimum Covariance Determinant (MCD) estimators. Even if they do not derive the exact distribution of the distance, they present two practical approaches that widely outperform the chi-square approximation.

Our objective here is to design an outlier rejection method in the context of clustering data based on Tyler's estimators. In order to do that, we study the distribution of the robust version of the Mahalanobis distance when replacing the true values by the robust estimators. When dealing with data points that are independent of the data used to build the robust clustering, this distribution can be well approximated by a Fisher distribution.

The rest of the paper is organized as follows. In Section II we present the clustering algorithm and the background on the distribution of the Mahalanobis distance and its robust variations. In Section III we describe the proposed outlier rejection methods. Section IV contains the experiments on the Mahalanobis distance distribution and outlier rejection. Finally, we discuss our conclusions in Section V.

Notation: Vectors (resp. matrices) are denoted by boldfaced lowercase letters (resp. uppercase letters). $\mathbf{A}^T$ represents the transpose of $\mathbf{A}$, $|\mathbf{A}|$ represents the determinant of $\mathbf{A}$ and tr($\mathbf{A}$) represents the trace of $\mathbf{A}$.

II. Theoretical Background

Given a data set, we want to cluster it and detect its multivariate outliers. In order to do that, we first apply the robust clustering presented in Section II-A. Once the mixture model is fitted, we reject outliers based on theoretical results for the distribution of the Mahalanobis distance and its variants that we develop in Section II-B.

A. Robust mixture model

We consider as model a mixture of K distributions with the following density:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_{\theta_k}(\mathbf{x}) \quad \text{with} \quad \sum_{k=1}^{K} \pi_k = 1, \tag{1}$$

where π_k represents the proportion of the $k^{\underline{\text{th}}}$ distribution in the mixture and f_{θ_k} its probability density function (pdf) depending on the parameters θ_k that characterize the distribution.

We restrict the possible distribution of the clusters to the compound-Gaussian class. This class is included in the wide family of Elliptically Symmetric distributions [10]. A distribution within this family is an affine transformation of a Gaussian process that is represented as follows

$$\mathbf{x}_i = \boldsymbol{\mu}_k + \sqrt{\tilde{\tau}_{ik}}\mathbf{A}_k\mathbf{g}_i, \tag{2}$$

where $\boldsymbol{\mu}_k$ corresponds to the mean, $\tilde{\tau}_{ik}$ is a positive random variable independent from $\mathbf{g}_i$, $\mathbf{g}_i \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{A}_k\mathbf{A}_k^T = \boldsymbol{\Sigma}_k$. A one-dimensional constraint on $\boldsymbol{\Sigma}_k$ (usually on the trace or the determinant) is assumed to avoid identifiability issues between $\tilde{\tau}_{ik}$ and $\boldsymbol{\Sigma}_k$. In this work, we assume $tr(\boldsymbol{\Sigma}_k) = m$ (See [11] for more details). We are interested in particular in the case where the pdf of $\tilde{\tau}_{ik}$ can be different from one cluster to another (which justifies the extra k index in the notation). In order to conserve a flexible model, we do not fix a particular distribution for $\tilde{\tau}_{ik}$ in Eq. (2). Instead, we assume as in [11] that each data point $\mathbf{x}_i \in \mathbb{R}^m$, from the cluster k of the mixture, is the result of multiplying a sample from a multivariate zero-mean Gaussian by a deterministic constant and a consequent translation. In consequence, each observation can be written as

$$\mathbf{x}_i = \boldsymbol{\mu}_k + \sqrt{\tau_{ik}}\mathbf{A}_k\,\mathbf{g}_i,$$

with the mean $\boldsymbol{\mu}_k$, τ_{ik} a deterministic constant or parameter, $\mathbf{g}_i \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{A}_k\mathbf{A}_k^T = \boldsymbol{\Sigma}_k$ with tr($\boldsymbol{\Sigma}_k$) $= m$.

Consequently, the data model considered for developing the clustering algorithm is the mixture model defined in Eq. (1) where $\theta_k = (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \{\tau_{ik}\}_{i=1,..,N})$ for $k = 1, ..., K$. When comparing to a classical EM algorithm of the GMM, this data modelling requires the estimation of all the τ_{ik}'s parameters. The derivation, when including τ_{ik} to the set of parameters, leads to different estimators for the mean and the scatter matrix that are expressed in terms of fixed-point equations. The resulting linked fixed-point equations for both estimators of the $k^{\underline{\text{th}}}$ cluster are the following:

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} \frac{p_{ik}\mathbf{x}_i}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}}{\sum_{i=1}^{n} \frac{p_{ik}}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}},$$

and

$$\widehat{\boldsymbol{\Sigma}}_k = m\sum_{i=1}^{n} \frac{w_{ik}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)},$$

with p_{ik} the probability of the data point $\mathbf{x}_i$ to belong to the $k^{\underline{\text{th}}}$ cluster,

$$w_{ik} = p_{ik} / \sum_{l=1}^{n} p_{lk},$$

and

$$\widehat{\tau}_{ik} = \frac{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}{m}.$$

The derivation of $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\boldsymbol{\Sigma}}_k$ is decoupled from $\widehat{\tau}_{ik}$ as in [12]. The solutions of these two equations are very similar to Tyler's estimators [13] with weights representing the membership of each data point to the corresponding cluster and with a different exponent in the Mahalanobis distance. The derivation of the algorithm can be found in [7], as well as comparisons of performance with other clustering algorithms.

B. The Mahalanobis distance and its variants

The Mahalanobis distance is a well-known measure in multivariate statistics. It measures the distance of a set of data points to a center by taking into account the dispersion of the data around that center. The squared Mahalanobis distance is defined as

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

If $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then it is easy to prove that $\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ follows a chi-square distribution with m degrees of freedom:

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m. \tag{3}$$

However, the real value of the parameters are usually unknown. It is the case in the clustering and the classification tasks where the distribution of each cluster has to be estimated. When replacing the real values by the estimators in the Mahalanobis distance definition, the distribution does change. In the case of the classical sample mean $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$,

$$\Delta(\bar{\mathbf{x}}, \mathbf{S}) = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}),$$

it has been proven in [14] that

$$\frac{n}{(n-1)^2} \Delta(\bar{\mathbf{x}}, \mathbf{S}) \sim \mathcal{B}\left(\frac{m}{2}, \frac{n-m-1}{2}\right), \tag{4}$$

with $\mathcal{B}(a, b)$ the beta distribution with parameters a and b.

On the other hand, as stated in [9] when $\mathbf{S}$ is independent of $\mathbf{x}$ the distribution $\Delta(\boldsymbol{\mu}, \mathbf{S})$ is exactly a scaled Fisher distribution:

$$\frac{(n-m)}{(n-1)m} \Delta(\boldsymbol{\mu}, \mathbf{S}) \sim F_{m,n-m}. \tag{5}$$

By using a Slutsky-type argument, we can prove that when $\mathbf{S}$ is independent of $\mathbf{x}$ the distribution of $(n - m)/\left[(n-1)m\right] \Delta(\bar{\mathbf{x}}, \mathbf{S})$ is approximately $F_{m,n-m}$.

In the algorithm presented in the section II-A, we compute $\widehat{\tau}_{ik}$ values that are, up to a scale factor, the squared Mahalanobis distance with Tyler's estimators of the mean and the scatter matrix

$$\Delta(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T) = (\mathbf{x} - \widehat{\boldsymbol{\mu}}_T)^T \widehat{\boldsymbol{\Sigma}}_T^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_T).$$

Even though the asymptotic distribution of the Tyler's estimator has been widely studied, its exact distribution is unknown. Similarly, the asymptotic distribution of $\Delta(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ can be proved to be chi-squared if simultaneous consistency is assumed but the true distribution remains unknown. Related to this particular estimator, in [15] the authors provide a study of the Mahalanobis distance (in the complex case) when the mean is known and the estimator of the parameter is done independently:

$$\Delta(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}}_T) = (\mathbf{x} - \boldsymbol{\mu})^T \widehat{\boldsymbol{\Sigma}}_T^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

When comparing $\Delta(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}}_T)$ to $\Delta(\bar{\mathbf{x}}, \mathbf{S})$ and $\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, they found that the asymptotic variance of the robust Mahalanobis $\Delta(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}}_T)$ when centering around the Wishart-based distance $\Delta(\bar{\mathbf{x}}, \mathbf{S})$ is smaller than the one when centering around the distance based on the true $\boldsymbol{\Sigma}$ parameter. They also showed that the $\Delta(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}}_T)$ distribution is better approximated with the equation (5) than the chi-square distribution, even when the dimension m is small.

III. Outlier rejection

In this section we propose three different procedures to detect multivariate outliers within the data, based on the theoretical results in Section II. These methods differ on the distribution assumption to construct the rejection test, depending on the nature of the problem. In the classification case, we have two independent sets of data A and B. On the other hand, in the classic clustering case we have only one data set C that needs to be labelled. In the following, we specify the procedures.

1) Classification case:

- We fit the robust model using the A dataset.
- We assign each data point of the set B to the cluster that minimizes the robust Mahalanobis distance.
- If a point x_i in B is assigned to the cluster k, we reject it as outlier based on the distribution of (5). If $\Delta(\widehat{\boldsymbol{\mu}}_k^{(A)}, \widehat{\boldsymbol{\Sigma}}_k^{(A)})$ is bigger than the α-quantile of the corresponding scaled Fisher distribution, we classify the observation as an outlier.

2) Clustering case using the chi-square distribution:

- We fit the robust model using the C dataset.
- We assign each data point of the set C to the cluster that minimizes the robust Mahalanobis distance.
- If a point x_i from C is assigned to the cluster k, we reject it as outlier based on (3).

3) Clustering case using the Fisher distribution:

- We divide the data set C into W folds, $C_1, ..., C_W$.
- For each fold C_l, we fit a model from Section II-A to the set $C \setminus C_l$.
- We assign the data point of C_l to the cluster that minimizes the robust Mahalanobis distance.
- If a point x_i from C_l is assigned to the cluster k, we reject it as outlier based on (5) with the respective estimators.

IV. Experiments

In this section, we begin by performing experiments with synthetic data in order to study which distribution is the closest to the robust Mahalanobis distance distribution when using Tyler's estimators. We then apply the method on a subset of the baseline dataset MNIST [16].

A. Simulations

In the following experiments, we show that using the chi-square distribution to approximate the distribution of the Mahalanobis distance can be misleading when the number of observations is not big enough. We simulate a zero-mean Gaussian distribution. We add samples drawn

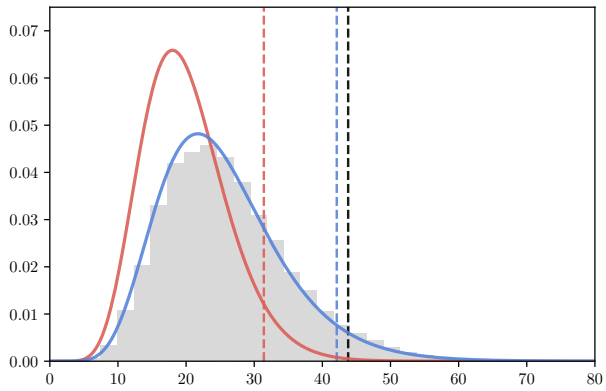

Fig. 1. Histogram of the empirical distribution of the Mahalanobis distance $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ in the classification problem ($n = 200, m = 20$). The theoretical distribution of the $\Delta(\widehat{\boldsymbol{\mu}}, \mathbf{S})$ Mahalanobis distance is plotted in blue and the asymptotic chi-square distribution is shown in red. The theoretical 0.95-quantiles of the distributions are represented with dashed vertical lines in the respective colors. The dashed black line represents the empirical quantile of the $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ distribution.

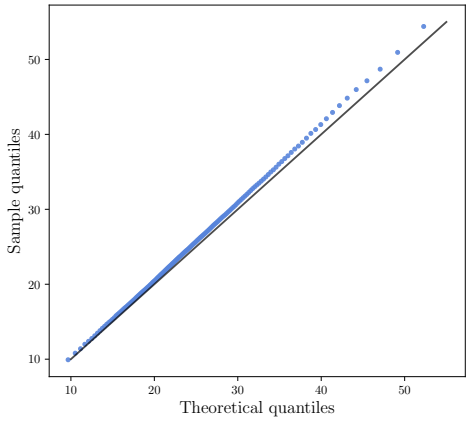

Fig. 2. QQ-plot of the empirical Mahalanobis distance $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ in the classification problem compared to the theoretical distribution in the Gaussian case of $\Delta(\widehat{\boldsymbol{\mu}}, \mathbf{S})$. The identity function is plotted in black.

from a Student's t-distribution variable with heavy tail that we use only on the fitting of the model. We consider the context of classification and the clustering context. In the clustering case, one data set is used to fit the model and each data point is then labelled into clusters. In the classification case, a new set of data independent from the original needs to be classified. Each new data point can be classified to one of the possible classes or as an outlier.

In the classification case, where the computation of the estimator of $\boldsymbol{\Sigma}$ is independent from the data to be classified, we see in Figure 1 that the empirical distribution of the robust Mahalanobis distance $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ is very close to the Fisher distribution in (5). Furthermore, the difference between the two theoretical distributions is clearly distinguishable and the thresholds for the same $\alpha-$quantile are very different. We verify the correspondence between the empirical distribution of the robust Mahalanobis distance and the theoretical distributions in (5) with the qq-plot in Figure 2 where the points can be fitted to a linear function very close to the identity function. Finally, we see in Table I that both accuracy and F1 score for the outlier detection task are much higher using the Fisher distribution than the asymptotic chi-squared.

In the clustering case, the set of data points to be classified is exactly the same as the set of data used for the estimation of the model. The Figure 3 shows that, when replacing the true parameter values by Tyler's estimators, the empirical distribution of the distance is heavier tailed than the compared theoretical densities. Nevertheless, we see that the empirical distance is closer to the asymptotic chi-square distribution than to the exact bounded beta distribution obtained when replacing the true value of the parameters by the classic sample estimators. The tail of the chi-square distribution is closer to the empirical tail and in consequence the decision threshold

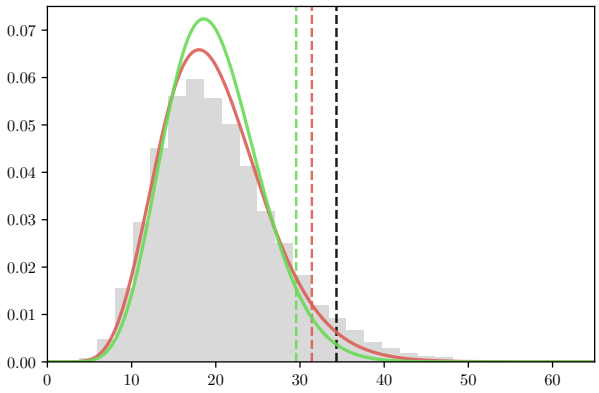

Fig. 3. Histogram of the empirical distribution of the Mahalanobis distance $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ in the clustering problem ($n = 200, m = 20$). The theoretical distribution of the $\Delta(\widehat{\boldsymbol{\mu}}, \mathbf{S})$ Mahalanobis distance is plotted in green and the asymptotic chi-square distribution is shown in red. The theoretical 0.95-quantiles of the distributions are represented with dashed vertical lines in the respective colors. The dashed black line represents the empirical quantile of the $\widehat{\Delta}(\widehat{\boldsymbol{\mu}}_T, \widehat{\boldsymbol{\Sigma}}_T)$ distribution.

is more accurate. Table I shows that even if the proposed W-folded method III-3 improves the accuracy of the model, it looses in power because of the smaller number of observations available which is reflected in the F1 score. This effect might be mitigated using bootstrapping. We leave this for future work.

TABLE I
PERFORMANCE FOR THE OUTLIER REJECTION TASKS FOR THE SIMULATED ZERO-MEAN GAUSSIAN DATA CONTAMINATED WITH OUTLIERS FROM A STUDENT'S DISTRIBUTION.

Case	**Rejection method**	**Performance measure**	
		Accuracy	***F1 score***
Classification	based on (3)	0.86	0.52
	III-1	**0.95**	**0.74**
Clustering	III-2	0.90	**0.68**
	III-3	**0.92**	0.60

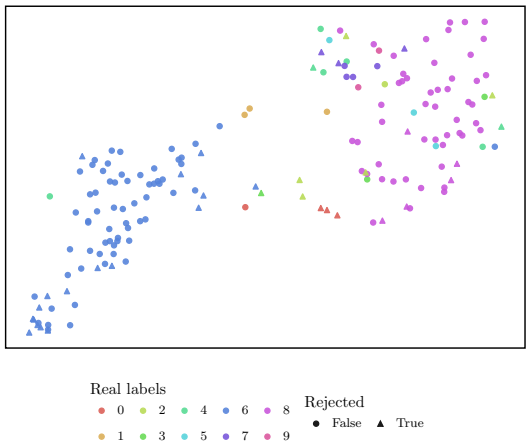

Fig. 4. UMAP visualization of the MNIST 6-8 data with noise in the clustering scenario. The shapes represent the outlier rejection and the colors represent the real digit label.

TABLE II
PERFORMANCE OF OUR PROPOSAL AND THE CHI-SQUARED TEST FOR THE OUTLIER REJECTION TASKS FOR THE CONTAMINATED MNIST DATASET.

Case	**Rejection method**	**Performance measure**	
		Accuracy	***F1 score***
Classification	based on (3)	0.63	**0.43**
	III-1	**0.75**	0.38
Clustering	III-2	0.65	**0.41**
	III-3	**0.72**	0.35

B. Tests on baseline data set

The MNIST dataset [16], a collection of labelled images of hand-written digits, has became a standard baseline to test classification and clustering methods. We test the method on a subset of the MNIST dataset composed for a large proportion of the digits 6 and 8, and a smaller proportion of other digits, representing outliers. Prior to the analysis, we first reduce the dimension of the data set in its original format with a PCA transformation. We perform this transformation based on a trade-off between number of dimensions and percentage of the variance explained, as it is common practice.

In Figure 4, we see a UMAP [17] embedding of the data set. The objective of the UMAP algorithm (as t-SNE but faster) is to visualize the data in 2 dimensions by preserving the local neighborhoods of the data points. The shape of the points represents the prediction outcome of the outlier rejection in each scenario after the model is fitted. The color of the points represents their real digit. We observe that the data points of the periphery of the clusters are often rejected by our method. The Table II shows the performance for the outlier detection task in terms of accuracy and F1 score. Even if we loose in F1 score, the general accuracy is improved.

V. CONCLUSIONS

We studied the distribution of the robust squared Mahalanobis distance using the Tyler's estimators in the context of data classification and clustering. We have seen in simulations that the Fisher distribution approximates much better the empirical distribution than using the chi-square distribution when the number of data points n is not big enough in comparison to the dimension m of the data. We use this study to design an outlier rejection method within a robust mixture model that improves the estimation, classification and outlier rejection. Finally, we tested our outlier rejection method on a variation of the real data set MNIST and we obtained interesting results even if the Gaussian distribution is not strictly followed as usual in real data. In the future, we plan to address the derivation of a better theoretical approximation for the robust Mahalanobis distance with Tyler's estimators as well as a study of bootstrapping methods involving the outlier detection proposal on several folds of the data.

REFERENCES

[1] V. Asanza, E. Pelaez, and F. Loayza, "EEG signal clustering for motor and imaginary motor tasks on hands and feet," in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, 2017, pp. 1–5.
[2] R. R. V. Gonçalves, J. Zullo, B. F. Amaral, P. P. Coltri, E. P. M. Sousa, and L. A. S. Romani, "Land use temporal analysis through clustering techniques on satellite image time series," in *2014 IEEE Geoscience and Remote Sensing Symposium*, 2014, pp. 2173–2176.
[3] M. E. Celebi, "Effective initialization of k-means for color quantization," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 1649–1652.
[4] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.
[5] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 339–348, Oct 2000.
[6] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar, "A general trimming approach to robust cluster analysis," *Ann. Statist.*, vol. 36, no. 3, pp. 1324–1345, 2008.
[7] V. Roizman, M. Jonckheere, and F. Pascal, "A flexible EM-like clustering algorithm for noisy data," *arXiv e-prints*, p. arXiv:1907.01660, Jul 2019.
[8] D. Ververidis and C. Kotropoulos, "Gaussian Mixture Modeling by Exploiting the Mahalanobis Distance," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2797–2811, July 2008.
[9] J. Hardin and D. M. Rocke, "The distribution of robust distances," *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 928–946, 2005.
[10] E. Ollila, D. Tyler, V. Koivunen, and H. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *Signal Processing, IEEE Transactions on*, vol. 60, no. 11, pp. 5597–5625, Nov 2012.
[11] F. Pascal, Y. Chitour, J.-P. Ovarlez, P. Forster, and P. Larzabal, "Covariance structure maximum-likelihood estimates in compound gaussian noise: Existence and algorithm analysis," *Trans. Sig. Proc.*, vol. 56, no. 1, pp. 34–48, Jan. 2008.
[12] E. Ollila and D. E. Tyler, "Distribution-free detection under complex elliptically symmetric clutter distribution," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2012, pp. 413–416.
[13] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.
[14] S. S. Wilks, *Mathematical Statistics*. Princeton University Press, 1943.
[15] G. Draškovic and F. Pascal, "New insights into the statistical properties of M-estimators," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4253–4263, Aug 2018.
[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[17] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018.