ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Robust Clustering of High-Dimensional Data

Anastasios Bellas[1], Charles Bouveyron[1], Marie Cottrell[1], Jérôme Lacaille[2] *

1- SAMM (EA 4543), Université Paris 1,
90, rue de Tolbiac, 75634 Paris Cedex 13, France
e-mail: anastasios.bellas@malix.univ-paris1.fr

2- Snecma, Groupe Safran,
77550 Moissy Cramayel, France

Abstract. We address the problem of robust clustering of high - dimensional data, which is recurrent in real-world applications. Existing robust clustering methods are unfortunately sensitive in high dimension, while existing approaches for high-dimensional data are in general not robust. We propose a hybrid iterative EM-based algorithm that combines an efficient high-dimensional clustering algorithm and the trimming technique. We test our algorithm on synthetic and real-world data from the domain of aircraft engine health monitoring and show its efficiency for high-dimensional noisy datasets.

1 Introduction

Given a set $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ of p-dimensional vectors, *clustering* refers to the task of partitioning $\mathbf{X}$ into K homogeneous groups. Clustering techniques proceed in an unsupervised manner, that is they do not possess any information on the true partition of the dataset. Many real-world datasets, for example those obtained by measuring some physical quantities, can often be "contaminated" by observations corresponding to extreme manifestations of the phenomenon being measured or even with no relation to it. These observations are known under the names "outliers", "noise", "anomalies", "novelties" and many more.

Unfortunately, most general clustering algorithms are not robust to the presence of outliers in the dataset. In the past few years, there have been efforts to develop robust clustering techniques that can be divided into two main families: mixture-based (probabilistic) and trimming-based. On the one hand, robust mixture-based methods consider an extra component admitting a uniform [1, 2], an improper uniform [3] or a t-Student [4] distribution to model outliers. On the other hand, there are methods that make use of trimming such as trimmed K-means [5] and TCLUST [6]. The latter is an iterative EM clustering type algorithm in which a predetermined proportion of outliers is being trimmed (removed) from the dataset.

Moreover, model-based clustering methods suffer from the *curse of dimensionality*, since they require an exponentially growing number of data points with increasing dimension. A direct consequence when using generative models is having to estimate a large number of parameters compared to the available

*This work was supported by Snecma, the French aircraft engine manufacturer.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

data. Parsimonious mixture models [1, 7] address this issue by imposing constraints on the covariance matrices in order to reduce the number of parameters that need to be estimated. Another reliable solution is subspace clustering methods like HDDC [8] and Fisher-EM [9], which model and cluster the data in low-dimensional subspaces.

The paper is structured as follows: Section 2 presents HDDC and its robust version called HDRC. In Section 3 we present results on synthetic and real data from the domain of aircraft engine health monitoring. Finally, in Section 4 we briefly discuss future work and perspectives.

2 High-Dimensional Robust Clustering

2.1 Gaussian models for HD data and the HDDC algorithm

As in the classical Gaussian mixture model framework [10], HDDC assumes that each of the K groups has a Gaussian density $\mathcal{N}_p(\mu_k, \Sigma_k)$ with means μ_k and covariance matrices Σ_k , for $k = 1, ..., K$. Let Q_k be the orthogonal matrix with the eigenvectors of Σ_k as columns and Δ_k be the diagonal matrix which contains the eigenvalues of Σ_k such that $\Delta_k = Q_k^t \, \Sigma_k \, Q_k$. The matrix Δ_k is therefore the covariance matrix of the kth class in its eigenspace. It is further assumed that Δ_k can be divided into two blocks:

$$\Delta_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & & \mathbf{0} \\ & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} & \\ \mathbf{0} & & \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (p - d_k) \end{matrix} \tag{1}$$

with $a_{kj} > b_k$, $j = 1, ..., d_k$, and where $d_k \in \{1, \ldots, p-1\}$ is unknown. From a practical point of view, one can say that the parameters $a_{k1}, ..., a_{kd_k}$ model the variance of the actual data of the kth class and the unique parameter b_k can be viewed as modeling the variance of the noise. The dimension d_k can be considered as well as the intrinsic dimension of the latent subspace of the kth group. By constraining model parameters between or across the groups, it is possible to obtain several submodels from this model. A list of all submodels is given in [8]. Notice that the family of submodels encompasses in particular the general GMM ($d_k = p - 1$) or the mixture of probabilistic PCA [11].

The intrinsic dimension and the number of groups cannot be estimated by maximum likelihood since they both control the model complexity. In [8], the authors proposed to estimate the dimensions d_k, $k = 1$, ..., K through the Cattell's scree-test [12] which looks for a break in the eigenvalues scree. The selected dimension is the one for which the subsequent eigenvalue differences are smaller than a threshold. The threshold can be provided by the user or selected using BIC [13]. The number of clusters K can be chosen thanks to the BIC criterion as well. In the specific case of models where $a_{kj} = a_k$ or $a_{kj} = a$ for

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

$k = 1, ..., K,\ j = 1, ..., d_k$, it has been recently proved by [14] that the maximum likelihood estimate of the intrinsic dimensions d_k is asymptotically consistent.

2.2 High-Dimensional Robust Clustering

We introduce a novel algorithm, called *HDRC* (High-Dimensional Robust Clustering), which combines HDDC and the trimming technique, therefore resulting in a robust clustering algorithm which is efficient for high-dimensional data. In particular, we extend HDDC by adding an intermediate trimming step (T-step) between the E- and M- steps of the EM part of the algorithm:

- The E step computes the posterior probabilities $t_{ik}^{(q)} = \mathbb{P}(Z = k|X = x_i)$ according to the model parameters estimated at iteration $q-1$ through the formula $t_{ik}^{(q)} = 1 \Big/ \sum_{\ell=1}^{K} \exp\left(\frac{1}{2}(\Gamma_k^{(q)}(x) - \Gamma_\ell^{(q)}(x))\right)$ where the classification function $\Gamma_k^{(q)}$ is as follows when $a_{kj} = a_k$, $k = 1, ..., K,\ j = 1, ..., d_k$:

$$\begin{aligned}\Gamma_k^{(q)}(x) &= \frac{1}{a_k}\|\mu_k - P_k(x)\|^2 + \frac{1}{b_k}\|x - P_k(x)\|^2 \\ &\quad + d_k \log(a_k) + (p - d_k)\log(b_k) - 2\log(\pi_k),\end{aligned}$$

  where P_k is the projection operator on the latent subspace of the kth class and models parameters are estimated in the M step at iteration $q-1$.

- The T step holds out of the dataset (trims) a fixed proportion of the data points with smallest values for $\max_{k=1,...,K} \pi_k^{(q)} f\left(x; \mu_k^{(q)}, \Sigma_k^{(q)}\right)$ which is equivalent to trimming the data points with the largest values for $\min_{k=1,...,K} \Gamma_k^{(q)}(x)$. Let $R^{(q)}$ be the set of the trimmed data points.

- The M step then updates the estimates of model parameters by maximizing the expectation of the *trimmed* complete likelihood conditionally to the posterior probabilities $t_{ik}^{(q)}$ for the data points $x_i \notin R^{(q)}$. Update formulas for the parameters can be found in [8].

3 Experiments

We tested HDRC on synthetic high-dimensional data and real data from the domain of aircraft engine health monitoring.

3.1 Application on synthetic data

For the synthetic dataset, we generated 3 multivariate Gaussians with a total of 1000 data points according to the HDRC model in the p-dimensional space, where $p = 10, \ldots, 100$. The means and covariance matrices were chosen such that the task would not be too challenging for the algorithms tested. In particular, we set the parameters as follows: $a_{1j} = 150$ and $b_1 = 15$, $a_{2j} = 75$ and $b_2 = 10$,

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

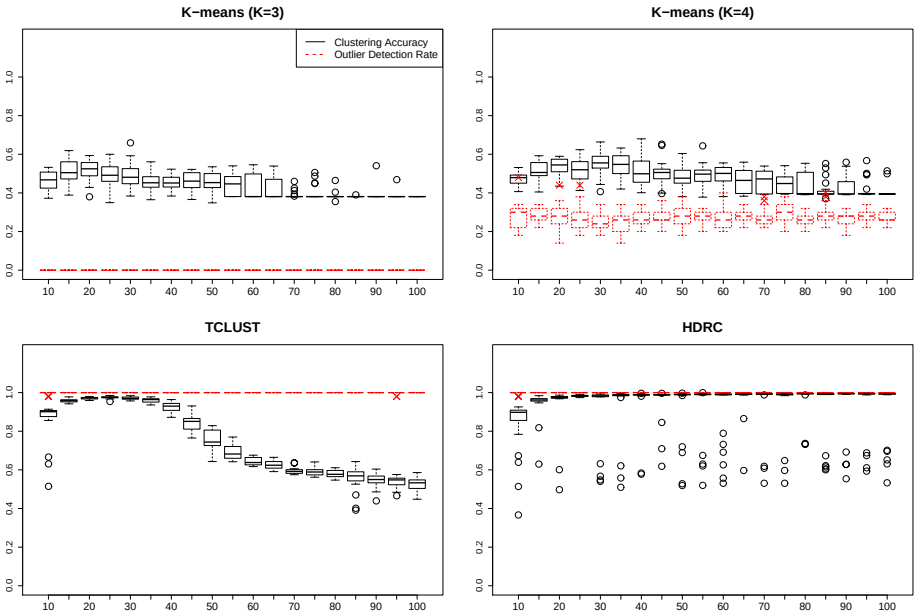

Fig. 1: Clustering accuracy (black solid boxes) and TPR of the outliers (red dashed boxes) for the synthetic dataset, plotted against the dimension of the data. Black circles correspond to extreme observations for the clustering accuracy, and red 'x' to those for the TPR of the outliers. For K-means with K=3 the TPR is artificially set to zero (see the text for details).

$a_{3j} = 50$ and $b_3 = 5$, respectively, for the three Gaussians, where $j = 1, \ldots, d_k$. The intrinsic dimensions d_k were $d_1 = 10$, $d_2 = 5$ and $d_3 = 2$ respectively. We added a quantity of outliers equal to 5% of the dataset size, uniformly distributed on the interval $[-40, 40]$ for each of the p variables.

In the experiments, we tested K-means with $K = 3$ and $K = 4$, TCLUST and HDRC. K-means was used as a baseline; we wanted to examine its behaviour when it has no knowledge of the existence of outliers ($K = 3$) and when this knowledge is given explicitly by adding an extra group ($K = 4$).

For TCLUST, we imposed a restriction on the eigenvalue ratio of the covariance matrices. More precisely, the maximum value for the ratio of the maximum to the minimum eigenvalue for the dimension $p = 10$ was set to 50, augmented by 50 for each dimension p thereafter. For HDRC we used a random initialization. For TCLUST and HDRC algorithms, we set the number of groups to $K = 3$ (not counting the group of the outliers). Moreover, we supplied them with the true outlier proportion, that is $\alpha = 0.05$. For all the algorithms, the number of initializations was set to 25 and the maximum number of iterations to 60.

For each value of p, we replicated the experiment 25 times. For each replication, we calculated the clustering accuracy (up to label switching) as the ratio of

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

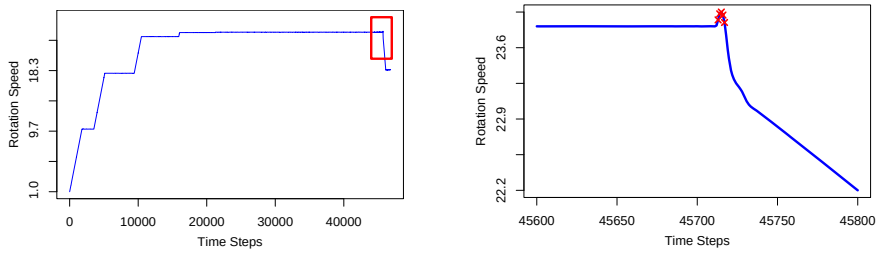

Fig. 2: On the left, plot of the engine speed variable. The red rectangle shows the area where the anomalies occur. On the right, a zoom-in on this area: the litte "bump" corresponds to an engine malfunction. HDRC with $K = 7$ detects succesfully the anomalous data points (red 'x'). We underline that clustering was performed on the p-dimensional space, where $p = 173$.

correct cluster assignments to the size of the dataset and the true positive rate for the outliers (TPR) as the ratio of data points correctly detected as outliers to the true number of outliers.

Figure 1 presents the clustering accuracy and the true positive rate for the outliers for our experiments. We observe that K-means with $K = 3$ fails in clustering correctly the data. Note that $K = 3$ indicates that K-means "naively" tries to cluster data with outliers without being aware of their presence and that is why we did not evaluate the TPR (artificially set to zero). We also observe that even when an extra group is added to model the outliers ($K = 4$), K-means does not do much better. As expected, TCLUST succeeds in detecting the outliers in all cases but appears to be sensitive in dimension. The way we simulated data, the outliers have a large variance and thus, it should be easy to detect them correctly. This means that the mediocre performance of TCLUST in clustering is, to a great extent, due to the high dimensionality of the data. Finally, we see that HDRC succesfully manages to cluster the data and detect the outliers even in high dimension.

3.2 Application to aircraft engine health monitoring data

In the aircraft engine domain, the monitoring of engine health is a crucial task. SNECMA performs such tests in a test chamber environment. A multitude of engine or test chamber parameters are measured, such as chamber pressure, engine temperature, engine speed etc. The goal here is to be able to issue a warning whenever there is a malfunction (anomaly) of the engine or the chamber, before significant damage occurs to any of the two.

The dataset we consider here consists of 46 830 observations and 173 variables. Each test is a sequence of alternating stationary and non-stationary phases at different levels. For this particular test series, we know that there has been a malfunction and that the test sequence was stopped abruptly . We apply HDRC to check if it will correctly detect the outlying observations (anomalies).

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

As we can see in Figure 2, HDRC succeeds in detecting the outliers (the anomalous behaviour). We underline that the clustering was performed in the p-dimensional space, where $p = 173$, and that we plotted only one of the variables in Figure 2 for visual clarity.

4 Discussion

We have presented a high-dimensional robust clustering algorithm, combining HDDC and the trimming concept. We have shown its efficiency on noisy high-dimensional synthetic datasets and on a concrete, real-world application. However, trimming-based robust clustering methods and HDRC assume that the true outlier proportion among the data is known. In practice, this is rarely true. Therefore, we need a procedure to select the proportion yielding the most satisfactory results. We are planning to address this problem using model selection techniques in future works. Moreover, preliminary experiments suggest that both TCLUST and HDRC give less satisfactory results when outliers are added in only some (not all) of the variables. We think that a variable selection technique adapted to this specific task could boost their performance in this case.

References

[1] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[2] C. Hennig and P. Coretto. The noise component in model-based cluster analysis. *Data Analysis, Machine Learning and Applications*, pages 127–138, 2008.

[3] C. Hennig. Breakdown points for maximum likelihood estimators of location–scale mixtures. *The Annals of Statistics*, 32(4):1313–1340, 2004.

[4] D. Peel and G.J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

[5] L.A. García-Escudero, A. Gordaliza, and C. Matrán. Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2):434–449, 2003.

[6] L.A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324–1345, 2008.

[7] P.D. Mcnicholas and T.B. Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

[8] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007.

[9] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, 2011.

[10] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[11] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neur. Comput.*, 11(2):443–482, 1999.

[12] R. Cattell. The scree test for the number of factors. *Multivariate Behav. Res.*, 1(2):245–276, 1966.

[13] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[14] C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.