Robust Estimation of Gaussian Mixture Models Using Anomaly Scores and Bayesian Information Criterion for Missing Value Imputation

Florian Mouret, Mohanad Albughdadi, Sylvie Duthoit, Denis Kouamé, Jean-Yves Tourneret

Robust Estimation of Gaussian Mixture Models Using Anomaly Scores and Bayesian Information Criterion for Missing Value Imputation

F. Mouret[(1,2)*], M. Albughdadi[(1)], S. Duthoit[(1)], D. Kouamé[(3)], J.-Y. Tourneret[(2)]
[(1)] TerraNIS, 12 Avenue de l'Europe, 31520 Ramonville-Saint-Agne, France
[(2)] University of Toulouse / IRIT-ENSEEIHT / TéSA, 2 Rue Charles Camichel, 31000 Toulouse, France
[(3)] University of Toulouse / IRIT-UPS, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France
[*] Correspondence: florian.mouret@irit.fr

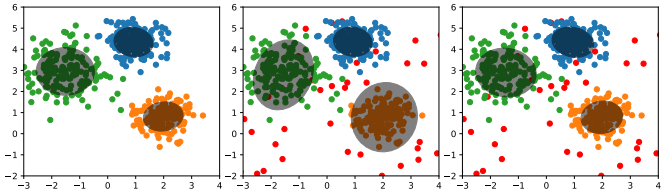
***Abstract*—The Expectation-Maximization algorithm is a very popular approach for estimating the parameters of Gaussian mixture models (GMMs). A known issue with GMM estimation is its sensitivity to outliers, which can lead to poor estimation performance depending on the dataset under consideration. A common approach to deal with this issue is robust estimation, which typically consists of reducing the influence of the outliers on the estimators by weighting the impact of some samples of the dataset considered as outliers. In an unsupervised context, it is difficult to know which sample from the database corresponds to a normal observation. To that extent, we propose to use within the EM algorithm an outlier detection step that attributes an anomaly score to each sample of the database in an unsupervised way. A modified Bayesian Information Criterion is also introduced to efficiently select the appropriate amount of outliers contained in a dataset. The proposed method is tested on a benchmark remote sensing dataset coming from the UCI Machine Learning Repository. The experimental results show the interest of the proposed robustification when compared to other benchmark imputation procedures.**

***Index Terms*—Imputation, Anomaly Detection, Gaussian Mixture Model, Robust estimation, Isolation Forest, One-Class SVM**

I. Introduction

Many real-world datasets have missing values, which explains why various approaches have been proposed to bypass this potential absence of data [1]. Missing data can be due to the data acquisition process *(e.g.*, presence of clouds in multispectral images) or to a change in the acquisition process (*e.g.*, addition of new variables). Missing Value Imputation (MVI) is a strategy commonly adopted to solve the missing data problem, in particular when using machine learning approaches, which generally require complete feature matrices. The various MVI techniques that have been studied in the literature can be grouped in two main categories, namely *statistical* and *machine learning* based techniques [2]. This includes the multiple imputation by chained equation (MICE) [3] for statistical approaches and the k-nearest neighbors (KNN) imputation [4] for machine learning based techniques.

Among these approaches, those based on the expectation–maximization (EM) algorithm [5] for Gaussian Mixture Models (GMMs) have received a considerable amount of attention [2]. GMMs are attractive mainly because: 1) they are able to model the statistical properties of many datasets, 2) they can be applied to a wide range of tasks such as clustering and classification and 3) they naturally handle missing data. Within the EM framework, missing data can be considered as latent variables that can be handled in a straightforward manner (after carefully deriving appropriate sufficient statistics). However, a known issue with GMM estimation is its sensitivity to outliers, which is illustrated in the toy example depicted in Fig. 1, where it can be observed that a classical GMM estimation is highly impacted by the presence of outliers (Fig. 1(b)). To overcome this issue, a classical approach is robust estimation. The main idea behind robust estimation is to estimate the model parameters by weighting the importance of outlier samples. Samples with small weights have a reduced influence on the parameter estimates whereas larger weights have a more important impact on the estimation. Fig. 1(c) shows an example of robust estimation obtained using the algorithm presented in this paper.



(a) (b) (c)

Fig. 1. Toy example with 3 Gaussian clusters (500 samples): (a) GMM estimation without outliers, (b) GMM estimation in the presence of 50 outliers (red points) and (c) Robust GMM estimation in the presence of 50 outliers.

In an unsupervised scenario such as the one considered in this paper, two main issues have to be considered: 1) no labeled samples are available (*i.e.*, representative examples of normal and abnormal behavior are not available) and 2) it is difficult to evaluate automatically which samples have to be considered as outliers when estimating the GMM parameters. This paper aims at addressing these two issues. To detect outliers in an unsupervised way, we propose to include the results of an unsupervised outlier detection algorithm directly within the EM framework. This follows the idea developed

in [6], which was specifically adapted to crop monitoring based on remote sensing. However, contrary to [6] who focused on the isolation forest (IF) algorithm [7], this paper also considers the one-class support vector machine (SVM) method, mainly to have a more generic method that can be adapted to the considered dataset. In addition, we also propose a way of tuning the parameters of the outlier detection method used within the EM algorithm for robust estimation. Based on a modification of the Bayesian Information Criterion (BIC) [8], this tuning procedure allows us to automatically choose both the appropriate outlier detection algorithm and the amount of samples to be considered as outliers within the EM framework.

II. Robust estimation of Gaussian mixture models using anomaly scores

In this section, we first recall the classical EM algorithm for GMM with missing data. In a second step, we introduce a strategy for robust estimation based on outlier scores computed using the one-class SVM or IF methods. Finally, a model selection strategy based on BIC is proposed to determine the best outlier detection approach to be used within the EM algorithm.

A. Standard EM algorithm for GMM with missing data

For GMM estimation, we suppose that each sample $\boldsymbol{x}_n \in \mathbb{R}^M$ is a row of the feature matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T$ (of size $N \times M$, with N the number of samples and M the number of features) drawn from a mixture of K multivariate normal distributions. In the presence of missing data, each sample can have missing values for specific features. More precisely, each sample can be decomposed into $\boldsymbol{x}_n = (\boldsymbol{x}_n^{o_n}, \boldsymbol{x}_n^{m_n})$, where $\boldsymbol{x}_n^{o_n}$ and $\boldsymbol{x}_n^{m_n}$ are the vectors of observed and missing features. More generally, the superscripts o_n and m_n denote the observed and missing components of the sample n and can be used for matrices too, *e.g.*, $\boldsymbol{\Sigma}_k^{o_n,m_n}$ refers to the elements of the matrix $\boldsymbol{\Sigma}_k$ in the rows and columns specified by o_n and m_n (and so on). For brevity, we will denote $o_n = o$ and $m_n = m$ in the following, but it is important to keep in mind that these subscripts are sample-dependent. The EM algorithm aims at maximizing the observed (or complete) log-likelihood $\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{z})$, or in brief $\log \mathcal{L}_c$, defined as:

$$\log \mathcal{L}_c = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right], \tag{1}$$

where $\boldsymbol{X}^o$ is the set of all observed variables, $\boldsymbol{X}^m$ is the set of all missing variables, $\mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the marginal multivariate Gaussian distribution of the observed sample $\boldsymbol{x}_n^o$ associated with the joint multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and $\boldsymbol{\theta} = \{\pi_1, ..., \pi_k, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}$ contains the set of parameters to be estimated. Note that π_k denotes the a priori probability of class k, whereas $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the kth component of the GMM. The latent variables z_{nk} (to be estimated) are introduced to define the cluster label of each observation (*i.e.*, $z_{nk} = 1$ if the sample n belongs to class k, and $z_{nk} = 0$ otherwise). The EM algorithm alternates between expectation (E-) and Maximization (M-) steps, which are detailed below, to find a local maximum of (1).

E-step: evaluate $E[\log \mathcal{L}_c | \boldsymbol{\theta}^{(t)}, \boldsymbol{x}^o]$ (the parameters at iteration t) requires to compute the following sufficient statistics:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_j^o, \boldsymbol{\Sigma}_j^{oo})}, \tag{2}$$

$$\hat{\boldsymbol{\mu}}_{nk}^m = \boldsymbol{\mu}_k^m + \boldsymbol{\Sigma}_k^{mo} (\boldsymbol{\Sigma}_k^{oo})^{-1} (\boldsymbol{x}_n^o - \boldsymbol{\mu}_k^o), \tag{3}$$

$$\hat{\boldsymbol{x}}_{nk}^m = (\boldsymbol{x}_n^o, \hat{\boldsymbol{\mu}}_{nk}^m), \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{nk}^{mm} = \boldsymbol{\Sigma}_k^{mm} - \boldsymbol{\Sigma}_k^{mo} (\boldsymbol{\Sigma}_k^{oo})^{-1} \boldsymbol{\Sigma}_k^{mo}, \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_{nk} = \begin{pmatrix} \boldsymbol{0}^{oo} & \boldsymbol{0}^{om} \\ \boldsymbol{0}^{mo} & \hat{\boldsymbol{\Sigma}}_{nk}^{mm} \end{pmatrix}. \tag{6}$$

The terms γ_{nk} are referred to as *responsibilities* and correspond to the probability that sample n is drawn from the kth class. More precisely, $\gamma_{nk} = E[z_{ik} | \boldsymbol{\theta}^{(t)}, \boldsymbol{x}^o]$, which is similar to the case without missing data except that it is evaluated on the observed data. The other terms are specific to the GMM estimation in the presence of missing data and results from the estimation of $E\left[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) | \boldsymbol{x}^o, \boldsymbol{\theta}^{(t)}\right]$ [5]. One can note that the missing values of sample n are imputed using the conditional expectation of the missing variables given that $\boldsymbol{x}_n$ has been generated by Gaussian $\#k$. Similarly, the conditional covariances of the missing values have to be computed, leading to $\hat{\boldsymbol{\Sigma}}_{nk}$ which is filled with zeros except for the missing components.

M-step: maximize the current expectation leads to the following new set of parameters:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N} = \frac{N_k}{N}, \tag{7}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \hat{\boldsymbol{x}}_n, \tag{8}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \left[(\hat{\boldsymbol{x}}_n - \boldsymbol{\mu}_k)(\hat{\boldsymbol{x}}_n - \boldsymbol{\mu}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk}\right], \tag{9}$$

where one can observe that the estimate of the mean vector is similar to the case without missing data, except that missing values have been imputed using the expressions resulting from the E-step, and that the estimate of the covariance matrix has been corrected by $\hat{\boldsymbol{\Sigma}}_{nk}$ to take into account missing values.

B. Robust GMM estimation

The estimation of the mixture parameters in the M-step is sensitive to outliers, which can be addressed with robust estimation. Following the works conducted in [9], a robust estimation of the mixture parameters can be obtained by introducing weights w_n associated with each sample of the dataset (the estimation of π_k is unchanged):

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} w_n \gamma_{nk} \hat{\boldsymbol{x}}_{nk}}{\sum_{n=1}^{N} w_n \gamma_{nk}}, \tag{10}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} w_n^2 \gamma_{nk} \left[(\hat{\boldsymbol{x}}_n - \boldsymbol{\mu}_k)(\hat{\boldsymbol{x}}_n - \boldsymbol{\mu}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk}\right]}{\sum_{n=1}^{N} w_n^2 \gamma_{nk}}, \tag{11}$$

where a small (resp. large) value of w_n means that the sample n has less (resp. more) influence on the estimation of the mean vectors and covariance matrices (note that w_n are global weights independent of k). As a consequence, one would like to have w_n close to 1 for inliers, and close to 0 for outliers. As explained before, in the unsupervised scenario, knowing the appropriate values for the weights w_n is difficult. For instance, the method proposed in [9] cannot be considered here since it relies on labeled data to separate inliers and outliers. Anomaly detection algorithms are particularly adapted to address this problem, since they (generally) provide an outlier score for each sample in a fully unsupervised manner. The first contribution of this work is to propose a way of injecting these outlier scores into the EM algorithm to make it robust to the presence of outliers. More precisely, we study two strategies allowing the weights w_n to be defined. These strategies are based on the Isolation Forest (IF) [7] and the One-Class Support Vector Machine (OCSVM) method [10]. Even if other approaches could be investigated, the main objective here is to present some insights about how to define the weights w_n depending on the outlier score provided by an outlier detection algorithm.

*1) **Isolation Forest**:* the IF algorithm [7] assumes that outliers can be isolated more easily by an isolation tree than normal instances. An isolation tree is a binary decision tree constructed by randomly choosing at each node a feature and a split value (chosen between the minimum and the maximum of the feature). The parameters of the IF algorithm are the number of trees, and the subsampling used to construct each tree (they were respectively fixed to 1000 and 256 during the experimental results conducted in this paper). The outlier score attributed by the IF algorithm is related to the average path length needed to isolate a given sample when using isolation trees. The final outlier score provided by the IF algorithm to sample x_n (denoted as $\text{score}_{\text{IF}}(\hat{\boldsymbol{x}}_n)$) varies in the range $[0, 1]$ (higher scores correspond to outliers). We then propose the following strategy to convert these scores into weights:

$$w_n = \frac{1}{1 + \exp\left[\alpha(\text{score}_{\text{IF}}(\hat{\boldsymbol{x}}_n) - \beta)\right]}, \tag{12}$$

where α and β are two parameters to be fixed by the user. Note that this function of the outlier score has a sigmoidal shape, with a unique inflection point at $\text{score}_{\text{IF}} = \beta$ and an inflection speed controlled by α (for high values of α, the function reduces to a hard thresholding operation).

*2) **One-Class SVM**:* OCSVM assumes that normal samples are part of the same class delimited by a separating boundary [11]. As in the classical SVM algorithm, the OCSVM approach may be used with a kernel for the decision function, which leads to learn a non-linear separating boundary. This paper concentrates on a radial basis function (RBF) kernel (see [12] for detail and derivations). Two hyperparameters control the behavior of the OCSVM algorithm with this kernel: ν, which is an upper bound for the maximum fraction of samples located outside the separating boundary and σ referred to as the kernel bandwidth, which has to be adjusted for each dataset. In this paper, we use the heuristic proposed in [13, p. 93] consisting of estimating σ as the median of the pairwise Euclidean distances between vectors from the learning set. A straightforward way to assign outlier scores with the OCSVM is to use the distance to the separating hyperplane, denoted as $D(\hat{\boldsymbol{x}}_n)$ ($D(\hat{\boldsymbol{x}}_n)$ is negative when the observation is within the learned boundaries and positive otherwise). The following function can then be used to convert the outlier scores to weights:

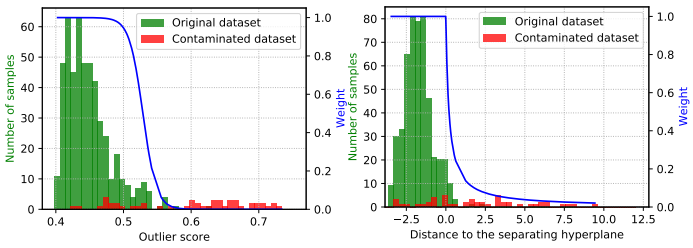
$$w_n = \frac{1}{1 + \alpha \times \text{score}_{\text{OCSVM}}(\hat{\boldsymbol{x}}_n)}, \tag{13}$$

where the value of α has an impact on the speed of decrease of the curve. In this work, we propose to define the value of $\text{score}_{\text{OCSVM}}(\hat{\boldsymbol{x}}_n)$ as follows:

$$\text{score}_{\text{OCSVM}}(\hat{\boldsymbol{x}}_n) = \begin{cases} 0 & \text{if } D(\hat{\boldsymbol{x}}_n) \leq 0, \\ D(\hat{\boldsymbol{x}}_n) & \text{if } D(\hat{\boldsymbol{x}}_n) > 0. \end{cases} \tag{14}$$

Thus, all the inliers samples have a weight equal to 1, whereas outliers have a weight whose value decreases proportionally to their distances to the separating boundary.

To have an easier appreciation of the two different strategies, we provide an illustration based on the dataset used in Fig. 1, where Fig. 2(a) displays the weights (in blue) attributed using (12) and Fig. 2(b) using (13).



(a) Isolation Forest (b) One-Class SVM

Fig. 2. Examples of weight variations (blue curve) with respect to the outlier scores (a) IF algorithm ($\alpha = 100$ and β fixed so that 10% of the samples are considered as outliers) and (b) OCSVM algorithm ($\nu = 10\%$ and $\alpha = 5$). The dataset used for this experiment was displayed in Fig. 1.

C. Model selection using a Bayesian Information Criterion

The values of the weights w_n are directly related to 1) the choice of the detection algorithm and 2) the threshold used to separate inliers and outliers. When using the IF algorithm, a natural threshold is $\beta = 0.5$. However, as illustrated in Fig. 2 this value of β is not always optimal. When using the OCSVM algorithm, the choice of the threshold (ν in that case) is more important since it directly influences the construction of the decision boundary.

In an unsupervised scenario, a manual tuning of the parameters defining the weights w_n can be difficult. To that extent, we propose to use the Bayesian Information Criterion (BIC) [14], which is a classical criterion used for model selection [15] defined as follows:

$$\text{BIC} = -2\log(L) + p\log(n), \tag{15}$$

where L is the likelihood of a given model, p is the number of unknown parameters and n is the number of samples used

to estimate the GMM parameters. The BIC aims at finding a compromise between the likelihood and the number of parameters in the model (the lower BIC, the better the model). Indeed, while the negative log-likelihood tends to decrease when adding parameters, having too many parameters can result in over-fitting and very poor estimation of the GMM. The penalization $p\log(n)$ in BIC counterbalances the effect of this increase of the likelihood when p increases. The value of p minimizing (15) is generally chosen for a particular experiment.

However, the use of BIC in the presence of outliers can be not appropriate if the best model is over-fitting the outlier samples, which was confirmed in our experiments. To that extent, we propose a simple modification that consists in evaluating the BIC only using the inlier samples, which correspond to samples with oulier scores lower than a given threshold (*i.e.*, defined with β when using the IF algorithm and ν when using the OCSVM algorithm). The value of L in (15) is then replaced by the likelihood of the inlier samples and n by the number of these inlier samples.

III. Experimental results

This section evaluates the proposed method on the "Statlog Landsat Satellite dataset", which is a benchmark dataset from the University of California at Irvine (UCI) database[1]. Each example of this dataset corresponds to a patch associated with a 3×3 neighborhood in a satellite image. Each pixel has been acquired with 4 different spectral bands, which makes a total of 36 features per sample. The training set of this database is composed of 4435 samples labeled using 6 different land cover classes (red soil, cotton crop, grey soil, etc.).

Missing data is a particularly important issue in remote sensing with multispectral images, since these images are sensitive to cloud coverage. Missing data were simulated by removing the values of pixels randomly selected within the dataset. When a pixel was declared as missing, all the 4 spectral bands values were removed from the corresponding vector. The percentage of mixing pixels was set to 40% (for conciseness, results obtained with different percentages of missing data are not presented here since they lead to similar conclusions). Finally, we added 10% of outlier samples to the dataset, with values randomly chosen between the minimum and maximum of each feature. These outlier pixels can occur in remote sensing, and typically correspond to undetected clouds, wrong parcel delineations or shadows [16].

For the GMM estimation algorithm, the number of components was set to $K=6$. The sigmoid parameter α was set to $\alpha=50$ for the IF algorithm and to $\alpha=2$ for the OCSVM algorithm. These values of α could also be adjusted using the BIC, but were fixed here for conciseness (we observed that the tuning of this parameter has less influence on the results).

A. Choose the appropriate GMM model

In a first experiment, we analyzed the effect of changing the threshold of the outlier detection algorithm used to attribute weights with the robust GMM imputation method. When using the IF algorithm, this consists in choosing for the parameter β a score that separates $x\%$ of the outliers, which corresponds to $\nu = x$ in the OCSVM algorithm. For each threshold value in the range $[1\%, 18\%]$, we computed the values of BIC, the values of the modified BIC (computed using the inliers samples only), and the Mean Absolute Percentage Error defined as follows:

$$\text{MAPE} = \frac{100}{N_{\text{miss}}} \sum_{i=1}^{N_{\text{miss}}} \frac{|f_i - \hat{f}_i|}{|f_i|}, \tag{16}$$

with N_{miss} is the number of missing values, f_i is the actual value of the ith feature and $\hat{f}_i$ is its imputation. The idea behind this experiment is to evaluate the interest of using the BIC to optimize, in an unsupervised scenario, the imputation of the missing features by choosing an optimal threshold to separate inliers and outliers. The results obtained using the same dataset with missing values are summarized in Fig. 3, which shows MAPE as a function of BIC (Fig. 3(a)) and MAPE as a function of the modified BIC (Fig. 3(b)), obtained by varying the threshold used within the outlier detection algorithms.

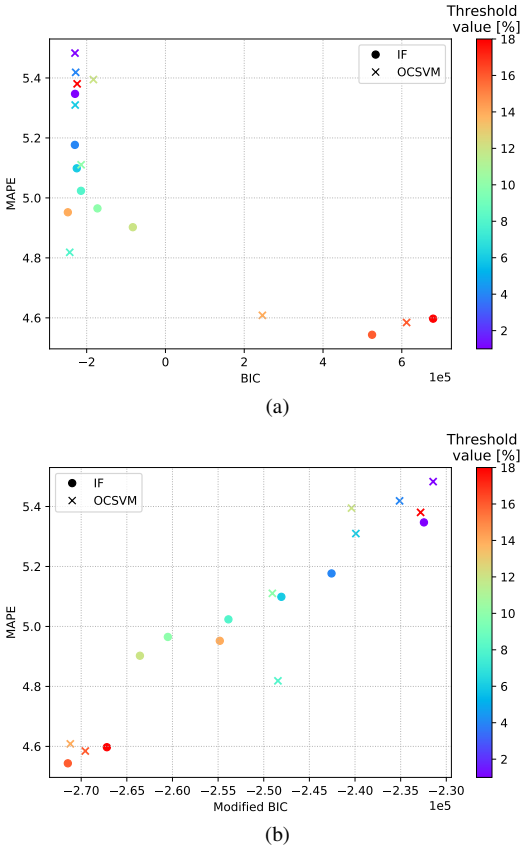


Fig. 3. MAPE vs BIC (a) MAP versus modified BIC (b) obtained by varying the threshold parameters used in the robust GMM framework (*i.e.*, the parameter β or ν, depending on the outlier detection algorithm). Crosses are for the OCSVM strategy and dots are obtained using the IF strategy.

Overall, two main conclusions can be drawn based on these experiments. First, the MAPE can be reduced by choosing an appropriate outlier detection strategy during the GMM

[1] https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)

estimation (*i.e.*, the MAPE varies approximately between 4.6% and 5.4% depending on the outlier detection algorithm and the percentage of outliers considered). Second, the BIC computed using the inliers samples only (Fig. 3(b)) can be used to select accurately an optimal imputation model since in this case, BIC is a function of MAPE satisfying "the lower BIC, the lower MAPE", contrary to the standard BIC displayed in Fig. 3(a). Finally, the two strategies proposed to attribute the outlier weights lead to similar results, which is interesting for datasets where a specific outlier detection method is more adapted.

B. Comparison with other imputation methods

The proposed imputation method was compared to the classical GMM imputation method, the KNN imputation method [4] and MICE [3]. We used the scikit learn implementations (version 0.24.2) of KNN and MICE algorithms with their default settings, and did our own python implementation of the EM algorithms. For the robust GMM algorithms, we used BIC (computed with inliers) to automatically choose the optimal outlier detection algorithm between various possible configurations (using IF and OCSVM with different outlier thresholds). Using the experiment setup presented below (*i.e.*, 40% of missing pixels and 10% of outliers added), we have run 50 Monte Carlo simulations on the landsat dataset and computed the MAPE for the 4 different imputation methods. The obtained results are summarized in Fig. 4. Overall, the proposed robust GMM imputation outperforms the other tested methods. In particular, while the standard GMM approach is more accurate than the KNN and MICE algorithms, using the robust extension always leads to better and more stable reconstructions. Using an i7-11850H processor, each MC run took approximately 5s with the KNN and MICE methods, whereas it took around 100s to fit a single GMM model (various GMM have to be fitted when tuning the model hyper-parameters with BIC). Thus, the better performances of the EM-based approaches is obtained at the price of a higher computational complexity.

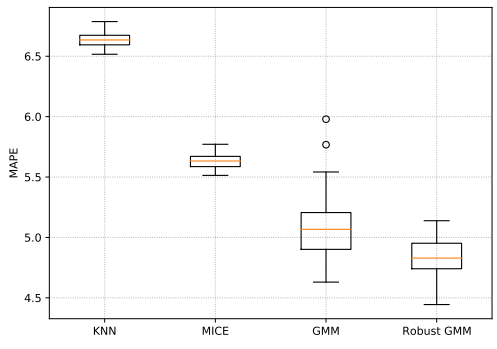


Fig. 4. Boxplots of 50 MC runs for the KNN, MICE, GMM and robust GMM algorithms. The quantity of missing data is 40% and there are 10% of outliers.

IV. Conclusion

This paper proposed a robust EM algorithm for GMM. This algorithm is based on the classical EM algorithm for GMM but uses a robust M-step for the estimation of the model parameters. More precisely, an outlier detection algorithm is used to attribute weights to each sample, reducing the impact of outliers on the parameter estimates. The proposed approach was used for data imputation, and showed interesting performance when compared to other classical methods, such as KNN imputation or MICE. The good imputation results obtained with the proposed method can be improved by carefully choosing the strategy used for the robust estimation, *i.e.*, by choosing the value of the threshold separating outliers and inliers. Another way of building robust estimation algorithms is to consider a mixture of distributions that take into account the presence of outliers, *e.g.*, compound Gaussian [17] or elliptical [18] distributions. It would be interesting to 1) compare the proposed approach with these methods and 2) investigate the interest of the modified BIC for these methods.

References

[1] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, 2nd ed. John Wiley & Sons, Inc. Hoboken, NJ, USA, 2002.

[2] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, pp. 1487–1509, 2020.

[3] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[4] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays ," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 06 2001.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, 1977.

[6] F. Mouret, M. Albughdadi, S. Duthoit, D. Kouamé, G. Rieu, and J.-Y. Tourneret, "Reconstruction of Sentinel-2 derived time series using robust Gaussian mixture models — Application to the detection of anomalous crop development," *Comput. Electron. Agric.*, vol. 198, p. 106983, 2022.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, Mar. 2012.

[8] E. Wit, E. v. d. Heuvel, and J.-W. Romeijn, "'all models are wrong...': an introduction to model uncertainty," *Stat. Neerl.*, vol. 66, no. 3, pp. 217–236, 2012.

[9] S. Tadjudin and D. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 1, pp. 439–445, 2000.

[10] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. NIPS*, vol. 12, Denver, CO, USA, Nov. 1999, pp. 582–588.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Survey of anomaly detection," *ACM Comput. Surveys*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[12] B. Schölkopf, K. Tsuda, and J.-P. Vert, *Kernel methods in computational biology*. Cambridge, Mass.: MIT Press, 2004.

[13] C. C. Aggarwal, *Outlier Analysis*, 2nd ed. Cham: Springer International Publishing, 2017.

[14] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Statist.*, vol. 6, no. 2, pp. 461 – 464, 1978.

[15] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 502–519, 2007.

[16] F. Mouret, M. Albughdadi, S. Duthoit, D. Kouamé, G. Rieu, and J.-Y. Tourneret, "Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and SAR time series," *Remote Sens.*, vol. 13, no. 5, p. 956, Mar 2021.

[17] A. Hippert-Ferrer, M. El Korso, A. Breloy, and G. Ginolhac, "Robust low-rank covariance matrix estimation with a general pattern of missing values," *Signal Processing*, vol. 195, p. 108460, 2022.

[18] F. Mouret, A. Hippert-Ferrer, F. Pascal, and J.-Y. Tourneret, "A robust and flexible EM algorithm for mixtures of elliptical distributions with missing data," *Under review*, 2022.