A Statistical Signal Processing Approach to Clustering over Compressed Data

Elsa DUPRAZ, Dominique PASTOR, Francois-Xavier SOCHELEAU
IMT Atlantique, Lab-STICC, Univ. Bretagne Loire, Brest, France

***Abstract*—In this paper, we consider a network of sensors in which a fusion center applies a clustering method over the sensor measurements. In order to limit their energy consumption, the sensors transmit their measurements in a compressed form. This paper proposes a novel clustering algorithm that applies directly over compressed data, and that does not require the knowledge of the number of clusters. The proposed algorithm is based on a new cost function for centroid estimation, and a theoretical analysis shows that the cluster centroids are the only minimizers of this cost function. The clustering algorithm then estimates the cluster centroids by looking for the minimizers of the cost function, even when their number is unknown. The proposed algorithm shows performance close to that of the K-means algorithm over compressed data, without need to know the number of clusters.**

I. INTRODUCTION

Wireless sensor networks are now employed in various applications in military, environmental, or telecommunication domains, see [1] for a survey. In most of these applications, a fusion center has to carry out a given processing or learning task over the collected sensor measurements. In this paper, we consider clustering as the learning task that should be realized by the fusion center. The objective of clustering is to separate the sensor measurements into clusters such that measurements assigned to the same cluster are close to each other and far from measurements belonging to other clusters. Several clustering algorithms such as DB-SCAN [2], OPTICS [3], and the very popular K-means [4], have been proposed in the literature.

In order to increase the lifetime of a network, the sensors must have very low energy consumption. The communication system that allows the sensors to transmit their data to the fusion center is responsible for the most important part of their energy consumption. In order to lower this energy consumption, the data should be transmitted in a compressed form. Since the objective of the fusion center is not to reconstruct all the sensor measurements but only to cluster them, we would like to apply the clustering algorithm directly over the compressed data. This would avoid complex decoding operations at the fusion center.

Clustering over compressed measurements was investigated recently in [5], [6] for the K-means algorithm only. However, one of the main limitations of the K-means algorithm is that it requires prior knowledge of the number of clusters. The objective of this paper is thus to propose a clustering algorithm that can be applied to compressed measurements, and that does not need to know the number of clusters.

In model-based clustering algorithms [7], the measurement vectors that belong to a cluster are modeled as the cluster centroid corrupted by additive Gaussian noise. It is further assumed that the cluster centroids and the noise variance are unknown, but that the number of clusters is known. The clustering algorithm we propose is based on the same Gaussian model as in [7]. However, as opposed to [7], our approach assumes that the number of clusters is unknown, and that the noise variance is known. This assumption is reasonable in many applications, for which the noise variance may be either determined from the physical characteristics of the sensors, or estimated locally by the sensors [8], [9].

From the Gaussian model, we introduce a new cost function for clustering over compressed data. Our cost function generalizes the one introduced in [10] for clustering over non-compressed data. By a theoretical analysis, we show that, under asymptotic conditions, the compressed cluster centroids are the only minimizers of the introduced cost function. It is worth mentioning that such a theoretical analysis is new and was not carried out in [10].

The cost function we introduce depends on a weight function and our theoretical analysis shows that this weight function has to satisfy certain properties. In this paper, we choose the weight function as the p-value of a Wald test [11]. This p-value is shown to satisfy the required properties. We then propose an algorithm that permits to estimate the cluster centroids by computing the minimizers of the cost function, even when the number of minimizers is *a priori* unknown. The full clustering algorithm we derive from this approach shows performance close to that of the K-means algorithm over compressed data and does not need to know the number of clusters.

The outline of the paper is as follows. Section II describes the signal model considered for the measurements. Section III introduces our cost function for clustering and provides the theoretical analysis. Section IV gives the expression of the considered weight function. Section V describes the clustering algorithm over compressed data. Section VI shows the simulation results.

II. SIGNAL MODEL

Consider a network of N sensors in which each sensor $n \in \{1, \cdots, N\}$ collects a measurement vector $\boldsymbol{Y}_n$. The vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ are assumed to be N independent and identically distributed (i.i.d.) d-dimensional random Gaussian vectors. We consider that the measurement vectors are split into K clusters defined by K deterministic centroids $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, with $\boldsymbol{\theta}_k \in \mathbb{R}^d$ for each $k \in \{1, 2, \ldots, K\}$. Accordingly, we assume that for each $n \in \{1, \ldots, N\}$, there exists $k \in \{1, 2, \ldots, K\}$ such

that $\boldsymbol{Y}_n \sim \mathcal{N}(\boldsymbol{\theta}_k, \sigma^2 \mathbf{I}_d)$ and we say that $\boldsymbol{Y}_n$ belongs to cluster k. In the above model, the noise variance σ^2 is the same for all the measurement vectors $\boldsymbol{Y}_n$. In the following, we assume that the value of σ^2 is known prior to clustering.

Each sensor applies a sensing matrix $A \in \mathbb{R}^{m \times d}$ to its measurement vector $\boldsymbol{Y}_n$. This produces compressed observations $\boldsymbol{Z}_n = A\boldsymbol{Y}_n$, $n \in \{1, \dots, N\}$, that are then transmitted to the fusion center. Here, as a first step, we assume that the fusion center directly observes the compressed observations $\boldsymbol{Z}_n$, while in practical situations, the received observations may be corrupted by some quantization or channel noise. More complex transmission models will be considered in future works. In the present setting, $\boldsymbol{Z}_n \sim \mathcal{N}(\boldsymbol{\phi}_k, \sigma^2 AA^T)$, where $\boldsymbol{\phi}_k = A\boldsymbol{\theta}_k$ represents the compressed centroids. Here, the matrix A is the same for all the sensors and performs compression whenever $m < d$. The theoretical analysis presented in the paper applies whatever the considered matrix, and in our simulations we will consider several different choices for A.

In the following, we assume that the centroids $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$, and their compressed versions $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K$, are unknown. The objective of this paper is to propose an algorithm that groups the N compressed measurement vectors $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_N$ into clusters. The first step of our algorithm consists of estimating the compressed centroids $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K$, as we now describe.

III. Centroid estimation

In this section, we introduce a new cost function for the estimation of the compressed centroids $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K$ from the measurement vectors $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$. We then present our theoretical analysis that shows that the compressed centroids $\boldsymbol{\phi}_k$ are the only minimizers of the introduced cost function.

A. Cost Function for centroid estimation

Consider an increasing, convex and differentiable function $\rho : \mathbb{R} \to \mathbb{R}$ that verifies $\rho(x) = 0 \Rightarrow x = 0$. Given an $m \times m$ positive-definite matrix C, define the Mahalanobis norm $\nu_C(\boldsymbol{z}) = \sqrt{\boldsymbol{z}^T C^{-1} \boldsymbol{z}}$ for any $\boldsymbol{z} \in \mathbb{R}^m$. First assume that the number K of clusters is known, and consider the following cost function for the estimation of the compressed centroids:

$$J(\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K) = \sum_{k=1}^{K} \sum_{n=1}^{N} \rho(\nu_C(\boldsymbol{z}_n - \boldsymbol{\phi}_k)). \tag{1}$$

This cost function generalizes the one introduced in [10] for centroid estimation when K is known. In [10], the clustering was performed over i.i.d. Gaussian vectors, and the particular case $C = I$ was considered. In contrast, our analysis assumes a general positive-definite matrix C, which will permit to take into account the correlation introduced by the compression. In addition, [10] only considers the particular case $\rho(x) = 1 - \exp(-\beta x)$, where β is a parameter that has to be chosen empirically. On the opposite, here, we consider a class of possible functions ρ, and the properties that these functions should verify will be exposed in the subsequent theoretical analysis. Note that the approach of [10] was inspired by the M-estimation theory [12].

In order to estimate the centroids, we want to minimize the cost function (1) with respect to the compressed centroids $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K$. Since J is convex, ρ is differentiable, and C is invertible, each centroid $\boldsymbol{\phi}_k$ should verify

$$\sum_{n=1}^{N} (\boldsymbol{z}_n - \boldsymbol{\phi}_k) w(\nu_C(\boldsymbol{z}_n - \boldsymbol{\phi}_k)) = 0 \tag{2}$$

where the function $w : \mathbb{R} \to \mathbb{R}$ is called the weight function and is given by $w = \rho'$. Solving (2) amounts to looking for the fixed-points $h_N(\boldsymbol{\phi}) = \boldsymbol{\phi}$ of the function h_N defined as

$$h_N(\boldsymbol{\phi}) = \frac{\sum_{n=1}^{N} w(\nu_C(\boldsymbol{z}_n - \boldsymbol{\phi}))\boldsymbol{z}_n}{\sum_{n=1}^{N} w(\nu_C(\boldsymbol{z}_n - \boldsymbol{\phi}))}, \boldsymbol{\phi} \in \mathbb{R}^m. \tag{3}$$

In the following, we show the following result: the centroids are the only fixed points of h_N under asymptotic conditions and given that the weight function w verifies certain properties. In addition and perhaps surprisingly, the expression of h_N (3) depends neither on the considered cluster k, nor on the number of clusters K. The foregoing suggests that, even when K is unknown, estimating the centroids can be performed by seeking the fixed points of h_N defined in (3). This claim is theoretically and experimentally verified below.

B. Fixed-point analysis

The following proposition shows that the compressed centroids $\boldsymbol{\phi}_k$ are the only fixed points of h_N (3).

Proposition 1. *Let $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K$ be K pairwise different elements of $\mathbb{R}^m$. For each $k \in [\![1, K]\!]$, suppose that $\boldsymbol{Z}_{k,1}, \dots, \boldsymbol{Z}_{k,N_k} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\phi}_k, \sigma^2 AA^T)$ and set $N = \sum_{k=1}^{K} N_k$. Assume that there exist $\alpha_1, \dots, \alpha_K \in (0, 1)$ such that $\lim_{N\to\infty} N_k/N = \alpha_k$. Also assume that the function w is such that if $\boldsymbol{Z}(\boldsymbol{\xi}) \sim \mathcal{N}(\boldsymbol{\xi}, \mathbf{I}_m)$ with $\boldsymbol{\xi} \in \mathbb{R}^m$, then*

$$\mathbb{E}\left[\, w(\nu_C(\boldsymbol{Z}(0)))\boldsymbol{Z}(0) \,\right] = 0, \tag{4}$$

$$\lim_{\|\boldsymbol{\xi}\|\to\infty} \mathbb{E}\left[\, w(\nu_C(\boldsymbol{Z}(\boldsymbol{\xi}))) \,\right] = 0, \tag{5}$$

$$\lim_{\|\boldsymbol{\xi}\|\to\infty} \mathbb{E}\left[\, w(\nu_C(\boldsymbol{Z}(\boldsymbol{\xi})))\boldsymbol{Z}(\boldsymbol{\xi}) \,\right] = 0. \tag{6}$$

Then, for any $i \in [\![1, K]\!]$ and any $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi}_i$,

$$\lim_{\forall k \neq i, \|\boldsymbol{\phi}_k - \boldsymbol{\phi}_i\| \to \infty} \Big(\lim_{N\to\infty} \big(h_N(\boldsymbol{\phi}) - \boldsymbol{\phi} \big) \Big) = 0 \quad \textit{iff} \quad \boldsymbol{\phi} = \boldsymbol{\phi}_i,$$

almost surely.

Proof: The proof is left for a longer version of the paper. ∎

Proposition 1 shows that the centroids are the unique fixed points of the function h_N, when the sample size and the distances between centroids tend to infinity. This result means that, at least asymptotically, no other vector than the centroids can be a fixed point of h_N. Based on this result and on the fact that the expression of h_N does not depend on k nor on K, we now assume that K is unknown. In the following, after introducing a particular weight function w, we propose an algorithm that estimates the centroids by computing all the fixed points of h_N.

IV. Weight function

The weight function $w(x) = \beta \exp(-\beta x)$ considered in [10] verifies the properties required in Proposition 1. However, in this weight function, the parameter β must be chosen empirically and its optimal value varies with the dimension m and with the noise parameters. A poor choice of β can dramatically impact the performance of the clustering algorithm proposed in [10]. On the opposite, here, we propose a new weight function whose expression is known whatever the dimension and noise parameters.

A. The Wald Test and its p-value

In addition to avoiding empirical parameters, we want to define a weight function that makes sense in a clustering problem. For this sake, consider the problem of testing whether an observation $\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{\phi}_{(i)}, \sigma^2 AA^T)$ belongs to cluster k, where $\boldsymbol{\phi}_{(i)}$ denotes the unknown centroid of the observation $\boldsymbol{Z}_i$. For this test, we usually do not have access to the true centroid $\boldsymbol{\phi}_k$, but only to an estimate $\hat{\boldsymbol{\phi}}_k$. For now, assume that $\widehat{\boldsymbol{\phi}}_k \sim \mathcal{N}(\boldsymbol{\phi}_k, r^2 AA^T)$, where r^2 represents the centroid estimation variance. The value of r^2 can be calculated analytically and its expression will be given when describing the clustering algorithm. In the following, we assume that $\widehat{\boldsymbol{\phi}}_k$ is independent of $\boldsymbol{Z}_i$. Now consider the following testing problem

$$\begin{cases} \text{Observation: } \boldsymbol{Z}_i - \widehat{\boldsymbol{\phi}}_k \sim \mathcal{N}(\boldsymbol{\phi}_{(i)} - \boldsymbol{\phi}_k, (\sigma^2 + r^2) AA^T), \\ \text{Hypotheses: } \begin{cases} \mathcal{H}_0 : \boldsymbol{\phi}_{(i)} - \boldsymbol{\phi}_k = 0, \\ \mathcal{H}_1 : \boldsymbol{\phi}_{(i)} - \boldsymbol{\phi}_k \neq 0. \end{cases} \end{cases} \tag{7}$$

This problem consists of testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\phi}_{(i)} - \boldsymbol{\phi}_k = 0$ (observation $\boldsymbol{Z}_i$ belongs to cluster k) against its alternative $\mathcal{H}_1 : \boldsymbol{\phi}_{(i)} - \boldsymbol{\phi}_k \neq 0$ (observation $\boldsymbol{Z}_i$ belongs to cluster k).

A test $\mathfrak{T}$ is any measurable map from $\mathbb{R}^m$ to $\{0, 1\}$. Given $\boldsymbol{z} \in \mathbb{R}^m$, the value $\mathfrak{T}(\boldsymbol{z})$ returned by $\mathfrak{T}$ is the index of the hypothesis considered to be true and we say that $\mathfrak{T}$ accepts $\mathcal{H}_0$ (resp. $\mathcal{H}_1$) at $\boldsymbol{z}$ if $\mathfrak{T}(\boldsymbol{z}) = 0$ (resp. $\mathfrak{T}(\boldsymbol{z}) = 1$). Given $\alpha \in (0, 1)$, let $\mu(\alpha)$ be the unique real value λ such that $Q_{m/2}(0, \lambda) = \alpha$ where $Q_{m/2}$ is the Generalized Marcum Function [14]. By setting $C = (\sigma^2 + r^2) AA^T$ and according to [11, Definition III & Proposition III, p. 450], the test defined for any $\boldsymbol{z} \in \mathbb{R}^m$ as

$$\mathfrak{T}_{\mu(\alpha)}(\boldsymbol{z}_i - \widehat{\boldsymbol{\phi}}_k) = \begin{cases} 0 & \text{if} \quad \nu_C(\boldsymbol{z}_i - \widehat{\boldsymbol{\phi}}_k) \leqslant \mu(\alpha) \\ 1 & \text{if} \quad \nu_C(\boldsymbol{z}_i - \widehat{\boldsymbol{\phi}}_k) > \mu(\alpha). \end{cases} \tag{8}$$

guarantees a false alarm probability α for the problem described by (7). The hypothesis test $\mathfrak{T}_{\mu(\alpha)}$ is a Wald test [11] and it was shown to be optimal with respect to several optimality criteria, see [15] for more details.

We can show that the p-value of the test $\mathfrak{T}_{\mu(\alpha)}$ is given for any $\boldsymbol{z} \in \mathbb{R}^m$ by:

$$\tilde{w}(\boldsymbol{z}) = Q_{m/2}\left(0, \nu_C(\boldsymbol{z})\right). \tag{9}$$

The proof is omitted due to the lack of space. A p-value function can be seen as a measure of the plausibility of the null hypothesis $\mathcal{H}_0$ given the observation [16, Sec. 3.3] As a result, in our case, the p-value $\tilde{w}(\boldsymbol{z}_i - \widehat{\boldsymbol{\phi}}_k)$ measures the plausibility that the measurement vector $\boldsymbol{z}_i$ belong to cluster k. It can be shown that the weight function defined from (9) by $w(x) = Q_{m/2}(0, x)$ verifies the properties required in Proposition 1, and this is why we will choose it as our weight function in the remaining of the paper. It is worth noting that the p-value does not depend on any empirical parameter, except the false alarm probability α. However, α does not depend on the dimension nor on the noise distribution, and in our simulations, we observed that this parameter does not influence much the performance of the clustering algorithm we now propose.

V. Clustering algorithm

This section describes our clustering algorithm CENTRE-X that applies to compressed data. The objective of the algorithm is to divide the set of received compressed vectors $\mathcal{Z} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_N\}$ into K clusters, when K is unknown *a priori*. The first step of the algorithm consists of estimating the cluster centroids by looking for all the fixed-points of the function h_N (3).

A. Initialization

Initialize by $\Phi = \{\varnothing\}$ the set of centroids estimated by the algorithm. Also, initialize by $\mathcal{M} = \{\varnothing\}$ the set of vectors $\boldsymbol{Z}_k$ that are considered as marked, where a marked vector cannot be used anymore to initialize the estimation of a new centroid, as described below.

B. Centroid estimation

The centroids are estimated one after the other, until $\mathcal{M} = \mathcal{Z}$. When the algorithm has already estimated k centroids, we have $\Phi = \{\widehat{\boldsymbol{\phi}}_1, \cdots, \widehat{\boldsymbol{\phi}}_k\}$. In order to estimate the $k+1$-th centroid, the algorithm picks a measurement vector $\boldsymbol{Z}_\star$ at random in the set $\mathcal{Z} \setminus \mathcal{M}$ and initializes the estimation process with $\widehat{\boldsymbol{\phi}}_{k+1}^{(0)} = \boldsymbol{Z}_\star$. In order to estimate $\widehat{\boldsymbol{\phi}}_{k+1}$ as a fixed point of h_N (3), the algorithm recursively computes $\widehat{\boldsymbol{\phi}}_{k+1}^{(\ell+1)} = h_N(\widehat{\boldsymbol{\phi}}_{k+1}^{(\ell)})$ [12]. At the first iteration, the function w in h_N is given from (9) calculated with $r = \sigma^2$, because $\widehat{\boldsymbol{\phi}}_{k+1}^{(0)}$ is initialized with a compressed observation. From iteration 2, w is calculated with $r = 0$, which implicitely assumes that $\widehat{\boldsymbol{\phi}}_{k+1}^{(\ell)}$ is very close to the true centroid $\boldsymbol{\phi}_{k+1}$. These choices of w are heuristic, but they lead to a good clustering performance in our simulations. The recursion stops when $\|\widehat{\boldsymbol{\phi}}_{k+1}^{(\ell+1)} - \widehat{\boldsymbol{\phi}}_{k+1}^{(\ell)}\|_2 \leq \epsilon$, where ϵ is a stopping condition. The newly estimated centroid is given by $\widehat{\boldsymbol{\phi}}_{k+1} = \widehat{\boldsymbol{\phi}}_{k+1}^{(L)}$, where L represents the final iteration. To finish, the set of estimated centroids is updated as $\Phi = \Phi \cup \{\widehat{\boldsymbol{\phi}}_{k+1}\}$.

Once the centroid $\widehat{\boldsymbol{\phi}}_{k+1}$ is estimated, the algorithm marks all the vectors that belong to cluster $k+1$. For this, the algorithm applies a Wald test to each $(\boldsymbol{Z}_i - \widehat{\boldsymbol{\phi}}_{k+1}) \sim \mathcal{N}(\boldsymbol{\phi}_{(i)} - \widehat{\boldsymbol{\phi}}_{k+1}, \sigma^2 AA^T)$, $i \in \{1, \cdots, N\}$. This corresponds to applying the Wald test (8) with $r = 0$. All the observations $\boldsymbol{Z}_i$ that accept the null hypothesis under this test are grouped into the set $\mathcal{M}_{k+1}$. The set of marked vectors is then updated as $\mathcal{M} \leftarrow \mathcal{M} \cup \{\boldsymbol{Z}_\star\} \cup \mathcal{M}_{k+1}$. Note that the measurement vector $\boldsymbol{Z}_\star$, which serves for initialization, is also marked in order to avoid initializing again with the same vector. If $\mathcal{M} \neq \mathcal{Z}$,

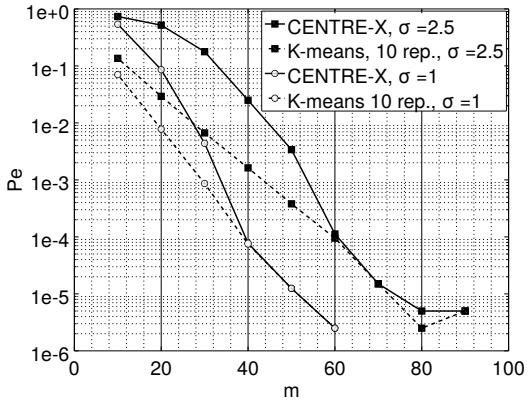

Fig. 1. Classification error probabilities of K-means with 10 replicates and CENTRE-X, for two values of σ, for non-sparse centroids

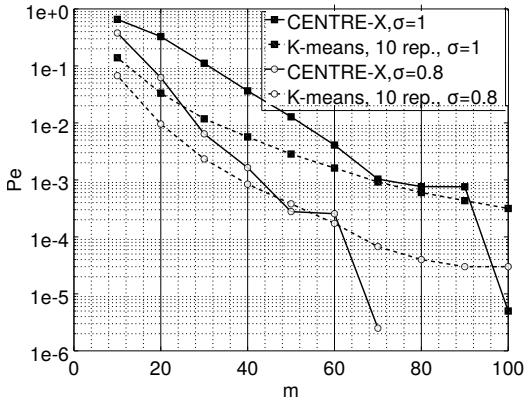

Fig. 2. Classification error probabilities of K-means with 10 replicates and CENTRE-X, for two values of σ, for sparse centroids

the algorithm estimates the next centroid $\widehat{\boldsymbol{\theta}}_{k+2}$. Otherwise, the algorithm moves to the fusion step.

C. Fusion

Once $\mathcal{M} = \mathcal{Z}$ and, say, K' centroids have been estimated, the algorithm applies a so-called fusion step to identify the centroids that may have been estimated several times from different initializations. At this step, the algorithm applies a Wald hypothesis test $\mathfrak{T}_{\mu(\alpha)}$ to $(\widehat{\boldsymbol{\phi}}_{k_1} - \widehat{\boldsymbol{\phi}}_{k_2}) \sim \mathcal{N}(\boldsymbol{\phi}_{k_1} - \boldsymbol{\phi}_{k_2}, 2\sigma_{k,l}^2 AA^T)$, for all pairs $(\widehat{\boldsymbol{\phi}}_{k_1}, \widehat{\boldsymbol{\phi}}_{k_2}) \in \Phi \times \Phi$ such that $k_1 < k_2$. This Wald test can be derive in a straightforward way form Section IV-A, and the expression of $\sigma_{k,l}$ is given in [13]. When $\mathfrak{T}_{\mu(\alpha)}(\widehat{\boldsymbol{\phi}}_{k_1} - \widehat{\boldsymbol{\phi}}_{k_2}) = 0$, the algorithm sets $\widehat{\boldsymbol{\phi}}_{k_1} = \frac{\widehat{\boldsymbol{\phi}}_{k_1} + \widehat{\boldsymbol{\phi}}_{k_2}}{2}$ and removes $\widehat{\boldsymbol{\phi}}_{k_2}$ from Φ. At the end, the number of centroids K is set as the cardinal of the final Φ and the elements are re-indexed in order to get $\Phi = \{\widehat{\boldsymbol{\phi}}_1, \cdots \widehat{\boldsymbol{\phi}}_K\}$.

D. Classification

Denote by $\mathcal{C}_k$ the set of measurement vectors assigned to cluster k. Each vector $\boldsymbol{Z}_i \in \mathcal{Z}$ is assigned to the cluster $\mathcal{C}_{k'}$ whose centroid $\widehat{\boldsymbol{\phi}}_{k'} \in \Phi$ is the closest to $\boldsymbol{Z}_i$, *i.e.*, $\widehat{\boldsymbol{\phi}}_{k'} = \arg\min_{\widehat{\boldsymbol{\phi}} \in \Phi} \|\boldsymbol{Z}_i - \widehat{\boldsymbol{\phi}}\|$. Here, using this condition instead of an hypothesis test forces each measurement vector to be assigned to a cluster.

VI. Simulation Results

This section evaluates the performance of the CENTRE-X algorithm from Monte Carlo simulations. In all our simulations, we consider $K = 4$, $d = 100$, $\alpha = 10^{-3}$, and the observation vectors $\boldsymbol{Y}_n$ that belong to cluster k are generated according to the model $\boldsymbol{Y}_n \sim \mathcal{N}(\boldsymbol{\phi}_k, \sigma^2 \mathbf{I}_d)$, where σ^2 is the noise variance. In the following, we consider various values of m, and, for each value of m, we measure the classification error probability over 1000 trials. In all the considered setups, we compare the performance or our algorithm against the performance of a K-means algorithm with 10 replicates (in order to lower initialization issues), and provided with the correct value of K. We consider the following two setups.

A. Non-sparse centroids

We first assume that new centroids are generated at each trial as $\boldsymbol{\theta}_k \sim \mathcal{N}(0, b^2 \mathbf{I}_d)$ with $b = 2$. In this case, the sensing matrix A is constructed so as to randomly select components of $\boldsymbol{Y}_n$, that is each row of A contains exactly one value 1, and 0 elsewhere. The classification error probabilities of CENTRE-X and of K-means with 10 replicates are represented in Figure 1, for $\sigma = 1$ and $\sigma = 2.5$. In both cases, we see that the error probability of CENTRE-X is close but a little degraded compared to K-means. However, in our experiments, K-means is evaluated in the most favorable case as it knows the number of clusters K and is repeated several times, which is not the case with our algorithm.

B. Sparse centroids

We now assume that the centroids are sparse. At each trial, a new set of centroids is generated, and each individual component of each centroid is generated as $\theta_{k,j} \sim \mathcal{N}(0, b^2)$ $(b = 2)$, with probability 0.2, and is equal to 0 otherwise. The matrix A is generated once for all the trials for each considered value of m. This corresponds to random projections, with $A_{i,j} \sim \mathcal{N}(0, md)$, $i = 1, \cdots, m$ and $j = 1 \cdots, d$. The classification error probabilities of CENTRE-X and of K-means with 10 replicates are represented in Figure 2, for $\sigma = 0.8$ and $\sigma = 1$. As for the case of non-sparse centroids, our algorithm only shows a limited performance loss with respect to K-means. Compared to K-means, CENTRE-X does not require prior knowledge of the number of clusters and does not suffer from initialization issues, at the price of a limited performance degradation.

VII. Conclusion

In this paper, we proposed a new clustering algorithm that applies over compressed data and that does not need to know the number of clusters. The clustering algorithm we proposed looks for the minimizers of a new cost function, and a theoretical analysis shows that the cluster centroids are the only minimizers of this cost function. Our clustering algorithm shows a little performance degradation compared to K-means, but without need to know the number of clusters.

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD*, 1996.

[3] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *ACM Sigmod record*, vol. 28, no. 2. ACM, 1999, pp. 49–60.

[4] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[5] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for K-means clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, 2015.

[6] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval, "Compressive K-means," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 6369–6373.

[7] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[8] P. Huber and E. Ronchetti, *Robust Statistics, second edition*. John Wiley and Sons, 2009.

[9] D. Pastor and F. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, 2012.

[10] K.-L. Wu and M.-S. Yang, "Alternative c-means clustering algorithms," *Pattern recognition*, vol. 35, no. 10, pp. 2267–2278, 2002.

[11] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, vol. 54, no. 3, pp. 426 – 482, Nov. 1943.

[12] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[13] D. Pastor, E. Dupraz, and F.-X. Socheleau, "Decentralized clustering based on robust estimation and hypothesis testing," *arXiv preprint arXiv:1706.03537*, 2017.

[14] Y. Sun, A. Baricz, and S. Zhou, "On the Monotonicity, Log-Concavity, and Tight Bounds of the Generalized Marcum and Nuttall Q-Functions," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1166 – 1186, Mar. 2010.

[15] D. Pastor and Q.-T. Nguyen, "Random Distortion Testing and Optimality of Thresholding Tests," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4161 – 4171, Aug. 2013.

[16] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses, 3rd edition*. Springer, 2005.