OUTLIER-AWARE ROBUST CLUSTERING

Pedro A. Forero, Vassilis Kekatos, and Georgios B. Giannakis

Dept. of Electrical & Computer Engrg.
University of Minnesota
Minneapolis, MN 55455, USA
Emails:{forer002,kekatos,georgios}@umn.edu

ABSTRACT

Clustering is a basic task in a variety of machine learning applications. Partitioning a set of input vectors into compact, well-separated subsets can be severely affected by the presence of model-incompatible inputs called outliers. The present paper develops robust clustering algorithms for jointly partitioning the data and identifying the outliers. The novel approach relies on translating scarcity of outliers to sparsity in a judiciously defined domain, to robustify three widely used clustering schemes: hard K-means, fuzzy K-means, and probabilistic clustering. Cluster centers and assignments are iteratively updated in closed form. The developed outlier-aware algorithms are guaranteed to converge, while their computational complexity is of the same order as their outlier-agnostic counterparts. Preliminary simulations validate the analytical claims.

Index Terms— K-means, robust clustering, convex relaxation, block coordinate descent, expectation maximization.

1. INTRODUCTION

Clustering aims to partition a set of observations into disjoint subsets, called clusters, such that observations assigned to the same cluster are similar in some sense. Working with unlabeled data and under minimal assumptions makes clustering a universal tool for revealing data relations in a gamut of applications such as DNA microarrays, bioinformatics, social networks, image processing, and data mining.

From the multitude of non-probabilistic clustering methods, the K-means algorithm is among the most popular ones [6, 7]. Either in its *hard* form or its *soft* (fuzzy) version, K-means is computationally simple. However, points lying relatively far from all other points can deteriorate its performance in terms of estimating the cluster centers and point assignments. Such points are called outliers and emerge either due to reading errors or because they belong to rarely-seen, and thus extremely informative, clusters. Past attempts to robustify K-means include: the noise clustering scheme, which adds an extra "outlier-cluster" and heuristically assumes its center to be equidistant from all input points [3]; and the α-cut fuzzy K-means algorithm that parsimoniously trims the core data of each cluster [9].

Probabilistic clustering assumes that the observed data are drawn from a probability density function (pdf) following a mixture model, where each component corresponds to a cluster [7]. Under the maximum likelihood (ML) framework, the expectation-maximization (EM) algorithm can be used to estimate the parameters of the data pdf and automatically provide the probability of a point being drawn from each cluster. In the presence of outliers, probabilistic clustering capabilities are still limited to mixture models with outlier-sensitive likelihood functions.

The contributions of this paper are: (i) a novel data model that explicitly includes outliers, and naturally lends itself to robust clustering criteria; (ii) outlier-aware K-means and probabilistic clustering algorithms based on a convex relaxation of the resultant non-smooth criteria, which permeates benefits of contemporary advances in compressive sensing to the clustering problem; and (iii) non-convex optimization solvers for robust clustering using block-coordinate descent (BCD) or EM iterations, with the iterates obtained in closed form, provably convergent to a stationary point, and capable of revealing the outliers present, at negligible excess cost relative to their non-robust counterparts. The simulations performed validate the performance of the clustering methods.

Notation: Lowercase (uppercase) boldface letters are reserved for column vectors (matrices), and calligraphic letters for sets; $(\cdot)^T$ denotes transposition; $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ stands for the multivariate Gaussian pdf with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$, while $\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{\Sigma})$ denotes the same pdf evaluated at $\mathbf{x}$.

2. ROBUSTIFYING CLUSTERING METHODS

2.1. Robustifying Hard K-Means

Given a set of p-dimensional vectors $\mathcal{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, let $\{\mathcal{X}_1, \ldots, \mathcal{X}_C\}$ be a *partition* of $\mathcal{X}$ with the subsets (clusters) $\mathcal{X}_c \subset \mathcal{X}$, $c = 1, \ldots, C$, being collectively exhaustive, mutually exclusive, and non-empty. Hard K-means seeks a partition for which vectors assigned to the same cluster are in some sense closer (in e.g., Euclidean distance) to each other when compared to vectors belonging to other clusters.

In the K-means setup, a cluster center $\mathbf{m}_c \in \mathbb{R}^p$ is introduced per cluster $\mathcal{X}_c$. Then, instead of comparing distances between pairs of points in $\mathcal{X}$, the distances $\|\mathbf{x}_n - \mathbf{m}_c\|_2^2$ are considered. Moreover, K-means introduces the unknown memberships u_{nc} defined to be 1 when $\mathbf{x}_n \in \mathcal{X}_c$ and 0 otherwise for all n, c. To guarantee a valid partition, the membership coefficients apart from being binary **(c1)** $u_{nc} \in \{0, 1\}$; they should also satisfy the constraints **(c2)** $\sum_{n=1}^{N} u_{nc} > 0$ for all c to preclude empty clusters; and **(c3)** $\sum_{c=1}^{C} u_{nc} = 1$ for all n, so that each vector is assigned to a cluster.

Under (c1)-(c3), consider the following data model which accounts explicitly for outliers

$$\mathbf{x}_n = \sum_{c=1}^{C} u_{nc}\mathbf{m}_c + \mathbf{o}_n + \mathbf{v}_n, \quad n = 1, \ldots, N \tag{1}$$

Work was supported by NSF grants CCF-0830480, 1016605, and ECCS-0824007, 1002180. Dr Kekatos' work was funded by the European Community's Seventh Framework Programme (grant FP7/2008-234914).

where the vector $\mathbf{o}_n$ is deterministically nonzero if $\mathbf{x}_n$ corresponds to an outlier, and $\mathbf{0}$ otherwise; and $\mathbf{v}_n$ is a zero-mean random vector capturing the deviation of $(\mathbf{x}_n - \mathbf{o}_n)$ from its cluster center.

The unknown parameters $\{u_{nc}\}$, $\{\mathbf{m}_c\}$, and $\{\mathbf{o}_n\}$ in (1) can be estimated using a least-squares (LS) approach as the minimizers of $\sum_{n=1}^{N} \left\| \mathbf{x}_n - \sum_{c=1}^{C} u_{nc}\mathbf{m}_c - \mathbf{o}_n \right\|_2^2$, which is equivalent to ML if $\mathbf{v}_n \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$. Even if u_{nc}'s were known, estimating $\{\mathbf{m}_c\}$ and $\{\mathbf{o}_n\}$ based solely on $\{\mathbf{x}_n\}$ would be a highly under-determined problem. Note however that most of the $\{\mathbf{o}_n\}$ are zero. This motivates the following criterion for hard clustering and identification of at most $s \in \{1, \ldots, N\}$ outliers

$$\min_{\mathbf{U}\in\mathcal{U}_1,\mathbf{M},\mathbf{O}} \sum_{n=1}^{N} \left\| \mathbf{x}_n - \sum_{c=1}^{C} u_{nc}\mathbf{m}_c - \mathbf{o}_n \right\|_2^2 \quad \text{s.to} \sum_{n=1}^{N} \mathrm{I}(\|\mathbf{o}_n\|_2 \neq 0) \leq s \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{N\times C}$ denotes the membership matrix with entries $[\mathbf{U}]_{n,c} := u_{nc}$; $\mathcal{U}_1$ is the set of all $\mathbf{U}$ matrices satisfying the constraints (c1)-(c3); $\mathbf{M} := [\mathbf{m}_1 \ \cdots \ \mathbf{m}_C]$; $\mathbf{O} := [\mathbf{o}_1 \ \cdots \ \mathbf{o}_N]$; and $\mathrm{I}(\cdot)$ denotes the indicator function.

Due to (c1) and (c3), each summand in the cost of (2) can be re-written as $\sum_{c=1}^{C} u_{nc}\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2$, which turns out to render (c2) redundant. The binary-alphabet constraint (c1) can be equivalently relaxed by the box constraint **(c4)** $u_{nc} \in [0, 1]$ for all n, c. Then, after defining $\mathcal{U}_2$ to be the set of all $\mathbf{U}$ matrices satisfying (c3)-(c4), the Lagrangian form of (2) becomes

$$\min_{\mathbf{U}\in\mathcal{U}_2,\mathbf{M},\mathbf{O}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc} \|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 + \lambda\, \mathrm{I}(\|\mathbf{o}_n\|_2 \neq 0) \tag{3}$$

where $\lambda \geq 0$ is an outlier-controlling parameter. For $\lambda = 0$, every $\mathbf{o}_n$ can be set to the (generally) nonzero value $\mathbf{x}_n - \mathbf{m}_c$ for any c and yield a zero optimum cost. Thus, all $\mathbf{x}_n$'s are declared as outliers for $\lambda = 0$. When $\lambda \to \infty$, the optimum $\mathbf{O}$ is zero, all the $\mathbf{x}_n$'s are deemed as inliers, and the problem in (3) interestingly reduces to the cost related to the *K-means algorithm*

$$\min_{\mathbf{U}\in\mathcal{U}_2,\mathbf{M}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc} \|\mathbf{x}_n - \mathbf{m}_c\|_2^2 . \tag{4}$$

Hence, K-means can be formulated based on model (1) once ignoring the $\mathbf{o}_n$ terms. Such a simplification together with the sensitivity of the square cost to large residuals explains K-means vulnerability to outliers [3].

Solving the clustering problem in (4) is known to be NP-hard, even for $C = 2$ [2]. Typically, a suboptimal solution is pursued using the K-means algorithm by alternately minimizing with respect to (wrt) one of the variables $\mathbf{M}$ and $\mathbf{U}$, while keeping the other one fixed, and iterating. K-means iterations are guaranteed to converge to a stationary point of (4).

As far as the more general, outlier-aware problem is concerned, iterations similar to K-means do not provide any convergence guarantees due to the non-smooth indicator function present in (3). Aiming at a practically feasible solver of (3), consider first that $\mathbf{U}$ is given. The optimization wrt $\{\mathbf{M}, \mathbf{O}\}$ remains non-convex due to $\sum_{n=1}^{N} \mathrm{I}(\|\mathbf{o}_n\|_2 \neq 0)$. Upon adopting the convex relaxation proposed in [4], this term can be approximated by $\sum_{n=1}^{N} \|\mathbf{o}_n\|_2$ to yield

$$\min_{\mathbf{U}\in\mathcal{U}_2,\mathbf{M},\mathbf{O}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc} \|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 + \lambda \sum_{n=1}^{N} \|\mathbf{o}_n\|_2 . \tag{5}$$

The robust hard K-means minimizes the cost in (5), which is convex in $\{\mathbf{M}, \mathbf{O}\}$, but jointly non-convex; see also Section 3.1 for an alternative solver. Note also that (5) resembles the group Lasso criterion used in [10] for recovering block-sparse vectors. This establishes a neat link between robust clustering and compressive sampling.

Remark 1. If the covariance matrix of $\mathbf{v}_n$ in (1) is known (call it $\mathbf{\Sigma}$), the Euclidean distance in (2)-(5) can be replaced by the Mahalanobis one, namely $\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_{\mathbf{\Sigma}^{-1}}^2$, where $\|\mathbf{z}\|_{\mathbf{H}}^2 := \mathbf{z}^T\mathbf{H}\mathbf{z}$.

2.2. Robustifying Soft Clustering Methods

The criteria in (3) and (5) yield *hard* membership assignments. However, soft memberships are well motivated because they: (i) can lead to improved clustering results [1]; (ii) identify ambiguous points lying at the intersection of clusters; and, (iii) can also provide a hard partition of $\mathcal{X}$ by assigning each $\mathbf{x}_n$ to the cluster $\widehat{c} := \arg\max_c u_{nc}$. To this end, two soft clustering approaches are robustified under the view of model (1): fuzzy K-means and probabilistic clustering.

The *fuzzy K-means* algorithm introduces a parameter $q \geq 1$ and by raising the u_{nc}'s in (4) to the q-th power can yield fractional memberships [1]. For $q = 1$, it boils down to the hard K-means. The outlier-aware fuzzy K-means scheme proposed here amounts to replacing $\mathbf{x}_n$ with its outlier-compensated counterpart $(\mathbf{x}_n - \mathbf{o}_n)$, and further promoting the sparsity constraints on $\mathbf{o}_n$. These steps lead to the following criterion

$$\min_{\mathbf{U}\in\mathcal{U}_2,\mathbf{M},\mathbf{O}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^q \left(\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 + \lambda \|\mathbf{o}_n\|_2 \right) \tag{6}$$

which can be optimized using the iterative scheme in Section 3.1.

An alternative approach to obtain soft memberships is via probabilistic clustering, which relies on a mixture model for the data pdf [7]. Suppose for simplicity that each $\mathbf{x}_n$ in (1) is drawn from a mixture of Gaussians, and view the u_{nc}'s as hidden random variables. Assume further that for each n, $p(\mathbf{x}_n|u_{nc} = 1) = \mathcal{N}(\mathbf{x}_n; \mathbf{m_c}+\mathbf{o}_n, \sigma^2\mathbf{I}_p)$, where σ^2 denotes the common variance. This implies that $p(\mathbf{x}_n) = \sum_{c=1}^{C} \pi_c \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c+\mathbf{o}_n, \sigma^2\mathbf{I}_p)$, where $\pi_c := \Pr(u_{nc} = 1)$. If the $\mathbf{x}_n$'s are independent, the log-likelihood of the entire data is

$$L(\mathcal{X}; \{\pi_c\}, \mathbf{M}, \mathbf{O}, \sigma^2) := \sum_{n=1}^{N} \log \left(\sum_{c=1}^{C} \pi_c \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c+\mathbf{o}_n, \sigma^2\mathbf{I}_p) \right).$$

To control the number of outliers (number of zero $\mathbf{o}_n$'s) suggests penalizing the negative log-likelihood to arrive at

$$\min_{\boldsymbol{\pi}\in\mathcal{P},\mathbf{M},\mathbf{O},\sigma^2>0} -L(\mathcal{X}; \boldsymbol{\pi}, \mathbf{M}, \mathbf{O}, \sigma^2) + \lambda \sum_{n=1}^{N} \|\mathbf{o}_n\|_2 \tag{7}$$

where $\mathcal{P}$ is the set of all $\boldsymbol{\pi} := [\pi_1 \cdots \pi_C]^T$ vectors satisfying $\boldsymbol{\pi}^T\mathbf{1}=1$, and $\boldsymbol{\pi}\geq\mathbf{0}$. The non-convex problem in (7) will be solved in Section 3.2 using EM iterations. Note that the u_{nc}'s in the probabilistic clustering criterion (1) are binary; while the posterior probabilities $\gamma_{nc} := \Pr(u_{nc}=1|\mathbf{x}_n)$ explicitly emerging in the expectation step of the algorithm, can be interpreted as soft memberships.

3. SOLVERS

3.1. Robust (Fuzzy) K-Means via Block Coordinate Descent

Consider first solving (6) for $q > 1$. Even though the cost in (6) is jointly non-convex, it is convex wrt each one of the block optimization variables $\mathbf{U}$, $\mathbf{M}$, and $\mathbf{O}$. This observation suggests a block-coordinate descent (BCD) solver. According to the BCD method, the cost is iteratively minimized wrt one of the three block variables at a time, while keeping the other two fixed. Specifically, let $\mathbf{U}^{(t)}$, $\mathbf{M}^{(t)}$, and $\mathbf{O}^{(t)}$ be the tentative solutions during the t-th iteration.

In the first step of the t-th iteration, (6) is minimized over $\mathbf{U}$, while $\mathbf{M}$ and $\mathbf{O}$ are set to $\mathbf{M}^{(t-1)}$ and $\mathbf{O}^{(t-1)}$, respectively. As in the fuzzy K-means update, the solution is provided in closed form for all n and c as (proofs are omitted due to space limitation)

$$u_{nc}^{(t)}=\left[\sum_{c'=1}^{C}\left(\frac{\|\mathbf{x}_n-\mathbf{m}_c^{(t-1)}-\mathbf{o}_n^{(t-1)}\|_2^2+\lambda||\mathbf{o}_n^{(t-1)}||_2}{\|\mathbf{x}_n-\mathbf{m}_{c'}^{(t-1)}-\mathbf{o}_n^{(t-1)}\|_2^2+\lambda||\mathbf{o}_n^{(t-1)}||_2}\right)^{\frac{1}{q-1}}\right]^{-1}. \quad (8)$$

In the second step, (6) is optimized wrt $\mathbf{M}$ for $\mathbf{U}=\mathbf{U}^{(t)}$ and $\mathbf{O}=\mathbf{O}^{(t-1)}$, while the problem decouples over the $\mathbf{m}_c$'s. Every $\mathbf{m}_c$ is the closed-form solution of a weighted least-squares problem

$$\mathbf{m}_c^{(t)}=\frac{\sum_{n=1}^{N}(u_{nc}^{(t)})^q\left(\mathbf{x}_n-\mathbf{o}_n^{(t-1)}\right)}{\sum_{n=1}^{N}(u_{nc}^{(t)})^q}. \quad (9)$$

In the third step, the task is to minimize (6) over $\mathbf{O}$ while $\mathbf{U}=\mathbf{U}^{(t)}$ and $\mathbf{M}=\mathbf{M}^{(t)}$. The optimization decouples over the index n, so that each $\mathbf{o}_n$ can be found as the minimizer of

$$\phi^{(t)}(\mathbf{o}_n):=\sum_{c=1}^{C}(u_{nc}^{(t)})^q\left(\|\mathbf{x}_n-\mathbf{m}_c^{(t)}-\mathbf{o}_n\|_2^2+\lambda\|\mathbf{o}_n\|_2\right). \quad (10)$$

The cost $\phi^{(t)}(\mathbf{o}_n)$ is convex and can be solved as a second-order cone program. Interestingly though, following the method in [4], we are able to show that its minimizer is provided in closed form too as

$$\mathbf{o}_n^{(t)}=\mathbf{r}_n^{(t)}\left[1-\frac{\lambda}{2\|\mathbf{r}_n^{(t)}\|_2}\right]_+, \text{ where} \quad (11)$$

$$\mathbf{r}_n^{(t)}:=\left(\sum_{c=1}^{C}(u_{nc}^{(t)})^q\left(\mathbf{x}_n-\mathbf{m}_c^{(t)}\right)\right)\left(\sum_{c=1}^{C}(u_{nc}^{(t)})^q\right)^{-1} \quad (12)$$

and $[x]_+:=\max\{x,0\}$ for all n. The update in (11) reveals two critical issues: (i) indeed $\phi^{(t)}(\mathbf{o}_n^{(t)})$ favors zero minimizers $\mathbf{o}_n^{(t)}$; and (ii) outliers can be identified. After updating the weighted residual $\mathbf{r}_n^{(t)}$, its norm is compared against the threshold $\lambda/2$. If it is larger, this input vector is deemed an outlier, and it is compensated by a nonzero $\mathbf{o}_n^{(t)}$. Otherwise, $\mathbf{o}_n^{(t)}$ is set to zero and $\mathbf{x}_n$ is being clustered as a regular point.

The developed *robust fuzzy K-means algorithm* (RFKM) entails the updates (8), (9), and (11). Regarding initialization, $\mathbf{M}$ is randomly chosen and $\mathbf{O}$ is set to zero. A careful counting of the computations involved shows that the total time-complexity is maintained in $\mathcal{O}(NCp)$, while $\mathbf{O}$ can be efficiently stored using sparse structures. The robust hard K-means (RHKM) algorithm corresponds to $q=1$, and can be derived by replacing (8) as follows: for every n, set the u_{nc} corresponding to $c_n^{(t)}:=\arg\min_c\|\mathbf{x}_n-\mathbf{m}_c^{(t-1)}-\mathbf{o}_n^{(t-1)}\|_2$ to 1, and the rest of the u_{nc}'s to zero.

The following proposition can be established by using the convergence results of: (i) the BCD method [8]; and, (ii) the majorization-minimization argument in [5].

Proposition 1. *The RKM algorithm converges to a stationary point of the cost function in* (6)*.*

3.2. An EM Algorithm for Robust Clustering

Problem (7) is approximately solved here using EM iterations. The algorithm, called hereafter *robust probabilistic clustering* (RPC), entails two steps during the t-th iteration. The *expectation step*, where the posterior probabilities γ_{nc} are updated for all $n,\ c$, as

$$\gamma_{nc}^{(t)}=\frac{\pi_c^{(t-1)}\mathcal{N}\left(\mathbf{x}_n;\mathbf{m}_c^{(t-1)}+\mathbf{o}_n^{(t-1)},\sigma^{2(t-1)}\mathbf{I}_p\right)}{\sum_{c'=1}^{C}\pi_{c'}^{(t-1)}\mathcal{N}\left(\mathbf{x}_n;\mathbf{m}_{c'}^{(t-1)}+\mathbf{o}_n^{(t-1)},\sigma^{2(t-1)}\mathbf{I}_p\right)} \quad (13)$$

and the *maximization step*, which involves the updates

$$\pi_c^{(t)}=\sum_{n=1}^{N}\gamma_{nc}^{(t)}/N, \text{ for all } c \quad (14)$$

$$\mathbf{m}_c^{(t)}=\left(\sum_{n=1}^{N}\gamma_{nc}^{(t)}\left(\mathbf{x}_n-\mathbf{o}_n^{(t-1)}\right)\right)\left(\sum_{n=1}^{N}\gamma_{nc}^{(t)}\right)^{-1}, \ \forall\, c \quad (15)$$

$$\sigma^{2(t)}=(Np)^{-1}\sum_{n=1}^{N}\sum_{c=1}^{C}\gamma_{nc}^{(t)}\|\mathbf{x}_n-\mathbf{m}_c^{(t)}-\mathbf{o}_n^{(t-1)}\|_2^2 \quad (16)$$

$$\mathbf{o}_n^{(t)}=\mathbf{r}_n^{(t)}\left[1-\frac{\lambda\sigma^{2(t)}}{\|\mathbf{r}_n^{(t)}\|_2}\right]_+, \text{ for all } n \text{ where} \quad (17)$$

$$\mathbf{r}_n^{(t)}:=\mathbf{x}_n-\sum_{c=1}^{C}\mathbf{m}_c^{(t)}\gamma_{nc}^{(t)}. \quad (18)$$

The algorithm cycling through updates (13)-(17) can be shown convergent to a stationary point of (7). Comparing the updates in (8)-(11) with those in (13)-(18), reveals that: (i) the probabilistic approach provides a solid interpretation of the soft memberships over the fuzzy one; and (ii) the thresholding operator is data-adaptive through $\sigma^{2(t)}$.

3.3. Weighted Solvers

The robust methods presented so far approximate $\mathrm{I}(\|\mathbf{o}_n\|_2\neq 0)$ by $\|\mathbf{o}_n\|_2$. It has been argued though that non-convex functions such as $\log(\|\mathbf{o}_n\|_2+\delta)$ for a small $\delta>0$ can be tighter approximants; see [4] and references therein. The same reasoning motivates replacing $\|\mathbf{o}_n\|_2$ by the penalty $\log(\|\mathbf{o}_n\|_2+\delta)$ in (5), (6), and (7) to enhance robustness. The involved optimizations wrt $\mathbf{o}_n$ are no longer convex, but single-iteration majorization-minimization updates (cf. [5]) can be derived [4]. It turns out that the resultant updates simply differ in the thresholding rules (11) and (17), where now λ becomes $\lambda_n^{(t)}=\lambda\left(\|\mathbf{o}_n^{(t-1)}\|_2+\delta\right)^{-1}$. The variability of $\lambda_n^{(t)}$ across iterations explains the adjective "weighted" in these solvers. Note that input vectors with rather large $\|\mathbf{o}_n^{(t-1)}\|_2$ here are compared to a smaller threshold, and are thus more likely to be declared as outliers.

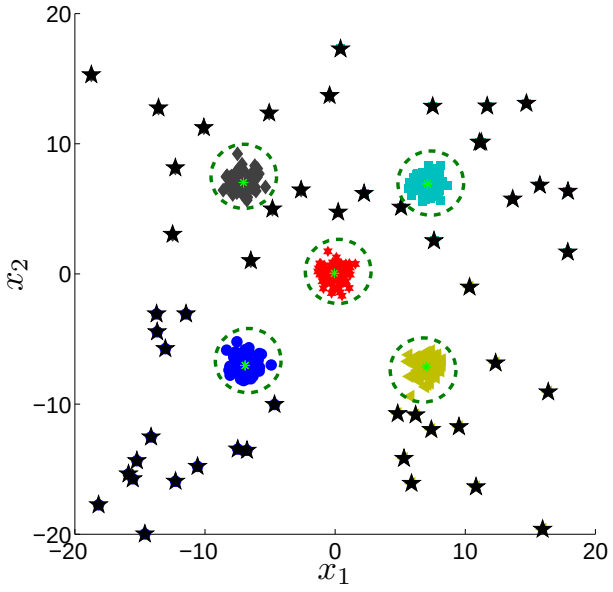

Fig. 1. Outlier-contaminated dataset.

4. SIMULATIONS

The developed robust clustering methods are evaluated through computer simulations. The synthetic dataset comprises vectors $\mathbf{x}_n \in \mathbb{R}^2$ belonging to $C = 5$ clusters with 100 vectors per cluster. The dataset is contaminated by adding a fixed set of outlier vectors shown as stars in Fig. 1. The figures of merit used are: (i) the squared-root of the mean-square error (RMSE) between the estimated cluster centers $\mathbf{M}$ and the cluster sample means; (ii) the probability of correctly identifying an outlier vector P_D; and (iii) the probability of falsely characterizing an inlier vector as an outlier P_{FA}.

The algorithms simulated were: hard K-means (HKM); fuzzy K-means (FKM); hard/fuzzy noise clustering HNC/FNC [3]; α-cut clustering [9]; probabilistic clustering under the assumptions of Section 3.2 (PC); robust hard/fuzzy K-means plain and weighted ((W-)RHKM/(W-)RFKM); and robust probabilistic clustering plain and weighted (W-)RPC. The parameter q is set to $q = 2$ for all fuzzy clustering algorithms. In each Monte Carlo run, the cluster centers were randomly initialized (at the same value for all algorithms), and $\mathbf{O}$ was set to zero. In all probabilistic clustering algorithms, the prior probabilities are set to $\boldsymbol{\pi}^{(0)} = \mathbf{1}/C$, and the variance is initialized to 10. The threshold λ as well as the parameters involved in HNC and FNC were tuned per algorithm such that the predetermined number of outliers was identified. For α-cut clustering, a grid of α values was used, and the one achieving the smallest RMSE was retained.

Table 1 shows the RMSE obtained by the algorithms for different levels of outlier contamination: 25 out of $N = 525$ points (approximately 5%), and 50 out of $N = 550$ (approx. 10%). The robust counterparts of the clustering algorithms simulated achieved a lower RMSE with an extra improvement by their weighted versions. The last two columns of Table 1 list the probabilities P_D and P_{FA} in the presence of 50 outliers. A dash indicates that the corresponding algorithm does not identify outliers inherently. In this case, W-RFKM correctly identified all the outliers introduced in the dataset.

Table 1. Performance of the clustering algorithms.

Outliers/N	25/525	50/550	50/550	
	RMSE		P_D	P_{FA}
HKM	6.4237	10.1763	–	–
FKM	0.1909	0.4283	–	–
HNC	2.2075	1.4610	0.950	0.005
FNC	0.0522	0.0607	0.980	0.002
α-cut	0.1781	0.4021	–	–
PC	0.7320	1.4322	–	–
RHKM	0.4344	0.6836	0.858	0.010
WRHKM	0.1899	0.2688	0.900	0.010
RFKM	0.0664	0.0932	0.958	0.000
WRFKM	0.0148	0.0147	1.000	0.000
RPC	0.2155	0.4720	0.980	0.002
WRPC	0.0073	0.0381	0.980	0.002

5. REFERENCES

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer, 1981.

[2] S. Dasgupta and Y. Freund, "Random projection trees for vector quantization," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3229–3242, Jul. 2009.

[3] R. N. Davé and R. Krishnapuram, "Robust clustering methods: a unified view," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 2, pp. 270–293, 1997.

[4] V. Kekatos and G. B. Giannakis, "From sparse signals to sparse residuals for robust sensing," *IEEE Trans. Signal Processing*, May 2010 (submitted).

[5] K. Lange, D. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions (with discussion)," *J. of Comp. and Graphical Stats*, vol. 9, pp. 1–59, 2000.

[6] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[7] X. Rui and D. Wunsch II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[8] P. Tseng, "Convergence of block coordinate descent method for nondifferentiable minimization," *Journal on Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[9] M. S. Yang, K. L. Wu, J. N. Hsieh, and J. Yu, "Alpha-cut implemented fuzzy clustering algorithms and switching regressions," *IEEE Trans. Syst., Man, Cybern. B*, vol. 38, pp. 588–603, 2008.

[10] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Stat. Society: Series B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.