Robust k-means: a Theoretical Revisit

Authored by:

ALEXANDROS GEORGOGIANNIS

Abstract

Over the last years, many variations of the quadratic k-means clustering procedure have been proposed, all aiming to robustify the performance of the algorithm in the presence of outliers. In general terms, two main approaches have been developed: one based on penalized regularization methods, and one based on trimming functions. In this work, we present a theoretical analysis of the robustness and consistency properties of a variant of the classical quadratic k-means algorithm, the robust k-means, which borrows ideas from outlier detection in regression. We show that two outliers in a dataset are enough to breakdown this clustering procedure. However, if we focus on ?well-structured? datasets, then robust k-means can recover the underlying cluster structure in spite of the outliers. Finally, we show that, with slight modifications, the most general non-asymptotic results for consistency of quadratic k-means remain valid for this robust variant.

1 Paper Body

Let ? : R ? R+ be a lower semi-continuous (lsc) and symmetric function with minimum value ?(0). Given a set of points X n = {x1 , . . . , xn } ? Rp , consider the generalized k-means problem (GKM) [7] n ! min Rn (c1 , . . . , ck) = min ?(——xi ? cl ——2) c1 ,...,ck 1?l?k (GKM) i=1 subject to cl ? Rp , l ? {1, . . . , k}. Our aim is to find a set of k centers {c1 , . . . , ck } that minimize the clustering risk Rn . These centers define a partition of X n into k clusters A = {A1 , . . . , Ak }, defined as " # Al = x ? X n : l = argmin1?j?k ?(——x ? cj ——2) , (1)

where ties are broken randomly. Varying ? beyond the usual quadratic function (?(t) = t2) we expect to gain some robustness against the outliers [9]. When ? is upper bounded by ?, the clusters are defined as follows. For l ? k, let " # Al = x ? X n : l = argmin1?j?k ?(——x ? cj ——2) and ?(——x ? cl ——2) ? ? , (2) and define the extra cluster

" # Ak+1 = x ? X n : min ?(——x ? cj ——2) ¿ ? . 1?j?k

(3)

This extra cluster contains points whose distance from their closest center, when measured according to ?(——x?cl ——2), is larger than ? and, as will

become clear later, it represents the set of outliers. From now on, given a set of centers {c1 , . . . , ck }, we write just A = {A1 , . . . , Ak } and implicitly mean A ? Ak+1 when ? is bounded.1 1

For a similar definition for the set of clusters induced by a bounded ? see also Section 4 in [2].

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Now, consider the following instance of (GKM), for the same set of points X n , n " # ! 1 min Rn? (c1 , . . . , ck) = min min ——xi ? cl ? oi ——22 + f? (——oi ——2) c1 ,...,ck oi 1?l?k 2 i=1 $ %& ' ?(——xi ?cl ——2)

p

subject to cl ? R , l = 1, . . . , k, oi ? Rp , i = 1, . . . , n,

(RKM)

where f? : R ? R+ is a symmetric, lsc, proper2 and bounded from below function, with minimum value f? (0), and ? a non-negative parameter. This problem is called robust k-means (RKM) and, as we show later, it takes the form of (GKM) when ? equals the Moreau envelope of f? . The problem (RKM) [5, 24] describes the following simple model: we allow each observation xi to take on an ?error? term oi and we penalize the errors, using a group penalty, in order to encourage most of the observations? errors to be equal to zero. We consider functions f? where the parameter ? ? 0 has the following effect: for ? = 0, all oi ?s may become arbitrary large (all observations are outliers), while, for ? ? ?, all oi ?s become zero (no outliers); non-trivial cases occur for intermediate values 0 ¡ ? ¡ ?. Our interest is in understanding the robustness and consistency properties of (RKM). Robustness: Although robustness is an important notion, it has not been given a standard technical definition in the literature. Here, we focus on the finite sample breakdown point [18], which counts how many outliers a dataset may contain without causing significant damage in the estimates of the centers. Such damage is reflected to an arbitrarily large magnitude of at least one center. In Section 3, we show that two outliers in a dataset are enough to breakdown some centers. On the other hand, if we restrict our focus on some ?well structured? datasets, then (RKM) has some remarkable robustness properties even if there is a considerable amount of contamination.

Consistency: Much is known about the consistency of (GKM) when the function ? is lsc and increasing [11, 15]. It turns out that this case also includes the case of (RKM) when f? is convex (see Section 3.1 for details). In Section 4, we show that the known non-asymptotic results about consistency of quadratic k-means may remain valid even when f? is non-convex.

2

Preliminaries and some technical remarks

We start our analysis with a few technical tools from variational analysis [19]. Here, we introduce the necessary notation and a lemma (the proofs are in the appendix). The Moreau envelope e?f (x) with parameter ? ¿ 0 (Definition 1.22 in [19]) of an lsc, proper, and bounded from below function f : Rp ? R and the associated (possibly multivalued) proximal map Pf? : Rp ? ? Rp are e?f (x) = minp z?R

1 1 ——x ? z——22 + f (z) and Pf? (x) = argminz?Rp ——x ? z——22 + f (z), 2? 2?

(4)

respectively. In order to simplify the notation, in the following, we fix ? to 1 and suppress the superscript. The Moreau envelope is a continuous approximation from below of f having the same set of minimizers while the proximal map gives the (possibly non-unique) minimizing arguments in (4). For (GKM), we define ? : Rp ? R as ?(x) := ?(——x——2). Accordingly, for (RKM), we define F? : Rp ? R as F? (x) := f? (——x——2). Thus, we obtain the following pairs: 1 (x ? o)2 + f? (o), Pf? (x) := argmino?R ef? (x), x ? R (5a) 2 1 eF? (x) := minp ——x ? o——22 + F? (o), PF? (x) := argmino?Rp eF? (x), x ? Rp . (5b) o?R 2 Obviously, (RKM) is equivalent to (GKM) when ?(x) = eF? (x). Every map P : R ? ? R throughout the text is assumed to be i) odd, i.e., P(?x) = ?P(x), ii) compact-valued, iii) non-decreasing, and iv) have a closed graph. We know that for any such map there exists at least one function f? such that P = Pf? (Proposition 3 in [26]).3 Finally, for our purposes (outlier detection), it is natural ef? (x) := min o?R

2 We call f proper if f (x) ¡ ? for at least one x ? Rn , and f (x) ¿ ?? for all x ? Rn ; in words, if the domain of f is a nonempty set on which f is finite (see page 5 in [19]). 3 Accordingly, for a general function ? : R ? [0, ?) to be a Moreau envelope, i.e., ?(?) = ef? (?) as defined in (5a) for some function f? , we require that ?(?) ? 12 — ? —2 is a concave function (Proposition 1 in [26]).

2

to require that v) P is a shrinkage rule, i.e., P(x) ? x, ?x ? 0. The following corollary is quite straightforward and useful in the sequel. Corollary 1. Using the notation in definitions (5a) and (5b), we have x PF? (x) = Pf (——x——2) and eF? (x) = ef? (——x——2). (6) ——x——2 ? Passing from a model of minimization in terms of a single problem, like (GKM), to a model in which a problem is expressed in a particular parametric form, like (RKM) with the Moreau envelope, the description of optimality conditions is opened to the incorporation of the multivalued map PF? . The next lemma describes the necessary conditions for a center cl to be (local) optimal for (RKM). Since we deal with the general case, well known results, such as smoothness of the Moreau envelope or convexity of its subgradients, can no longer be taken for granted. ? Remark 1. Let ?(?) = eF? (?). The usual subgradient, denoted as ??(x), is not sufficient to characterize the differentiability properties of Rn? in (RKM). Instead, we use the (generalized) subd? ifferential ??(x) (Definition 8.3 in [19]). For all x, we have ??(x) ? ??(x). Usually, the previous two sets coincide at a point x. In this case, ? is called regular at x. However, it is common in ? practice that the sets ??(x) and ??(x) are different (for a detailed exposition on subgradients see Chapter 8 in [19]; see also Example 1 in Appendix A.9). Lemma 1. Let PF? : Rp ? ? Rp be a proximal map and set ?(?) = eF? (?). The necessary (generalized) first order conditions for the centers {c1 , . . . , ck } ? Rp to be optimal for (RKM) are "! # ! ! 0?? ?(xi ? cl) ? ??(xi ? cl) ? (cl ? xi + PF? (xi ? cl)) , l ? {1, . . . , k}. i?Al

i?Al

i?Al

(7)

The interpretation of the set inclusion above is the following: for any center cl ? Rp , every subgradient vector in ??(xi ? cl) must be a vector associated with a vector in PF? (xi ? cl) (Theorem 10.13 in [19]). However, in general, the converse does not hold true. We note that when the proximal map is single-valued and continuous, which happens for example not only when f? is convex, but also for many popular non-convex penalties, both set inclusions become equalities and the converse holds, i.e., every vector in PF? (xi ? cl) is a vector associated with a subgradient in ??(xi ? cl) (Theorem 10.13 in [19] and Proposition 7 in [26]). We close this section with some useful remarks on the properties of the Moreau envelope as a map between two spaces of functions. There exist cases where two different functions, f? ?= f?? , have equal Moreau envelopes, ef? = ef?? (Proposition 1 in [26]), implying that two different forms of (RKM) correspond to the same ? in (GKM). For example, the proximal hull of f? , defined as h?f? (x) := ?e?(?e?) (x), is a function different from f? but has the same Moreau envelope as f? f?

(see also Example 1.44 in [19], Proposition 2 and Example 3 in [26]). This is the main reason we preferred the proximal map instead of the penalty function point of view for the analysis of (RKM).

3

On the breakdown properties of robust k-means

In this section, we study the finite sample breakdown point of (RKM) and, more specifically, its universal breakdown point. Loosely speaking, the breakdown point measures the minimum fraction of outliers that can cause excessive damage in the estimates of the centers. Here, it will become clear how the interplay between the two forms, (GKM) and (RKM), helps the analysis. Given a dataset n X n = {x1 , . . . , xn } and a nonnegative integer m ? n, we say that Xm is an m-modification if it arises from X n after replacing m of its elements by arbitrary elements x?i ? Rp [6]. Denote as r(?) the non-outlier samples, as counted after solving (RKM), for a dataset X n and some ? ? 0, i.e., 4 ((((r(?) := ({xi ? X n : ——oi ——2 = 0, i = 1, . . . , n}(. (8)

Then, the number of estimated outliers is q(?) = n ? r(?). In order to simplify notation, we drop the dependence of r and q on ?. With this notation, we proceed to the following definition. 4

More than one ? can yield the same r, but this does not affect our analysis.

3

Definition 1 (universal breakdown point for the centers [6]). Let n, r, k be such that n ? r ? k + 1. n Given a dataset Xm in Rp , let {c1 , . . . , ck } denote the (global) optimal set of centers for (RKM). The universal breakdown value of (RKM) is "m # ?(n, r, k) := min min : sup max ——c —— = ? . (9) l 2 n 1?l?k X n 1?m?n n Xm n Here, X n = {x1 , . . . , xn } ? Rp while Xm ? Rp runs over all m-modifications of X n .

According to the concept of universal breakdown point, (RKM) breaks down at the first integer m for which there exists a set X n such that the estimates of the cluster centers become arbitrarily bad n for a suitable modification Xm . Our

analysis is based on Pf? and considers two cases: those of biased and unbiased proximal maps. The former corresponds to the class of convex functions f? , while the latter corresponds to a class of non-convex f? . 3.1

Biased proximal maps: the case of convex f?

If f? is convex, then ? = eF? is also convex while PF? is continuous, single-valued, and satisfies [19] ——x ? PF? (x)——2 ? ? as ——x——2 ? ?. (10) Proximal maps with this property are called biased since, as the l2 -norm of x increases, so does the norm of the difference in (10). In this case, for each xi ? Al , from Lemma 1 and expression (10), we have ——??(xi ?cl)——2 = ——?eF? (xi ?cl)——2 = ——cl ?xi +PF? (xi ?cl)——2 ? ? as ——xi ?cl ——2 ? ?. (11) The supremum value of ——??(x ? cl)——2 is closely related to the gross error sensitivity of an estimator [9]. It is interpreted as the worst possible influence which a sample x can have on cl [7]. In view of (11) and the definition of the clusters in (1), (RKM) is extremely sensitive. Although it can detect an outlier, i.e., a sample xi with a nonzero estimate for ——oi ——2 , it does not reject it since the influence of xi on its closest center never vanishes.5 The l1 -norm, f? (x) = ?—x—, which has Moreau envelope equal to the Huber loss-function [24], is the limiting case for the class of convex penalty functions that, although it keeps the difference ——x ? PF? (x)——2 in (10) constant and equal to ?, introduces a bias term proportional to ? in the estimate cl . The following proposition shows that (RKM) with a biased PF? has breakdown point equal to n1 , i.e., one outlier suffices to breakdown a center. Proposition 1. Assume k ? 2, k + 1 ¡ r ? n. Given a biased proximal map, there exist a dataset X n and a modification X1n such that (RKM) breaks down. 3.2

Unbiased proximal maps: the case of non-convex f? 2

Consider now the l0 -(pseudo)norm on R, f? (z) := ?—z—0 = ?2 {z?=0} , thresholding proximal operator P?—?—0 : R ? ? R, ? ?0, P?—?—0 (x) = arg minz?R 12 (x ? z)2 + f? (z) = {0, x}, ? x,

According to Lemma 1, for p = 1 (scalar case), we have

and the associated hard—x— ¡ ?, —x— = ?, —x— ¿ ?.

(12)

??(xi ? cl) ? cl ? xi + P?—?—0 (xi ? cl) = {0} for —xi ? cl — ¿ ?, xi ? Al ,

(12)

(13)

implying that ?(xi ? cl), as a function of, cl , remains constant for —xi ? cl — ¿ ?. As a consequence of (13), if cl is local optimal, then 0 ? ?{ i?Al ?(xi ? cl)} and ! ! . 0? (cl ? xi) + cl ? xi + P?—?— (xi ? cl) . (14) i?Al , i?Al , —xi ?cl —¡?

—xi ?cl —=?

Depending on the value of ?, (RKM) with the l0 -norm is able to ignore samples with distance from their closest center larger than ?. This is done since P?—?—0 (xi ? cl) = xi ? cl whenever —xi ? cl — ¿ ? 5

See the analysis in [7] about the influence function of (GKM) when ? is convex.

4

and the influence of xi vanishes. In fact, there is a whole family of non-convex f? ?s whose proximal map Pf? satisfies Pf? (x) = x, for all —x— ¿ ?,

(15)

for some ? ¿ 0. These are called unbiased proximal maps [13, 20] and have the useful property that, as one observation is arbitrarily modified, all estimated cluster centers remain bounded by a constant that depends only on the remaining unmodified samples. Under certain circumstances, the proof of the following proposition reveals that, if there exists one outlier in the dataset, then robust k-means will reject it. Proposition 2. Assume k ? 2, k + 1 ¡ r ? n, and consider the dataset X n = {x1 , . . . , xn } along with its modification by one replacement y, X1n = {x1 , . . . , xn?1 , y}. If we solve (RKM) with X1n and an unbiased proximal map satisfying (15), then all estimates for the cluster centers remain bounded by a constant that depends only on the unmodified samples of X n . Next, we show that, even for this class of maps, there always exists a dataset that causes one of the estimated centers to breakdown as two particular observations are suitably replaced. Theorem 1 (Universal breakdown point for (RKM)). Assume k ? 2 and n ? r ? k + 2. Given an unbiased proximal map satisfying (15), there exist a dataset X n and a modification X2n , such that (RKM) breaks down. Hence, the universal breakdown point of (RKM) with an unbiased proximal map is n2 . In Figure 1, we give a visual interpretation of Theorem 1. The top subfigure depicts the unmodified initial dataset X 9 = {x1 , . . . , x9 } (black circles) with a clear two-cluster structure; the bottom subfigure shows the modification X29 (dashed line arrows). Theorem 1 states that (RKM) on X29 fails to be robust since, every subset of X29 with r = 8 points has a cluster containing an outlier. 3.3

Figure 1: The top subfigure is the unmodified dataset X 9 . Theorem 1 states that every subset of the modification X29 (bottom subfigure) with size 8 contains an outlier.

Restricted robustness of robust k-means for well-clustered data

The result of Theorem 1 is disappointing but it is not (RKM) to be blamed for the poor performance but the tight notion of the definition about the breakdown point [6, 7]; allowing any kind of contamination in a dataset is a very general assumption. In this section, we place two restrictions: i) we consider datasets where inlier samples can be covered by unions of balls with centers that are ?far apart? each other, and ii) we ask a question different from the finite sample breakdown point. We want to exploit as much as possible the results of [2] concerning a new quantitative measure of noise robustness which compares the output of (RKM) on a contaminated dataset to its output on the uncontaminated version of the dataset. Our aim is to show that (RKM), with a certain class of proximal maps and datasets that are well-structured ignores the influence of outliers when grouping the inliers. First, we introduce Corollary 2 which states the form that Pf? should have in order the results of [2] to apply to (RKM) and, second, we give details about the datasets which we consider as wellstructured. Using this corollary we are able to design proximal maps for which Theorems 3, 4, and 5 in [2] apply; otherwise, it is not clear how the

analysis of [2] is valid for (RKM). Let h : R ? R be a continuous function with the following properties: 1. h is odd and non-decreasing (h+ (?) is used to denote its restriction on [0, ?)); 2. h is a shrinkage rule: 0 ? h+ (x) ? x, ?x ? [0, ?); 3. the difference x ? h+ (x) is non-decreasing, i.e., for 0 ? x1 ? x2 we have x1 ? h+ (x1) ? x2 ? h+ (x2). 5

Define the map

? ?h(x), Pf? (x) := {h(x), x}, ? x,

—x— ¡ ?, —x— = ?, —x— ¿ ?.

(16)

Multivaluedness of Pf? at —x— = ? signals that ef? is non-smooth at these points. An immediate consequence for the Moreau envelope associated with the previous map is the following. Corollary 2. Let the function g : [0, ?) ? [0, ?) be defined as / x g(x) := (u ? h(u))du, x ? [0, ?). (17) 0

Then, the Moreau envelope associated with Pf? in (16) is

ef? (x) = min{g(—x—), g(?)} = g(min{—x—, ?}).

(18)

Next, we define what it means for a dataset to be (?1 , ?2)-balanced; this is the class of datasets which we consider to be well-structured. Definition 2 ((?1 , ?2) balanced dataset [2]). Assume that a set X n ? Rp has a subset I (inliers), with at least n2 samples, and the following properties: 1. I =

0k

l=1

Bl , where Bl = B(bl , r) is a ball in Rp with bounded radius r and center bl ;

2. ?1 —I— ? —Bl — ? ?2 —I— for every l, where —Bl — is the number of samples in Bl and ?1 ,?2 ¿ 0; 3. ——bl ? bl? ——2 ¿ v for every l ?= l? , i.e., the centers of the balls are at least v ¿ 0 apart. Then, X n is a (?1 , ?2)-balanced dataset. We now state the form that Theorem 3 in [2] takes for (RKM). Theorem 2 (Restricted robustness of (RKM)). If i) ef? is as in Corollary 2, i.e., ef? (——x——2) = g(min{——x——2 , ?}), ii) X n has a (?1 , ?2)-balanced subset of samples I with k balls, and 2 iii) the centers of the balls are at least v ¿ 4r + 2g ?1 (?1?+? g(r)) apart, then for ? ? 1 1 2 33 —I— v v ?1 the set of outliers X n I has no effect on the 2, g —X n I— (?1 g(2 ? 2r) ? (?1 + ?2)g(r))

grouping of inliers I. In other words, if {x, y} ? Bl and {c1 , . . . , ck } are the optimal centers when solving (RKM) for a ? as described before, then l = argmin1?j?k ef? (——x ? cj ——2) = argmin1?j?k ef? (——y ? cj ——2). For the sake of completeness, we give a proof of this theorem in the appendix. In a similar way, we can recast the results of Theorems 4 and 5 in [2] to be valid for (RKM).

4

On the consistency of robust k-means

Let X n be a set with n independent and identically distributed random samples xi from a fixed but unknown probability distribution ?. Let C? be the empirical optimal set of centers, i.e., C? := argminc1 ...,ck ?Rp Rn? (c1 , . . . , ck).

(19)

C ? := argminc1 ...,ck ?Rp R? (c1 , . . . , ck),

(20)

The population optimal set of centers is the set

where R? is the population clustering risk, defined as / " # 1 R? (c1 , . . . , ck) := min minp ——x ? cl ? o——22 + f? (——o——2) ?(dx). 1?l?k o?R 2 $ %& '

(21)

?(——x?cl ——2)=ef? (——x?cl ——2)

Loss consistency and (simply) consistency for (RKM) require, respectively, that n?? ? n?? Rn? (C) ?? R? (C ?) and C? ?? C ? .

6

(22)

? converges In words, as the size n of the dataset X n increases, the empirical clustering risk Rn? (C) ? ? almost surely to the minimum population risk R (C) and (for n large enough) C? can effectively replace the optimal set C ? in quantizing the unknown probability measure ?. For the case of convex f? , non-asymptotic results describing the rate of convergence of Rn? to R in (22) are already known ([11], Theorem 3). Noting that the Moreau envelope of a non-convex f? belongs to a class of functions with polynomial discrimination [16] (the shatter coefficient of this class is bounded by a polynomial) we give a sketch proof of the following result. Theorem 3 (Consistency of (RKM)). Let the samples xi ? X n , i ? {1, . . . , n}, come from a fixed but unknown probability measure ?. For any k ? 1 and any unbiased proximal map, we have ? ? R? (C ?) and lim ER? (C)

n??

lim C? ? C ? (convergence in probability).

n??

(23)

Theorem 3 reads like an asymptotic convergence result. However, its proof (given in the appendix) uses combinatorial tools from Vapnik-Chervonenkis 4theory, revealing that the non-asymptotic rate ? to R? (C ?) is of order O(log n/n) (see Corollary 12.1 in [4]). of convergence of ER? (C)

5

Relating (RKM) to trimmed k-means

As the effectiveness of robust k-means on real world and synthetic data has already been evaluated [5, 24], the purpose of this section is to relate (RKM) to trimmed k-means (TKM) [7]. Trimmed kmeans is based on the methodology of ?impartial trimming?, which is a combinatorial problem fundamentally different from (RKM). Despite their differences, the experiments show that, both (RKM) and (TKM) perform remarkably similar in practice. The solution of (TKM) (which is also a set of k centers) is the solution of quadratic k-means on the subsample containing ?n(1 ? ?)? points with the smallest mean deviation (0 ¡ ? ¡ 1). The only common characteristic of (RKM) and (TKM) is that they both have the same universal breakdown point, i.e., n2 , for arbitrary datasets. Trimmed k-means takes as input a dataset X n , the number of clusters k, and

a proportion of outliers a ? (0, 1) to remove.6 A popular heuristic algorithm for (TKM) is the following. After the initialization, each iteration of (TKM) consists of the following steps: i) the distance of each observation from its closest center is computed, ii) the top ?an? observations with larger distance from its closest center are removed, iii) the remaining points are used to update the centers. The previous three steps are repeated untill the centers converge.7 As for robust k-means, we solve the (RKM) problem with a coordinate optimization procedure (see Appendix A.9 for details). The synthetic data for the experiments come from a mixture of Gaussians with 10 components and without any overlap between them.8 The number of inlier samples is 500 and each inlier xi ? [?1, 1]10 for i ? {1, . . . , 500}. On top of the inliers lie 150 outliers in R10 distributed uniformly in general positions over the entire space. We consider two scenarios: in the first, the outliers lie in [?3, 3]10 (call it mild-contamination), while, in the second, the outliers lie in [?6, 6]10 (call it heavy-contamination). The parameter a in trimmed k-means (the percentage of outliers) is set to a = 0.3, while the value of the parameter ? for which (RKM) yields 150 outliers is found through a search over a grid on the set ? ? (0, ?max) (we set ?max as the maximum distance between two points in a dataset). Both algorithms, as they are designed, require as input an initial set of k points; these points form the initial set of centers. In all experiments, both (RKM) and (TKM) take the same k vectors as initial centers, i.e., k points sampled randomly from the dataset. The statistics we use for the comparison are: i) the rand-index for clustering accuracy [17] ii) the cluster estimation error, i.e., the root mean square error between the estimated cluster centers and the sample mean of each cluster, iii) the true positive outlier detection rate, and finally, iv) the false positive outlier detection rate. In Figures 2-3, we plot the results for a proximal map Pf like the one in (16) with h(x) = ?x and ? = 0.005; with this choice for h, we mimic the hard-thresholding operator. The results for each scenario (accuracy, cluster estimation error, etc) are averages over 150 runs of the experiment. As seen, both algorithms share almost the same statistics in all cases. 6

We use the implementation of trimmed k-means in the R package trimcluster [10]. The previous three steps are performed also by another robust variant of k-means, the k-means? (see [3]). 8 We use the R toolbox MixSim [14] that guarantees no overlap among the 10 mixtures. 7

7
0.5
? ? ? ? ? ?
?
7.5
5.0
?
?
0.950
0.925 ?
?
robust k?means trimmed k?means

robust k?means trimmed k?means
0.015
0.010
?
0.005
?
?
0.000
robust k?meanstrimmed k?means
Cluster Radius Estimation Error
? ? ? ? ? ?
10.0
0.975
False Positive Error Rate
? ?
?
True Positive Error Rate
Accuracy
0.7
12.5
?
Center Estimation Error
? ?
0.9
? ? ?
9
? ? ?
6
3
? ? ?
?
0
robust k?meanstrimmed k?means
robust k?means trimmed k?means
0.4
? ? ?
?
? ? ? ? ? ?
0.2
? ?
15.0
12.5
10.0
?
robust k?means trimmed k?means
?

0.75 ? ?
0.50 ? ?
0.25 ? ? ? ?
0.00
?
0.3
? ?
? ?
False Positive Error Rate
0.6
1.00
?
True Positive Error Rate
Accuracy
0.8
Center Estimation Error
17.5
robust k?means trimmed k?means
? ?
0.2
Cluster Radius Estimation Error

Figure 2: Performance of robust and trimmed k-means on a mixture of 10 Gaussians without overlap. On top of the 500 samples from the mixture there are 150 outliers uniformly distributed in [?1, 1]10 . ? ? ? ? ? ?

? ?
? ?
0.1 ?
0.0
robust k?means trimmed k?means
? ? ?
4
3
2
1
robust k?means trimmed k?means
robust k?means trimmed k?means
? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ?
? ?
20
1.000
?
? ?
10
?
robust k?means trimmed k?means

?
?
? ?
?
0.975
?
?
0.950
0.925
0.900 robust k?means trimmed k?means
?
? ?
?
?
?
? ?
?
?
robust k?meanstrimmed k?means
False Positive Error Rate
0.6
Center Estimation Error
Accuracy
0.8
?
True Positive Error Rate
30
1.0
0.04 ? ?
?
0.03 ? ?
0.02 0.01
? ?
?
?
?
?
? ? ?
0.00 robust k?means trimmed k?means

Figure 4: Results on two spherical clusters with equal radius r, each one with 150 samples, and centers are at least 4r apart. On top of the samples lie 150 outliers uniformly distributed in [?6, 6]10 . In Figure 4, we plot the results for the case of two spherical clusters in R10 with equal radius r, each one with 150 samples, and centers that are at least 4r apart from each other. The inlier samples are in [?3, 3]10 . The outliers are 150 (half of the dataset is contaminated) and are uniformly distributed in [?6, 6]10 . The results (accuracy,

cluster estimation error, etc) are averages over 150 runs of the experiment. This configuration is a heavy contamination scenario but, due to the structure of the dataset, as expected from Theorem 2, (RKM) performs remarkably well; the same holds for (TKM).

6

Conclusions

We provided a theoretical analysis for the robustness and consistency properties of a variation of the classical quadratic k-means called robust k-means (RKM). As a by-product of the analysis, we derived a detailed description of the optimality conditions for the associated minimization problem. In most cases, (RKM) shares the computational simplicity of quadratic k-means, making it a ?computationally cheap? candidate for robust nearest neighbor clustering. We show that (RKM) cannot be robust against any type of contamination and any type of datasets, no matter the form of the proximal map we use. If we restrict our attention to ?well-structured? datasets, then the algorithm exhibits some desirable noise robustness. As for the consistency properties, we showed that most general results for consistency of quadratic k-means still remain valid for this robust variant. Acknowledgments The author would like to thank Athanasios P. Liavas for useful comments and suggestions that improved the quality of the article.

8

Cluster Radius Estimation Error

Figure 3: The same setup as in Figure 2 except that the coordinates of each outlier lie in [?3, 3]10 . 7.5

5.0

2.5

0.0 robust k?means trimmed k?means

2 References

[1] Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. Journal of the American Statistical Association, 2011. [2] Shai Ben-David and Nika Haghtalab. Clustering in the presence of background noise. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 280?288, 2014. [3] Sanjay Chawla and Aristides Gionis. k-means-: A unified approach to clustering and outlier detection. SIAM. [4] L. Devroye, L. Gy?orfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Stochastic Modelling and Applied Probability. Springer New York, 1997. [5] Pedro A Forero, Vassilis Kekatos, and Georgios B Giannakis. Robust clustering using outlier-sparsity regularization. Signal Processing, IEEE Transactions on, 60(8):4163?4177, 2012. [6] Mar??a Teresa Gallegos and Gunter Ritter. A robust method for cluster analysis. Annals of Statistics, pages 347?380, 2005. ? [7] Luis Angel Garc??a-Escudero and Alfonso Gordaliza. Robustness properties of k-means and trimmed k-means. Journal of the American Statistical Association, 94(447):956?969, 1999. [8] Michael R. Garey and David S. Johnson.

Computers and Intractability: A Guide to the Theory of NPCompleteness. W. H. Freeman & Co., New York, NY, USA, 1979. [9] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. Robust statistics: the approach based on influence functions, volume 114. John Wiley & Sons, 2011. [10] Christian Hennig. trimcluster: Cluster analysis with trimming, 2012. R package version 0.1-2. [11] Tam?as Linder. Learning-theoretic methods in vector quantization. In Principles of nonparametric learning, pages 163?210. Springer, 2002. [12] Stuart P Lloyd. Least squares quantization in pcm. Information Theory, IEEE Transactions on, 28(2):129? 137, 1982. [13] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association, 2012. [14] Volodymyr Melnykov, Wei-Chen Chen, and Ranjan Maitra. MixSim: An R package for simulating data to study performance of clustering algorithms. Journal of Statistical Software, 51(12):1?25, 2012. [15] David Pollard. Strong consistency of k-means clustering. The Annals of Statistics, 9(1):135?140, 1981. [16] David Pollard. Convergence of stochastic processes. Springer Science & Business Media, 1984. [17] William M Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846?850, 1971. [18] G. Ritter. Robust Cluster Analysis and Variable Selection. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2014. [19] R Tyrrell Rockafellar and Roger J-B Wets. Variational analysis, volume 317. Springer Science & Business Media, 2009. [20] Yiyuan She et al. Thresholding-based iterative selection procedures for model selection and shrinkage. Electronic Journal of statistics, 3:384?415, 2009. [21] Marc Teboulle. A unified continuous optimization framework for center-based clustering methods. The Journal of Machine Learning Research, 8:65?102, 2007. [22] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications, 109(3):475?494, 2001. [23] Sara Van De Geer. Empirical processes in m-estimation. June 13, 2003. Handout at New Directions in General Equilibrium Analysis (Cowles Workshop, Yale University). [24] Daniela M Witten. Penalized unsupervised learning with outliers. Statistics and its Interface, 6(2):211, 2013. [25] Stephen J Wright. Coordinate descent algorithms. Mathematical Programming, 151(1):3?34, 2015. [26] Yaoliang Yu, Xun Zheng, Micol Marchetti-Bowick, and Eric P Xing. Minimizing nonconvex nonseparable functions. In AISTATS, 2015.

9