Centroidal Clustering of Noisy Observations by Using rth Power Distortion Measures

Erdem Koyuncu, *Member, IEEE*

***Abstract*— We consider the problem of clustering a dataset through multiple noisy observations of its members. The goal is to obtain a clustering that is as faithful to the clustering of the original dataset as possible. We propose a centroidal approach whose distortion measure is the sum of rth powers of the distances between the cluster center and the noisy observations. For $r = 2$, our scheme boils down to the well-known approach of clustering the average of noisy samples. First, we provide a mathematical analysis of our clustering scheme. In particular, we find formulas for the average distortion and the spatial distribution of the cluster centers in the asymptotic regime where the number of centers is large. We then provide an algorithm to numerically optimize the cluster centers in the finite regime. We extend our method to automatically assign weights to noisy observations. Finally, we show that for various practical noise models, with a suitable choice of r, our algorithms can outperform several other existing techniques over various datasets.**

***Index Terms*— Centroidal clustering, high-resolution theory, noisy clustering, quantization.**

I. Introduction

A. Ordinary Centroidal Clustering

The goal of clustering is to partition an unlabeled dataset into disjoint sets or clusters such that each cluster only consists of similar members of the dataset [1]–[3]. Of particular interest to this work are center-based or centroidal clustering methods, as described in the following. Let $D = \{y_{1,i}\}_{i=1}^m \subset \mathbb{R}^d$ be a dataset of d-dimensional vectors, whose elements are drawn according to a random vector X_1. In classical k-means clustering [4]–[6], one is interested in finding the optimal cluster centers $u_1, \ldots, u_n$ that minimize the average distortion

$$(u_1, \ldots, u_n) \mapsto \frac{1}{m} \sum_{i=1}^{m} \min_k \ \|u_k - y_{1,i}\|^2. \tag{1}$$

Distinct clusters can then be identified via the Voronoi cells $\{y \in D : \|y - u_i\| \leq \|y - u_j\|, \forall j\}$, $i = 1, \ldots, n$ (ties are broken arbitrarily). Several variations to the basic formulation in (1) have been studied. For example, the squared-error distortion measure $(u, y) \mapsto \|u - y\|^2$ between the cluster center u_k and the dataset sample $y_{1,i}$ in (1) can be replaced with the rth power distortion measure $(u, y) \mapsto \|u - y\|^r$ [7], a quadratic distortion $(u, y) \mapsto (u - y)^T A(u - y)$, where A is a positive semidefinite matrix [8], [9] or Bregman divergence [10], [11].

Finding a (globally) optimal solution to (1) is known to be an NP-hard problem [12]. Nevertheless, locally optimal solutions can be found using the k-means algorithm or its extensions such as the generalized Lloyd algorithm [6], [13]. Moreover, vector quantization theory [14], [15] provides a precise description of the structure of optimal solutions and the corresponding minimum average distortions in the asymptotic regimes $n, m \to \infty$ [16], [17].

B. Clustering Noisy Observations and Related Work

In this work, we will study the following practical variant of the clustering problem. Consider a physical process that generates a (noiseless) dataset $D' = \{y'_i\}_{i=1}^m$, and suppose that our goal is cluster D'. In practice, we may only access a noisy version of D' through, for example, sensor measurements. For any given sample index $i \in \{1, \ldots, m\}$, given that there are L sensors in total, sensor ℓ can provide a noisy version $y_{\ell,i}$ of the true data sample y'_i. In such a scenario, one only has the noisy dataset $D = \{[y_{1,i}, \ldots, y_{L,i}]\}_{i=1}^m \subset \mathbb{R}^{d \times L}$ available and cannot access D'. We wish to find a clustering of D that is as faithful to the clustering of D' as possible.

The noisy clustering formulation above is very well studied, at least for the case of a single observation $L = 1$ under many different formulations [18]–[20]. The multiple observations case $L > 1$ appears prominently in the area of bioinformatics [21]–[23]. In fact, many types of biological data such as gene expressions are prone to random measurement errors during the acquisition or measurement phase. A commonly utilized technique is thus to measure the same biological sample multiple times, thus resulting in multiple noisy versions of the actual data. A notable feature of these measurements is that the corresponding measurement noises are often observed to be heavy-tailed, following, e.g., a (Student's) t-distribution with a low degree of freedom, rather than a Gaussian distribution [24], [25]. In fact, heavy-tailed noise appears in a variety of practical phenomena and thus has been the subject of many recent publications in the context of machine learning [26], [27]. Another primary application area for the case $L > 1$ is sensor networks, where each sensor observes a different noisy version of the underlying process [28].

There have been many different proposed techniques to effectively utilize the multiple observations for clustering purposes in the aforementioned studies. A basic method is averaging, where one simply clusters the dataset of averages $\{(1/L) \sum_{\ell=1}^L y_{\ell,i}\}_{i=1}^m$ of noisy observations [22] using the k-means algorithm. In the co-clustering approach, one instead clusters all mL noisy observations $\bigcup_{\ell=1}^L \{y_{\ell,i}\}_{i=1}^m$ together [23]. For every sample index i, a majority vote is then cast among the clusters of $\{y_{\ell,i}\}_{\ell=1}^L$ to determine the cluster of sample index i. Another approach is to merely concatenate the L noisy observation vectors $\{y_{\ell,i}\}_{\ell=1}^L$ into one dL-dimensional vector and cluster the resulting dataset $\{[y_{1,i}^T, \ldots, y_{L,i}^T]^T\}_{\ell=1}^L$ of dL-dimensional vectors. This becomes a special case of the multimodal or multiview clustering [29], [30], where each noisy measurement corresponds to one individual view. Ensemble methods can also be utilized. In [23], each noisy dataset $\{y_{\ell,i}\}_{i=1}^m$ is first clustered individually, resulting in a co-occurrence matrix M_ℓ. The (i, j)th entry of the sum $\sum_{\ell=1}^L M_\ell$ then indicates for how many observations the sample indices i and j belong to the same cluster. Agglomerative hierarchical clustering is then applied to $\sum_{\ell=1}^L M_\ell$ to estimate the

Manuscript received 19 August 2021; revised 24 March 2022; accepted 11 June 2022. Date of publication 22 June 2022; date of current version 5 January 2024. This work was supported in part by the Army Research Lab (ARL) under Grant W911NF-21-2-0272, in part by NSF under Award CCF–1814717 and Award CNS–2148182, and in part by the University of Illinois at the Chicago Discovery Partners Institute Seed Funding Program.

The author is with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: ekoyuncu@uic.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3183294.

Digital Object Identifier 10.1109/TNNLS.2022.3183294

clustering of the noiseless dataset. In [28], a Kalman-type filter is proposed for a scenario where each observation is corrupted by an additive Gaussian noise whose variance is time-varying but known.

C. Main Contributions and Brief Organization

Despite various practical applications, some of which have been mentioned above, the optimal method to combine multiple noisy observations for clustering purposes remains a wide open problem. Previous work has made significant progress by identifying various combining methods and evaluating the performance of these methods over various datasets. However, a mathematical analysis of the available methods is generally not available. Another drawback of the existing methods is that they may not be necessarily tailored for practically relevant cases of noise, which may be heavy-tailed or stem from outliers, as discussed in Section I-B. In this work, we introduce a new centroid-based method for clustering noisy observations. Through various numerical simulations over different datasets, we show that our method outperforms various existing methods. We also provide a mathematical analysis of our clustering method by finding the corresponding optimal centroid distribution and the asymptotic distortion. Beyond clustering, our analytical results generalize some of the fundamental classical results of quantization theory and also find applications in the area of facility location optimization.

The rest of this brief is organized as follows. In Section II, we introduce our new clustering method. In Section III, we introduce some well-known results from quantization theory to aid in analyzing the method. In Section IV, we provide a theoretical analysis of our clustering scheme. In Section VI, we provide numerical results over different datasets. In Section V, we describe an algorithm to optimize observation weights. Finally, in Section VII, we draw our main conclusions.

II. Description of the Method

In this section, we describe our new technique to cluster noisy observations and discuss its motivations. A detailed analysis of the scheme and numerical results will be provided in Section IV.

A. Formulation of the Distortion Measure

Our method is inspired by and generalizes the basic method of averaging, which was described in Section I-B. In order to present our method in full generality, we consider the generalized averaging scheme [24], [31], which relies on clustering the dataset of a weighted sum of observations $\{(1/c_1)(\lambda_1 y_{1,i} + \cdots + \lambda_L \lambda_{L,i})\}_{i=1}^m$. Here, $\{\lambda_\ell\}_{\ell=1}^L$ are the weights that govern the reliability of each observation set, and $c_1 = \lambda_1 + \cdots + \lambda_L$. Equivalently, one designs the clustering so as to minimize the average distortion [c.f. (1)]

$$(u_1, \ldots, u_n) \mapsto \frac{1}{m} \sum_{i=1}^{m} \min_k \left\| u_k - \frac{1}{c_1} \sum_{\ell=1}^{L} \lambda_\ell y_{\ell,i} \right\|^2. \tag{2}$$

A first step is to rewrite the mean-squared term as a sum of squares through the equality

$$\sum_{\ell=1}^{L} \lambda_\ell \|u_k - y_{\ell,i}\|^2 = c_1 \left\| u_k - \frac{1}{c_1} \sum_{\ell=1}^{L} \lambda_\ell y_{\ell,i} \right\|^2 + \sum_{\ell=1}^{L} \lambda_\ell \|y_{\ell,i}\|^2 - \frac{1}{c_1} \left\| \sum_{\ell=1}^{L} \lambda_\ell y_{\ell,i} \right\|^2. \tag{3}$$

Identity (3) can easily be verified by expanding the squared Euclidean norms on both sides via the formula $\|\alpha+\beta\|^2 = \|\alpha\|^2 + \|\beta\|^2 + 2\alpha^T\beta$, where α and β are arbitrary vectors. Noting that the last two terms in (3) are independent of k, the problem (2) is equivalent to finding a clustering that minimizes the average distortion

$$(u_1, \ldots, u_n) \mapsto \frac{1}{m} \sum_{i=1}^{m} \min_k \sum_{\ell=1}^{L} \lambda_\ell \|u_k - y_{\ell,i}\|^2. \tag{4}$$

Our proposed method is to focus on a general power $r \geq 1$ of the Euclidean norms $\|u_k - y_{\ell,i}\|$ that appear in (4). In other words, we propose to find a clustering that minimizes

$$(u_1, \ldots, u_n) \mapsto \frac{1}{m} \sum_{i=1}^{m} \min_k \sum_{\ell=1}^{L} \lambda_\ell \|u_k - y_{\ell,i}\|^r \tag{5}$$

where $r \geq 1$ is some real number. For the special case $r = 2$, our formulation boils down to the averaging scheme. We shall shortly discuss the motivation behind considering a general power $r \neq 2$ and why such a generalization should help. Let us first describe the solution to the minimization of (5) and the resulting overall clustering algorithm.

Algorithm 1 Algorithm for Clustering Noisy Observations

1: Initialize the cluster centers $U = (u_1, \ldots, u_n)$ arbitrarily.
2: Iterate (6) and (7) until convergence of the cost (5).
3: For $i = 1, \ldots, m$, set the final clustering $\zeta_i = j$, where $i \in V_j$.

B. Generalized Lloyd Algorithm

Since (5) generalizes the classical k-means problem (1), which itself is NP-hard, finding the globally optimal solution to (5) is also a hopeless problem in general. However, a locally optimal solution can be found via the following variant of the generalized Lloyd algorithm. First, one initializes some arbitrary U and then iterates between the two steps of calculating the generalized Voronoi cells

$$V_j \leftarrow \left\{ i : \sum_{\ell=1}^{L} \lambda_\ell \|u_j - y_{\ell,i}\|^r \leq \sum_{\ell=1}^{L} \lambda_\ell \|u_k - y_{\ell,i}\|^r \ \forall k \right\} \tag{6}$$

for $j = 1, \ldots, n$, and the generalized centroids

$$u_j \leftarrow \arg\min_u \sum_{i \in V_j} \sum_{\ell=1}^{L} \lambda_\ell \|u - y_{\ell,i}\|^r. \tag{7}$$

It is easily seen that the resulting algorithm results in a nonincreasing average distortion at every iteration and thus converges in a cost-function sense. Moreover, the calculation of (7) can be accomplished in a computationally efficient manner as it is convex for any $r \geq 1$ (also see [32] for fast gradient-based algorithms). In fact, for $r = 1$, (7) reduces to the calculation of the so-called geometric median, and Weiszfeld's algorithm provides an efficient solution.

The overall algorithm for clustering a dataset of noisy observations is provided in Algorithm 1. Note that we have allowed a different weighting of observations to present the algorithm in full generality. The weights $\lambda_1, \ldots, \lambda_L$ can all be set to the same value (e.g., unity) when the observation noise statistics are known to be independent and identically distributed. Another possibility is to optimize the weights with respect to the statistics of observations; an automatic weighting algorithm will be presented later in Section V. The power r should typically be set depending on noise statistics; some guidelines are provided in the following.

C. Motivations for the Method

We conclude the description of our clustering scheme by providing the main motivations behind its formulation. At first sight, replacing the second power $\|\cdot\|^2$ in (4) with the rth power $\|\cdot\|^r$ to transition from the averaging scheme to our new method seems like an arbitrary choice. Our key motivation is that in many practical applications, as discussed in Section I-B, noisy observations are corrupted by heavy-tailed noise, which creates many outliers. On the other hand, it is a well-known fact that squared-error distortion typically overamplifies the effects of outliers. It may thus be more suitable for using a smaller power $r < 2$ or even $r = 1$ for robustness in practical applications. In addition, we expect $r > 2$ to be relevant for noise distributions with lighter-than-exponential tails. This includes noises with finite support.

We would also like to note that the objective function (5) can also be interpreted in the context of facility location optimization, at least for the special case $L = 2$ (and $L = 1$). In fact, many facility location optimization problems can be formulated as clustering or quantization problems [33], [34]. For our scenario, consider packages at locations $y_{1,1}, \ldots, y_{1,m}$, which are to be processed at one of the facilities $u_1, \ldots, u_n$ and then conveyed to their destinations $y_{2,1}, \ldots, y_{2,m}$, respectively. The cost of conveying a package from one location to another can be modeled to be proportional to the rth power of the distance between the two locations [35]–[38]. Thus, the total cost of conveying the ith package through the facility at u_k is given by $\|y_{1,i} - u_k\|^r + \|u_k - y_{2,i}\|^r$. The minimum average cost of conveying all packages is then given by (5) with $L = 2$ and $\lambda_\ell = 1, \forall \ell$. Minimizing (5) corresponds to optimizing the facility locations.

III. Quantization-Theoretical Tools

A commonly utilized technique in analyzing various centroidal clustering schemes is to assume having $m \to \infty$ observations from the dataset D [11]. In particular, for the simple k-means scenario in (1), this allows one to replace the empirical sum with the integral

$$\delta_1(U) \triangleq \int \min_k \|u_k - x\|^2 f_{X_1}(x)\mathrm{d}x \tag{8}$$

where $U = (u_1, \ldots, u_n)$ is the quantizer codebook and f_{X_1} represents the probability density function (pdf) of X_1. We omit the domain of integration when it is clear from the context.

For $n > 1$, we have the asymptotic result [16], [39]

$$\min_U \ \delta_1(U) = \kappa_d n^{-\frac{2}{d}} \|f_{X_1}\|_{\frac{d}{d+2}} + o\big(n^{-\frac{2}{d}}\big) \tag{9}$$

where the constant κ_d depends only on the dimension d and $\|f_{X_1}\|_p \triangleq (\int (f_{X_1}(x))^p \mathrm{d}x)^{(1/p)}$ is the p-norm of the density f_{X_1}. The sequence of quantizer codebooks that achieve the performance in (9) has the following property. There exists a continuous function $\lambda(x)$ such that at the cube $[x, x + \mathrm{d}x]$, optimal quantizer codebooks contain $n\lambda(x)\mathrm{d}x$ points as $n \to \infty$. Hence, $\lambda(x)$ can be thought as a "point density function" and obeys the normalization $\int \lambda(x)\mathrm{d}x = 1$. For the squared-error distortion, the optimal point density function depends on the input distribution through

$$\lambda(x) = f_{X_1}^{\frac{d}{d+2}}(x) \Big/ \int f_{X_1}^{\frac{d}{d+2}}(y)\mathrm{d}y. \tag{10}$$

Equivalently, we say that λ is proportional to $f_{X_1}^{(d/(d+2))}$. The question is now how the $n\lambda(x)\mathrm{d}x$ quantization points are to be deployed optimally inside each cube $[x, x + \mathrm{d}x]$. Since the underlying density f_{X_1} is approximately uniform on $[x, x+\mathrm{d}x]$, the question is equivalent to finding the structure of an optimal quantizer for a uniform distribution. For one and two dimensions, the optimal quantizers are known to be the uniform and the hexagonal lattice quantizers, respectively (thus, the $n\lambda(x)\mathrm{d}x$ points should follow a hexagonal lattice on the square $[x, x + \mathrm{d}x]$ in an optimal quantizer). We thus have $[b]\kappa_1 = (1/12)$ and $\kappa_2 = (5/(18\sqrt{3}))$, corresponding to the normalized second moment of the interval and the regular hexagon, respectively. For a set $A \subset \mathbb{R}^d$ with $\int_A x\mathrm{d}x = 0$, the normalized second moment is defined as $\kappa(A) \triangleq \int_A \|x\|^2 \mathrm{d}x / (\int_A \mathrm{d}x)^{((d+2)/d)}$. For $d \geq 3$, the optimal quantizer and κ_d remain unknown.

IV. Analysis of the Clustering Scheme

Let us now consider our problem of quantizing multiple sources to a common cluster center, as formulated in (5). For each ℓ, we assume that $\{y_{\ell,i}\}_{i=1}^m$ is drawn according to a random vector X_ℓ. Following the formulation in Section III, we replace the empirical sum in (5) with the integral

$$\delta(U) \triangleq \int \min_k \sum_{\ell=1}^{L} \lambda_\ell \|u_k - x_\ell\|^r f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \tag{11}$$

where $\mathbf{x} = [x_1, \ldots, x_L]$ represents a realization of the random matrix $\mathbf{X} = [X_1, \ldots, X_L]$ and $f_{\mathbf{X}}(\mathbf{x})$ is the pdf of $\mathbf{X}$.

A. Squared-Error Distortions

We first consider squared-error distortions $r = 2$, which allows a simple characterization of the optimal clustering scheme. In fact, for $r = 2$, the decomposition (3) leads to

$$\delta(U) = c_2 + c_1 \int \min_k \|u_k - z\|^2 f_Z(z)\mathrm{d}z \tag{12}$$

where $c_1 = \sum_{\ell=1}^L \lambda_\ell$ is defined in Section I-B, $Z \triangleq (1/c_1)\sum_{\ell=1}^L \lambda_\ell X_\ell$ is the average of noisy samples, and $c_2 \triangleq \mathrm{E}[\sum_{\ell=1}^L \lambda_\ell \|X_\ell\|^2 - (1/c_1)\|\sum_{\ell=1}^L \lambda_\ell X_\ell\|^2]$ is a constant. We observe that the integral in (12) is merely the average squared-error distortion of a quantizer given a source with density Z. Therefore, when $r = 2$, the optimal quantization of multiple sources to a common cluster center is equivalent to the optimal quantization of the single source Z with the usual squared-error distortion measure. It follows that the results of Section III are directly applicable, and we have the following result.

Proposition 1: Let $r = 2$. As $n \to \infty$, we have

$$\min_U \ \delta(U) = c_2 + c_1\kappa_d n^{-\frac{2}{d}} \|f_Z\|_{\frac{d}{d+2}} + o\big(n^{-\frac{2}{d}}\big). \tag{13}$$

Moreover, the optimal point density function that achieves (13) is proportional to $f_Z^{(d/(d+2))}$.

Proof: Formula (13) follows immediately from (10) and (12). The optimal point density function is given by (9). □

This precisely characterizes the asymptotic average distortion for $r = 2$. When $L = 1$, which corresponds to ordinary quantization with squared-error distortion, the average distortion decays to zero as the number of quantizer centers n grows to infinity. Proposition 1 demonstrates that when $L > 1$, the average distortion converges to c_2, which is in general nonzero. The reason is that when $L > 1$, a single quantizer center is used to reproduce multiple sources, which makes zero distortion impossible to achieve whenever the sources are not identical.

We conclude this section with a few remarks on the asymptotic nature of the analysis. We also discuss how to interpret and make use of the asymptotic average distortion and point density function formulas provided by the analysis. First, although the distortion formula in (1) is asymptotic in nature, the first two terms in the right-hand side of (13) can be used to approximate the performance at any finite n. The numerical results in Section VI shall demonstrate that even when n is as low as 4 or 8, this provides a very accurate approximation of the average distortion. The average distortion itself characterizes the representation accuracy of the clustering

process. The fact that the average distortion can never be made zero (i.e., bounded below by c_2) is a testament to the fact that finding the true clustering is impossible in general. This is because the observations have been corrupted by continuous noise. The optimal point density $f_Z^{(d/(d+2))}$ provides important global importance about the structure of the optimal clustering scheme as it indicates the spatial location of the cluster centers. It can be directly used to design optimal clustering schemes without going through the time-consuming Lloyd-based numerical design process. An example will be provided in Section VI.

B. Distortions With Arbitrary Powers of Errors

We now consider the achievable performance for a general $r \neq 2$. We also consider the case of $L = 2$ observations. Without loss of generality, let $\lambda_1 = 1$. In this case, the objective function in (11) takes the form

$$\delta(U) = \iint \min_k \{\|x_1 - u_k\|^r + \lambda_2 \|x_2 - u_k\|^r\} \times f_{X_1,X_2}(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2. \quad (14)$$

The main difficulty for $r \neq 2$ is that an algebraic manipulation of the form (3) is not available. Nevertheless, it turns out that an analysis in the high-resolution regime $n \to \infty$ is still feasible. First, we need the following basic lemma.

Lemma 1: Let $\xi(u) = \|x_1 - u\|^r + \lambda_2 \|x_2 - u\|^r$. The global minimizer of ξ is $z = ((x_1 + \alpha x_2)/(1 + \alpha))$, where $\alpha \triangleq \lambda_2^{(1/(r-1))}$. The corresponding global minimum is $\xi(z) = (\alpha^{r-1}/((1 + \alpha)^{r-1}))\|x_1 - x_2\|^r$.

Proof: Note that ξ is convex in u so that it is has a global minimum. Observe that this global minimum should be located on the line l that connects x_1 and x_2. In fact, suppose that the global minimizer z does not belong to l. We can project z to l to come up with a new point z' that satisfies $\|x_1 - z'\| < \|x_1 - z\|$ and $\|x_2 - z'\| < \|x_2 - z\|$. This implies $\xi(z') < \xi(z)$ and thus contradicts the optimality of z. Given that z should be on l, it can be written as $z = ((x_1 + \alpha x_2)/(1 + \alpha))$ for some $\alpha \in \mathbb{R}$. We have

$$\xi(z) = \left\|x_1 - \frac{x_1 + \alpha x_2}{1 + \alpha}\right\|^r + \lambda_2 \left\|x_2 - \frac{x_1 + \alpha x_2}{1 + \alpha}\right\|^r = \frac{\lambda_2 + \alpha^r}{(1 + \alpha)^r}\|x_1 - x_2\|^r.$$

Let us now calculate $(\partial\xi(z)/\partial\alpha) = ((r(1 + \alpha)^{r-1}\|x_1 - x_2\|^r)/((1 + \alpha)^{2r}))(\alpha^{r-1} - \lambda_2)$. According to the derivative, $\xi(z)$ is decreasing for $\alpha^{r-1} - \lambda_2 < 0$ and increasing for $\alpha^{r-1} - \lambda_2 > 0$. The global minimum is thus achieved for $\alpha = \lambda_2^{(1/(r-1))}$. Substituting this optimum value for α, we obtain the same expression for $\xi(z)$ as in the statement of the lemma. This concludes the proof. □

According to Lemma 1, given one data point at $X_1 = x_1$, and the other at $X_2 = x_2$, the minimum cost can be achieved by using a cluster center at $z = ((x_1 + \alpha x_2)/(1 + \alpha))$. Then, given a hypothetically infinite number of cluster centers, one can achieve the optimal performance by placing the centers at every possible location imaginable. On the other hand, given only finitely many centers, one has no choice but be content with choosing the center that is close to the optimal location z. In such a scenario, it makes sense to analyze the behavior of the function $\xi(u)$ near its optimal value $u = z$. For this purpose, we utilize the following lemma.

Lemma 2: Let $v \triangleq x_1 - x_2$, $\vec{v} = v/\|v\|$, and $\Delta(u, z) \triangleq (u - z)^T(\mathbf{I} + (r - 2)\vec{v}\vec{v}^T)(u - z)$. We have

$$\xi(u) = \frac{\alpha^{r-1}\|v\|^r}{(1 + \alpha)^{r-1}} + \frac{\alpha^{r-2} r \|v\|^{r-2}}{2(1 + \alpha)^{r-3}}\Delta(u, z) + o(\|u - z\|^2). \quad (15)$$

Proof: Let $u = z + \epsilon$, where $\|\epsilon\|$ is small. We have

$$\xi(z + \epsilon) = \left\|x_1 - \frac{x_1 + \alpha x_2}{1 + \alpha} - \epsilon\right\|^r + \lambda \left\|x_2 - \frac{x_1 + \alpha x_2}{1 + \alpha} - \epsilon\right\|^r = \frac{1}{(1 + \alpha)^r}\left(\alpha^r \left\|x_1 - x_2 - \epsilon\frac{1 + \alpha}{\alpha}\right\|^r + \lambda\|x_1 - x_2 + \epsilon(1 + \alpha)\|^r\right). \quad (16)$$

Now, let $\nabla f(w)$ and $\nabla^2 f(w)$ denote the gradient and the Hessian of a multivariate function f evaluated at w, respectively. We have the generic multivariate Taylor series expansion

$$f(w + \epsilon) = f(w) + \epsilon^T \nabla f(w) + \frac{1}{2}\epsilon^T \nabla^2 f(w)\epsilon + o(\|\epsilon\|^2). \quad (17)$$

In order to expand (16) for small ϵ, we need to find the Taylor expansion of the function $w \mapsto \|w\|^r$. For this purpose, for any vector w, we let $\vec{w} = w/\|w\|$. The gradient and the Hessian of $w \mapsto \|w\|^r$ can then be calculated as $\nabla\|w\|^r = r\|w\|^{r-2}w$ and $\nabla^2\|w\|^r = r\|w\|^{r-2}(\mathbf{I} + (r - 2)\vec{w}\vec{w}^T)$, respectively, where $\mathbf{I}$ is the identity matrix. Substituting to (17), we obtain $\|w + \epsilon\|^r = \|w\|^r + r\|w\|^{r-2}w^T\epsilon + ((r\|w\|^{r-2})/2)\epsilon^T(\mathbf{I} + (r - 2)\vec{w}\vec{w}^T)\epsilon + o(\|\epsilon\|^2)$. Using this expansion in (16) leads to $\xi(z + \epsilon) = (\alpha^{r-1}/((1 + \alpha)^{r-1}))\|v\|^r + ((r\alpha^{r-2})/(2(1 + \alpha)^{r-3}))\|v\|^{r-2}\epsilon^T(\mathbf{I} + (r - 2)\vec{v}\vec{v}^T)\epsilon + o(\|\epsilon\|^2)$. Substituting $u - z$ in place of ϵ concludes the proof. □

Now, let $c_3 = ((r\alpha^{r-2})/(2(1 + \alpha)^{r-3}))$ and $c_4 = (\alpha/(1 + \alpha))^{r-1}E\|X_1 - X_2\|^r$. Substituting (15) into (14) and using $\min_k(\phi_k + o(\phi_k)) = \min_k \phi_k + o(\min_k \phi_k)$ for arbitrary functions ϕ_k yield

$$\delta(U) = c_3 \iint \min_k \|v\|^{r-2}\Delta(u_k, z) f_{\left[\begin{smallmatrix}X_1\\X_2\end{smallmatrix}\right]}\left(\left[\begin{smallmatrix}x_1\\x_2\end{smallmatrix}\right]\right)\mathrm{d}x_1\mathrm{d}x_2 + c_4 + \iint o\left(\min_k \|u_k - z\|^2\right) f_{X_1,X_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2. \quad (18)$$

According to (9), the last term is $o(n^{-(2/d)})$. For the second term, a change of variables $v = x_1 - x_2$, $z = ((x_1 + \alpha x_2)/(1 + \alpha))$ yields

$$\delta(U) = c_4 + c_3 \int \min_k (z - u_k)^T B(z)(z - u_k)\mathrm{d}z + o\left(n^{-\frac{2}{d}}\right)$$

$$B(z) = \int \|v\|^{r-2}(\mathbf{I} + (r - 2)\vec{v}\vec{v}^T) f_{X_1,X_2} \times \left(z + \frac{\alpha v}{1 + \alpha}, z - \frac{v}{1 + \alpha}\right)\mathrm{d}v. \quad (19)$$

Note that for any v and $r \geq 1$, the matrix $\mathbf{I} + (r - 2)\vec{v}\vec{v}^T$ is positive semidefinite. This implies that the matrix $B(z)$ is positive semidefinite for every $r \geq 1$.

The function $(z, u) \mapsto (z - u)^T B(z)(z - u)$ defines an input-weighted quadratic distortion measure. The structure and distortion of the optimal quantizers corresponding to such distortion measures have been studied in [9]. As discussed above, the matrix $B(z)$ is positive semidefinite for every z so that the results of [9] are applicable. In particular, we have

$$\min_U \int \min_k (z - u_k)^T B(z)(z - u_k)\mathrm{d}z = \kappa_d n^{-\frac{2}{d}} \left\|(\det B(z))^{\frac{1}{d}}\right\|_{\frac{d}{d+2}} + o\left(n^{-\frac{2}{d}}\right) \quad (20)$$

and the optimal point density that achieves (20) is

$$z \mapsto (\det B(z))^{\frac{1}{d+2}} \Big/ \int (\det B(z))^{\frac{1}{d+2}}\mathrm{d}z. \quad (21)$$

For our specific problem, we obtain the following theorem.

Theorem 1: Let $r \geq 1$ and $L = 2$. As $n \to \infty$, we have

$$\min_{U} \ \delta(U) = c_4 + c_3 \kappa_d n^{-\frac{2}{d}} \left\| (\det B(z))^{\frac{1}{d}} \right\|_{\frac{d}{d+2}} + o\left(n^{-\frac{2}{d}}\right). \tag{22}$$

The optimal point density function that achieves (22) is proportional to $(\det B(z))^{(1/(d+2))}$.

Proof: Formula (22) follows immediately via (19) and (20). The optimal point density function is provided by (21). □

For ordinary center-based quantization with the rth power distortion measure, classical results [7], [16] imply that the average distortion decays as $n^{-(r/d)}$. These results only consider the reproduction of one source sample with one quantization point. Our analysis extends the classical results to the case where a common reproduction point is used to recover two sources simultaneously. It is interesting to note that the average distortion in this case decays as $n^{-(2/d)}$, independently of r.

Let us now discuss certain special cases of the conclusions of Theorem 1 above.

Example 1: For $r = 2$, we have $\alpha = \lambda_2$, and we can easily verify $c_4 = c_3$, $c_2 = c_1$. Moreover, $B(z) = \int f_{X_1,X_2}(z + (\alpha v/(1+\alpha)), z - (v/(1+\alpha)))\mathrm{d}v\mathbf{I} = ((1+\alpha)/\alpha) \int f_{X_1,X_2}(v, ((z(1+\alpha)-v)/\alpha))\mathrm{d}v\mathbf{I} = f_Z(z)\mathbf{I}$. The second equality follows from a change of variables $z + (\alpha v/(1+\alpha)) \leftarrow v$, and the last equality follows once we view the pdf of $Z = ((X_1 + \alpha X_2)/(1+\alpha))$ as a convolution. Substituting the derived equalities, we find that the conclusions of Proposition 1 and Theorem 1 become identical. □

Example 2: Let $d = 1$, $\alpha > 1$. Suppose X_1 and X_2 are independent and uniform on [0, 1]. After some basic calculations, we obtain

$$B(z) = \begin{cases} z^{r-1}(1+\alpha)^{r-1}\left(1 + \dfrac{1}{\alpha^{r-1}}\right), & z \in \left[0, \dfrac{1}{1+\alpha}\right] \\ \left(\dfrac{1+\alpha}{\alpha}\right)^{r-1} ((1-z)^{r-1} + z^{r-1}) & \\ \qquad z \in \left[\dfrac{1}{1+\alpha}, \dfrac{\alpha}{1+\alpha}\right] & \\ (1-z)^{r-1}(1+\alpha)^{r-1}\left(1 + \dfrac{1}{\alpha^{r-1}}\right) & \\ \qquad z \in \left[\dfrac{\alpha}{1+\alpha}, 1\right] & \\ 0, \quad z \notin [0,1]. \end{cases} \tag{23}$$

According to Theorem 1, the optimal point density at z is proportional to the cube root of $B(z)$. The normalizing constant can be calculated to be

$$\begin{aligned} &\int_0^1 (B(z))^{\frac{1}{3}} \mathrm{d}z \\ &\quad = \frac{6}{(r+2)(1+\alpha)}\left(1 + \frac{1}{\alpha^{r-1}}\right)^{\frac{1}{3}} \\ &\qquad + \left(1 + \frac{1}{\alpha}\right)^{\frac{r-1}{3}} \int_{\frac{1}{1+\alpha}}^{\frac{\alpha}{1+\alpha}} ((1-z)^{r-1} + z^{r-1})^{\frac{1}{3}} \mathrm{d}z. \end{aligned} \tag{24}$$

The integral in (24) cannot be expressed in terms of elementary functions but can easily be evaluated numerically. Also, for the special case of $\alpha = 1$, the integral vanishes so that we have simply $\int_0^1 (B(z))^{(1/3)}\mathrm{d}z = ((2^{(1/3)}3)/(r+2))$. Also, when X_1 and X_2 are independent and uniform on [0, 1], the random variable $\|X_1 - X_2\| = |X_1 - X_2|$ has pdf $f_{|X_1-X_2|}(z) = 2(1-z)$, $z \in [0,1]$. Therefore, $\mathrm{E}\|X_1 - X_2\|^r = \int_0^1 z^r f_{|X_1-X_2|}(z)\mathrm{d}z = (2/((r+1)(r+2)))$. A closed-form asymptotic expression for the optimal asymptotic distortion can then obtained via Theorem 1. One only needs to numerically evaluate the integral in (24). In particular, for $\alpha = 1$, we obtain $\delta(U) = (2^r/((r+1)(r+2))) + (18r/(2^r(r+2)^3))(1/n^2) + o(1/n^2)$. According to (23) and Theorem 1, the corresponding optimal point density function is proportional to $(1 - |2z-1|)^{r-1}$ on [0, 1] and vanishes everywhere else. □

We leave a mathematical analysis for the general case of $L > 2$ sources as future work. Note, however, that the numerical design of the quantizer [i.e., a minimization of (11)] in the general case is always possible via Algorithm 1.

Algorithm 2 Weighting Algorithm for Noisy Clustering

1: Initialize weights $\lambda_\ell = L^{-\beta}$ uniformly.
2: Initialize the cluster centers $U = (u_1, \ldots, u_n)$ arbitrarily.
3: Iterate (6), (7), and (26) until convergence of (5).
4: For $i = 1, \ldots, m$, set the final clustering $\zeta_i = j$, where $i \in V_j$.

V. Automatic Weight Optimization

We now describe an algorithm to optimize the weights λ_ℓ. When the observation noises are independent and identically distributed, one can use equal weights for samples. For the remaining cases, we describe an automated weight update algorithm inspired by the previous work [40]–[43]. The idea is to minimize the cost function (5) over the clustering U as well as weights $\lambda_1, \ldots, \lambda_L$ that satisfy the constraint $\sum_{\ell=1}^L \lambda_\ell^{1/\beta} = 1$, where $\beta > 1$ is some hyperparameter. In the following, we discuss how to perform this minimization for a fixed cluster assignment of samples. Letting V_j denote the Voronoi cell for clutter center j as in (6), we have

$$\begin{aligned} \delta(U) &= \frac{1}{m}\sum_{i=1}^m \min_k \sum_{\ell=1}^L \lambda_\ell \|u_k - y_{\ell,i}\|^r \\ &= \frac{1}{m}\sum_{j=1}^n \sum_{i \in V_j} \sum_{\ell=1}^L \lambda_\ell \|u_j - y_{\ell,i}\|^r = \frac{1}{m}\sum_{\ell=1}^L \lambda_\ell \theta_\ell \end{aligned}$$

where $\theta_\ell \triangleq \sum_{j=1}^n \sum_{i \in V_j} \|u_j - y_{\ell,i}\|^r$. By reverse Hölder's inequality, $\delta(U) \geq (1/m)(\sum_{\ell=1}^L \lambda_\ell^{(1/\beta)})^\beta \ (\sum_{\ell=1}^L \theta_\ell^{-(1/(\beta-1))})^{-(\beta-1)} = (1/m)(\sum_{\ell=1}^L \theta_\ell^{-(1/(\beta-1))})^{-(\beta-1)}$, with equality if and only if $\lambda_\ell^{(1/\beta)} = c\theta_\ell^{-(1/(\beta-1))}$ for some constant c. We can solve for c using the constraint on λ_ℓ to find the optimal weights as

$$\lambda_\ell = \theta_\ell^{-\frac{\beta}{\beta-1}} \Big/ \left(\sum_{\ell=1}^L \theta_\ell^{-\frac{1}{\beta-1}}\right)^\beta. \tag{25}$$

One can now extend Algorithm 1 to alternate between the three stages of Voronoi diagram assignments, centroid calculation, and weight optimization. A naive implementation of such an algorithm leads to convergence to undesirable to local minima, where more noisy samples are assigned zero weights. We thus utilize a momentum update for the weights. Specifically, given that the local distortions θ_ℓ are calculated using the "old" weights λ_ℓ, let $\lambda_{\ell,0}$ be as defined in (25). Then, λ_ℓ is updated using a momentum rule as

$$\lambda_\ell \leftarrow \left(\mu \lambda_\ell^{\frac{1}{\beta}} + (1-\mu)\lambda_{\ell,0}^{\frac{1}{\beta}}\right)^\beta. \tag{26}$$

The full algorithm is summarized in Algorithm 2. For $r = 2$ and without the momentum update, Algorithm 4 is essentially the same as the algorithm presented in [40]. Also, if the noise statistics are identical, the algorithm essentially boils down to Algorithm 1 as uniform weights are optimal.

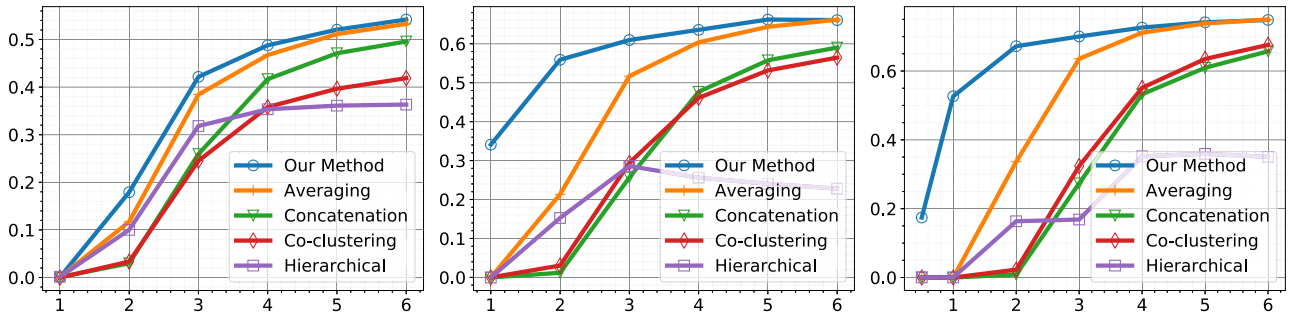


Fig. 1. Performance of our method for $r = 1$ compared to different clustering methods for t-distributed noise with different degrees of freedom τ over the Iris dataset and $n = 3$ cluster centers. The leftmost, middle, and right figures represent the cases of $L = 2, 4$, and 8 observations, respectively. For each figure, the horizontal axis represents τ, and the vertical axis represents the ARI.

VI. Numerical Results

In this section, we provide numerical experiments to show the performance of our algorithms over various datasets and to verify our analytical results.

A. Performance Evaluation Over Existing Datasets

In order to evaluate our clustering scheme, we conducted experiments over UCI benchmark datasets [44] and the 32-D Gaussian mixture dataset [45]. For more discussion on the size of clusters, we refer to [46, Table I].

Let $\mathcal{D}' = \{y'_i\}_{i=1}^m$ be one of the (noiseless) datasets. We assume that the observations are given by the additive perturbations $y_{\ell,i} = y_{\ell,i} + \eta_{\ell,i}$, $\ell = 1, \ldots, L$, $i = 1, \ldots, m$, where $\eta_{\ell,i}$s are the noise terms. Using only the resulting noisy dataset $D = \{[y_{1,i}, \ldots, y_{L,i}]\}_{i=1}^m \subset \mathbb{R}^{d\times L}$, we wish to obtain a clustering that is as close to the clustering of $\mathcal{D}'$ as possible. We have used the adjusted Rand index (ARI) [47] to quantify the closeness of clusterings and thus to compare different methods. Specifically, we utilize our clustering method for different values of r and equal weighting across observations unless otherwise specified. We compare our method with the benchmark methods of averaging, co-clustering, concatenation, and hierarchical clustering, as described in Section I-B. Note that averaging is equivalent to our method when $r = 2$. We evaluate different methods by generating a sufficiently large number of random observation instances and averaging out the resulting ARI measures. We generate enough samples so that a Monte Carlo similarity measure average γ is accurate within $[\gamma - 0.01, \gamma + 0.01]$ with 95% confidence. Unless otherwise specified, we consider independent and identically distributed noises across observations. Therefore, the weights λ_ℓ corresponding to each observation can be set to unity without loss of optimality, and Algorithm 1 can be utilized as the clustering algorithm.

In Fig. 1, we show the numerical results for the Iris dataset and evaluate our method for the case $r = 1$. We consider t-distributed noise samples with pdf $x \mapsto ((\Gamma((1+\nu)/2))/(\sqrt{\nu\pi}\,\Gamma(\nu/2)))(1 + (x^2/\nu))^{-(1/2)(1+\nu)}$, $x \in \mathbb{R}$, where $\nu > 0$ is the degree of freedom and $\Gamma(\cdot)$ is the gamma function. Practical values for ν include $\nu \in [1, 4]$ [25]. As mentioned in Section I-B, heavy-tailed distributions create many observations that deviate significantly from the ground truth observation. They can thus accurately model various real-life phenomena that include outliers such as in biology. We also consider Gaussian and uniformly distributed noise.

Fig. 1 shows the numerical results for the Iris dataset and the comparison of our method for $r = 1$ with state-of-the-art competing methods. We can observe that our method outperforms all existing methods for all different values of the degree of freedom ν and number of observations L. In fact, the performance gain becomes very significant as the noise becomes more heavy-tailed and severe (lower values of ν) when a large number of observations are available. In fact, we can observe that the ARI with our method gradually increases from 0 to 0.53 as one increases the number of observations from 2 to 8, while the ARI with other methods remains zero. We can also observe that as the noise becomes less severe, the performance of averaging (which corresponds to $r = 2$ with our method) approaches to the performance of our method.

TABLE I

Performance Results for the Iris Dataset. Observation Components Are Either Perturbed by Additive Gaussian Noises With Zero Mean and Variance σ^2 (Columns Captioned With σ^2s) or Additive Noises That Are Uniform on $[-\epsilon, +\epsilon]$ (Columns Captioned With ϵs)

	$L = 2, n = 3$				$L = 8, n = 6$			
	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\epsilon = 0.25$	$\epsilon = 1$	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\epsilon = 0.25$	$\epsilon = 1$
Our Method, $r = 1$	0.868	**0.622**	0.914	0.754	0.734	0.497	0.769	0.633
Our Method, $r = 3$	0.909	0.621	**0.954**	**0.777**	0.728	0.502	0.755	**0.646**
Averaging ($r = 2$)	0.897	**0.622**	0.938	0.762	**0.735**	**0.508**	0.762	0.634
Concatenation	**0.900**	0.606	0.947	0.762	0.726	0.461	0.765	0.606
Co-clustering	0.844	0.501	0.920	0.687	0.713	0.364	**0.770**	0.512
Hierarchical	0.836	0.424	0.908	0.664	0.598	0.239	0.696	0.404

TABLE II

Results for Handwritten Digits Dataset

	$L = 2$			$L = 4$			$L = 8$		
	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
Our Method, $r = 1$	**0.000**	**0.060**	0.767	**0.164**	**0.727**	0.803	**0.114**	**0.732**	**0.814**
Averaging ($r = 2$)	**0.000**	0.043	**0.773**	0.000	0.242	**0.804**	0.000	0.489	**0.814**
Concatenation	**0.000**	0.004	0.716	0.000	0.002	0.738	0.000	0.001	0.759
Co-clustering	**0.000**	0.002	0.617	0.000	0.000	0.511	0.000	0.000	0.316
Hierarchical	**0.000**	0.008	0.738	0.000	0.015	0.673	0.000	0.033	0.113

In Table I, we provide the results for the Iris dataset for uniform and Gaussian noises with different variances. We can observe that the huge advantage of using our method with $r = 1$ over other methods has largely disappeared. In fact, $r = 2$ outperforms $r = 1$ for almost all the scenarios considered. This is expected as the considered noises are light-tailed or even have finite support. For light-tailed noises, with the intuition that the mean-squared error ($r = 2$) should be optimized for the Gaussian distribution, our idea is to increase r beyond $r = 2$ so that the quantizer is potentially better "matched" to the noise. In fact, using $r = 3$ with our method can still provide a modest gain compared with other methods, especially for the case of uniform noise. This suggests that r can be optimized further depending on the dataset and the noise statistics for better performance. We leave such optimizations as future work.

Tables II and III show the results for Handwritten Digits and Gaussian Mixture datasets, respectively. Similar to the results in

TABLE III
RESULTS FOR THE GAUSSIAN MIXTURE DATASET

	$L=2$					$L=4$					$L=8$				
	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$
Hierarchical	0.002	0.031	**0.311**	**0.722**	**0.998**	0.007	0.070	0.597	**0.916**	**1.000**	0.012	0.137	**0.760**	**0.967**	**1.000**
Co-clustering	0.001	0.001	0.008	0.105	0.969	0.000	0.000	0.000	0.007	0.931	0.000	0.000	0.000	0.000	0.087
Averaging ($r=2$)	0.001	. 0.020	0.193	0.487	0.993	0.001	0.029	0.261	0.575	0.997	0.002	0.039	0.321	0.648	0.998
Our Method, $r=1$	**0.003**	**0.023**	0.190	0.481	0.993	**0.616**	**0.575**	**0.653**	0.780	0.997	**0.676**	**0.625**	0.700	0.815	0.998
Concatenation	0.001	0.003	0.045	0.246	0.984	0.000	0.001	0.010	0.128	0.982	0.000	0.000	0.000	0.052	0.977

TABLE IV
RESULTS FOR THE BREAST CANCER DATASET

	$L=2$					$L=4$					$L=8$				
	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$	$\nu=1$	$\nu=1.1$	$\nu=1.2$	$\nu=1.3$	$\nu=2$
Hierarchical	−0.005	**0.483**	**0.789**	**0.932**	**1.000**	−0.006	0.733	0.930	**0.932**	0.998	0.405	0.903	**0.995**	**0.998**	**1.000**
Co-clustering	**0.000**	0.103	0.412	0.570	0.999	0.000	0.000	0.140	0.570	**1.000**	0.000	0.000	0.000	0.362	**1.000**
Averaging ($r=2$)	−0.002	. 0.339	0.656	0.752	0.999	0.001	0.334	0.643	0.752	0.997	0.062	0.314	0.723	0.848	**1.000**
Our Method, $r=1$	−0.002	0.338	0.669	0.718	0.940	**0.938**	**0.939**	**0.940**	0.940	0.940	**0.940**	**0.940**	0.940	0.940	0.940
Concatenation	−0.002	0.212	0.521	0.676	0.999	0.001	0.135	0.418	0.676	0.997	0.000	0.007	0.248	0.662	**1.000**

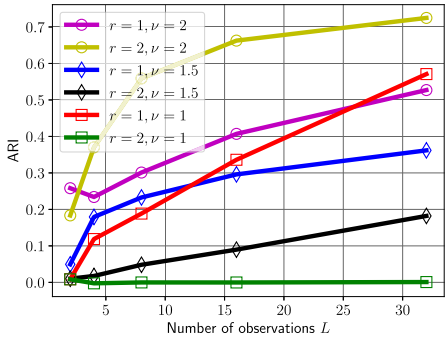


Fig. 2. Results for the Glass dataset.

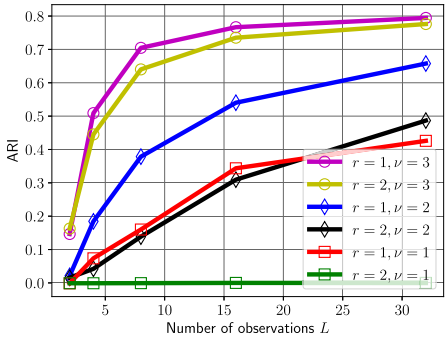


Fig. 3. Results for the Thyroid dataset.

Fig. 1 for the Iris dataset, our method for $r=1$ is superior to the existing methods, especially for small degrees of freedom ν. For the Gaussian mixture dataset, we provide results for a "condensed" set of degrees of freedom $\nu \in \{1.0, 1.1, 1.2, 1.3, 2\}$ to demonstrate how quickly the performance can vary depending on ν for certain datasets. While our method performs the best for small ν, as ν becomes larger, hierarchical clustering outperforms all centroid-based methods. This suggests hierarchical clustering as an alternative for certain datasets with well-separated clusters. A similar phenomenon can be observed for the results for the Breast Cancer dataset, as shown in Table IV. Also, for the Gaussian Mixture dataset, although it is not shown in the table, for $\nu=3$, all methods provide an ARI of 1.

In Fig. 2, we show the numerical results for the Glass dataset. The horizontal axis represents the number of observations, and the vertical axis represents the ARI. Here, we only show the results for $r=1$ and $r=2$ (averaging) as the other competitor methods provided strictly inferior performance than either scheme under all conditions considered. We can observe that improving the number of observations generally improves the similarity measure. Moreover, it is more appropriate to use $r=2$ when the noise distribution has lighter tail (increasing values of ν). In fact, the choice $r=1$ generally outperforms $r=2$ if $\nu \in \{1, 1.5\}$, while $r=2$ outperforms $r=1$ if $\nu=2$. We have observed similar behavior in the previous datasets as well, for different cutoff values for ν. In particular, the results for the Thyroid dataset in Fig. 3 suggest that the cutoff value for ν is slightly above 3.

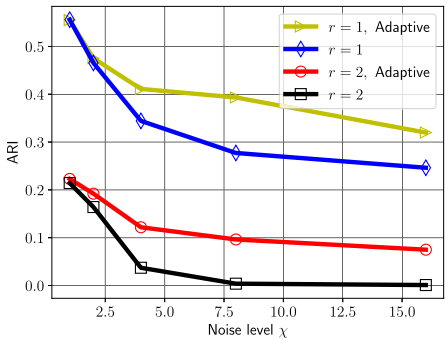
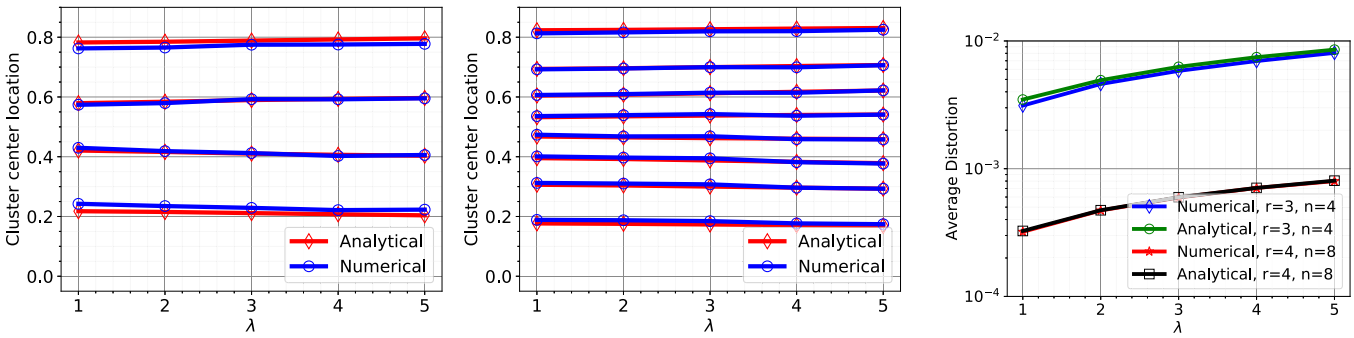


Fig. 4. Results for Algorithm 2.

We now validate the performance of our automatic weighting scheme provided by Algorithm 2. For this purpose, we consider the Iris dataset with four observations. We assume that all four observations are perturbed by the t-distributed noise with parameter $\nu=2$. In addition, the second and the third observation noises are scaled by a factor of χ. We have applied Algorithm 4 with $\beta=2$ and the momentum parameter $\mu=0.5$, for $r \in \{1, 2\}$. In Fig. 4, the horizontal axis represents χ, and the vertical axis denotes the resulting ARI. We can observe that when $\chi=1$ (all noise levels are the same), both equal weighting and the adaptive weighting strategies provide the same performance. This verifies that for equal noise levels, equal weightings are optimal. As the noise level increases at the second and third observations, the ARI performance generally decreases.



(a) (b) (c)

Fig. 5. Analytical versus numerical cluster center locations and average distortions for the uniform sources in Example 2. (a) Cluster center locations for $n = 4$ and $r = 3$. (b) Cluster center locations for $n = 8$ and $r = 4$. (c) Average distortions of quantizers in (a) and (b).

TABLE V
EXECUTION TIMES IN MILLISECONDS OF DIFFERENT SCHEMES

		Hier.	Maj.	$r = 2$	$r = 1$	Con.
Iris, $L = 4$	$\nu = 1$	7.9	2.3	1.4	24.5	1.2
	$\nu = 2$	9.9	2.6	1.8	18.4	1.2
	$\nu = 3$	14.0	4.3	2.1	16.5	1.7
	$\nu = 4$	16.3	5.2	1.9	14.9	1.8
	$\nu = 5$	17.6	5.3	2.3	14.0	2.0
Glass, $L = 8$	$\nu = 1$	36.5	7.0	2.0	71.1	4.4
	$\nu = 2$	43.0	8.1	3.1	52.1	4.5
	$\nu = 3$	43.1	15.9	3.1	38.4	5.1
	$\nu = 4$	45.5	17.4	2.8	33.6	5.1
	$\nu = 5$	42.8	17.8	2.8	29.1	4.7
Thyroid, $L = 32$	$\nu = 1$	49.6	5.9	1.7	49.2	2.7
	$\nu = 2$	52.2	6.1	2.3	50.1	2.7
	$\nu = 3$	65.0	7.4	2.5	46.4	3.2
	$\nu = 4$	83.9	10.6	2.5	45.3	3.5
	$\nu = 5$	91.0	13.0	2.4	45.6	3.6

However, we observe that the adaptive weighting strategies are more robust to heterogeneous noises across observations compared to equal weighting strategies. Also, strategies where $r = 1$ generally outperform those with $r = 2$ as before.

In Table V, we show the execution times of different clustering schemes for different datasets and parameters. We can observe that $r = 1$ is considerably slower than $r = 2$, with the main bottleneck being the centroid calculation. In fact, for $r = 2$, the centroid calculation is a mere average of noisy samples, while for $r = 1$, one has to perform a computationally expensive gradient descent step. Among clustering schemes with simple centroid calculation (i.e., majority voting, averaging, and concatenation), majority voting has the highest complexity. This is because it factors the effective dataset size by L requiring more time for the Lloyd algorithm to converge. Also, concatenation has slightly higher complexity than averaging ($r = 2$) as it factors the effective dataset dimension by L, again resulting in longer convergence times. Hierarchical clustering and our method with $r = 1$ have comparable time complexity, with our method often running faster. In general, we observe that if the resulting ARI performance is high, our algorithm typically converges faster. This can be verified by comparing with the corresponding ARI results in Figs. 2 and 3. Hence, although our algorithm is somewhat slower than the ordinary ℓ_2-norm-based clustering methods, the performance gains justify the cost in execution time.

B. Validation of High-Resolution Analysis

We now provide numerical experiments that verify our high-resolution analysis. We consider the same scenario as in Example 2 for different values of r and n. We compare the cluster centers obtained using our generalized Lloyd algorithm (labeled "Numerical") with those provided by Theorem 1 (labeled "Analytical") for $n = 4$ and $r = 3$ in Fig. 5(a) and $n = 8$ and $r = 4$ in Fig. 5(b). The horizontal axis represents λ, and the vertical axis represents the cluster center or the quantization point locations. Note that Theorem 1 provides the optimal quantizer point density function, not the individual quantization points or cluster centers. We may use, however, inverse transform sampling to obtain a sequence of quantization points that will be faithful to the quantizer point density function. Namely, if the desired point density function is $\lambda(x)$, we use the quantization points $\Lambda^{-1}((2i-1)/2n)$, $i = 1, \ldots, n$, where $\Lambda(y) \triangleq \int_{-\infty}^{y} \lambda(x)\mathrm{d}x$ is the cumulative point density function.

Although the results of Theorem 1 are only valid asymptotically, we can observe that they still provide a very accurate description of the optimal quantizers for a number of cluster centers as low as $n = 4$. There is only slight mismatch for centers that are close to the boundaries 0 and 1. A similar phenomenon can be observed in Fig. 5(b), but the amount of mismatch is lower. Also, the theory can precisely predict very subtle changes in the optimal quantization points, e.g., the movement of the optimal location for the third quantization point from 0.4 from 0.38 as λ grows from 1 to 5. In Fig. 5(c), we show the average distortion performances corresponding to the cluster centers in Fig. 5. Again, the quantization theory can precisely predict the average distortion performance even when n is as small as 4. The difference between the analytical and the numerical results becomes indistinguishable for the case of $n = 8$ cluster centers.

VII. CONCLUSION

We have introduced a new centroid-based clustering method for data with multiple noisy observations. Our method has performed remarkably well in the case of heavy-tailed noise and outperformed various existing methods in the literature. We have also analyzed the asymptotic centroid distribution and the average distortion with our clustering scheme. Future directions include the extension of our analytical results to the case of more than two noisy observations. Designing an automatic power optimization scheme, similar to an automatic weighting scheme is also of great interest.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.
[2] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2008.

[3] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA, USA: SIAM, 2007, vol. 20.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, no. 14, Oakland, CA, USA, 1967, pp. 281–297.

[5] H. Steinhaus, "Sur la division des corps matériels en parties," *Bull. de l'Académie Polonaise des Sciences, Classe III*, vol. 4, no. 12, pp. 801–804, 1956.

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

[7] J. Bucklew and G. Wise, "Multidimensional asymptotic quantization theory with rth power distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 239–247, Mar. 1982.

[8] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 367–381, Sep. 1995.

[9] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1082–1091, May 1999.

[10] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 1705–1749, 2005.

[11] A. Fischer, "Quantization and clustering with Bregman divergences," *J. Multivariate Anal.*, vol. 101, no. 9, pp. 2207–2221, Oct. 2010.

[12] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Mach. Learn.*, vol. 75, no. 2, pp. 245–248, 2009.

[13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[14] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[15] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. New York, NY, USA: Kluwer, 1992.

[16] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 139–149, Mar. 1982.

[17] C. Liu and M. Belkin, "Clustering with Bregman divergences: An asymptotic analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2351–2359.

[18] N. Farvardin, "A study of vector quantization for noisy channels," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 799–809, Jul. 1990.

[19] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Recognit. Lett.*, vol. 12, no. 11, pp. 657–664, Nov. 1991.

[20] Y.-X. Wang and H. Xu, "Noisy sparse subspace clustering," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 89–97.

[21] E. R. Dougherty *et al.*, "Inference from clustering with application to gene-expression microarrays," *J. Comput. Biol.*, vol. 9, no. 1, pp. 105–126, Jan. 2002.

[22] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biol.*, vol. 4, no. 5, Apr. 2003, Art. no. R34.

[23] R. Sloutsky, N. Jimenez, S. J. Swamidass, and K. M. Naegle, "Accounting for noise when clustering biological data," *Briefings Bioinf.*, vol. 14, no. 4, pp. 423–436, Jul. 2013.

[24] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 16, pp. 8961–8965, Jul. 2001.

[25] A. Posekany, K. Felsenstein, and P. Sykacek, "Biological assessment of robust noise models in microarray data analysis," *Bioinformatics*, vol. 27, no. 6, pp. 807–814, Mar. 2011.

[26] D. Hsu and S. Sabato, "Loss minimization and parameter estimation with heavy tails," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 543–582, 2016.

[27] Y. Cherapanamjeri, S. B. Hopkins, T. Kathuria, P. Raghavendra, and N. Tripuraneni, "Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond," in *Proc. 52nd Annu. ACM SIGACT Symp. Theory Comput.*, Jun. 2020, pp. 601–609.

[28] K. K. Saab, "Estimation of cluster centroids in presence of noisy observations," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Nov. 2016, pp. 1–4.

[29] G. Chao, S. Sun, and J. Bi, "A survey on multi-view clustering," 2017, *arXiv:1712.06246*.

[30] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2004, pp. 19–26.

[31] M. Bittner *et al.*, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[32] R. Chen, "Location problems with costs being sums of powers of Euclidean distances," *Comput. Oper. Res.*, vol. 11, no. 3, pp. 285–294, Jan. 1984.

[33] A. Okabe and A. Suzuki, "Locational optimization problems solved through Voronoi diagrams," *Eur. J. Oper. Res.*, vol. 98, no. 3, pp. 445–456, May 1997.

[34] R. Farahani, M. SteadieSifi, and N. Asgari, "Multiple criteria facility location problems: A survey," *Appl. Math. Model.*, vol. 34, no. 7, pp. 1689–1709, Jul. 2010.

[35] L. A. A. Meira, F. K. Miyazawa, and L. L. C. Pedrosa, "Clustering through continuous facility location problems," *Theor. Comput. Sci.*, vol. 657, pp. 137–145, Jan. 2017.

[36] L. A. A. Meira and F. K. Miyazawa, "A continuous facility location problem and its application to a clustering problem," in *Proc. ACM Symp. Appl. Comput.*, Jan. 2008, pp. 1826–1831.

[37] A. Czumaj, C. Lammersen, M. Monemizadeh, and C. Sohler, "$(1+\varepsilon)$-approximation for facility location in data streams," in *Proc. 24th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, Jan. 2013, pp. 1710–1728.

[38] E. Koyuncu, "Power-efficient deployment of UAVs as relays," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[39] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, Jul. 1948.

[40] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.

[41] Y.-M. Cheung, "Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 750–761, Jun. 2005.

[42] G. Gan and M. K.-P. Ng, "K-means clustering with outlier removal," *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, Apr. 2017.

[43] K. Song, X. Yao, F. Nie, X. Li, and M. Xu, "Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107560.

[44] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[45] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Appl. Intell.*, vol. 48, no. 12, pp. 4743–4759, 2018. [Online]. Available: http://cs.uef.fi/sipu/datasets/

[46] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Self-adaptive multiprototype-based competitive learning approach: A k-means-type algorithm for imbalanced data clustering," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1598–1612, Mar. 2021.

[47] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.