Recovery of a mixture of Gaussians by sum-of-norms clustering

Tao Jiang[*]   Stephen Vavasis[†]   Chen Wen Zhai[‡]

Abstract

Sum-of-norms clustering is a method for assigning n points in $\mathbf{R}^d$ to K clusters, $1 \leq K \leq n$, using convex optimization. Recently, Panahi et al. proved that sum-of-norms clustering is guaranteed to recover a mixture of Gaussians under the restriction that the number of samples is not too large. The purpose of this note is to lift this restriction, i.e., show that sum-of-norms clustering with equal weights can recover a mixture of Gaussians even as the number of samples tends to infinity. Our proof relies on an interesting characterization of clusters computed by sum-of-norms clustering that was developed inside a proof of the agglomeration conjecture by Chiquet et al. Because we believe this theorem has independent interest, we restate and reprove the Chiquet et al. result herein.

1 Introduction

Clustering is perhaps the most central problem in unsupervised machine learning and has been studied for over 60 years [10]. The problem may be stated informally as follows. One is given n points, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ lying in $\mathbf{R}^d$. One seeks to partition $\{1, \ldots, n\}$ into K sets $C_1, \ldots, C_K$ such that the $\boldsymbol{a}_i$'s for $i \in C_m$ are closer to each other than to the $\boldsymbol{a}_i$'s for $i \in C_{m'}$, $m' \neq m$.

Clustering is usually posed as a nonconvex optimization problem, and therefore prone to nonoptimal local minimizers, but Pelckmans et al. [8], Hocking et al. [5], and Lindsten et al. [6] proposed the following convex formulation for the clustering problem:

$$\min_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbf{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{a}_i\|^2 + \lambda \sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| . \tag{1}$$

---

[*]Department of Combinatorics & Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, `tao.jiang@uwaterloo.ca`.

[†]Department of Combinatorics & Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, `vavasis@uwaterloo.ca`. Research supported in part by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada.

[‡]Department of Combinatorics & Optimization and Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1, `sabrina.zhai@edu.uwaterloo.ca`.

This formulation is known in the literature as sum-of-norms clustering, convex clustering, or clusterpath clustering. Let $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$ be the optimizer. (Note: (1) is strongly convex, hence the optimizer exists and is unique.) The assignment to clusters is given by the $\boldsymbol{x}_i^*$'s: for i, i', if $\boldsymbol{x}_i^* = \boldsymbol{x}_{i'}^*$ then i, i' are assigned to the same cluster, else they are assigned to different clusters. It is apparent that for $\lambda = 0$, each $\boldsymbol{a}_i$ is assigned to a different cluster (unless $\boldsymbol{a}_i = \boldsymbol{a}_{i'}$ exactly), whereas for λ sufficiently large, the second summation drives all the $\boldsymbol{x}_i$'s to be equal (and hence there is one big cluster). Thus, the parameter λ controls the number of clusters produced by the formulation.

Throughout this paper, we assume that all norms are Euclidean, although (1) has also been considered for other norms. In addition, some authors insert nonnegative weights in front of the the terms in the above summations. Our results, however, require all weights identically 1.

Panahi et al. [7] developed several recovery theorems as well as a first-order optimization method for solving (1). Other authors, e.g., Sun et al. [11] have since extended these results. One of Panahi et al.'s results pertains to a mixture of Gaussians, which is the following generative model for producing the data $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. The parameters of the model are K means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbf{R}^d$, K variances $\sigma_1^2, \ldots, \sigma_k^2$, and K probabilities $w_1, \ldots, w_k$, all positive and summing to 1. One draws n i.i.d. samples as follows. First, an index $m \in \{1, \ldots, K\}$ is selected at random according to probabilities $w_1, \ldots, w_K$. Next, a point $\boldsymbol{a}$ is chosen according to the spherical Gaussian distribution $N(\boldsymbol{\mu}_m, \sigma_m^2 I)$.

Panahi et al. proved that for the appropriate choice of λ, sum-of-norms clustering formulation (1) will exactly recover a mixture of Gaussians (i.e., each point will be labeled with m if it was selected from $N(\boldsymbol{\mu}_m, \sigma_m^2 I)$) provided that for all m, m', $1 \le m < m' \le K$,

$$\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| \ge \frac{CK\sigma_{\max}}{w_{\min}}\text{polylog}(n). \tag{2}$$

One issue with this bound is that as the number of samples n tends to infinity, the bound seems to indicate that distinguishing the clusters becomes increasingly difficult (i.e., the $\boldsymbol{\mu}_m$'s have to be more distantly separated as $n \to \infty$).

The reason for this aspect of their bound is that their proof technique requires a gap of positive width (i.e., a region of $\mathbf{R}^d$ containing no sample points) between $\{\boldsymbol{a}_i : i \in C_m\}$ and $\{\boldsymbol{a}_i : i \in C_{m'}\}$ whenever $m \ne m'$. Clearly, such a gap cannot exist in the mixture-of-Gaussians distribution as the number of samples tends to infinity.

The purpose of this note is to prove that (1) can recover a mixture of Gaussians even as $n \to \infty$. This is the content of Theorem 2 in Section 4 below. Naturally, under this hypothesis we cannot hope to correctly label all samples since, as $n \to \infty$, some of the samples associated with one mean will be placed arbitrarily close to another mean. Therefore, we are content in showing that (1) can correctly cluster the points lying within some fixed number of standard-deviations for each mean. Radchenko and Mukherjee [9] have previously analyzed the special case of mixture of Gaussians with $K = 2$, $d = 1$ under slightly different hypotheses.

Our proof technique requires a cluster characterization theorem for sum-of-norms clustering derived by Chiquet et al. [3]. This theorem is not stated by these authors as a theorem, but instead appears as a sequence of steps inside a larger proof in a "supplementary material" appendix to their paper. Because we believe that this theorem is of independent interest,

we restate it below and for the sake of completeness provide the proof (which is the same as the proof appearing in Chiquet et al.'s supplementary material). This material appears in Section 2.

2 Cluster characterization theorem

The following theorem is due to Chiquet et al. [3] but is not stated as a theorem by these authors; instead it appears as a sequence of steps in a proof of the agglomeration conjecture. Refer to the next section for a discussion of the agglomeration conjecture. We restate the theorem here because it is needed for our analysis and because we believe it is of independent interest.

Theorem 1. *Let $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$ denote the optimizer of (1). For notational ease, let $\boldsymbol{x}^*$ denote the concatenation of these vectors into a single nd-vector. Suppose that C is a nonempty subset of $\{1, \ldots, n\}$.*

(a) Necessary condition: If for some $\hat{\boldsymbol{x}} \in \mathbf{R}^d$, $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$ and $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$ (i.e., C is exactly one cluster determined by (1)), then there exist $\boldsymbol{z}_{ij}^*$ for $i, j \in C$, $i \neq j$, which solve

$$\boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l = \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* \quad \forall i \in C,$$

$$\left\| \boldsymbol{z}_{ij}^* \right\| \leq 1 \qquad \forall i, j \in C, i \neq j,$$

$$\boldsymbol{z}_{ij}^* = -\boldsymbol{z}_{ji}^* \qquad \forall i, j \in C, i \neq j. \tag{3}$$

(b) Sufficient condition: Suppose there exists a solution $\boldsymbol{z}_{ij}^*$ for $j \in C - \{i\}$, $i \in C$ to the conditions (3). Then there exists an $\hat{\boldsymbol{x}} \in \mathbf{R}^d$ such that the minimizer $\boldsymbol{x}^*$ of (1) satisfies $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$.

Note: This theorem is an almost exact characterization of clusters that are determined by formulation (1). The only gap between the necessary and sufficient conditions is that the necessary condition requires that C be exactly all the points in a cluster, whereas the sufficient condition is sufficient for C to be a subset of the points in a cluster. The sufficient condition is notable because it does not require any hypothesis about the other $n - |C|$ points occurring in the input.

Proof. (Chiquet et al.) <u>Proof for Necessity (a)</u>
As $\boldsymbol{x}^*$ is the minimizer of the problem (1), and this objective function, call it $f(\boldsymbol{x})$, is convex, it follows that $\boldsymbol{0} \in \partial f(\boldsymbol{x}^*)$, where $\partial f(\boldsymbol{x}^*)$ denotes the subdifferential, that is, the set of subgradients of f at $\boldsymbol{x}^*$. (See, e.g., [4] for background on convex analysis). Written explicitly in terms of the derivative of the squared-norm and subdifferential of the norm, this means that $\boldsymbol{x}^*$ satisfies the following condition:

$$\boldsymbol{x}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij}^* = \boldsymbol{0} \qquad \forall i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{w}_{ij}^*$, $i = 1, \ldots, n$, $j = 1, \ldots, n$, $i \neq j$, are subgradients of the Euclidean norm function satisfying

$$\boldsymbol{w}_{ij}^* = \begin{cases} \frac{\boldsymbol{x}_i^* - \boldsymbol{x}_j^*}{\|\boldsymbol{x}_i^* - \boldsymbol{x}_j^*\|}, & \text{for } \boldsymbol{x}_i^* \neq \boldsymbol{x}_j^*, \\ \text{arbitrary point in } B(\boldsymbol{0}, 1), & \text{for } \boldsymbol{x}_i^* = \boldsymbol{x}_j^*, \end{cases}$$

with the requirement that $\boldsymbol{w}_{ij}^* = -\boldsymbol{w}_{ji}^*$ in the second case. Here, $B(\boldsymbol{x}, r)$ is notation for the closed Euclidean ball centered at $\boldsymbol{x}$ of radius r. Since $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$, $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$, the KKT condition for $i \in C$ is rewritten as

$$\hat{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\|} + \lambda \sum_{j \in C - \{i\}} \boldsymbol{w}_{ij}^* = \boldsymbol{0}, \tag{5}$$

Define $\boldsymbol{z}_{ij}^* = \boldsymbol{w}_{ij}^*$ for $i, j \in C$, $i \neq j$. Then

$$\|\boldsymbol{z}_{ij}^*\| \leq 1, \boldsymbol{z}_{ij}^* = -\boldsymbol{z}_{ji}^*, \forall i, j \in C, i \neq j.$$

Substitute $\boldsymbol{w}_{ij}^* = \boldsymbol{z}_{ij}^*$ into the equation (5) to obtain

$$\hat{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\|} + \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* = \boldsymbol{0}, \tag{6}$$

Sum the preceding equation over $i \in C$, noticing that the last term cancels out, leaving

$$|C|\hat{\boldsymbol{x}} - \sum_{i \in C} \boldsymbol{a}_i + \lambda |C| \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\|} = \boldsymbol{0},$$

which is rearranged to (renaming i to l):

$$\lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\|} = -\hat{\boldsymbol{x}} + \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l. \tag{7}$$

Subtract (7) from (6), simplify and rearrange to obtain

$$\boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l = \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* \qquad \forall i \in C, \tag{8}$$

as desired.

Proof for Sufficiency (b)

We will show that at the solution of (1), all the $\boldsymbol{x}_i^*$'s for $i \in C$ have a common value under the hypothesis that $\boldsymbol{z}_{ij}^*$ is a solution to the equation (3) for $i, j \in C$, $i \neq j$.

First, define the following intermediate problem. Let $\tilde{\boldsymbol{a}}$ denote the centroid of $\boldsymbol{a}_i$ for $i \in C$:

$$\tilde{\boldsymbol{a}} = \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l.$$

Consider the weighted problem sum-of-norms clustering problem with unknowns as follows: one unknown $\boldsymbol{x} \in \mathbf{R}^d$ is associated with C, and one unknown $\boldsymbol{x}_j$ is associated with each $j \notin C$ (for a total of with $n - |C| + 1$ unknown vectors):

$$\min_{\boldsymbol{x}; \boldsymbol{x}_j} \frac{|C|}{2} \cdot \|\boldsymbol{x} - \tilde{\boldsymbol{a}}\|^2 + \frac{1}{2} \sum_{j \notin C} \|\boldsymbol{x}_j - \boldsymbol{a}_j\|^2 + \lambda |C| \sum_{j \notin C} \|\boldsymbol{x} - \boldsymbol{x}_j\| + \lambda \sum_{\substack{i,j \notin C \\ i < j}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \quad (9)$$

This problem, being strongly convex, has a unique optimizer; denote the optimizing vectors $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}_j$ for $j \notin C$.

First, let us consider the optimality conditions for (9), which are:

$$|C|(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{a}}) + \lambda |C| \sum_{j \notin C} \boldsymbol{g}_j = \mathbf{0}, \quad (10)$$

$$\tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i - \lambda |C| \boldsymbol{g}_i + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} = \mathbf{0} \qquad \forall i \notin C, \quad (11)$$

with subgradients defined as follows:

$$\boldsymbol{g}_j = \begin{cases} \frac{\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_j}{\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_j\|}, & \text{for } \tilde{\boldsymbol{x}}_j \neq \tilde{\boldsymbol{x}}, \\ \text{arbitrary in } B(\mathbf{0}, 1), & \text{for } \tilde{\boldsymbol{x}}_j = \tilde{\boldsymbol{x}}, \end{cases} \qquad \forall j \notin C,$$

and

$$\boldsymbol{y}_{ij} = \begin{cases} \frac{\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j}{\|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\|}, & \text{for } \tilde{\boldsymbol{x}}_i \neq \tilde{\boldsymbol{x}}_j, \\ \text{arbitrary in } B(\mathbf{0}, 1), & \text{for } \tilde{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_j, \end{cases} \qquad \forall i, j \notin C, i \neq j,$$

with the proviso that in the second case, $\boldsymbol{y}_{ij} = -\boldsymbol{y}_{ji}$.

We claim that the solution for (1) given by defining $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}$ for $i \in C$ while keeping the $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$ for $j \notin C$, where $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}_j$ are the optimizers for (9) as in the last few paragraphs, is optimal for (1), which proves the main result. To show that this solution is optimal for (1), we need to provide subgradients to establish the necessary condition. Define $\boldsymbol{w}_{ij}$ to be the subgradients of $\boldsymbol{x}_i \mapsto \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_j^*\|$ evaluated at $\tilde{\boldsymbol{x}}_i^*$ as follows:

$$\begin{aligned} \boldsymbol{w}_{ij} &= \boldsymbol{g}_j & & \text{for } i \in C, j \notin C, \\ \boldsymbol{w}_{ij} &= \boldsymbol{y}_{ij} & & \text{for } i, j \notin C, \, i \neq j, \\ \boldsymbol{w}_{ij} &= \boldsymbol{z}_{ij}^* & & \text{for } i, j \in C, \, i \neq j, \end{aligned}$$

Before confirming that the necessary condition is satisfied, we first need to confirm that these are all valid subgradients. In the case that $i \in C$, $j \notin C$, we have constructed $\boldsymbol{g}_j$ to be a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated at $\tilde{\boldsymbol{x}}$, and we have taken $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}$, $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$.

In the case that $i, j \notin C$, we have construct $\boldsymbol{y}_{ij}$ to be a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated at $\tilde{\boldsymbol{x}}_i$, and we have taken $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}_i$, $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$.

In the case that $i, j \in C$, by construction $\boldsymbol{x}_i^* = \boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}$, so any vector in $B(\mathbf{0}, 1)$ is a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated $\tilde{\boldsymbol{x}}_i$. Note that since $\boldsymbol{z}_{ij}^* \in B(\mathbf{0}, 1)$, then $\boldsymbol{w}_{ij}$ defined above also lies in $B(\mathbf{0}, 1)$.

Now we check the necessary conditions for optimality in (1). First, consider an $i \in C$:

$$
\begin{aligned}
\tilde{\boldsymbol{x}}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij} &= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \in C - \{i\}} \boldsymbol{w}_{ij} + \lambda \sum_{j \notin C} \boldsymbol{w}_{ij} \\
&= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* + \lambda \sum_{j \notin C} \boldsymbol{g}_j \\
&= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l + \lambda \sum_{j \notin C} \boldsymbol{g}_j \qquad \text{(by (3))} \\
&= \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{a}} + \lambda \sum_{j \notin C} \boldsymbol{g}_j \\
&= \boldsymbol{0} \qquad \text{(by (10))}.
\end{aligned}
$$

Then we check for $i \notin C$:

$$
\begin{aligned}
\tilde{\boldsymbol{x}}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij} &= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i + \lambda \sum_{j \in C} \boldsymbol{w}_{ij} + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{w}_{ij} \\
&= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i + \lambda \sum_{j \in C} (-\boldsymbol{g}_i) + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} \\
&= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i - \lambda |C| \boldsymbol{g}_i + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} \\
&= \boldsymbol{0} \qquad \text{(by (11))}.
\end{aligned}
$$

$\square$

3   Agglomeration Conjecture

Recall that when $\lambda = 0$, each $\boldsymbol{a}_i$ is in its own cluster in the solution to (1) (provided the $\boldsymbol{a}_i$'s are distinct), whereas for sufficiently large λ, all the points are in one cluster. Hocking et al. [5] conjectured that sum-of-norms clustering with equal weights has the following agglomeration property: as λ increases, clusters merge with each other but never break up. This means that the solutions to (1) as λ ranges over $[0, \infty)$ induce a tree of hierarchical clusters on the data.

This conjecture was proved by Chiquet et al. [3] using Theorem 1. Consider a $\bar{\lambda} \geq \lambda$ and its corresponding sum-of-norms cluster model:

$$
\min_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n} \frac{1}{2} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{a}_i\|^2 + \bar{\lambda} \sum_{i < j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{12}
$$

Corollary 1.1. *(Chiquet et al.) If there is a C such that minimizer $\boldsymbol{x}^*$ of (1) satisfies $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$, $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$ for some $\hat{\boldsymbol{x}} \in \mathbf{R}^d$, then there exists an $\hat{\boldsymbol{x}}' \in \mathbf{R}^d$ such that the minimizer of (12), $\bar{\boldsymbol{x}}^*$, satisfies $\bar{\boldsymbol{x}}_i^* = \hat{\boldsymbol{x}}'$ for $i \in C$.*

The corollary follows from Theorem 1. If C is a cluster in the solution of (1), then by the necessary condition, there exist multipliers $\boldsymbol{z}_{ij}^*$ satisfying (3) for λ. If we scale each of these multipliers by $\lambda/\bar{\lambda}$, we now obtain a solution to (3) for with λ replaced by $\bar{\lambda}$, and the theorem states that this is sufficient for the points in C to be in the same cluster in the solution to (12).

It should be noted that Hocking et al. construct an example of unequally-weighted sum-of-norms clustering in which the agglomeration property fails. It is still mostly an open question to characterize for which norms and for which families of unequal weights the agglomeration property holds. Refer to Chi and Steinerberger [2] for some recent progress.

4 Mixture of Gaussians

In this section, we present our main result about recovery of mixture of Gaussians. As noted in the introduction, a theorem stating that every point is labeled correctly is not possible in the setting of $n \to \infty$, so we settle for a theorem stating that points within a constant number of standard deviations from the means are correctly labeled.

Theorem 2. *Let the vertices $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbf{R}^d$ be generated from a mixture of K Gaussian distributions with parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, $\sigma_1^2, \ldots, \sigma_K^2$, and $w_1, \ldots, w_K$. Let $\theta > 0$ be given, and let*

$$V_m = \{\boldsymbol{a}_i : \|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \le \theta \sigma_m\}, \quad m = 1, \ldots, K.$$

Let $\epsilon > 0$ be arbitrary. Then for any $m = 1, \ldots, K$, with probability exponentially close to 1 (and depending on ϵ) as $n \to \infty$, for the solution $\boldsymbol{x}^*$ computed by (1), the points in V_m are in the same cluster provided

$$\lambda \ge \frac{2\theta\sigma_m}{(F(\theta, d)w_m - \epsilon)n}. \tag{13}$$

Here, $F(\theta, d)$ denotes the cumulative density function of the chi-squared distribution with d degrees of freedom (which tends to 1 rapidly as θ increases). Furthermore, the cluster associated with V_m is distinct from the cluster associated with $V_{m'}$, $1 \le m < m' < k$, provided that

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n-1)}. \tag{14}$$

Proof. Let $\epsilon > 0$ be fixed. Fix an $m \in \{1, \ldots, K\}$. First, we show that all the points in V_m are in the same cluster. The usual technique for proving a recovery result is to find subgradients to satisfy the sufficient condition, which in this case is Theorem 1 taking C in the theorem to be V_m. Observe that conditions (3) involve equalities and norm inequalities. A standard technique in the literature (see, e.g., Candès and Recht [1]) is to find the least-squares solution to the equalities and then prove that it satisfies the inequalities. This is the technique we adopt herein. The conditions (3) are in sufficiently simple form that we can write down the least-squares solution in closed form; it turns out to be:

$$\boldsymbol{z}_{ij}^* = \frac{1}{\lambda|V_m|}(\boldsymbol{a}_i - \boldsymbol{a}_j) \quad \forall i, j \in V_m, i \ne j.$$

7

It follows by construction (and is easy to check) that this formula satisfies the equalities in (3), so the remaining task is to show that the norm bound $\|\boldsymbol{z}_{ij}^*\| \leq 1$ is satisfied. By definition of V_m, $\|\boldsymbol{a}_i - \boldsymbol{a}_j\| \leq 2\theta\sigma_m$. The probability that an arbitrary sample $\boldsymbol{a}_i$ is associated with mean $\boldsymbol{\mu}_m$ is w_m. Furthermore, with probability $F(\theta, d)$, this sample satisfies $\|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \leq \theta\sigma_m$, i.e., lands in V_m. Since the second choice in the mixture of Gaussians is conditionally independent from the first, the overall probability that $\boldsymbol{a}_i$ lands in V_m is $F(\theta, d)w_m$. Therefore, $E[|V_m|] = F(\theta, d)w_m n$. By the Chernoff bound for the tail of a binomial distribution, it follows that the probability that $|V_m| \geq (F(\theta, d)w_m - \epsilon)n$ is exponentially close to 1 for a fixed $\epsilon > 0$. Thus, provided $\lambda \geq 2\theta\sigma_m/((F(\theta, d)w_m - \epsilon)n)$, we have constructed a solution to (3) with probability exponentially close to 1.

For the second part of the theorem, suppose $1 \leq m < m' \leq K$. For each sample $\boldsymbol{a}_i$ associated with $\boldsymbol{\mu}_m$ satisfying $\|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \leq \theta\sigma_m^2$ (i.e., lying in V_m), the probability is $1/2$ that

$$(\boldsymbol{a}_i - \boldsymbol{\mu}_m)^T(\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m) \leq 0$$

by the fact that the spherical Gaussian distribution has mirror-image symmetry about any hyperplane through its mean. Therefore, with probability exponentially close to 1 as $n \to \infty$, we can assume that at least one $\boldsymbol{a}_i \in V_m$ satisfies the above inequality. Similarly, with probability exponentially close to 1, at least one sample $\boldsymbol{a}_{i'} \in V_{m'}$ satisfies

$$(\boldsymbol{a}_{i'} - \boldsymbol{\mu}_{m'})^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) \leq 0.$$

Then

$$\begin{aligned}
\|\boldsymbol{a}_i - \boldsymbol{a}_{i'}\|^2 &= \|\boldsymbol{a}_i - \boldsymbol{\mu}_m - \boldsymbol{a}_{i'} + \boldsymbol{\mu}_{m'} + \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2 \\
&= \|\boldsymbol{a}_i - \boldsymbol{\mu}_m - \boldsymbol{a}_i + \boldsymbol{\mu}_{m'}\|^2 + 2(\boldsymbol{a}_i - \boldsymbol{\mu}_m)^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) \\
&\quad - 2(\boldsymbol{a}_{i'} - \boldsymbol{\mu}_{m'})^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) + \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2 \\
&\geq \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2,
\end{aligned} \tag{15}$$

where, in the final line, we used the two inequalities derived earlier in this paragraph.

Consider the first-order optimality conditions for equation (1), which are given by (4). Apply the triangle inequality to the summation in (4) to obtain,

$$\|\boldsymbol{x}_i^* - \boldsymbol{a}_i\| \leq \lambda(n - 1), \text{ and} \tag{16}$$
$$\|\boldsymbol{x}_{i'}^* - \boldsymbol{a}_{i'}\| \leq \lambda(n - 1). \tag{17}$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i'}^*\| &= \|\boldsymbol{a}_i - \boldsymbol{a}_{i'} + \boldsymbol{x}_i^* - \boldsymbol{a}_i - \boldsymbol{x}_{i'}^* + \boldsymbol{a}_{i'}\| \\
&\geq \|\boldsymbol{a}_i - \boldsymbol{a}_{i'}\| - \|\boldsymbol{x}_i^* - \boldsymbol{a}_i\| - \|\boldsymbol{x}_{i'}^* - \boldsymbol{a}_{i'}\| && \text{(by the triangle inequality)} \\
&\geq \|\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m\| - 2\lambda(n - 1) && \text{(by (15), (16), and (17)).}
\end{aligned}$$

Therefore, we conclude that $\boldsymbol{x}_i^* \neq \boldsymbol{x}_{i'}^*$, i.e., that V_m and $V_{m'}$ are not in the same cluster, provided that the right-hand side of the preceding inequality is positive, i.e.,

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n - 1)}.$$

This concludes the proof of the second statement. $\qquad\square$

Clearly, there exists a λ so that the solution to (1) can simultaneously place all points in V_m into the same cluster for each $m = 1, \ldots, K$ while distinguishing the clusters provided that the right-hand side of (14) exceeds the right-hand side of (13). In order to obtain a compact inequality that guarantees this condition, let us fix some values. For example, let us take $\theta = 2d$ and let $c_d = F(2d, d)$. The Chernoff bound implies that $c_d \to 1$ exponentially fast in d. Let $w_{\min}$ be the minimum weight in the mixture of Gaussians. Let $\sigma_{\max}$ denote the maximum standard deviation in the distribution. Finally, let us take $\epsilon = c_d w_{\min}/2$. Then the above theorem states there is a λ such that with probability tending to 1 exponentially fast in n, the points in V_m, for any $m = 1, \ldots, K$ are each in the same cluster, and these clusters are distinct, provided that

$$\min_{1 \leq m < m' \leq K} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \frac{16 d \sigma_{\max}}{c_d w_{\min}}. \tag{18}$$

Compared to the Panahi et al. bound (2), we have removed the dependence of the right-hand side on n as well as the factor of K. (The dependence of the Panahi et al. bound on d is not made explicit so we cannot compare the two bounds' dependence on d. Note that there is still an implicit dependence on K in (18) since necessarily $w_{\min} \leq 1/K$.)

5   Discussion

The analysis of the mixture of Gaussians in the preceding section used only standard bounds and simple properties of the normal distribution, so it should be apparent to the reader that many extensions of this result (e.g., Gaussians with a more general covariance matrix, uniform distributions, many kinds of deterministic distributions) are possible. The key technique is Theorem 1, which essentially decouples the clusters from each other so that each can be analyzed in isolation. Such a theorem does not apply to most other clustering algorithms, or even to sum-of-norm clustering in the case of unequal weights, so obtaining similar results for other algorithms remains a challenge.

References

[1] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009.

[2] E. Chi and S. Steinerberger. Recovering trees with convex clustering. `https://arxiv.org/abs/1806.11096`, 2018.

[3] J. Chiquet, P. Gutierrez, and G. Rigaill. Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26:205–216, 2017.

[4] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer, 2012.

[5] T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: An algorithm for clustering using convex fusion penalties. In *International Conference on Machine Learning*, 2011.

[6] F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2011.

[7] A. Panahi, D. Dubhashi, F. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. *Journal of Machine Learning Research*, 70, 2017.

[8] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Convex cluster shrinkage. Available on-line at `ftp://ftp.esat.kuleuven.ac.be/sista/kpelckma/ccs_pelckmans2005.pdf`, 2005.

[9] Peter Radchenko and Gourab Mukherjee. Convex clustering via l1 fusion penalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1527–1546, 2017.

[10] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge University Press, 2014.

[11] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: model, theoretical guarantees and efficient algorithm. `https://arxiv.org/abs/1810.02677`, 2018.