PROPERTIES OF WEIGHTED SUM-OF-NORMS CLUSTERING ON SMALL DATASETS

MICHAEL THOMAS

Abstract. This paper investigates the practical application of the weighted sum-of-norms clustering method under computational constraints and limited data. We provide guidelines for parameter selection and compare the performance of our approach with the unweighted sum-of-norms method and the k-means method. We also discuss the limitations of using the mean squared error between the cluster representatives found by the algorithm and the true cluster centroids as a sole metric for evaluating the performance of a clustering algorithm.

1. Introduction and Background

Clustering is a fundamental task in unsupervised machine learning that aims to partition data points into groups based on some measure of similarity. Among various clustering methods, the sum-of-norms clustering method has gained attention due to its convex formulation and ability to adapt to complex cluster shapes [4]. The sum-of-norms clustering method aims to find a set of cluster representatives (i.e. centroids) that minimize a combination of two terms: the sum of squared distances between each data point and its assigned cluster representative, and a weighted sum of pairwise distances between cluster representatives. The first term ensures that the cluster representatives are close to their assigned data points, while the second term encourages nearby points to cluster together. A weight function in the second term determines the strength of the penalty on the pairwise distances between cluster representatives, based on the distance between the corresponding data points. Specifically, for datapoints $(x_1, \dots, x_n) \in (\mathbf{R}^d)^N$, weighted sum-of-norms minimizes the loss function

$$(y_1, \dots, y_N) \mapsto \frac{1}{N} \sum_{n=1}^{N} \|y_n - x_n\|_2^2 + \frac{\lambda}{N^2} \sum_{m,n=1}^{N} w(\|x_m - x_n\|_2)\|y_m - y_n\|_2$$

over $(y_1, \dots, y_n) \in (\mathbf{R}^d)^N$ for some clustering parameter λ and nonincreasing weight function w. Another common and equivalent formulation of this problem is given in [1].

Recently, Dunlap and Mourrat proved theoretical properties of weighted sum-of-norms clustering that incorporates a specific weight function, $w(r) = \gamma^{(d+1)}e^{(-\gamma r)}$, where γ is a parameter that can be tuned based on the number of data points. This localized approach allows for better separation of nearby clusters under certain regularity assumptions on the data distribution and cluster shapes. The authors provided theoretical results showing that the mean squared distance between the cluster representatives found by the algorithm and the true cluster

centroids can be bounded asymptotically under certain regularity conditions [3].

In particular, we assume that the data points are uniformly distributed on a set of disjoint, bounded, and effectively star-shaped open sets with boundaries sufficiently smooth that they can be locally approximated by a Lipschitz continuous function. In the context of the paper, A set is star-shaped if there exists a point in the set such that for any other point in the set, the line segment connecting these two points is entirely contained within the set; a set is effectively star-shaped if allowance for sufficiently small non-zero perturbations can always be made for the line segment. This is a relaxation of the convexity assumption and allows for a wider range of cluster shapes.

These regularity conditions collectively ensure that the data points are drawn from well-behaved clusters that are sufficiently separated and have a relatively smooth distribution within each cluster. Under these assumptions, the localized sum-of-norms clustering algorithm can be shown to recover the true cluster centroids with high accuracy, as stated in the following theorem presented by Dunlap and Mourrat:

Theorem 1.1. *Let μ be a probability measure on $\mathbb{R}^d$ such that the support of μ is the union over the closure of $\overline{U}_\ell$ with ℓ from 1 to L, where $U_1, \dots, U_L$ are bounded, effectively star-shaped open sets with Lipschitz boundaries, such that their closures are pairwise disjoint. Assume that μ admits a density with respect to the Lebesgue measure, and that this density is Lipschitz and bounded away from zero on* $\operatorname{supp}\mu$. *Then there exist $\lambda_c, C < \infty$ such that for every $\lambda \geq \lambda_c$, the following holds:*

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables distributed according to μ, $N \geq 1$ be an integer, $\mu_N := \frac{1}{N}\sum n = 1^N_{n=1}\delta_{X_n}$ be the empirical measure of the datapoints, and

$$A_N^{(\ell)} := n \in \{1, \dots, N\} \mid X_n \in U_\ell, \quad \ell \in 1, \dots, L$$

be the set of indices of datapoints in U_ℓ. For every $\gamma \geq 1$, the MSE between the clustering algorithm and the centroids of the clusters is bounded as follows:

$$\mathbb{E}\left[\frac{1}{N}\sum_{\ell=1}^{L}\sum_{n\in A_N^{(\ell)}} |u_{\mu_N,\lambda,\gamma}(X_n) - cent_\mu(U_\ell)|^2\right] \leq C\left(\gamma N^{-1/(d\vee 2)}(\log N)^{1/d'} + (1+\lambda)\gamma^{-1/3}\right) \tag{1.1}$$

Dunlap and Mourrat also present a bound on the difference between the minimizers of the original sum-of-norms clustering objective and its truncated version, where the weight function $r \mapsto e^{-\gamma r}$ is replaced by $r \mapsto e^{-\gamma r}\mathbf{1}_{r\leq\omega}$.

Theorem 1.2. *Let $\gamma, \lambda, \omega > 0$ and let μ be a probability measure on $\mathbb{R}^d$ with compact support. Let $M := \operatorname{diam}\operatorname{supp}\mu$. Then we have*

$$\int |\bar{u}_{\mu,\lambda,\gamma,\omega}(x) - u_{\mu,\lambda,\gamma}(x)|^2 d\mu(x) \leq 2M\lambda\gamma^{d+1}e^{-\gamma\omega}. \tag{1.2}$$

where $\bar{u}_{\mu,\lambda,\gamma,\omega}(x)$ is the minimizer of the truncated loss function and $u_{\mu,\lambda,\gamma}$ is the minimizer of the non-truncated loss function. Proofs for these theorems are found in [3].

This result suggests that the truncated version of the algorithm where distant points are given weight zero can provide a good approximation to the original problem. The truncation technique can lead to significant computational savings, especially when dealing with large datasets, as it reduces the number of pairwise interactions that need to be considered [5]. This makes the truncated sum-of-norms clustering algorithm more suitable for real-world applications where computational resources are limited. The bound in equation 1.2 provides a theoretical justification for using the truncated version of the algorithm and offers guidance on how to choose the truncation parameter ω based on the desired level of approximation.

While the theoretical properties of the weight function presented for the localized sum-of-norms clustering algorithm hold stead in the limiting asymptotic case, their empirical performance on smaller data sets have not been explored. Moreover, the choice of parameters, such as the weight function parameter γ and the regularization parameter λ, can significantly impact the clustering results, and guidelines for selecting these parameters are needed.

In this paper, we investigate these aforementioned empirical aspects of the localized sum-of-norms clustering method. We evaluate the algorithm's performance under limited data availability, providing insights into its scalability and robustness. We also propose guidelines for parameter selection, aiming to help practitioners choose appropriate values for γ and λ based on the characteristics of their datasets. Furthermore, we compare the performance of the localized sum-of-norms clustering with other popular clustering algorithms, such as the unweighted sum-of-norms and K-means.

Through experiments on various datasets, we demonstrate the advantages of the localized approach in separating nearby clusters and adapting to complex cluster shapes. Finally, we discuss the limitations of using the MSE between the cluster representatives found by the algorithm and the true cluster centroids as a sole metric for evaluating the performance of a clustering algorithm.

2. Methodology and Results

2.1. **Choice of γ and λ.** We first consider the problem of choosing the optimal value of the weight function parameter γ, a naive approach is to find the γ that minimizes the expected mean squared error (MSE) upper bound; we differentiate the bound with respect to γ for different values of the dimension d. The motivation behind this approach is that the γ value that minimizes the expected MSE upper bound could potentially be the optimal choice for the localized sum-of-norms clustering algorithm. However, our subsequent empirical results from

a grid search suggest that this γ value may not always be optimal in practice. For the case where $d = 1$, the derivative of the expected MSE upper bound with respect to γ is given by the equation solver sympy as:

$$\frac{\partial}{\partial\gamma}C\left(\gamma N^{-1/(d\vee 2)}(\log N)^{1/d'} + (1+\lambda)\gamma^{-1/3}\right) = C\left(\frac{\gamma}{N^{1/2}} + \frac{\lambda+1}{\gamma^{1/3}}\right) \tag{2.1}$$

Setting this derivative to zero and solving for γ yields four roots, two of which are imaginary, one negative, and one positive. The only real positive root is given by:

$$\gamma_1^* := 0.438691337650831(N^{0.5}(\lambda+1.0))^{0.75} \tag{2.2}$$

This positive root represents the γ value that minimizes the expected MSE upper bound for the case where $d = 1$ as the second derivative is positive everywhere. Similarly, we can derive the only positive roots for the cases of d = 2:

$$\gamma_2^* := 0.438691337650831\left(\frac{N^{0.5}\left(\lambda+1.0\right)}{\log\left(N\right)^{0.75}}\right)^{\frac{3}{4}} \tag{2.3}$$

and d $\geq$ 3:

$$\gamma_{\geq 3}^* := 0.438691337650831\left(N^{\frac{1}{d}}\left(\lambda+1.0\right)\log\left(N\right)^{-\frac{1}{d}}\right)^{\frac{3}{4}} \tag{2.4}$$

To investigate the optimal choice of the weight function parameter γ in practice, we then conduct a grid search over a range of values for the regularization parameter λ and γ. Our initial assumption is that the γ value obtained by solving for the root of the derivative of the expected mean squared error (MSE) upper bound (inequality 1.1) is optimal. However, we also consider alternative γ values that are scaled versions of this root, namely $\frac{\gamma_2^*}{10}$, $\frac{\gamma_2^*}{2}$, γ_2^*, and $2\gamma_2^*$.

The grid search is performed on a synthetic dataset generated according to the stochastic ball model. In this model, data points are sampled uniformly at random from the union of two disjoint balls of radius 1 with $d = 2$. For our experiments, we set the centers of the balls at $(1.05, 0)$ and $(-1.05, 0)$ to ensure that their closures are disjoint, satisfying the assumptions of Theorem 1.2. We generate a dataset consisting of 800 points using this model. For each value of λ in the grid, we evaluate the clustering performance of the localized sum-of-norms algorithm using five different γ values: the root of the derivative of inequality 1.1, and the four scaled versions mentioned above. The evaluation is based on the MSE between the cluster representatives found by the algorithm and the true cluster centroids.

The grid search results reveal that, in most cases, the best performing γ value is $\frac{\gamma_2^*}{2}$, rather than the root of the derivative of inequality 1.1, γ_2^*. This finding suggests that the theoretical analysis based on the MSE upper bound may not always provide the optimal γ value in practice even though it may provide us with a good initial guess. The discrepancy between the theoretical and empirical results highlights the importance of empirical validation and the need to consider

a range of γ values. Below we present a table of our results where the columns are different lambdas and the rows are different gamma multiples:

TABLE 1. MSE for different choices of γ and λ

λ	1	6	11	16	21	26	31	36	41	46
$2\gamma^*$	0.5031	0.0007	0.3526	0.2967	0.0775	0.3371	0.4617	0.219	0.335	0.3325
γ^*	0.5049	0.0007	0.0007	0.2569	0.0007	0.3556	0.0007	0.0007	0.3527	0.46
$0.5\gamma^*$	0.5049	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
$0.1\gamma^*$	0.5049	0.5049	0.5049	0.0007	1.1029	1.1029	1.1029	1.1029	1.1029	1.1029

It is worth noting that the grid search procedure is computationally intensive, requiring approximately 8 minutes to evaluate all five γ values for each λ in the grid, on a dataset of 800 points. This computational overhead underscores the importance of developing efficient methods for selecting the optimal γ value, especially when dealing with larger datasets.

2.2. **Choice of ω and N.** To empirically investigate the effect of truncation on the localized sum-of-norms clustering algorithm, we generate synthetic datasets using the stochastic ball model, where data points are sampled uniformly at random from the union of two disjoint balls of equal radius. In this experiment, we set the centers of the balls at $+1.01$ and -1.01 to ensure that their closures are disjoint, while being closer to each other compared to the previous experiment.

We focus on the truncation parameter ω, which determines the radius of the local ball within which the objective function is optimized. By varying the value of ω, we can explore the trade-off between computational efficiency and clustering accuracy. For each value of ω, we choose the number of data points N such that the algorithm runs in approximately 15 seconds. This allows us to assess the scalability of the algorithm under different truncation settings.

Table 2 presents the results of our experiments for different values of ω, ranging from 0.03 to none, where none implies no truncation. The table includes the number of data points N used for each ω value, the MSE between the cluster representatives found by the algorithm and the true cluster centroids and the time taken for the algorithm to converge (cluster time).

TABLE 2. MSE for different choices of ω and N

ω	none	0.4	0.2	0.13	0.03
N	772	3917	7828	11785	15719
MSE	0.001927	0.000338	0.000060	0.000341	0.498450
Runtime	18.12	18.44	14.92	20.26	11.02

The results in the table 2 demonstrate that the localized sum-of-norms clustering algorithm remains robust under truncation, allowing us to consider a large number of data points. As the truncation parameter ω decreases, the algorithm

can handle a larger number of data points N while still converging within the desired time constraint of approximately 15 seconds. The MSE values decrease until ω equals 0.2, indicating that the clustering performance initially benefits from larger N with corresponding ω, but after a certain point, the truncation trade-off for larger N becomes detrimental to the clustering performance.

This finding suggests that when dealing with large datasets, it may be more beneficial to discard a portion of the data points rather than excessively reducing the value of ω. By throwing away some data, we can effectively reduce the computational burden while maintaining a reasonable level of clustering accuracy. In contrast, setting ω to a very small value in an attempt to accommodate a large number of data points can lead to a significant loss of information and a deterioration in the algorithm's performance.

These empirical findings support the theoretical bounds on the truncated version of the algorithm, as presented in inequality 1.2. The results highlight the scalability of the localized sum-of-norms clustering algorithm under truncation, making it suitable for applications involving large datasets where computational efficiency is crucial.

2.3. **Comparison to Unweighted sum-of-norms and K-means.** When compared to other popular clustering algorithms, such as the unweighted sum-of-norms clustering and the K-means algorithm, the localized weighted sum-of-norms clustering demonstrates superior performance in handling complex cluster structures. The localized approach, with its incorporation of a weight function, allows for a more flexible and adaptive clustering that can capture intricate cluster shapes and overlapping convex hulls [2].

For each algorithm, parameters were selected to yield low MSE. For the weighted SON case, we selected λ and γ according to equation (2.2) and $\omega = 0.2$. For the unweighted SON, λ was selected to minimize MSE, which meant assigning all points to one cluster for these examples. For k-means, the number of clusters was set to 2, which is the true value.

The unweighted sum-of-norms clustering, which is a special case of the weighted version with a constant weight function, has a significant limitation in its ability to capture non-convex clusters. By definition, the unweighted sum-of-norms clustering can only produce clusters whose convex hulls are disjoint [2]. This means that if the true underlying clusters have overlapping convex hulls, the unweighted version will fail to accurately recover the cluster structure. In contrast, the localized weighted sum-of-norms clustering, by incorporating a weight function that allows for local adaptivity, can effectively handle overlapping convex hulls and recover the true cluster structure.

On the other hand, the K-means algorithm, which is one of the most widely used clustering techniques, also has its own limitations. K-means partitions the

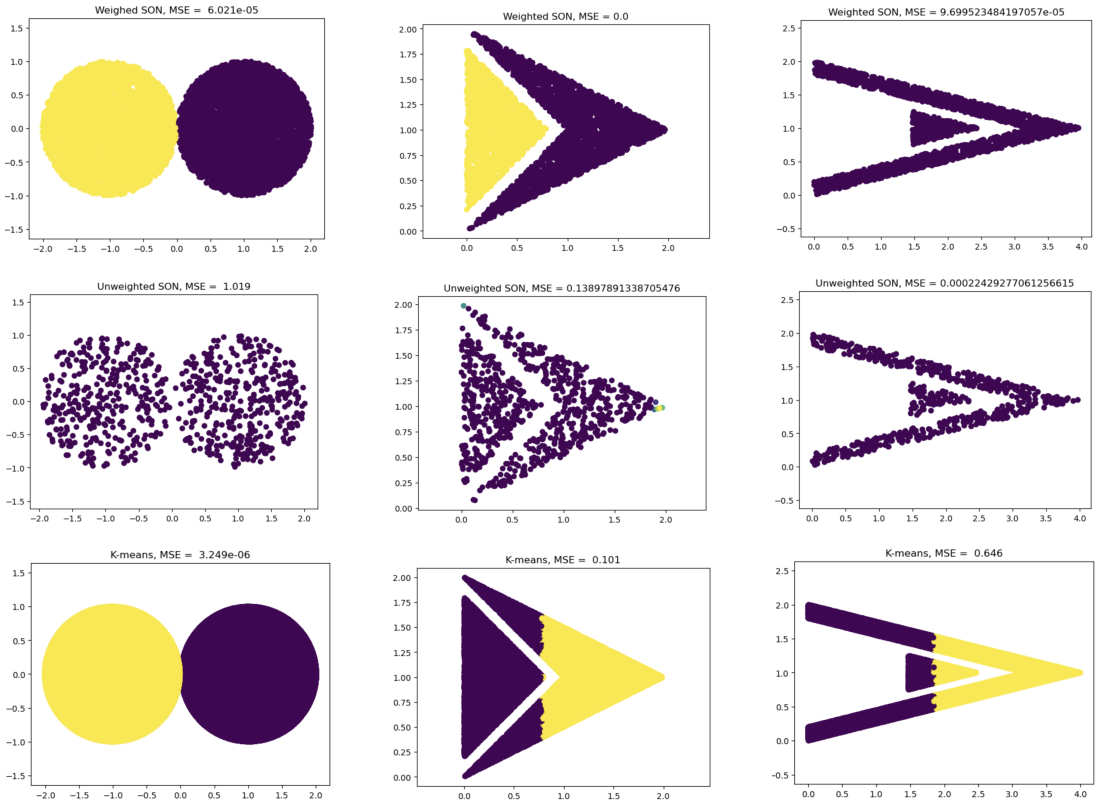

FIGURE 1. Results of weighted SON, unweighted SON and K-means algorithms on three artificial datasets. The number of datapoints used in each algorithm is maximized given computational constraints.

data space into Voronoi cells, where each data point is assigned to the cluster corresponding to the nearest centroid. While K-means can work well for clusters that have similar sizes and a spherical shape, it struggles when the clusters are anisotropically distributed or have non-spherical shape [4]. In the case of clusters with overlapping convex hulls, K-means tends to split the overlapping region and assign data points to different clusters based on their proximity to the centroids, leading to suboptimal clustering results.

The localized weighted sum-of-norms clustering, in comparison, is more robust to the presence of nearby clusters. By incorporating a weight function that takes into account the distances between data points, the localized approach can effectively separate nearby clusters and recover the true cluster structure. The weight function allows for a more nuanced assignment of data points to clusters, considering not only the proximity to the cluster representatives but also the overall cluster density and shape.

Empirical evidence supports the superiority of the localized weighted sum-of-norms clustering over the unweighted version and K-means in handling complex

cluster structures due to how it incorporates local information. Experiments conducted on synthetic datasets generated demonstrate that the localized approach consistently outperforms the other two algorithms in terms of clustering accuracy and the ability to recover overlapping clusters. These results highlight the effectiveness of the localized weighted sum-of-norms clustering in capturing the true underlying cluster structure, even in the presence of overlapping convex hulls and nearby clusters.

2.4. **Limitations of localized sum-of-norms.** Despite its ability in handling complex cluster structures and overlapping convex hulls, the localized weighted sum-of-norms clustering algorithm has some limitations that should be considered when applying it in practice. One significant limitation of the localized algorithm is its behavior when the centroids of different clusters are close to each other. In such cases, the algorithm may correctly identify the cluster centroids but fail to accurately assign data points to their respective clusters.

This limitation arises from the fact that the localized approach relies on the distances between data points and cluster representatives to determine cluster assignments. When the centroids are nearby, the distances between data points and multiple centroids may be similar, leading to ambiguity in cluster assignments. As a result, the algorithm may struggle to distinguish between data points belonging to different clusters in the overlapping region, even though it can identify the cluster centroids accurately. This limitation can have practical implications in scenarios where the true underlying clusters are not well-separated and have nearby centroids. In such cases, the localized weighted sum-of-norms clustering algorithm may provide a good estimate of the cluster centroids but may not yield satisfactory results in terms of cluster assignments.

The above limitation also leads us to another point: while the MSE between the cluster representatives found by a clustering algorithm and the true cluster centroids is commonly used as a performance metric to evaluate the algorithm's effectiveness, it is important to recognize that MSE alone may not be the best criterion for assessing the quality of the clustering results. Indeed, when the centroids of different clusters are close to each other, the algorithm may achieve a low MSE by accurately identifying the cluster centroids, even if it fails to correctly assign data points to their respective clusters. See the case of two disjoint star-shaped clusters with identical centroids in the top right of Figure 1, where the algorithm achieves a low MSE by recovering the centroids, but it fails to recover the clusters. In such cases, the low MSE can give a false impression of the algorithm's performance, as it does not capture the inaccuracies in cluster assignments. Moreover, the MSE does not take into account the complex shapes and overlapping nature of the clusters, which are crucial aspects of the clustering problem. Therefore, relying solely on MSE as a performance metric can be misleading and may not provide a comprehensive assessment of the algorithm's effectiveness in recovering the true underlying cluster structure.

Another limitation of the localized algorithm is the difficulty in selecting the optimal values for its hyperparameters, particularly the weight function parameter γ and the regularization parameter λ. The performance of the algorithm is sensitive to the choice of these parameters, and finding the right combination can be challenging. The optimal values of γ and λ may vary depending on the characteristics of the dataset, such as the number of clusters, the cluster shapes, and the level of noise. In practice, selecting the optimal hyperparameters often requires a grid search or other optimization techniques, which can be computationally expensive, especially for large datasets. Moreover, the range of values to be searched and the granularity of the search grid can have a significant impact on the quality of the resulting clustering. A poorly chosen range or an insufficiently fine-grained search may lead to suboptimal hyperparameter values and, consequently, suboptimal clustering results.

3. Conclusions and Further Research

In this paper, we have investigated the empirical application of the localized weighted sum-of-norms clustering algorithm, focusing on its performance under computational constraints and limited data availability. We have provided guidelines for selecting the key parameters, namely the weight function parameter γ and the regularization parameter λ, and have shown that the optimal choice of γ based on the theoretical MSE upper bound may not always align with the empirical results obtained through a grid search.

Our experiments have also demonstrated the robustness of the localized algorithm under truncation, where the objective function is optimized over a local ball of radius ω. We have shown that as the truncation parameter ω decreases, the algorithm can handle a larger number of data points while maintaining a reasonable level of clustering accuracy. However, we have also observed that excessively reducing ω can lead to a significant deterioration in performance, suggesting that discarding a portion of the data points may be more beneficial than severely truncating the weight function in scenarios with large datasets.

Compared to other popular clustering algorithms, such as the unweighted sum-of-norms clustering and the K-means algorithm, the localized weighted sum-of-norms clustering has demonstrated superior performance in handling complex cluster structures and overlapping convex hulls. The incorporation of a weight function allows for a more flexible and adaptive clustering that can effectively separate nearby clusters and recover the true underlying cluster structure.

Despite its strengths, the localized algorithm has some limitations that should be considered. When the centroids of different clusters are close to each other, the algorithm may struggle to accurately assign data points to their respective clusters, even though it can identify the cluster centroids correctly. Additionally, the performance of the algorithm is sensitive to the choice of hyperparameters, and finding the optimal combination can be computationally expensive.

Furthermore, we have discussed the limitations of using the MSE between the cluster representatives and the true cluster centroids as the sole performance metric for evaluating clustering algorithms. While MSE can provide insights into the accuracy of centroid recovery, it may not capture the nuances of cluster assignments and the complex shapes of the clusters.

Future research directions could include investigating the performance of the localized algorithm on real-world datasets. While our experiments have focused on synthetic datasets generated from the stochastic ball model due to the theoretical properties guaranteed by [3], it would be valuable to assess the algorithm's performance on a variety of real-world datasets with different characteristics and complexities. This would provide further insights into the algorithm's strengths and limitations in practical applications.

Alternatively, we could consider extending the localized algorithm to handle well-behaved data that is synthetically noisy and / or incomplete to attain heuristics and intuition on how it may respond to noise and missing data in the real world. By addressing these research directions, we can further enhance the understanding and applicability of the localized weighted sum-of-norms clustering algorithm, making it a more powerful and reliable tool for data analysis and pattern discovery.

References

[1] Eric C. Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

[2] Alexander Dunlap and Jean-Christophe Mourrat. Sum-of-norms clustering does not separate nearby balls. April 2021.

[3] Alexander Dunlap and Jean-Christophe Mourrat. Local versions of sum-of-norms clustering. *SIAM Journal on Mathematics of Data Science*, 4(4):1250–1271, 2022.

[4] Qiying Feng, C. L. Philip Chen, and Licheng Liu. A review of convex clustering from multiple perspectives: Models, optimizations, statistical properties, applications, and connections. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2023.

[5] Yancheng Yuan, Defeng Sun, and Kim-Chuan Toh. An efficient semismooth newton based algorithm for convex clustering. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5718–5726. PMLR, 10–15 Jul 2018.