

# Sublinear Projective Clustering with Outliers

Nina Mishra\*      Rajeev Motwani †      Sergei Vassilvitskii‡

Given a set of  $n$  points in  $\mathbb{R}^d$ , a family of shapes  $\mathcal{S}$  and a number of clusters  $k$ , the projective clustering problem is to find a collection of  $k$  shapes in  $\mathcal{S}$  such that the maximum distance from a point to its nearest shape is minimized. Some special cases of the problem include the  $k$ -line center problem where the goal is to cover the points with minimum radius hypercylinders and the  $k$ -hyperplane center problem where the goal is to cover the points with minimum width slabs.

In practice, projective clustering algorithms are often used as a dimension reduction technique to enable more effective data representation for indexing and data mining purposes on massively large datasets (See, for example, [8, 9]). In typical applications the number of points  $n$  is extremely large, the dimensionality  $d$  is large, the data possesses some outliers, while the number of clusters,  $k$  is small. Consequently, the emphasis of this paper will be on the running times of the algorithms. We present for the first time *sublinear* time randomized algorithms for the  $k$ -line and hyperplane center problems, where the running times of our algorithms are independent of  $n$ .

**Related Work** Both the  $k$ -line and  $k$ -hyperplane center problems are computationally difficult to solve in any exact or single-criteria approximation sense. Megiddo and Tamir [7] show that it is NP-hard to decide whether a set of points in the plane can be covered by  $k$  lines, i.e., cylinders with radius 0. Consequently, no constant factor approximation on the width is possible unless P=NP. In terms of approximating the number of lines, the problem is known to be APX-hard [5], and a slight generalization of the problem was recently shown to be NP-hard to approximate to a factor better than  $\log n$  [10].

Consequently, bicriteria approximations have been suggested. An  $(a, b)$ -*approximation* algorithm computes  $ak$  shapes of width  $b$  times the optimum width. Current results can be divided into two approaches. The series of papers by Har Peled, Varadarajan [6, 2] and others develop a coresets framework for this problem. These algorithms find exactly  $k$  cylinders of radius arbitrarily close to optimal. However, their running times are exponential either in the dimension,  $d$  or in the total number of clusters  $k$ . The other approach is employed by Agarwal and Procopiuc [1], who give a  $(O(d \log k), 8)$ -approximation in expected time of  $O(dnk^3 \log^4 n)$ . Note, however that the approximation guarantees are not strong enough for the algorithm to be used in a dimension reduction application, as the algorithm may return more than  $d$  cylinders.

**Our Results** Our main contribution is the development of very simple algorithms that can be easily implemented, are fast enough to be used on very large high dimensional datasets, and improve the approximation ratios of best known polynomial time algorithms. Our algorithms differ from the previous work in one important way. While previous approaches produced a complete covering of the points, the results of our algorithms will cover  $(1 - \epsilon)n$  of the points. While this is a significant

---

\*HP Labs/Stanford University, nmishra@cs.stanford.edu. Supported in part by NSF Grant EIA-0137761.

†Stanford University, rajeev@cs.stanford.edu. Supported in part by NSF Grants EIA-0137761 and ITR-0331640, and grants from Media-X and SNRC.

‡Stanford University, sergei@cs.stanford.edu. Supported in part by an NDSEG Fellowship.

Reference	Num of clusters	Width Approx	Outliers in Soln,OPT	Time
[1]	$kd \log k$	8	0, 0	$O(dnk^3 \log^4 n)$
[2]	$k$	$(1 + \alpha)$	0, 0	$O(n \log n + \frac{k^{O(dk^2)}}{\alpha^{O(dk(d+k))}})$
[6]	$k$	$(1 + \alpha)$	0, 0	$O(dn \frac{k}{\alpha^5} \log \frac{1}{\alpha})$
[6]	$k$	$(1 + \alpha)$	$\epsilon n, \epsilon n$	$O(dn \frac{k}{\alpha^5} \log \frac{1}{\alpha})$
This Work	$O(k \log(\frac{kd}{\epsilon}))$	2	$\epsilon n, 0$	$O(d \frac{k^4}{\epsilon} \log^4 \frac{kd}{\epsilon})$
This Work	$O(k \log(\frac{kd}{\epsilon}))$	2	$(\epsilon + \gamma)n, \gamma n$	$O(d \frac{k^4}{\epsilon^2(1-\gamma)^2} \log^4 \frac{kd}{\epsilon})$

Table 1:  $k$ -Line Center Results

relaxation of the problem, the benefits are dramatic. Moreover, in practice the data often contain outliers, which need not be considered in the overall clustering.

For the  $k$ -line center problem, we give an  $\tilde{O}(\frac{dk^4}{\epsilon})$  time algorithm that identifies a collection  $O(k \log \frac{kd}{\epsilon})$  cylinders of radius at most twice the optimum that cover  $(1 - \epsilon)n$  of the points. Since real data often contains outliers, we strengthen our algorithms to handle the case when the optimum solution covers only a  $(1 - \gamma)$  fraction of the points. Our algorithms with high probability find a collection of  $O(k \log \frac{kd}{\epsilon})$  cylinders that cover  $(1 - \epsilon - \gamma)n$  of the points in time  $\tilde{O}(\frac{dk^4}{\epsilon^2(1-\gamma)^2})$ . We then present a general framework in which sublinear projective clustering results may be obtained for any shape. In this framework, we prove that the  $k$   $q$ -dimensional hyperplane center problem can be solved in time  $\tilde{O}\left(d \left(\frac{kq}{\epsilon}\right)^{q+1}\right)$  where the algorithm finds a collection of  $O(k \log \frac{kdq}{\epsilon})$  slabs of width at most  $2^q$  times the optimum and cover all but an  $\epsilon$  fraction of the points.

**Techniques** Our algorithms have the following natural flavor: (1) Draw a sample of points. (2) Consider the hypercylinders of radius  $r$  whose central axis is formed by pairs of points in the sample. (3) Run the greedy set cover algorithms.

We begin by proving that it is sufficient to restrict the space of cylinders to those defined by pairs of points in the data. This fact has previously been noted in the literature [1], although we give a simpler proof that yields an improved factor 2 approximation on the width (from a factor 8).

We next demonstrate why drawing a sample and considering cylinders formed by pairs of points in the sample is sufficient for identifying an approximately good clustering. Our analysis for this algorithm draws upon the theory of  $\epsilon$ -nets and VC-dimension.

With these two tools in mind we can phrase the problem as a set cover problem and apply the greedy set cover algorithm. We show that while this already achieves a better approximation factor in polynomial time, the running time has a cubic dependence on  $od$ . We introduce a subsampling method to further speed up the greedy set cover algorithm while keeping the approximation factor the same.

The above algorithms assume that the optimum solution has no outliers and our algorithm is allowed to ignore an  $\epsilon$ -fraction of the points. We next generalize this setting to the case where the optimum solution has a  $\gamma$ -fraction of outliers and our goal is to find a clustering with at most  $(\gamma + \epsilon)$  fraction of outliers.

Our results can be extended to the case where dist is the  $L_1$  distance metric, i.e., where the goal is to identify a collection of  $k$  lines such that the maximum  $L_1$  distance from a point to its nearest line is minimized.

Our work both introduces new techniques and uses techniques already known in the community to achieve easily implementable algorithms with running time independent of  $n$ . For instance, while

the work of Alon et al [3] establishes that sublinear-sized samples are sufficient for k-line-center clustering, their algorithm requires time exponential in the sample size. Similarly, while Indyk and Har Peled demonstrated that random sampling can effectively handle outliers [4], the running time of their algorithm depends on  $n$ . Our work builds upon these known properties of random samples to simplify the projective clustering problem and then creates new techniques to obtain for the first time *sublinear time* algorithms that can be applied on large, high-dimensional datasets commonly found in data mining applications.

## References

- [1] P. Agarwal and C. Procopiuc. Approximation algorithms for projective clustering. *Journal of Algorithms*, 46:115–139, 2003.
- [2] P. Agarwal, C. Procopiuc, and K. Varadarajan. Approximation algorithms for k-line center. In *ESA*, 2002.
- [3] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. *SIAM Journal of Discrete Math*, 16, 2003.
- [4] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.
- [5] B. Broden, M. Hammar, and B. Nilsson. Guarding lines and 2-link polygons is apx-hard. In *13th Canadian Conf. on Computational Geometry*, pages 45–48, 2001.
- [6] S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. In *SOCG*, pages 312–318, 2002.
- [7] N. Megiddo and Tamir. On the complexity of locating linear facilities in the plane. *Operations Research Letters*, 1:194–197, 1982.
- [8] E. Ng, A. Fu, and R. Wong. Projective clustering by histograms. *IEEE TKDE*, 17(3):369–383, 2005.
- [9] C. Procopiuc, M. Jones, P. Agarwal, and T. Murali. A Monte Carlo algorithm for fast projective clustering. In *SIGMOD*, pages 418–427, 2002.
- [10] R. Hariharan V.S.A Kumar, S. Arya. Hardness of set covering with intersection 1. In *27th International Colloquium on Automata, Languages and Programming*, pages 624–635, 2000.