

Analysis of Incomplete Data and an Intrinsic-Dimension Helly Theorem*

Jie Gao[†] Michael Langberg[‡] Leonard J. Schulman[†]

Abstract

The analysis of incomplete data is a long-standing challenge in practical statistics. When, as is typical, data objects are represented by points in \mathbb{R}^d , incomplete data objects correspond to affine subspaces (lines or Δ -flats). With this motivation we study the problem of finding the *minimum intersection radius* $r(\mathcal{L})$ of a set of lines or Δ -flats \mathcal{L} : the least r such that there is a ball of radius r intersecting every flat in \mathcal{L} . Known algorithms for finding the minimum enclosing ball for a point set (or clustering by several balls) do not easily extend to higher-dimensional flats, primarily because “distances” between flats do not satisfy the triangle inequality. In this paper we show how to restore geometry (i.e., a substitute for the triangle inequality) to the problem, through a new analog of Helly’s theorem. This “intrinsic-dimension” Helly theorem states: for any family \mathcal{L} of Δ -dimensional convex sets in a Hilbert space, there exist $\Delta + 2$ sets $\mathcal{L}' \subseteq \mathcal{L}$ such that $r(\mathcal{L}) \leq 2r(\mathcal{L}')$. Based upon this we present an algorithm that computes a $(1 + \varepsilon)$ -core set $\mathcal{L}' \subseteq \mathcal{L}$, $|\mathcal{L}'| = O(\Delta^4/\varepsilon^2)$, such that the ball centered at a point c with radius $(1 + \varepsilon)r(\mathcal{L}')$ intersects every element of \mathcal{L} . The running time of the algorithm is $O(n^{\Delta+1}d \text{poly}(1/\varepsilon))$. For the case of lines or line segments ($\Delta = 1$), the (expected) running time of the algorithm can be improved to $O(nd \text{poly}(1/\varepsilon))$. We note that the size of the core set depends only on the dimension of the input objects and is independent of the input size n and the dimension d of the ambient space.

*Full version to appear in ACM-SIAM Symposium on Discrete Algorithms (SODA), Jan, 2006.

[†]Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794. Email: jgao@cs.sunysb.edu. Work was done when the author was with Center for the Mathematics of Information, California Institute of Technology.

[‡]Department of Computer Science, California Institute of Technology, Pasadena, CA 91125. Email: [mikel,schulman@caltech.edu](mailto:{mikel,schulman}@caltech.edu). M. Langberg is supported in part by NSF grant CCF-0346991. L. Schulman is supported in part by an NSF ITR and the Okawa Foundation.