Computing Projective Clusters via Certificates

Cecilia Procopiuc
AT&T Labs
(joint work with Pankaj Agarwal and Kasturi Varadarajan)
Applications

- Shape Fitting
- Database Indexing
- Information Retrieval
- Data Compression
- Image Processing
Example
- $S$: set of $n$ points
- $k$: integer

**$k$-Line-Center:** Find $k$ lines $l_1, \ldots, l_k$ that minimize

$$\max_{p \in S} \min_{1 \leq j \leq k} d(p, l_i).$$

$w^* = \text{minimum value so that } S \text{ can be covered by } k \text{ hyper-cylinders of diameter } w^*.$

**Projective Clustering:** Find $q$-dimensional flats $h_1, \ldots, h_k$, for some integer $q$. 
Results

1. Most variants of projective clustering problems are **NP-Hard**: Meggido and Tamir ’82.

2. $d = 2, 3, k = 1, 2$: Houle & Toussaint ’98, Agarwal & Sharir ’96, Jaromczyk & Kowaluk ’95.

3. $k = 1$, general $d$, $(1 + \varepsilon)$-approx.:
   - $q = d - 1$ (width): Duncan et al. ’97, Chan ’00.
   - $q = 1$ (enclosing cyl.): Har-Peled & Varadarajan ’01, Bădoiu et al. ’02.
   - general $q$: Har-Peled & Varadarajan ’03.

4. General $k$ and $d$:
   - $O(dk \log k)$ hyper-cylinders of diameter $8w^*$ in $\tilde{O}(d nk^3)$ time: Agarwal & Procopiuc ’00
   - $k$ hyper-cylinders of diameter $(1 + \varepsilon)w^*$ in $\tilde{O}(nf(k, d, \varepsilon))$ time: Agarwal, Procopiuc & Varadarajan ’02.
   - $k$ $q$-flats of diameter $(1 + \varepsilon)w^*$ in $dn^{O(g(k, q, \varepsilon))}$ time: Har-Peled & Varadarajan ’02.
For each flat $h$ in optimal cover, there exists small subset $Q_h$ s.t. $\text{subspace}(Q_h)$ contains $\varepsilon$-approx. flat.

$Q_h$: core-set of $h$.

$|\bigcup_h Q_h| = f(k, q, \varepsilon)$: independent of $n$ and $d$!

1. Find core-sets $Q_h$ (brute force enumeration).
2. Compute $\varepsilon$-approx. solution (brute force).
Certificates (Agarwal, Procopiuc & Varadarajan)

There exists small subset $Q$ s.t. $Q$ covered by $k$ congruent hyper-cylinders $\Rightarrow S$ covered by the $\varepsilon$-expanded hyper-cylinders.

$Q$: certificate of $S$.

$|Q| = f(k, \varepsilon, d)$: independent of $n$!

1. Find certificate $Q$ (iterative sampling).
2. Compute optimal solution on $Q$ (brute force).
3. Expand to solution on $S$. 
1-Strip Certificate

Computing Projective Clusters via Certificates
1-Strip Certificate

Computing Projective Clusters via Certificates
1-Strip Certificate
1-Strip Certificate

Computing Projective Clusters via Certificates
1-Strip Certificate

Computing Projective Clusters via Certificates
1-Strip Certificate
1-Strip Certificate
1-Strip Certificate

Computing Projective Clusters via Certificates
1-Strip Certificate
2-Strip Certificate
2-Strip Certificate
2-Strip Certificate
2-Strip Certificate

Computing Projective Clusters via Certificates
• $P$: set of points on \textit{real line}.

• $Q \subseteq P$: \textit{$k$-certificate} if \textit{any $k$ intervals} that cover $Q$ can be $\varepsilon$-expanded to cover $P$.

\textbf{Claim:} A $k$-strip certificate can be obtained from the union of $k$-certificates of all grid lines.
Line Certificate

\[ \varepsilon \Delta / 2 \]

\[ \Delta \]

\[ k = 2 \]
Line Certificate

\[ k = 2 \]
Line Certificate

Computing Projective Clusters via Certificates

\[ k = 2 \]
Lemma 1: For any set of points in $\mathbb{R}$, there exists a line certificate of size $(k/\varepsilon)^{O(k)}$.

Lemma 2: For any set of points in $\mathbb{R}^d$, there exists a certificate of size $k^{O(k)}/\varepsilon^{O(d+k)}$.

- Iterative random sampling
Open Problems

1. Certificates of smaller size?
2. Constructive proof for certificates.
3. Extensions to $q$-flats.