Shape Fitting With Outliers

Sariel Har-Peled and Yusu Wang
ACM SoCG 2003
Paper Goal, I

- Find the shape (of a given type) which approximates the best fit to a point set
  - Minimize the maximum distance from any point to the shape
  - Very sensitive to outliers
• Allow the fitting algorithm to ignore a fixed number, k, of outliers
• Find a coreset to allow efficient ($O(n)$) approximation
• Coreset construction algorithm is the paper’s major contribution
k-Extent for Hyperplanes

• Shape fitting with outliers is related to this problem

• Given a set of $n$ hyperplanes in $\mathbb{R}^d$, find the shortest “vertical” line segment which intersects all but $k$ of the hyperplanes
Coresets for the k-Extent of Hyperplanes

• Can be solved approximately using a coreset of the hyperplanes
• A good coreset for this problem can be used to find a good coreset of points for shape fitting
For each of the definitions that follow, assume the following primitives:

- \( x \) : Point in \( \mathbb{R}^{d-1} \)
- \( A \) : Surface in \( \mathbb{R}^d \), \((x,f_A(x))\) over all \( x \)
- \( B \) : Surface in \( \mathbb{R}^d \), \((x,f_B(x))\) over all \( x \)
- \( H \) : Set of \( n \) hyperplanes in \( \mathbb{R}^d \)
- \( \varepsilon \) : Constant \( > 0 \)
- \( \delta \) : Constant \( \geq 0 \)
Definition: Vertical, Above/Below

- The **vertical** direction is parallel to the $x_d$-axis in $\mathbb{R}^d$
- For any function $f(x)$ defined for $x$ in $\mathbb{R}^{d-1}$, the surface $(x,f(x))$ in $\mathbb{R}^d$ intersects any vertical line exactly once
- A is **above** B at $x$ if $f_A(x) \geq f_B(x)$
- A is **below** B at $x$ if $f_A(x) \leq f_B(x)$
Definition: Level

- The **level** of a point \((x, f(x))\) relative to \(H\) is the number of hyperplanes in \(H\) below that point.
The **k-level** of $H$ (written $L_{H,k}$) is the closure of all points in $H$ whose level is $k$

$L_{H,k}(x) = x_d$, such that $(x, x_d)$ is in $L_{H,k}$
Definition: Top k-Level

- The **top k-level** of $H$ (written $U_{H,k}$) is the closure of all points in $H$ whose level is $n-k-1$
- $U_{H,k}(x) = x_d$, such that $(x, x_d)$ is in $U_{H,k}$
Definition: \((k,r)\)-extent

- The \((k,r)\)-extent of \(H\) at \(x\) (written \(H|^{k}_{r}(x)\)) is the vertical distance between the \(r\)-level and the top \(k\)-level of \(H\).
Definition: k-extent

- The **k-extent** of H at x (written $E_{H,k}(x)$) is shorthand for the (k,k)-extent
- Notation $\Delta_{opt}(H,k)$ refers to the smallest k-extent over all x
Definition: $\delta$-sheaf

- A set of hyperplanes, $J$, is a $\delta$-sheaf if $\Delta^{\text{opt}}(J,0) \leq \delta$.
- A 0-sheaf is simply a sheaf.
Definition: \( \varepsilon \)-approximation to \( H|^{k}r \)

- An \( \varepsilon \)-approximation to \( H|^{k}r \) is a function \( E_\varepsilon : \mathbb{R}^{d-1} \rightarrow \mathbb{R} \) such that
  \[
  (1-\varepsilon)H|^{k}r(x) \leq E_\varepsilon(x) \leq H|^{k}r(x)
  \]
A \((k, \varepsilon, \delta)\)-coreset of \(H\) at \(x\) is a subset \(H'\) of \(H\) such that for any \(r<k\),

\[
L_{H,r}(x) \leq L_{H',r}(x) \leq L_{H,r}(x) + \varepsilon E_{H,r}(x) + \delta
\]

\[
U_{H,r}(x) - \varepsilon E_{H,r}(x) - \delta \leq U_{H',r}(x) \leq U_{H,r}(x)
\]

A \((k, \varepsilon)\)-coreset is one with \(\delta=0\)
Overview: Computing \((k, \varepsilon)\)-Coresets for Sets of Hyperplanes

- Start with hyperplane set \(H\)
- Select \(O(1)\) vertical segments of bounded length which collectively intersect all but \(O(k)\) hyperplanes in \(H\)
- Use segments to subdivide \(H\) into \(O(1/\varepsilon)\) disjoint subsets
- Compute \(O(k/\varepsilon^{d-2})\)-sized coreset for each subset
- Merge subset coresets to get a coreset of size \(O(k/\varepsilon^{d-1})\) for \(H\)
Lemma 3.1

Let \( J \) be a set of \( n \) hyperplanes in \( \mathbb{R}^d \).
In \( O(n) \) time, one can compute a set \( S \) of \( O(1) \) vertical segments, such that with high probability:

- All segments are no longer than \( \Delta_{\text{opt}}^{\text{opt}}(J,k) \)
- \( O(k) \) hyperplanes are not stabbed by \( S \).
Lemma 3.1 Proof: Algorithm

• Randomized, iterative algorithm
• At step $i$ of the algorithm, let:
  – $S_i = \text{Set of segments created}$
  – $Q_i = \text{Set of hyperplanes not intersected by } S_i$
• $S_1 = \text{empty set, } Q_1 = J$
Lemma 3.1 Proof: Algorithm

- **Case 1:** If $|Q_i| = O(k)$
  - then done

- **Case 2:** If $|Q_i| \leq n^{1/(3d)}$
  - compute k-level and top k-level
  - find shortest vertical segment connecting the two levels
  - add that segment to $S_{i+1}$
    - computable in $O(|Q_i|^{2d+1}) = O(n)$ time

- Otherwise...
Lemma 3.1 Proof: Algorithm

• Case 3 : If \( |Q_i| \geq n^{1/(3d)} \)
  – pick random sample, \( R_i \), from \( Q_i \) of size \( O(n^{1/(3d)} \log n) \)
  – compute the \( K_i \)-level and \( K_i \) top-level of \( R_i \), where \( K_i = \left\lceil \frac{|R_i| (k/|Q_i| + \tau)}{n} \right\rceil \) and \( \tau = 1/n^{1/6d} \)
  – find the shortest segment connecting the two levels
  – add that segment to \( S_{i+1} \)
  – computable in \( O(|Q_i|^{2d+1}) = O(n) \) time
Lemma 3.1 Proof: Bounding the Number of Case 3 Iterations

- The number of distinct hyperplane sets that can be crossed by any vertical segment is bounded by $O(n^{d+1})$
- The range space has a bounded Vapnik-Chervonenkis dimension, so the $\varepsilon$-sample theorem holds:

$$\frac{|Q_i \cap s|}{|Q_i|} - \tau \leq \frac{|R_i \cap s|}{|R_i|} \leq \frac{|Q_i \cap s|}{|Q_i|} + \tau,$$

- (for any vertical segment $s$, and $\tau = 1/n^{1/6d}$)
Lemma 3.1 Proof: Bounding the Number of Case 3 Iterations

- With high probability
  - segment created is shorter than $\Delta^{opt}(J,k)$
  - intersects at least $|Q_i|(1-4\tau)-2k$ hyperplanes of $Q_i$
    where $\tau = \frac{1}{n^{1/6d}}$
  - $|Q_{i+1}| \leq 4\tau|Q_i| + 2k$

- For $|Q_i| \geq k/\tau$
  $|Q_{i+1}| \leq 6\tau|Q_i| \leq (6/n^{1/6d})|Q_i|$

- For $|Q_i| \leq k/\tau$
  $|Q_{i+1}| \leq 6k$ (final iteration)

- Total number of iterations: $O(\log_{1/\tau} n) = O(6d) = O(1)$
• Create set of $O(1)$ vertical segments, $S$
• Divide each segments $S$ into $\lceil 4/\epsilon \rceil$ equal subsegments
• Subsegment lengths $\leq \epsilon \Delta^{opt}(H,k)/4$
• Associate each hyperplane of $H$ that intersects $S$ with exactly one of its intersecting subsegments
• Each of the $O(1/\epsilon)$ subsets is a $(\epsilon \Delta^{opt}(H,k)/4)$-sheaf
Lemma 3.4

- Let $J$ and $J'$ be hyperplane sets
- Let $J_0, \ldots, J_m$ be a partition of $J$
- Let $J'_0, \ldots, J'_m$ be a partition of $J'$
- If $J'_i$ is a $(k, \varepsilon, \delta)$-coreset for $J_i$ for $i = 0, \ldots, m$, then $J'$ is a $(k, \varepsilon, \delta)$-coreset for $J$
Lemma 3.4 Proof

- At any point $x$ in $\mathbb{R}^{d-1}$, let $i_0, \ldots, i_m$ be the number of hyperplanes in $J_0, \ldots, J_m$ below $(x, L_{J,r}(x))$
- Note that $L_{J,r}(x) = \max_j L_{J_j,r}(x)$

\[
L_{J,r}(x) \leq L_{J',r}(x) \leq \max_{j=1}^m L_{J'_j,r}(x) \\
\leq \max_{j=1}^m (L_{J_j,r}(x) + \varepsilon E_{J,j,k}(x) + \delta) \\
\leq \left( \max_{j=1}^m L_{J_j,r}(x) \right) + \varepsilon E_{J,k}(x) + \delta \\
= L_{J,r}(x) + \varepsilon E_{J,k}(x) + \delta.
\]

- Analogous argument bounds $U_{J',r}(x)$
• Each of the subsets of $H$ is a 2D $\delta$-sheaf
• Select coreset of $O(k)$ lines which approximately preserve the first $k$ levels and top levels
Lemma 3.5

• Given:
  – $J$: Set of hyperplanes in $\mathbb{R}^d$
  – $J^*$: Hyperplanes of $J$, each independently shifted down by some distance $0 \leq \delta_h \leq \delta$

• For any $x$ in $\mathbb{R}^{d-1}$ and $1 \leq r, k \leq n$:
  – $L_{J,r}(x) - \delta \leq L_{J^*,r}(x) \leq L_{J,r}(x)$
  – $U_{J,r}(x) - \delta \leq U_{J^*,r}(x) \leq U_{J,r}(x)$
  – $E_{J,k}(x) - \delta \leq E_{J^*,k}(x) \leq E_{J,k}(x) + \delta$

• $J^*$ is a $(k, 0, \delta)$-coreset of $J$
• Consider sheaf with lines parallel to $\delta$-sheaf’s
• Keep only $k$ lines with max and min slope
• Shift kept lines up by $\leq \delta$ to return to normal positions
• Result is $(k, 0, \delta)$-coreset to $\delta$-sheaf
Observation 3.6

• **Given:**
  - \( H : \) set of hyperplanes in \( \mathbb{R}^d \)
  - \( H' : (k,\varepsilon,\delta) \)-coreset for \( H \)

• \( H' \) is a \((k,\mu)\)-coreset for \( H \), where
  \[ \mu = \varepsilon + \delta / \Delta^{\text{opt}}(H,k) \]
Coresets for the 2D Case, III

- Partition creates $\delta$-sheafs, where
  \[ \delta = \left( \varepsilon \Delta^{\text{opt}}(H,k) / 4 \right) \]
- Create $(k,0,\delta)$-coresets to $\delta$-sheafs
- By observation 3.6, coresets are $(k,\varepsilon)$-coresets, since
  \[ \mu = 0 + \delta / \Delta^{\text{opt}}(H,k) = \varepsilon / 4 \]
• Algorithm uses result from 2D case
• Computes coresets of size $O(k/\varepsilon)$ in $O(n+k/\varepsilon)$ time
Theorem 3.13: Coresets for Higher Dimensions

• Given:
  – $H$: Set of $n$ hyperplanes in $\mathbb{R}^d$
  – $k, \varepsilon > 0$

• With high probability, one can compute a subset $H'$ of $H$ such that:
  – $H'$ is a ($k, \varepsilon$)-coreset of $H$
  – $|H'| = O(k/\varepsilon^{d-1})$

• Requires $O(n+k/\varepsilon^{d-1})$ time
Linearization of Multivariate Polynomials

• Let $F = \{f_1, \ldots, f_n\}$ be polynomials in $\mathbb{R}^d$

• Linearizing the polynomials
  – Replace each monomial in $F$ with a new, distinct variable, to create hyperplane set $H$
  – Variable remappings define a “linearization image” from old variables to new ones

• For example:
  – $f_i(x) = a_i x^2 + b_i x + c_i \Rightarrow h_i(u,v) = a_i u + b_i v + c_i$
  – Linearization Image: $\eta(x) = (u,v) = (x,x^2)$
  – $H$ corresponds to $F$ only where $v = u^2$
Coresets for Polynomials

- $H$ only accurately represents $F$ within a (nonconvex) subset of points, $X$
- $H$ is not directly useful for finding the minimum extent of $F$
• Within $X$, levels and extent of $F$ is faithfully represented in $H$
• Coresets for $H$ imply coresets for $F$
Theorem 4.4: Coresets for Square Roots of Polynomials

Given:
- $F = \{f_1^{1/2}, \ldots, f_n^{1/2}\}$ (the $f_i$'s are polynomials on $\mathbb{R}^d$)
- $0 \leq \varepsilon \leq 1$

Let $s$ be the number of distinct monomials present in $F$

Can compute subset $F'$ of $F$ such that:
- $F'$ is $(k, \varepsilon)$-sensitive for $F$
- $|F'| = O(k/\varepsilon^{2s})$

Requires $O(n+k/\varepsilon^{2s})$ time
Application: Min-Width Annulus

- Given a set of points in $\mathbb{R}^d$
- Find the thinnest spherical shell that contains the points
Approach: Min-Width Annulus

- Parametrize over annulus center, $c$, in $\mathbb{R}^d$
- Squared distance from $c$ to $p_i$ defines a polynomial function $f_i(c)$ over $\mathbb{R}^d$
- Define $F = \left\{ f_1^{1/2}, \ldots, f_{|P|}^{1/2} \right\}$
- Optimal center occurs at the point of min extent of $F$
Application: Min-Width Annulus
Application: Min-Width Annulus
Approach: Min-Width Annulus

- Define \( F = \{ f_1^{\frac{1}{2}}, \ldots, f_{|P|}^{\frac{1}{2}} \} \)
- Compute a \((k, \varepsilon)\)-coreset, \( F' \), for \( F \)
- Find the point of min extent for \( F' \) by brute force
- Requires \( O(n + k^{2d+1}/\varepsilon^{4d(d+1)}) \) time
• For many shape-fitting problems, we can build coresets whose size depends only on $\varepsilon$ and $k$
• Can create coresets in $O(n)$ time
• Allows shape-fitting complexity to be independent of $n$
• Limitations
  – $k$ must be very small ($O(n^{1/2d})$)
  – $k$ must be specified before processing