Approximating the Silhouette

IN PRACTICE

* $\mathcal{C} = (C_1, C_2, \ldots, C_k)$ $\quad P = C_1 \cup C_2 \cup \ldots \cup C_k$
* t = sample size (per cluster)
* $t_i = \min\{t, |C_i|\}$ $\quad 1 \leq i \leq k$

1) For each C_i $(1 \leq i \leq k)$ compute a sample $S_i \subseteq C_i$ selecting each point $x \in C_i$ with probability $t_i / |C_i|$

Obs. If $|C_i| < t \Rightarrow t_i = |C_i| \Rightarrow S_i = C_i$

Approximating the Silhouette

Only one sample S_i for each cluster C_i must be computed!

2) For each $p \in P$ (say $p \in C_i$)

$$\tilde{a}_p = \frac{1}{|C_i|} \cdot \underbrace{\frac{|C_i|}{t_i} \cdot \sum_{x \in S_i} d(p,x)}_{\tilde{d}_{sum}(p, C_i, t_i)}$$

$$= \frac{1}{\min\{t, |C_i|\}} \sum_{x \in S_i} d(p,x)$$

Approximating the Silhouette

$$\tilde{b}_p = \min_{j \neq i} \frac{1}{|C_j|} \frac{|C_j|}{t_j} \sum_{x \in S_j} d(p,x)$$

$$= \min_{j \neq i} \frac{1}{\min\{t, |C_j|\}} \sum_{x \in S_j} d(p,x)$$

$$\tilde{s}_p = \frac{\tilde{b}_p - \tilde{a}_p}{\max\{\tilde{a}_p, \tilde{b}_p\}}$$

$$\tilde{s}_Q = \frac{1}{|P|} \cdot \sum_{p \in P} \tilde{s}_p$$