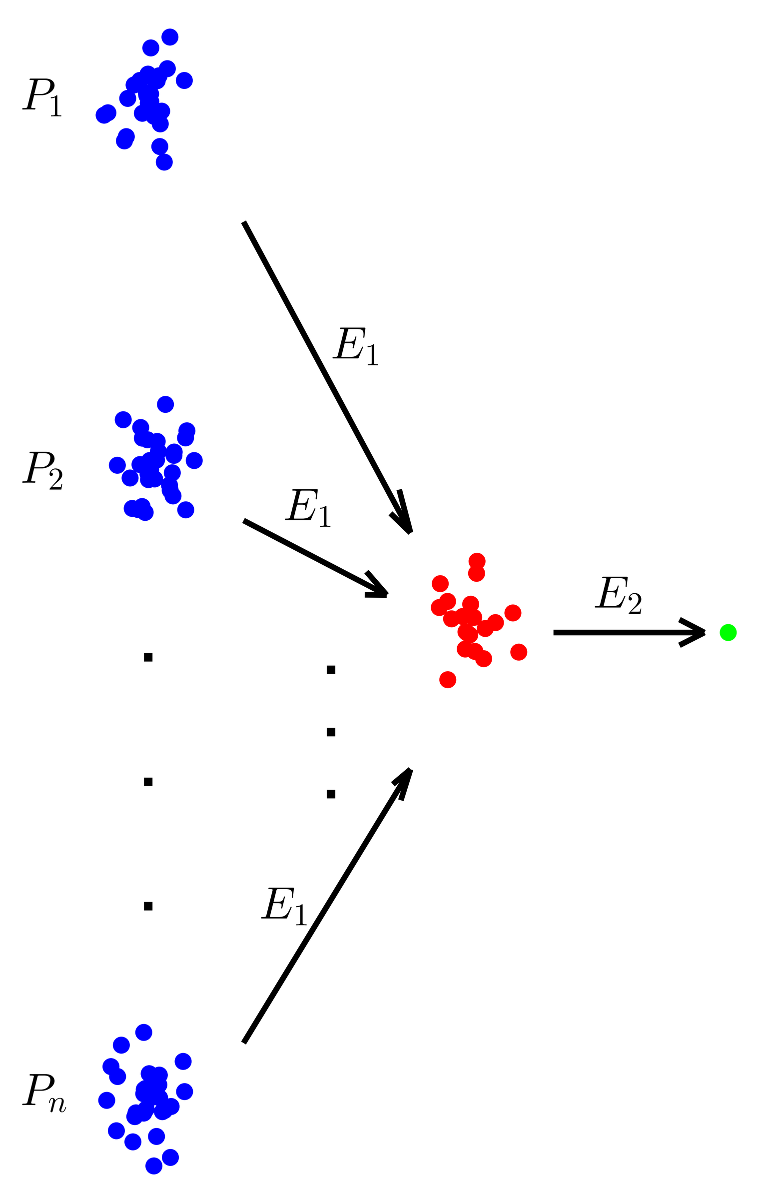
The Robustness of Estimator Composition

Pingfan Tang and Jeff M. Phillips | School of Computing, University of Utah

Motivation of Estimator composition

Composite estimators arise in many scenarios in data analysis.



- **Uncertain Data.** Considering tracking n people. Get k readings of ith person's location P_i. Estimate the location of ith person: $x_i \leftarrow E_1(P_i)$. Then summarize the entire group $E_2(x_1, x_2, \ldots, x_n)$.
- **Data Analysis Pipeline.** Estimators or analysis is performed on data at several stages, each composing estimation from prior stages. How robust is the composition of these estimators?

Definition of Breakdown Point

Informally, the **breakdown point** is the proportion of data which must be moved to infinity so that the estimator will do the same.

We define an *estimator* E as a function from the collection of some finite subsets of a metric space $(\mathscr{X}, d)$ to another metric space $(\mathscr{X}', d')$:

$$E: \mathscr{A} \subset \{X \subset \mathscr{X} \mid 0 < |X| < \infty\} \mapsto \mathscr{X}'.$$

Its *finite sample breakdown point* $g_E(n)$ (n is a positive integer) is

$$g_E(n) = \max(M) \text{ if } M \neq \emptyset \text{ and } g_E(n) = 0 \text{ if } M = \emptyset$$

with $\rho(x', X) = \max_{x \in X} d(x', x)$ and $M = \{m \in [0, n] \mid \forall X \in \mathscr{A}, |X| = n, \forall\, G_1 > 0, \exists\, G_2 = G_2(X, G_1) \text{ s.t. } \forall X' \in \mathscr{A}, \text{ if } |X'| = n \text{ and } |\{x' \in X' \mid \rho(x', X) > G_1\}| \leq m \text{ then } d'(E(X), E(X')) \leq G_2\}$.

Asymptotic Breakdown Point:

$$\beta = \lim_{n \to \infty} \frac{g_E(n)}{n}.$$

Asymptotic Onto-Breakdown Point:
Informally, the proportion of data which must be moved to change the estimator to *any value*. Technically defined:

$$\lim_{n \to \infty} \frac{f_E(n)}{n},$$

where $f_E(n) = \min(\widetilde{M})$ and $\widetilde{M} = \{0 \leq m \leq n \mid \forall X \in \mathscr{A}, |X| = n, \forall y \in \mathscr{X}', \exists X' \in \mathscr{A} \text{ s.t. } |X'| = n, |X \cap X'| = n - m, E(X') = y\}$.

Main Theorems

Definition of $E1$-$E2$ Estimators, and their Robustness
For two estimators:

$$E_1: \mathscr{A}_1 \subset \{X \subset \mathscr{X}_1 \mid 0 < |X| < \infty\} \mapsto \mathscr{X}_2,$$

$$E_2: \mathscr{A}_2 \subset \{X \subset \mathscr{X}_2 \mid 0 < |X| < \infty\} \mapsto \mathscr{X}_2',$$

suppose $P_i \in \mathscr{A}_1$, $|P_i| = k$ for $i = 1, 2, \cdots, n$ and $P_{\text{flat}} = \uplus_{i=1}^n P_i$, where $\uplus$ means the union of multisets. We define

$$E(P_{\text{flat}}) = E_2\left(E_1(P_1), E_1(P_2), \cdots, E_1(P_n)\right).$$

Theorem. *Consider estimators E_1, E_2, and E.*

- *Let β_1 be the asymptotic breakdown point **and** the asymptotic onto-breakdown point of E_1.*
- *Let β_2 be the asymptotic breakdown point of E_2.*

Then the asymptotic breakdown point of E (the E_1-E_2 estimator) is

$$\beta = \beta_1 \beta_2.$$

- **Onto Requirement.** Without the introduction of asymptotic onto-breakdown point (and a few other omitted conditions) in the above theorem, we can only obtain $\beta_1\beta_2 \leq \beta$.
- **Multi-level Composition.** Suppose $\beta_1, \beta_2, \beta_3$ and β are the asymptotic breakdown points of E_1, E_2, E_3 and E_1-E_2-E_3 respectively. If E_1, E_2 and E_3 satisfies some similar conditions as above, then $\beta = \beta_1\beta_2\beta_3$.

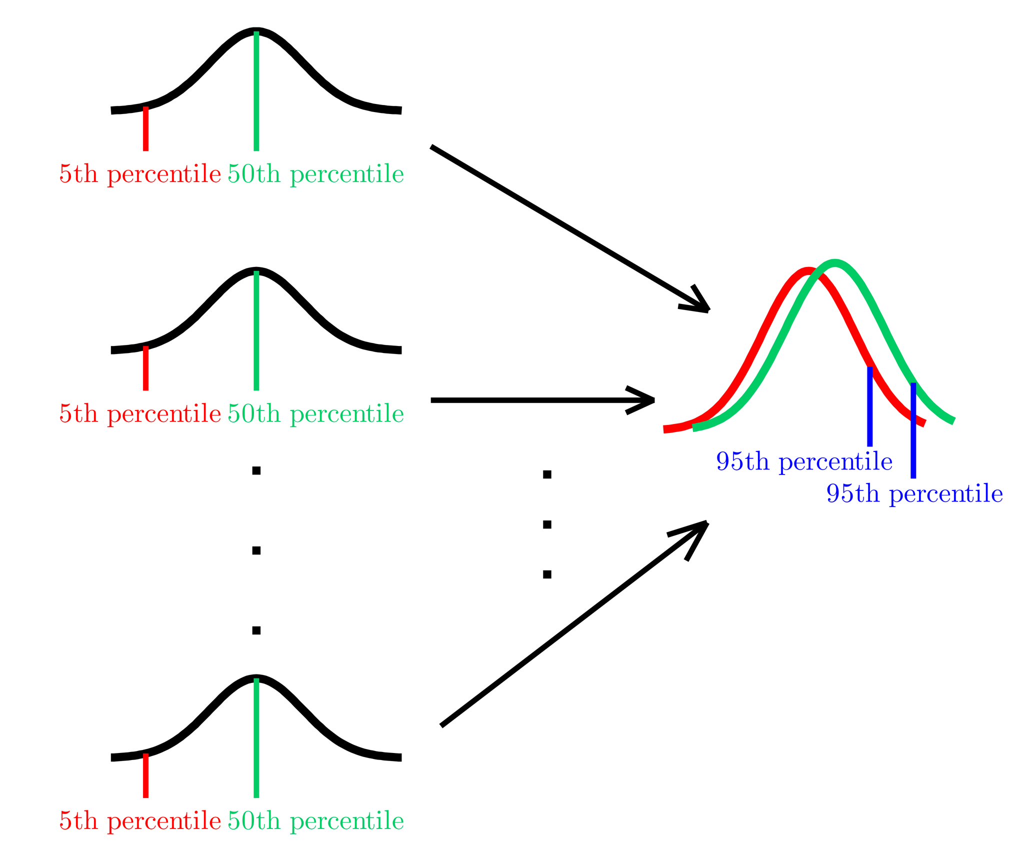
Composing Quantiles

What happens without the *onto* condition on β_1?

- E_1: 0.25 quantile, asymptotic breakdown point $\beta_1 = 0.25$,
  asymptotic onto-breakdown point$= 0.75$.
- E_2: 0.75 quantile, asymptotic breakdown point $\beta_2 = 0.25$,
  asymptotic onto-breakdown point$= 0.75$.
- $E = E_1$-E_2: $E_2(E_1(P_1), E_1(P_2), \cdots, E_1(P_n))$,
  $|P_1| = |P_2| = \cdots = |P_n|$.
- $\beta_1\beta_2 = 0.25 \cdot 0.25 = 0.0625$.
- Breakdown point of E is $\beta = 0.75 \cdot 0.25 = 0.1875$.
- For E_1, asymptotic breakdown point is not equal its asymptotic onto-breakdown point, so we only have $\beta_1\beta_2 < \beta$.

Application: Significance Thresholds

In hypothesis testing, we desire the 0.05 significance level.



- $E(P_{\text{flat}}) = E_2(E_1(P_1), E_1(P_2), \cdots, E_1(P_n))$, $P_{\text{flat}} = \uplus_{i=1}^n P_i$
  E_2: 95th percentile, from hypothesis testing.
- E_1: 50th percentile, the the breakdown point of E_1-E_2 estimator is $0.5 \cdot 0.05 = 0.025$.
- E_1: 5th percentile, the the breakdown point of E_1-E_2 estimator is $0.95 \cdot 0.05 = 0.0475$.

Preprocessing initial data (the E_1 estimator) with other quantile (e.g., 5th percentile) is more stable than the median. May introduce bias.

Application: L_1-Median of L_1-Medians

The L_1-median (point which minimizes sum of distance to data set) has asympotitic onto breakdown point of 0.5.

- E_1: L_1-median ($\beta_1 = 0.5$), E_2: L_1-median ($\beta_2 = 0.5$).
  $E = E_1$-E_2 estimator has $\beta = \beta_1\beta_2 = 0.25$.
- Consider $n = 5$ sets of $k = 8$ points each. Given an target point p_0, we only need to modify $\lceil \frac{n}{2} \rceil \lceil \frac{k}{2} \rceil$ points (in this case 12 points) so the estimator is equal to p_0.