Interval data and sample variance

Ondřej Sokol Elena Kuchina

University of Economics, Prague

ROBUST 2016

Interval data and statistics

- One-dimensional dataset of exact values is unobservable.
- Observable is a collection of intervals.
- There is no other information about data but the lower and upper bound.
- Under these weak assumptions, the only information we can infer about statistics from the observable data is the lower and upper bound.
- In this work we deal with the upper bound of sample variance.

Intervals

Interval

Given a *center* $x^c \in \mathbb{R}$ and a *radius* $x^\Delta \in \mathbb{R}^+$, the *interval* $\boldsymbol{x}$ is the set $\{\xi : x^c - x^\Delta \le \xi \le x^c + x^\Delta\}$.

Iterval with lower bound $\underline{x}$ and upper bound $\overline{x}$ will be written as $[\underline{x}, \overline{x}]$.

Narrowed interval

Given a center x^c and radius x^Δ of interval $\boldsymbol{x}$ and positive real α, the α-*narrowed* interval $\boldsymbol{x}$, denoted $\boldsymbol{x}^\alpha$ is interval with center x^c and radius αx^Δ, i.e. $[x^c - \alpha x^\Delta, x^c + \alpha x^\Delta]$.

In this talk, we need only $\alpha \le 1$ – this explains the term "narrowing".

Problem formulation

Our problem

Input: intervals $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, given as centers $x_1^c, \dots, x_n^c$ and radii $x_1^\Delta, \dots, x_n^\Delta$.

Output: minimal and maximal variance among samples of crisp values $(x_1, \dots, x_n)$ chosen from $\boldsymbol{x}_1 \times \cdots \times \boldsymbol{x}_n$.

It consists of solving

$$\begin{aligned} \text{optimize} \quad & \sigma^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \\ \text{subject to} \quad & x_i \in \boldsymbol{x}_i \quad \text{for } i = 1, \dots, n. \end{aligned}$$

Properties of our problem

- Sample variance σ^2 is convex in x_i, the set of all x_i is a convex set.
- The lower bound of sample variance over interval data can be found in polynomial time.
- Computation of the upper bound of sample variance over interval data is known to be **NP-hard problem**.
- We studied the behaviour of specialized algorithms (by Ferson (2005) and Xiang (2007)) on "common" randomly generated instances of this problem, exploiting their polynomial behaviour on "good" instances.
- Experiments show that **random instances are usually solvable in reasonable time**.

Complexity of Ferson's algorithm

We focus on behaviour of Ferson's algorithm.

- If the **the $\frac{1}{n}$-narrowed intervals do not intersect**, the algorithm computes $\overline{\sigma^2}$ in **quadratic time** in n.
- If the $\frac{1}{n}$-narrowed intervals have a common point, then the computational complexity of the algorithm is $O(2^k n^2)$, where k is the maximal number of the narrowed intervals that have at least one common point.
- Formally, define $G_n = (\{1, \ldots, n\}, E)$, where

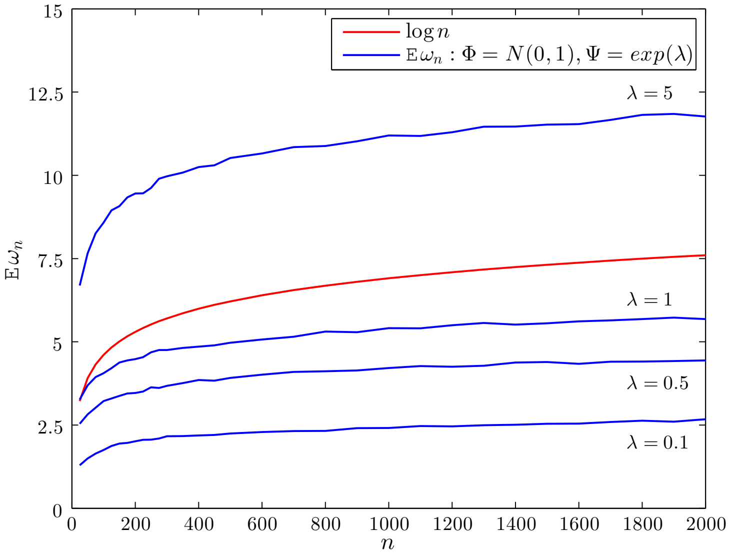
$$E := \{\{i, j\} : \boldsymbol{x}_i^{\frac{1}{n}} \cap \boldsymbol{x}_j^{\frac{1}{n}} \neq \emptyset\}.$$

- Let ω_n be the size of the largest clique in G_n, then $k = \omega_n$.
- The instances with **"small"** k **are of interest**. But how frequent are these instances?

Experiments

- The experiments tested the size of $k (= \omega_n)$ on randomly generated intervals.
  - Denote by Φ the distribution of centers of the intervals.
  - Denote by Ψ the (nonnegative) distribution of radii of the intervals.
  - The samples were independent.
- The experiments suggest that if the intervals come from *reasonable* distributions Φ and Ψ, **the size of the largest clique of the average case can be approximated by function of** $\log n$.

Results for $\Phi = N(0,1)$ and various λ of $\Psi = \text{Exp}(\lambda)$



Conjecture

The conclusions of the experiments is formalized in the following:

Conjecture

If Φ is a continuous distribution with finite first and second moments and its density function is limited from above and Ψ has finite first and second moments, then $\mathrm{E}\omega_n = O(\log n)$ and $Var(\omega_n) = O(1)$.

Our goal

- If the conjecture is true, then ***the algorithm runs in polynomial time on the random data***.
- Our goal is **to decide the conjecture**.
- It appears to be hard in its full generality. We restrict ourselves to the following stochastic setup:
  - centers are uniformly distributed on $[0, 1]$,
  - radii are constant and equal to 1.

Idea

- We subdivide the whole domain $[0, 1]$ by $\lfloor \frac{n}{2} \rfloor + 1$ equidistant points $t_0, \ldots, t_{\lfloor \frac{n}{2} \rfloor}$.
- In every such point, say point t, we express the distribution of the (random) number $A_n(t)$ of $\frac{1}{n}$-narrowed intervals containing t.
- It is sufficient to compute $\max_i A_n(t_i)$. Unfortunately, random variables $A_n(t_i)$ are not independent, however, they have negative covariance vanishing with $n \to \infty$.
- Now, it is sufficient to overcome the dependency – we suggest to approximate $A_n(t_i)$ with (independent) Poisson variables, however, we are not able to do this yet.

Transformation

- Define the indicator variable

$$Z_i^{\frac{1}{n}}(t) = \begin{cases} 1, & \text{if } t \in \boldsymbol{x}_i^{\frac{1}{n}}, \\ 0, & \text{otherwise.} \end{cases}$$

- Let $A_n(t)$ denote the number of $\frac{1}{n}$-narrowed intervals intersecting t.
- As $Z_i^{\frac{1}{n}}(t)$ for $i = 1, \dots, n$ has alternative distribution and $Z_i^{\frac{1}{n}}(t)$ and $Z_j^{\frac{1}{n}}(t)$ are independent for $i \neq j$, then $A_n(t)$ has binomial distribution $Bi(n, \frac{2}{n})$ as $A_n(t) = \sum_{i=1}^{n} Z_i^{\frac{1}{n}}(t)$.
- Now, $A_n(t)$ has approximately Poisson distribution with parameter 2.

- Let choose $\lfloor \frac{n}{2} \rfloor + 1$ equidistant points on interval $[0, 1]$.

Lemma

For every k such that $i < k < j$ and $|i - j| \leq \frac{2}{n}$ it follows that $A_n(k) \leq A_n(i) + A_n(j)$.

- With this placement of points, the covariance of $A_n(t)$ and $A_n(s)$ for $t \neq s$ is diminishing with $n \to \infty$ as $cov(A_n(t), A_n(s)) = -\frac{4}{n}$.
- Now, we need to compute the distribution of the maximum of $\lfloor \frac{n}{2} \rfloor + 1$ correlated variables with identical binomial (Poisson?) distribution.

Lemma (Kimber (1983))

Let $X_n(j) \sim Pois(\lambda)$, are independent for $j = 1, \ldots, n$, $\lambda > 0$ and $M = \max(X(j) : j \in \{1, \ldots, n\})$. Then for $n \to \infty$, $M \approx \log n / \log \log n$.

Summary

- We deal with computation **of maximal variance over interval data** – an **NP-hard problem** in general.
- Computational experiments suggest that **Ferson's algorithm** runs in **polynomial time** for most instances.
- We propose an approache to provide theoretical reasoning for what we observed empirically.
- However, some open "hard" steps remain unresolved.

Thank you for your attention.

FERSON, Scott; GINZBURG, Lev; KREINOVICH, Vladik; AVILES, Monica. Exact Bounds on Sample Variance of Interval Data. *Reliable Computing*. 2005, Vol. 11, No. 3, pp. 207–233.

KIMBER, Alan C. A note on Poisson maxima. *Probability Theory and Related Fields*. 1983, Vol. 63, No. 4, pp. 551–552. ISSN 0044-3719, 1432–2064.