Estimating Risk under Interval Uncertainty: Sequential and Parallel Algorithms

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
vladik@utep.edu

Hung T. Nguyen
Department of Mathematical Sciences
New Mexico State University

Songsak Sriboonchita
Faculty of Economics, Chiang Mai University

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

1. Computing statistics is important

- *Problem:* estimating the quality of of an individual investment – and of the investment portfolio.
- *Traditional econometrics approach:* use expected return and its risk (variance).
- *How to estimate these characteristics:*
  - trace the past returns $x_1, \ldots, x_n$ of a given (and/or similar) investment;
  - compute the statistical characteristics based on these returns.
- *The expected return:* $E = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i.$
- *The risk:* $V = \dfrac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2.$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

2. Additional problem: interval uncertainty

- The return (per unit investment) is defined as
  - the selling price of the corresponding financial instrument at the end of, e.g., a one-year period,
  - divided by the buying price of this instrument at the beginning of this period.
- It is usually assumed that we know the exact values $x_1, \ldots, x_n$ of the returns.
- In practice, however, both the selling and the buying prices unpredictably fluctuate within a single day.
- These minute-by-minute fluctuations are not always recorded.
- What we usually have recorded is the daily range of prices $[\underline{x}_i, \overline{x}_i]$.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

3. Traditional approach to solving the problem of interval uncertainty

- *Traditional approach:*
  - take the average $\widetilde{x}_i = \dfrac{\underline{x}_i + \overline{x}_i}{2}$ and
  - compute the characteristics based on these averages.
- *Resulting estimate for the expected return:*

$$\widetilde{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \widetilde{x}_i,$$

- *Resulting estimate for the risk:*

$$\widetilde{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\widetilde{x}_i - \widetilde{E})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\widetilde{x}_i)^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \widetilde{x}_i \right)^2.$$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

4. Traditional approach: limitations

- *In the bull market:*
  - there may be dips leading to a small value of x_i,
  - but overall, the values are increasing and
  - therefore, $\overline{x}_i$ is a reasonable estimate for x_i, and $\widetilde{x}_i$ underestimates the high price x_i.
- *In the bear market:*
  - spikes are accidental but lower values are typical,
  - therefore, $\underline{x}_i$ is a reasonable estimate for x_i, and $\widetilde{x}_i$ overestimates the low price x_i.
- So, we underestimate the low prices and underestimate the high prices.
- Thus we underestimate the variance (the measure of price variation).

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

5. Estimating statistics under interval uncertainty: a computational problem

- *Traditional assumption:* we know the true values $x_1, \ldots, x_n$.
- *Traditional computations:* estimate the value of a statistical characteristic $C(x_1, \ldots, x_n)$.
- *Interval uncertainty:* we only know the intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$ that contain x_i.
- *Fact:* different values $x_i \in \mathbf{x}_i$ lead, in general, to different values of $C(x_1, \ldots, x_n)$.
- *Conclusion:* we need to estimate the range

  $$C(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- *Computational challenge:* modify the existing statistical algorithms so that they compute these ranges.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

6. Estimating expected return under interval uncertainty

- *Fact:* the expected return (arithmetic average) E is a monotonically increasing function of $x_1, \ldots, x_n$.
- *Conclusions:*
  - the smallest possible value $\underline{E}$ is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$);
  - the largest possible value is attained when $x_i = \overline{x}_i$ for all i.
- In other words, the range $\mathbf{E}$ of E is equal to

$$[E(\underline{x}_1, \ldots, \underline{x}_n), E(\overline{x}_1, \ldots, \overline{x}_n)].$$

- In other words, $\underline{E} = \frac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E} = \frac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)$.

Computing statistics is ...
Additional problem: ...
Traditional approach ...
Estimating statistics ...
Linearized techniques
The exact estimation ...
Need for parallelization
Acknowledgments

7. Linearized techniques

- *Idea:* when the daily fluctuations are small, we can use the linearization techniques:
  - we represent the values x_i as $x_i = \widetilde{x}_i + \Delta x_i$, where the differences $\Delta x_i \stackrel{\text{def}}{=} x_i - \widetilde{x}_i$ are small, and
  - we ignore quadratic terms in the formula for the variance.
- *Details:* the condition that $x_i \in [\underline{x}_i, \overline{x}_i]$ means that $\Delta x_i \in [-\Delta_i, \Delta_i]$, where $\Delta_i \stackrel{\text{def}}{=} \dfrac{\overline{x}_i - \underline{x}_i}{2}$.
- *General case:*

$$C(x_1, \ldots, x_n) \approx C(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum_{i=1}^{n} \frac{\partial C}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n) \cdot \Delta x_i.$$

- *Case study:* the variance $V = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i\right)^2.$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

8. Linearization (cont-d)

- *Formula:* $V = \widetilde{V} + 2\sum\limits_{i=1}^{n} (\widetilde{x}_i - \widetilde{E}) \cdot \Delta x_i$.
- The expression for V is monotonic in each $\Delta x_i \in [-\Delta_i, \Delta_i]$:
  - it is increasing when $\widetilde{x}_i \ge \widetilde{E}$ and
  - it is decreasing when $\widetilde{x}_i \le \widetilde{E}$.
- *When $\widetilde{x}_i \ge \widetilde{E}$:* maximum is attained when $\Delta x_i = \Delta_i$; the corresponding term in V is $(\widetilde{x}_i - \widetilde{E}) \cdot \Delta_i$.
- *When $\widetilde{x}_i \le \widetilde{E}$:* maximum is attained when $\Delta x_i = -\Delta_i$; the corresponding term in V is $-(\widetilde{x}_i - \widetilde{E}) \cdot \Delta_i$.
- *General expression:* $|\widetilde{x}_i - \widetilde{E}| \cdot \Delta_i$.
- *Conclusion:* the range of V is $[\widetilde{V} - 2\Delta, \widetilde{V} + 2\Delta]$, where

$$\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{n} |\widetilde{x}_i - \widetilde{E}| \cdot \Delta_i.$$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

9. Linearization approximation is not always adequate

- In finance, the gain is often obtained by a small (often $< 1\%$) advantage.
- From this viewpoint, it is desirable to have estimates which are as accurate as possible.
- When the situation is stable, the daily fluctuations are low, and quadratic terms can be reasonable ignored.
- However, the whole purpose of estimating risk is to cover situations with high volatility.
- In such situations, the daily fluctuations $\overline{x}_i - \underline{x}_i = 2\Delta_i$ can also be sizeable.
- Thus, terms quadratic in Δ_i cannot be ignored if we want accurate estimates.
- In such situations, we need the *exact* range of the variance (risk) V.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

10. The exact estimation of risk under interval uncertainty is, in general, an NP-hard problem

- *Computational problem* (reminder):
  - *given:* interval data $x_i \in [\underline{x}_i, \overline{x}_i]$;
  - *compute:* the exact range $\mathbf{V} = [\underline{V}, \overline{V}]$ for the risk (variance) V.
- *Fact:* this problem is, in general, computationally difficult (NP-hard).
- *Specifically:*
  - there is a $O(n \cdot \log(n))$ time algorithm for computing $\underline{V}$, but
  - computing $\overline{V}$ is, in general, NP-hard.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

11. Sequential algorithm for computing $\overline{V}$ in the no-proper-subset case

- *Good news:* in many practical situations, there are efficient algorithms for computing $\overline{V}$.
- *Auxiliary notion:* "narrowed" intervals are defined as

$$[x_i^-, x_i^+] \stackrel{\text{def}}{=} \left[\widetilde{x}_i - \frac{\Delta_i}{n}, \widetilde{x}_i + \frac{\Delta_i}{n}\right].$$

- *Example when an efficient algorithm exists:* when no two are proper subsets of one another, i.e.,

$$[x_i^-, x_i^+] \not\subseteq (x_j^-, x_j^+) \text{ for all } i \text{ and } j.$$

- *In this case:* there exists a $O(n \cdot \log(n))$ time algorithm.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

12. Algorithm: general structure

1. First, we sort the values $\widetilde{x}_i$ into an increasing sequence:

$$\widetilde{x}_1 \le \widetilde{x}_2 \le \ldots \le \widetilde{x}_n.$$

2. Then, for every k from 0 to n, we compute the value $V^{(k)} = M^{(k)} - (E^{(k)})^2$ of the variance V for

$$x^{(k)} = (\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n).$$

3. Finally, we compute $\overline{V}$ as the largest of $n+1$ values

$$V^{(0)}, \ldots, V^{(n)}.$$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

13. Algorithm: details of Stage 2

- *Main idea:* use previous values of $M^{(k)}$ and $E^{(k)}$ to compute the next values $M^{(k+1)}$ and $E^{(k+1)}$.

- *First:* compute $M^{(0)} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\overline{x}_i)^2$, $E^{(0)} = \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i$,

  and

  $$V^{(0)} = M^{(0)} - (E^{(0)})^2.$$

- *Then:* once we know the values $M^{(k)}$ and $E^{(k)}$, we compute

$$M^{(k+1)} = M^{(k)} + \frac{1}{n} \cdot (\underline{x}_{k+1})^2 - \frac{1}{n} \cdot (\overline{x}_{k+1})^2;$$

$$E^{(k+1)} = E^{(k)} + \frac{1}{n} \cdot \underline{x}_{k+1} - \frac{1}{n} \cdot \overline{x}_{k+1}; \text{ and}$$

$$V^{(k+1)} = M^{(k+1)} - (E^{(k+1)})^2.$$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

14. Sequential algorithm: number of computation steps

- Sorting requires $O(n \cdot \log(n))$ steps.
- Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V^{(0)}$ requires linear time $O(n)$.
- For each $k = 0, \ldots, n-1$, we need a constant number of steps to compute the next values

$$M^{(k+1)}, E^{(k+1)}, \text{ and } V^{(k+1)}.$$

- Finally, finding the largest of $n+1$ values $V^{(k)}$ also requires $O(n)$ steps.
- Thus, overall, we need

$$O(n \cdot \log(n)) + O(n) + O(n) + O(n) = O(n \cdot \log(n))$$

steps.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

15. Comment about the possibility of linear-time algorithms

- In the $O(n \cdot \log(n))$ algorithm, the main computation time is used on *sorting*.
- It is possible to avoid sorting and use instead the known fact that we can compute the *median* in linear time.
- *Asymptotically:* the linear time algorithm for computing the median is faster than sorting.
- *In practice:*
  - the median computing algorithm is still rather complex
  - so, for reasonable size n, sorting is faster than computing the median.
- Thus, sorting-based algorithms are actually faster than median-based ones.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

16. Need for parallelization

- *Traditional algorithms* for computing the variance V from the exact values $x_1, \ldots, x_n$ take linear time $O(n)$.
- *Interval uncertainty:* we need a larger amount of computation time – e.g., time $O(n \cdot \log(n))$.
- *In financial applications:* it is often very important to produce the result as fast as possible.
- One way to speed up computations is to perform these algorithms *in parallel* on several processors.
- Let us we show how the algorithms for estimating variance under interval uncertainty can be parallelized.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

17. Possibility of parallelization

- *Reminder:* for large n,
  - we may want to further speed up computations
  - if we have several processors working in parallel.
- In the general case, all the stages of the above algorithm can be parallelized by known techniques.
- In particular, the computation of $M^{(k)}$, $E^{(k)}$ on Stage 2 is a particular case of a general *prefix-sum* problem:
  - we must compute the values

$$a_1, \quad a_1 * a_2, \quad a_1 * a_2 * a_3, \ldots,$$

  - for some associative operation $*$.
- In our case, $* = +$.

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

18. Case of potentially unlimited number of processors

- *Case:* we have a potentially unlimited number of processors.
- *Stage 1:* we can sort the values $\widetilde{x}_i$ in time $O(\log(n))$.
- *Stage 2:* we can compute the values $V^{(k)}$ (i.e., solve the prefix-sum problem) in time $O(\log(n))$.
- *Stage 3:* we can compute the maximum of $V^{(k)}$ in time $O(\log(n))$.
- *As a result:* we can compute $\overline{V}$ time

$$O(\log(n)) + O(\log(n)) + O(\log(n)) = O(\log(n)).$$

Computing statistics is . . .
Additional problem: . . .
Traditional approach . . .
Estimating statistics . . .
Linearized techniques
The exact estimation . . .
Need for parallelization
Acknowledgments

19. Case when we have $p < n$ processors

- *Stage 1:* sort n values in time
$$O\left(\frac{n\cdot \log(n)}{p}+\log(n)\right).$$
- *Stage 2:* compute the values $V^{(k)}$ in time
$$O\left(\frac{n}{p}+\log(p)\right).$$
- *Stage 3:* compute the maximum of $V^{(i)}$ in time
$$O\left(\frac{n}{p}+\log(p)\right).$$
- *Overall:* we thus need time
$$O\left(\frac{n\cdot \log(n)}{p}+\log(n)\right)+O\left(\frac{n}{p}+\log(p)\right)+O\left(\frac{n}{p}+\log(p)\right)=$$
$$O\left(\frac{n\cdot \log(n)}{p}+\log(n)+\log(p)\right).$$

20. Acknowledgments

This work was supported in part:

- by NSF grant HRD-0734825 and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.

Computing statistics is ...
Additional problem: ...
Traditional approach ...
Estimating statistics ...
Linearized techniques
The exact estimation ...
Need for parallelization
Acknowledgments