Some applications of interval computation in statistics

Michal Černý

University of Economics in Prague, Faculty of Computer Science and Statistics,
Winston Churchill Square 4, 13067 Prague, Czech Republic

cernym@vse.cz

Keywords: interval-valued data, interval linear regression, interval computation

Introduction

One of the main goals of interval analysis is to determine the range of a given continuous function over a given (multidimensional) interval. This talk is devoted to some particular functions important in statistics.

One-dimensional data

First we consider the case of one-dimensional data. We assume that there is a dataset $x = (x_1, \ldots, x_n)$ (a random sample from some distribution, say) and a continuous function (statistic) $S(x)$. The dataset x is unobservable; what is observable is a collection of intervals $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ such that we are guaranteed that $x \in \boldsymbol{x}$ a.s. If we do not make any stronger assumptions on the distribution of $(x, \boldsymbol{x})$, then the maximum information we can infer about $S(x)$ from the observable data $\boldsymbol{x}$ is the pair of bounds $\underline{S} = \min\{S(\xi) : \xi \in \boldsymbol{x}\}$ and $\overline{S} = \max\{S(\xi) : \xi \in \boldsymbol{x}\}$.

Only a few statistics can be evaluated by the interval arithmetic, such as the sample mean or variance $n^{-1} \sum_i (x_i - \mu)^2$, when the true mean μ is known. More often, the arithmetical expressions suffer from the dependency problem.

One of the best understood statistics is the sample variance $\sigma^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - n^{-1}\sum_{j=1}^n x_j)^2$. It is directly seen that $\underline{\sigma^2}$ can be computed in weakly polynomial time; however, there even exists a strongly polynomial method. On the other hand, computation of $\overline{\sigma^2}$ is NP-hard and inapproximable with an arbitrary absolute error. It is an open problem whether it is efficiently approximable with some "reasonable" relative error. A good news is that $\overline{\sigma^2}$ can be computed in pseudopolynomial time. Furthermore, many special cases solvable in polynomial time are known. We will study the algorithm by Ferson et al. [5], which works in time $O(2^{\omega_n} n^2)$, where ω_n is the size of the largest clique of the undirected graph $G(V,E)$ with $V = \{1,\ldots,n\}$ and $\{i,j\} \in E$ iff $[\boldsymbol{x}_i^C \pm n^{-1}\boldsymbol{x}_i^\Delta] \cap [\boldsymbol{x}_j^C \pm n^{-1}\boldsymbol{x}_j^\Delta] \neq \emptyset$. In general, ω_n can be large, but in many reasonable and natural stochastic setups it seems that $\omega_n = O(\log n)$ on average, making the algorithm practically very useful. Moreover, it seems that $\mathrm{var}(\omega_n) = O(1)$, showing that hard instances occur very rarely.

We will deal with other statistics of one-dimensional data, such as higher moments or the coefficient of variation, from a similar perspective. We will also mention statistics important in testing hypotheses. We will also deal with simultaneous regions for dependent statistics, such as the joint region $\{(n^{-1}\sum_{i=1}^n \xi_i,\ (n-1)^{-1}\sum_{i=1}^n [\xi_i - n^{-1}\sum_{j=1}^n \xi_j]^2) \in \mathbb{R}^2 : \xi \in \boldsymbol{x}\}$ for sample mean and variance.

Linear regression

In the multivariate setup we discuss the linear regression model $y = X\beta + \varepsilon$, where the data (X, y) are unobservable and we can observe only intervals $\boldsymbol{X}, \boldsymbol{y}$ such that $X \in \boldsymbol{X}$ and $y \in \boldsymbol{y}$ a.s. Here, the most important statistics are estimators of the regression coefficients β and goodness-of-fit measures. We will study minimum-norm estimators based on L_p-norms and their associated loss functions, such as Ordinary Least Squares, Generalized Least Squares, Least Absolute Deviations and Chebyshev Approximation. We show that the orthant decomposition of the parameter space based on Oettli-Prager Theorem yields useful algorithms, which are exponential in the number of

regression parameters, but *not* in the number of observations.

We will also show how the orthant decomposition method applies to a form of the Errors-In-Variables model. In particular, we assume that the observations of both X and y are contaminated by random errors with a bounded support with a common radius. Then, the orthant decomposition method allows us to construct a consistent estimator of the regression parameters and the error radius.

***

The talk summarizes some well-known results, some new results as well as research challenges on the border between interval theory and statistics.

Acknowledgment. Many thanks to my colleagues and friends (alphabetically): Jaromír Antoch, Milan Hladík, Miroslav Rada and Ondřej Sokol.

This work was supported by the Czech Science Foundation under Grant P402/12/G097.

References

[**1**] M. Černý, J. Antoch and M. Hladík, On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data, *Information Sciences* 244: 26–47, 2013.

[**2**] M. Černý and M. Hladík, The complexity of computation and approximation of the t-ratio over one-dimensional interval data, *Computational Statistics & Data Analysis* 80: 26–43, 2014.

[**3**] E. Dantsin, V. Kreinovich, A. Wolpert and G. Xiang, Population variance under interval uncertainty: A new algorithm, *Reliable Computing* 12(4): 273–280, 2006.

[**4**] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré and M. Aviles, Computing variance for interval data is NP-hard, *ACM SIGACT News* 33(2): 108–118, 2002.

[5] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact bounds on finite populations of interval data, *Reliable Computing* 11(3): 207–223, 2005.

[6] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf and L. Ginzburg, *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories, Albuquerque, 2007.

[7] M. Hladík and M. Černý, Linear regression with interval data: Computational issues, Technical report II/2015, Department of Econometrics, University of Economics in Prague, `http://nb.vse.cz/~cernym/lri.pdf`.

[8] M. Hladík and M. Černý, Total least squares and Chebyshev norm, *Proceedings of International Conference on Computational Science (ICCS) 2015*, Reykjavik, to appear in *Procedia Computer Science*, Preprint: `http://nb.vse.cz/~cernym/tls.pdf`.

[9] J. Horowitz, C. Manski, C. Ponomareva and J. Stoye, Computation of bounds on population parameters when the data are incomplete, *Reliable Computing* 9(6): 419–440, 2003.

[10] H. T. Nguyen, V. Kreinovich, B. Wu and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Studies in Computational Intelligence 393, Springer, 2012.

[11] S. Schön and H. Kutterer, Using zonotopes for overestimation-free interval least-squares: Some geodetic applications, *Reliable Computing* 11: 137–155, 2005.

[12] E. Tamer, Partial identification in econometrics, *Annual Review of Economics* 2: 167–195, 2010.

[13] G. Xiang, M. Ceberio and V. Kreinovich, Computing population variance and entropy under interval uncertainty: Linear-time algorithms, *Reliable computing* 13(6): 467–488, 2007.