Uncertainty quantification for mean estimates

Noémi Takács

Supervisor: Ambrus Tamás, SZTAKI, ELTE

1 Introduction

In this project I study confidence region estimates of several parameters. In the first semester I dealt with the one-dimensional case, i.e., built confidence intervals for the mean or median of a random variable. In this report I empirically analize the Sign-Perturbed Sums (SPS) algorithm [1, 2] which constructs confidence intervals for the median of symmetric variables. I show its advantages through many examples compare to other asymptotic methods. Finally, I present a reformulation of the resampling framework to a simple discrete mean estimation problem.

In the scalar case, the general linear regression model looks as shown below:

$$Y_t = X_t \cdot \vartheta^* + \varepsilon_t \quad (t = 1, \dots, n),$$

where $Y_t \in \mathbb{R}$ is the output, $X_t \in \mathbb{R}$ is the regressor, $\varepsilon_t \in \mathbb{R}$ is the noise and $\vartheta^* \in \mathbb{R}$ is the true parameter, which we want to estimate. In my simulations I deal with the case when $X \equiv 1$. My only assumption is

(a1) The noise term, $\{\varepsilon_t\}$ is a sequence of independent random variables, and each of them has a symmetric probability distribution about zero.

When we talk about confidence intervals, we can discuss two important properties: whether they contain the "true" expected value corresponding to the confidence level and their lengths. Obviously, we want an exact inclusion rate and in addition the length of the interval to be as short as possible.

2 The SPS method

A major advantage of the SPS algorithm is that it requires mild statistical assumptions, (a1), hence it is distribution-free. The distribution of the noise can be arbitrary as long as it is symmetric about zero. Unlike other asymptotic methods, e.g., which use the central limit theorem (CLT), SPS provides finite sample guarantees, that is SPS can be used to construct non-asymptotic confidence regions (intervals) for any finite sample, so it provides reliable results even for small sample sizes.

The SPS algorithm starts with the initialization part. For a user-chosen confidence level $p \in (0, 1)$ set integers $m > q > 0$ such that $p = 1 - q/m$. Generate $n(m - 1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with $\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = 0,5$ for all integers $1 \leq i \leq m - 1$ and $1 \leq t \leq n$. Then generate a permutation π of the set $\{0, \dots, m-1\}$ randomly, where each of the $m!$ possible permutations has the same probability $1/(m!)$ to be selected. The algorithm for deciding whether a ϑ parameter is included in the confidence region is shown in Table 1.

```
1. For a given $\vartheta$, compute the prediction errors
       $\varepsilon_t(\vartheta) \doteq Y_t - \vartheta \quad \text{for} \quad 1 \leq t \leq n.$
2. Evaluate
       $S_0(\vartheta) \doteq \sum_{t=1}^{n} \varepsilon_t(\vartheta), \quad \text{and} \quad S_i(\vartheta) \doteq \sum_{t=1}^{n} \alpha_{i,t}\varepsilon_t(\vartheta),$
   for all indices $1 \leq i \leq m - 1$.
3. Order scalars $\{S_i^2(\vartheta)\}$ according to $\succ_\pi$, where
   "$\succ_\pi$" is "$>$" with random tie-breaking, cf. [1]
4. Compute the rank of $S_0^2(\vartheta)$ by
       $\mathcal{R}(\vartheta) \doteq \left[ 1 + \sum_{i=1}^{m-1} \mathbb{I}\left( S_0^2(\vartheta) \succ_\pi S_i^2(\vartheta) \right) \right].$
5. Return 1 if $\mathcal{R} \leq m - q$, otherwise return 0.
```

Table 1: Pseudocode: SPS-Indicator(θ)

Definition 1 *The p-level SPS confidence region:*

$$C_p \doteq \{\vartheta \in \mathbb{R} : \textit{ SPS-indicator}(\vartheta) = 1\}.$$

Theorem 1 *Assuming (a1) the coverage probability of the SPS confidence interval is exactly p, i.e.,*

$$\mathbb{P}\left(\vartheta^* \in C_p\right) = 1 - \frac{q}{m} = p.$$

3 Simulations

First, I implemented the algorithm of the SPS method. To identify its properties, I chose different distributions and generated a sample from each of them. I constructed a confidence interval for the medians of these symmetric distributions, using the SPS and a method based on the CLT. For each sample I repeated the previous steps 10 000 times for different number of observations: from 10 to 100 increasing by 10. Finally I measured the proportion of cases, where the interval contained the true parameter, and the average length of the intervals. The examined distributions were as follows:

- Standard normal distribution.
- Mixture of two normal distribution: I randomly generate each data point from $\mathcal{N}(-m, 1)$ or $\mathcal{N}(m, 1)$ with probability 1/2 for each distributions, for $m = 2, 10, 20$.
- Student's t distribution with 2 degrees of freedom.
- Standard Cauchy distribution: it has a property that both its expected value and its variance are undefined. It causes a huge difference in the two methods, i.e., the asymptotic CLT type method fails in this case.

- Symmetrized Pareto distribution with different values of α: I consider three values of α, first 2.5, when both the expected value and the variance exist, then 1.5, when there is no variance but the expected value exists and finally 0.5, when neither the expected value nor the variance exist.

My results are shown in Figure 1, 2 and 3. From the simulations one can conclude that SPS produces stable exact confidence levels, unlike the CLT-based method, which can be less useful when there is no variance and/or the sample size is too small. Although the CLT intervals are generally shorter, the ones generated by the SPS are shrinking also at a similar rate to the other method. Thus a shorter confidence interval can be misleading.

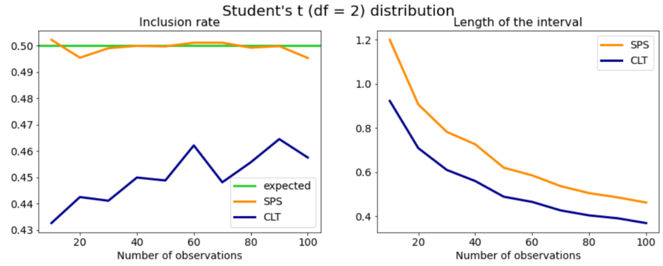

Figure 1: Simulation results for t distribution

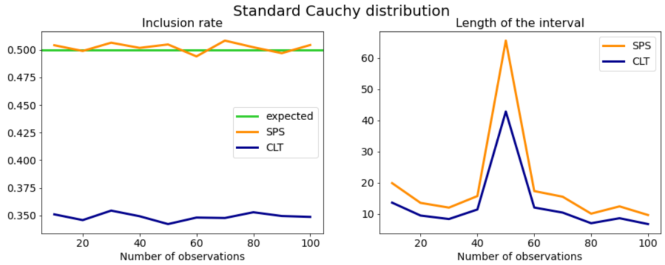

Figure 2: Simulation results for Cauchy distribution

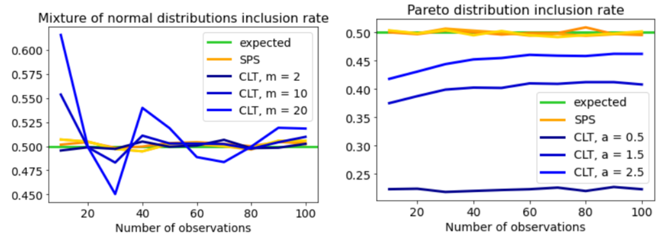

Figure 3: Simulation results for normal mixtures and Pareto distributions

4 Classification

In classification problems we want to predict the output using explanatory variables. The main difference w.r.t. regression problems is that in regression the output variable is continuous, but in classification it is discrete, in the simplest binary case Y_t can take only two values, so let $Y_t \in \{0,1\}$. I deal with the case where there is no explanatory variable, that is Y_t is an indicator variable. The aim is to estimate the unknown probability p. Here we have the benefit that we know the distribution family of Y_t and $n\bar{Y}$, which is binomial with order n and parameter θ^*, where $\bar{Y}$ is the sample mean. The „best" method to construct confidence intervals for θ^* uses the binomial distribution. One can test any candidate θ by generating variables $Y_{i,t}$ from an indicator distribution with probability θ for $i = 1, \dots, m-1$ and $t = 1, \dots, n$. The SPS-like method can be used with $S_i(\theta) \doteq \sum_{t=1}^{n}(Y_{i,t} - \theta)$ for $i = 1, \dots, m-1$.

I simulated i.i.d. samples for $\theta^* = 0.8$, and constructed three type of confidence intervals based on the SPS, CLT and binomial distribution. I repeated the experiments 10 000 times and examined the inclusion rates of the true parameter and the length of the intervals. The results are shown in Figure 4.

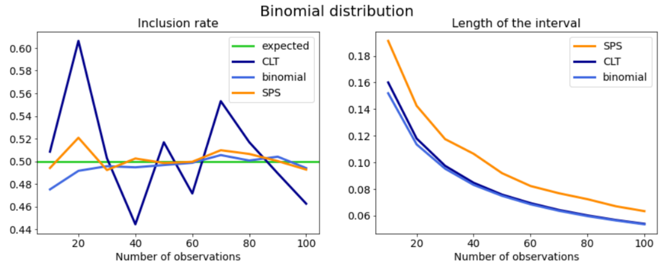

Figure 4: Simulation results for binomial distribution

My simulation results are similar to the previous ones, but now we can say that the construction based on the binomial distribution performs the best. This gave the shortest confidence interval and at the same time the inclusion rate was close to 0.5. It is not surprising, this was to be expected. But in general, when we have one or more classifiers, we do not know any theoretical distribution. In those cases can be very useful a non-asymptotic, distribution-free method.

5 Conclusions

In this report I presented the SPS method and supported its theoretical guarantees with simulations. In the following semesters, the plan is to deal with the generalization of this method for multivariate classification problems.

References

[1] Csáji, B. Cs., Campi, M. C., Weyer, E. (2015). Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models. *IEEE Transactions on Signal Processing*, 63(1), 169–181.

[2] Szentpéteri, Sz., Csáji, B. Cs. (2023). Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case*. *IFAC-PapersOnLine*, 56(2), 10363-10370.