Stat4DS / Homework 03

Pierpaolo Brutti

Due anytime before February 15, 2020, 23:59 PM on Moodle

General Instructions

I expect you to upload your solutions on Moodle as a **single running** `R Markdown` file (`.rmd`) + its `.html` output, named with your surnames.

R Markdown Test

To be sure that everything is working fine, start `RStudio` and create an empty project called `HW3`. Now open a new `R Markdown` file (`File > New File > R Markdown...`); set the output to `HTML mode`, press `OK` and then click on `Knit HTML`. This should produce a web page with the knitting procedure executing the default code blocks. You can now start editing this file to produce your homework submission.

Please Notice

- For more info on `R Markdown`, check the support webpage that explains the main steps and ingredients: R Markdown from RStudio.
- For more info on how to write math formulas in LaTex: Wikibooks.
- Remember our **policy on collaboration**: *Collaboration on homework assignments with fellow students is* ***encouraged****. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had discussions concerning your solution. You may* ***not****, however, share written work or code after discussing a problem with others. The solutions should be written by* ***you****.*

Exercise: Estimating a population mean. . . in 2020. . .

1. Introduction

Given $\{X_1, \ldots, X_n\}$ IID from some (*univariate* for now) distribution, in this exercise we consider a seemingly trivial goal:

♡ ESTIMATE THE POPULATION MEAN $\mu = \mathbb{E}(X)$ ♡

An obvious choice would be the plug-in estimator, the *empirical mean* that you all know and love. . .

$$\widehat{\mu}_n \overset{\text{def.}}{=} \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This estimator is computationally attractive, requires no prior knowledge and automatically scale with the population variance σ. In addition, tweaking a bit the *Central Limit Theorem*, we also know that

$$\boxed{\mathbb{P}(|X| > t) \leq C e^{-v t^2}}$$

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\sqrt{n}\,|\widehat{\mu}_n - \mu|}{\sigma} \leqslant \sqrt{2\,\log\left(\tfrac{2}{\alpha}\right)}\right) = \lim_{n\to+\infty} \mathbb{P}\left(|\widehat{\mu}_n - \mu| \leqslant \sigma\sqrt{\tfrac{2}{n}\,\log\left(\tfrac{2}{\alpha}\right)}\right) \geqslant 1 - \alpha,$$

Sub-Gausian
Gausian

result that also holds *non-asymptotically* under some suitable technical conditions. If these conditions are not met, we still have Chebyshev's inequality, which says that with probability at least $1 - \alpha$

$$|\widehat{\mu}_n - \mu| \leqslant \sigma\sqrt{\frac{2}{n\,\alpha}},$$

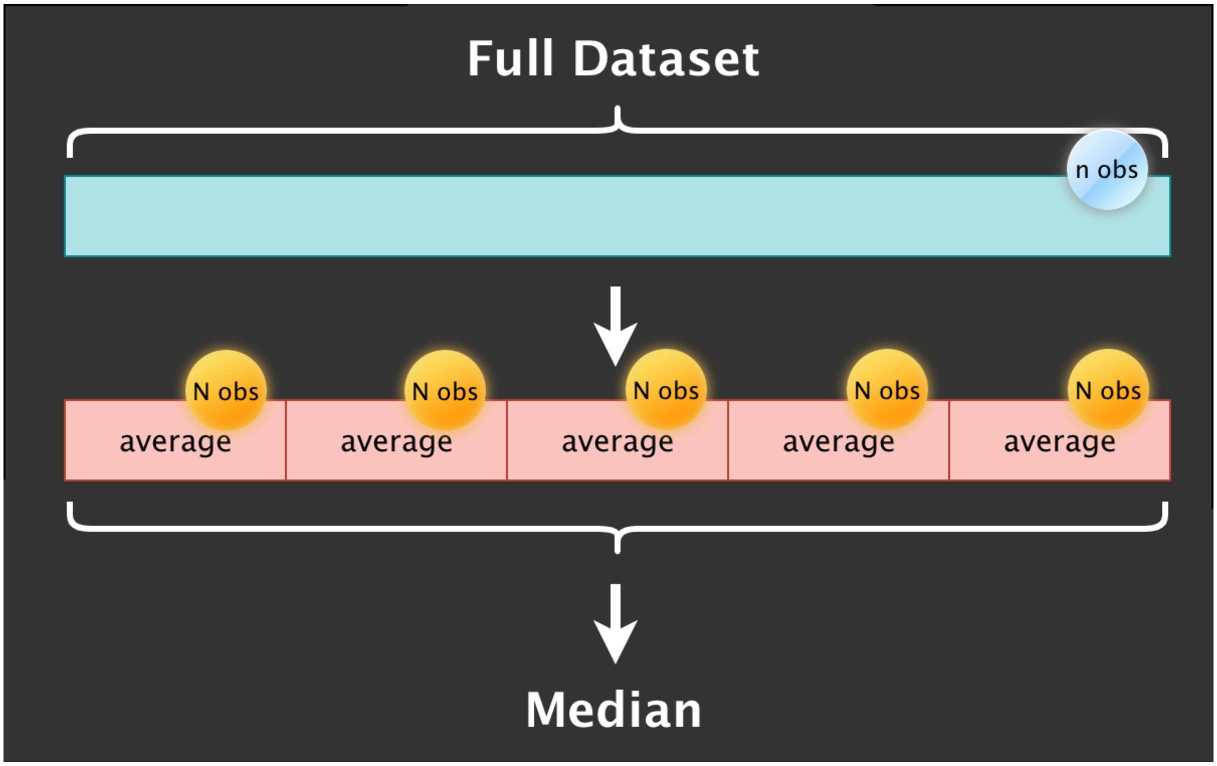
an exponentially weaker bound that will especially hurt in modern applications where many means have to be estimated simultaneusly (e.g. empirical risk minimization methods).

2. Our question: can we do better?

The sample average $\widehat{\mu}_n$ is our gold standard, but there is an interesting alternative: the **median-of-means** (`MoM`) estimator.

To define the `MoM`, assume that we chop the original n observations in k independent blocks of size N (approximately), then

$$\widehat{\mu}_n^{\texttt{MM}}(k) = \{\text{median of the } k \text{ block-means}\} = \text{median}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i, \ldots, \frac{1}{N}\sum_{i=(k-1)N}^{kN} X_i\right\}.$$



This new estimator is in general biased but, if we carefully choose the block number k, then for any distribution with finite variance σ (and also in some infinite variance case) with probability at least $1-\alpha$ we have

$$\left|\widehat{\mu}_n^{\texttt{MM}}(k) - \mu\right| \leqslant 8\,\sigma\sqrt{\tfrac{1}{n}\log\left(\tfrac{2}{\alpha}\right)},$$

an inequality exactly of the form we like.

The theoretical optimal block number is then $k^\star = \lceil 8\log(1/\alpha)\rceil$ where $\lceil\cdot\rceil$ denotes the ceiling function.

3. Multivariate extension

In this section we briefly discuss extensions of the mean estimation problem to the multivariate setting, that is, when one is interested in estimating the mean of a random vector.

Let $\boldsymbol{X}$ be a random vector taking values in $\mathbb{R}^d$. Assume that the mean vector $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ and covariance matrix $\Sigma = \mathbb{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathtt{T}}$ exist.

Given n independent, identically distributed samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ drawn from the distribution of $\boldsymbol{X}$, one wishes to estimate the mean vector.

Just like in the univariate case, a natural choice is the sample mean but only with

$$\widehat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The sample mean has a near-optimal behavior whenever the distribution is sufficiently **light-tailed**. However, whenever **heavy tails** are a concern, the sample mean is to be avoided as it may have a sub-optimal performance: **robust statistical analysis** are almost unavoidable in this case, and the median strikes back!

One may in fact try to extend the MoM estimator to the multivariate case, only one problem: what is a *median* in the multivariate case? Given n points $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ in $\mathbb{R}^d$, the center of the smallest ball that contains *at least* half of the points may be considered as a notion of a multivariate median. Computing such a median is totally a nontrivial problem.

Instead, we may leverage the *variatinal* definition of the median as the values that minimize the MAE to get the so-called **geometric median** defined as

$$\vec{\boldsymbol{m}} = \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^d} \sum_{i=1}^{n} \|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{a}}\|, \qquad a \in \mathbb{R}^d$$

where $\|\cdot\|$ denotes the euclidean norm/distance between d dimensional vectors.

The multivariate MoM estimator may be defined as the geometric median of the sample means of the k blocks defined before. As in the univariate case, the *theoretical* optimal block number is $k^\star = \lceil 8\log(1/\alpha)\rceil$. As a quick final remark, let me notice that, for the purpose of experimenting a bit, we could replace this last summary with any robust multivariate location estimator.

4. On robust procedures and heavy-tailed distribution in Data Science

As mentioned, MoM estimators should have an edge on the beloved sample average in multivariate, heavy-tailed cases. At this point in time, *heavy-tailed distributions* have been accepted as realistic models for various phenomena:

- www-session characteristics (e.g. sizes and durations of sub-sessions; sizes of responses inter-response time intervals)
- on/off-periods of packet traffic
- file sizes
- service-time in queueing model
- flood levels of rivers
- major insurance claims
- extreme levels of ozon concentrations
- high wind-speed values
- wave heights during a storm
- low and high temperatures

But there's more. As you probably know, recent technological developments have allowed companies and state organizations to collect and store huge datasets. Big datasets have also challenged scientists in statistics and computer science to develop new methods. In fact, because of the very "unstructured" way in which these datasets are collected, oftentimes they tend to be corrupted by nasty outliers and/or exhibit heavy tails.

The need for robust statistical procedures in data science can be appreciated by the recently posted challenges on kaggle: the 1.5 million dollars problem *Passenger Screening Algorithm Challenge* is about to find terrorist activity from 3D images, whereas *The NIPS 2017: Defense Against Adversarial Attack* regards constructing algorithms robust to adversarial data.

There are mainly two types of outliers in practice: those corrupting a dataset which are **not** interesting (outliers can appear in datasets due to storage issues, they can also be adversarial data as fake news, false declarative data, etc.), and those that are rare but important observations like frauds, terrorist activities, tumors in medical images, etc. Two famous examples of the latter type of outliers discovered unexpectedly were the CMB by Penzias and Wilson in 1964, and the ozone hole by Farman and Gardiner in 1985. In the latter case, the challenge is to detect outliers, whereas in the former the main problem is to construct predictions as sharp as if the dataset was clean: MoM in this business!

Finally, as a quick reminder, here's some family of heavy/light tailed distributions we mentioned along the way:

- **Light-tailed distributions**
  - Exponential
  - Gamma
  - Weibull (with shape parameter larger than 1)
  - Normal
- **Heavy-tailed distributions**
  - Subexponential (e.g. Pareto, Lognormal, Weibull with shape parameter lesser than 1)
  - With regularly varying tails (e.g. Pareto, Cauchy, Burr, Zipf-Mandelbrot)

⇝ Your job ⇜

1. Setup a sensible simulation study to compare $\widehat{\mu}_n = \bar{X}_n$ and $\widehat{\mu}_n^{\texttt{MM}}(k)$ in terms of their bias, variance and, then, MSE:

$$\text{MSE}\big(\widehat{\mu}\big) \overset{\texttt{def.}}{=} \mathbb{E}\big(\mu - \widehat{\mu}\big)^2 = \mathbb{V}\text{ar}\big(\widehat{\mu}\big) + \text{bias}^2\big(\widehat{\mu}\big) \;\text{ where }\; \widehat{\mu} \;\text{ is a generic estimator of }\; \mu.$$

   You should carry out the simulation under at least two population models, one with light tails, and one with fatter tails (explaining your choice).

   For each scenario, fix the variance/noise level σ to some value, and vary the sample size n (starting quite small) and, for the *median-of-means*, the block number k. Make nice plots to explain your findings, and carefully comment the results.

2. Repeat the previous exercise for the multivariate case. Notice that, in this multivariate setup, the MSE of a generic estimator $\widehat{\boldsymbol{\mu}} = \big[\widehat{\mu}_1 \cdots \widehat{\mu}_j \cdots \widehat{\mu}_d\big]^{\texttt{T}}$ for the d-dimensional mean vector $\boldsymbol{\mu} = [\mu_1 \cdots \mu_j \cdots \mu_d]^{\texttt{T}}$ is simply defined as

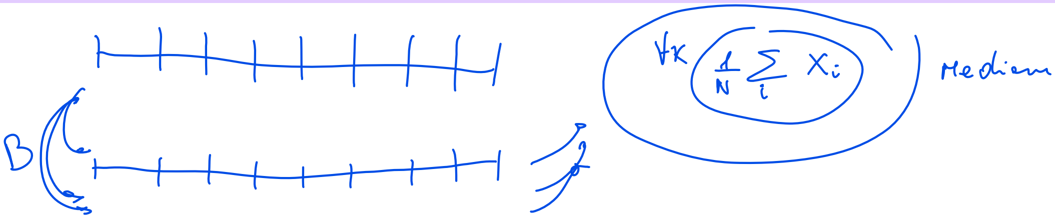
$$\text{MSE}\big(\widehat{\boldsymbol{\mu}}\big) \overset{\texttt{def.}}{=} \mathbb{E}\big\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\big\|^2 \overset{\texttt{def.}}{=} \sum_{j=1}^{d} \mathbb{E}\big(\mu_j - \widehat{\mu}_j\big)^2 = \sum_{j=1}^{d} \text{MSE}\big(\widehat{\mu}_j\big) = \sum_{j=1}^{d} \Big(\mathbb{V}\text{ar}\big(\widehat{\mu}_j\big) + \text{bias}^2\big(\widehat{\mu}_j\big)\Big).$$

   Use some of the functions provided by the package `heavy` to simulate from the multivariate light/heavy tailed distribution of your choice (pick the dimension d you prefer **or**, even better, let d vary to), and then use the package `Gmedian` to get the *geometric median*.

3. Consider the `CRSPday` data. In this dataset there are 4 variables, daily returns from January 3, 1969, to December 31, 1998, on 3 stocks, `GE`, `IBM` and `Mobil`, and on the `CRSP` value-weighted index, including dividends. CRSP is the Center for Research in Security Prices at the University of Chicago.

   First of all, numerically and graphically explore the distributions of each single variable trying to quantify how "heavy" their tails are (histograms, boxplot, normal-quantile plot, kurtosis estimator, or even the more specialized Hill's estimator, can all be useful here). Then, since the correlations between each individual stock and a market index are a key component of finance theory (e.g. Capital Asset Pricing Model) explore also this aspect of the dataset.

   Finally we want to estimate the (4-dimensional) mean vector using the `MoM`, but we need to choose the block number k. To this end, consider a sequence of candidate values $k_{\texttt{seq}} = \{k_1, k_2, \ldots, k_\ell\}$. For each one of them use the *nonparametric bootstrap* to estimate the MSE of $\widehat{\mu}_n^{\texttt{MM}}(k)$. Pick the block size with associated the smallest (bootstrap-estimated) MSE.


B
∀k
1/N Σ_i X_i
Median