Fundamentals of Statistical Learning

December 17-18 2025

Classwork-02

Question 1

An old friend! The file `CRSPday.csv` contains information on the returns for specific stocks (e.g. `IBM`) and the `CRSP` index.

Tomorrow I'll select two of them **and** two competing copula models: your task will be to use a suitable model selection technique to pick the best one.

The main `R` packages to use will be `copula` and `fCopulae`. This vignette for the `copula` package could also be useful.

Question 2

Background: Robust procedures and heavy-tailed distribution in data analysis

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 can be also appreciated by looking at some past challenges on `kaggle` – surely there are more, these are just examples – like 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.

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)

Material: Estimating a population mean in 2025

Given $\{X_1, \dots, X_n\}$ IID from some (*univariate* for now) distribution F_X, here we consider a seemingly trivial goal:

♡ NONPARAMETRICALLY 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 $\widehat{\mu}_n \overset{\texttt{def.}}{=} \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. 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 could also show that

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

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 $1-\alpha$

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

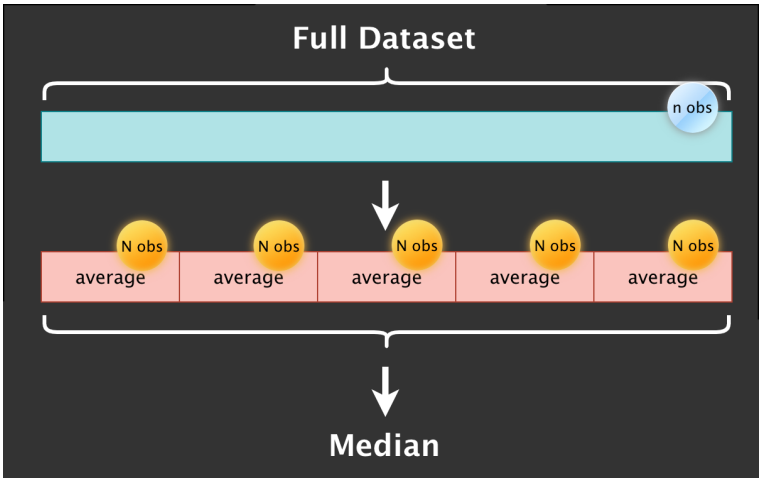
an exponentially weaker bound that will especially hurt in modern applications where many means have to be estimated simultaneusly. All in all, the sample mean has a near-optimal behavior whenever the distribution is sufficiently **light-tailed**. However, whenever **heavy tails** are a concern, the 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 *medians* strike back!

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

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



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

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

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.

But why stop here?

Under the `MoM` framework, we could also get a (potentially) more robust density estimator. Assume that the data are IID from an unknown density f_X supported on $\mathcal{X} = [0, 1]$. Then a classic nonparametric estimator is the *histogram* with binwidth h, defined as (pls, check our October-notes under **Other Random Stuff**):

$$\widehat{f}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\theta}_j}{h^d} \mathbb{I}(x \in B_j).$$

Now, imagine to split the sample into K random subsamples and compute the histogram within each sample, leading to

$$\big\{\widehat{f}_{1,h_1}(x), \dots, \widehat{f}_{K,h_K}(x)\big\}.$$

The `MoM` histogram, or `MoM-H` is

$$\widehat{f}_{\texttt{MM}}(x) = \text{median}\big\{\widehat{f}_{1,h_1}(x), \dots, \widehat{f}_{K,h_K}(x)\big\}.$$

Please Note: this estimator may *not* be a density (it may not integrate to 1), but we can simply rescale it to fix this issue.

Goal: tomorrow you'll play around this method by comparing it with the usual histogram on simulated data (I'll pick the distributions to sample from).

The metric/loss will be the square distance:

$$L_2(f_X, \widehat{f}_h) \overset{\texttt{def.}}{=} \|\widehat{f}_h - f_X\|_2 \overset{\texttt{def.}}{=} \int \big(\widehat{f}_h(x) - f_X(x)\big)^2 \mathrm{d}x,$$

with associated risk equal to

$$R(f_X, \widehat{f}_h) = \mathbb{E}\big(L_2(f_X, \widehat{f}_h)\big) \overset{\texttt{Fubini}}{=} \int \mathbb{E}\Big[\big(\widehat{f}_h(x) - f_X(x) \pm \bar{f}_h(x)\big)^2\Big] \mathrm{d}x = \int \mathbb{V}\mathrm{ar}\big(\widehat{f}_h(x)\big)\mathrm{d}x + \int \mathrm{bias}^2\big(\widehat{f}_h(x)\big)\mathrm{d}x,$$

where $\bar{f}_h(x) = \mathbb{E}\big(\widehat{f}_h(x)\big)$ is the average "output" of our algorithm (in repeated sampling from f_X), and $\mathrm{bias}\big(\widehat{f}_h(x)\big) = f_X(x) - \bar{f}_h(x)$ is its bias.

Bonus Question (optional)

Uhmmmm... we spent quite some time in getting confidence intervals and bands for the `CDF` $\theta = F_X(\cdot)$... what about the density f_X?

Starting from the histogram $\widehat{f}_h$, any "simple" way to get a confidence interval (allegedly!) targeting the parameter $\theta = f_X(x_0)$; that is, an interval that trap the true value of the density at a specific point, say x_0, with a specific level $1 - \alpha$?

Look at the structure of $\widehat{f}_h$ and assume that h is a fixed constant and not data-driven (as typically is).
