Non-Parametric Statistics Workshop 1

David Plazas Escudero
Juan Pablo Vidal
Juan Sebastián Cárdenas-Rodríguez
Mathematical Engineering, Universidad EAFIT

May 7, 2020

1 Workshop Exercises

The code done to solve this workshop can be found in a Jupyter notebook in this link.

Exercise 1 (10%)

The data used is the average daily temperatures in Canada for the past 35 years. The empirical cumulative distributions (ECDFs) for each year are presented in Figure 1, where the yellow-most curve represents the ECDF of the data recorded for the first year, and the blue-most is the last year.

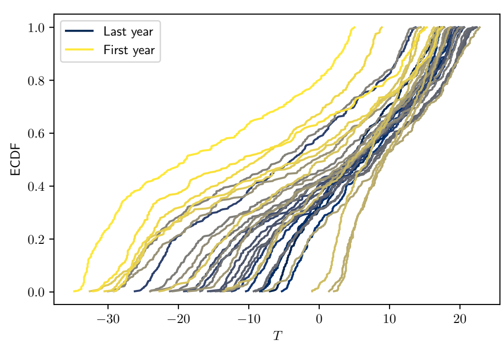


Figure 1: Empirical distributions by years.

Based on this plots, it can be observed a climate change effect over the years. The plots of the first years arise for lower values of the temperature T, whereas the last years tend to be on the right of the plot. This can be interpreted as follows: since the plots for the early years are on the left, these years reported more data on lower average daily temperature than the most recent years. This means that recent years have been hotter in average daily temperature.

Exercise 2 (Extra)

The plug-in principle is a technique used in probability theory and statistics to estimate a feature of a probability distribution (e.g., the expected value, the variance, a quantile) that cannot be

computed exactly. In general, the plug-in principle says that a feature of a given distribution can be approximated by the same feature of the empirical distribution of a sample of observations drawn from the given distribution [7].
The feature of the empirical distribution is called a plug-in estimate of the feature of the given distribution. For example, a quantile of a given distribution can be approximated by the analogous quantile of the empirical distribution of a sample of draws from the given distribution.

The following is a formal definition of plug-in estimate.

A statistical functional $T(F)$ is any function of F.
The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T\left(\widehat{F}_n\right)$$

A functional of the form $\int a(x)dF(x)$ is called a linear functional.
The plug-in estimator for linear functional $T(F) = \int a(x)dF(x)$ is:

$$T\left(\widehat{F}_n\right) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a\left(X_i\right)$$

It is important to note that $T\left(F_n\right)$ converges to $T(F)$ as the sample size n increases.
In the practice, the limited area is calculated in the first quadrant for each empirical curve of the temperature data, this allows to extract the plug-in estimator of the mean, since the difference between the area with positive values and the area with negative values is extracted from the empirical curve, which allows estimating the mean of the series. The plug-in estimation of the mean for each year is presented in Figure 2, where a comparison with the natural estimator of the mean is presented.

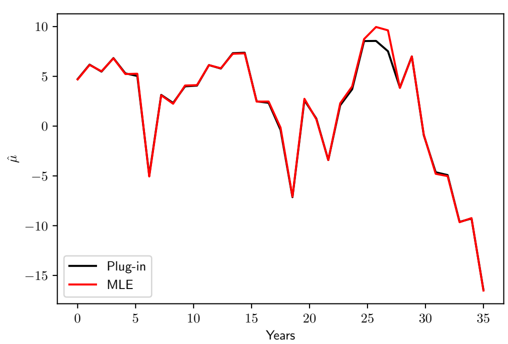


Figure 2: Mean estimation of the series.

From Figure 2 an almost identical estimate can be seen between the two estimators, however Figure 3 shows a slight difference between the mean estimates, this is because the only case in which both estimates are identical is when the sample size is infinite.

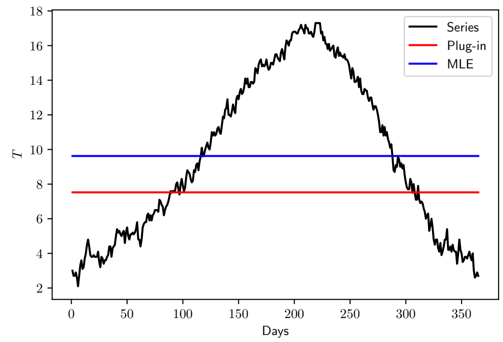


Figure 3: Comparison between the natural estimator and the plug-in estimator.

Exercise 3 (10%)

Calculate and plot the confidence bands for the empirical continuous distribution function (ECDF) of the coldest and hottest year in average with a confidence of 95 % . Are there any sectors that are not enclosed in the bands?

Proof. Let n be the size of the sample and $1-\alpha$ the desired confidence for the bands. To calculate the confidence bands for the ECDF, we first define ϵ_n by the following formula:

$$\epsilon_n = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}$$

Let $\hat{F}_n(x)$ be the ECDF. Then, for each x in the ECDF we define the lower $(L(\cdot))$ and upper $(U(\cdot))$ bound by:

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\}$$
$$U(x) = \min\{\hat{F}_n(x) + \epsilon_n, 1\}$$

The results obtained by using the temperatures of the coldest and hottest year are seen in Figure 4.

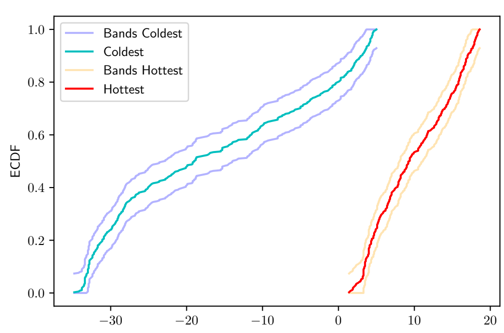


Figure 4: Bands for the coldest and hottest year.

It can be seen that the upper and lower bands fully enclose the ECDF. Nevertheless, there exists two points where the function and the bands meets. This happens in the lower band at the beginning and the upper band at the final point.
This phenomenon is due to the full certainty at those points in a sense that the lower bound, at the start, has to be the same point as it cannot go lower than 0. A similar reasoning can explain the upper bound and the final point. □

Exercise 4 (10%)

Write and execute a code that allows to visualize the Glivenko Cantelli for a Weibull distribution.

Proof. The Glivenko Cantelli shows the relationship of how the difference between the empirical and theoretical distribution changes depending of the number of the sample.
In this manner, if n is the size of the sample, $\hat{F}_n(\cdot)$ is the ECDF of that sample and $F(\cdot)$ is the theoretical distribution the theorem states that:

$$\sup_x \mid \hat{F}_n(x) - F(x) \mid \xrightarrow{\text{a.s.}} 0$$

In these manner, we generated Weibull random variables of different sizes n_i calculated with:

$$n_i = 2^i, \quad \text{for } i = 1, \ldots, 20$$

The results of the experiment can be found in Figure 5.

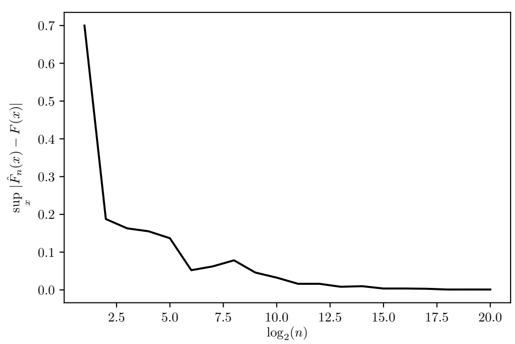


Figure 5: Glivenko Cantelli theorem visualization.

It is clear that the biggest difference between the ECDF and the theoretical distribution starts to go to 0 when n gets bigger. Furthermore, although the behavior in the graph does not show a monotone decrease, it does display the asymptotic behavior described in the theorem. □

Exercise 5 (5 %)

Theorem 1.1 (Jensen's Inequality)*.* Let $g(\cdot)$ be a concave function and let X be a random variable, then

$$\mathbb{E}\left[g(X)\right] \leq g(\mathbb{E}\left[X\right]).$$

Proof. Let $g(\cdot)$ be a concave function and $L(x)$ be its tangent line at a fixed point x_0. It holds that (see Theorem 6 from [5])

$$g(x) \leq L(x). \tag{1}$$

If $x_0 = \mathbb{E}[X]$. From (1) and given that $\mathbb{E}[\cdot]$ preserves order, we have

$$\mathbb{E}[g(X)] \leq \mathbb{E}[L(X)] = \mathbb{E}[a + bX] = a + b\mathbb{E}[X] = L(\mathbb{E}[X]) = g(\mathbb{E}[X]).$$

$\square$

Exercise 7 (Extra)

Deduce the distribution and density of the j-th ordered statistic. Explain with detail what would be a simple procedure to simulate the j-th ordered statistic. Simulate 1000 observations of some ordered statistic of a sample of size n that comes from a Weibull distribution. Draw in a same graph the ECDF and theoretical distribution.

Proof. Let's first deduce the distribution and density of the j-th ($X_{[j]}$) ordered statistic. Let $X_1, \ldots, X_n$ be independent random variables that come from a distribution $F(\cdot)$. Hence, the probability that $X_i \leq t$ is given by:

$$\mathrm{P}\left(X_i \leq t\right) = F(t)$$

Let Z_t be a random variable that represents the number of variables whose value are less than t. Hence:

$$Z_t \in \{0, 1, \ldots, n\}$$

Hence:

$$\begin{aligned}
\mathrm{P}\left(Z_t = 0\right) &= \mathrm{P}\left(X_1 > t \wedge \ldots \wedge X_n > t\right) \\
&= \mathrm{P}\left(X_1 > t\right) \ldots \mathrm{P}\left(X_n > t\right) \\
&= (1 - F(t)) \ldots (1 - F(t)) \\
&= (1 - F(t))^n \\
\mathrm{P}\left(Z_t = 1\right) &= \binom{n}{1} \mathrm{P}\left(X_1 \leq t \wedge X_2 > t \wedge \ldots \wedge X_n > t\right) \\
&= \binom{n}{1} F(t)(1 - F(t)) \ldots (1 - F(t)) \\
&= \binom{n}{1} F(t)(1 - F(t))^{n-1} \\
&\vdots \\
\mathrm{P}\left(Z_t = j\right) &= \binom{n}{j} F(t)^j (1 - F(t))^{n-j}
\end{aligned}$$

Hence, the distribution j-th ordered statistic is given by:

$$\mathrm{P}\left(X_{[j]} \leq t\right) = \mathrm{P}\left(Z_t \geq j\right) = \sum_{i=j}^{n} \binom{n}{i} F(t)^i (1 - F(t))^{n-i}$$

In these manner, we obtained the distribution for the j-th ordered statistic. Then, to obtain the density ($f_{[j]}(\cdot)$) we just differentiate hence:

$$
\begin{aligned}
f_{[j]}(t) &= \frac{d}{dt}\mathrm{P}\left(X_{[j]} \leq t\right) \\
&= \sum_{i=j}^{n}\binom{n}{i}\left[iF(t)^{i-1}(1-F(t))^{n-i}f(t)-(n-i)F(t)^{i}(1-F(t))^{n-i-1}f(t)\right] \\
&= f(t)\sum_{i=j}^{n}\binom{n}{i}F(t)^{i-1}(1-F(t))^{n-i-1}\left[i(1-F(t))-(n-i)F(t)\right] \\
&= f(t)\sum_{i=j}^{n}\binom{n}{i}F(t)^{i-1}(1-F(t))^{n-i-1}(i-nF(t)) \\
&= f(t)\left[\sum_{i=j}^{n} i\binom{n}{i}F(t)^{i-1}(1-F(t))^{n-i-1}-\sum_{i=j}^{n} n\binom{n}{i}F(t)^{i}(1-F(t))^{n-i-1}\right] \\
&= nf(t)\left[\sum_{i=j}^{n}\binom{n-1}{i-1}F(t)^{i-1}(1-F(t))^{n-i-1}-\sum_{i=j}^{n}\binom{n}{i}F(t)^{i}(1-F(t))^{n-i-1}\right]
\end{aligned}
$$

Furthermore, simulating a j-th statistic can be done easily. In first place, generate samples of size n and obtain the j-th statistic of that sample. These generates one data of the j-th ordered statistic. Repeat the previous process until the desired number of data is obtained.
The result of the simulation of the 5-th ordered statistic for a Weibull distribution can be found in Figure 6.

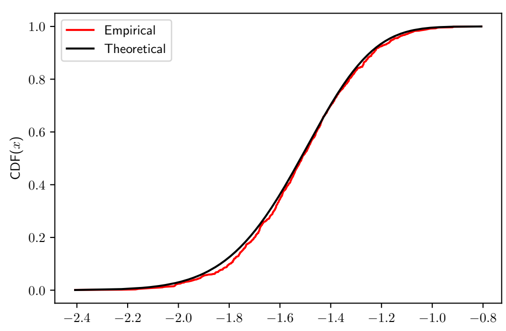


Figure 6: Comparison between ECDF and theoretic distribution.

□

Exercise 8 (10%)

Suppose X is an exponentially random variable of parameter β. Calculate:

$$P(|X-\mu|>k\sigma)$$

for $k>1$.

Proof. Recall Chebyshev's inequality: let Y be a random variable and let $\mathbb{E}[Y] = \mu$ and $\operatorname{Var}[Y] = \sigma^2$, then

$$P(|Y-\mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Let

$$\begin{aligned}
P(|X-\mu| > k\sigma) &= P(-k\sigma > X-\mu > k\sigma) \\
&= P(-k\sigma+\mu > X > k\sigma+\mu) \\
&= P(X < \mu - k\sigma) + P(X > k\sigma+\mu) \\
&= P\left(X < \frac{1-k}{\beta}\right) + 1 - P\left(X < \frac{k+1}{\beta}\right) \qquad \text{but } 1-k<0 \\
&= 1 - \left(1 - e^{-\frac{k+1}{\beta^2}}\right) \\
&= P(|X-\mu| > k\sigma) \\
&\leq \frac{\sigma^2}{(k\sigma)^2} \\
&\leq \frac{1}{k^2}
\end{aligned}$$

Clearly $e^{-\frac{k+2}{\beta^2}}$ is always lesser than 1 and because $k > 1$ then $\frac{1}{k^2}$ is also lesser than one. Thus, $P(|X-\mu| > k\sigma)$ is bounded by Chebyshev's inequality. $\square$

Exercise 9 (5 %)

Prove that if $X \sim \text{Poisson}(\lambda)$, then

$$P(X \geq 2\lambda) \leq \frac{1}{\lambda}.$$

Proof. Let $X \sim \text{Poisson}(\lambda)$. Recall Chebyshev's inequality: let Y be a random variable and let $\mathbb{E}[Y] = \mu$ and $\operatorname{Var}[Y] = \sigma^2$, then

$$P(|Y-\mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Clearly, $\mathbb{E}[X] = \operatorname{Var}[X] = \lambda$. If we set $t = \lambda$, then

$$\begin{aligned}
P(|X-\lambda| \geq \lambda) &\leq \frac{1}{\lambda} \\
P\left[(X-\lambda \geq \lambda) \cup (X-\lambda \leq -\lambda)\right] &\leq \frac{1}{\lambda} \\
P\left[(X \geq 2\lambda) \cup (X \leq 0)\right] &\leq \frac{1}{\lambda} \\
P(X \geq 2\lambda) + P(X \leq 0) &\leq \frac{1}{\lambda} \\
P(X \geq 2\lambda) &\leq \frac{1}{\lambda}
\end{aligned}$$

$\square$

Exercise 10 (5 %)

Definition 1.1. The sequence $\{X_n\}$ of random variables is said to converge in probability to X if

$$\forall \epsilon > 0, \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0$$

Definition 1.2. The sequence $\{X_n\}$ of random variables is said to converge in mean-square to X if

$$\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$$

Theorem 1.2. The convergence in mean-square implies convergence in probability.

Proof. Let $\{X_n\}$ be a sequence of random variables that converges in mean-square to X. Recall Markov's inequality: Let Y be a non-negative random variable and suppose $\mathbb{E}[Y]$ exists. Then for any $t > 0$,

$$P(Y > t) \leq \frac{\mathbb{E}[Y]}{t}.$$

Take $Y = |X_n - X|$ and $t = \epsilon$, then

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P\left[(X_n - X)^2 > \epsilon^2\right] \\ &\leq \frac{\mathbb{E}\left[(X_n - X)^2\right]}{\epsilon^2} \end{aligned}$$

Since $\{X_n\}$ converges in mean-square to X
, $\mathbb{E}\left[(X_n - X)^2\right] \to 0$ as $n \to \infty$, which directly implies that $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. □

Exercise 11 (7 %)

Show that the ECDF converges in probability to the theoretical continuous distribution function.

Proof. Let X_i, for $i = 1, \ldots, n$, be a independent data sample. Then, the empirical distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}$$

Hence, to see that it converges in probability lets see if the MSE tends to 0 when the number of samples is bigger. Let's calculate the expectancy of the estimator.

$$\begin{aligned} \mathbb{E}\left[\hat{F}_n(x)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[I(X_i \leq x)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}\left(X_i \leq x\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} F(x) \\
&= F(x)
\end{aligned}
$$

In this manner, it is a non-biased estimator for the theoretical distribution. Hence, the MSE is calculated by the variance. Then:

$$
\begin{aligned}
\text{MSE} &= \text{Var}\left[\hat{F}_n(x)\right] + \text{Bias}\left(\hat{F}_n(x), F(x)\right) \\
&= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left[I(X_i \leq x)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left[\text{P}\left(X_i \leq x\right)\left(1 - \text{P}\left(X_i \leq x\right)\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left[F(x)(1 - F(x))\right] \\
&= \frac{F(x)(1 - F(x))}{n}
\end{aligned}
$$

Hence, when $n \to \infty$ the MSE tends to 0. Therefore,

$$
\hat{F}_n(x) \xrightarrow{P} F(x)
$$

□

Exercise 12 (12 %)

Consider the daily temperatures of the hottest year in average. Calculate a confidence interval for the maximum temperature. Calculate the bias of $T_{[n]}$ and the variance.

Proof. To calculate the interval for the maximum temperature, we applied the normal bootstrap method. The normal bootstrap method consists of, in a sample $X_1, \ldots, X_n$ and a confidence of $1 - \alpha$:

1. First, extract k samples of the size n with repetition and calculate the desired statistic to find it's confidence intervals. Let S_i be the sample of each statistic calculated.

2. Calculate:
$$
v_{\text{boot}} = \frac{1}{k}\sum_{i=1}^{k}\left(S_i - \frac{1}{n}\sum_{i=1}^{k} S_i\right)^2
$$

3. Then, the interval of confidence for the statistic is, with S the statistic calculated in the original sample:
$$
(S - z_{\alpha/2}\sqrt{v_{\text{boot}}}, S + z_{\alpha/2}\sqrt{v_{\text{boot}}})
$$
where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution.

The result for the maximum temperatures interval is, with $\alpha = 0.05$ and $k = 1000$:

$$T_{[n]} \in (18.486, 18.714)$$

On the other hand, to calculate the variance and bias the Jackknife method was used. The Jackknife method consists of:

1. First, generate new samples by removing one data from the original sample. Then, calculate the statistic for this new sample. Let S_i be the sample of each statistic calculated and S be statistic calculated in the original sample.

2. To calculate the bias, use:
$$\text{Bias} = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} S_i\right)$$

□

Exercise 13 (13 %)

In this section, a sample of a uniform distribution in the interval [0,1] is generated. Then a bootstrap method to calculate the variance and a Jackknife method to calculate the bias of this sample are implemented.
The bootstrap and the jackknife are non-parametric methods for computing standard errors and confidence intervals [6].

The Jackknife

Jackknife is a simple method for approximating the bias and variance of an estimator. Let $T_n = T(X_1, \ldots, X_n)$ be an estimator of some quantity θ and let $\text{bias}(T_n) = \mathbb{E}(T_n) - \theta$ denote the bias. Let $T_{(-i)}$ denote the statistic with the i^{th} observation removed. The jackknife bias estimate is defined by
$$b_{\text{jack}} = (n-1)\left(\bar{T}_n - T_n\right)$$
where $\bar{T}_n = n^{-1}\sum_i T_{(-i)}$. The bias-corrected estimator is $T_{\text{jack}} = T_n - b_{\text{jack}}$

The Bootstrap

Bootstrap is a method for estimating the variance and the distribution of a statistic $T_n = g(X_1, \ldots, X_n)$. Let $\mathbb{V}_F(T_n)$ denote the variance of T_n. We have added the subscript F to emphasize that the variance is a function of F. If we knew F we could, at least in principle, compute the variance. For example, if $T_n = n^{-1}\sum_{i=1}^{n} X_i$ then

$$\mathbb{V}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}$$

which is clearly a function of F.
With the bootstrap, we estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}_n}(T_n)$. In other words, we use a plug-in estimator of the variance. since, $\mathbb{V}_{\widehat{F}_n}(T_n)$ may be difficult to compute, we approximate it with a simulation

estimate denoted by v_{boot}

Its important to know that The jackknife is a linear approximation of the bootstrap.

In the simulation, the Variance Bootstrap obtained from the uniform sample of the Order statistic is $v_{\text{boot}} = 3.5815\text{x}10^{-8}$, this indicates that almost all of the Order data values are nearly identical. Moreover, the theoretical bias and the Jackknife bias generated similar results, $b_{\text{jack}} = -1.6166\text{x}10^{-5}$ and $b_{\text{theoretical}} = -9.9990\text{x}10^{-5}$, which shows that the bias of the estimator is unbiased, hence the sampling distribution has a mean that is equal to the parameter being estimated.
It should be noted that the jackknife method is less computationally expensive, but the bootstrap has some statistical advantages, such as its simplicity to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution.

Exercise 14 (Extra)

Show the differences between the parametric and non-parametric bootstrap. Investigate robust versions of the bootstrap method and show examples about their performance.

Proof. Following the ideas from [8] and [4], Bootstrap is performed over a sample $X_1, ..., X_n$. In the case of the nonparametric Bootstrap, we generate new samples (resampling) based on the ECDF $\hat{F}_n$ of the sample; the generation of samples from the ECDF is equivalent to draw samples $X_1^j, ..., X_n^j$ from the original data **with replacement** for $j = 1, ..., N$, where N is the number of Bootstrap resamples. On the other hand, the parametric Bootstrap takes into consideration that the data comes from an specific distribution F_θ that, clearly, depends on an unknown parameter θ; instead of drawing from $\hat{F}_n$, we draw from $F_{\hat{\theta}}$, where $\hat{\theta}$ is an estimator of θ based on the sample. This method is just as accurates as the nonparametric, but under certain scenarios could not behave properly. An excellent example of the parametric and nonparametric Bootstrap is presented in point 11 of page 40 from [8].

□

Exercise 15 (13 %)

The data used in this section is bivariate, taking the temperature from the years that are, in average, the coldest and hottest. Let $\mathbf{Y} \in \mathbb{R}^{n\times 2}$ be the set of bivariate samples of size n (data matrix). In our case, $n = 365$ days.
The usual cutout of outliers in elliptical data (e.g. bivariate normally distributed data) is made using the χ^2 distribution. However, the bivariate data obtained from the temperatures is not guaranteed to follow such elliptical behavior. Therefore, the usual cutout will be done with the same methodology as the robust variations. The methods for estimating the covariance matrix will be first presented, then the methodology to remove outliers will be described and finally the results will be presented.

Usual Estimation

Definition 1.3. Let $\mathbf{x}$ be an multivariate observation from a set of observations with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. The Mahalanobis distance of the observation $\mathbf{x}$ is

$$d^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{2}$$

Clearly, the usual cutout is performed using the natural estimation of the Mahalanobis distance, using the standard unbiased estimators of the mean vector and covariance matrix from the data matrix $\mathbf{Y}$, presented respectively in Equation (3), where $\mathbf{y}_i$ is the i-th row of the data matrix $\mathbf{Y}$.

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}} = \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \right]^T , \quad \hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{y}_i^T - \bar{\mathbf{y}})(\mathbf{y}_i^T - \bar{\mathbf{y}})^T \tag{3}$$

Comedian Estimation

The first robust calculation of the Mahalanobis distance is based on a robust estimation of the covariance matrix following the ideas from [2], using the following definition.

Definition 1.4. Let X and Y be random variables. The comedian between X and Y is defined as

$$\mathrm{Com}(X, Y) = \mathrm{Med}\left[(X - \mathrm{Med}(X))(Y - \mathrm{Med}(Y))\right]$$

The covariance matrix is then estimated by applying the comedian to each entry. Then the Mahalanobis distance formula is applied using this matrix and the median vector instead of the mean.

Kendall Estimation

This method uses the Kendall rank correlation coefficient, usually known as Kendall's τ coefficient, originally proposed in [3]. Each entry of the covariance matrix is estimated using

$$\mathrm{Cov}(X, Y) = \rho_k S_X S_Y,$$

where ρ_k is Kendall's τ coefficient, and S_X and S_Y are the respective standard deviations.

Spearman Estimation

This estimation was performed similarly to Kendall's. The covariance matrix is estimated using

$$\mathrm{Cov}(X, Y) = \rho_s S_X S_Y,$$

where ρ_s is Spearman's correlation coefficient (see [1]), and S_X and S_Y are the respective standard deviations.

Cutout Procedure

As previously mentioned, the cutout method is performed with the same procedure for all the covariance matrix estimations. The procedure is presented as follows:

1. Estimate the covariance matrix accordingly to the estimation method.
2. Apply equation (2) to each bivariate data, with the estimation of the covariance matrix from step 1 and the corresponding estimated mean vector.
3. Fit a continuous distribution to the vector of squared distances: take the available continuous distributions from Python's package SciPy, fit each distribution to the squared distances and apply a Kolmogorov-Smirnov goodness-of-fit test in order to rank the fitted distributions. Keep the best fit.

4. Calculate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the fitted distibution.

5. Mark the outliers as the data points that are beyond the quantiles obtained on step 4.

Results

In Figure 7 shows the scatter plot of the bivariate data used in this section.

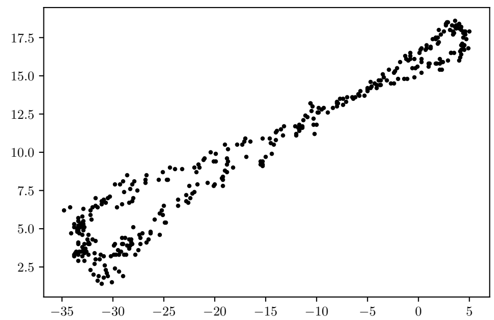
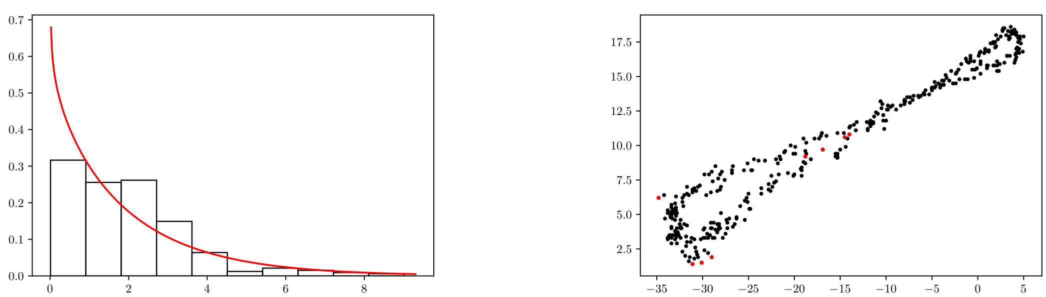


Figure 7: Scatter plot of bivariate data.

The first cut method was done with the usual estimation of the covariance matrix. In Figure 8, the results for this procedure are presented. The best-fit was a Mielke's Beta-Kappa distribution, whose probability density function(p.d.f.) is

$$f(t) = \frac{kt^{k-1}}{(1+t^s)^{1+\frac{k}{s}}}, \quad t > 0$$

with parameters $k = 0.96$, $s = 3.85$, loc=0.01 and scale=3.37, and it was not rejected with a p-value of 0.34.



(a) Fitted distribution.

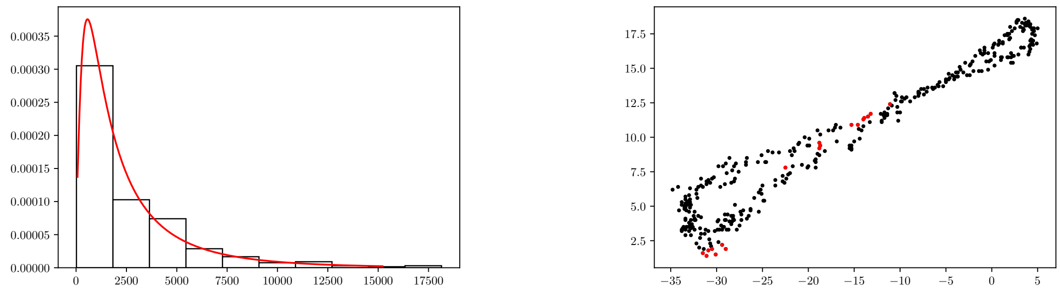
(b) Outliers cutout.

Figure 8: Results using the usual estimation.

On the other hand, the first robust approach is usig the Comedian procedure above described, where each entry is estimated using the comedian and the mean vector is replaced with the median vector. The results are shown in Figure 9. The best-fit was an inverse gaussian distribution, whose density is given by

$$f(t) = \frac{1}{\sqrt{2\pi t^3}} e^{-\frac{(t-\mu)^2}{2t\mu^2}}, \quad t > 0$$

with parameters $\mu = 1.11$, loc=-254.79 and scale=2667.77, and it not rejected with a p-value of 0.28.



(a) Fitted distribution.

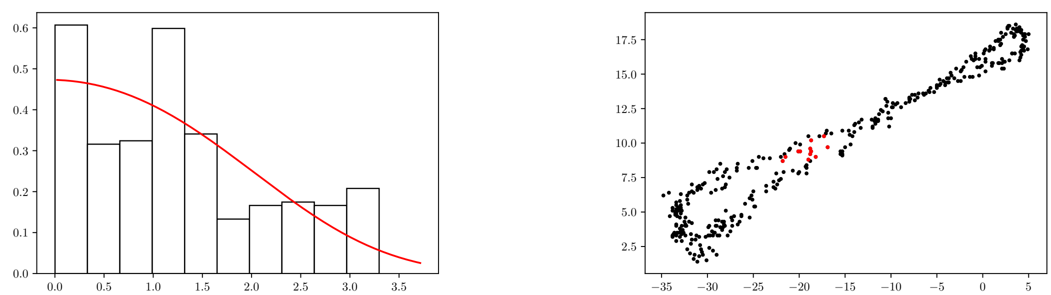
(b) Outliers cutout.

Figure 9: Results using the comedian estimation.

Furthermore, the same procedure was carried out using the covariance matrix estimation based on Kendall's τ correlation coefficient. Figure 10 shows the obtained results for the bivariate data. The best-fit was a folded normal distribution, whose p.d.f is given by

$$f(t) = \sqrt{\frac{2}{\pi}} \cosh{(ct)} e^{-\frac{(t^2+c^2)}{2}}, \quad t > 0$$

with parameters $\mu = 1.01$, loc=0 and scale=2.19, and it was not rejected with a p-value of 0.45.



(a) Fitted distribution.

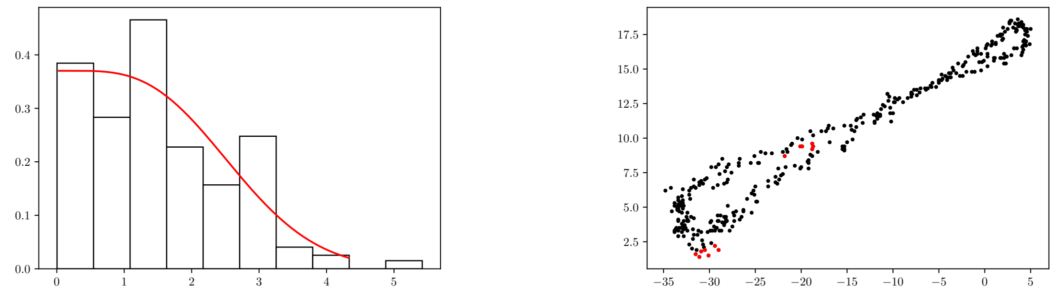
(b) Outliers cutout.

Figure 10: Results using Kendall estimation.

Finally, the last robust estimation of the covariance matrix was constructed using Spearman's correlation coefficient and the outlier detection results are presented in Figure 11. The best-fit was a Gompertz distribution, whose p.d.f. is given by

$$f(t) = ce^t e^{-ce^t - 1}, \quad t > 0$$

with parameters $c = 1.04$, loc=0 and scale=2.19, and it was not rejected with a p-value of 0.31.

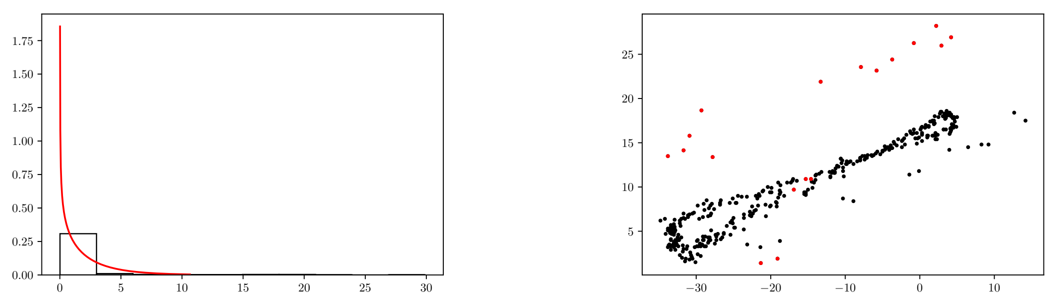


(a) Fitted distribution.

(b) Outliers cutout.

Figure 11: Results using Spearman estimation.

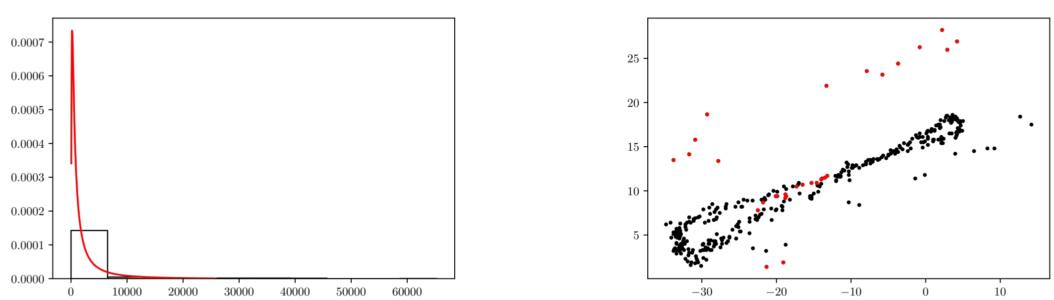
In the following results, the same procedures of outlier cutout is presented, but taking a "damaged" sample. This new sample takes 30 random bivariate observations and adds normally distributed observations, with mean 10 and standard deviation of 0.5. The results for the usual, the comedian, the Kendall and the Spearman cutouts are presented, respectively, in Figures 12, 13, 14 and 15. It is important to highlight that the fitting was done assuming the same distributions from the previous procedures.



(a) Fitted distribution.

(b) Outliers cutout.

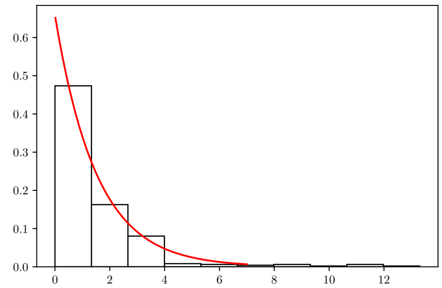
Figure 12: Results using the usual estimation on contaminated data.



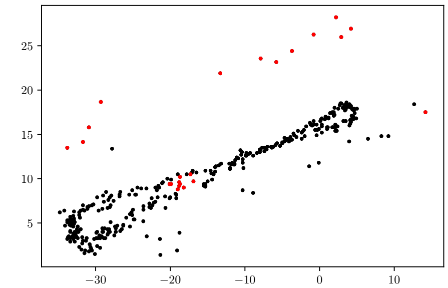
(a) Fitted distribution.

(b) Outliers cutout.

Figure 13: Results using the comedian estimation on contaminated data.

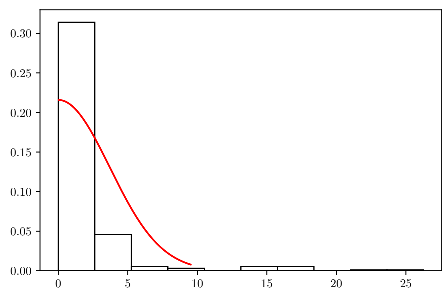


(a) Fitted distribution.

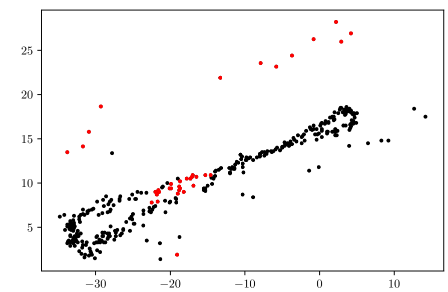


(b) Outliers cutout.

Figure 14: Results using Kendall estimation on contaminated data.



(a) Fitted distribution.



(b) Outliers cutout.

Figure 15: Results using Spearman estimation on contaminated data.

2 Book Exercises (Extra)

2.1 Page 41, Exercise 11

All exercises in this section are extracted from [8]. Let $X_1, ..., X_n \sim \text{Uniform}(0, 1)$, the maximum likelihood estimator (MLE) for $\theta = 1$ is $\hat{\theta} = X_{\max} = \max\{X_1, ..., X_n\}$. The distribution of $\hat{\theta}$ is a direct consequence of the proof made in Subsection 1 and it is given by

$$F(t) = t^n.$$

In Figure 16 shows the theoretical density of $X_{\max}$ and the approximation in histograms by parametric and nonparametric Bootstrap.

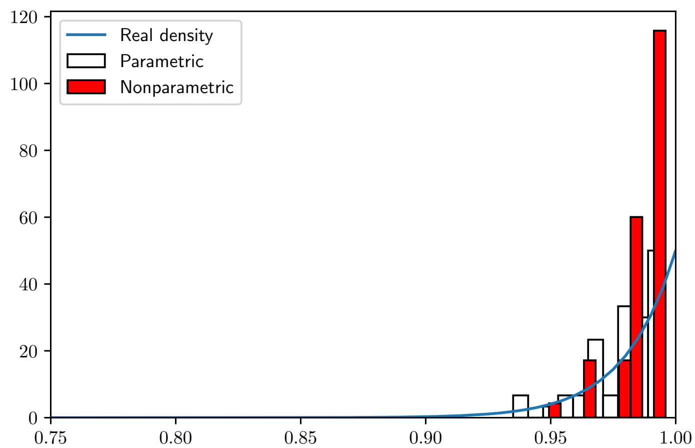


Figure 16: Density of $X_{\max}$, parametric and nonparametric histograms of $X_{\max}$.

Let $\hat{\theta}^*$ be the estimation of the parameter θ for each Bootstrap replication. For a parametric Bootstrap, one could draw samples from $F_{\hat{\theta}}$, instead of drawing samples from the empirial distribuion $\hat{F}_n$. Therefore, we would draw a N (number of Bootstrap iterations) new samples $X_1^j, ..., X_n^j \sim$ Uniform$(0, X_{\max})$, for $j = 1, ..., N$. Clearly, $P(\hat{\theta}^* = \hat{\theta}) = 0$, since it is the probability of a point on a continuous random variable, which is obvious that it has a probability measure 0.
On the other hand, for a nonparametric Bootstrap, the procedure would be to draw new samples from the empirical distribution, i.e. simply draw samples with replacement from the initial sample $X_1, ..., X_n$ and calculate the new estimation. In this case, $P(\hat{\theta}^* = \hat{\theta}) \approx 1 - (1 - \frac{1}{n})^n$, since the event $\hat{\theta}^* = \hat{\theta}$ could be interpreted as the probability of $\hat{\theta}^*$ appearing at least once within the j-th resample for the nonparametric Bootstrap.
Now, in order to calculate the true probabilty of the event $\hat{\theta}^* = \hat{\theta}$, take the limit as $n \to \infty$. It is well known that $(1 - \frac{1}{n})^n \to e^{-1}$ as $n \to \infty$. Then, $P(\hat{\theta}^* = \hat{\theta}) = 1 - e^{-1} \approx 0.632$.

References

[1] Spearman CE. "General Intelligence, Objectively Determined and Measured". In: *American Journal of Psychology* 15 (1904), pp. 201–293.

[2] Michael Falk. "On MAD and Comedians". In: *Annals of the Institute of Statistical Mathematics* 49.4 (1997), pp. 615–644.

[3] Maurice G Kendall. "A New Measure of Rank Correlation". In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[4] R.J. Kulperger. "Parametric and Nonparametric Bootstrap". In: *Handouts for Introduction to the Theory of Statistics* (2019). URL: http://fisher.stats.uwo.ca/faculty/kulperger/SS3858/Handouts/Bootstrap.pdf.

[5] John Nachbar. "Concave and Convex Functions". In: *Lecture Notes for Economics* 4111 (2018).

[6] Joseph Lee Rodgers. "The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy". In: *Multivariate Behavioral Research* 34.4 (1999), pp. 441–456.

[7] Aad W Van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge university press, 2000.

[8] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.