MAD 2020-21, Assignment 3

Bulat Ibragimov, Kim Steenstrup Pedersen

hand in until: 13.12.2021 at 23:59

General comments: The assignments in MAD must be completed and written individually. You are allowed (and encouraged) to discuss the exercises in small groups. If you do so, you are required to list your group partners in the submission. The report must be written completely by yourself. In order to pass the assignment, you will need to get at least 40% of the available points. The data needed for the assignment can be found in the assignment folder that you download from Absalon.

Submission instructions: Submit your report as a PDF, not zipped up with the rest. Please add your source code to the submission, both as executable files and as part of your report in appendix. To include it in your report, you can use the `lstlisting` environment in LaTeX, or you can include a "print to pdf" output in your pdf report. In some exercises we will ask you to include a code snippet as part of your solution text - a code snippet is only the most essential lines of code needed for solving the problem, this does not include import statements, other forms of boiler plate code, as well as plotting code.

Principal Components Analysis

Exercise 1 (**Implement PCA, 8 points**). See the Jupyter notebook file `pca_StudentVersion.ipynb` for the detailed questions and hints.

a) Implement PCA on the diatoms database. You are supposed to write your own implementation and thus is not allowed to use implementations found in various libraries. Please output the proportion of variance explained by each of the first 10 components (5 points)
b) Visualize fourth component of the PCA (3 points)

Deliverables. a) The jupyter notebook with your solution and write in the report the proportion of variance explained by the first 10 components, b) include the visualisations in the report and comment on the results.

Solution:
See the TA version of the Jupyter notebook file.
Instructions:

a) 1 point can be given if the student fail to compute PCA, but was trying to use the correct components: covariance matrix and eigh solver. 2 points can be deduced if student forgot to normalize data to the zero mean. 1 point can be deduced if variances are slightly miscalculated in the plot.
c) One point can be given if the student tried to combine two out of three components including 1) mean diatom shape; 2) eigenvalue 3) eigenvector, but failed to combine them properly. Two points can be given if student combined three components properly but did not use [-3,-2,...., 2, 3] mulitpliers properly.

Statistics

Exercise 2 (**2 points. (based on Blitzstein & Hwang Exercise 10.7.6) Inequalities).** Let X be a random variable with mean μ and variance σ^2. Show that

$$E\left[(X-\mu)^4\right] \geq \sigma^4 \ .$$

Hint: Consider if you can use Jensen's inequality.

Deliverables. Include the steps and argumentation of the proof of the expression.

Solution:

The solution uses Jensen's inequality and goes like this. Introduce an auxillary r.v. $Y = (X-\mu)^2$, then we have that $E[Y] = E[(X-\mu)^2] = \sigma^2$. Next we use Jensen's inequality on Y and the convex function $g(z) = z^2$ (they don't need to prove that g is convex) and get

$$E[g(Y)] \geq g(E[Y]) \Longrightarrow E[Y^2] \geq (E[Y])^2 \ .$$

Finally, we substitute the expression for Y into the right-hand version of Jensen's inequality and $E[Y] = \sigma^2$ to reach the answer:

$$E[\left((X-\mu)^2\right)^2] \geq (\sigma^2)^2 \Longrightarrow E[(X-\mu)^4] \geq \sigma^4 \ .$$

Instructions:
2 points - assign based on percentage of steps in derivation being carried out and how correct they are.

Exercise 3 (**4 points. Confidence Intervals).** Let $\gamma \in \mathbb{R}$ be fixed and let $X_1, \ldots, X_n$ be i.i.d. with Normal distribution $\mathcal{N}(\mu, \sigma^2)$. We estimate μ by the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. In the lecture, we have seen that

$$\sqrt{n}\,\frac{\hat{\mu}-\mu}{\sigma} \sim \mathcal{N}(0,1). \tag{1}$$

a) Pretend that (1) holds, even if we replace σ by the estimator $\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu})^2}$ (i.e. the sample standard deviation). Construct a γ-confidence interval for μ by using the procedure explained in the lecture.

b) We can make a simulation of many experiments at a fixed n, and compute the probability of the correct μ value being outside the confidence interval – if our estimated confidence interval is a good fit then this probability should be $< (1-\gamma)$. Modify the code in `confidenceinterv.py` (here, $n = 9$) and report, how often (out of 10000 experiments) the correct parameter lies outside the confidence interval. (Hint: Numpy's `np.var` divides by n, not by $n-1$. To correct for this, set the parameter `ddof=1` of `np.var`.)

c) In fact, (1) does not hold if we replace σ with $\hat{\sigma}$. Instead, we have

$$\sqrt{n}\,\frac{\hat{\mu}-\mu}{\hat{\sigma}} \sim t_{n-1}$$

where t_{n-1} is a student-t distribution with $n-1$ degrees of freedom. Again, report the corresponding confidence interval, modify the notebook, and report, how often the correct parameter is not covered. (Hint: Use `scipy.stats.t.ppf(q, n-1)` to compute the critical value by reverse look up in the CDF of the t distribution of $n-1$ degrees of freedom.)

Deliverables. a) Write the correct expression for the γ-confidence interval under these assumptions, b) your modified version of the code and your answer to the question, c) write the correct expression for the γ-confidence interval under these assumptions, and include your modified code and your answer to the question.

Solution:

a) The full derivation (this is not necessary for the solution) is (remember we are assuming Normal distribution and know variance, hence the interval is symmetric)

$$\begin{aligned} & \mathbb{P}(-c \leq \sqrt{n}\,\frac{\hat{\mu}-\mu}{\hat{\sigma}} \leq c) = \gamma \\ \Rightarrow \quad & \mathbb{P}(-(\hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n}}) \leq -\mu \leq c\frac{\hat{\sigma}}{\sqrt{n}} - \hat{\mu}) = \gamma \\ \Rightarrow \quad & \mathbb{P}(\hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n}} \geq \mu \geq \hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}}) = \gamma. \end{aligned}$$

and thus the confidence interval (specifying this is considered a correct answer) is

$$[\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}} \ , \ \hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n}}]$$

and the critical value is computed from the CDF of the Normal distribution $Phi(c)$

$$\Phi(c) = \frac{1+\gamma}{2} \Rightarrow c = \Phi^{-1}(\frac{1+\gamma}{2}) \ .$$

b) The simulation shows that this confidence interval does not have the correct coverage, see `confidenceinterv-solution.py`.

c) The same as for a) and the confidence interval is

$$[\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}} \ , \ \hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n}}]$$

and the critical value is computed from the CDF of the Student t distribution $F(c)$

$$F(c) = \frac{1+\gamma}{2} \Rightarrow c = F^{-1}(\frac{1+\gamma}{2}) \ .$$

For the code and answer, see `confidenceinterv-solution.py`.

Instructions:
a) 1 point b) 1 point c) 2 points (1 point the correct confidence interval, 1 point the code)

Exercise 4 (4 points. Hypothesis Testing). A scientist claims that he has found a single gene that has an influence on the flowering time of a plant. In order to see whether his claim is true, he obtains five pairs $(X_1, Y_1), \ldots, (X_5, Y_5)$ of two genetically identical replicates. In each second replicate $(Y_1, \ldots, Y_5)$ he has knockedout the gene. The following table shows the flowering time (in days).

Plant	1	2	3	4	5
Replicate 1 without knockout	4.1	4.8	4.0	4.5	4.0
Replicate 2 with knockout	3.1	4.3	4.5	3.0	3.5

Assume that the differences $X_i - Y_i$ (that is, flowering time replicate X minus flowering time replicate Y) are normally distributed with mean μ and variance σ^2.

a) Choose the null hypothesis and briefly justify your answer.
b) Perform the corresponding t-test to the level 0.05 (Hint: Use the "six steps" from the lecture).
c) Can the scientist change the test result by (illegally) copying the data set, that is, by writing down each data point k times and pretending that he has investigated $5 \cdot k$ independent pairs of plants? Justify your answer.

Deliverables. a) Your justified answer, b) explain the six steps you go through in the t-test and whether or not you can reject the null hypothesis, c) your justified answer.

Solution:

a)

$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0$$

with μ being the mean of $D_i = X_i - Y_i$. That is, the null hypothesis is that there is no difference in flowering time and the alternative hypothesis is that flowering time is different. Flowering of a plant is a complex procedure, furthermore, there are thousands of genes. If a scientist claims, he has found a gene with causal influence, one wants to be sure that this is correct. Thus, we want to bound the error probability of 'The gene has no measurable influence but we decided that it does.'

b) Based on the six steps from the slides (it is also acceptable if they use the steps from Kreyszig sec. 25.4, page 1059)

1. Model: $D_i \sim \mathcal{N}(\mu, \sigma^2)$
2. Null hypothesis: $H_0 : \mu = 0$
   Alternative: $H_A : \mu \neq 0$
3. Test statistic:

$$T = \frac{\sqrt{n}\bar{D}}{\hat{\sigma}}$$

Distribution of T under H_0: $T \sim t_{n-1} = t_4$.

4. Level: $\alpha = 0.05$
5. Rejection region:

$$\begin{aligned} K_1 &= (-\infty, -t_{n-1;1-\alpha/2}] \cup [t_{n-1;1-\alpha/2}; \infty) \\ &= (-\infty, -t_{4;0.975}] \cup [t_{4;0.975}; \infty) = (-\infty, -2.776] \cup [2.776; \infty) \end{aligned}$$

6. Computing T: With $n = 5$ we get

$$\begin{aligned} \bar{D} &= \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{5}(1 + 0.5 - 0.5 + 1.5 + 0.5) = 0.6 \\ \hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \\ &= \frac{1}{4}\left((0.4)^2 + (-0.1)^2 + (-1.1)^2 + (0.9)^2 + (-0.1)^2\right) \\ &= \frac{1}{4}\big(0.16 + 2 \cdot 0.01 + 1.21 + 0.81\big) = \frac{1}{4}2.2 = 0.55 \end{aligned}$$

Thus

$$T = \frac{\sqrt{5} \cdot (0.6)}{\sqrt{0.55}} \approx 1.81$$

test decision: $T \notin K_1 \Rightarrow H_0$ not rejected.

c) Yes.
The mean $\bar{D}_k$ does not change, the sd $\hat{\sigma}_k$ becomes smaller with increasing k:

$$\hat{\sigma}_k^2 = \frac{1}{nk-1}k(n-1)\hat{\sigma}_1^2 \leq \hat{\sigma}_1^2$$

(Note: $\hat{\sigma}_k^2 \to \hat{\sigma}_1^2$ for $k \to \infty$.) The test statistic $T = \sqrt{kn}\frac{\bar{D}_k}{\hat{\sigma}_k}$ follows a t-distribution with $k \cdot n - 1$ degrees of freedom and thus converges towards $z_{1-\alpha/2}$ with increasing k. Therefore, at some point, the test statistic exceeds $t_{k \cdot n-1,1-\alpha/2}$ and the test is rejected.
(NB: The assumption that the data are iid is certainly violated after copying the data.)
An alternative solution is to perform a simulation and show what happens when k increases.

Instructions:
a) 0.5 point, b) 2.5 point for specifiying the 6 steps and correct answer, c) 1 point for the correct and justified answer, give fractional point based on quality of justification (0.5 point for correct answer, 0.5 for justification).