Stat 3005 – Nonparametric Statistics

Assignment 1

Due: 27 September (Friday) at 1800

Fall 2024

Exercise 1.1 (★★☆ — Parametric and nonparametric procedures (60%)). Let $X_1, \ldots, X_n$ be the data that are generated from the signal-plus-noise model:

$$X_i = \mu_i + \varepsilon_i \qquad \text{where} \qquad i = 1, \ldots, n,$$

where $\mu_1, \ldots, \mu_n \in \mathbb{R}$ are some deterministic values, and $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{IID}}{\sim} [0, \sigma^2]$. The quantity of interest is σ^2.

1. In each of the following three cases, explain why the proposed estimator is a sensible estimator of σ^2. Determine whether the estimator is parametric or nonparametric, and use no more than 20 words to explain in each case.
   (a) (10%) Amy assumes that $\mu_1 = \cdots = \mu_n = 0$. Her proposed estimator of σ^2 is
   $$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$
   (b) (10%) Billy assumes that $\mu_1 = \cdots = \mu_n$ and $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$. His proposed estimator of σ^2 is
   $$\widehat{\sigma}_2^2 = \left\{ \frac{X_{(\lfloor 0.75n \rfloor)} - X_{(\lfloor 0.25n \rfloor)}}{2\Phi^{-1}(0.75)} \right\}^2,$$
   where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of $X_{1:n}$, and $\Phi^{-1}(\cdot) = \texttt{qnorm}(\cdot)$ is the quantile function of $\mathrm{N}(0, 1)$.
   (c) (10%) Cindy assumes that $\mu_i = g(i/n)$ for $i = 1, \ldots, n$, where $g : [0, 1] \to \mathbb{R}$ is an unknown differentiable function that is free of the data and n. Her proposed estimator of σ^2 is
   $$\widehat{\sigma}_3^2 = \frac{1}{2(n-1)}\sum_{i=2}^{n} (X_i - X_{i-1})^2.$$
2. Set $n = 30$. Consider the following three data generating processes (DGPs):
   - DGP (i): $X_i \overset{\text{IID}}{\sim} \mathrm{N}(10, 5^2)$ for $i = 1, \ldots, n$.
   - DGP (ii): $X_i \overset{\text{IID}}{\sim} \mathrm{Unif}(-1, 1)$ for $i = 1, \ldots, n$.
   - DGP (iii): $X_i \overset{\perp\!\!\!\perp}{\sim} 100(i/n - 0.5)^2 + t_4$ for $i = 1, \ldots, n$.

   Note that t_v denotes the Student's t distribution with v degree of freedom. We write $X \sim \mu + t_v$ if X can be represented as $X = \mu + Z$, where $Z \sim t_v$.
   (a) (10%) Write down the true value of σ^2 under each DGP.
   (b) (10%) Perform a simulation experiment to find out the root mean squared errors of $\log \widehat{\theta}_1$, $\log \widehat{\theta}_2$ and $\log \widehat{\theta}_3$ as estimators of $\log \sigma^2$ under each DGP. Use 2^{10} replications for each case. Report your results in the following form:

	$\mathsf{RMSE}(\log \widehat{\sigma}_1^2)$	$\mathsf{RMSE}(\log \widehat{\sigma}_2^2)$	$\mathsf{RMSE}(\log \widehat{\sigma}_3^2)$
DGP (i)			
DGP (ii)			
DGP (iii)			

   (c) (10%) Use no more than 100 words to describe your findings in the simulation study.

Hints: See Remark 1.1. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 1.2 (★★☆ — Simulation experiments (40%)). Assume the conditions in Theorem 2.2 of the lecture note. Suppose $n \geq 2$. Let

$$T = \sum_{i=2}^{n} \frac{i}{n+1}(R_i - R_{i-1}).$$

1. (20%) Prove that

$$\mathsf{E}(T) = 0 \qquad \text{and} \qquad \mathsf{Var}(T) = \frac{(n^2+n+2)n}{12(n+1)} - \frac{n^2+n+2}{12(n+1)^2(n-1)}\mathsf{E}(A), \quad \text{[Updated]}$$

   where A is defined in Theorem 2.2.
2. (20%) Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \lfloor 5\text{Exp}(1) \rfloor$, i.e., the distribution of the floor of an exponential RV with mean 5, where $\lfloor \cdot \rfloor$ is the floor function. For $n = 2, 4, 6, 8$, use simulation to verify your answer in part 1.

Hints: See Remark 1.2. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 1.3 (★★☆ — ✂ Properties of ranks). Let $X_1, \ldots, X_n$ be IID continuous RVs. Denote their ranks by $R_1, \ldots, R_n$.

1. Find $\mathsf{Cov}(R_1 + R_2, R_1 - R_2)$.
2. Are $R_1 + R_2$ and $R_1 - R_2$ independent?

Remark 1.1 (Hints for Exercise 1.1).

1. (a) Under the assumption, prove that $\sigma^2 = \mathsf{E}(X_1^2)$. Section 1.2 of the lecture note provide some guidelines for determining whether an estimator is parametric or nonparametric; see also Examples 1.1–1.2. In particular, how many parameters are in the model?

   (b) Let the common mean be μ, i.e., $\mu_1 = \cdots = \mu_n = \mu$. Argue that

$$X_{(\lfloor np \rfloor)} \xrightarrow{\text{pr}} \mu + \sigma\Phi^{-1}(p),$$

   where the probability limit is the pth quantile of $\text{N}(\mu, \sigma^2)$, and $p \in (0, 1)$; see Theorem 2.5.

   (c) Argue that

$$\mu_i - \mu_{i-1} = g(i/n) - g((i-1)/n) \approx g'(i/n)/n \approx 0,$$

   where g' is the derivative of g. Also, argue that

$$(X_i - X_{i-1})^2 = (\varepsilon_i - \varepsilon_{i-1})^2 + 2(\varepsilon_i - \varepsilon_{i-1})(\mu_i - \mu_{i-1}) + (\mu_i - \mu_{i-1})^2$$

   and that $\mathsf{E}(X_i - X_{i-1})^2 \approx 2\sigma^2$. Then use the method of moment to suggest an estimator for σ^2.

2. (b) The root mean squared error of $\log \widehat{\sigma}^2$ as an estimator of $\log \sigma^2$ is

$$\mathsf{RMSE}(\log \widehat{\sigma}^2) = \left\{\mathsf{E}(\log \widehat{\sigma}^2 - \log \sigma^2)^2\right\}^{1/2}.$$

   You may wish to use the following R-code.

```
nRep = 2^10
n = 30
out = array(NA, dim=c(nRep,3,3))
dimnames(out) = list(paste0("iRep=",1:nRep),
                     paste0("DGP",c("(i)","(ii)","(iii)")),
                     paste0("estimator",1:3))
# theta is a vector storing the true values of theta in DGP (i), (ii) and (iii)
estimand = c(5^2 ,    ,    )      # <<< Complete this line
for(iRep in 1:nRep){
    for(iDGP in 1:3){
        if(iDGP==1) x = rnorm(n,10,5)
        if(iDGP==2) x =                    # <<< Complete this line
        if(iDGP==3) x =                    # <<< Complete this line
        out[iRep,iDGP,1] = mean(x^2)
        out[iRep,iDGP,2] =                 # <<< Complete this line
        out[iRep,iDGP,3] =                 # <<< Complete this line
        out[iRep,iDGP,] = log(out[iRep,iDGP,])-log(estimand[iDGP])
    }
}
sqrt(apply(out^2,2:3,mean))
```

Remark 1.2 (Hints for Exercise 1.2)**.**

1. Similar ideas are also used in Example 2.5 in the lecture note. You may also need to use the results in Theorem 2.2. You may wish to prove that

$$T = \frac{1}{n+1}\left(nR_n - 2R_1 - \sum_{i=2}^{n-1} R_i\right).$$

2. The value of $n_1, \ldots, n_k$ could be computed by `table(x)`. You may wish to use the following R-code.

```
nRep  = 2^12
n =                                 # <<< Complete this line
out = array(NA,dim=c(nRep,length(n),2))
dimnames(out) = list(paste0("iRep=",1:nRep),paste0("n=",n),c("T","A"))
for(iRep in 1:nRep){
    for(i.n in 1:length(n)){
        n0 = n[i.n]
        x =                         # <<< Complete this line
        R =                         # <<< Complete this line
        nj =                        # <<< Complete this line
        A =                         # <<< Complete this line
        out[iRep,i.n,"T"] =         # <<< Complete this line
        out[iRep,i.n,"A"] = A
    }
}
m  = apply(out[,,"T"],2,mean)
v  = apply(out[,,"T"],2,var)
m0 = 0
v0 = (n^3+2*n^2-3*n+6) * (1/(12*(n+1)) - 1/(12*(n+1)^2*(n-1)*n) * apply(out[,,"A"],2,mean))
...                                 # <<< Finish the remaining parts of the program
```

Instructions

1. There are two types of questions.
   - Compulsory questions are marked with percentage points. Complete all of them.
   - Optional questions (if any) are marked with ✂. Complete them to get at most $b = 0.5$ bonus point.
2. The formats of your submitted file:
   - For the written parts, either (i) handwrite and scan your answers, or (ii) type your answers in LaTeX. Use A4-size papers. Submit your answers in one single `.pdf` file. Name the file in the form: `S3005_A#_sid_name.pdf`. E.g., `S3005_A3_1155001234_ChanKinWai.pdf` for assignment 3.
   - For the programming parts, solely use `R` to complete the problems. Submit one single `.R` file. Name the file in the form: `S3005_A#_sid_name.R`.
   - If the question requires you to produce graphs, compute values or perform experiments, you must include the results in the `.pdf` file. Graders may not compile your codes and grade the corresponding outputs.
   - For the optional parts (if any), there is no special format unless otherwise stated. Additional files (in any format) could be submitted for supplementary purpose.
   - Submission with format other than `.pdf` or `.R` may not be graded.
3. Submit your completed assignment to Blackboard. Compulsory and optional parts should be submitted to two separate submission links. Email submission is not accepted in general.
4. You may submit your work as many time as you wish, however, only the last submission will be graded.
5. You may discuss the assignment questions with others. However, your final answers must be written by you and fully reflect your own intellectual effort. Please also list the names of your collaborators on the first page of your assignments. No credit will be deducted for appropriate collaborations. All answers will receive a grade of zero without proper declaration of collaborators and may be subject to disciplinary action.
6. Generative AI tools (e.g., ChatGPT) can be used for auxiliary purposes. However, your final answers must be written by you and fully reflect your own intellectual ideas and independent effort. The graders have the ultimate authority to decide the appropriateness and level of originality of the submitted answers. Please also list the names of the AI tools you have used on the first page of your assignments. No credit will be deducted for appropriate use of AI. All answers will receive a grade of zero without proper declaration of AI usage and may be subject to disciplinary action.
7. Please always read the assignment questions stated in the `.pdf` file downloaded from the course website.
8. Late submission will receive the following penalty:
   - For late submission on the same day, the total score will be multiplied by 0.95.
   - For late submission on any day after the deadline, the total score will be multiplied by 0.9^t where t is the number of days after the deadline.
   - Submission after the solution is posted will receive zero.