Exam-2019-20 with solutions

Problem 1
Let $\mathbf{X} = (X_1, \ldots, X_d)$ be multivariate normally distributed with unknown expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and known covariance matrix Σ. The d hypotheses are

$$H_1 : \mu_1 \leq 0, \ldots, H_d : \mu_d \leq 0.$$

Based on a sample $(X_{11}, \ldots, X_{1d}), \ldots, (X_{n1}, \ldots, X_{nd})$ of size n from $\mathbf{X}$ we reject hypothesis i, $1 \leq i \leq d$, at level α if

$$\sum_{\ell=1}^{n} X_{\ell i} > \sqrt{n}\, \sigma_i q_{1-\alpha},$$

where σ_i is the standard deviation of X_i and $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal.

(i) For $d = 15$ which multiple testing procedure could you use to test the d hypotheses?

(ii) For $d = 500$ which multiple testing procedure would you use to test the d hypotheses?

Please motivate your answers in (i) and (ii) briefly.

Solution: (i) Because $d = 15$ is relatively small we could use the traditional procedures of Bonferroni and Holm because the assumption underlying these procedures is fulfilled in the setting considered here; see Exercise 7. Preference would be given to Holm's procedure as it controls the same criterion but has a higher power.
We could also use a k-FWER procedure with $k \geq 2$ as the underlying assumption is the same. Although k-FWER procedures with k bigger than 1 were designed for large d, they can be applied for d small as well. The only difference is that they control a different criterion than the traditional procedures, i.e. Bonferroni and Holm.
Moreover, we could use Benjamini and Yekutieli as it works whatever the unknown dependence structure is. Notice that Benjamini and Hochberg cannot be used because it requires independence or the PRDS property which may hold here or not but we cannot decide based on the information given.

(ii) As $d = 500$ is pretty large we should not use Bonferroni or Holm even if the assumption is fulfilled because their power is so small for $d = 500$ that we would not be able to detect deviations from the null. We could use either Benjamini and Yekutieli or k-FWER with d bigger than 1. But not Benjamini and Hochberg for the reasons outlined above.

Problem 2
Assume that X is exponentially distributed with parameter λ, i.e. the cumulative distribution function of X is given by

$$F_\lambda(x) = 1 - \exp(-\lambda\, x),\ x > 0,\ \text{and zero otherwise.}$$

For testing

$$H : \lambda \leq 1, \quad A : \lambda > 1,$$

we use a sample X of size 1. Our test statistic is then simply $T(X) = X$ and our p-value is simply $\hat{p}(X) = 1 - \exp(-X)$.

(i) Show that $\hat{p}(X)$ is uniformly distributed on $(0,1)$ if the cumulative distribution function of X is

$$F_1(x) = 1 - \exp(-x),\ x > 0,\ \text{and zero otherwise.}$$

(ii) Find the cumulative distribution function of $\hat{p}(X)$ if the cumulative distribution function of X is F_2.

Solution: (i) Note first that $\hat{p}(X) = 1 - \exp(-X)$ takes only values from the interval (0,1) because X takes values in $(0, \infty)$. We have for $t \in (0,1)$

$$\mathbb{P}(1 - \exp(-X) \leq t) = \mathbb{P}(X \leq -\log(1-t)) = 1 - \exp(-(-\log(1-t))) = t.$$

(ii) We have almost as before

$$\mathbb{P}(1 - \exp(-X) \leq t) = \mathbb{P}(X \leq -\log(1-t)) = 1 - \exp(-2(-\log(1-t))) = 1 - (1-t)^2.$$

Problem 3

Assume that Y_i, $1 \leq i \leq n$, is Poisson distributed given the explanatory variables x_{i1}, x_{i2} and x_{i3}, $1 \leq i \leq n$, with parameter $\lambda_i = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$, $1 \leq i \leq n$. For a sample $y_1, \ldots, y_n$ and explanatory variables $((x_{11}, x_{12}, x_{13}), (x_{21}, x_{22}, x_{23}), \ldots, (x_{n1}, x_{n2}, x_{n3}))$ give the log-likelihood function and find the first order conditions for β_1, β_2, and β_3.

Solution: The general case was discussed in lecture 6. Here we have $d = 3$.

Problem 4

In Problem 3 we related λ_i, i.e. the expectation of Y_i, to $\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ using the function $h : \mathbb{R} \to \mathbb{R}_+$ defined by

$$h(x) := \exp(x).$$

For each of the following alternative choices, argue if it is meaningful to use them instead of the above h. Explain your answer.

(i) $h_1(x) = x^3$, $x \in \mathbb{R}$;

(ii) $h_2(x) = \frac{x^2}{1+|x|}$, $x \in \mathbb{R}$.

Solution: (i) it not a valid choice as the link function can take negative values in contrast to the parameter of a Poisson distribution.
(ii) h_2 maps to $[0, \infty)$. Arguing that it is not a good choice because it can be equal to zero whereas $\lambda > 0$ would have been ok. Arguing that it is a suitable choice would have been ok too. In class we consider the link function $h(x) = x^2$ which has the 'same problem'. Yet the idea is if we have random design with continuous covariates and non-degenerate distribution (i.e. here we have a density) the probability that $\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ equals zero is zero.

Problem 5

Assume our data come from the linear model

$$Y_i = \sum_{j=1}^{d} \beta_j X_{ij} + \epsilon_i,\ i = 1, \ldots, n,$$

with ϵ_i, $1 \leq i \leq n$, independent and normally distributed with expectation zero and variance σ^2. Consider the following choices for d as a function of the sample size

(i) $d = n^n$;

(ii) $d = n^{15}\log(n)$;

(iii) $d = \sqrt{n}\,\exp\left(n^{0.8}\right)$.

For which of these choices do we have consistency of the LASSO estimator as $n \to \infty$ (you can assume that the design matrix fulfills the restricted eigenvalue condition and that β has only k non-zero entries for all d and n)?

Please motivate your answers in (i), (ii) and (iii) briefly.

Solution: In all three cases we need to check if

$$\sqrt{\frac{k\log(d)}{n}} \to 0 \left(\text{ or } \frac{k\log(d)}{n} \to 0 \right) \text{ as } n \to \infty, \text{ and } \log(d) \to \infty \text{ as } n \to \infty;$$

(see Theorem (upper bound on error of LASSO probabilistic version)) and notice that there σ and γ are only constants and we can choose $\tau > \sqrt{8}$.

(i) Since $\log(d) = \log(n^n) = n\log(n)$ and $k\log(n) \to \infty$ as $n \to \infty$ we do not have consistency according to the Theorem.

(ii) As $\log(d) = \log(n^{15}\log(n)) = 15\log(n) + \log(\log(n))$ and $\log(n)/n \to 0$ as $n \to \infty$ (which implies $\log(\log(n))/n \to 0$ as $n \to \infty$) the first condition is met. For the second condition it is enough to note that $d > 15\log(n) \to \infty$ as $n \to \infty$.

(iii) As $\log(d) = 0.5\log(n) + n^{0.8}$ and $n^{0.8}/n = n^{-0.2} \to 0$ as well as $\log(n)/n \to 0$ as $n \to \infty$ the first condition is met. Clearly $\log(d) > n^{0.8} \to \infty$ as $n \to \infty$ such that the second condition is met too.

Problem 6

Consider the following set-up:

- X_i, $1 \leq i \leq 50$, is binomially distributed with success probability $p_i = p$, $1 \leq i \leq 50$;
- The hypotheses H_i, $1 \leq i \leq 50$, are $H_i : p_i = 0.5$;
- From each X_i we have a sample $(X_{i1}, \ldots, X_{i10})$ of size 10, and reject H_i if

$$\left| \frac{\sqrt{10}}{0.5} (\hat{p}_i - 0.5) \right| > 1.645,$$

  where $\hat{p}_i = \frac{1}{10}\sum_{j=1}^{10} X_{ij}$.
- If H_i is rejected we construct the following confidence interval for p_i

$$CI_i = \left[\hat{p}_i - 1.645\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{10}}; \hat{p}_i + 1.645\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{10}} \right].$$

Assume $p = 0.6$, put $S = \{i \in \{1, \ldots, 50\} \,|\, H_i \text{ was rejected } \}$ and let $p_S = (p_{i_1}, \ldots, p_{i_{|S|}})$, $i_1 < \ldots < i_{|S|}$, $i_j \in S$, be the vector of selected parameters (selected here means associated with the

hypotheses rejected). Write pseudo code (or code) that could be used to calculate the probability that p_S is contained in $CI_{i_1} \times \ldots \times CI_{i_{|S|}}$.

Solution:

```
sum=0
for (i in 1:M){
p=rep(0,50)
sumcovered=0
ind=c()
for (j in 1:50){
p[j]=(1/10)*rbinom(1,10,0.6)
if (abs((p[j]-0.5)*10^0.5/0.5)>1.645){
ind=cbind(ind,j)
}
}
for (k in ind){ if (0.6>=(p[k]-(1/10^0.5)*1.645*(p[k]*(1-p[k]))^0.5) & 0.6 <= (p[k]+(1/10^0.5)*1.645*(p[k]*(1-p[k]))^0.5)){
sumcovered=sumcovered+1
}
}
if (sumcovered==length(ind)){
sum=sum+1
}
print(sum/M)
}
```

Comments:

1. line 1: Initializing sum which will count how often p_S is contained in $CI_{i_1} \times \ldots \times CI_{i_{|S|}}$
2. line 2: M number of repetitions;
3. line 3: initializing $\hat{p}$;
4. line 5: vector that will be used to store indices of selected parameters;
5. line 7: generating the $\hat{p}$'s;
6. line 8 and 9 checking whether we reject and if we reject we add the associated index to ind;
7. line 12, 13, and 14 checking whether the true parameter is covered in case of rejection and if it is we increase sumcovered by 1;
8. line 17, 18 checking if all selected parameters were covered and if this is true we increase sum by 1;
9. line 20 diving all coverages by number of repetitions.