Homework #01

SMDS-2022-2023

M.Sc. in Data Science

A. Simulation

1. Consider the following joint discrete distribution of a random vector (Y, Z) taking values over the bi-variate space:

$$\mathcal{S} = \mathcal{Y} \times \mathcal{Z} = \{(1,1);(1,2);(1,3);\\ (2,1);(2,2);(2,3);\\ (3,1);(3,2);(3,3)\}$$

The joint probability distribution is provided as a matrix J whose generic entry `J[y,z]` $= Pr\{Y = y, Z = z\}$

```
J
```

```
     1    2    3
1 0.06 0.17 0.10
2 0.10 0.12 0.11
3 0.14 0.02 0.18
```

```
S
```

```
      row col
(1,1)   1   1
(1,2)   1   2
(1,3)   1   3
(2,1)   2   1
(2,2)   2   2
(2,3)   2   3
(3,1)   3   1
(3,2)   3   2
(3,3)   3   3
```

You can load the matrix `S` of all the couples of the states in $\mathcal{S}$ and the matrix `J` containing the corresponding bivariate probability masses from the file "Hmwk.RData". How can you check that J is a probability distribution?

To check that J is a probability distribution we simply have to check the following property: $\sum_{(y,z)\in\mathcal{S}} J(y,z) = 1$

```
load("Hmwk.RData")
# print the sum of all the elements of the J matrix
sum(J)
```

```
[1] 1
```

2. How many *conditional distributions* can be derived from the joint distribution `J`? Please list and derive them.

:

For *discrete distributions* we can derive the *marginal distributions* for a specific variable directly from the *joint distribution* in the following way:

> - $p_Y(y) = \sum_{z\in\mathcal{Z}} J(y,z)$
> - $p_Z(z) = \sum_{y\in\mathcal{Y}} J(y,z)$

The joint distribution is bivariate and I can derive the marginal distributions of the two variables Y and Z. Then I evaluate the conditional distributions of the two variables each represented by a 3×3 matrix using the joint and the marginals:

> - $p_{Y|Z}(y|z) = \frac{J(y,z)}{p_Z(z)}$
> - $p_{Z|Y}(z|y) = \frac{J(y,z)}{p_Y(y)}$

```
   p.Y  p.Z
1 0.33 0.30
2 0.33 0.31
3 0.34 0.39
[1] "---------------------------"
[1] "p.Z.given.Y"
          1          2         3
1 0.1818182 0.51515152 0.3030303
2 0.3030303 0.36363636 0.3333333
3 0.4117647 0.05882353 0.5294118
[1] "---------------------------"
[1] "p.Y.given.Z"
          1          2         3
1 0.2000000 0.54838710 0.2564103
2 0.3333333 0.38709677 0.2820513
3 0.4666667 0.06451613 0.4615385
```

3. Make sure they are probability distributions.

```
[1] "verify that each row of the p.Z.given.Y matrix sums to 1:"
1 2 3
1 1 1
[1] "verify that each column of the p.Y.given.Z matrix sums to 1:"
1 2 3
1 1 1
```

4. Can you simulate from this `J` distribution? Please write down a working procedure with few lines of R code as an example. Can you conceive an alternative approach? In case write down an alternative working procedure with few lines of R

```
# the code is also available in the "scripts.R" file but I prefer report it explicitly in markdown

# first simulation method
sim1 <- function(J, n){
  probs <- c(J) # flattening the probability matrix
  # well define the support from which we sample
  support <- as.data.frame(t(expand.grid(y=1:3, z=1:3)))
  samples.list <- as.data.frame(t(sample(support, n, replace=T, prob=probs))) # sample
  rownames(samples.list) <- NULL # reset the row indexes
  return(samples.list)
}

# second simulation method
sim2 <- function(J, n){
  distros <- derive.distr(J) # derive marginals and conditionals
  p.Y <- distros$p.Y # distr. of Y
  p.Z.given.Y <- distros$p.Z.given.Y # distr. of Z|Y
  y.sample <- sample(1:3, n, replace=T, prob=p.Y) # sample y from the marginal of Y
  z.sample <- rep(NA,n) # init the z sample
  for(idx in 1:n){
    y <- y.sample[idx]
    p.Z.given.y <- p.Z.given.Y[,y] # choose the correct conditional distr.
    # sample from the conditional distr. given the sampled Y=y
    z.sample[idx] <- sample(1:3, 1, prob=p.Z.given.y)
  }
  sample.list <- data.frame(y=y.sample, z=z.sample)
  return(sample.list)
}
```

B. Bulb lifetime: a conjugate Bayesian analysis of exponential data

You work for Light Bulbs International. You have developed an innovative bulb, and you are interested in characterizing it statistically. You test 20 innovative bulbs to determine their lifetimes, and you observe the following data (in hours), which have been sorted from smallest to largest.

Based on your experience with light bulbs, you believe that their lifetimes Y_i can be modeled using an exponential distribution conditionally on θ where $\psi = 1/\theta$ is the average bulb lifetime.

1. Write the main ingredients of the Bayesian model.

2. Choose a conjugate prior distribution $\pi(\theta)$ with mean equal to 0.003 and standard deviation 0.00173.

In order to choose a proper prior distribution we can start by providing a short analysis of the *likelihood function*. Moreover we have to assume the independence of the bulbs lifetimes' distributions when conditioned on θ, i.e: consider the bulbs lifetimes as random variables $Y_i|\theta \sim Exp(\theta), i = \{1,2,\dots,20\}$ we assume that $f_{Y_1,\dots,Y_{20}|\theta=\theta_0}(y_1,\dots,y_{20}) = \prod_{i=1}^n f_{Y_i|\theta=\theta_0}(y_i)$.

$$L_{Y_1,\dots,Y_{20}}(\theta) = \prod_{i=1}^n f_{Y_i|\theta=\theta_0}(y_i) =\\ = \prod_i \theta e^{-\theta y_i} =\\ = \theta^n \cdot e^{-(\theta\sum_i y_i)}$$

We can easily note a similarity with a well known distribution: the *Gamma*, which is characterized by the generic shape $f_\theta = g(c_1,c_2)\cdot\theta^{c_1}e^{-c_2\theta}$. With this naive intuition I decide to use a Gamma distribution as prior conjugate to the exponential distribution (later I'll provide a complete proof about this choice).

Now I can proceed solving the system of two equations in order to satisfy the requested properties about mean and standard deviation (having selected the distribution):

$$\begin{cases}\alpha/\beta = 0.003\\ \sqrt{\alpha}/\beta = 0.00173\end{cases} \leftrightarrow \begin{cases}\alpha = (0.003/0.00173)^2\\ \beta = 0.003/(0.00173)^2\end{cases} \rightarrow \theta \sim Gamma(3.007, 1002.372)$$

3. Argue why with this choice you are providing only a vague prior opinion on the average lifetime of the bulb.

In the exercise 4 is reported a proof about the conjugacy class gamma to the exponential likelihood and the resulting updated hyperparameters of the posterior $\pi(\theta|y_1,\dots,y_2)$ given by the formula :

$$\begin{cases}\alpha^* = \alpha + n\\ \beta^* = \beta + \sum_i y_i\end{cases}$$

Now I can proceed with the analysis of some of the features of the updated θ distribution to make some consideration about the obtained result.

First of all notice that the gamma distribution is the one that I choose for θ and for $\psi = 1/\theta$ (that parametrizes the mean of the exponential distr.) it can be easily proved that the equivalent distribution is an *Inverse-Gamma*, thus $\psi \sim InvGamma(\alpha+n, \beta+\sum_i y_i)$.

Explicitly deriving the expected value for ψ it can be shown that it is a convex combination of the sample mean (corresponding to the MLE) and the mean of the prior Inverse-Gamma ($\frac{\beta}{(\alpha-1)}$):

$$E(\psi|y_1,\dots,y_n) = \frac{\beta^*}{\alpha^*-1} = \frac{\beta+\sum_i y_i}{\alpha+n-1} =\\ = \frac{\beta}{\alpha+n-1}\cdot\frac{\sum_i y_i}{\alpha+n-1} =\\ = \frac{(\alpha-1)}{(\alpha+n-1)}\cdot\frac{\beta}{(\alpha-1)} + \frac{n}{\alpha+n-1}\cdot\bar{y}$$

Note that for $n \to \infty$,$E(\psi) \to \bar{y}$ going back to the frequentist framework and obtaining the *maximum likelihood estimator* itself for the mean value of an exponential distribution. In this way the two results/frameworks are comparable in some way, in particular, considering the weights $w_1 = \frac{(\alpha-1)}{(\alpha-1)+n}$ and $w_2 = \frac{n}{(\alpha-1)+n}$ it is possible to control the effect of our prior belief and how much it actually impacts on the learning process. In this case the exercise opts for fairly weak assumptions given that $w_1 << w_2$.
In general, a Bayesian estimator may be preferred to the maximum likelihood estimator because of its lower estimation variability. In fact, although the former is biased, the latter suffers from higher variance and the overall performance in terms of MSE may suffer greatly. The difference between the two performances comes from a number of factors including: sample size and precision of the a priori assumptions.
In this case the prior beliefs about the average lifetime are too vague since the variance of the estimator of the mean seems to be larger than the variance of the sample mean. This could lead to an increase in Loss by introducing an estimator with high variability and biased at the same time! In point 5 of this homework I'll explicitly evaluate the variance of the bayesian estimator of ψ, here I provide a qualitative comparison between ψ_{Bayes} and ψ_{mle} estimators:

$$Bias[\psi_{Bayes}] = w_1\frac{\beta}{(\alpha-1)} + (w_2-1)\psi_{true} \rightarrow biased\\ Bias[\psi_{mle}] = \psi_{true} - \psi_{true} = 0 \rightarrow unbiased$$

As demonstrated the Bayesian estimator must justify a loss in performance in terms of precision (bias) with better variance than other estimators so as to reduce this component of the MSE.

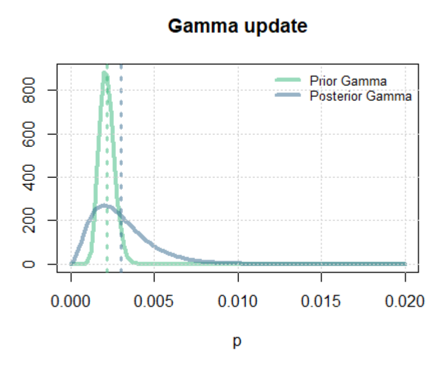
4. Show that this setup fits into the framework of the conjugate Bayesian analysis.

In order to show the fit of this setup into the conjugate Bayesian framework I can plug-in the selected prior distribution into the Bayes' formula and verify that it returns a posterior distribution of the same type (a Gamma).

$$\pi(\theta|y_1,\dots,y_n) = \pi(\theta)\cdot\frac{f_{Y_1,\dots,Y_n}(y_1,\dots,y_n|\theta)}{f_{Y_1,\dots,Y_n}(y_1,\dots,y_n)} =\\ = \theta^{(\alpha-1)}\cdot e^{-\beta\theta}\cdot\theta^n\cdot e^{(-\theta\sum_i y_i)}\cdot c(y,\alpha,\beta)\\ \rightarrow \pi(\theta|y_1,\dots,y_n) \propto \theta^{(\alpha+n-1)}\cdot e^{-(\beta+\sum_i y_i)\theta}$$

Recall that the shape of the gamma distribution (in the shape-rate reparametrization) is $f_\theta \propto \theta^{c_1-1}\cdot e^{-c_2\theta}$ and, looking at the above proportionality equation, recognize newely a gamma distribution as posterior with updated parameters ($\alpha^* = \alpha+n, \beta^* = \beta+\sum_i y_i$).

To summarize, if we look at a sample of random variables distributed as an exponential and consider a gamma prior distribution for the parameter of the exponential, then the posterior will also be a gamma distribution.



5. Based on the information gathered on the 20 bulbs, what can you say about the main characteristics of the lifetime of your innovative bulb? Argue that we have learnt some relevant information about the θ parameter and this can be converted into relevant information about the unknown average lifetime of the innovative bulb $\psi = 1/\theta$.

First, from the obtained information, it is possible to make an estimate of the average life time of a bulb based on the expected value of the mean ψ reported and analyzed in Step 3 of this exercise. Given observations above $y_1,\dots,y_{20}$ we know that $E(\psi|y_1,\dots,y_{20}) = 481.72$. The simple sample mean is $\hat{\psi}_{mle} = 479.95$: it is slightly different from the previous conditional expectation (but not that much given the weak a priori assumptions) but these results would converge for larger and larger sample sizes.

I report further information like the variance (that I mentioned earlier):

$$Var(\psi|y_1,..,y_{20}) = \frac{(\beta^*)^2}{(\alpha^*-1)^2(\alpha^*-2)} = \frac{(\beta+\sum_i y_i)^2}{(\alpha+n-1)^2(\alpha+n-2)} = 11046.85$$

So the variance literally explodes (more than I expected) and it seems that our result can vary extremely for different observations showing high uncertainty in the estimation of θ and ψ!

One last information never touched in the previous Steps is the *posterior predictive distribution* which can be derived using all the other components of this Bayesian conjugate analysis :

$$f_{Y_{new}}(y_{new}|y_{obs}) = \int_0^\infty f(y_{new}|\theta,y_{obs})\cdot f(\theta|y_{obs})d\theta =\\ = \int_0^\infty f(y_{new}|\theta)\cdot f(\theta|y_{obs})d\theta =\\ = \int_0^\infty dexp(y_{new},\theta)\cdot dgamma(\theta,\alpha+\beta,\beta+\sum_i y_i)d\theta =\\ = \int_0^\infty \theta e^{-\theta y_{new}}\cdot\frac{(\beta+\sum_i y_i)^{\alpha+n}}{\Gamma(\alpha+n)}\cdot\theta^{(\alpha+n+1)}\cdot e^{-(\beta+\sum_i y_i)\theta}d\theta =\\ = \frac{(\beta+\sum_i y_i)^{(\alpha+n)}}{\Gamma(\alpha+n)}\int_0^\infty \theta^{(\alpha+n+2)}\cdot e^{-(y_{new}+\sum_i y_i+\beta)\theta}d\theta =\\ = \frac{(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)}\cdot\frac{\Gamma(\alpha^*+1)}{(y_{new}+\beta^*)^{(\alpha^*+1)}} =\\ \propto (\beta^*+y_{new})^{-(\alpha^*+1)} =\\ = (\sum_i y_i+\beta+y_{new})^{-(\alpha+n+1)}\\ \rightarrow y_{new}|y_{obs} \sim ParetoII(\alpha+n, \sum_i y_i+\beta)$$

6. However, your boss would be interested in the probability that the average bulb lifetime $1/\theta$ exceeds 550 hours. What can you say about that after observing the data? Provide her with a meaningful Bayesian answer.

I'm simply going to evaluate in R the following value:

$$P(\psi > 550) = 1 - P(\psi \le 550) = 1 - pinvgamma(550, \alpha^*, \beta^*)$$

Where the parameters for the Inverse-Gamma are the same as those of the posterior Gamma:

$$\begin{cases}\alpha^* = \alpha+n = 23.007\\ \beta^* = \beta+\sum_i y_i = 10601.372\end{cases}$$

```
[1] "The probability that the average bulb lifetime exceeds 550 hours is :"
[2] "0.22541922584114"
```

I therefore infer a low probability of encountering a bulb with the average lifetime that exceeds 550 hours.

C. Exchangeability

Let us consider an infinitely exchangeable sequence of binary random variables $X_1, \dots, X_n, \dots$

1. Provide the definition of the distributional properties characterizing an infinitely echangeable binary sequence of random variables $X_1, \dots, X_n, \dots$. Consider the De Finetti representation theorem relying on a suitable distribution $\pi(\theta)$ on $[0,1]$ and show that

$$\begin{aligned} E[X_i] &= E_\pi[\theta] \\ E[X_i X_j] &= E_\pi[\theta^2] \\ Cov[X_i X_j] &= Var_\pi[\theta] \end{aligned}$$

A stichastic process $X_1, .., X_n, \dots$ is *infinitely exchangeable if we can take for each tuple $(n_1, \dots, n_k)$ and any permutation of the first k integers $\sigma = (\sigma_1, \dots, \sigma_k)$ the following rule holds: $(X_{n_1}, .., X_{n_k})$ have the same distribution of $(X_{\sigma_1}, .., X_{\sigma_k})$.
This condition means that the order of the observation of a sequence of random variables has no role in the definition of the joint distribution of the sequence itself and a second distributional property implied by the exchangeability condition is the fact that $X_1, .., X_n, \dots$ are identically distributed and we have a sort of conditional independence of $X_1, .., X_n, \dots$ given θ.
It can be proved that $\overline{X_n} = \frac{\sum_i X_i}{n} \approx \pi(\theta)$.
Throughout the lectures we have shown that the beta is a conjugate distribution of the binomial. In this case it would be reasonable to choose as the density $\pi(\theta)$ a beta.

1. Let's verify the reported properties:

> - by assumption $X_i \sim Ber(\theta)$, thus it's easy to prove that $E_{X_i|\theta}[X_i|\theta] = \theta$ and the *law of total expectation* claims that
>
> $$E_{X_i}[X_i] = E_\theta[E_{X_i|\theta}[X_i|\theta]] = E_\theta[\theta]$$
>
> - the exchangeability condition implies that $X_1, .., X_n|\theta$ are i.i.d and this in turn implies that:

$$\begin{aligned} &E[X_i \cdot X_j|\theta] = E[X_i|\theta] \cdot E[X_j|\theta] = \theta^2, \forall i \neq j \\ &\rightarrow E_{X_i X_j}[X_i X_j] = E_\theta[E_{X_i X_i|\theta}[X_i X_i|\theta]] = E_\theta[\theta^2] \end{aligned}$$

> - by simply using the above properties and the definition of variance and covariance it is possible to prove that:
>
> $$\begin{aligned} Cov(X_i X_j) &= E[X_i X_j] - E[X_i] \cdot E[X_j] = \\ &= E_\theta[\theta^2] - (E_\theta[\theta])^2 = Var_\theta(\theta) \end{aligned}$$

2. Prove that any couple of random variabes in that sequence must be non-negatively correlated.

Starting from the definition of *correlation* $Cor[X_i X_j] = \frac{Cov(X_i X_j)}{sd(X_i) \cdot sd(X_j)}$, it's possible to use the above properties again since:

> - $Cov(X_i X_j) = Var_\theta(\theta) \geq 0$ by definition of variance;
> - $sd(X_i) \geq 0 \; \forall i$ by definition of standard deviation.

This implies that the sequence is non-negatively correlated.

3. Find what are the conditions on the distribution $\pi(\cdot)$ so that $Cor[X_i X_j] = 1$.

Let's derive the $Var_{X_j}(X_j)$ with the *law of total variance*:

$$\begin{aligned} Var_{X_j}(X_j) &= E_\theta[Var_{X_j|\theta}(X_j|\theta)] + Var_\theta(E_{X_j|\theta}[X_j|\theta]) = \\ &= E_\theta[\theta \cdot (1-\theta)] + Var_\theta(\theta) = \\ &= E_\theta[\theta] - E_\theta[\theta^2] + Var_\theta(\theta) \end{aligned}$$

Remembering the formula for correlation given in the previous point and using the *law of total variance* :

$$\begin{cases} Cor(X_i X_j) = \frac{Var_\theta(\theta)}{Var(X_j)} \\ Var_{X_j}(X_j) = E_\theta[\theta] - E_\theta[\theta^2] + Var_\theta(\theta) \end{cases} \rightarrow Cor(X_i X_j) = \frac{Var_\theta(\theta)}{E_\theta[\theta] - E_\theta[\theta^2] + Var_\theta(\theta)}$$

Thus the distribution $\pi(\cdot)$ must respect all over the previous properties and the following last condition of equality between first and second moment:

$$Cor(X_i X_j) = 1 \leftrightarrow E_\theta[\theta] = E_\theta[\theta^2]$$

4. What do these conditions imply on the type and shape of $\pi(\cdot)$? (make an example).

The shape of the distribution has to respect the condition of equality between first and second moments and the fact that the support is $[0,1]$. In particular we need to select a good distribution for the parameter θ of a Bernoulli distribution s.t. the sequence of infinitely echangeable r.V.s $\{X_i\}_{i=0}^\infty$ are identically distributed as a $Ber(\theta)$ and the condition $Cor(X_i X_j) = 1$ is satisfied $\forall i \neq j$. A distribution of this type is a degenerative one, for example a *two-point distribution* with possible outcomes 0 and 1. In this case it is then possible to simply refer to a Bernoulli random variable for θ and the distribution $\pi(.)$ is parameterized over p, assigning probability p to the point mass at 1 and $(1-p)$ at the point mass at 0.

It' easy to verify that $E[\theta] = E[\theta^2] = p$. By this construction of $\pi(\theta)$ and considering the De Finetti representation theorem we ensure unit correlation between each pair of random variables in the sequence by guaranteeing an obvious sequence result of only zeros or only ones.

```
Last update by LT: Thu Apr 20 22:03:04 2023
```