Second block exercises

Author: Aleix López Pascual
Statistics and Data Analysis

December 31, 2017

Exercise 1

We are going to compute the confidence interval (CI) of an estimator. The CI is an alternative method of reporting the statistical error of a measurement. In our case, the parameter to estimate will be the mean μ.

First of all, we evaluate an estimator $\hat{\mu}_{obs}$ from the 6 independent observations distributed as a Gaussian $\mathcal{N}(\mu, \sigma^2)$, where σ is known and μ is the true value, which is unknown. As an estimator of μ, we use the arithmetic mean $\hat{\mu} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, which is an unbiased estimator. We obtain $\hat{\mu}_{obs} = 1.19117$.

On the other hand, we know σ of $\mathcal{N}(\mu, \sigma^2)$, i.e. the standard deviation of only one measurement. But we want to know $\sigma_{\hat{\mu}}$, i.e. the standard deviation of the arithmetic mean. As we know, $\sigma_{\hat{\mu}} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} = 0.30619$. Note that this is not an estimator, it is well known.

Now we can define the pdf of the estimator. Since the measurements follow a normal distribution, then the sample mean will also follow a normal distribution $\mathcal{N}(\hat{\mu}; \mu, \sigma_{\hat{\mu}})$.

We are interested in computing the symmetric confidence interval for μ with confidence level $1-\alpha$, i.e. consider the probabilities $P(a(\hat{\mu}) \geq \mu)$ and $P(b(\hat{\mu}) \leq \mu)$ equal and with value $\frac{\alpha}{2}$. $a(\hat{\mu}_{obs})$ and $b(\hat{\mu}_{obs})$ determine the limits of the confidence interval $[a, b]$. The values of a and b are obtained solving the following equations:

$$\begin{aligned} \frac{\alpha}{2} &= \int_{\hat{\mu}_{obs}}^{\infty} g(\hat{\mu}; a) d\hat{\mu} = 1 - G(\hat{\mu}_{obs}; a) \\ \frac{\alpha}{2} &= \int_{-\infty}^{\hat{\mu}_{obs}} g(\hat{\mu}; b) d\hat{\mu} = G(\hat{\mu}_{obs}; b) \end{aligned} \tag{1}$$

In our case, we have said that the estimator $\hat{\mu}$ is Gaussian distributed $\mathcal{N}(\hat{\mu}; \mu, \sigma_{\hat{\mu}})$. Therefore, Eq. 1 becomes

$$\begin{aligned} \frac{\alpha}{2} &= 1 - \Phi\left(\frac{\hat{\mu}_{obs} - a}{\sigma_{\hat{\mu}}}\right) \\ \frac{\alpha}{2} &= \Phi\left(\frac{\hat{\mu}_{obs} - b}{\sigma_{\hat{\mu}}}\right) \end{aligned} \tag{2}$$

where Φ is the cumulative distribution of the standard Gaussian. Solving Eq. 2 we obtain

$$\begin{aligned} a &= \hat{\mu}_{obs} - \sigma_{\hat{\mu}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ b &= \hat{\mu}_{obs} + \sigma_{\hat{\mu}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned} \tag{3}$$

where Φ^{-1} are the quantiles of the standard Gaussian and their values can be found in some tables. Our values of interest are found in Table 1.

$1-\alpha$	$\Phi^{-1}(1-\alpha/2)$
0.68	1.000
0.90	1.645
0.95	1.960

Table 1: Quantiles of the standard Gaussian Φ^{-1} for different confidence levels $1-\alpha$ of central intervals.

Now we have all the ingredients to compute $[a, b]$ from Eq. 3. We obtain the following results:

$1-\alpha$	a	b
0.68	0.88498	1.49735
0.90	0.68749	1.69484
0.95	0.59104	1.79129

Table 2: Symmetric confidence intervals $[a, b]$ for μ for different confidence levels $1-\alpha$.

We know that for the conventional 68.3% central confidence interval, one should obtain

$$[a, b] = [\hat{\mu}_{obs} - \sigma_{\hat{\mu}},\ \hat{\mu}_{obs} + \sigma_{\hat{\mu}}] \tag{4}$$

Since $\hat{\mu}_{obs} = 1.19117$ and $\sigma_{\hat{\mu}} = 0.30619$, we indeed find $[a, b] = [0.88498, 1.49735]$, which coincide with the computed results of Table 2.

CI for unknown standard deviation

In case σ is not known, we have to estimate it. As an estimator of σ we use the sample variance $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$, which is unbiased. We obtain $\hat{\sigma} = 0.75778$. Then $\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{N}} = 0.30936$. As we can observe $\hat{\sigma}_{\hat{\mu}}$ depends on $\hat{\mu}$. Then, it will not be so simple to relate the cumulative distribution $G(\hat{\mu}; \mu, \hat{\sigma}_{\hat{\mu}})$ to the cumulative distribution of the standard Gaussian Φ. Only if we can consider that $\hat{\sigma}_{\hat{\mu}}$ is a sufficiently good approximation of $\sigma_{\hat{\mu}}$, we can still use Eq. 2. This could be the case, for example, for a large enough data sample. Nevertheless, it is clear not our situation since we only have 6 observations. Instead, in our case (small sample) where in fact $\hat{\mu}$ is representing the mean of 6 Gaussian variables of unknown $\sigma_{\hat{\mu}}$, we can relate $G(\hat{\mu}; \mu, \hat{\sigma}_{\hat{\mu}})$ to the Student's t-distribution with $(N-1)$ degrees of freedom. In such case, we obtain the limits of the confidence interval as

$$\begin{aligned} a &= \hat{\mu}_{obs} - \sigma_{\hat{\mu}} G^{-1}\left(1 - \frac{\alpha}{2}\right) \\ b &= \hat{\mu}_{obs} + \sigma_{\hat{\mu}} G^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned} \tag{5}$$

where G^{-1} are the quantiles of the Student's t-distribution. Our values of interest for 5 degrees of freedom are found in Table 3.

$1-\alpha$	$1-\alpha/2$	$G^{-1}(1-\alpha/2)$
0.68	0.840	1.004
0.90	0.950	2.015
0.95	0.975	2.571

Table 3: Quantiles of the Student's t-distribution G^{-1} of 5 degrees of freedom for different confidence levels $1-\alpha$ of central intervals.

Therefore we obtain the following results:

$1-\alpha$	a	b
0.68	0.84963	1.53270
0.90	0.56780	1.81453
0.95	0.39580	1.98654

Table 4: Symmetric confidence intervals $[a, b]$ for μ for different confidence levels $1-\alpha$ in the case of unknown standard deviation.

As we can observe, these intervals are larger than those obtained when σ was assumed known.

Exercise 2

I. A pdf always has to be normalized in accordance to the Kolmogorov axioms of probability. Taking into account that the Poisson distribution is a discreet probability distribution and $N = 0, 1, 2...$, we have

$$\sum_{\Omega} P(N;\nu) = \sum_{N=0}^{\infty} \frac{\nu^N}{N!} e^{-\nu} = e^{-\nu} \sum_{N=0}^{\infty} \frac{\nu^N}{N!} = e^{-\nu} e^{\nu} = 1 \tag{6}$$

In order to find a frequentist estimator for ν we can use the maximum likelihood method. In our case, since our sample only contains one single measurement ($N = N_{obs}$, $n = 1$), the likelihood function is simply:

$$L(\nu) = \frac{\nu^N}{N!} e^{-\nu} \tag{7}$$

Then the ML estimator is

$$\frac{\partial \log L}{\partial \nu} = 0 \;\Rightarrow\; \hat{\nu}_{ML} = N = N_{obs} \tag{8}$$

The expected value of the estimator $\hat{\nu}$ is

$$\begin{aligned} \langle \hat{\nu} \rangle = \langle N \rangle &= \sum_{N=0}^{\infty} N \frac{\nu^N}{N!} e^{-\nu} = \sum_{N=1}^{\infty} N \frac{\nu^N}{N!} e^{-\nu} = \sum_{N=1}^{\infty} \frac{\nu^N}{(N-1)!} e^{-\nu} \\ &= \nu e^{-\nu} \sum_{N=1}^{\infty} \frac{\nu^{N-1}}{(N-1)!} = \nu e^{-\nu} \sum_{N=0}^{\infty} \frac{\nu^N}{N!} = \nu e^{-\nu} e^{\nu} = \nu \end{aligned} \tag{9}$$

Therefore, the estimator $\hat{\nu}$ is unbiased ($b = \langle \hat{\nu} \rangle - \nu = 0$), which means that it is independent of the sample size n. The variance of the estimator $\hat{\nu}$ is

$$\begin{aligned} \mathrm{Var}(\hat{\nu}) = \mathrm{Var}(N) &= \langle N^2 \rangle - \langle N \rangle^2 = \langle N(N-1) + N \rangle - \langle N \rangle^2 \\ &= \langle N(N-1) \rangle + \langle N \rangle - \langle N \rangle^2 = \langle N(N-1) \rangle + \nu - \nu^2 \end{aligned} \tag{10}$$

We need to calculate $\langle N(N-1) \rangle$:

$$\begin{aligned} \langle N(N-1) \rangle &= \sum_{N=0}^{\infty} N(N-1) \frac{\nu^N e^{-\nu}}{N!} = \sum_{N=2}^{\infty} N(N-1) \frac{\nu^N e^{-\nu}}{N!} = \sum_{N=2}^{\infty} \frac{\nu^N e^{-\nu}}{(N-2)!} \\ &= \nu^2 e^{-\nu} \sum_{N=2}^{\infty} \frac{\nu^{N-2}}{(N-2)!} = \nu^2 e^{-\nu} \sum_{N=0}^{\infty} \frac{\nu^N}{N!} = \nu^2 e^{-\nu} e^{\nu} = \nu^2 \end{aligned} \tag{11}$$

Thus

$$\mathrm{Var}(\hat{\nu}) = \nu^2 + \nu - \nu^2 = \nu \tag{12}$$

II. We assume that the pdf for the estimator $\hat{\nu}$ is Gaussian $\mathcal{N}(\hat{\nu}; \mu = \langle \hat{\nu} \rangle, \sigma^2 = \mathrm{Var}(\hat{\nu}))$ and we consider an observed value of $N_{obs} = 150$. Therefore $\hat{\nu} = N_{obs} = 150$. As we have seen, the expected value and variance of the estimator $\hat{\nu}$ are functions of the true (and unknown) parameter ν, i.e. $a = a(\nu)$. However, because of the transformation invariance of ML estimators, the ML estimator of a function a of a parameter ν is simply $\hat{a} = a(\hat{\nu})$. This is true because

$$\frac{\partial L}{\partial \nu} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial \nu} = 0 \qquad \text{where } \frac{\partial a}{\partial \nu} \neq 0 \tag{13}$$

Therefore

$$\hat{E}(\hat{\nu}) = \hat{\nu} = N_{obs} = 150 \tag{14}$$

$$\hat{\mathrm{Var}}(\hat{\nu}) = \hat{\nu} = N_{obs} = 150 \tag{15}$$

Taking into account this, now we can plot the corresponding pdf (Fig. 1).

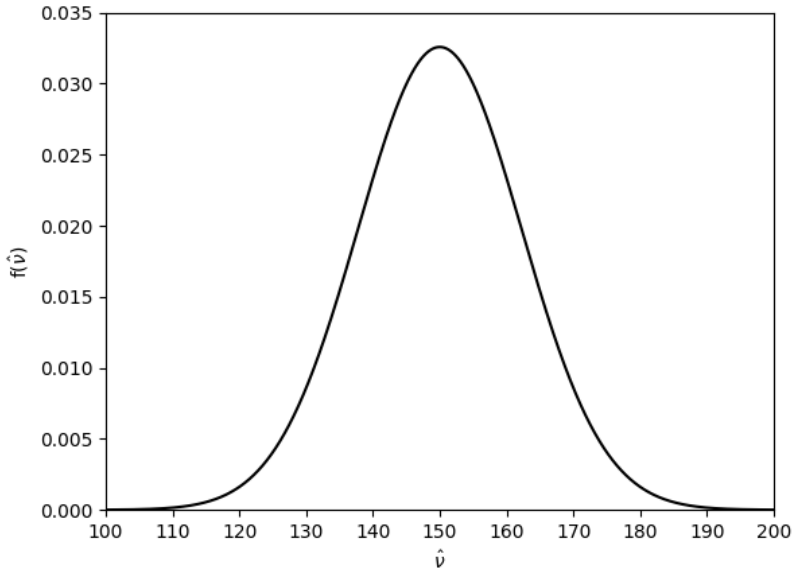

Figure 1: Pdf for the estimator $\hat{\nu}$ as a function of $\hat{\nu}$. It corresponds to a Gaussian $\mathcal{N}(\mu = 150, \sigma^2 = 150)$.

III. Now, instead, we are going to construct the pdf for $\hat{\nu}$ using a Monte Carlo simulation. In order to do this, one must simulate a large number of experiments, compute the ML estimates each time and look at how the resulting values are distributed.

Firstly, regarding the first experiment for which we obtained $q = \hat{\nu} = N_{obs} = 150$ as the "real" one, we now generate 10^6 Monte Carlo experiments using q of the first experiment as the true value of the parameter ν. We want the measurement of each MC to be Poisson distributed, and select one of the measurements per each one. However, since the Poisson distribution is a well-known distribution, we already have a module, so-called numpy.random.poisson, which generates random numbers from a Poisson distribution directly. Then, we compute a new estimator $q_i = \hat{\nu}_i = N_i$ per each experiment and we plot a histogram of these values (Fig. 2).

As we can observe from Fig. 2, the histogram fits with the Gaussian $\mathcal{N}(150, 150)$. This is a general property of the ML estimators for the large sample limit (10^6), known as asymptotic normality. In fact, now that we have a sample of $\hat{\nu}_i$, we can compute an estimator for the mean and the variance using

$$\overline{\hat{\nu}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\nu}_i \tag{16}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\nu}_i - \overline{\hat{\nu}})^2 \tag{17}$$

and we obtain $\overline{\hat{\nu}} = 150$ and $s^2 = 150$, which coincides with the "true" parameter used $\nu = 150$ as expected since $\hat{\nu}$ is unbiased.

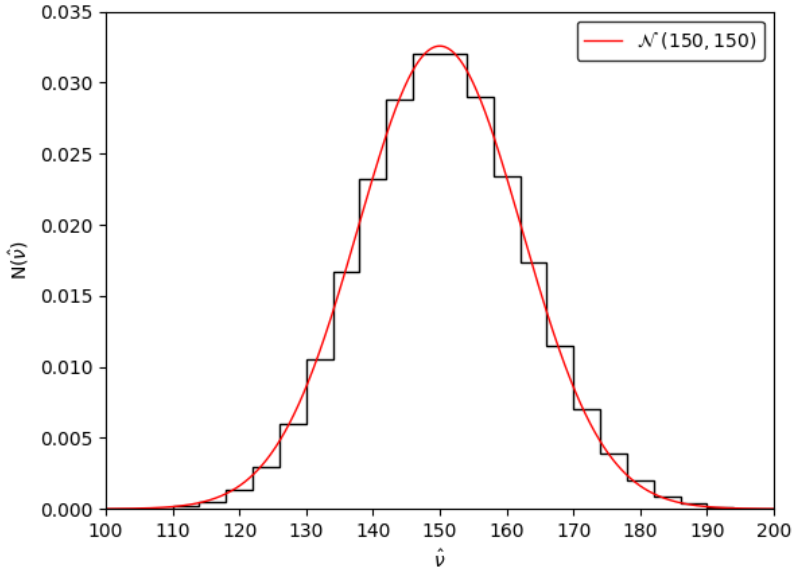

Figure 2: Histogram of the estimator $\hat{\nu}$ from 10^6 Monte Carlo experiments using $\nu = 150$ as the "true" parameter. We compare it with the result in part II.

IV. So far we have been working with frequentist statistics. From now on, we are going to use the Bayesian interpretation. We are going to compute the Bayesian posterior pdf for ν, i.e. $P(\nu|N)$. In order to do so, we apply the Bayes theorem:

$$P(\nu|N) = \frac{L(N|\nu)\pi(\nu)}{\int d\nu L(N|\nu)\pi(\nu)} \tag{18}$$

where $L(N|\nu)$ is the likelihood function and $\pi(\nu)$ is the prior probability density for ν, which reflects the state of knowledge of ν before consideration of the data. Once we know $P(\nu|N)$, i.e. the pdf for ν, we can compute everything (mean, variance, CI, ...). The likelihood $L(N|\nu)$ is

$$L(N|\nu) = \frac{\nu^N}{N!}e^{-\nu} \tag{19}$$

On the other hand, we must define the prior. Since we know that $\nu \geq 0$, we can easily incorporate this knowledge by setting the prior $\pi(\nu) = 0$ in the excluded region ($\nu < 0$). Then Bayes theorem gives a posterior pdf $P(\nu|N) = 0$ for $\nu < 0$. However, we do not know anything else about ν and we still have the range $[0, \infty]$ undefined. What do we do? We assume that a priori all values of $\nu \geq 0$ are equally likely (uniform). This is expressed by $\pi(\nu) = constant$ for $\nu \geq 0$. Therefore our non-informative prior is defined as

$$\begin{cases} \pi(\nu)d\nu = Cd\nu & \text{if } \nu \geq 0 \\ \pi(\nu)d\nu = 0 & \text{if } \nu < 0 \end{cases} \tag{20}$$

Note that the range of $\nu \geq 0$ is infinite. Then, the constant cannot be normalized. This is called an improper prior. However it is not a problem in our case since the constants cancel out. Replacing all this we obtain

$$P(\nu|N) = \frac{\frac{\nu^N}{N!}e^{-\nu}C}{\int_0^\infty d\nu \frac{\nu^N}{N!}e^{-\nu}C} = \frac{\nu^N e^{-\nu}}{\int_0^\infty d\nu\ \nu^N e^{-\nu}} \tag{21}$$

We solve the integral

$$\int_0^\infty d\nu\ \nu^N e^{-\nu} = \Gamma(N+1) = N! \tag{22}$$

which implies that the normalization constant of $P(\nu|N)$, i.e. the denominator, is equal to 1. So at the end, we have

$$\begin{cases} P(\nu|N) = \dfrac{\nu^N e^{-\nu}}{N!} & \text{if } \nu \geq 0 \\ P(\nu|N) = 0 & \text{if } \nu < 0 \end{cases} \tag{23}$$

As we can observe from Eq. 23, the posterior pdf for ν does not look as a Gaussian. However when we plot it for a given observed value of $N = N_{obs} = 150$, we actually obtain approximately a Gaussian (Fig. 3). Notice that when we compute Eq. 23 for $N = 150$ we obtain overflow problems in ν^N. In order to solve this we have computed $\nu^{N/2} e^{-\nu} \nu^{N/2}$ in such order.

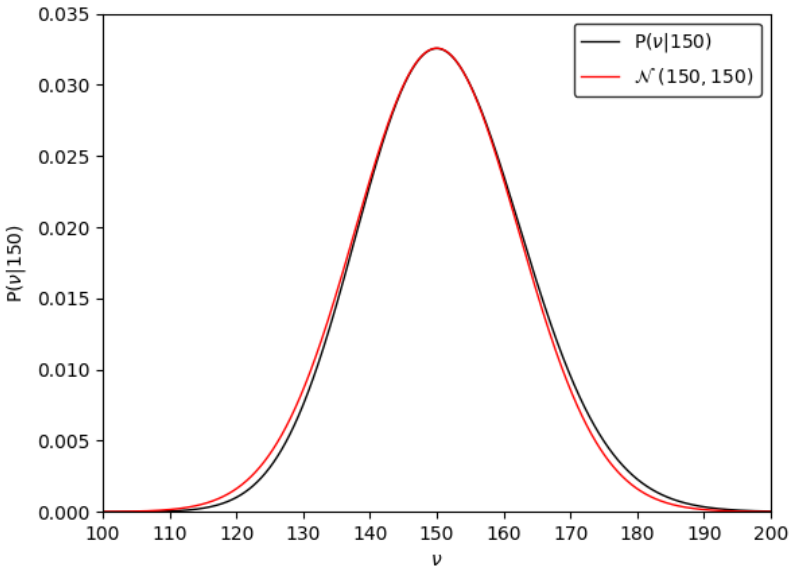

Figure 3: Posterior pdf for the parameter ν given the data $N = N_{obs} = 150$ as a function of ν. We compare it with the Gaussian $\mathcal{N}(150, 150)$ from II.

As we can observe from Fig. 3, the posterior pdf for ν fits approximately with the Gaussian $\mathcal{N}(150, 150)$. The small differences are found in the tails and not in the peak. This is important because we usually define the Bayesian estimator as the value of ν at which $P(\nu|N)$ is a maximum, i.e. the posterior mode. Then, considering that the prior $\pi(\nu)$ has been taken to be constant, $P(\nu|N)$ is proportional to the likelihood $L(N|\nu)$. Therefore, the Bayesian estimator and the ML estimator should coincide. As we can observe, this is in fact the case $\hat{\nu}_{ML} = \hat{\nu}_{Bayes} = 150$.

V. Now we are going to construct the Bayesian posterior pdf for ν using a Monte Carlo simulation. Firstly, we generate a set of values ν in accordance to our prior $\pi(\nu)$ (Eq. 20). This corresponds to a uniform distribution $\mathcal{U}(0, \infty)$ where we have considered the infinite upper limit as $N_{obs} + 10\sqrt{N_{obs}}$. As we know from Eq. 18, and taking into account that the normalization constant is equal to 1, $P(\nu|N) = L(N|\nu)\pi(\nu)$. Then, in order to consider the contribution of $L(N|\nu)$, we generate one value of N according to a Poisson distribution for each value of ν, i.e. one MC experiment per each ν. Finally, since we want to compute the pdf for ν given the data $N_{obs} = 150$, we only accept the value of ν if $N = N_{obs}$.

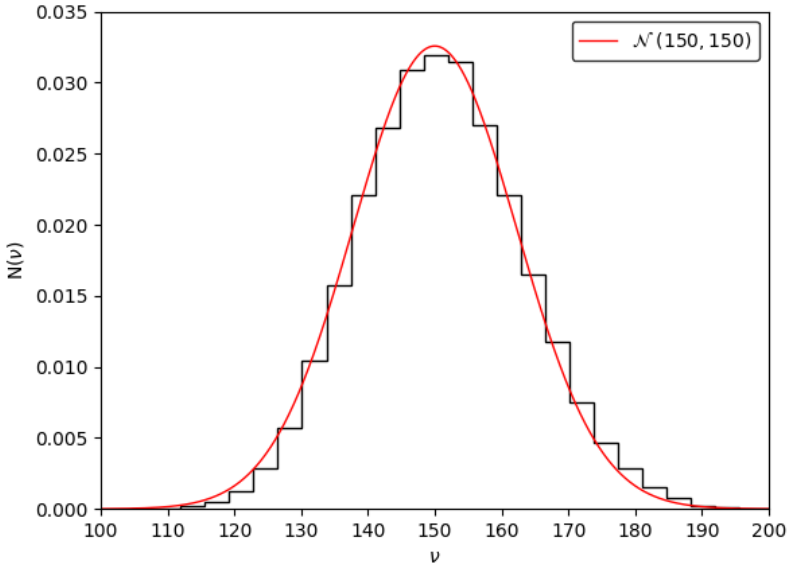

Figure 4: Histogram of the parameter ν from 10^7 Monte Carlo experiments using a different ν per each one. We compare it with the Gaussian $\mathcal{N}(150, 150)$ from II.

As we can observe, the histogram fits approximately with the Gaussian $\mathcal{N}(150, 150)$. However, again we find small differences in the tails as in Fig. 3. Remember that these differences did not appear in the frequentist case (Fig. 2), for which we obtained a sample mean of 150 and a sample variance of 150. Instead, if now we compute these estimators, we find out $\overline{\nu} = 151$ and $s^2 = 151$. Thus, we should expect the obtained histogram to fit better with a Gaussian $\mathcal{N}(151, 151)$. In fact, this is the case (Fig. 5).

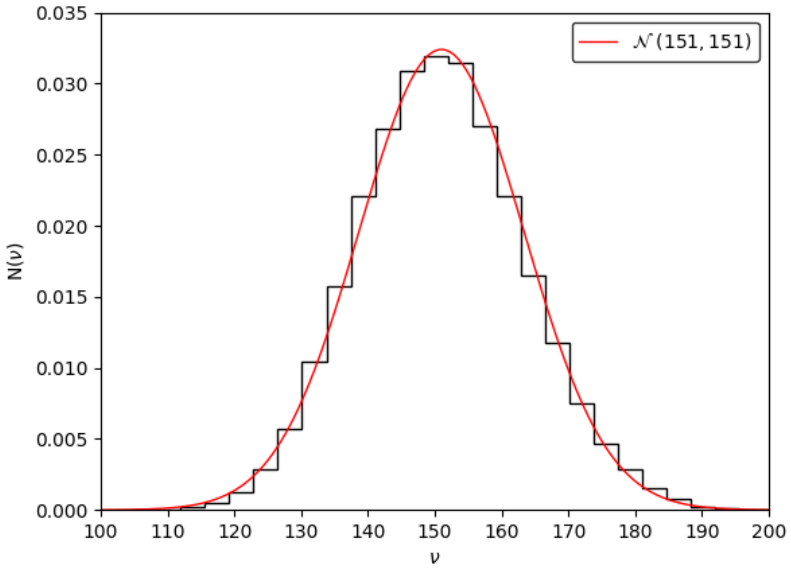

Figure 5: Histogram of the parameter ν from 10^7 Monte Carlo experiments using a different ν per each one. We compare it with the Gaussian $\mathcal{N}(\overline{\nu} = 151, s^2 = 151)$.

We have seen that the posterior mode is 150 and then $\hat{\nu}_{ML} = \hat{\nu}_{Bayes} = 150$. But on the other hand we have obtained the unexpected $\overline{\nu} = 151$, $s^2 = 151$. Why? The reason is because in the frequentist case, we computed the pdf of $\hat{\nu}$ assuming an initial "true" value ν of 150 in order to perform the MC simulation. Instead, in the Bayesian case, we are not assuming any true value for ν. We are just using our knowledge of ν to define a prior $\pi(\nu)$. Since $\pi(\nu)$ covers a wider range of values, we have obtained different results.

VI. and VII. We are going to repeat the calculations for different values of the data N_{obs} in order to observe how the behaviour of the corresponding posterior pdf $P(\nu|N_{obs})$ changes.

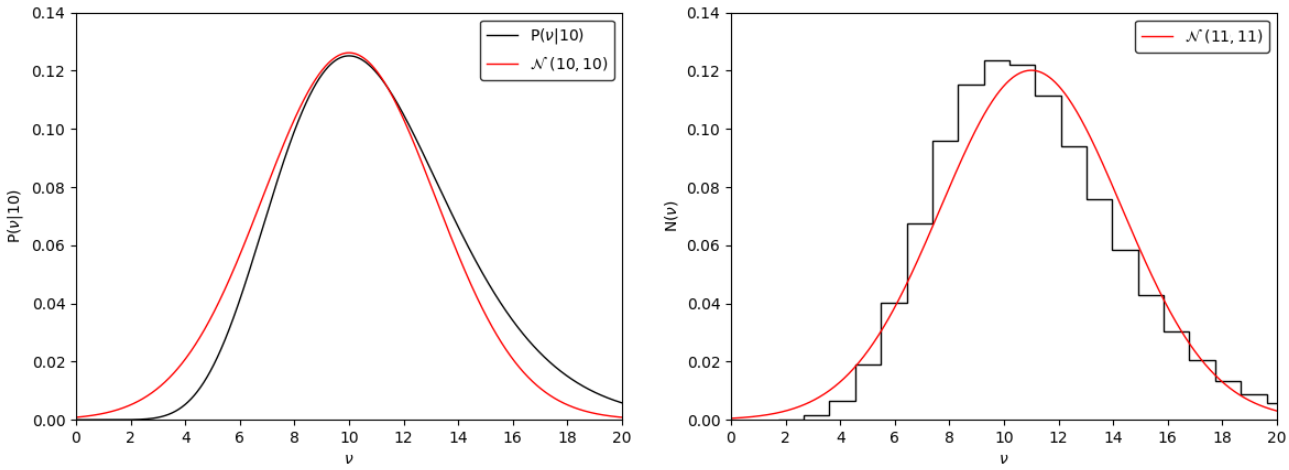

Figure 6: On the left, posterior pdf for ν given $N_{obs} = 10$ and using the prior flat in ν. We compare it with the Gaussian $\mathcal{N}(10, 10)$. On the right, histogram of ν compared with the Gaussian $\mathcal{N}(\overline{\nu} = 11, s^2 = 11)$.

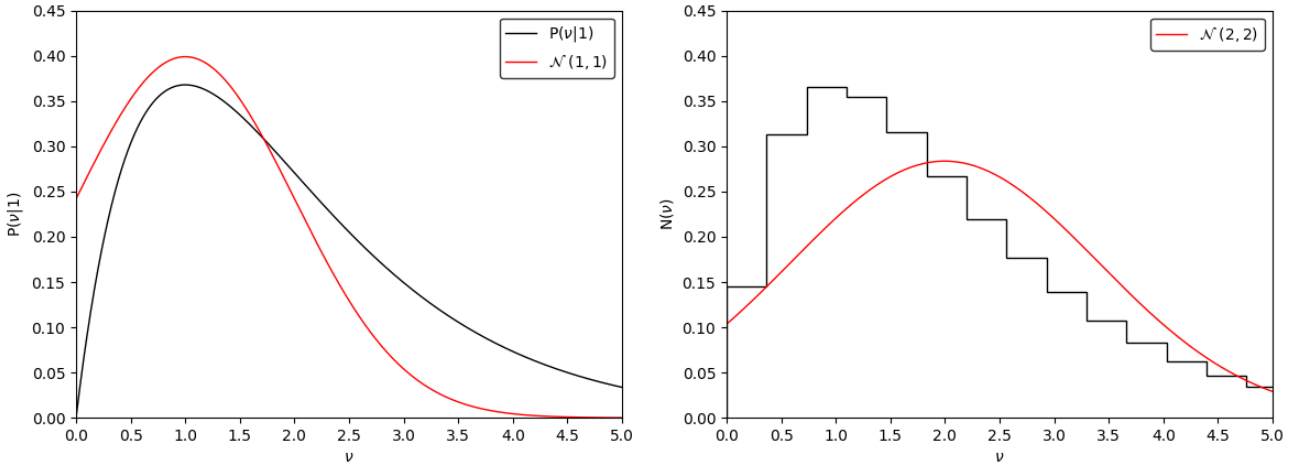

Figure 7: On the left, posterior pdf for ν given $N_{obs} = 1$ and using the prior flat in ν. We compare it with the Gaussian $\mathcal{N}(1, 1)$. On the right, histogram of ν compared with the Gaussian $\mathcal{N}(\overline{\nu} = 2, s^2 = 2)$.

As we can observe from Fig. 6 and Fig. 7, when N_{obs} is decreased, $P(\nu|N_{obs})$ stops resembling a Gaussian and it starts to look like a Poisson. Why? The reason is because we are using a flat non-informative prior, i.e. a prior with minimal influence on the inference. As a consequence $P(\nu|N_{obs}) \propto L(N_{obs}|\nu)$ and we have seen that $\hat{\nu}_{Bayes} = \hat{\nu}_{ML} = N_{obs}$. Since $L(N_{obs}|\nu)$ corresponds to a Poisson distribution for a single measurement N_{obs}, then we should expect that $P(\nu|N_{obs})$ behaves as such.

As we know, for values $\nu > 10$, the Gaussian distribution $\mathcal{N}(\nu, \nu)$ starts to be a good approximation for the Poisson distribution. Particularly, for sufficiently large values of ν ($\nu > 1000$), the Gaussian becomes an excellent approximation. Having said that and recalling $\hat{\nu}_{Bayes} = \hat{\nu}_{ML} = N_{obs}$, the behaviour of Fig. 6 and Fig. 7 is justified.

VIII. So far we have used a non-informative prior based on the translational invariance (Eq. 20), and we have obtained unexpected results. However, there is another type of non-informative prior for continuous variables: a prior based on the scale invariance. This prior is named Jeffreys prior and is defined as

$$\begin{cases} \pi(\nu)d\nu = C\dfrac{d\nu}{\nu} = C\ d\log\nu & \text{if } \nu \geq 0 \\ \pi(\nu)d\nu = 0 & \text{if } \nu < 0 \end{cases} \tag{24}$$

Now we are going to repeat the previous computations using this prior. First, we calculate the Bayesian posterior pdf for ν analytically using the Bayes theorem:

$$P(\nu|N) = \frac{L(N|\nu)\pi(\nu)}{\int d\nu L(N|\nu)\pi(\nu)} = \frac{\frac{\nu^N}{N!}e^{-\nu}\frac{C}{\nu}}{\int_0^\infty d\nu \frac{\nu^N}{N!}e^{-\nu}\frac{C}{\nu}} = \frac{\nu^{N-1}e^{-\nu}}{\int_0^\infty d\nu\ \nu^{N-1}e^{-\nu}} \tag{25}$$

We solve the integral

$$\int_0^\infty d\nu\ \nu^{N-1}e^{-\nu} = \Gamma(N) = (N-1)! \tag{26}$$

So at the end, we have

$$\begin{cases} P(\nu|N) = \dfrac{\nu^{N-1}e^{-\nu}}{(N-1)!} & \text{if } \nu \geq 0 \\ P(\nu|N) = 0 & \text{if } \nu < 0 \end{cases} \tag{27}$$

which differs from Eq. 23 by the change $N \to N-1$. Therefore, for large N we should expect no distinction between the results of both priors.

As we can observe from Fig. 8, using this prior we also find differences between the pdf and the Gaussian as in Fig. 3. Note that the differences in the tails have been corrected, but now new differences in the peak have appeared, which in fact, imply $\nu_{mode} = 149 \neq \hat{\nu}_{ML}$.

Now we construct the pdf for ν using a Monte Carlo simulation. We proceed as before, except that we have to generate a set of values ν in accordance to a log uniform distribution $\pi(\nu)$. We can generate values distributed in such way just using the uniform distribution and a change of variables:

$$\log\mathcal{U}(a,b) \sim \exp(\mathcal{U}(\log(a), \log(b))) \tag{28}$$

Notice that we want to generate values of $\nu \geq 0$. It is clear that if we drawn values from $\exp(\mathcal{U}(\log(0), \log(N_{obs} + 10\sqrt{N_{obs}})))$, our lower limit will be $\nu = 1$. The correct distribution is then $\exp(\mathcal{U}(-100, \log(N_{obs} + 10\sqrt{N_{obs}})))$.

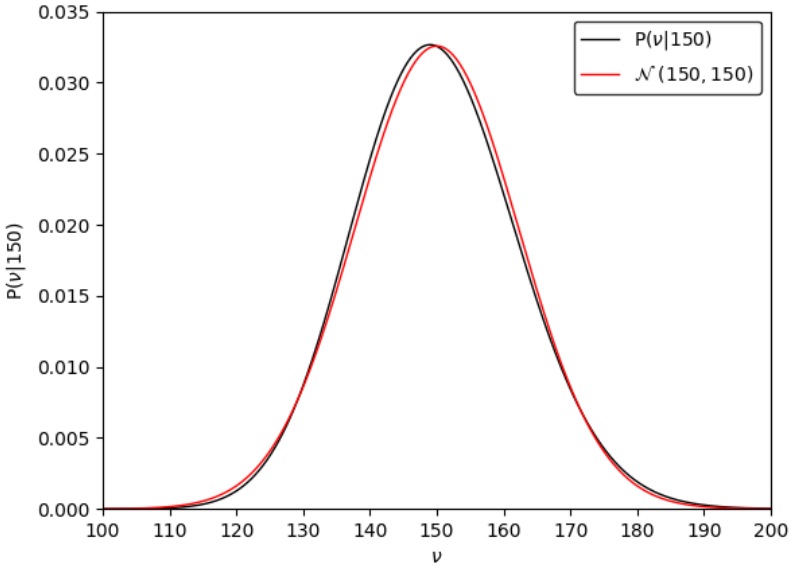

Figure 8: Posterior pdf for the parameter ν given the data $N = N_{obs} = 150$ as a function of ν. We have used the Jeffreys prior. We compare it with the Gaussian $\mathcal{N}(150, 150)$ from II.

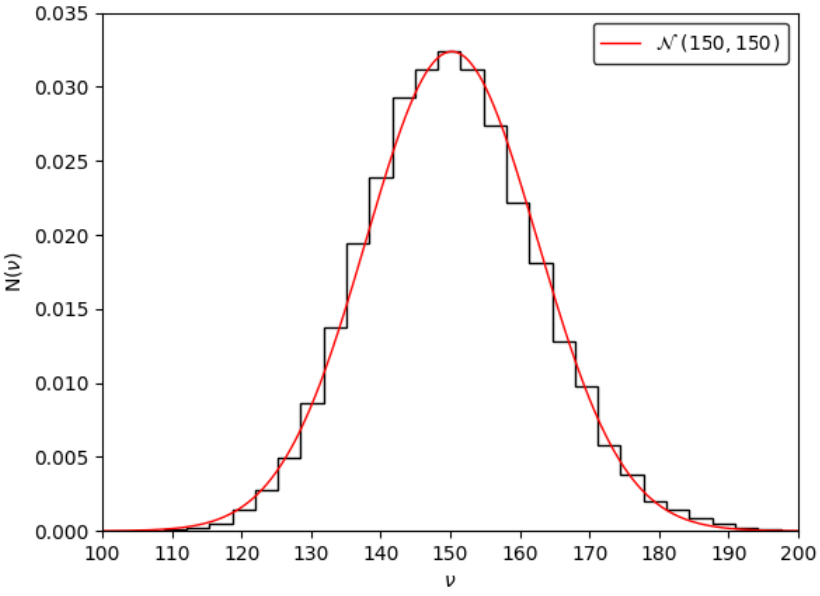

Figure 9: Histogram of the parameter ν from 10^7 Monte Carlo experiments using the Jeffreys prior and $N_{obs} = 150$. We compare it with the Gaussian $\mathcal{N}(\overline{\nu} = 150, s^2 = 150)$.

As we have said, using the Jeffreys prior we have found out that $\nu_{mode} = 149 \neq \hat{\nu}_{ML}$. Instead, we find that the sample mean is $\overline{\nu} = 150$ and the sample variance is $s^2 = 150$. Therefore, we prefer to report $\overline{\nu}$ and s^2 as estimators.

Now we repeat the calculations for different values of the data N_{obs} and we obtain Fig. 10 and Fig. 11.

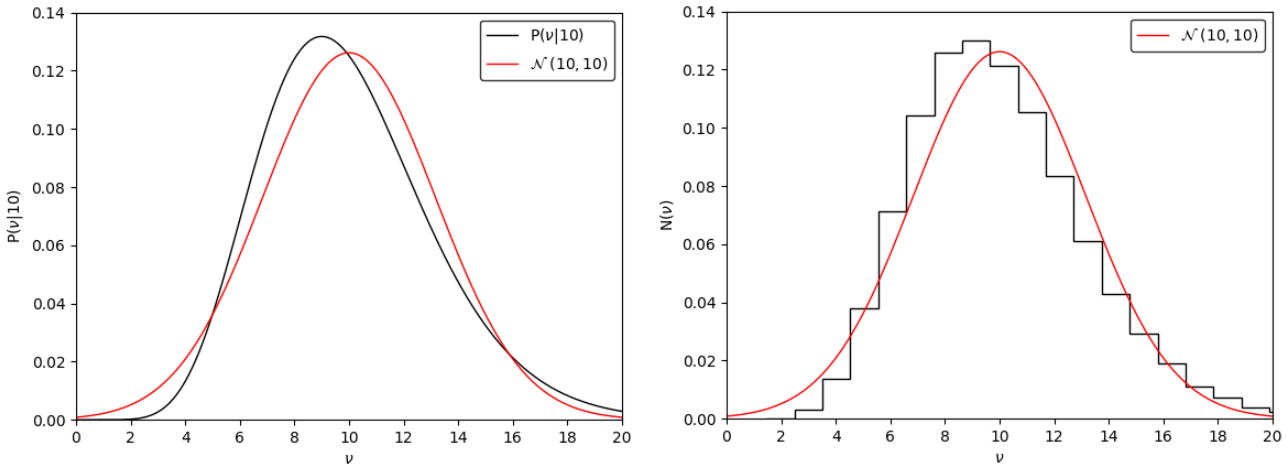

Figure 10: On the left, posterior pdf for ν given $N_{obs} = 10$ and using the Jeffreys prior. We compare it with the Gaussian $\mathcal{N}(10, 10)$. On the right, histogram of ν compared with the Gaussian $\mathcal{N}(\overline{\nu} = 10, s^2 = 10)$.

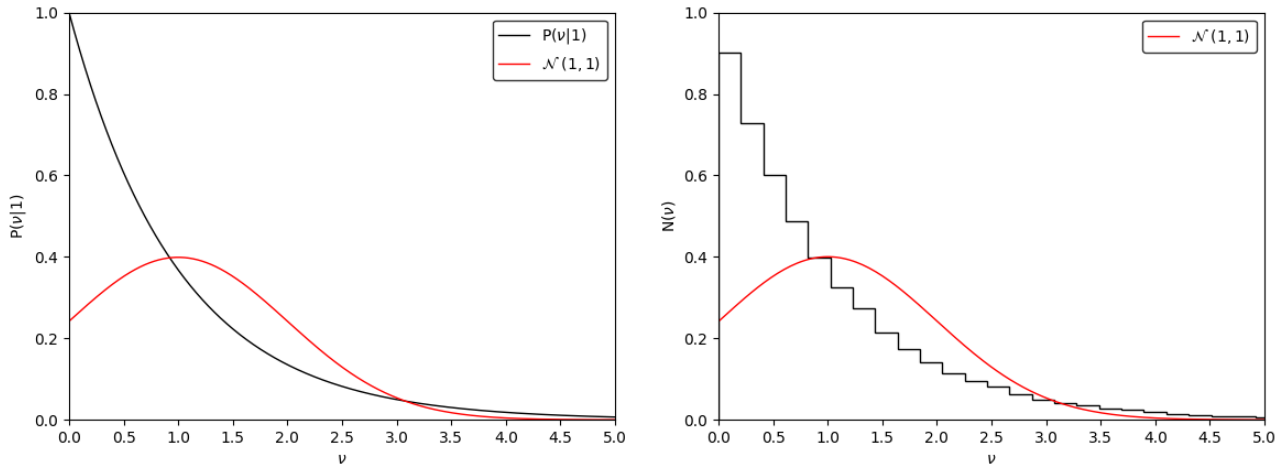

Figure 11: On the left, posterior pdf for ν given $N_{obs} = 1$ and using the Jeffreys prior. We compare it with the Gaussian $\mathcal{N}(1, 1)$. On the right, histogram of ν compared with the Gaussian $\mathcal{N}(\overline{\nu} = 1, s^2 = 1)$.

Again, we find the same conclusion that we found in VI and VII. We also obtain that $\nu_{mode} \neq \hat{\nu}_{ML}$ and $\overline{\nu} = \hat{\nu}_{ML}$ as we thought.

IX. Similitudes and differences between the frequentist and Bayesian results: As we have seen, in the frequentist case the histogram of the parameter ν fits exactly with the Gaussian due to the asymptotic normality (large sample limit). Instead, in the Bayesian interpretation we find differences. The reason is because in the frequentist case we compute the pdf of $\hat{\nu}$ assuming an initial "true" value ν. In the Bayesian case, we are not assuming any true value for ν. We are just using our knowledge of ν to define a prior $\pi(\nu)$ which covers a wider range of values. Furthermore, we have found that such differences increase when N_{obs} decreases. Then we conclude that a difficult of the Bayesian interpretation is to specify the prior correctly.

On the other hand, we have seen that the frequentist estimator for ν coincides with the Bayesian estimator. This is in fact a characteristic of flat priors. However, we have used a different Bayesian estimator depending on the prior used: ν_{mode} for the prior flat on ν, and $\overline{\nu}$ for the Jeffreys prior.

Importance of choosing Jeffreys prior: The Poisson distribution is characterized by the rate parameter ν. A rate parameter is simply the reciprocal of a scale parameter. Instead, it does not have location parameters.

Throughout the exercise, we have used two priors with particular invariances: the prior flat on ν was location invariant and the Jeffreys prior was scale invariant. The first reason of choosing the Jeffreys prior is because it is more physical for the Poisson case. We should not give the same probability to the range of values $1-2$ than $15698-15699$. We prefer to consider logarithmic ranges. The second reason is because is scale invariant. This implies that we can compute the posterior pdf of a different parameter just performing a reparametrization (change of variables). This is important because you can be interested in a different parameter rather than the usual one.

In fact, we could define a more general Jeffreys prior which fulfils both invariances. In our case is defined as:

$$\begin{cases} \pi(\nu)d\nu = \nu^{-1/2}\, d\nu & \text{if } \nu \geq 0 \\ \pi(\nu)d\nu = 0 & \text{if } \nu < 0 \end{cases} \tag{29}$$

Validity of the Gaussian approximation: We have seen that in the frequentist case the pdf for ν fits exactly with the Gaussian, while in the Bayesian case we find differences with the two priors used. These differences increase when N_{obs} decreases. The reason of that was justified in VI. and VII. As we discussed there, for sufficiently large values of ν ($\nu > 1000$), the Gaussian becomes an excellent approximation.

Exercise 3

I. We are going to generate a simulated experimental data set D of 10^4 points from a 2D Gaussian with parameters $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 2$ and $\rho = 0.8$. In order to do so, we are going to make use of the Cholesky decomposition. This decomposition is commonly used in the Monte Carlo method for simulating systems with multiple correlated variables.

The statement of the Cholesky decomposition is that given a hermitian positive-definite matrix A, we can decompose it uniquely as $A = LL^\dagger$, where L is a lower triangular matrix with real and positive diagonal entries. In our case, the matrix to decompose is the covariance matrix V, which is real. In order to compute the L matrix, we use the module numpy.linalg.cholesky, which takes V as an argument. Once we have L, we multiply it to a vector of uncorrelated variables $\vec{y}$ and we obtain a vector $L\vec{y}$ with the covariance properties of the initial variables $\vec{x}$.

In our case, we want to generate the correlated variables $\vec{x} = (x_1, x_2)$ distributed as a 2D Gaussian $\mathcal{N}(\vec{\mu}, V)$. Then, we consider the uncorrelated variables $\vec{y} = (y_1, y_2)$ distributed as a Gaussian $\mathcal{N}(0, 1)$. As a consequence, the transformation between both variables is $\vec{x} = \vec{\mu} + L\vec{y}$. The covariance matrix is

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{30}$$

After applying the Cholesky decomposition we obtain

$$L = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{1-\rho^2}\sigma_2 \end{pmatrix} \tag{31}$$

Then the transformation gives

$$\begin{aligned} x_1 &= \mu_1 + \sigma_1 y_1 \\ x_2 &= \mu_2 + \rho\sigma_2 y_1 + \sqrt{1-\rho^2}\sigma_2 y_2 \end{aligned} \tag{32}$$

After generating 10^4 observations (x_1, x_2) following this transformation we obtain the Fig. 12

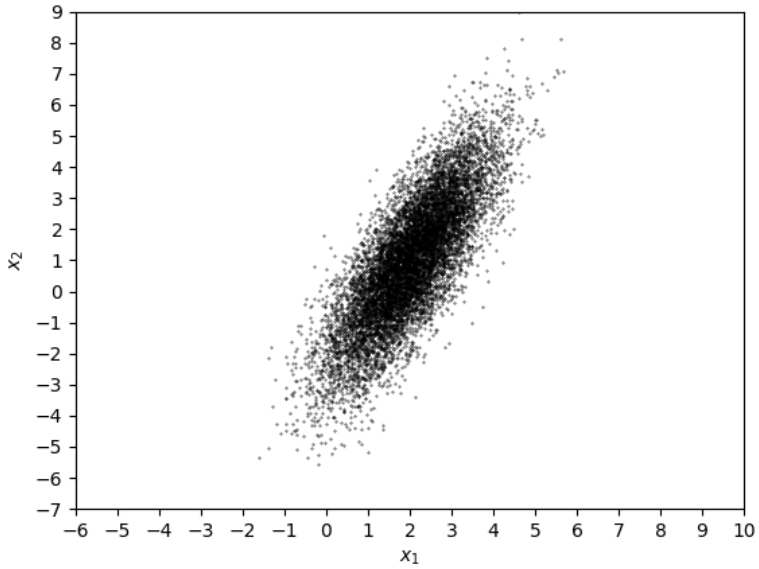

Figure 12: Scatter plot of the random variables x_1 and x_2 distributed as a 2D Gaussian $\mathcal{N}(\vec{\mu}, V)$ with positive correlation $\rho = 0.8$. $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 2$.

As we can observe (Fig. 12), the two variables are still far from having a linear relationship with each other. Also, we can note that the direction of the relationship is the expected for a positive correlation.

Alternatively, there is a multivariate normal function in the scipy.stats subpackage which can also be used to obtain a multivariate Gaussian pdf directly. Both methods are correct.

The point of the exercise will be to determine the parameters μ_1 and μ_2 using the Bayesian interpretation. At the end, we should obtain that these parameters coincide with the true values used to generate the initial data set.

II. We are going to compute the Bayesian posterior pdf for (μ_1, μ_2), i.e. $P(\mu_1, \mu_2|D)$. We proceed as we did in exercise 2, but now we have two unknown parameters instead of one. Also, this time we are not going to worry about the normalization constant, since in the next parts of the exercise it will cancel out. Then, the Bayes theorem reads as

$$P(\mu_1, \mu_2|D) \propto L(D|\mu_1, \mu_2)\pi(\mu_1, \mu_2) \tag{33}$$

Firstly, we start building the non-normalized likelihood $L(D|\mu_1, \mu_2)$:

$$\begin{aligned} L(D|\mu_1, \mu_2) &= \prod_{i=1}^{N} f(x_{i,1}, x_{2,i}; \mu_1, \mu_2) \\ &= \prod_{i=1}^{N} \frac{1}{2\pi|det\ V|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x_i} - \vec{\mu})^T V^{-1} (\vec{x_i} - \vec{\mu}) \right) \end{aligned} \tag{34}$$

where we have used the definition of the bivariate normal distribution and $N = 10^4$ is the number of observations. We remove the normalization factor. Therefore

$$L(D|\mu_1, \mu_2) \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2} \begin{pmatrix} x_{1,i} - \mu_1 \\ x_{2,i} - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_{1,i} - \mu_1 \\ x_{2,i} - \mu_2 \end{pmatrix} \right] \tag{35}$$

The inverse of V is

$$V^{-1} = \frac{1}{|V|} adj(V) = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \tag{36}$$

Then

$$\begin{aligned} L(D|\mu_1, \mu_2) &\propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} x_{1,i} - \mu_1 \\ x_{2,i} - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x_{1,i} - \mu_1 \\ x_{2,i} - \mu_2 \end{pmatrix} \right] \\ &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} x_{1,i} - \mu_1 \\ x_{2,i} - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_2^2(x_{1,i} - \mu_1) - \rho\sigma_1\sigma_2(x_{2,i} - \mu_2) \\ \sigma_1^2(x_{2,i} - \mu_2) - \rho\sigma_1\sigma_2(x_{1,i} - \mu_1) \end{pmatrix} \right] \\ &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \Big(\sigma_2^2(x_{1,i} - \mu_1)^2 - \rho\sigma_1\sigma_2(x_{1,i} - \mu_1)(x_{2,i} - \mu_2) \right. \\ &\qquad \left. + \sigma_1^2(x_{2,i} - \mu_2)^2 - \rho\sigma_1\sigma_2(x_{1,i} - \mu_1)(x_{2,i} - \mu_2) \Big) \right] \\ &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x_{1,i} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_{1,i} - \mu_1)(x_{2,i} - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_{2,i} - \mu_2)^2}{\sigma_2^2} \right) \right] \\ &= \exp\left[-\frac{1}{2(1-\rho^2)} \sum_{i=1}^{N} \left(\frac{(x_{1,i} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_{1,i} - \mu_1)(x_{2,i} - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_{2,i} - \mu_2)^2}{\sigma_2^2} \right) \right] \end{aligned} \tag{37}$$

Notice that since there is correlation between both variables (x_1, x_2), we cannot recognize this equation as the product of two independent Gaussian distributions.

On the other hand, we must define the prior pdf for μ_1, μ_2. Since we do not know anything about μ_1, μ_2, we assume that all the values are equally likely (uniform). This is expressed by $\pi(\mu_1, \mu_2)d\mu_1 d\mu_2 = Cd\mu_1 d\mu_2 \ \forall \mu_1, \mu_2$. Again, since C is a constant, it will cancel out. Then $P(\mu_1, \mu_2|D) \propto L(D|\mu_1, \mu_2)$, so the non-normalized posterior pdf for μ_1, μ_2 is

$$P(\mu_1, \mu_2|D) \propto \exp\left[-\frac{1}{2(1-\rho^2)}\sum_{i=1}^{N}\left(\frac{(x_{1,i}-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_{1,i}-\mu_1)(x_{2,i}-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_{2,i}-\mu_2)^2}{\sigma_2^2}\right)\right] \tag{38}$$

As we can observe, this Bayesian posterior pdf is elaborate and it has underflow problems. As a consequence, it is not possible and computationally efficient to sample (μ_1, μ_2) directly from $P(\mu_1, \mu_2|D)$. In order to deal with this, in the next parts III. and IV. we are going to use the Markov chain simulation method.

III. We are going to sample $P(\mu_1, \mu_2|D)$ using the Markov chain Monte Carlo (MCMC) with a Metropolis-Hastings sampler. The MCMC is a general method based on drawing values of random variables θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution $P(\theta|x)$. It is called a Markov chain because the sampling is done sequentially with the distribution of the sampled draws depending on the last value drawn. Then, at each step of the simulation, we expect to draw from a distribution that becomes closer to $p(\theta|x)$. The Metropolis-Hastings algorithm proceeds as follow:

1. Draw a starting point $\vec{\mu}^0 = (0, 0)$
2. For each iteration $t = 1, 2, ...$:
   (a) Sample a proposal $\vec{\mu}^*$ from a proposal distribution $J_t(\vec{\mu}^*; \vec{\mu}^{t-1})$ (this proposal do not have to be symmetric). As a proposal we are going to choose a bivariate normal distribution uncorrelated, i.e. the product of two independent Gaussian:

$$J_t(\vec{\mu}^*|\vec{\mu}^{t-1}) = \mathcal{N}\left(\vec{\mu}^*; \vec{\mu}^{t-1}, c^2\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right) \tag{39}$$

   Note that this distribution has the means of the last iteration. On the other hand, it presents the fixed variance matrix corresponding to the known values σ_1, σ_2. However, this variance matrix is scaled with the scale factor c. This factor has to be chosen properly to optimize the efficiency of the Metropolis algorithm. This is usually done by calculating the acceptance rate. The desired acceptance rate depends on the target distribution. It has been shown that the ideal acceptance for one-dimensional Gaussian distribution is approx 0.44, decreasing to 0.23 for a N-dimensional Gaussian target distribution. We will show the results for different values of c.

   (b) Calculate the ratio of densities

$$r = \frac{P(\vec{\mu}^*|D)}{P(\vec{\mu}^{t-1}|D)}\frac{J_t(\vec{\mu}^{t-1}|\vec{\mu}^*)}{J_t(\vec{\mu}^*|\vec{\mu}^{t-1})} \tag{40}$$

   Note that some terms cancel out. This is why we did not consider normalization constants before. Nevertheless, we must be careful when we are computing this expression. The posterior pdf $P(\vec{\mu}|D)$ alone has numerical underflow problems. Therefore, we must rearrange the computations in order to evade this. What do we do? We compute the logarithm of $P(\vec{\mu}^*|D)$ and $P(\vec{\mu}^{t-1}|D)$ separately and we subtract them. Then we apply the exponential to recover the original equality $(\exp(\log(x)) = x)$. This new result do not have underflow. After this, we multiply by the drawn values from the proposal distributions.

(c) Compute the acceptance probability

$$\alpha = \min(1, r) \tag{41}$$

It is clear that α can be 1 or less. If it is 1, then the acceptance ratio will be 1.00. Since we have $\vec{\mu} = (\mu_1, \mu_2)$, we will have two ratios (r_1, r_2) and then (α_1, α_2).

(d) Accept or reject. We generate a uniform random number u. If $u \leq \alpha_1$ and $u \leq \alpha_2$, we accept and set $\vec{\mu}^t = \vec{\mu}^*$. Otherwise we reject and set $\vec{\mu}^t = \vec{\mu}^{t-1}$.

We run this algorithm for a sufficiently large number of iterations (t). At each step we save the obtained (μ_1, μ_2) and we count the accepted steps. Then, we can compute the acceptance fraction as

$$\text{Acceptance ratio} = \frac{\text{accepted iterations}}{\text{number of iterations}} \tag{42}$$

Once we have the sample of (μ_1, μ_2), we still have to apply an appropriate burn-in period. This refers to the practice of discarding the early values in the Markov chain, i.e. the values that have not yet converged. This is done to diminish the influence of the starting value. Therefore, it is clear that the length of this period depends on the starting point and the scale factor c. Since we have chosen $\vec{\mu}^0 = (0, 0)$ which is close to the true value $\vec{\mu} = (2, 1)$, we expect a rapid convergence. Analysing the draws at each iteration, we find out how many iterations we have to discard. Finally, we compute the sample means and standard deviations of μ_1 and μ_2 from the remaining sample. The results are shown in Table 5.

c	Burn-in	Acc. ratio	$\overline{\mu}_1$	$\overline{\mu}_2$	$\hat{\sigma}_{\mu_1}$	$\hat{\sigma}_{\mu_2}$
0.1	120	0.0148	2.003	0.999	0.010	0.019
0.06	200	0.0365	2.005	1.000	0.010	0.021
0.01	1000	0.4102	2.005	1.001	0.010	0.020
0.001	7500	0.6907	2.004	0.998	0.007	0.016

Table 5: Values of the sample means and standard deviations of μ_1 and μ_2 for different values of the scale factor c and its corresponding acceptance ratio. We have applied a different burn-in for each c. Number of iterations = 10000. Starting point $\vec{\mu}^0 = (0, 0)$. All computations are performed with the same seed.

As we can observe (Table 5), if c is too large ($c = 0.1$) the chain converges rapid (small burn-in), but the acceptance ratio is very small so we recollect very few different values in the sample. On the other hand, if c is to small ($c = 0.001$) the acceptance ratio is high but the chain converges slowly (large burn-in). In any case, the results fit with the true values $\mu_1 = 2, \mu_2 = 1$ within the corresponding standard deviations.

Remember that the optimal acceptance ratio was between 0.44 and 0.23. In fact, $c = 0.01$ is within this range and it gives correct results, so we are going to choose this scale factor. It is considered the most optimal in the sense that we have a rapid convergence and a large acceptance ratio, so at the end it gives a sample with many different values. The plot of the sample of $P(\mu_1, \mu_2|D)$ is done with this c and is shown in Fig. 13.

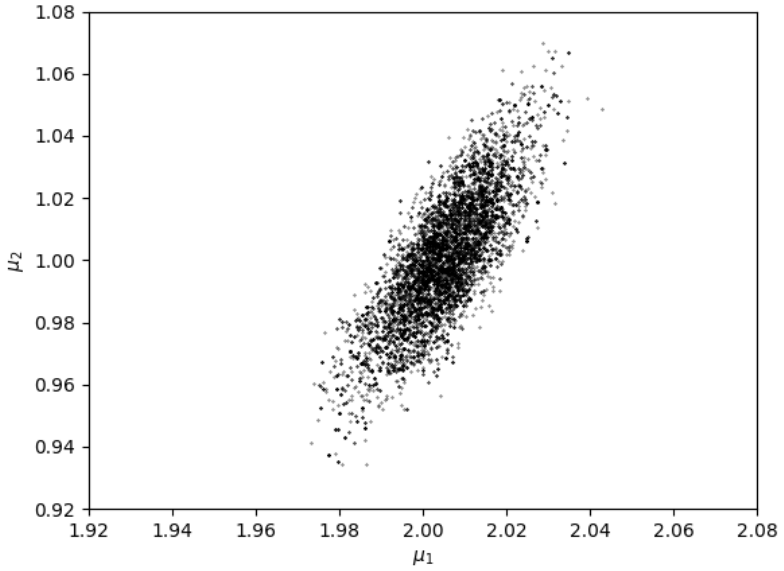

Figure 13: Scatter plot of μ_1 and μ_2 distributed as $P(\mu_1, \mu_2|D)$ using the Metropolis-Hastings sampler. Number of iterations = 10000. $c = 0.01$. It shows the points after the burn-in period.

IV. We are going to sample $P(\mu_1, \mu_2|D)$ using the Markov chain Monte Carlo (MCMC) with a Gibbs sampler. It is also known as conditional sampling. We apply this method when it is difficult to sample directly from the joint distribution, but it is easier to sample from the conditional distributions of each variable.

Given a parameter vector $\vec{\theta} = (\theta_1, ...\theta_d)$, at each iteration t an ordering of θ is chosen and each θ_j^t is sampled from the conditional distribution given all the other components of θ:

$$P(\theta_j^t|\theta_1^t, ...\theta_{j-1}^t, \theta_{j+1}^{t-1}, ...\theta_d^{t-1}, y) \tag{43}$$

In our case, the Gibbs sampler algorithm proceeds as follow:

1. Draw a starting point $\vec{\mu}^0 = (0, 0)$
2. For each iteration $t = 1, 2, ...$:
   (a) Choose an order $\vec{\mu} = (\mu_1, \mu_2)$
   (b) Sample μ_1^t from $P(\mu_1^t|\mu_2^{t-1}, D)$ and μ_2 from $P(\mu_2^t|\mu_1^t, D)$.

It is clear that we need the conditional posterior distributions. We are going to deduce them form the joint distribution (Eq. 38). We start deducing $P(\mu_1|\mu_2, D)$. In order to have a one dimensional pdf for μ_1, we fix μ_2. Then, the terms without μ_1 are constants, so we are not interested in them. We obtain

$$P(\mu_1|\mu_2, D) \propto \exp\left[-\frac{1}{2(1-\rho^2)}\sum_{i=1}^{N}\left(\frac{(x_{1,i}-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_{1,i}-\mu_1)(x_{2,i}-\mu_2)}{\sigma_1\sigma_2}\right)\right] \tag{44}$$

We take out σ_1^2 and we expand all terms:

$$P(\mu_1|\mu_2, D) \propto \exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\sum_{i=1}^{N}\left((x_{1,i})^2 - 2x_{1,i}\mu_1 + \mu_1^2 - \frac{2\rho\sigma_1}{\sigma_2}(x_{1,i}x_{2,i} - x_{1,i}\mu_2 - \mu_1 x_{2,i} + \mu_1\mu_2)\right)\right] \tag{45}$$

Again, we discard terms that do not depend on μ_1, we perform the sum in each term and we write in terms of the means $\overline{x}_1$ and $\overline{x}_2$:

$$P(\mu_1|\mu_2, D) \propto \exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\left(-2N\overline{x}_1\mu_1 + N\mu_1^2 - \frac{2\rho\sigma_1}{\sigma_2}(-\mu_1 N\overline{x}_2 + N\mu_1\mu_2)\right)\right] \tag{46}$$

Now we want to rewrite it as a Gaussian:

$$P(\mu_1|\mu_2, D) \propto \exp\left[-\frac{N}{2\sigma_1^2(1-\rho^2)}\left(\mu_1^2 - 2\mu_1\left(\overline{x}_1 + \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - \overline{x}_2)\right)\right)\right] \tag{47}$$

We add a constant:

$$\begin{aligned} P(\mu_1|\mu_2, D) &\propto \exp\left[-\frac{N}{2\sigma_1^2(1-\rho^2)}\left(\mu_1^2 - 2\mu_1\left(\overline{x}_1 + \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - \overline{x}_2)\right) + \left(\overline{x}_1 + \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - \overline{x}_2)\right)\right)\right] \\ &= \exp\left[-\frac{N}{2\sigma_1^2(1-\rho^2)}\left(\mu_1 - \left(\overline{x}_1 + \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - \overline{x}_2)\right)\right)^2\right] \end{aligned} \tag{48}$$

We can identify this last expression as a Gaussian. Therefore

$$P(\mu_1|\mu_2, D) \sim \mathcal{N}\left(\overline{x}_1 + \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - \overline{x}_2), \frac{\sigma_1^2(1-\rho^2)}{N}\right) \tag{49}$$

Analogously,

$$P(\mu_2|\mu_1, D) \sim \mathcal{N}\left(\overline{x}_2 + \frac{\rho\sigma_2}{\sigma_1}(\mu_1 - \overline{x}_1), \frac{\sigma_2^2(1-\rho^2)}{N}\right) \tag{50}$$

Once we have $P(\mu_1|\mu_2, D)$ and $P(\mu_2|\mu_1, D)$, we run the algorithm for a sufficiently large number of iterations (t). At each iteration, we save the obtained (μ_1, μ_2). Once we have the sample of (μ_1, μ_2), we apply an appropriate burn-in period. Analysing the draws at each iteration, we conclude that we only need to discard the first 50 iterations. Finally, we compute the sample means and standard deviations of μ_1, μ_2. The results are shown in Table 6.

$\overline{\mu}_1$	$\overline{\mu}_2$	$\hat{\sigma}_{\mu_1}$	$\hat{\sigma}_{\mu_2}$
2.005	1.001	0.010	0.020

Table 6: Values of the sample means and standard deviations of μ_1 and μ_2 using the Gibbs sampler. Number of iterations = 10000. Burn-in = 50. Starting point $\vec{\mu}^0 = (0, 0)$. We have used the same seed as before.

As we can observe (Table 6), the results fit with the true values $\mu_1 = 2, \mu_2 = 1$ within the corresponding standard deviations. Furthermore, we notice that the results are equal to the ones obtained using the Metropolis-Hastings sampler with $c = 0.01$. We can also conclude that this method converges faster than the Metropolis-Hastings sampler, since we have only required to discard 50 iterations. The plot of the sample of $P(\mu_1, \mu_2|D)$ is shown in Fig. 14.

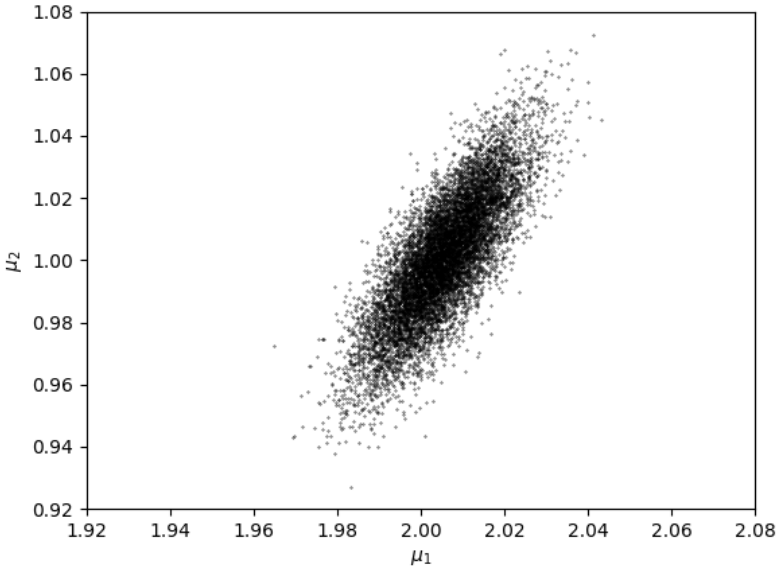

Figure 14: Scatter plot of μ_1 and μ_2 distributed as $P(\mu_1, \mu_2|D)$ using the Gibbs sampler. Number of iterations = 10000. It shows the points after the burn-in period.

V. We are going to repeat all the calculations with a different correlation coefficient $\rho = 0.999$. Since this correlation is approx 1, we should expect that the two variables x_1 and x_2 are practically having a linear relationship.

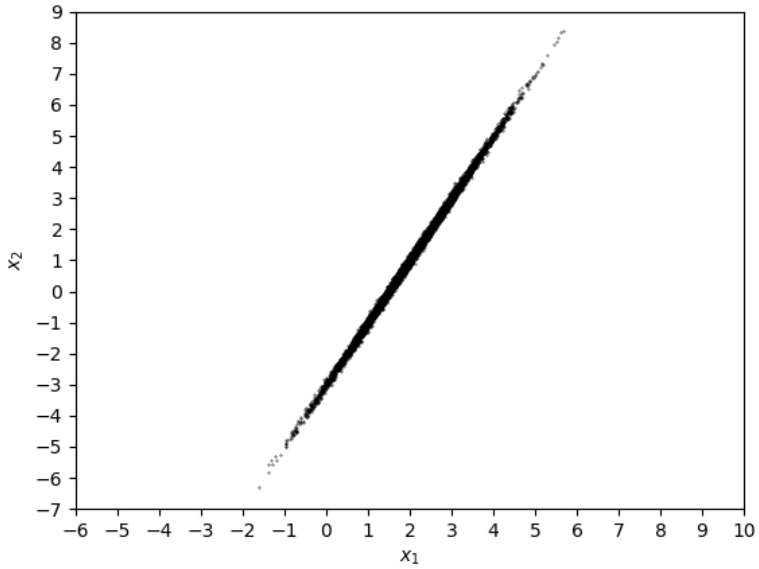

Figure 15: Scatter plot of the random variables x_1 and x_2 distributed as a 2D Gaussian $\mathcal{N}(\vec{\mu}, V)$ with positive correlation $\rho = 0.999$. $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 2$.

Metropolis-Hastings sampler: We have to readjust the scale factor c in order to have an optimal acceptance ratio. We have chosen $c = 0.001$. Since c is small we expect a slow convergence (large burn-in). However, despite the value of c, we notice that the convergence with $\rho = 0.999$ is slower than the convergence with $\rho = 0.8$. As we can observe (Table 7), the results fit with the true values $\mu_1 = 2, \mu_2 = 1$ within the corresponding standard deviations. Notice (Fig. 16) that μ_1 and μ_2 are practically having a linear relationship as expected.

c	Burn-in	Acc. ratio	$\overline{\mu}_1$	$\overline{\mu}_2$	$\hat{\sigma}_{\mu_1}$	$\hat{\sigma}_{\mu_2}$
0.001	15000	0.3758	2.003	1.006	0.007	0.015

Table 7: Values of the sample means and standard deviations of μ_1 and μ_2 using the Metropolis-Hastings sampler. Number of iterations = 30000. Starting point $\vec{\mu}^0 = (0, 0)$. We have used the same seed as before.

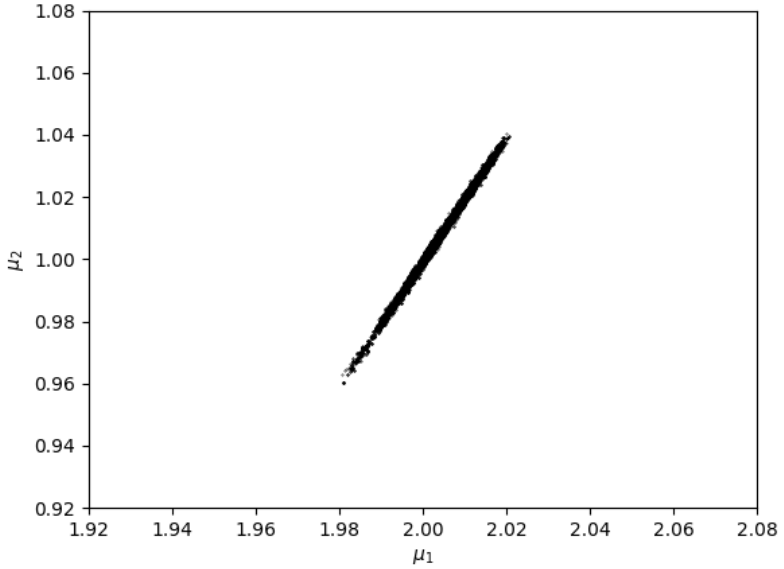

Figure 16: Scatter plot of μ_1 and μ_2 distributed as $P(\mu_1, \mu_2|D)$ using the Metropolis-Hastings sampler. Number of iterations = 30000. $c = 0.001$. It shows the points after the burn-in period.

Gibbs sampler: Again, we have to apply a larger burn-in period. The results (Table 8) fit with the true values $\mu_1 = 2, \mu_2 = 1$ within the corresponding standard deviations and coincide with the ones obtained with the Metropolis-Hastings sampler.

$\overline{\mu}_1$	$\overline{\mu}_2$	$\hat{\sigma}_{\mu_1}$	$\hat{\sigma}_{\mu_2}$
2.001	1.001	0.008	0.015

Table 8: Values of the sample means and standard deviations of μ_1 and μ_2 using the Gibbs sampler. Number of iterations = 30000. Burn-in = 15000. Starting point $\vec{\mu}^0 = (0, 0)$. We have used the same seed as before.

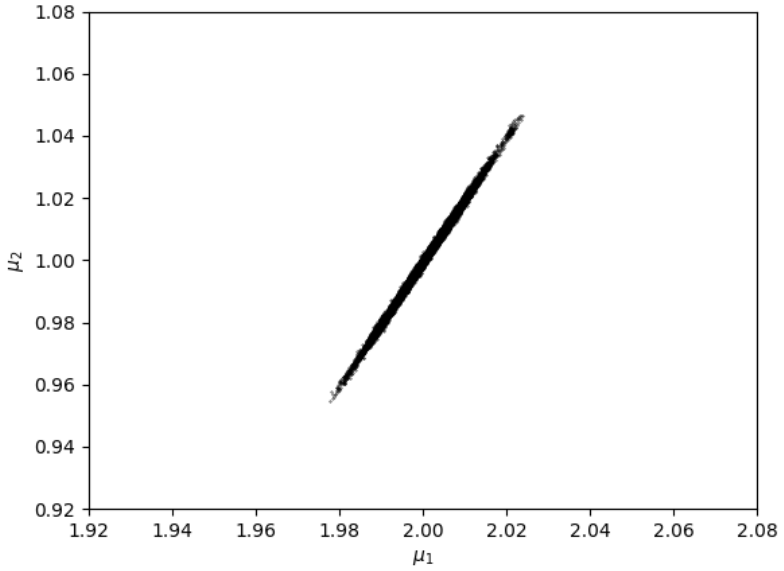

Figure 17: Scatter plot of μ_1 and μ_2 distributed as $P(\mu_1, \mu_2|D)$ using the Gibbs sampler. Number of iterations = 30000. It shows the points after the burn-in period.