Technical University of Denmark

Bayesian machine learning

Project Bayesian Machine Learning

Students :
Emmanuel Minois-Genin s233427
Hongjin Chen s232289
Jialu Chen Christiansen s194175
Chuang Sun Hembo s233427
Changrong Xue s232242

February 25, 2024

Contents

1 Part 1: The beta-binomial model **2**

2 Part 2: The sum and product rule for a simple toy model **4**

3 Part 3: A simple Linear Gaussian system **6**

4 Part 4: Bayesian inference for two-parameter model **8**

1 Part 1: The beta-binomial model

Task 1.1 We use a beta-binomial model with an uniform prior distrubtion for the parameter $\mu \in [0, 1]$ which parametrize the binomial distribution y. Therefore we have $\mu \sim \text{Beta}(a_0, b_0)$ and $y \sim \text{Binomial}(N, \mu)$, where N is the number of potential customers who visited the website and y is the number of resuling purchasers. Since we are using a uniform prior on $[0, 1]$ we now that $a_0 = 1$ and $b_0 = 1$ for the beta distribution μ. The posterior distribution $p(\mu|y)$ is calculated thanks to Bayes rule, we have:

$$p(\mu|y) = \frac{p(y|\mu)p(\mu)}{p(y)}, \tag{1}$$

$$p(\mu|y) = \frac{\binom{N}{y}\mu^y(1-\mu)^{N-y}\frac{1}{B(a_0,b_0)}\mu^{a_0-1}(1-\mu)^{b_0-1}}{p(y)} \tag{2}$$

$$= \frac{\binom{N}{y}\mu^y(1-\mu)^{N-y}\mu^0(1-\mu)^0}{B(1,1)p(y)} \tag{3}$$

Where $B(1, 1)$ is a scale constant derived from the Gamma function to ensure that the probability integrates to 1. We can see that we indeed have a uniform prior $\mu^0(1-\mu)^0 = 1$, the prior is uninformative, it does not influence the posterior (it only scale it so it integrates to 1 over μ). In the end we have:

$$p(\mu|y) = \frac{\mu^y(1-\mu)^{N-y}}{const} \tag{4}$$

Where $const = \frac{B(1,1)p(y)}{\binom{N}{y}}$, we recognize a beta distribution with $a_0 = y + 1 = 2$ and $b_0 = N - y + 1 = 17$, since the probability must integrate to 1 w.r.t μ we therefore have $const = \frac{B(1,1)p(y)}{\binom{N}{y}} = B(2, 17)$. The posterior distribution of μ is $p(\mu|y) = \text{Beta}(2, 17)$.

Task 1.2 We can use the scipy package to get the 95% posterior interval for μ, and the mean of the beta distribution is given by $\frac{a_0}{a_0+b_0}$, we find the posterior mean is given by:

$$\mathbb{E}[\mu|y] = \frac{a}{a+b} = \frac{a_0+y}{a_0+b_0+N} = \frac{2}{19} \approx 0.1053 \tag{5}$$

and $p(0.0138 < \mu < 0.2729) = 0.95$.

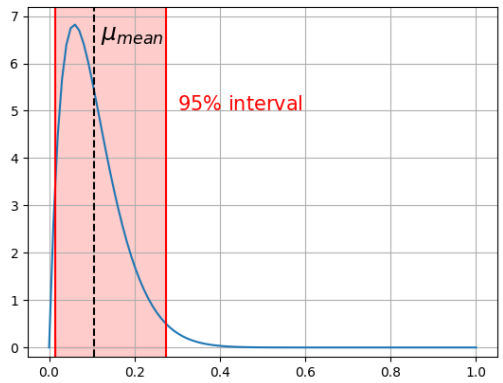

Figure 1: 95% credibility interval and posterior mean for Beta$(2, 17)$

Task 1.3 Our prior is now the posterior distribution from Task 1.1, ie $p(\mu) \sim \text{Beta}(2, 17)$, and our new knowledge is $N_2 = 20$ and $y_2 = 4$, using the same formulas as in Task 1.1, we find that the posterior distribution is $p(\mu|y) \sim \text{Beta}(6, 33)$ (notice that this time the prior is informative, modifying the posterior distribution), we can compute the new mean and the new interval, we get $\mu_{mean} = \mathbb{E}[\mu] = \frac{6}{39} \approx 0.1538$, and $p(0.0602 < \mu < 0.2809) = 0.95$.

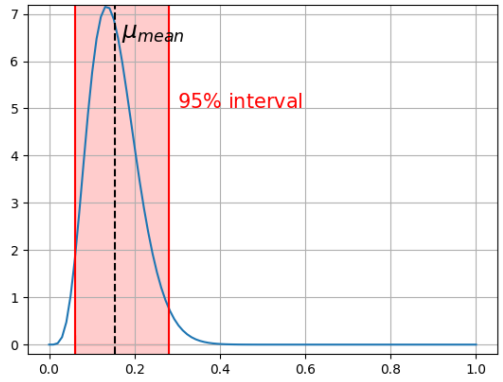

Figure 2: 95% credibility interval and posterior mean for Beta$(6, 33)$

Task 1.4 We have the log probability density function which is defined over the unit interval, we can use exponential properties to gind the probability density function:

$$\ln p(\mu) = 95 \ln \mu + 10 \ln(1 - \mu) + c, \tag{6}$$

$$p(\mu) = \mathrm{e}^{\mathrm{c}} \times \mu^{95}(1 - \mu)^{10} \tag{7}$$

Where $\exp\{c\}$ is a constant independent of μ, since this probability should integrate to 1 w.r.t μ we can regonize a scaling constant and even find that $\exp\{c\} = \frac{1}{B(96,11)}$, indeed we can recognize the functional form of a beta distribution, since we are looking at $\mu \in [0, 1]$

and the functional form of a beta distribution is $p(y) \propto y^{a_0-1}(1-y)^{b_0-1}$. We therefore have $\mu \sim \text{Beta}(96, 11)$, we can find the mean $\mu_{mean} = \mathbb{E}[\mu] = \frac{a_0}{a_0+b_0} = \frac{96}{107} \approx 0.8972$.

2 Part 2: The sum and product rule for a simple toy model

Task 2.1 The prior is defined as a Bernoulli distribution with parameter α, we can therefore find the mean of this discrete distribution:

$$\mathbb{E}[x] = \sum xp(x) = 1 \times \alpha + 0 \times (1-\alpha) = \alpha \tag{8}$$

Hence, $\mathbb{E}[x] = \alpha$.

Task 2.2 Using the product rule we have:

$$p(y, x) = p(y|x)p(x) = p(x|y)p(y) \tag{9}$$

we see the lileihood $p(y|x)$ and the prior $p(x)$ appearing in the first formula.

Task 2.3 We can use the sum rule to marginalize $p(y) = \sum_x p(y, x) = \sum_x p(y|x)p(x)$, using **Task 2.2**. We can replace with the distributions of x and y,

$$p(y) = \sum_x \text{Ber}(x|\alpha)\mathcal{N}(y|2x, \sigma_y^2) \tag{10}$$

$$= \alpha\mathcal{N}(y|2, \sigma_y^2) + (1-\alpha)\mathcal{N}(y|0, \sigma_y^2) \tag{11}$$

The distribution of y is a linear combination of 2 normal distributions with different means.

Task 2.4 We can use the linearity of expectation to get

$$\mathbb{E}[y] = \mathbb{E}[\alpha\mathcal{N}(y|2, \sigma_y^2) + (1-\alpha)\mathcal{N}(y|0, \sigma_y^2)] \tag{12}$$

$$= \alpha\mathbb{E}[\mathcal{N}(y|2, \sigma_y^2)] + (1-\alpha)\mathbb{E}[\mathcal{N}(y|0, \sigma_y^2)] \tag{13}$$

$$= 2 \times \alpha \tag{14}$$

as $\mathbb{E}[\mathcal{N}(\mu, \sigma^2)] = \mu$.

Task 2.5 We want to find the second moment of y,

$$\mathbb{E}[y^2] = \mathbb{E}[\alpha^2\mathcal{N}(y|2, \sigma_y^2)^2 + 2\alpha(1-\alpha)\mathcal{N}(y|2, \sigma_y^2)\mathcal{N}(y|0, \sigma_y^2) + (1-\alpha^2)^2\mathcal{N}(y|0, \sigma_y^2)^2] \tag{15}$$

we can use the formula for the second moment of a normal distribution $\mathbb{E}[\mathcal{N}(y|\mu, \sigma_y^2)^2] = \mu^2 + \sigma_y^2$. Therefore

$$\mathbb{E}[y^2] = \alpha^2(4 + \sigma_y^2) + (1-\alpha)^2\sigma_y^2 \tag{16}$$
$$= (4 + 2\sigma_y^2)\alpha^2 - 2\sigma_y^2\alpha + \sigma_y^2 \tag{17}$$

where we used the fact that the two normal disitrubtions are independent $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Task 2.6 The variance is

$$\mathbb{V}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 \tag{18}$$
$$= (4 + 2\sigma_y^2)\alpha^2 - 2\sigma_y^2\alpha + \sigma_y^2 - 4\alpha^2 \tag{19}$$
$$= 2\sigma_y^2\alpha^2 - 2\sigma_y^2\alpha + \sigma_y^2 \tag{20}$$
$$= \sigma_y^2(1 - 2\alpha(1-\alpha)) \tag{21}$$

Task 2.7

$$p(x=1|y) = \frac{p(x=1,y)}{p(y)} \tag{22}$$
$$= \frac{\alpha\mathcal{N}(y|2,\sigma_y^2)}{\alpha\mathcal{N}(y|2,\sigma_y^2) + (1-\alpha)\mathcal{N}(y|0,\sigma_y^2)} \tag{23}$$

we divide the numerator and the denominator with $\alpha\mathcal{N}(y|2,\sigma_y^2)$, and we get

$$p(x=1|y) = \frac{1}{1 + \frac{(1-\alpha)\mathcal{N}(y|0,\sigma_y^2)}{\alpha\mathcal{N}(y|2,\sigma_y^2)}} \tag{24}$$

Task 2.8 Using the result from the previous task we have

$$p(x=1|y=1.5) = \frac{1}{1 + \frac{(1-0.5)\mathcal{N}(1.5|0,0.5)}{0.5\mathcal{N}(1.5|2,0.5)}} \tag{25}$$
$$\approx 0.982 \tag{26}$$

If the noise variance $\sigma_y^2 = 5$ we get $p(x=1|y=1.5) = 0.510$, which makes sense because if we increase the noise in the normal distributions, it becomes difficult to know from which distribution the result is coming from.

3 Part 3: A simple Linear Gaussian system

Task 3.1 This time, the prior disitribution is replace by a normal disitribution. Therefore

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{27}$$

$$= \frac{\mathcal{N}(y|2x, \sigma_y^2)\mathcal{N}(x|m_x, \sigma_x^2)}{p(y)} \tag{28}$$

Since $p(y)$ is independent of x, taking the log of $p(x|y)$ gives

$$\log p(x|y) = \log \mathcal{N}(y|2x, \sigma_y^2) + \log \mathcal{N}(x|m_x, \sigma_x^2) + K \tag{29}$$

Where $K = -\log p(y)$ is a constant with respect to x.

Task 3.2 Since, for a normal distribution we have

$$p(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{30}$$

We can use this expression in $p(x|y)$, we get

$$\log p(x|y) = \log \frac{1}{\sigma_y\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-2x)^2}{2\sigma_y^2}} + \log \frac{1}{\sigma_x\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-m_x)^2}{2\sigma_x^2}} + K \tag{31}$$

$$= -\frac{(y-2x)^2}{2\sigma_y^2} - \frac{(x-m_x)^2}{2\sigma_x^2} - \log \sigma_y\sqrt{2\pi} - \log \sigma_x\sqrt{2\pi} + K \tag{32}$$

since σ_x and σ_y are constants independent of x we have

$$\log p(x|y) = -\frac{(y-2x)^2}{2\sigma_y^2} - \frac{(x-m_x)^2}{2\sigma_x^2} + K_1 \tag{33}$$

$$K_1 = -\log \sigma_y\sqrt{2\pi} - \log \sigma_x\sqrt{2\pi} + K \tag{34}$$

Task 3.3 We develop the parentheses

$$\log p(x|y) = -\frac{(y-2x)^2}{2\sigma_y^2} - \frac{(x-m_x)^2}{2\sigma_x^2} + K_1 \tag{35}$$

$$= -\frac{y^2 - 4yx + 4x^2}{2\sigma_y^2} - \frac{x^2 - 2m_x x + m_x^2}{2\sigma_x^2} + K_1 \tag{36}$$

$$= -\frac{1}{2}x^2\left(\frac{2^2}{\sigma_y^2} + \frac{1}{\sigma_x^2}\right) + x\left(\frac{2y}{\sigma_y^2} + \frac{m_x}{\sigma_x^2}\right) - \frac{y^2}{2\sigma_y^2} - \frac{m_x^2}{2\sigma_x^2} + K_1 \tag{37}$$

since $\frac{y^2}{2\sigma_y^2}$ and $\frac{m_x^2}{2\sigma_x^2}$ are constants with respect to x, we have

$$\log p(x|y) = -\frac{1}{2}x^2(\frac{2^2}{\sigma_y^2} + \frac{1}{\sigma_x^2}) + x(\frac{2y}{\sigma_y^2} + \frac{m_x}{\sigma_x^2}) + K_2 \tag{38}$$

$$K_2 = -\frac{y^2}{2\sigma_y^2} - \frac{m_x^2}{2\sigma_x^2} + K_1 \tag{39}$$

Task 3.4 We recognize that $\log p(x|y)$ is a concave quadratic function, which is the functional form of a Gaussian distribution, there must be some m and v such that $p(x|y) \sim \mathcal{N}(m, v)$.

Task 3.5 We can take the general functional form of a univariate Gaussian $p(x|m, v) = -\frac{1}{2v}x^2 + \frac{m}{v}x + K$ and identify the coefficients for $p(x|y)$, therefore we have $v = (\frac{2^2}{\sigma_y^2} + \frac{1}{\sigma_x^2})^{-1}$, which is equivalent to

$$v^{-1} = \frac{2^2}{\sigma_y^2} + \frac{1}{\sigma_x^2} \tag{40}$$

Task 3.6 We use the same procedure for m, indeed $\frac{m}{v} = \frac{2y}{\sigma_y^2} + \frac{m_x}{\sigma_x^2}$, we can isolate m and develop the expression

$$m = (\frac{2y}{\sigma_y^2} + \frac{m_x}{\sigma_x^2})v \tag{41}$$

$$= \frac{\frac{2y}{\sigma_y^2} + \frac{m_x}{\sigma_y^2}}{\frac{2^2}{\sigma_y^2} + \frac{1}{\sigma_x^2}} \tag{42}$$

We multiply both the numerators and the denominators with σ_y^2,

$$m = \frac{2y + \frac{\sigma_y^2 m_x}{\sigma_x^2}}{2^2 + \frac{\sigma_y^2}{\sigma_x^2}} \tag{43}$$

$$= \frac{2}{2^2 + \frac{\sigma_y^2}{\sigma_x^2}}y + \frac{\frac{\sigma_y^2}{\sigma_x^2}}{2^2 + \frac{\sigma_y^2}{\sigma_x^2}}m_x \tag{44}$$

$$= \frac{2}{2^2 + \frac{\sigma_y^2}{\sigma_x^2}}y + \frac{1}{\frac{2^2\sigma_x^2}{\sigma_y^2} + 1}m_x \tag{45}$$

Task 3.7 If $\sigma_y^2 \to \infty$ and σ_x^2 is fixed, we have $\frac{\sigma_y^2}{\sigma_x^2} \to \infty$ and $\frac{\sigma_x^2}{\sigma_y^2} \to 0$. Which means that the coefficient before y will tend toward 0 whereas the coefficient before m_x will tend

toward 1, ie $m \to m_x$. It is what we expect because if the variance of the data is high our model would rely mainly on the prior distribution.

Task 3.8 Same reasonning as before but with swapped roles between σ_x^2 and σ_y^2, this time the coefficient before m_x tends toward 0 and the coefficient before y tends toward $\frac{1}{2}$ which means that the posterior $m \to \frac{y}{2}$. It is what we expect when the variance of the prior is too high we only rely on the information coming from the data (the prior is uninformative).

4 Part 4: Bayesian inference for two-parameter model

Task 4.1 We implemented a function to evaluate the log prior, the likelihood and the joint distribution and made a contour plots of the three distributions as can be seen from figure 3. Note that the posterior distribution has the same contour plot as the joint distribution since they are equal up to a constant (the evidence $p(y)$).

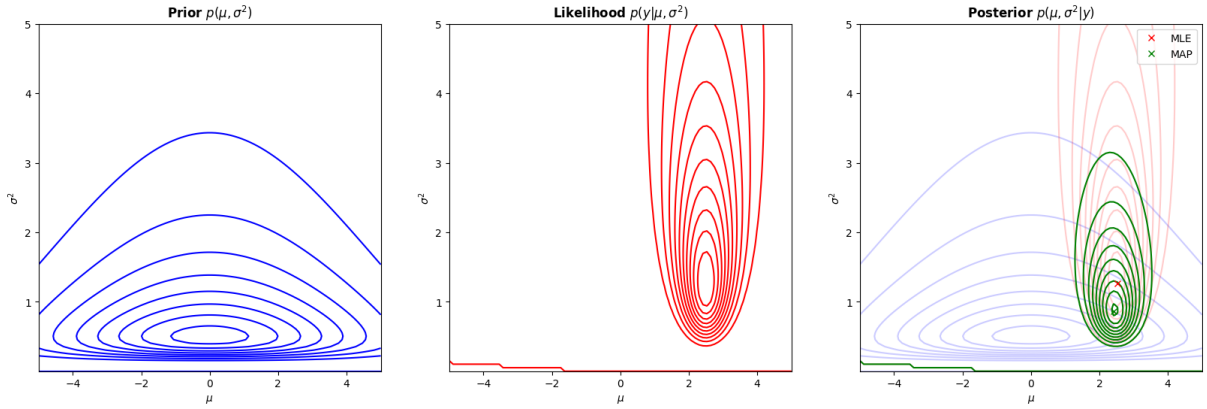

Figure 3: Contour plots of the prior, the likelihood and the posterior distribution

Task 4.2 Since we have no analytical solution for the integral of the product of a Gaussian function and an inverse gamma function, we have to use numerical methods to approximate the posterior. We implemented grid approximation to approximate the posterior $p(\mu, \sigma^2|y)$, as shown in figure 4.

The idea behind grid approximation is simple, we specify an interval of σ^2 and μ that we think covers most of the posterior distribution (we can use the contour plot of the posterior) and then evaluate the joint probability at a large number of coordinates $[\mu, \sigma^2]$ (hence the term grid). Once we have these values we normalize them so that they add up to 1 to get the approximate posterior distribution. We can sum up the method with the following equation, for $[\mu_k, \sigma_k^2]$ in the grid we have:

$$q(\mu_k, \sigma_k^2) \approx p(\mu_k, \sigma_k^2|y) \tag{46}$$

$$= \frac{p(\mu_k, \sigma_k^2)}{\sum_{[\mu_i, \sigma_j^2] \in \text{Grid}} p(\mu_i, \sigma_j^2)} \tag{47}$$

Where Grid is the ensemble of coordinates we are using for the approximation.

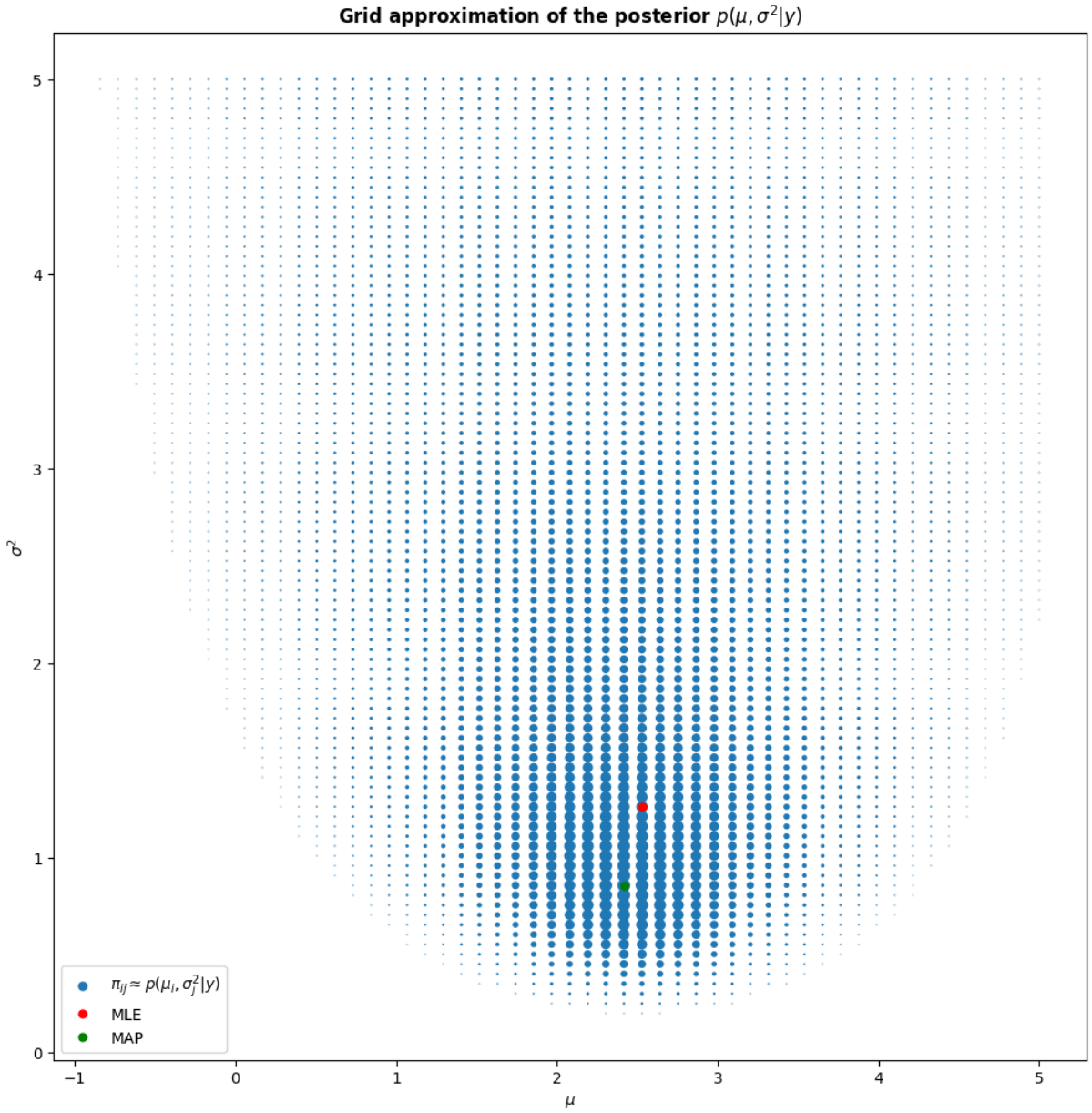

Figure 4: Grid approximation of the posterior distribution $p(\mu, \sigma^2|y)$ (only the probabilities $\geq 1e-6$ are plotted)

From the approximate posterior $q(\mu, \sigma^2)$ we can also approximate $p(\mu|y)$ and $p(\sigma^2|y)$

by marginalizing out the other parameter:

$$p(\mu_k|y) = \sum_{i \in Grid} q(\mu_k, \sigma_i^2) \tag{48}$$

$$p(\sigma_k^2|y) = \sum_{i \in Grid} q(\mu_i, \sigma_k^2) \tag{49}$$

Where we are summing over all the possible σ^2 values in the grid in 48 (same thing with μ values in 49). The resulting distributions are plotted in figure 5.

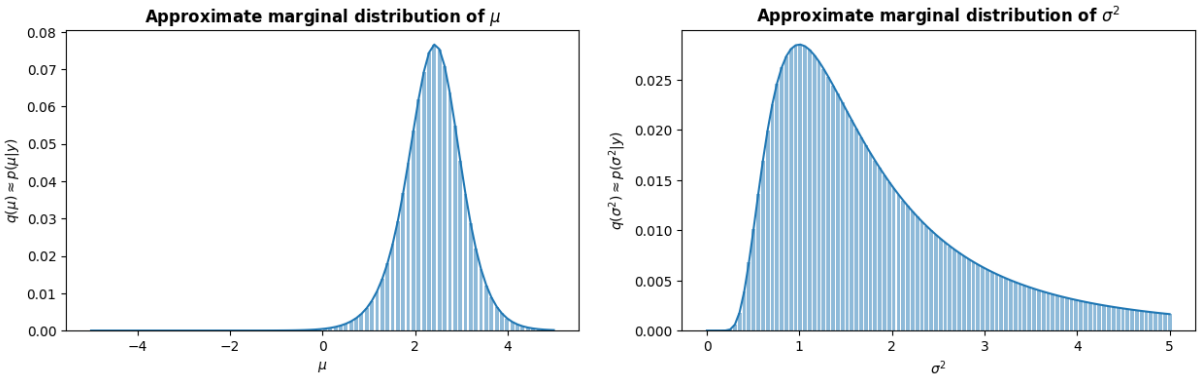

Figure 5: Approximate marginal posterior distributions (the result from grid approximation is a discrete distribution, we also plot the kernel density estimation to better vizualize the shape of the distribution)

Task 4.3 We can compute the mean of the approximate marginal posterior distributions as well as 95% credibility intervals. Using the results of the grid approximation. Since we are working with discrete distributions, we have $\mathbb{E}[\mu|y] = \sum \mu q(\mu|y)$, we use the quantile function Q to get the (centered) credibility interval as $p(Q(0.025) < \mu < Q(0.975)) = 0.95$. We can do the same operations with $sigma^2$. We obtain the results shows in table 1:

Variable	Mean	95% credibility interval
μ	2.39	[1.07, 3.65]
σ^2	1.78	[0.56, 4.39]

Table 1: Summary of posterior mean and credibility interval

Task 4.4 We can get our posterior predictive distribution for our new observation y^*:

$$p(y^*|y) = \iint p(y^*|\mu, \sigma^2) p(\mu, \sigma^2|y) \mathrm{d}\mu \mathrm{d}\sigma^2 \tag{50}$$

We can see that this integration averages the likelihood of the new observation y^* across all possible parameter values, weighted by their posterior probability given the observed data y. We can use MLE and MAP estimates to perform plugin approximations (which

means we approximate the posterior distribution of μ and σ^2 as a product of two Diracs):

$$p(y^*|y) = p(y^*|\mu_{MLE}, \sigma^2_{MLE}) \tag{51}$$
$$p(y^*|y) = p(y^*|\mu_{MAP}, \sigma^2_{MAP}) \tag{52}$$

But since we have an approximate of the posterior distribution $q(\mu, \sigma^2)$ we can actually perform a full Bayesian approach: 6.

$$p(y^*|y) = \sum_{\mu,\sigma^2} p(y^*|\mu, \sigma^2) q(\mu, \sigma^2) \tag{53}$$

We can see in 53 that we are summing instead of integrating, that is because our approximate posterior is a discrete distribution (wich could be seen as a Dirac comb) resulting in the given sum. The results we get for the different modelling approach for $p(y^*|y)$ are reported in fig 6:

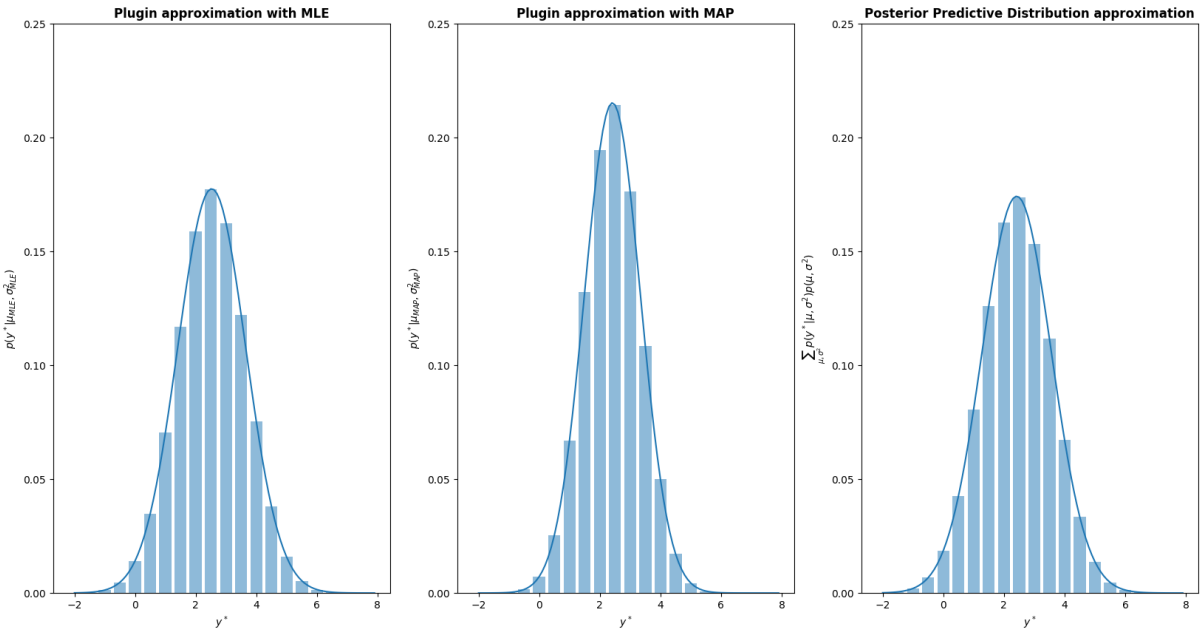

Figure 6: Posterior predictive density for y^* using plugin approximations (MLE and MAP) as well as a fully Bayesian approach (approximated as we used grid approximation to get the posterior distribution)

As we can see in the resulting distributions, the plugin approximation with MAP is the more confident one (sharpest distribution) which could indicate that the prior is in line with the data (thus this would increase the predictor confidence). Whereas both the MLE distribution and the Posterior Predictive Distribution are less confident in their predictions. We can see that the distribution most open to new data points is the posterior predictive distribution, as would be expected in a fully Bayesian approach. Furthermore we can see the influence of the prior in both the MAP approximation and the posterior

predictive distribution (the shape is similar although the variance differs) whereas the MLE approximation is different (because no priors).

To better understand the difference between the distributions we can for example visualize the difference between two distributions, for example in fig 7:

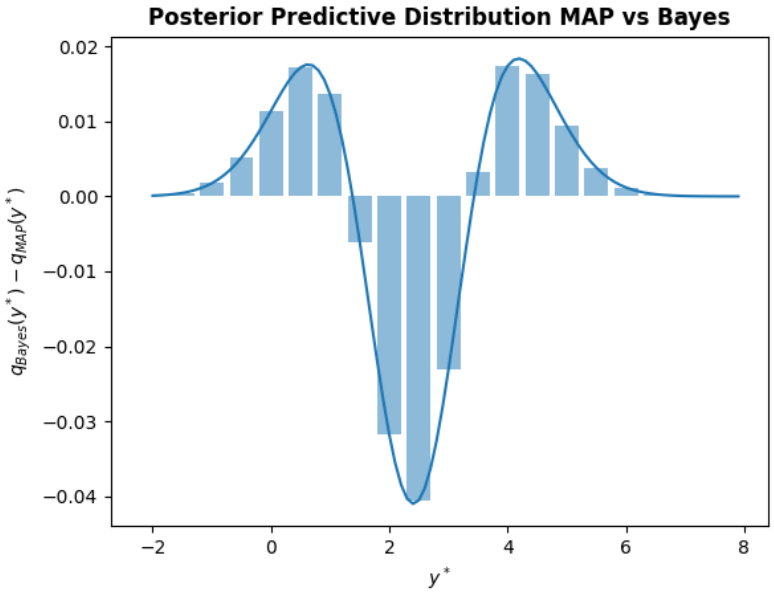

Figure 7: Difference between the posterior predictive distribution and the MAP approximation

We can clearly see how the MAP approximation gives a really high probability for values already observed in the dataset (data within $[1, 4]$), whereas posterior predictive distribution gives higher probability (compared to MAP approximation) for points lying outside of the data range values.