Chapter 5

Bayesian approach

In Bayesian approach, we will write $p(x|\theta)$ instead of $p_\theta(x)$, which is to reflect the fact that θ should be regarded as a random variable itself and not a fixed but unknown parameter as in the frequentist approach. Within the Bayesian approach, apart from $p(x|\theta)$, we need to specify prior distribution $p(\theta)$, which reflects our knowledge of the parameter which we have prior to performing any experiment.

In frequentist approach our goal was to minimize the estimation variance, as given in Eq. (4.3), with local unbiasedness condition imposed. In Bayesian approach the goal is to minimize the average variance:

$$\Delta^2\tilde{\theta} = \int \mathrm{d}\theta\, p(\theta) \int \mathrm{d}x \, \left(\tilde{\theta}(x) - \theta\right)^2 p(x|\theta). \tag{5.1}$$

In this case there is no need to impose any additional requirements such as unbiasedness. We simply look for an estimator $\tilde{\theta}$ that minimizes the above quantity.

5.1 Optimal Bayesian estimator

Let us rewrite the formula for the average variance, using the Bayes rule $p(x|\theta)p(\theta) = p(\theta|x)p(x)$, as follows:

$$\Delta^2\tilde{\theta} = \int \mathrm{d}x\, p(x) \int \mathrm{d}\theta, \left(\tilde{\theta}(x) - \theta\right)^2 p(\theta|x). \tag{5.2}$$

Since $p(x) \geq 0$, and $\tilde{\theta}(x)$ for different x can be treated as independent variables, minimization over $\tilde{\theta}$ amounts to minimization of $\int \mathrm{d}\theta, \left(\tilde{\theta}(x) - \theta\right)^2 p(\theta|x)$ quantity independently for each x over $\tilde{\theta}(x)$. This is a quadratic function in $\tilde{\theta}(x)$ and hence minimization is straightforward, as

$$\frac{\mathrm{d}}{d\tilde{\theta}(x)} \int \mathrm{d}\theta \, \left(\tilde{\theta}(x) - \theta\right)^2 p(\theta|x) = 0 \tag{5.3}$$

implies

$$\tilde{\theta}(x) = \int \mathrm{d}\theta\, p(\theta|x)\theta = \langle\theta\rangle_{p(\theta|x)}. \tag{5.4}$$

Hence the optimal Bayesian estimator corresponds to the the mean of the posteriori distribution $p(\theta|x)$. The corresponding minimal cost reads:

$$\Delta^2\tilde{\theta} = \int \mathrm{d}x p(x) \int \mathrm{d}\theta, \left(\langle\theta\rangle_{p(\theta|x)} - \theta\right)^2 p(\theta|x) =$$
$$= \int \mathrm{d}x\, p(x)\, \Delta^2\theta\Big|_{p(\theta|x)} \tag{5.5}$$

and amount to the average variance of the posteriori distribution.

Figure 5.1: Bayesian update of a prior distribution to a posteriori distribution based on data obtained

It is therefore clear that the fundamental object in the Bayesian approach is the posteriori distribution $p(\theta|x)$. This can be calculated via Bayes rule:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}. \tag{5.6}$$

Note, however, that $p(x)$ is not explicitly given, and calculating it requires performing the following integral: $p(x) = \int \mathrm{d}\theta\, p(x|\theta)p(\theta)$.

$p(x)$ plays a role of a normalization factor for the distribution, while the θ dependence is determined by the product of $p(x|\theta)$ and $p(\theta)$. To get a better intuitive understanding of the Bayesian approach, observe that $p(\theta)$ represents just the prior knowledge while all the information that we get from the data is captured by $p(x|\theta)$. Analyzing these two function one my easily understand what is the relative role of the prior information vs data. If $p(\theta)$ varies much slower with θ compared to $p(x|\theta)$ it means that the prior is largely irrelevant. In the opposite case the the prior dominates our inference strategy.

Example 5.1 Let us again reconsider the gaussian estimation model, where our observations are modeled as N i.i.d. random variables $x_i \sim \mathcal{N}(\theta, \sigma^2)$, from which we want to estimate θ. However, this time we further assume that we have a gaussian prior distribution of the θ parameter itself $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$. For the model presented we want to find the optimal Bayesian estimator and the resulting estimation uncertainty.

For this model:

$$p(\boldsymbol{x}|\theta)p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2}G_{\boldsymbol{x}}(\theta)}, \tag{5.7}$$

where $G_{\boldsymbol{x}}(\theta) = \frac{1}{\sigma_\theta^2}(\theta - \mu_\theta)^2 + \frac{1}{\sigma^2}\sum_i (x_i - \theta)^2$.

We immediately see that the posteriori distribution $p(\theta|\boldsymbol{x}) \sim p(\boldsymbol{x}|\theta)p(\theta)$ will also be Gaussian. As a result we can easily normalize it and arrive at the final form of the posteriori distribution:

$$p(\theta|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma_{\theta|\boldsymbol{x}}^2}} e^{-\frac{1}{2\sigma_{\theta|\boldsymbol{x}}^2}\left(\theta - \mu_{\theta|\boldsymbol{x}}\right)^2}, \tag{5.8}$$

where

$$\sigma_{\theta|\boldsymbol{x}}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}, \tag{5.9}$$

$$\mu_{\theta|\boldsymbol{x}} = \left(\frac{\sum_i x_i}{\sigma^2} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\sigma_{\theta|\boldsymbol{x}}^2 \tag{5.10}$$

are respectively the variance and the mean of the posteriori distribution.

The optimal Bayesian estimator which is the mean of the posteriori distribution $\mu_{\theta|\boldsymbol{x}}$ may be rewritten in a more appealing form

$$\tilde{\theta}(\boldsymbol{x}) = \alpha\bar{x} + (1-\alpha)\mu_\theta, \tag{5.11}$$

where $\bar{x} = \sum_i x_i/N$, $\alpha = \frac{N}{\sigma^2}/\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right)$. The above form clearly shows that the optimal Bayesian estimator arises as a result of compromise between what the data suggest (in this case the mean of observed values) and the prior information (in this case the mean of the prior μ_θ) and α represents the weight of the information part.

According to (5.5) the resulting cost will be the average of the posteriori variance. In our model the variance of the posteriori distribution does not depend on $\boldsymbol{x}$ and hence we may immediately write:

$$\Delta^2\tilde{\theta} = \sigma_{\theta|x}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}. \tag{5.12}$$

In the limit $N \to \infty$ the role of the prior becomes irrelevant and as a result $\tilde{\theta}(\boldsymbol{x}) \to \bar{x}$ and $\Delta^2\tilde{\theta} \to \sigma^2/N$, which coincides with the results obtained within the frequentist approach.

5.2 Bayesian Cramér-Rao bound

Even though, unlike in the frequentist approach, the recipe for the optimal Bayesian estimator is explicit, it may still be useful to have an easily computable lower bound on achievable estimation uncertainty within the Bayesian framework in the spirit of the Cramér-Rao bound derived within the frequentist approach.

Theorem 5.1 (Bayesian Cramér-Rao bound—van Trees inequality)**.** Given a Bayesian estimation problem with $p(\theta)$ as priori distribution and $p(x|\theta)$ as conditional distribution for observation of an event x, than assuming standard regularity conditions and vanishing of the prior on the ends of the interval over which parameter θ is considered ($\theta \in [\theta_-, \theta_+]$), the following bound on the average variance holds irrespectively of the estimator function chosen

$$\Delta^2\tilde{\theta} \geq \frac{1}{\bar{F} + I}, \tag{5.13}$$

where $\bar{F} = \int \mathrm{d}\theta\, F(\theta)$ is the FI averaged with the prior distribution, while $I =$

$\int \mathrm{d}\theta \, \frac{1}{p(\theta)} \left(\frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta}\right)^2$ represents the information contribution coming from the prior distribution.

Proof. Let us define two functions

$$f(\theta, x) = \sqrt{p(x|\theta)p(\theta)}(\tilde{\theta}(x) - \theta), \tag{5.14}$$

$$g(\theta, x) = \frac{1}{\sqrt{p(x|\theta)p(\theta)}} \frac{\mathrm{d}p(x|\theta)p(\theta)}{\mathrm{d}\theta}. \tag{5.15}$$

First observe that $\Delta^2\tilde{\theta} = \int \mathrm{d}\theta \mathrm{d}x \, f(\theta, x)^2$, so the average variance may be viewed as the squared norm of the the function f. Further note:

$$\int \mathrm{d}\theta \mathrm{d}x \, g(\theta, x)^2 = \int \mathrm{d}\theta \mathrm{d}x \frac{p(\theta)}{p(x|\theta)} \left(\frac{\mathrm{d}p(x|\theta)}{\mathrm{d}\theta}\right)^2 + \frac{p(x|\theta)}{p(\theta)} \left(\frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta}\right)^2 + 2\frac{\mathrm{d}p(x|\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta} = \int \mathrm{d}\theta p(\theta) F(\theta) + \int \mathrm{d}\theta \frac{1}{p(\theta)} \left(\frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta}\right)^2 = \bar{F} + I, \tag{5.16}$$

where we have used regularity assumptions thanks to which $\int \mathrm{d}\theta \mathrm{d}x \, \frac{\mathrm{d}p(x|\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta} = 0$. Moreover:

$$\begin{aligned} &\int \mathrm{d}\theta \mathrm{d}x \, f(\theta, x) g(\theta, x) = \\ &= \int \mathrm{d}\theta \mathrm{d}x \, (\tilde{\theta}(x) - \theta) \frac{\mathrm{d}p(x|\theta)p(\theta)}{\mathrm{d}\theta} = \\ &= \int \mathrm{d}x \;\, p(x|\theta)p(\theta)|_{\theta_-}^{\theta_+} - \int \mathrm{d}\theta \, \theta \frac{\mathrm{d}p(\theta)}{\mathrm{d}\theta} = \\ &= -\theta p(\theta)|_{\theta_-}^{\theta_+} + \int \mathrm{d}\theta \, p(\theta) = 1, \end{aligned} \tag{5.17}$$

where in the last step we performed integration by parts, and we have used the fact that $p(\theta_+) = p(\theta_-) = 0$. Applying now the Cauchy-Schwarz inequality

$$\int \mathrm{d}\theta \mathrm{d}x \, f(\theta, x)^2 \int \mathrm{d}\theta' \mathrm{d}x' \, g(\theta', x')^2 \geq \left(\int \mathrm{d}x \mathrm{d}\theta f(\theta, x) g(\theta, x)\right)^2, \tag{5.18}$$

we prove the theorem. □

The above inequality clearly illumianates the role of the data and the prior in Bayesian inference. The $\bar{F}$ quantity corresponds to the information coming from data while I represents the information due to prior. In the large number of experiment repetition limit, we expect $\bar{F}$ to grow linearly with number of repetitions, while I remains constant. Hence in this limit we will recover standard CR inequality as I will be negligible compared to $\bar{F}$.

Example 5.1 (continued) For the gaussian Bayesian model considered before, we may calculate the quantities appearing in the Bayesian Cramér-Rao inequality. FI does not depend on θ and hence $\bar{F} = N/\sigma^2$, while $I = 1/\sigma_\theta^2$. We see that the Bayesian Cramer-Rao inequality implies:

$$\Delta^2\tilde{\theta} \geq \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}, \tag{5.19}$$

which is exactly the precision achieved by the optimal Bayesian estimator. In this case the inequality is saturated.

Problems

Problem 5.1 Consider a Bayesian estimation problem, but with a different cost function than the mean squared error. In case when we want to estimate a phase (or some other angle-like parameter) $\theta \in [0, 2\pi]$, a more practical cost function is a function of the form $C(\theta, \tilde{\theta}) = 4\sin^2\left(\frac{\theta-\tilde{\theta}}{2}\right)$, which for small deviations between θ and $\tilde{\theta}$ is equivalent to the variance but respects that fact, that the 2π difference is not relevant. Average cost is then given by:

$$\bar{C} = \int \mathrm{d}\theta \mathrm{d}x \; 4\sin^2\left(\frac{\theta - \tilde{\theta}(x)}{2}\right) p(x|\theta)p(\theta). \tag{5.20}$$

Find the optimal Bayesian estimator for this cost function.

Problem 5.2 Analyze the conditions for saturation of the Bayesian Cramér-Rao inequality and check if the gaussian model consider during the lecture is the only one for which the inequality is actually saturated.