CS 8850: Advanced Machine Learning **Fall 2017**

Topic 6: Bayesian Inference

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN © COPYRIGHT 2017

6.1 Bayesian Hypothesis Testing

Recall from past lectures that in hypothesis testing we observe data x and want to decide between two hypotheses:

$$\begin{aligned} H_0 : \ x \ &\sim \ \mathsf{p}(\mathrm{x}|H_0), \\ H_1 : \ x \ &\sim \ \mathsf{p}(\mathrm{x}|H_1), \end{aligned}$$

where $\mathsf{p}(\mathrm{x}|H_\mathrm{h})$ is just an other way to write $\mathsf{p}_\mathrm{h}(\mathrm{x})$, $\mathrm{h} \in \{0,1\}$. We saw that one way to do this was using the *likelihood ratio test* (LRT):

$$\frac{\mathsf{p}(\mathrm{x}|H_1)}{\mathsf{p}(\mathrm{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

In bayesian hypothesis testing we also want to decide between two hypotheses. The only difference is that we have some *prior* knowledge of the probabilities that H_0 or H_1 are true. That is, we know the *priors* $\mathsf{p}(H_\mathrm{h})$. As we will see, this will allow us to use the *posterior* probabilities $\mathsf{p}(H_\mathrm{h}|\mathrm{x})$ rather than the *likelihoods* $\mathsf{p}(\mathrm{x}|H_\mathrm{h})$. This results in the *posterior ratio test*.

Definition 6.1 (Posterior ratio test (PRT))**.**

$$\frac{\mathsf{p}(H_1|\mathrm{x})}{\mathsf{p}(H_0|\mathrm{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

Luckily, a simple application of Bayes rule allows us to express the *posteriors* in the PRT in terms of the *likelihoods* and the *priors*, thus obtaining the following practical result. In words, it tells us that to obtain the PRT we simply need to scale the LRT by the priors (see Figure 6.1 to build some intuition).

Proposition 6.1. The PRT is equivalent to:

$$\frac{\mathsf{p}(\mathrm{x}|H_1)\mathsf{p}(H_1)}{\mathsf{p}(\mathrm{x}|H_0)\mathsf{p}(H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

Proof. By Bayes rule,

$$\frac{\mathsf{p}(H_1|\mathrm{x})}{\mathsf{p}(H_0|\mathrm{x})} = \frac{\frac{\mathsf{p}(\mathrm{x}|H_1)\mathsf{p}(H_1)}{\cancel{\mathsf{p}(\mathrm{x})}}}{\frac{\mathsf{p}(\mathrm{x}|H_0)\mathsf{p}(H_0)}{\cancel{\mathsf{p}(\mathrm{x})}}} = \frac{\mathsf{p}(\mathrm{x}|H_1)\mathsf{p}(H_1)}{\mathsf{p}(\mathrm{x}|H_0)\mathsf{p}(H_0)}.$$

□

Example 6.1. During the Cold War, the U.S. Navy developed *boomer* submarines that could launch nuclear missiles at other nations. Their main purpose was to help maintain the Cold War equilibrium by assuring mutual destruction in case somebody started a nuclear war. A secure bunker would constantly transmit a signal y to these submarines:

$$y = \begin{cases} 0 & \text{indicating to stand by,} \\ 1 & \text{indicating to launch attack.} \end{cases}$$

The submarine would receive a signal $x = y + z$ (where $z \sim \mathcal{N}(0, \sigma^2)$ represents noise), and would have to decide between:

$$\begin{aligned} H_0 :\ x &\sim \mathcal{N}(0, \sigma^2) & &\Rightarrow \text{stand by,} \\ H_1 :\ x &\sim \mathcal{N}(1, \sigma^2) & &\Rightarrow \text{launch attack.} \end{aligned}$$

Soldiers at the submarine know that a $(1|0)$ error would have catastrophic consequences, so they want to minimize p_{10}. Furthermore, they know that it is very unlikely that someone would start a nuclear war, so they decided to use this information in a bayesian test with a prior $\mathsf{p}(H_1) = {}^1\!/_5$:

$$\frac{\mathsf{p}(H_1|\mathrm{x})}{\mathsf{p}(H_0|\mathrm{x})} = \frac{\mathsf{p}(\mathrm{x}|H_1)\mathsf{p}(H_1)}{\mathsf{p}(\mathrm{x}|H_0)\mathsf{p}(H_0)} = \frac{\mathsf{p}(\mathrm{x}|H_1)1/\cancel{5}}{\mathsf{p}(\mathrm{x}|H_0)4/\cancel{5}} = \frac{\cancel{\frac{1}{\sqrt{2\pi}\sigma}}e^{-\frac{1}{2}\left(\frac{\mathrm{x}-1}{\sigma}\right)^2}}{4\cancel{\frac{1}{\sqrt{2\pi}\sigma}}e^{-\frac{1}{2}\left(\frac{\mathrm{x}-0}{\sigma}\right)^2}} = \frac{1}{4}e^{-\frac{1}{2\sigma^2}(\cancel{\mathrm{x}^2}-2\mathrm{x}+1-\cancel{\mathrm{x}^2})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

Equivalently,

$$\begin{aligned} \frac{1}{2\sigma^2}(2\mathrm{x}-1) &\underset{H_0}{\overset{H_1}{\gtrless}} \log(4\tau) \\ \mathrm{x} &\underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2\log(4\tau) + {}^1\!/_2. \end{aligned}$$

The prior knowledge that H_1 is *unlikely* is reflected in the *shrinkage* of the region of the PRT where we choose H_1. For instance, the LRT would have a larger region for H_1. See Figure 6.1 to build some intuition.

Notice that the PRT has the same form as the LRT, except that τ is *scaled* by the priors. So an other way to interpret Proposition 6.1 is as follows:

Proposition 6.2. The PRT $\frac{\mathsf{p}(H_1|\mathrm{x})}{\mathsf{p}(H_0|\mathrm{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$ is equivalent to the LRT $\frac{\mathsf{p}(\mathrm{x}|H_1)}{\mathsf{p}(\mathrm{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau'$, with $\tau' = \tau\frac{\mathsf{p}(H_0)}{\mathsf{p}(H_1)}$.

Proof. Follows directly from Proposition 6.1 □

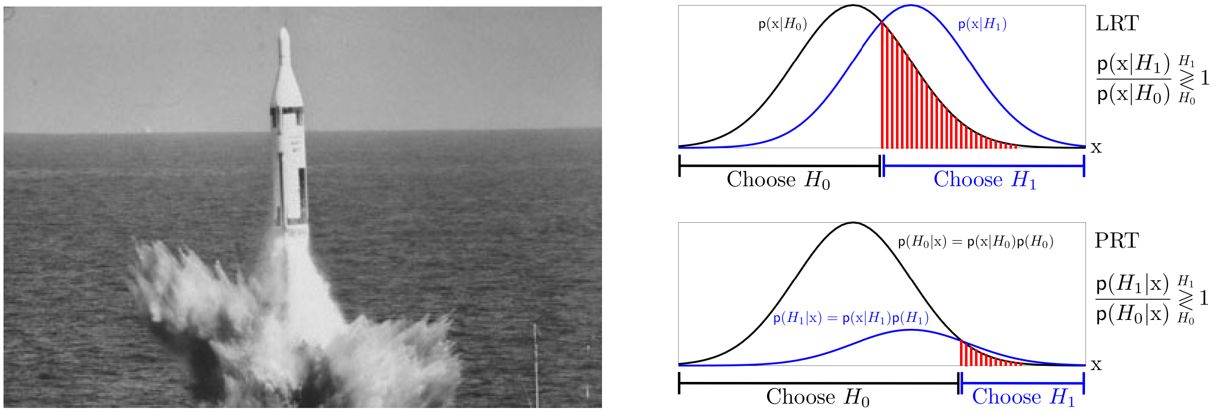

Figure 6.1: **Left**: A Polaris missile is launched from a submarine in a demonstration of an advanced weapons system. The submarine receives a signal x and must decide whether to H_0: stand by, or H_1: launch an attack. See Example 6.1 **Top**: LRT with $\tau = 1$, so that this test selects the hypothesis with the highest *likelihood* $p(x|H_h)$. **Bottom**: PRT with $\tau = 1$, so that this test selects the hypothesis with the highest *posterior* $p(H_h|x)$. To obtain the PRT we simply need to *scale* the likelihoods by their corresponding *priors* $p(H_h)$. In this case it was known that H_1 was *unlikely*, which resulted in a *shrinkage* of $p(x|H_1)$, as we can see in this figure.

6.2 Bayesian Parameter Estimation

In past lectures we studied the estimation problem where we observe data

$$x \sim p(x|\theta^\star), \ \theta^\star \in \Theta,$$

where $\theta^\star$ is an unknown parameter that we want to estimate. The only difference in bayesian estimation is that we treat $\theta^\star$ as a random variable, and we assume that we know the *prior* $p(\theta)$ that generated $\theta^\star$. This allows us (through Bayes rule) to use the *posterior* probability $p(\theta|x)$ rather than the *likelihood* $p(x|\theta)$. This results in the *maximum a posteriori estimator*, whose main idea is to find the θ that maximizes $p(\theta|x)$.

Definition 6.2 (Maximum a posteriori (MAP) estimator)**.**

$$\hat{\theta}_{MAP} \ := \ \arg\max_{\theta \in \Theta} \ p(\theta|x).$$

Using Bayes rule, we obtain the following practical result, which tells us that the posterior $p(\theta|x)$ is simply the likelihood $p(x|\theta)$ scaled by $p(\theta)$.

Proposition 6.3. The MAP estimator is given by

$$\hat{\theta}_{MAP} \ = \ \arg\max_{\theta \in \Theta} \ p(x|\theta)p(\theta).$$

Proof. By Bayes rule,

$$\hat{\theta}_{MAP} \ := \ \arg\max_{\theta\in\Theta} \ \mathsf{p}(\theta|\mathrm{x}) \ = \ \arg\max_{\theta\in\Theta} \ \frac{\mathsf{p}(\mathrm{x}|\theta)\mathsf{p}(\theta)}{\underbrace{\mathsf{p}(\mathrm{x})}_{\text{constant}}} \ = \ \arg\max_{\theta\in\Theta} \ \mathsf{p}(\mathrm{x}|\theta)\mathsf{p}(\theta).$$

□

Example 6.2 (Pharmaceuticals loooove bayesian estimation)**.** Scientists at a big pharmaceutical company have designed a new cancer treatment, and want to estimate its probability of success $\rho^\star$. To this end they will conduct a clinical trial where they will test their treatment on N individuals, and record whether they react favorably. This can be modeled as

$$x_1, \dots, x_{\mathrm{N}} \overset{iid}{\sim} \text{Bernoulli}(\rho^\star),$$

where they want to estimate $\rho^\star$.

Pharmaceuticals design many treatments. Testing them on humans is difficult and expensive. Hence they first experiment in-vitro or with animals to find the most effective ones. The particular treatment that we are studying has already been tested in vitro, mice, rabbits and chimpanzees, and has proven to be very effective. Hence we expect *a priori* that $\rho^\star$ will be closer to 1 than to 0. Thus, a good model for the prior $\mathsf{p}(\rho)$ would be the Beta density

$$\mathsf{p}(\rho) \ = \ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \, \rho^{\alpha-1}(1-\rho)^{\beta-1}$$

with parameters $\alpha > \beta > 1$, so that the density is skewed towards 1 (see Figure 6.2 for some intuition). Using this prior information, the pharmaceutical will use a bayesian approach to estimate $\rho^\star$. The likelihood function is

$$\mathsf{p}(\mathbf{x}|\rho) \ = \ \rho^{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}} \, (1-\rho)^{\mathrm{N}-\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}} \ = \ \rho^{\mathbf{1}^\mathsf{T}\mathbf{x}} \, (1-\rho)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}}.$$

Hence

$$\begin{aligned} \mathsf{p}(\rho|\mathbf{x}) \ &= \ \frac{\mathsf{p}(\mathbf{x}|\rho)\mathsf{p}(\rho)}{\mathsf{p}(\mathbf{x})} \propto \mathsf{p}(\mathbf{x}|\rho)\mathsf{p}(\rho) \ = \ \left(\rho^{\mathbf{1}^\mathsf{T}\mathbf{x}} \, (1-\rho)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}}\right)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \, \rho^{\alpha-1}(1-\rho)^{\beta-1}\right) \\ &\propto \ \rho^{\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha-1} \, (1-\rho)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}+\beta-1}, \end{aligned}$$

and so we recognize $\mathsf{p}(\rho|\mathbf{x})$ to be the Beta(α', β') density with parameters $\alpha' = \mathbf{1}^\mathsf{T}\mathbf{x} + \alpha$ and $\beta' = \mathrm{N} - \mathbf{1}^\mathsf{T}\mathbf{x} + \beta$ (here we are omitting the normalization factor $\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}$, which we know $\mathsf{p}(\rho|\mathbf{x})$ must have, because it is a density and must integrate to 1). It follows that $\hat{\rho}_{MAP} = \arg\max_\rho \mathsf{p}(\rho|\mathbf{x})$ is the point that maximizes the Beta(α', β') density, i.e., its mode: if $\alpha', \beta' > 1$, this is given by $\frac{\alpha'-1}{\alpha'+\beta'-2}$; if α' or $\beta' < 1$, it is one of the extreme points $\{0, 1\}$; if $\alpha' = \beta' = 1$, then Beta(α', β') = Uniform$[0, 1]$.

Definition 6.3 (Conjugate prior)**.** Whenever $\mathsf{p}(\theta)$ has the same form (but possibly different parameters) as $\mathsf{p}(\theta|\mathrm{x})$, we say that $\mathsf{p}(\theta)$ is a *conjugate prior* of $\mathsf{p}(\mathrm{x}|\theta)$.

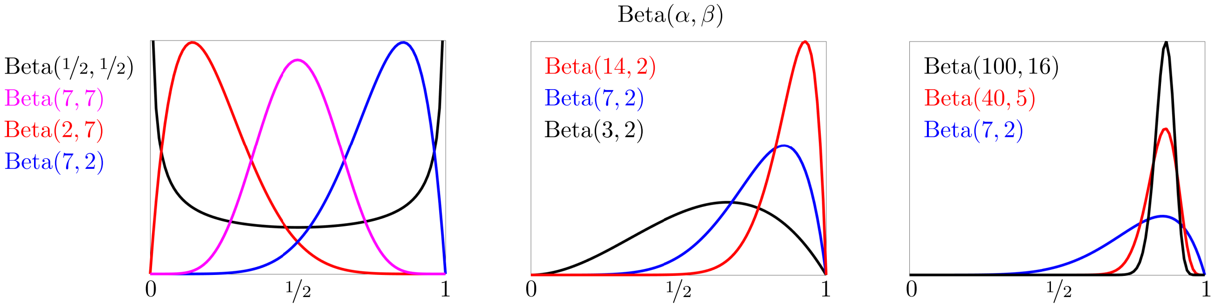

Figure 6.2: Beta(α,β) densities are good models for prior distributions of proportions (like the probability of success $\rho^\star$). Loosely speaking, the gap between α and β determines where we a priori *think* $\rho^\star$ is; the magnitudes of α and β determine how *confident* we are. This way the parameters α and β determine our a priori *bias* and *certainty*. For instance, in Example 6.2, if we believe the probability of success ρ to be closer to 1, we can model it as Beta(α,β) with $\alpha > \beta > 1$, so that $\mathsf{p}(\rho)$ is *biased* towards 1. If we are *somewhat* certain, we can take $\alpha = 7$ and $\beta = 2$. If we are *extremely* certain, we can choose α and/or β to be much larger.

Example 6.3. In Example 6.2, $\mathsf{p}(\rho)$ =Beta(α,β) is a conjugate prior of $\mathsf{p}(\mathbf{x}|\rho)$ =Bernoulli(ρ) because $\mathsf{p}(\rho|\mathbf{x})$ =Beta(α',β')

6.3 MAP vs MLE: Which is Better?

We now discuss some relations between the MAP estimator and the MLE.

Proposition 6.4. If there is no prior information, the MAP is equal to the MLE.

Proof. Having *no* prior information is equivalent to having θ ~Uniform(Θ). In this case $\mathsf{p}(\theta) = \frac{1}{|\Theta|}$ for every θ, i.e., $\mathsf{p}(\theta)$ is a constant. Hence

$$\hat{\theta}_{MAP} \;:=\; \arg\max_{\theta\in\Theta}\; \mathsf{p}(\theta|\mathrm{x}) \;=\; \arg\max_{\theta\in\Theta}\; \mathsf{p}(\mathrm{x}|\theta)\; \underbrace{\mathsf{p}(\theta)}_{\text{constant}} \;=\; \arg\max_{\theta\in\Theta}\; \mathsf{p}(\mathrm{x}|\theta) \;=:\; \hat{\theta}_{ML},$$

where the second equality follows by Proposition 6.3. □

The next theorem states that the MAP converges to the MLE as the number of samples N grows.

Theorem 6.1. Let $x_1, \ldots, x_{\mathrm{N}} \overset{iid}{\sim} \mathsf{p}(\mathrm{x}|\theta)$. Then $\hat{\theta}_{MAP} \to \hat{\theta}_{ML}$ as $\mathrm{N} \to \infty$.

Proof. Write:

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta\in\Theta}\ \mathsf{p}(\theta|\mathbf{x})\mathsf{p}(\theta) \;=\; \arg\max_{\theta\in\Theta}\ \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{p}(\theta|\mathrm{x_i})\mathsf{p}(\theta) \;=\; \arg\max_{\theta\in\Theta}\ \log\left(\prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{p}(\theta|\mathrm{x_i})\mathsf{p}(\theta)\right) \\
&= \arg\max_{\theta\in\Theta}\ \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\Big(\mathsf{p}(\theta|\mathrm{x_i})\mathsf{p}(\theta)\Big) \;=\; \arg\max_{\theta\in\Theta}\ \underbrace{\frac{1}{\mathrm{N}}}_{\text{constant}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\Big(\mathsf{p}(\theta|\mathrm{x_i})\mathsf{p}(\theta)\Big) \\
&= \arg\max_{\theta\in\Theta}\ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\theta|\mathrm{x_i}) \;+\; \frac{1}{\mathrm{N}}\log\mathsf{p}(\theta) \;=\; \arg\max_{\theta\in\Theta}\ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\theta|\mathrm{x_i}) \;+\; \underbrace{\frac{1}{\mathrm{N}}\log\mathsf{p}(\theta)}_{\xrightarrow{\mathrm{N}\to\infty} 0} \\
&\longrightarrow \arg\max_{\theta\in\Theta}\ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\theta|\mathrm{x_i}) \;=\; \arg\max_{\theta\in\Theta}\ \frac{1}{\mathrm{N}}\log\prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{p}(\theta|\mathrm{x_i}) \;=\; \arg\max_{\theta\in\Theta}\ \underbrace{\frac{1}{\mathrm{N}}}_{\text{constant}}\ \mathsf{p}(\theta|\mathbf{x}) \\
&= \arg\max_{\theta\in\Theta}\ \mathsf{p}(\theta|\mathbf{x}) \;=\; \hat{\theta}_{ML}.
\end{aligned}
$$

□

Example 6.4. In Example 6.2, first observe that

$$
\hat{\rho}_{ML} \;=\; \arg\max_{\rho\in[0,1]} \mathsf{p}(\mathbf{x}|\rho) \;=\; \arg\max_{\rho\in[0,1]}\ \rho^{\mathbf{1}^\mathsf{T}\mathbf{x}}\,(1-\rho)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}} \;=\; \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i},
$$

where I'll let you fill in the details of the last equality. On the other hand, since $\mathsf{p}(\rho|\mathbf{x}) = \mathrm{Beta}(\alpha',\beta')$,

$$
\begin{aligned}
\mathsf{var}(\rho|\boldsymbol{x}) &= \frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)} \;=\; \frac{(\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha)(\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}+\beta)}{(\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\alpha+\mathrm{N}-\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\beta)^2(\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\alpha+\mathrm{N}-\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\beta+1)} \\
&= \frac{(\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha)(\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}+\beta)}{(\mathrm{N}+\alpha+\beta)^2(\mathrm{N}+\alpha+\beta+1)} \;\longrightarrow\; \frac{\overbrace{(\mathbf{1}^\mathsf{T}\mathbf{x})}^{\le\mathrm{N}}\,\overbrace{(\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x})}^{\le\mathrm{N}}}{\mathrm{N}^3} \;\le\; \frac{1}{\mathrm{N}} \;\xrightarrow{\mathrm{N}\to\infty}\; 0.
\end{aligned}
$$

It follows that

$$
\hat{\rho}_{MAP} \;=\; \mathsf{E}[p|\mathbf{x}] \;=\; \frac{\alpha'}{\alpha'+\beta'} \;=\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha}{\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\alpha+\mathrm{N}-\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\beta} \;=\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha}{\mathrm{N}+\alpha+\beta} \;\longrightarrow\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}}{\mathrm{N}} \;=\; \hat{\rho}_{ML}.
$$

Here is an alternative derivation: observe that since $\alpha' = \mathbf{1}^\mathsf{T}\mathbf{x}+\alpha$ and $\beta' = \mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{x}+\beta$, if $\rho^\star \neq 0, 1$, for large N we will have $\alpha',\beta' > 1$, whence $\hat{\rho}_{MAP}$ is the mode of the $\mathrm{Beta}(\alpha',\beta')$ distribution, i.e.,

$$
\hat{\rho}_{MAP} \;=\; \frac{\alpha'-1}{\alpha'+\beta'-2} \;=\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha-1}{\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\alpha+\mathrm{N}-\cancel{\mathbf{1}^\mathsf{T}}\mathbf{x}+\beta-2} \;=\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}+\alpha-1}{\mathrm{N}+\alpha+\beta-2} \;\longrightarrow\; \frac{\mathbf{1}^\mathsf{T}\mathbf{x}}{\mathrm{N}} \;=\; \hat{\rho}_{ML}.
$$

On the other hand, if $\rho^\star = 0$ or 1, then so will $\hat{\rho}_{MAP}$ and $\hat{\rho}_{ML}$.

Experiments

Experiments are not only good to verify our results. They are also good to build intuition, test theories and draw conclusions. In this section we will further study Example 6.2 to compare the MAP and the MLE.

In short, our conclusion will be that if our prior is accurate, the MAP will be better, but if our prior is inaccurate, the MAP will be worse.

Recall that the treatment in Example 6.2 has already shown great results on other organisms, so we believe its probability of success on humans $\rho^\star$ to be closer to 1 than to 0. With this in mind we will use a $\mathrm{Beta}(7, 2)$ as prior, so that the density is skewed towards 1. (see Figure 6.2 to build some intuition).

Next we will generate a random vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{N}}$ with i.i.d. $\mathrm{Bernoulli}(\rho^\star)$ entries. The i^{th} entry in $\boldsymbol{x}$ simulates whether the i^{th} patient reacted favorably to the treatment. We showed in Example 6.2 that the posterior $\mathsf{p}(\rho|\mathbf{x}) = \mathrm{Beta}(\alpha', \beta')$ with $\alpha' = \mathbf{1}^{\mathsf{T}}\mathbf{x} + \alpha$ and $\beta' = \mathrm{N} - \mathbf{1}^{\mathsf{T}}\mathbf{x} + \beta$.

Let us now consider two scenarios:

(i) $\rho^\star = 0.7$. This would be a case when our prior is correct (Figure 6.3—left). We can see that even with a few samples ($\mathrm{N} = 3$), our estimate $\hat{\rho}_{MAP} = \arg\max_{\rho \in [0,1]} \mathsf{p}(\rho|\mathbf{x})$ would be very close to $\rho^\star$. The code for this simulation is in Appendix A.

(ii) $\rho^\star = 0.3$. This would be a case when our prior is incorrect (Figure 6.3—right). We can see that unless we have a lot of samples (N large), our estimate $\hat{\rho}_{MAP}$ could be very far from $\rho^\star$! In words, the prior is *pulling* the posterior towards it. This is one of the dangers of bayesian estimation: the bias induced by the prior might make it harder to see the truth. What do you think would happen if we use a *stronger* prior, like $\mathsf{p}(\rho) = \mathrm{Beta}(100, 16)$? How would this affect the posterior $\mathsf{p}(\rho|\mathbf{x})$? Try it out and see; you only need to change a few lines of code. Do the results match your intuition?

Case (*ii*) shows one of the risks of bayesian estimation. Now the question is: is it worth it? In other words, *if* our prior is correct, do we really have that much to gain? Let us find out by comparing the MAP with the MLE. Since $x_i \overset{iid}{\sim} \mathrm{Bernoulli}(\rho^\star)$, it follows that $\hat{\rho}_{ML} = \sum_{\mathrm{i}=1}^{\mathrm{N}} x_i$ has a $\mathrm{Binomial}(\mathrm{N}, \rho^\star)$ distribution (scaled by $^{1}/_{\mathrm{N}}$). Figure 6.4 shows a comparison of the distributions of $\hat{\theta}_{MAP}$ and $\hat{\theta}_{ML}$ for different scenarios.

These experiments show that the prior is essentially *biasing* us towards our *beliefs*. If our beliefs are correct, the MAP will be more accurate than the MLE (given the same number of samples N). In contrast, if our beliefs are incorrect, it will take more samples to *correct* the posterior, and so the MAP will be more inaccurate.

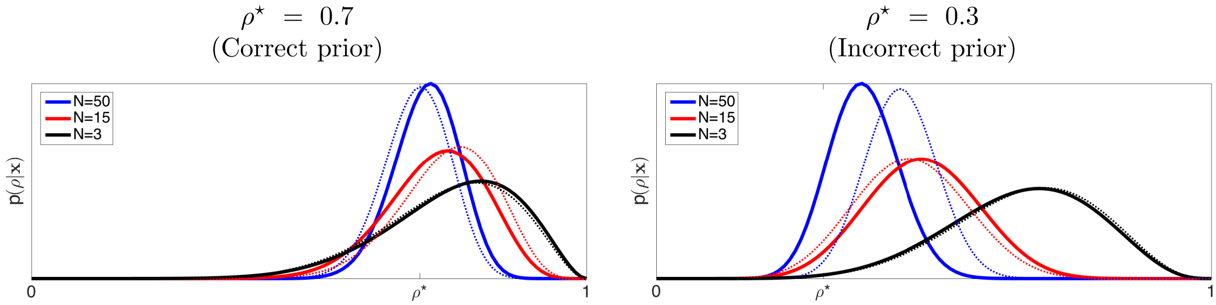

Figure 6.3: Posterior distribution $\mathsf{p}(\rho|\mathbf{x})$ in Example 6.2. In expectation, $\mathbf{1}^{\mathsf{T}}\mathbf{x} = \mathrm{N}\rho^\star$, so the expected posterior is distributed $\mathrm{Beta}(\mathrm{N}\rho^\star + \alpha, \mathrm{N}(1 - \rho^\star) + \beta)$, plotted in solid colors. Dotted lines are the posterior distributions given a particular sample. The code for this simulation is in Appendix A.

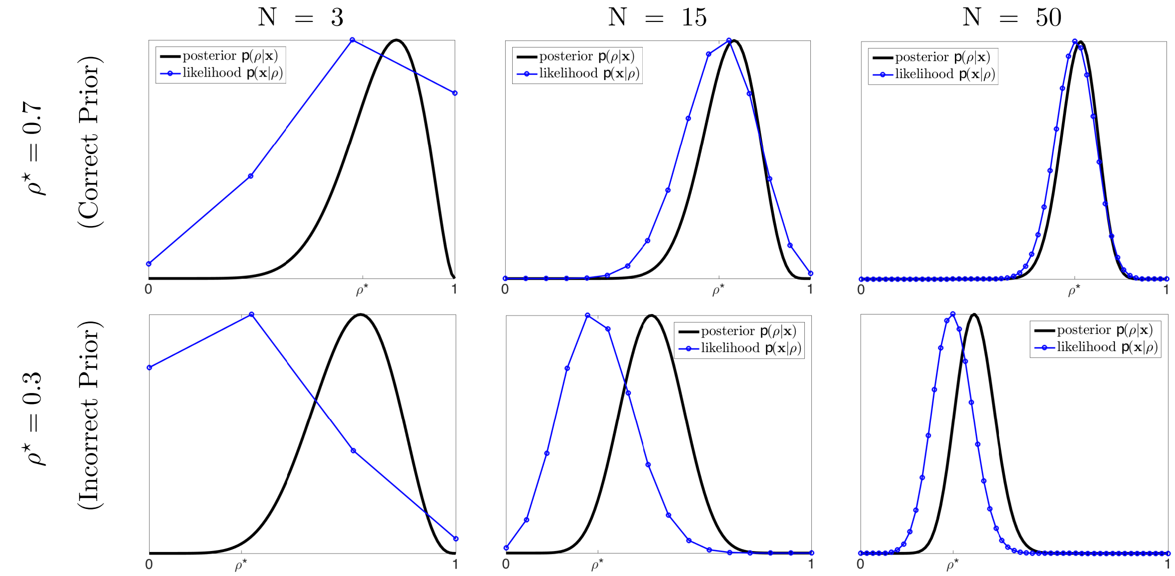

Figure 6.4: Distribution of the MAP and the MLE for different values of $\rho^\star$ and N. $\hat{\theta}_{MAP} \sim \text{Beta}(\text{N}\rho^\star + \alpha, \text{N}(1-\rho^\star)+\beta)$ and $\hat{\theta}_{ML} \sim \text{Binomial}(\text{N}, \rho^\star)$. The code for this is in Appendix B.

A Code for Simulation of Example 6.2

```
clear all; close all; clc; warning('off','all');

% === Code to simulate the posterior distribution of rho ===
% ===== See Example 6.2.

rho_star = 0.7;         % True probability of success.
alpha = 7;              % Parameter of prior distribution.
beta = 2;               % Parameter of prior distribution.
NN = [50,15,3];         % Sample sizes we will try.

% Create figure.
figure(1);
axes('Box','on');
hold on;

% Plot expected posterior distribution.
rho = 0:0.01:1;           % all possible values of rho.
color = ['b','r','k'];    % For plotting.
for n=1:length(NN)

    N = NN(n);                              % Number of samples.
    alpha_prime = N*rho_star + alpha;   % Parameter of expected posterior distribution.
    beta_prime = N*(1-rho_star) + beta; % Parameter of expected posterior distribution.
    posterior = betapdf(rho,alpha_prime,beta_prime);   % Expected posterior distribution.
    plot(rho,posterior,color(n),'LineWidth',4);

end

% Legends.
legend(['N=',num2str(NN(1))],['N=',num2str(NN(2))],['N=',num2str(NN(3))],...
    'Interpreter','latex','fontsize',20,'Location','Northwest');

% Plot posterior distributions for a particular sample [X_1,...X_N].
for n=1:length(NN)

    N = NN(n);                            % Number of samples.
    X = rand([N,1]) < rho_star;       % Sample.
    alpha_prime = sum(X) + alpha;     % Parameter of posterior distribution based on sample.
    beta_prime = N - sum(X) + beta;   % Parameter of posterior distribution based on sample.
    posterior = betapdf(rho,alpha_prime,beta_prime);   % Posterior distribution based on ...
        sample.
    plot(rho,posterior,[color(n),':'],'LineWidth',2);

end

% Make figure look sexy.
axis tight;
ylabel('$\mathsf{p}(\rho|\textbf{x})$','Interpreter','latex','fontsize',20);
xlabel('','Interpreter','latex','fontsize',20);
set(gca,'XTick',[0,rho_star,1],'xticklabel',{'0','\rho*','1'},'fontsize',20);
set(gca,'YTick',[],'yticklabel',[]);
set(gcf,'PaperUnits','centimeters','PaperSize',[30,10],'PaperPosition',[0,0,30,10]);

% Save figure.
set(gcf, 'renderer','default');
figurename = 'MAP.pdf';
saveas(gcf,figurename);
```

B Code for Comparison of MAP and MLE in Figure 6.4

```
clear all; close all; clc; warning('off','all');

% === Code to simulate the posterior distribution of rho ===
% ===== See Example 6.2 and Figure 6.4.

rho_star = 0.7;         % True probability of success.
alpha = 7;              % Parameter of prior distribution.
beta = 2;               % Parameter of prior distribution.
NN = [50,15,3];         % Sample sizes we will try.

% Plot distributions of the MAP and the MLE.
for n=1:length(NN)

    N = NN(n);  % Number of samples.

    % Create figure
    figure(n);
    axes('Box','on');
    hold on;

    % MAP distribution.
    rho = 0:0.01:1;                     % all possible values of rho (continuous).
    alpha_prime = N*rho_star + alpha;   % Parameter of expected posterior distribution.
    beta_prime = N*(1-rho_star) + beta; % Parameter of expected posterior distribution.
    posterior = betapdf(rho,alpha_prime,beta_prime);   % Expected posterior distribution.
    h1 = plot(rho,posterior,'k','LineWidth',4);

    % MLE distribution.
    rho = 0:N;                                % all possible values of rho (discrete).
    likelihood = binopdf(rho,N,rho_star);     % Distribution of the MLE.
    likelihood = likelihood/max(likelihood)*max(posterior);
    h2 = plot(rho/N,likelihood,'b-o','LineWidth',2);

    % Make figure look sexy.
    axis tight;
    ylabel('','Interpreter','latex','fontsize',20);
    xlabel('','Interpreter','latex','fontsize',20);
    set(gca,'XTick',[0,rho_star,1],'xticklabel',{'0','\rho*','1'},'fontsize',20);
    set(gca,'YTick',[],'yticklabel',[]);
    title(['N$ = ',num2str(N),'$'],'Interpreter','latex','fontsize',25);
    legend({'posterior $\mathsf{p}(\rho|\textbf{x})$',...
        'likelihood $\mathsf{p}(\textbf{x}|\rho)$'},...
        'Interpreter','latex','fontsize',20,'Location','Northwest');
    set(gcf,'PaperUnits','centimeters','PaperSize',[20,15],'PaperPosition',[0,0,20,15]);

    % Save figure.
    set(gcf, 'renderer','default');
    figurename = ['MAPvsMLE_',num2str(N),'.pdf'];
    saveas(gcf,figurename);

end
```