Chapter 10

Bayesian Statistics

In the frequentist paradigm we model the data as realizations from a distribution that is fixed. In particular, if the model is parametric, the parameters are *deterministic* quantities. In contrast, in Bayesian parametric modeling the parameters are modeled as **random variables**. The goal is to have the flexibility to quantify our uncertainty about the underlying distribution beforehand, for example in order to integrate available prior information about the data.

10.1 Bayesian parametric models

In this section we describe how to fit a parametric model to a data set within a Bayesian framework. As in Section 9.6, we assume that the data are generated by sampling from known distributions with unknown parameters. The crucial difference is that we model the parameters as being random instead of deterministic. This requires selecting their prior distribution before fitting the data, which allows to quantify our uncertainty about the value of the parameters beforehand. A Bayesian parametric model is specified by:

1. The **prior** distribution is the distribution of $\vec{\Theta}$, which encodes our uncertainty about the model before seeing the data.
2. The **likelihood** is the conditional distribution of $\vec{X}$ given $\vec{\Theta}$, which specifies how the data depend on the parameters. In contrast to the frequentist framework, the likelihood is *not* interpreted as a deterministic function of the parameters.

Our goal when learning a Bayesian model is to compute the **posterior distribution** of the parameters Θ given $\vec{X}$. Evaluating this posterior distribution at the realization $\vec{x}$ allows to update our uncertainty about Θ using the data.

The following example fits a Bayesian model to iid samples from a Bernoulli random variable.

Example 10.1.1 (Bernoulli distribution)**.** Let $\vec{x}$ be a vector of data that we wish to model as iid samples from a Bernoulli distribution. Since we are taking a Bayesian approach we choose a prior distribution for the parameter of the Bernoulli. We will consider two different Bayesian estimators Θ_1 and Θ_2:

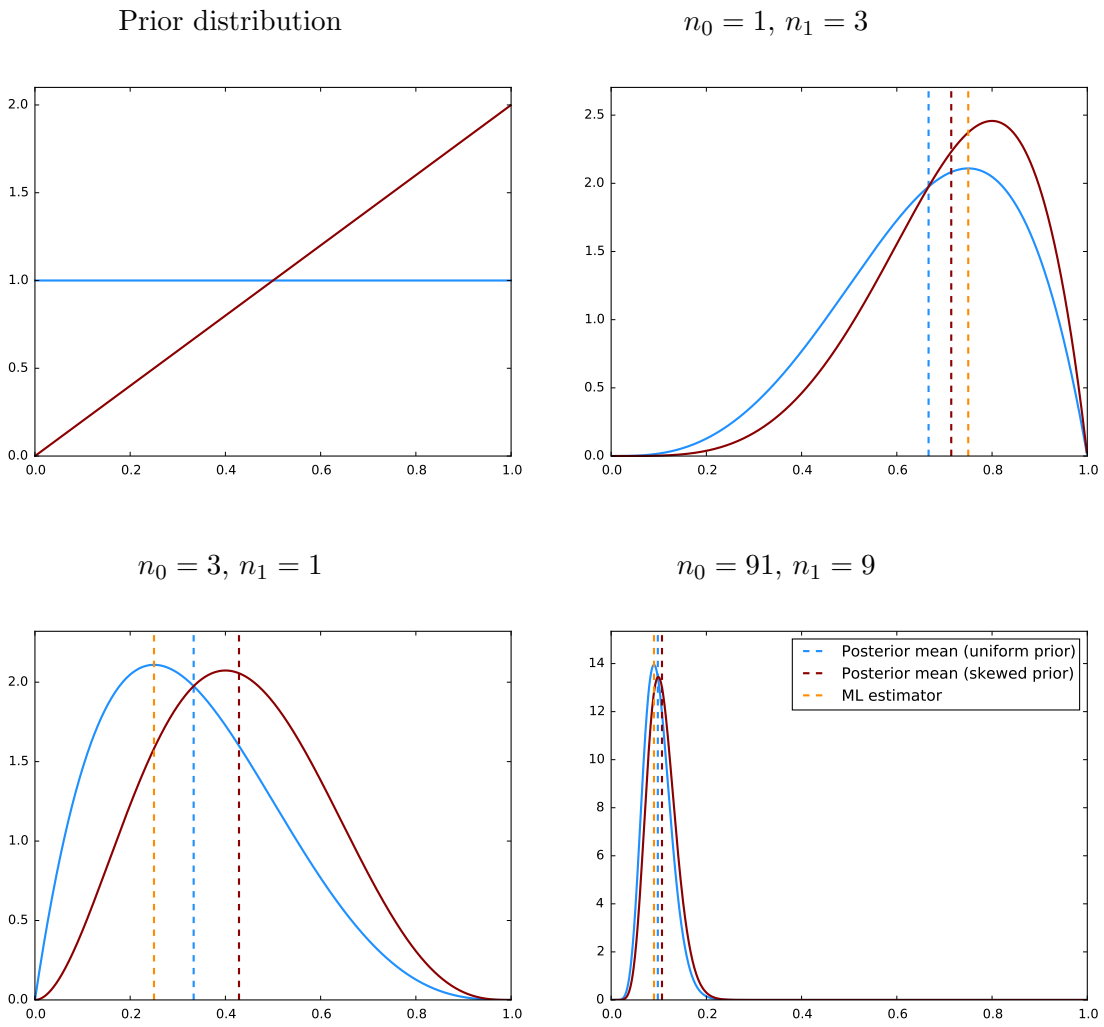

Figure 10.1: The prior distribution of Θ_1 (blue) and Θ_2 (dark red) in Example 10.1.1 are shown in the top-left graph. The rest of the graphs show the corresponding posterior distributions for different data sets.

1. Θ_1 represents a conservative estimator in terms of prior information. We assign a uniform pdf to the parameter. Any value in the unit interval has the same probability density:

$$f_{\Theta_1}(\theta) = \begin{cases} 1 & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.1}$$

2. Θ_2 is an estimator that assumes that the parameter is closer to 1 than to 0. We could use it for instance to capture the suspicion that a coin is biased towards heads. We choose a skewed pdf that increases linearly from zero to one,

$$f_{\Theta_2}(\theta) = \begin{cases} 2\,\theta & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.2}$$

By the iid assumption, the likelihood, which is just the conditional pmf of the data given the parameter of the Bernoulli, equals

$$p_{\vec{X}|\Theta}\left(\vec{x}|\theta\right) = \theta^{n_1}\left(1-\theta\right)^{n_0}, \tag{10.3}$$

where n_1 is the number of ones in the data and n_0 the number of zeros (see Example 9.6.3). The posterior pdfs of the two estimators are consequently equal to

$$\begin{aligned}
f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_1}\left(\theta\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.4)\\
&= \frac{f_{\Theta_1}\left(\theta\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{\int_u f_{\Theta_1}\left(u\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|u\right)\,\mathrm{d}u} && (10.5)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\int_u u^{n_1}\left(1-u\right)^{n_0}\,\mathrm{d}u} && (10.6)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+1, n_0+1\right)}, && (10.7)\\
f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_2}\left(\theta\right) p_{\vec{X}|\Theta_2}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.8)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\int_u u^{n_1+1}\left(1-u\right)^{n_0}\,\mathrm{d}u} && (10.9)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+2, n_0+1\right)}, && (10.10)\\
& && (10.11)
\end{aligned}$$

where

$$\beta\left(a,b\right) := \int_u u^{a-1}\left(1-u\right)^{b-1}\,\mathrm{d}u \tag{10.12}$$

is a special function called the beta function or Euler integral of the first kind, which is tabulated.

Figure 10.1 shows the plot of the posterior distribution for different values of n_1 and n_0. It also shows the maximum-likelihood estimator of the parameter, which is just $n_1/\left(n_0+n_1\right)$ (see Example 9.6.3). For a small number of flips, the posterior pdf of Θ_2 is skewed to the right with respect to that of Θ_1, reflecting the prior belief that the parameter is closer to 1. However for a large number of flips both posterior densities are very close.

△

10.2 Conjugate prior

Both posterior distributions in Example 10.1.1 are beta distributions (see Definition 2.3.12), and so are the priors. The uniform prior of Θ_1 is beta with parameters $a=1$ and $b=1$, whereas the skewed prior of Θ_2 is beta distribution with parameters $a=2$ and $b=1$. Since the prior and the posterior belong to the same family, computing the posterior is equivalent to just updating the parameters. When the prior and posterior are guaranteed to belong to the same family of distributions for a particular likelihood, the distributions are called conjugate priors.

Definition 10.2.1 (Conjugate priors)**.** *A conjugate family of distributions for a certain likelihood satisfies the following property: if the prior belongs to the family, then the posterior also belongs to the family.*

Beta distributions are conjugate priors when the likelihood is binomial.

Theorem 10.2.2 (The beta distribution is conjugate to the binomial likelihood)**.** *If the prior distribution of Θ is a beta distributions with parameters a and b and the likelihood of the data X given Θ is binomial with parameters n and x, then the posterior distribution of Θ given X is a beta distribution with parameters $x+a$ and $n-x+b$.*

Proof.

$$f_{\Theta \,|\, X}\left(\theta \,|\, x\right) = \frac{f_\Theta\left(\theta\right) p_{X \,|\, \Theta}\left(x \,|\, \theta\right)}{p_X\left(x\right)} \tag{10.13}$$

$$= \frac{f_\Theta\left(\theta\right) p_{X \,|\, \Theta}\left(x \,|\, \theta\right)}{\int_u f_\Theta\left(u\right) p_{X \,|\, \Theta}\left(x \,|\, u\right) \,\mathrm{d}u} \tag{10.14}$$

$$= \frac{\theta^{a-1}\left(1-\theta\right)^{b-1}\binom{n}{x}\theta^x\left(1-\theta\right)^{n-x}}{\int_u u^{a-1}\left(1-u\right)^{b-1}\binom{n}{x}u^x\left(1-u\right)^{n-x}\,\mathrm{d}u} \tag{10.15}$$

$$= \frac{\theta^{x+a-1}\left(1-\theta\right)^{n-x+b-1}}{\int_u u^{x+a-1}\left(1-u\right)^{n-x+b-1}\,\mathrm{d}u} \tag{10.16}$$

$$= f_\beta\left(\theta; x+a, n-x+b\right). \tag{10.17}$$

□

Note that the posteriors obtained in Example 10.1.1 follow immediately from the theorem.

Example 10.2.3 (Poll in New Mexico)**.** In a poll in New Mexico for the 2016 US election, 429 participants, 227 people intend to vote for Clinton and 202 for Trump (the data are from a real poll[1], but for simplicity we are ignoring the other candidates and people that were undecided). Our aim is to use a Bayesian framework to predict the outcome of the election in New Mexico using these data.

We model the fraction of people that vote for Trump as a random variable Θ. We assume that the n people in the poll are chosen uniformly at random with replacement from the population, so given $\Theta = \theta$ the number of Trump voters is a binomial with parameters n and θ. We don't have any additional information about the possible value of Θ, so we assume it is uniform or equivalently a beta distribution with parameters $a := 1$ and $b := 1$.

By Theorem 10.2.2 the posterior distribution of Θ given the data that we observe is a beta distribution with parameters $a := 203$ and $b := 228$, depicted in Figure 10.2. The corresponding probability that $\Theta \geq 0.5$ is 11.4%, which is our estimate for the probability that Trump wins in New Mexico.

△

[1]The poll results are taken from
https://www.abqjournal.com/883092/clinton-still-ahead-in-new-mexico.html

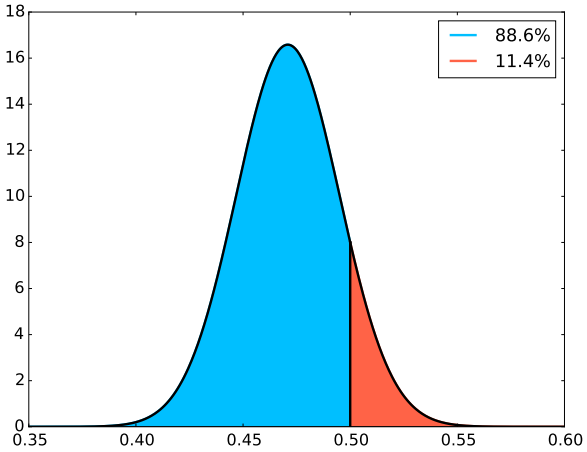

Figure 10.2: Posterior distribution of the fraction of Trump voters in New Mexico conditioned on the poll data in Example 10.2.3.

10.3 Bayesian estimators

The Bayesian approach to learning probabilistic models yields the whole posterior distribution of the parameters of interest. In this section we describe two alternatives for deriving a single estimate of the parameters from the posterior distribution.

10.3.1 Minimum mean-square-error estimation

The mean of the posterior distribution is the conditional expectation of the parameters given the data. Choosing the posterior mean as an estimator for the parameters $\vec{\Theta}$ has a strong theoretical justification: it is guaranteed to achieve the minimum mean square error (MSE) among *all possible estimators*. Of course, this only holds if all of the assumptions hold, i.e. the parameters are generated according to the prior and the data are then generated according to the likelihood, which may not be the case for real data.

Theorem 10.3.1 (The posterior mean minimizes the MSE)**.** *The posterior mean is the minimum mean-square-error (MMSE) estimate of the parameter $\vec{\Theta}$ given the data $\vec{X}$. To be more precise, let us define*

$$\theta_{\text{MMSE}}\left(\vec{x}\right) := \text{E}\big(\vec{\Theta}|\vec{X} = \vec{x}\big). \tag{10.18}$$

For any arbitrary estimator $\theta_{\text{other}}\left(\vec{x}\right)$,

$$\text{E}\left(\left(\theta_{\text{other}}(\vec{X}) - \vec{\Theta}\right)^2\right) \geq \text{E}\left(\left(\theta_{\text{MMSE}}(\vec{X}) - \vec{\Theta}\right)^2\right). \tag{10.19}$$

Proof. We begin by computing the MSE of the arbitrary estimator conditioned on $\vec{X} = \vec{x}$ in

terms of the conditional expectation of Θ given $\vec{X}$,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\right) \tag{10.20}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})+\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\right) \tag{10.21}$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\Big) \tag{10.22}$$

$$+2\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)\mathrm{E}\left(\theta_{\text{MMSE}}(\vec{x})-\mathrm{E}\left(\vec{\Theta}\middle|\vec{X}=\vec{x}\right)\right)$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\Big). \tag{10.23}$$

By iterated expectation,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\Theta\right)^2\right)=\mathrm{E}\left(\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}\right.\right)\right) \tag{10.24}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}\right.\Big)\right)$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right) \tag{10.25}$$

$$\geq\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right), \tag{10.26}$$

since the expectation of a nonnegative quantity is nonnegative. □

Example 10.3.2 (Bernoulli distribution (continued))**.** In order to obtain point estimates for the parameter in Example 10.1.1 we compute the posterior means:

$$\mathrm{E}\left(\Theta_1|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.27}$$

$$=\frac{\int_0^1\theta^{n_1+1}\left(1-\theta\right)^{n_0}\,\mathrm{d}\theta}{\beta\left(n_1+1,n_0+1\right)} \tag{10.28}$$

$$=\frac{\beta\left(n_1+2,n_0+1\right)}{\beta\left(n_1+1,n_0+1\right)}, \tag{10.29}$$

$$\mathrm{E}\left(\Theta_2|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.30}$$

$$=\frac{\beta\left(n_1+3,n_0+1\right)}{\beta\left(n_1+2,n_0+1\right)}. \tag{10.31}$$

Figure 10.1 shows the posterior means for different values of n_0 and n_1.

△

10.3.2 Maximum-a-posteriori estimation

An alternative to the posterior mean is the posterior mode, which is the maximum of the pdf or the pmf of the posterior distribution.

Definition 10.3.3 (Maximum-a-posteriori estimator)**.** *The maximum-a-posteriori (MAP) estimator of a parameter $\vec{\Theta}$ given data $\vec{x}$ modeled as a realization of a random vector $\vec{X}$ is*

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.32}$$

if $\vec{\Theta}$ is modeled as a discrete random variable and

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.33}$$

if it is modeled as a continuous random variable.

In Figure 10.1 the ML estimator of Θ is the mode (maximum value) of the posterior distribution when the prior is uniform. This is not a coincidence, under a uniform prior the MAP and ML estimates are the same.

Lemma 10.3.4. *The maximum-likelihood estimator of a parameter Θ is the mode (maximum value) of the pdf of the posterior distribution given the data $\vec{X}$ if its prior distribution is uniform.*

Proof. We prove the result when the model for the data and the parameters is continuous, if any or both of them are discrete the proof is identical (in that case the ML estimator is the mode of the pmf of the posterior). If the prior distribution of the parameters is uniform, then $f_{\vec{\Theta}}\left(\vec{\theta}\right)$ is constant for any $\vec{\theta}$, which implies

$$\arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}|\vec{x}\right) = \arg\max_{\vec{\theta}} \frac{f_{\vec{\Theta}}\left(\vec{\theta}\right) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right)}{\int_u f_{\vec{\Theta}}\left(u\right) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|u\right) \mathrm{d}u} \tag{10.34}$$

$$= \arg\max_{\vec{\theta}} f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right) \qquad \text{(the rest of the terms do not depend on } \vec{\theta}\text{)}$$

$$= \arg\max_{\vec{\theta}} \mathcal{L}_{\vec{x}}\left(\vec{\theta}\right). \tag{10.35}$$

□

Note that uniform priors are only well defined in situations where the parameter is restricted to a bounded set.

We now describe a situation in which the MAP estimator is optimal. If the parameter Θ can only take a discrete set of values, then the MAP estimator minimizes the probability of making the wrong choice.

Theorem 10.3.5 (MAP estimator minimizes the probability of error)**.** *Let $\vec{\Theta}$ be a discrete random vector and let $\vec{X}$ be a random vector modeling the data. We define*

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{X} = \vec{x}\right). \tag{10.36}$$

For any arbitrary estimator $\theta_{\text{other}}\left(\vec{x}\right)$,

$$\mathrm{P}\left(\theta_{\text{other}}(\vec{X}) \neq \vec{\Theta}\right) \geq \mathrm{P}\left(\theta_{\text{MAP}}(\vec{X}) \neq \vec{\Theta}\right). \tag{10.37}$$

In words, the MAP estimator minimizes the probability of error.

Proof. We assume that $\vec{X}$ is a continuous random vector, but the same argument applies if it is discrete. We have

$$\mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{X})\right) = \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, \mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{x}) \,|\, \vec{X} = \vec{x}\right) \mathrm{d}\vec{x} \tag{10.38}$$

$$= \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\,|\,\vec{X}}(\theta_{\text{other}}(\vec{x}) \,|\, \vec{x})\, \mathrm{d}\vec{x} \tag{10.39}$$

$$\leq \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\,|\,\vec{X}}(\theta_{\text{MAP}}(\vec{x}) \,|\, \vec{x})\, \mathrm{d}\vec{x} \tag{10.40}$$

$$= \mathrm{P}\left(\Theta = \theta_{\text{MAP}}(\vec{X})\right), \tag{10.41}$$

where (10.40) follows from the definition of the MAP estimator as the mode of the posterior. □

Example 10.3.6 (Sending bits)**.** We consider a very simple model for a communication channel in which we aim to send a signal Θ consisting of a single bit. Our prior knowledge indicates that the signal is equal to one with probability 1/4.

$$p_\Theta(1) = \frac{1}{4}, \qquad p_\Theta(0) = \frac{3}{4}. \tag{10.42}$$

Due to the presence of noise in the channel, we send the signal n times. At the receptor we observe

$$\vec{X}_i = \Theta + \vec{Z}_i, \quad 1 \leq i \leq n, \tag{10.43}$$

where $\vec{Z}$ contains n iid standard Gaussian random variables. Modeling perturbations as Gaussian is a popular choice in communications. It is justified by the central limit theorem, under the assumption that the noise is a combination of many small effects that are approximately independent.

We will now compute and compare the ML and MAP estimators of Θ given the observations.

The likelihood is equal to

$$\mathcal{L}_{\vec{x}}(\theta) = \prod_{i=1}^{n} f_{\vec{X}_i|\Theta}(\vec{x}_i \,|\, \theta) \tag{10.44}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\vec{x}_i - \theta)^2}{2}}. \tag{10.45}$$

It is easier to deal with the log-likelihood function,

$$\log \mathcal{L}_{\vec{x}}(\theta) = -\sum_{i=1}^{n} \frac{(\vec{x}_i - \theta)^2}{2} - \frac{n}{2} \log 2\pi. \tag{10.46}$$

Since Θ only takes two values, we can compare directly. We will choose $\theta_{\text{ML}}(\vec{x}) = 1$ if

$$\log \mathcal{L}_{\vec{x}}(1) = -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i + 1}{2} - \frac{n}{2} \log 2\pi \tag{10.47}$$

$$\geq -\sum_{i=1}^{n} \frac{\vec{x}_i^2}{2} - \frac{n}{2} \log 2\pi \tag{10.48}$$

$$= \log \mathcal{L}_{\vec{x}}(0). \tag{10.49}$$

Equivalently,

$$\theta_{\mathrm{ML}}\left(\vec{x}\right) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.50}$$

The rule makes a lot of sense: if the sample mean of the data is closer to 1 than to 0 then our estimate is equal to 1. By the law of total probability, the probability of error of this estimator is equal to

$$\begin{aligned} \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X})\right) &= \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X}) \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X}) \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\ &= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i > \frac{1}{2} \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i < \frac{1}{2} \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\ &= Q\left(\sqrt{n}/2\right), \end{aligned} \tag{10.51}$$

where the last equality follows from the fact that if we condition on $\Theta = \theta$ the empirical mean is Gaussian with variance σ^2/n and mean θ (see the proof of Theorem 6.2.2).

To compute the MAP estimate we must find the maximum of the posterior pdf of Θ given the observed data. Equivalently, we find the maximum of its logarithm (this is equivalent because the logarithm is a monotone function),

$$\log p_{\Theta|\vec{X}}\left(\theta|\vec{x}\right) = \log \frac{\prod_{i=1}^{n} f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right) p_{\Theta}\left(\theta\right)}{f_{\vec{X}}\left(\vec{x}\right)} \tag{10.52}$$

$$= \sum_{i=1}^{n} \log f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right) p_{\Theta}\left(\theta\right) - \log f_{\vec{X}}\left(\vec{x}\right) \tag{10.53}$$

$$= -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i\theta + \theta^2}{2} - \frac{n}{2}\log 2\pi + \log p_{\Theta}\left(\theta\right) - \log f_{\vec{X}}\left(\vec{x}\right). \tag{10.54}$$

We compare the value of this function for the two possible values of Θ: 0 and 1. We choose $\theta_{\mathrm{MAP}}\left(\vec{x}\right) = 1$ if

$$\log p_{\Theta|\vec{X}}\left(1|\vec{x}\right) + \log f_{\vec{X}}\left(\vec{x}\right) = -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i + 1}{2} - \frac{n}{2}\log 2\pi - \log 4 \tag{10.55}$$

$$\geq -\sum_{i=1}^{n} \frac{\vec{x}_i^2}{2} - \frac{n}{2}\log 2\pi - \log 4 + \log 3 \tag{10.56}$$

$$= \log p_{\Theta|\vec{X}}\left(0|\vec{x}\right) + \log f_{\vec{X}}\left(\vec{x}\right). \tag{10.57}$$

Equivalently,

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i > \frac{1}{2} + \frac{\log 3}{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.58}$$

The MAP estimate shifts the threshold with respect to the ML estimate to take into account that Θ is more prone to equal zero. However, the correction term tends to zero as we gather more evidence, so if a lot of data is available the two estimators will be very similar.

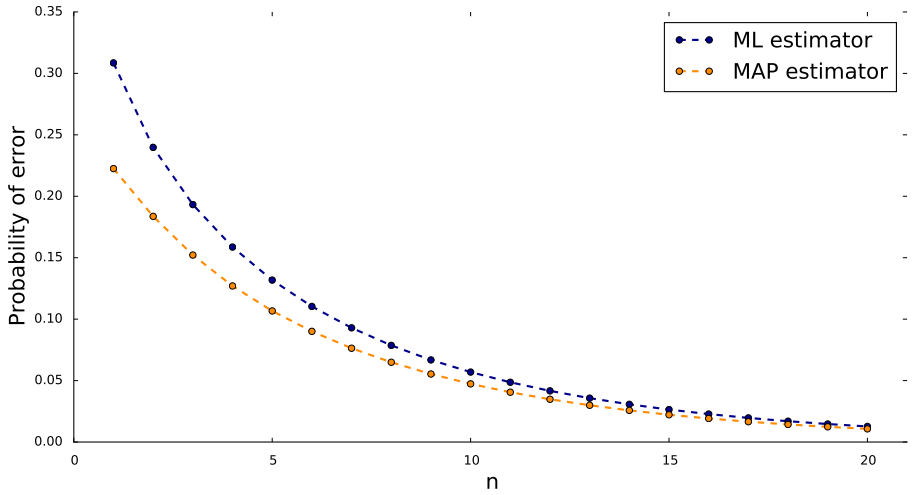

Figure 10.3: Probability of error of the ML and MAP estimators in Example 10.3.6 for different values of n.

The probability of error of the MAP estimator is equal to

$$\begin{aligned}
\mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X})\right) &= \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X}) | \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X}) | \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\
&= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i > \frac{1}{2} + \frac{\log 3}{n} \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) \qquad (10.59) \\
&\quad + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i < \frac{1}{2} + \frac{\log 3}{n} \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\
&= \frac{3}{4} Q\left(\sqrt{n}/2 + \frac{\log 3}{\sqrt{n}}\right) + \frac{1}{4} Q\left(\sqrt{n}/2 - \frac{\log 3}{\sqrt{n}}\right). \qquad (10.60)
\end{aligned}$$

We compare the probability of error of the ML and MAP estimators in Figure 10.3. MAP estimation results in better performance, but the difference becomes small as n increases.

△