General Bayesian Inference I

Outline:

- Basic concepts,
- One-parameter models,
- Noninformative priors.

Reading: Chapters 10 and 11 in Kay-I.

(Occasional) Simplified Notation. When there is no potential for confusion, we may use

$$f(x \,|\, \theta)$$

instead of the more cumbersome

$$f_{X \,|\, \Theta}(x \,|\, \theta).$$

Basic Concepts

X is the observable random variable and θ is the "true state of nature";

$f_{X \,|\, \Theta}(x \,|\, \theta)$ or $p_{X \,|\, \Theta}(x \,|\, \theta)$ denote the data model, likelihood;

$\pi(\theta)$ or $f_{\Theta}(\theta)$ is the prior distribution of θ (epistemic probability), i.e. our knowledge about the true state of nature.

In the Bayesian approach, assign a prior distribution on parameter θ. Here, note that θ is often *not* really random, but the epistemic argument justifies the use of a probability distribution. We apply the Bayes' rule and base our inference on the posterior distribution of θ:

$$f_{\Theta \,|\, X}(\theta \,|\, x) = \frac{f_{X,\Theta}(x,\theta)}{\underbrace{\int f_{X,\Theta}(x,\vartheta)\, d\vartheta}_{\text{does not depend on } \theta}} \propto f_{X,\Theta}(x,\theta)$$

and, therefore,

$$f_{\Theta \,|\, X}(\theta \,|\, x) \propto \underbrace{f_{X \,|\, \Theta}(x \,|\, \theta)\, \pi(\theta)}_{f_{X,\Theta}(x,\theta)}.$$

Note:

- $f_{\Theta \,|\, X}(\theta \,|\, x)$ is also an epistemic probability.
- It is important to master the use of $\propto$.

A Side Comment (Exercise). Make sure that you understand the following:

$$\begin{aligned} f(\theta \,|\, x_1, x_2) &\propto f(\theta, x_1, x_2) \\ &\propto f(\theta, x_1 \,|\, x_2) \\ &\propto f(\theta, x_2 \,|\, x_1). \end{aligned}$$

More Exercise. Consider

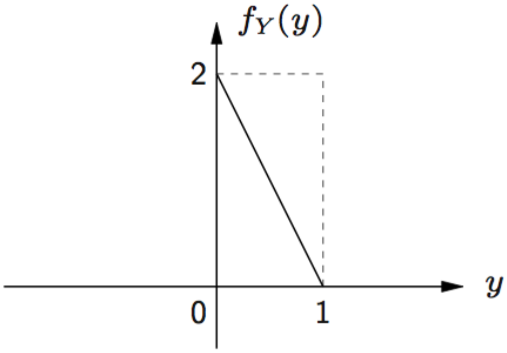
$$f_{X,Y}(x, y) = \begin{cases} 2 & x, y \geq 0, x + y \leq 1 \\ 0, & \text{otherwise} \end{cases} .$$

Now, computing $f_{X \,|\, Y}(x \,|\, y)$ can be done two ways.

The first (harder) way requires the computation of $f_Y(y)$:

To find $f_Y(y)$, we use the law of total probability

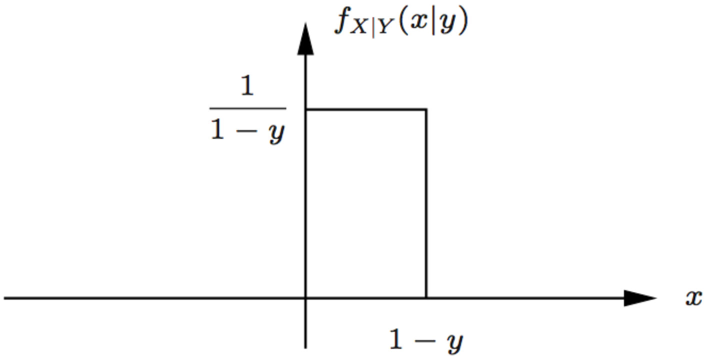
$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\
&= \begin{cases} \int_0^{(1-y)} 2\,dx & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2(1-y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$



Then,

$$
f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{1-y} & 0 \le y < 1,\, 0 \le x \le 1-y \\ 0 & \text{otherwise} \end{cases}
$$

In other words, $X \,|\, \{Y = y\} \sim \mathrm{U}[\,0, 1-y\,]$



The second (easier) way employs the $\propto$ notation:

$$\begin{aligned} f_{X \mid Y}(x \mid y) &\propto f_{X,Y}(x,y) \\ &= \left\{ \begin{array}{ll} \underbrace{\text{function}(y)}_{\text{flat in } x}, & 0 \le x \le 1-y \\ 0, & \text{otherwise} \end{array} \right. \cdot i_{(0,1)}(y). \end{aligned}$$

Now, $\text{function}(y)$ is just a normalizing constant satisfying

$$\int_{-\infty}^{+\infty} f_{X \mid Y}(x \mid y)\, dx = 1$$

which yields

$$f_{X \mid Y}(x \mid y) = \left\{ \begin{array}{ll} \frac{1}{1-y}, & 0 \le x \le 1-y \\ 0, & \text{otherwise} \end{array} \right. \cdot i_{(0,1)}(y).$$

Conjugate Priors

If F is a class of measurement models and P a class of prior distributions, then P is *conjugate* for F if $\pi(\theta) \in P$ and $f_{X \mid \Theta}(x \mid \theta) \in F$ implies $f_{\Theta \mid X}(\theta \mid x) \in P$. It is convenient to choose conjugate priors: they allow finding analytically tractable posteriors.

Important special case: If F is the exponential family of distributions, then we have natural conjugate priors. Consider

$$f_{X \mid \Theta}(x_n \mid \boldsymbol{\theta}) = h(x_n)\, q(\boldsymbol{\theta})\, \exp[\boldsymbol{\eta}^T(\boldsymbol{\theta})\, \boldsymbol{t}(x_n)] \quad n = 1, 2, \ldots, N.$$

For conditionally independent, identically distributed (i.i.d.) x_n given $\boldsymbol{\Theta} = \boldsymbol{\theta}$, the likelihood function is

$$f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \Big[\prod_{n=1}^{N} h(x_n)\Big] \cdot \underbrace{q^N(\boldsymbol{\theta})}_{[q(\boldsymbol{\theta})]^N}\, \exp[\boldsymbol{\eta}^T(\boldsymbol{\theta})\, \boldsymbol{T}(\boldsymbol{x})]$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$ and the natural sufficient statistic is

$$\boldsymbol{T}(\boldsymbol{x}) = \sum_{n=1}^{N} \boldsymbol{t}(x_n).$$

Consider the following prior pdf/pmf:

$$\pi(\boldsymbol{\theta}) \propto q^{\xi}(\boldsymbol{\theta})\, \exp[\boldsymbol{\eta}^T(\boldsymbol{\theta})\, \boldsymbol{\nu}]. \tag{1}$$

Then, the posterior pdf/pmf is

$$f_{\Theta\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto q^{N+\xi}(\boldsymbol{\theta})\ \exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\,[\boldsymbol{T}(\boldsymbol{x})+\boldsymbol{\nu}]\}$$

and hence $\pi(\boldsymbol{\theta})$ is indeed the conjugate prior for $f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\boldsymbol{\theta})$.

Example. Consider the following binomial data model:

$$\begin{aligned} p_{X\,|\,\Theta}(x\,|\,\theta) &= \mathrm{Bin}(x\,|\,N,\theta) \\ &= \binom{N}{x}\,\theta^x\,(1-\theta)^{N-x}\,i_{(0,1)}(\theta) \\ &\overset{\text{exp. fam.}}{=} \underbrace{\binom{N}{x}}_{h(x)}\ \underbrace{[(1-\theta)\,i_{(0,1)}(\theta)]^N}_{q(\theta)}\ \exp\Big(\underbrace{\ln\big(\frac{\theta}{1-\theta}\big)}_{\eta(\theta)}\ \underbrace{x}_{t(x)}\Big)\end{aligned}$$

where N is a known constant (number of coin flips, say) and $\theta\in[0,1]$ is the parameter (probability of heads, say). A conjugate prior family of pdfs for θ follows by using (1):

$$\begin{aligned}\pi(\theta) &\propto \underbrace{(1-\theta)^{N\,\xi}\,i_{(0,1)}(\theta)}_{[q(\theta)]^\xi}\ \underbrace{\exp\Big(\ln\big(\frac{\theta}{1-\theta}\big)\,\nu\Big)}_{\exp[\eta(\theta)\,\nu]} \\ &= (1-\theta)^{N\,\xi-\nu}\,\theta^\nu\,i_{(0,1)}(\theta)\end{aligned}$$

where we recognize the kernel of the family of beta pdfs from the table of distributions; this family is traditionally parametrized as follows:

$$\pi(\theta) = \mathrm{Beta}(\theta \,|\, \alpha, \beta) \propto \theta^{\alpha-1} \, (1-\theta)^{\beta-1} \, i_{(0,1)}(\theta).$$

Sequential-Bayesian Idea

Suppose that we have observed x_1 and x_2, where x_1 comes first (e.g. the subscript is a time index). We wish to make inference about θ. Then, conditioning on $X_1 = x_1$ yields

$$f_{X_2,\Theta \,|\, X_1}(x_2, \theta \,|\, x_1) = f_{X_2 \,|\, \Theta, X_1}(x_2 \,|\, \theta, x_1) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1) \quad (2)$$

where

$f_{X_2 \,|\, X_1,\Theta}(x_2 \,|\, x_1, \theta)$ new, updated likelihood for θ based on x_2

and

$f_{\Theta \,|\, X_1}(\theta \,|\, x_1)$ new, updated prior for θ.

Now, (2) implies

$$f_{\Theta \,|\, X_1, X_2}(\theta \,|\, x_1, x_2) \propto f_{X_2 \,|\, \Theta, X_1}(x_2 \,|\, \theta, x_1) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1). \quad (3)$$

Conditionally independent observations X_1 and X_2 given Θ. In the special case where X_1 and X_2 are conditionally independent given $\Theta = \theta$, we have

$$f_{X_1, X_2 \,|\, \Theta}(x_1, x_2 \,|\, \theta) = f_{X_1 \,|\, \Theta}(x_1 \,|\, \theta) \cdot f_{X_2 \,|\, \Theta}(x_2 \,|\, \theta) \quad (4)$$

and, consequently, by the definition of conditional independence,

$$f_{X_2 \,|\, X_1, \Theta}(x_2 \,|\, x_1, \theta) = f_{X_2 \,|\, \Theta}(x_2 \,|\, \theta). \quad (5)$$

[We can also go from (4) to (5) as follows:

$$\begin{aligned} f(x_2 \,|\, x_1, \theta) &\propto f(x_2, x_1 \,|\, \theta) = f(x_2 \,|\, \theta) \cdot f(x_1 \,|\, \theta) \\ &\propto f(x_2 \,|\, \theta). \end{aligned}$$

This exercise is a good practice for familiarizing with the $\propto$ notation.]

Substituting (5) into (3) yields

$$f_{\Theta \,|\, X_1, X_2}(\theta \,|\, x_1, x_2) \propto \underbrace{f_{X_2 \,|\, \Theta}(x_2 \,|\, \theta)}_{\text{ordinary likelihood based on } x_2} \cdot \underbrace{f_{\Theta \,|\, X_1}(\theta \,|\, x_1)}_{\text{new prior}}. \quad (6)$$

On Prediction

We continue with the scenario described on the last two pages. Suppose that we have observed $X_1 = x_1$ and wish to predict X_2. For this purpose, we use the *predictive distribution* (say a pdf, for simplicity of exposition):

$$f_{X_2 \,|\, X_1}(x_2 \,|\, x_1) \tag{7}$$

We derive this pdf as follows. Recall (2):

$$f_{X_2,\Theta \,|\, X_1}(x_2, \theta \,|\, x_1) = f_{X_2 \,|\, \Theta, X_1}(x_2 \,|\, \theta, x_1) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1). \tag{8}$$

Now, marginalize the pdf in (8) with respect to the unknown parameter Θ, i.e. integrate Θ out:

$$\begin{aligned} f_{X_2 \,|\, X_1}(x_2 \,|\, x_1) &= \int f_{X_2,\Theta \,|\, X_1}(x_2, \theta \,|\, x_1)\, d\theta \\ &= \int f_{X_2 \,|\, \Theta, X_1}(x_2 \,|\, \theta, x_1)\, f_{\Theta \,|\, X_1}(\theta \,|\, x_1)\, d\theta. \end{aligned} \tag{9}$$

Conditionally independent observations $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ given $\boldsymbol{\Theta = \theta}$. In the special case where X_1 and X_2 are conditionally independent given $\Theta = \theta$, (4) and (5) hold, and, therefore, (9) simplifies to

$$f_{X_2 \,|\, X_1}(x_2 \,|\, x_1) = \int f_{X_2 \,|\, \Theta}(x_2 \,|\, \theta) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1)\, d\theta. \tag{10}$$

The First (Ever) Bayesian Model: Binomial Measurements

Suppose that X_1 and X_2 are independent given $\Theta = \theta$, coming from

$$\{X_i \,|\, \Theta = \theta\} \sim \mathrm{Bin}(N_i, \theta) \quad i = 1, 2$$

i.e. the joint likelihood of x_1 and x_2 is

$$p_{X_1,X_2 \,|\, \Theta}(x_1, x_2 \,|\, \theta) = p_{X_1 \,|\, \Theta}(x_1 \,|\, \theta) \cdot p_{X_2 \,|\, \Theta}(x_2 \,|\, \theta)$$
$$= \binom{N_1}{x_1} \theta^{x_1} (1-\theta)^{N_1 - x_1} \cdot \binom{N_2}{x_2} \theta^{x_2} (1-\theta)^{N_2 - x_2} \cdot i_{(0,1)}(\theta).$$

As we have seen on p. 7, the conjugate prior pdf family for θ under this data model is

$$\pi(\theta) = \mathrm{Beta}(\theta \,|\, \alpha, \beta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1} i_{(0,1)}(\theta).$$

Therefore, the posterior pdf of θ is

$$\begin{aligned} f_{\Theta \,|\, X_1,X_2}(\theta \,|\, x_1, x_2) &\propto p_{X_1,X_2 \,|\, \Theta}(x_1, x_2 \,|\, \theta) \, \pi(\theta) \\ &\propto \quad \theta^{x_1+x_2+\alpha-1} (1-\theta)^{N_1+N_2-x_1-x_2+\beta-1} \cdot i_{(0,1)}(\theta) \end{aligned}$$

which is the kernel of the $\mathrm{Beta}(x_1+x_2+\alpha, \beta+N_1-x_1+N_2-x_2)$ pdf, see the table of distributions. Hence,

$$\begin{aligned} &f_{\Theta \,|\, X_1,X_2}(\theta \,|\, x_1, x_2) \\ &\quad = \quad \mathrm{Beta}(\theta \,|\, x_1 + x_2 + \alpha, \beta + N_1 - x_1 + N_2 - x_2). \quad (11) \end{aligned}$$

How about the posterior pdf $f_{\Theta \,|\, X_1}(\theta \,|\, x_1)$ given only $X_1 = x_1$? Now,

$$\begin{aligned} f_{\Theta \,|\, X_1}(\theta \,|\, x_1) &\propto p_{X_1 \,|\, \Theta}(x_1 \,|\, \theta)\, \pi(\theta) \\ &\propto \theta^{x_1+\alpha-1}\, (1-\theta)^{N_1-x_1+\beta-1} \cdot i_{(0,1)}(\theta) \end{aligned}$$

which is the kernel of the $\mathrm{Beta}(x_1+\alpha, \beta+N_1-x_1)$ pdf; therefore,

$$f_{\Theta \,|\, X_1}(\theta \,|\, x_1) = \mathrm{Beta}(\theta \,|\, x_1+\alpha, \beta+N_1-x_1).$$

Since X_1 and X_2 are conditionally independent given $\Theta = \theta$, we apply (6) (for practice, to verify its validity):

$$\begin{aligned} & f_{\Theta \,|\, X_1, X_2}(\theta \,|\, x_1, x_2) \\ & \qquad \underbrace{\propto}_{\text{keep track of } \theta} \quad p_{X_2 \,|\, \Theta}(x_2 \,|\, \theta) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1) \\ & = \binom{N_2}{x_2} \theta^{x_2}\, (1-\theta)^{N_2-x_2} \\ & \cdot \frac{\Gamma(\overbrace{x_1+\alpha+\beta+N_1-x_1}^{\alpha+\beta+N_1})}{\Gamma(x_1+\alpha)\Gamma(\beta+N_1-x_1)} \theta^{x_1+\alpha-1}\, (1-\theta)^{\beta+N_1-x_1-1} \cdot i_{(0,1)}(\theta) \\ & \qquad \propto \underbrace{\theta^{x_1+x_2+\alpha-1}\, (1-\theta)^{N_1+N_2-x_1-x_2+\beta-1} \cdot i_{(0,1)}(\theta)}_{\text{kernel of } \mathrm{Beta}(\theta \,|\, x_1+x_2+\alpha, \beta+N_1-x_1+N_2-x_2)} \end{aligned}$$

which corresponds to (11). **Therefore, we can either perform sequential or batch inference: both approaches lead to the same answer.**

How about predicting X_2 after observing $X_1 = x_1$? Since X_1 and X_2 are independent given $\Theta = \theta$, we apply (10):

$$p_{X_2 \,|\, X_1}(x_2 \,|\, x_1) = \int_0^1 p_{X_2 \,|\, \Theta}(x_2 \,|\, \theta) \cdot f_{\Theta \,|\, X_1}(\theta \,|\, x_1) \, d\theta$$

$$= \binom{N_2}{x_2} \cdot \frac{\Gamma(\alpha + \beta + N_1)}{\Gamma(x_1 + \alpha)\Gamma(\beta + N_1 - x_1)}$$

$$\cdot \underbrace{\int_0^1 \theta^{x_1 + x_2 + \alpha - 1} \, (1 - \theta)^{\beta + N_1 - x_1 + N_2 - x_2 - 1} \, d\theta}_{\frac{\Gamma(x_1 + x_2 + \alpha)\,\Gamma(\beta + N_1 - x_1 + N_2 - x_2)}{\Gamma(\alpha + \beta + N_1 + N_2)}}$$

$$= \binom{N_2}{x_2} \cdot \frac{\Gamma(\alpha + \beta + N_1)}{\Gamma(x_1 + \alpha)\,\Gamma(\beta + N_1 - x_1)}$$

$$\cdot \frac{\Gamma(x_1 + x_2 + \alpha)\,\Gamma(\beta + N_1 - x_1 + N_2 - x_2)}{\Gamma(\alpha + \beta + N_1 + N_2)}$$

which is the predictive pmf of X_2 given $X_1 = x_1$.

Comments:

- Here, we have used the fact that $\mathrm{Beta}(\alpha, \beta)$ pdf of a random variable Θ has the following form (see the distribution table):

$$f_\Theta(\theta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}}_{\text{normalizing constant}} \cdot \theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

implying that

$$\int_0^1 \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}\,d\theta = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

- Bayes used a special case of the above model.
- Laplace computed posterior probabilities under a special case of this model. In particular, he considered a single observation x_1 (the number of girls born in Paris over a time period in the 18th century) coming from

$$\{X_1 \,|\, \Theta = \theta\} \sim \text{Bin}\Big(N_1, \underbrace{\theta}_{\text{prob. that a newborn child is a girl}}\Big)$$

and set the following prior pdf:

$$\pi(\theta) = \text{uniform}(\theta \,|\, 0, 1) = \text{Beta}(\theta \,|\, 1, 1).$$

Here is the measurement:

$$x_1 = 241,945$$

and $N_1 = 241,945 + 251,527$. Laplace computed

$$\Pr\{\Theta \geq 0.5 \,|\, X_1 = x_1\} \approx 10^{-42}.$$

An Example of Bayesian Inference: DC-level Estimation in AWGN with Known Variance

Single Observation. Choose the data model:

$$f_{X\,|\,\Theta}(x\,|\,\theta) = \mathcal{N}(x\,|\,\theta, \sigma^2)$$

where we assume that σ^2 is known. Hence, the likelihood for one measurement is

$$f_{X\,|\,\Theta}(x\,|\,\theta) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \cdot \exp\left[-\frac{1}{2\,\sigma^2}\,(x-\theta)^2\right]. \qquad (12)$$

To obtain the conjugate prior pdf for θ, we view $f_{X\,|\,\Theta}(x\,|\,\theta)$ as a function of θ:

$$\begin{aligned} f_{X\,|\,\Theta}(x\,|\,\theta) \quad & \underbrace{\propto}_{\text{keep track of } \theta} \quad \exp\left[-\underbrace{\frac{1}{2\,\sigma^2}\,(\theta^2 - 2\,x\,\theta + x^2)}_{\text{quadratic in } \theta}\right] \\ \propto \quad & \underbrace{\exp\left[-\frac{1}{2\,\sigma^2}\,(\theta^2 - 2\,x\,\theta)\right]}_{\text{kernel of the Gaussian pdf with mean } x \text{ and variance } \sigma^2} \qquad (13) \\ \propto \quad & \underbrace{\exp\left(-\frac{1}{2\,\sigma^2}\,\theta^2 + \frac{x\,\theta}{\sigma^2}\right)}_{\text{kernel of the Gaussian pdf with mean } x \text{ and variance } \sigma^2} \end{aligned}$$

Then, according to the results from p. 6,

$$\begin{aligned} q(\theta) &= \exp\Big(-\frac{1}{2\,\sigma^2}\,\theta^2\Big) \\ t(x) &= x \\ \eta(\theta) &= \frac{\theta}{\sigma^2} \end{aligned}$$

and the conjugate prior pdf for θ has the following form:

$$\pi(\theta) \propto \underbrace{\exp(-\frac{\xi}{2\,\sigma^2}\,\theta^2)}_{q(\theta)^\xi}\ \underbrace{\exp(\frac{\theta}{\sigma^2}\,\nu)}_{\exp[\eta(\theta)\,\nu]}$$

which can be reparametrized as

$$\pi(\theta) \propto \exp\Big[-\frac{1}{2\,\tau_0^2}\,(\theta-\mu_0)^2\Big]$$

and we conclude that the conjugate prior pdf for the likelihood function in (12) is

$$\pi(\theta) = \mathcal{N}(\theta\,|\,\mu_0, \tau_0^2).$$

Here, μ_0 and τ_0^2 are *known hyperparameters*. (Of course, we can continue and assign a prior joint pdf for the hyperparameters, which would lead to a hierarchical Bayesian model.) We now compute the posterior pdf by collecting the terms that contain θ and θ^2:

$$
\begin{aligned}
f_{\Theta\,|\,X}(\theta\,|\,x) \propto f_{X\,|\,\Theta}(x\,|\,\theta)\,\pi(\theta) & \\
\propto \quad & \exp\Big[-\frac{1}{2}\cdot\Big(\frac{x^2 - 2\,x\,\theta + \theta^2}{\sigma^2} + \frac{\theta^2 - 2\,\mu_0\,\theta + \mu_0^2}{\tau_0^2}\Big)\Big] \\
\propto \quad & \exp\Big[-\frac{1}{2}\cdot\frac{(\tau_0^2 + \sigma^2)\,\theta^2 - 2\,(x\,\tau_0^2 + \mu_0\,\sigma^2)\,\theta}{\sigma^2\,\tau_0^2}\Big] \\
\propto \quad & \exp\Big[-\frac{1}{2}\cdot\underbrace{\frac{\sigma^2 + \tau_0^2}{\sigma^2\,\tau_0^2}}_{1/\tau_1^2}\cdot\Big(\theta^2 - 2\,\underbrace{\frac{x\,\tau_0^2 + \mu_0\,\sigma^2}{\sigma^2 + \tau_0^2}}_{\mu_1}\,\theta\Big)\Big]
\end{aligned}
$$

implying that $f_{\Theta\,|\,X}(\theta\,|\,x)$ is a Gaussian pdf with mean and variance

$$
\begin{aligned}
\mu_1 \quad & = \quad \mu_1(x) = \frac{x\,\tau_0^2 + \mu_0\,\sigma^2}{\sigma^2 + \tau_0^2} \\
\tau_1^2 \quad & = \quad \frac{\sigma^2\,\tau_0^2}{\sigma^2 + \tau_0^2}.
\end{aligned}
$$

compare with (13). We will generalize the above expressions to multiple measurements.

Comments on the Single Observation Case:

- The posterior mean is a weighted average of the observation and the prior mean:

$$\begin{aligned}\mu_1 &= \mu_1(x) = \frac{x\,\tau_0^2 + \mu_0\,\sigma^2}{\sigma^2 + \tau_0^2} = \frac{\frac{1}{\sigma^2}\,x + \frac{1}{\tau_0^2}\,\mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}} \\ &= \frac{\text{likelihood precision} \cdot x + \text{prior precision} \cdot \mu_0}{\text{likelihood precision} + \text{prior precision}}.\end{aligned}$$

- We will show that the posterior mean is the (Bayesian) minimum mean-square error (MMSE) estimate of θ.

- Here, the weights are given by *precisions* $\frac{1}{\sigma^2}$ and $\frac{1}{\tau_0^2}$. (The inverse of the variance of a Gaussian distribution is called *precision*.)

- As the likelihood precision $\frac{1}{\sigma^2}$ increases, we have

$$\mu_1(x) \to x.$$

- As the prior precision $\frac{1}{\tau_0^2}$ increases, we have

$$\mu_1(x) \to \mu_0.$$

- The posterior mean is the measurement x *shifted* towards the prior mean (the right word in *shrunk* when the prior mean is zero, due to the magnitude reduction):

$$\mu_1(x) = x - \frac{\sigma^2}{\sigma^2 + \tau_0^2}\,(x - \mu_0)$$

or the prior mean adjusted towards the measurement x:

$$\mu_1(x) = \mu_0 + \frac{\tau_0^2}{\sigma^2 + \tau_0^2}\,(x - \mu_0).$$

- Posterior precision is the sum of the prior and likelihood precisions:

$$\frac{1}{\tau_1^2} = \frac{\sigma^2 + \tau_0^2}{\sigma^2\,\tau_0^2} = \frac{1}{\sigma^2} + \frac{1}{\tau_0^2}.$$

Multiple Conditionally I.I.D. Observations given the mean parameter $\boldsymbol{\theta}$. Consider now N conditionally i.i.d. observations $X[0], X[1], \ldots, X[N-1]$ given θ:

$$\begin{aligned} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) &\propto \pi(\theta)\, f_{\boldsymbol{X} \,|\, \Theta}(\boldsymbol{x} \,|\, \theta) \\ &\propto \quad \exp\big[-\frac{1}{2\,\tau_0^2}\,(\theta - \mu_0)^2\big] \cdot \prod_{n=0}^{N-1} \exp\Big[-\frac{1}{2\,\sigma^2}(x[n] - \theta)^2\Big] \end{aligned}$$

where $\boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]^T$. This posterior pdf depends on $\boldsymbol{x}$ only through the sample mean

$$\overline{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

i.e. $\overline{x}$ is the *sufficient statistic* for θ in this model. Note that

$$\{\overline{X} \,|\, \Theta = \theta\} \quad \sim \quad \underbrace{\mathcal{N}(\theta, \sigma^2/N)}_{\substack{\text{new likelihood} \\ \text{(using sufficiency)}}} \quad .$$

By employing sufficiency, we reduce our problem to the single-observation case, where $\overline{x}$ is our equivalent single observation. Hence,

$$\begin{aligned} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) &\overset{\text{sufficiency}}{=} f_{\Theta \,|\, \overline{X}}(\theta \,|\, \overline{x}) \propto \pi(\theta)\, f_{\overline{X} \,|\, \Theta}(\overline{x} \,|\, \theta) \\ &= \mathcal{N}(\theta \,|\, \mu_N(\overline{x}), \tau_N^2) \qquad (14) \end{aligned}$$

with

$$\mu_N(\overline{x}) = \frac{\frac{N}{\sigma^2}\,\overline{x} + \frac{1}{\tau_0^2}\,\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}} \qquad \frac{1}{\tau_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau_0^2} \qquad (15)$$

see also Example 10.2 in Kay-I.

Comments:

- If N is large, the influence of the prior pdf disappears and the posterior pdf effectively depends only on $\overline{x}$ and σ^2.

- If $\tau_0^2 = \sigma^2$, the prior has the same weight as adding one more observation with value μ_0.

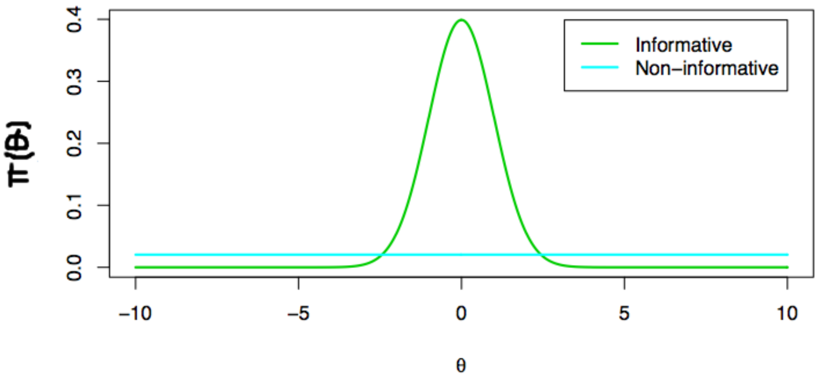
- When $\tau_0^2 \nearrow +\infty$ with N fixed or $N \nearrow +\infty$ with τ_0 fixed, we have
$$f_{\Theta \,|\, \overline{X}}(\theta \,|\, \overline{x}) \to \mathcal{N}\Big(\theta \,\Big|\, \overline{x}, \frac{\sigma^2}{N}\Big) \tag{16}$$
which is a good general approximation whenever our prior knowledge about θ is vague or the number of observations N is large. In this scenario, the influence of the prior disappears. Furthermore, $\tau_0^2 \nearrow +\infty$ corresponds to
$$\pi(\theta) \propto 1 \tag{17}$$
and leads to the posterior pdf proportional to the likelihood:
$$f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) \overset{\text{sufficiency}}{=} f_{\Theta \,|\, \overline{X}}(\theta \,|\, \overline{x}) \propto \underbrace{\pi(\theta)}_{\propto 1, \text{ see (17)}} f_{\overline{X} \,|\, \Theta}(\overline{x} \,|\, \theta)$$
$$\propto \underbrace{f_{\overline{X} \,|\, \Theta}(\overline{x} \,|\, \theta)}_{\text{likelihood}}.$$

The prior choice (17) does not describe a valid probability density, since

$$\int_{-\infty}^{+\infty} 1 = +\infty.$$

Hence, (17) is an *improper prior*. However, we can still use it because the posterior pdf in (16) is proper.

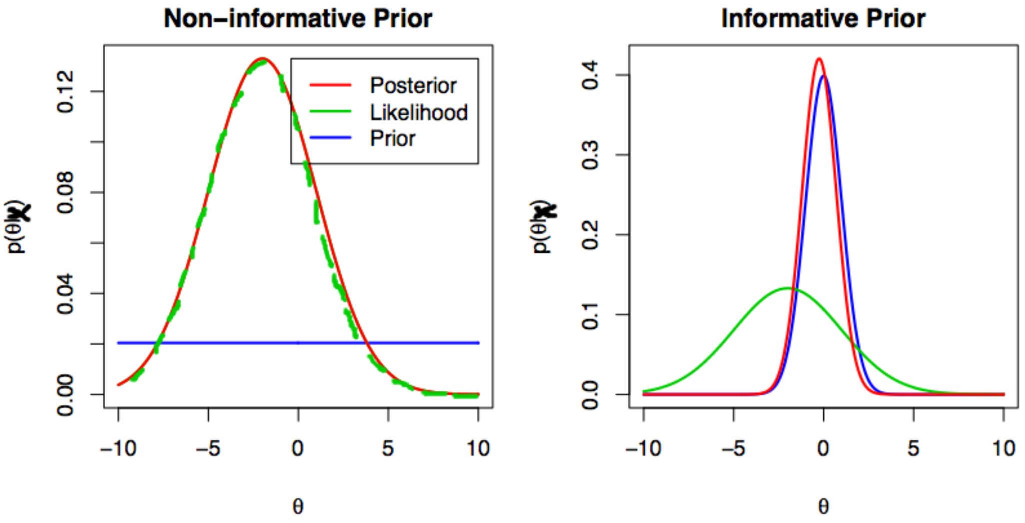
If τ_0^2 is large, we obtain a noninformative prior:



Recall: The posterior mean and precision are

$$\mu_N(\overline{x}) = \frac{\frac{N}{\sigma^2}\,\overline{x} + \frac{1}{\tau_0^2}\,\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\frac{1}{\tau_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau_0^2}$$



Sufficiency and Bayesian Models

Since we have just applied sufficiency to simplify our Bayesian calculations, perhaps it is a good idea to formally state the following (Kolmogorov's) result:

Theorem 1. *If a statistic $T(\mathbf{X})$ is sufficient for a parameter θ, then*

$$f_{\Theta \,|\, T(\boldsymbol{X})}(\theta \,|\, T(\boldsymbol{x})) = f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}).$$

For Bayesians, the statement

$$f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) = f_{\Theta \,|\, T(\boldsymbol{X})}(\theta \,|\, T(\boldsymbol{x}))$$

is the definition of a sufficient statistics $T(\boldsymbol{x})$ for θ. Note that the factorization theorem applies to the posterior $f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x})$ the same way it does to the likelihood $f_{\boldsymbol{X} \,|\, \Theta}(\boldsymbol{x} \,|\, \theta)$.

Back to DC-level Estimation in AWGN with Known Variance: Predictive Distribution

Suppose that we have collected N conditionally i.i.d. observations $X[0], X[1], \ldots, X[N-1]$ given θ, following the DC-level model from p. 21:

$$\{X[n] \,|\, \Theta = \theta\} \sim \mathcal{N}(\theta, \sigma^2).$$

We wish to predict the next observation, denoted by $X_\star$, which is conditionally independent of $X[0], X[1], \ldots, X[N-1]$ given $\Theta = \theta$, and

$$\{X_\star \,|\, \Theta = \theta\} \sim \mathcal{N}(\theta, \sigma^2). \tag{18}$$

Recall that

$$\overline{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$$

is a sufficient statistic for θ based on $X[0], X[1], \ldots, X[N-1]$ and that

$$f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) = f_{\Theta \,|\, \overline{X}}(\theta \,|\, \overline{x})$$

where $\boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]^T$. Then, we apply the identity (10) to obtain

$$f_{X_\star \,|\, \overline{X}}(x_\star \,|\, \overline{x}) = \int \underbrace{f_{X_\star \,|\, \Theta}(x_\star \,|\, \vartheta)\, f_{\Theta \,|\, \overline{X}}(\vartheta \,|\, \overline{x})}_{f_{X_\star,\Theta \,|\, \overline{X}}(x_\star,\vartheta \,|\, \overline{x})} \, d\vartheta. \tag{19}$$

The fact that $X_\star$ and $X[0], X[1], \ldots, X[N-1]$ are conditionally independent given $\Theta = \theta$ also implies

- $$f_{X_\star \,|\, \Theta, \overline{X}}(x_\star \,|\, \theta, \overline{x}) = f_{X_\star \,|\, \Theta}(x_\star \,|\, \theta) \tag{20}$$

  which is analogous to (5), and

- based on (19):

  $$\mathrm{E}_{X_\star,\Theta \,|\, \overline{X}} = \mathrm{E}_{\Theta \,|\, \overline{X}}\{\mathrm{E}_{X_\star \,|\, \Theta}[\cdot \,|\, \Theta] \,|\, \overline{X}\} \tag{21}$$

  which we will use to apply the laws of iterated expectations and conditional variances.

We focus on the integrand of (19):

$$f_{X_\star,\Theta \,|\, \overline{X}}(x_\star, \theta \,|\, \overline{x}) = \underbrace{f_{X_\star \,|\, \Theta}(x_\star \,|\, \theta)}_{\mathcal{N}(x_\star \,|\, \theta,\sigma^2)} \underbrace{f_{\Theta \,|\, \overline{X}}(\theta \,|\, \overline{x})}_{\mathcal{N}(\theta \,|\, \mu_N,\tau_N^2)}$$

and, therefore,

$$\begin{aligned} f_{X_\star,\Theta \,|\, \overline{X}}(x_\star, \theta \,|\, \overline{x}) \quad &\propto \quad \exp\Big[-\frac{1}{2\,\sigma^2}\,(x_\star-\theta)^2\Big] \\ &\qquad \cdot \exp\Big\{-\frac{1}{2\,\tau_N^2}\,[\theta-\mu_N(\overline{x})]^2\Big\} \end{aligned}$$

which is kernel of a bivariate Gaussian pdf, see p. 15 in handout # 0b. Hence,

$$f_{X_\star,\Theta \,|\, \overline{X}}(x_\star, \theta \,|\, \overline{x})$$

is a bivariate Gaussian pdf. We wish to find the predictive pdf

$$f_{X_\star \,|\, \overline{X}}(x_\star \,|\, \overline{x}).$$

Now, integrating θ out (i.e. marginalizing with respect to θ) in (19) is easy, see Property 3 on p. 25 of handout # 0b. Since we know that the predictive pdf $f_{X_\star \,|\, \overline{X}}(x_\star \,|\, \overline{x})$ must be Gaussian, we just need to find its mean:

$$\begin{aligned} &\underbrace{\mathrm{E}_{\,X_\star \,|\, \overline{X}}[X_\star \,|\, \overline{x}]}_{\mathrm{E}_{\,X_\star,\Theta \,|\, \overline{X}}[X_\star \,|\, \overline{x}]} \overset{\text{iter. exp.}}{=} \underbrace{\mathrm{E}_{\,\Theta \,|\, \overline{X}}[\mathrm{E}_{\,X_\star \,|\, \Theta}(X_\star \,|\, \Theta) \,|\, \overline{x}]}_{\text{see (21)}} \\ &\overset{\text{see (18)}}{=} \quad \mathrm{E}_{\,\Theta \,|\, \overline{X}}(\Theta \,|\, \overline{x}) \overset{\text{see (14)}}{=} \mu_N(\overline{x}) \end{aligned}$$

and variance [where we use the law of conditional variances based on (21)]:

$$\underbrace{\text{var}_{X_\star,\Theta\,|\,\overline{X}}(X_\star\,|\,\overline{x})}_{\text{var}_{X_\star\,|\,\overline{X}}(X_\star\,|\,\overline{x})} \overset{\text{cond. var.}}{=} \mathrm{E}_{\Theta\,|\,\overline{X}}\Big[\underbrace{\text{var}_{X_\star\,|\,\Theta}(X_\star\,|\,\theta)}_{\sigma^2,\ \text{see (18)}}\ \Big|\,\overline{x}\Big]$$

$$+\text{var}_{\Theta\,|\,\overline{X}}\Big[\underbrace{\mathrm{E}_{X_\star\,|\,\Theta}(X_\star\,|\,\theta)}_{\Theta,\ \text{see (18)}}\ \Big|\,\overline{x}\Big]$$

$$\overset{\text{see (14)}}{=} \quad \sigma^2 + \tau_N^2$$

see the probability review in handout # 0b. Therefore

$$f_{X_\star\,|\,\overline{X}}(x_\star\,|\,\overline{x}) = \mathcal{N}(\mu_N(\overline{x}), \sigma^2 + \tau_N^2).$$

Proper vs. Improper Priors

A prior $\pi(\theta)$ is called *proper* if it is a valid probability distribution:

$$\pi(\theta) \geq 0 \quad \forall\theta, \qquad \int \pi(\theta)\, d\theta = 1.$$

A prior $\pi(\theta)$ is called *improper* if

$$\pi(\theta) \geq 0 \quad \forall\theta, \qquad \int \pi(\theta)\, d\theta = +\infty.$$

If a prior is proper, so is the posterior

$$f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) \propto \pi(\theta)\, f_{\boldsymbol{X} \,|\, \boldsymbol{\Theta}}(\boldsymbol{x} \,|\, \boldsymbol{\theta})$$

and everything is fine.

If a prior is improper, the posterior may or may not be proper. For many common problems, popular improper noninformative priors (e.g. Jeffreys' priors, to be discussed later in this handout) lead to proper posteriors, assuming enough data have been collected. But, this has to be checked.

Regarding "propriety," all that we really care about is that the posterior is proper: a valid posterior pdf/pmf is key to Bayesian inference.

Conjugate Prior for the Variance of a Gaussian Distribution with Known Mean

See also Section 2.7 in

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, 2nd ed. New York: Chapman & Hall, 2004.

Data model:

$$f_{X \,|\, \Sigma^2}(x \,|\, \sigma^2) = \mathcal{N}(X \,|\, \mu, \sigma^2)$$

where $\sigma^2 \geq 0$ is now the parameter of interest and μ is a known constant.

Example. We now find the conjugate prior family of pdfs for σ^2 under this model. First, write $f_{X \,|\, \Sigma^2}(x \,|\, \sigma^2)$ explicitly:

$$f_{X \,|\, \Sigma^2}(x \,|\, \sigma^2) = \exp\Big[\underbrace{-\frac{1}{2\,\sigma^2}}_{\eta(\sigma^2)} \; \underbrace{(x-\mu)^2}_{t(x)} \Big] \cdot \underbrace{\frac{1}{\sqrt{2\,\pi\,\sigma^2}} \cdot i_{[0,+\infty)}(\sigma^2)}_{q(\sigma^2)} \, .$$

A conjugate prior family of pdfs for σ^2 follows by using (1):

$$\pi(\theta) \quad \propto \quad \underbrace{(2\,\pi\,\sigma^2)^{-\xi/2}\, i_{[0,+\infty)}(\sigma^2)}_{[q(\sigma^2)]^{\xi}} \; \underbrace{\exp\Big[-\frac{1}{2\,\sigma^2}\,\nu \Big]}_{\exp[\eta(\sigma^2)\,\nu]} \, .$$

How does this pdf *look like*? By looking up the table of distributions we see that it “looks like” (and therefore is) an inverse-gamma pdf:

$$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2) \cdot i_{[0,+\infty)}(\sigma^2)$$

where α and β are *known hyperparameters*. (Note that this distribution is used as a prior distribution for the variance parameter in Example 10.3 of Kay-I.)

For ease of interpretation, use an equivalent prior pdf: a *scaled inverted χ^2 distribution* with scale σ_0^2 and ν_0 degrees of freedom; here σ_0^2 and ν_0 are the known hyperparameters. In other words, we take the prior distribution of σ^2 to be the distribution of

$$\frac{\sigma_0^2\,\nu_0}{X}$$

where X is a $\chi^2_{\nu_0}$ random variable (see the underlined part of the distribution handout). We use the following notation for this distribution:

$$\Sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

or

$$\begin{aligned} f_{\Sigma^2}(\sigma^2) &= \text{Inv-}\chi^2(\sigma^2 \,|\, \nu_0, \sigma_0^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\Big(-\frac{\nu_0\,\sigma_0^2}{2\,\sigma^2}\Big) \cdot i_{[0,+\infty)}(\sigma^2). \end{aligned}$$

Note: From the table of distributions, we also obtain the following facts:

- the mean of $f_{\Sigma^2}(\sigma^2)$ is

$$\mathrm{E}_{\Sigma^2}(\sigma^2) = \frac{\sigma_0^2 \, \nu_0}{\nu_0 - 2} \tag{22}$$

  and

- when ν_0 is large, the variance behaves like $(\sigma_0^2)^2/\nu_0$, implying that large ν_0 yields high precision.

Example: Estimating the Variance of a Gaussian Distribution with Known Mean

For conditionally i.i.d. $X[0], X[1], \ldots, X[n-1]$ given σ^2, the likelihood function is

$$\begin{aligned} f_{\boldsymbol{X} \,|\, \Sigma^2}(\boldsymbol{x} \,|\, \sigma^2) &= (2\,\pi\,\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\,\sigma^2} \sum_{n=0}^{N-1} (x[n]-\mu)^2 \Big] \\ &= (2\,\pi\,\sigma^2)^{-N/2} \exp\Big(-\frac{N\,T(\boldsymbol{x})}{2\,\sigma^2} \Big) \end{aligned}$$

where

$$T(\boldsymbol{x}) \stackrel{\triangle}{=} \frac{1}{N} \sum_{n=0}^{N-1} (x[n]-\mu)^2$$

is the natural sufficient statistic; note that the above likelihood function is in the exponential-family form. Choose the conjugate prior family of pdfs:

$$\begin{aligned} \pi(\sigma^2) &= \text{Inv-}\chi^2(\sigma^2 \,|\, \nu_0, \sigma_0^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\Big(-\frac{\nu_0\,\sigma_0^2}{2\,\sigma^2} \Big) \cdot i_{[0,+\infty)}(\sigma^2) \end{aligned}$$

i.e. scaled inverted χ^2 distribution.

Now,

$$\begin{aligned} f_{\Sigma^2 \,|\, \boldsymbol{X}}(\sigma^2 \,|\, \boldsymbol{x}) &\propto \pi(\sigma^2)\, f_{\boldsymbol{X} \,|\, \sigma^2}(\boldsymbol{x} \,|\, \sigma^2) \\ &\propto (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\Big(- \frac{\nu_0\,\sigma_0^2}{2\,\sigma^2}\Big) \\ &\quad \cdot (\sigma^2)^{-N/2} \cdot \exp\Big(- \frac{N\,T(\boldsymbol{x})}{2\,\sigma^2}\Big) \\ &\propto (\sigma^2)^{-(\frac{\nu_N}{2}+1)} \cdot \exp\Big(- \frac{\nu_N\,\sigma_N^2}{2\,\sigma^2}\Big) \end{aligned}$$

with

$$\nu_N = \nu_0 + N$$

and

$$\sigma_N^2 = \sigma_N^2(\boldsymbol{x}) = \frac{N\,T(\boldsymbol{x}) + \nu_0\,\sigma_0^2}{N + \nu_0}.$$

Therefore, $f_{\Sigma^2 \,|\, \boldsymbol{X}}(\sigma^2 \,|\, \boldsymbol{x})$ is also a scaled inverted χ^2 distribution. Now, the posterior mean (and the MMSE estimate of σ^2, to be shown later) is

$$\mathrm{E}_{\,\Sigma^2 \,|\, \boldsymbol{X}}(\Sigma^2 \,|\, \boldsymbol{X} = \boldsymbol{x}) = \frac{\sigma_N^2\,\nu_N}{\nu_N - 2} = \frac{N\,T(\boldsymbol{x}) + \nu_0\,\sigma_0^2}{N + \nu_0 - 2}$$

obtained by using (22), but now for the posterior pdf.

Comments:

- The MMSE estimate of σ^2 is a weighted average of the prior guess and a data estimate:

$$\mathrm{E}_{\Sigma^2 \,|\, \boldsymbol{X}}(\Sigma^2 \,|\, \boldsymbol{X} = \boldsymbol{x}) = \frac{N\,T(\boldsymbol{x}) + \nu_0\,\sigma_0^2}{N + \nu_0 - 2}$$

where the weights are obtained using the prior and sample degrees of freedom.

- **Interpretation of the prior information:** the chosen prior provides information equivalent to ν_0 observations with average variance equal to σ_0^2.

- As $N \nearrow +\infty$, $\sigma_N^2 \to T(\boldsymbol{x})$ and

$$\mathrm{E}_{\Sigma^2 \,|\, \boldsymbol{X}}(\Sigma^2 \,|\, \boldsymbol{X} = \boldsymbol{x}) \to T(\boldsymbol{x}).$$

- As $\nu_0 \nearrow +\infty$, $\sigma_N^2 \to \sigma_0^2$ and

$$\mathrm{E}_{\Sigma^2 \,|\, \boldsymbol{X}}(\Sigma^2 \,|\, \boldsymbol{X} = \boldsymbol{x}) \to \sigma_0^2.$$

Noninformative Priors

Although it may seem that picking a noninformative prior distribution is easy, (e.g. just use a uniform pdf or pmf), it is not quite that straightforward.

Example: Estimating the Standard Deviation and Variance of a Gaussian Distribution with Known Mean. Consider now N conditionally i.i.d. observations $X[0], X[1], \ldots, X[N-1]$ given the standard deviation $\Sigma = \sigma$, where

$$\begin{aligned} \{X[n] \,|\, \Sigma = \sigma\} &\sim \mathcal{N}(0, \sigma^2) && (23) \\ \pi(\sigma) &\propto i_{[0,\infty)}(\sigma) && (24) \end{aligned}$$

i.e. we assume a uniform prior (from zero to infinity) for the standard deviation σ.

Question: What is the equivalent prior for the variance σ^2?

Reminder: Consider a random variable Θ with pdf $f_\Theta(\theta)$ and a one-to-one transformation

$$\phi = h(\theta).$$

Then, the pdf of ϕ satisfies

$$f_\Phi(\phi) = f_\Theta(\theta) \cdot |d\theta/d\phi| = f_\Theta(\theta) \cdot |h'(\theta)|^{-1}\Big|_{\theta=h^{-1}(\phi)}.$$

We now apply the above change-of-variables formula to our problem: $h(\sigma) = \sigma^2$, $h'(\sigma) = 2\,\sigma$, yielding

$$\pi(\sigma^2) \propto \frac{1}{2\,\sigma}\, i_{[0,\infty)}(\sigma) \tag{25}$$

which is *not uniform*. Therefore, (24) implies (25), which means that our prior belief is that the variance σ^2 is small.

Then, the following prior for the standard deviation σ

$$\pi(\sigma) \propto 2\,\sigma\, i_{[0,\infty)}(\sigma^2)$$

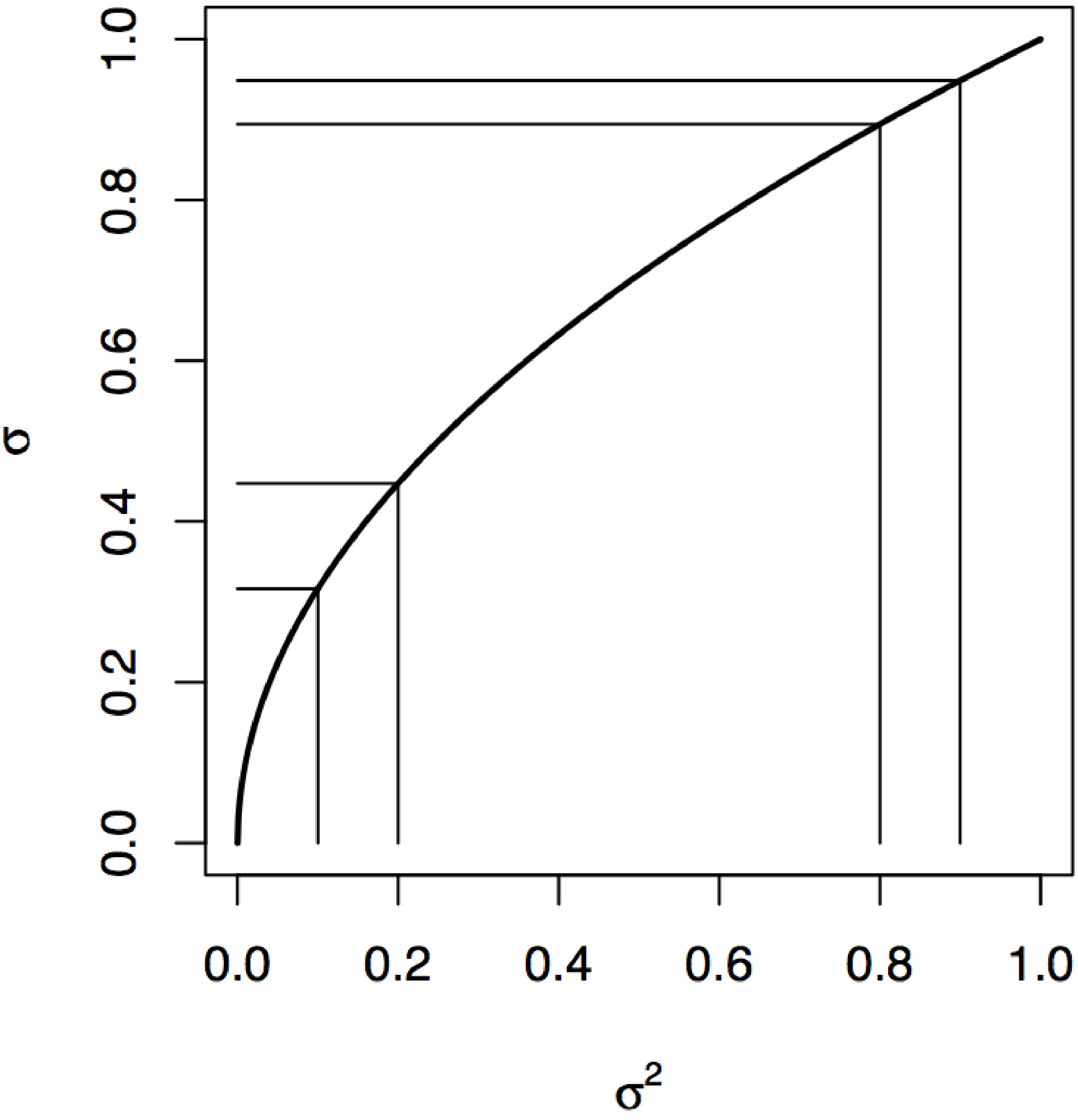
implies that we believe that the standard deviation σ is large, but is equivalent to the uniform prior on the variance σ^2:

$$\pi(\sigma^2) \propto i_{[0,\infty)}(\sigma^2).$$

Problems of this kind are the main reasons why R.A. Fisher, the father of statistics, had a distaste for the Bayesian approach.

One way to visualize the observed phenomenon observed is to look at what happens to intervals of equal measure.

In the case of σ^2 being uniform, an interval $[a, a + 0.1]$ must have the same prior measure as the interval $[0.1, 0.2]$. When we transform to σ, the corresponding prior measure must have intervals $[\sqrt{a}, \sqrt{a + 0.1}]$ having equal measure. But, the length of $[\sqrt{a}, \sqrt{a+0.1}]$ is a decreasing function of a, which agrees with the increasing density in σ.



Therefore, when talking about non-informative priors, we need to think about scale.

Jeffreys' Priors

Can we pick a prior where the scale of the parameter does not matter?

Jeffreys' general principle states that any rule for determining the prior density

$$f_\Theta(\theta)$$

for parameter θ should yield an equivalent result if applied to the transformed parameter ($\phi = h(\theta)$, say, where $h(\cdot)$ is a one-to-one transformation). Therefore, applying the prior

$$f_\Phi(\phi) \quad = \quad \{\pi_\theta(\theta) \cdot |h'(\theta)|^{-1}\}\big|_{\theta=h^{-1}(\phi)}$$

for Φ should give the same answer as dealing directly with the transformed model,

$$f_{\boldsymbol{X},\Phi}(\boldsymbol{x},\phi) = f_\Phi(\phi)\, f_{\boldsymbol{X}\,|\,\Phi}(\boldsymbol{x}\,|\,\phi).$$

Jeffreys' suggestion: choose

$$f_\Theta(\theta) \propto \sqrt{\mathcal{I}(\theta)} \tag{26}$$

where $\mathcal{I}(\theta)$ is the Fisher information for θ. Why is this choice good? If we make a one-to-one transformation $\phi = h(\theta)$, then the Jeffreys' prior for the transformed model is

$$f_\Phi(\phi) \propto \sqrt{\mathcal{I}(\phi)}. \tag{27}$$

Recall the change-of-variables formula for CRB in (12) of handout #2:

$$\mathcal{I}(\phi) \quad = \quad \mathcal{I}(\theta) \cdot \left|\frac{d\theta}{d\phi}\right|^2 \bigg|_{\theta=h^{-1}(\phi)}$$

implying

$$\underbrace{\sqrt{\mathcal{I}(\phi)}}_{\propto f_\Theta(\theta), \quad \text{see (26)}} \quad = \underbrace{\sqrt{\mathcal{I}(\theta)} \cdot \left|\frac{d\theta}{d\phi}\right| \bigg|_{\theta=h^{-1}(\phi)}}_{\propto f_\Phi(\phi), \quad \text{see (27)}}$$

where, recall the Jacobian transformation applied to the prior pdf (or pmf)

$$f_\Phi(\phi) = \{f_\Theta(\theta) \cdot |d\theta/d\phi|\}\big|_{\theta=h^{-1}(\phi)}.$$

Example. Estimating the Variance of a Gaussian Distribution with Known Mean: Recall that the Fisher information for σ^2 is [see eq. (30) in handout # 2]

$$\mathcal{I}(\sigma^2) = \frac{N}{2\,(\sigma^2)^2}.$$

Therefore, the Jeffreys' prior for σ^2 is

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \cdot i_{[0,+\infty)}(\sigma^2). \tag{28}$$

Alternative descriptions under different parameterizations for the variance parameter are (for $\sigma^2 > 0$)

$$\pi(\sigma) \propto \frac{1}{\sigma} i_{[0,+\infty)}(\sigma), \quad \pi(\ln \sigma^2) \propto 1 \quad \text{uniform on } (-\infty, +\infty).$$

Here, $\pi(\ln \sigma^2) \propto 1$ means that

$$\pi_Q(q) \propto 1 \quad \text{uniform on } (-\infty, +\infty)$$

where $Q = \ln \Sigma^2$.

Example. Estimating the Mean of a Gaussian Distribution with Known Variance. Consider N conditionally i.i.d. observations $X[0], X[1], \ldots, X[N-1]$ given $\Theta = \theta$, following

$$\{X[n] \mid \Theta = \theta\} \sim \mathcal{N}(\theta, \sigma^2)$$

where σ^2 is a known constant. Here

$$\mathcal{I}(\theta) = \frac{N}{\sigma^2} = \text{const}$$

and, therefore, the clearly improper Jeffreys' prior for θ is

$$\pi(\theta) \propto 1 \quad \text{uniform on } (-\infty, +\infty).$$

Bayesian Estimation

Suppose that we need to provide a point estimate of the parameter of interest. How do we do that in the Bayesian setting? Here, we first consider the most popular *squared-error loss scenario* and then discuss the general scenario (for an arbitrary loss).

Bayesians construct estimators

$$\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$$

based on the posterior distribution $f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x})$. Hence, a *Bayesian approach* to solving the above problem is, say, to obtain $\widehat{\theta}$ by minimizing a *posterior expected (squared, say) loss*:

$$\begin{aligned} \rho(\widehat{\theta} \,|\, \boldsymbol{x}) &= \mathrm{E}_{\Theta \,|\, \boldsymbol{X}}\{[\widehat{\theta}(\boldsymbol{X}) - \Theta]^2 \,|\, \boldsymbol{x}\} \\ &= \int \underbrace{[\widehat{\theta}(\boldsymbol{x}) - \theta]^2}_{\text{squared-error loss}} \; f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) \, d\theta \end{aligned}$$

with respect to $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$. This is easy to do: decompose

$\rho(\widehat{\theta}\,|\,\boldsymbol{x})$ as

$$\begin{aligned}
\rho(\widehat{\theta}\,|\,\boldsymbol{x}) = \int [\widehat{\theta} - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x}) + \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x}) - \theta]^2 \, f_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{x})\, d\theta \\
= \quad & \int \left[\widehat{\theta} - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x})\right]^2 f_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{x})\, d\theta \\
& + \int \left[\theta - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{X} = \boldsymbol{x})\right]^2 f_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{x})\, d\theta \\
& + 2\left[\widehat{\theta} - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x})\right] \\
& \cdot \underbrace{\int \left[\theta - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{X} = \boldsymbol{x})\right] f_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{x})\, d\theta}_{\mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x}) - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x}) = 0}
\end{aligned}$$

and the optimal $\widehat{\theta}$ follows by minimizing the first term (which has a quadratic form, hence the minimization is trivial):

$$\arg\min_{\widehat{\theta}} \rho(\widehat{\theta}\,|\,\boldsymbol{x}) = \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x}).$$

Clearly, the posterior mean of the parameter θ minimizes its posterior expected squared loss; the minimum posterior expected squared loss is

$$\min_{\widehat{\theta}} \rho(\widehat{\theta}\,|\,\boldsymbol{x}) = \int \left[\theta - \mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\Theta\,|\,\boldsymbol{x})\right]^2 f_{\Theta\,|\,\boldsymbol{X}}(\theta\,|\,\boldsymbol{x})\, d\theta.$$

Mean-square error measures.

1. **Classical:**

$$\begin{aligned}\mathrm{MSE}\{\widehat{\theta}\} &= \mathrm{E}_{\boldsymbol{X}\,|\,\Theta}([\widehat{\theta}(\boldsymbol{X}) - \theta]^2 \,|\, \Theta = \theta) \\ &= \int [\widehat{\theta}(\boldsymbol{x}) - \theta]^2 \, f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\theta)\, d\boldsymbol{x}\end{aligned}$$

see also (4) in handout # 1.

2. **"Bayesian" (preposterior) MSE:**

$$\begin{aligned}\mathrm{BMSE}\{\widehat{\theta}\} &= \mathrm{E}_{\boldsymbol{X},\Theta}([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2) \\ &= \int\int [\widehat{\theta}(\boldsymbol{x}) - \theta]^2 \, f_{X\,|\,\Theta}(x\,|\,\theta)\, \pi(\theta)\, dx\, d\theta \\ &\overset{\text{iter. exp.}}{=} \mathrm{E}_{\Theta}[\mathrm{MSE}\{\widehat{\theta}\}]\end{aligned}$$

see also (2) in handout # 0. The preposterior MSE (BMSE) is obtained by averaging the squared-error loss over both the noise *and* parameter realizations. It is computable before the data has been collected, hence the name preposterior.

- The classical MSE generally depends on the true value of the parameter θ. Therefore, classical MMSE "estimates" usually depend on θ; hence, classical MMSE estimates do not exist.

- Since Θ is integrated out, the preposterior BMSE *does not* depend on θ; hence, Bayesian MMSE estimates exist.

Which $\widehat{\theta}$ minimizes BMSE? Since

$$\begin{aligned} \mathrm{BMSE}\{\widehat{\theta}\} &= \mathrm{E}_{\boldsymbol{X},\Theta}([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2) \\ &= \mathrm{E}_{\boldsymbol{X}}\{\underbrace{\mathrm{E}_{\Theta \,|\, \boldsymbol{X}}([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2 \,|\, \boldsymbol{X})}_{\rho(\widehat{\theta} \,|\, \boldsymbol{X})}\} \end{aligned}$$

and, for every given $\boldsymbol{x}$, we know that

$$\widehat{\theta} = \mathrm{E}_{\Theta \,|\, \boldsymbol{X}}(\Theta \,|\, \boldsymbol{x}) \quad \text{posterior mean of } \Theta$$

minimizes the posterior expected squared loss $\rho(\widehat{\theta} \,|\, \boldsymbol{x}) = \mathrm{E}_{\Theta \,|\, \boldsymbol{X}}\{[\widehat{\theta}(\boldsymbol{X}) - \Theta]^2 \,|\, \boldsymbol{x}\}$.

Loss (Cost) Functions for Bayesian Estimation and the Corresponding Optimal Estimators

Define the estimation error

$$\epsilon = \epsilon(\boldsymbol{x}, \theta) = \widehat{\theta}(\boldsymbol{x}) - \theta$$

and assign a loss (cost) function $\mathrm{L}(\epsilon)$. We may choose $\widehat{\theta}(\boldsymbol{x})$ to minimize the *preposterior risk*:

$$\mathrm{E}_{\boldsymbol{X},\Theta}[\mathrm{L}(\epsilon)] = \mathrm{E}_{\boldsymbol{X},\Theta}[\mathrm{L}(\widehat{\theta}(\boldsymbol{X}) - \Theta)]$$

but this is equivalent to minimizing the *posterior expected loss*:

$$\rho(\widehat{\theta} \,|\, \boldsymbol{x}) = \mathrm{E}_{\Theta \,|\, \boldsymbol{X}}[\mathrm{L}(\epsilon) \,|\, \boldsymbol{x}] = \int \mathrm{L}(\widehat{\theta}(\boldsymbol{x}) - \theta)\, f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x})\, d\theta$$

for each $\boldsymbol{X} = \boldsymbol{x}$, which is a Bayesian criterion. The proof is the same as for the squared-error loss:

$$\mathrm{E}_{\boldsymbol{X},\Theta}[\mathrm{L}(\widehat{\theta}(\boldsymbol{X}) - \Theta)] \overset{\text{iter. exp.}}{=} \mathrm{E}_{\boldsymbol{X}}\{\underbrace{\mathrm{E}_{\Theta \,|\, \boldsymbol{X}}[\mathrm{L}(\widehat{\theta}(\boldsymbol{X}) - \Theta) \,|\, \boldsymbol{X}]}_{\rho(\widehat{\theta}(\boldsymbol{X}) \,|\, \boldsymbol{X})}\}. \tag{29}$$

Here are a few popular loss functions:

1. $\mathrm{L}(\epsilon) = \epsilon^2$ (squared-error loss, accurate, most popular),

2. $\mathrm{L}(\epsilon) = |\epsilon|$ (absolute-error loss, robust to outliers),

3. $\mathrm{L}(\epsilon) = \begin{cases} 0, & |\epsilon| \le \Delta/2 \\ 1, & |\epsilon| > \Delta/2 \end{cases}$ (0-1 loss, tractable)

corresponding to:

1. **MMSE estimator:**

$$\widehat{\theta} = \widehat{\theta}(\boldsymbol{x}) = \mathrm{E}_{\Theta \,|\, \boldsymbol{X}}[\Theta \,|\, \boldsymbol{X} = \boldsymbol{x}]$$

   the *posterior mean of θ given $\boldsymbol{X} = \boldsymbol{x}$* (as proved earlier in this handout).

2. **Posterior median,** i.e. the optimal $\widehat{\theta}$ satisfies:

$$\int_{-\infty}^{\widehat{\theta}} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x})\, d\theta = \int_{\widehat{\theta}}^{+\infty} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x})\, d\theta$$

   HW: check this.

3. **MAP (maximum *a posteriori*) estimator,** i.e. the optimal $\widehat{\theta}_{\rm MAP} = \widehat{\theta}_{\rm MAP}(\boldsymbol{x})$ satisfies:

$$\widehat{\theta}_{\rm MAP}(\boldsymbol{x}) = \arg\max_{\theta} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) \tag{30}$$

also known as the *posterior mode*.

We now show the MAP estimator result in 3.

MAP Estimation. Start from (29):

$$\mathrm{E}_{\,X}\{\mathrm{E}_{\,\Theta \,|\, X}[\mathrm{L}(\epsilon)]\} = \mathrm{E}_{\,X}\Big\{1 - \int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, x)\, d\theta\Big\}.$$

To minimize this expression with respect to $\widehat{\theta}$, we maximize

$$\int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, x)\, d\theta$$

with respect to $\widehat{\theta}$ which, for small Δ, reduces to maximizing

$$f_{\Theta \,|\, \boldsymbol{X}}(\widehat{\theta} \,|\, x)$$

with respect to $\widehat{\theta}$ and (30) follows. Since

$$f_{\Theta \,|\, \boldsymbol{X}}(\theta \,|\, \boldsymbol{x}) \propto f_{\boldsymbol{X} \,|\, \Theta}(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)$$

we have

$$\widehat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta}[\ln f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\theta) + \ln \pi(\theta)].$$

Note that $\ln f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\theta)$ is the log-likelihood function of θ. Thus, for a *flat prior*

$$\pi(\theta) \propto 1$$

the MAP estimator *coincides with* the classical maximum-likelihood (ML) estimator.