5 Bayesian Statistics

5.1 Bayesian Model

In a Bayesian model, our parameter θ is also a random variable. Hence, $\theta \sim \Lambda$, where the probability measure Λ is called the prior distribution of θ. The parameter for the prior distribution is called a hyperparameter. Our data are generated from the conditional distribution $X \mid \theta \sim \mathrm{P}_\theta$. That is, P_θ is the conditional distribution of X given θ, and the density $p(x \mid \theta)$ is viewed as a conditional density of x given θ.

The Bayesian statistics is based on the Bayes formula

1. If A and E are two events, then

$$\mathrm{P}(A \mid E) = \frac{\mathrm{P}(E \mid A)\,\mathrm{P}(A)}{\mathrm{P}(E)} = \frac{\mathrm{P}(E \mid A)\,\mathrm{P}(A)}{\mathrm{P}(E \mid A)\,\mathrm{P}(A) + \mathrm{P}(E \mid A^c)\,\mathrm{P}(A^c)}.$$

2. If X and Y are two random variables, then

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)} = \frac{p(x \mid y)\,p(y)}{\int p(x \mid y)\,p(y)\,d\mu(y)},$$

where $p()$ is used as the generic symbol.

Bayesian reasoning:

1. The prior summarizes our prior information about θ.
   - From similar experiences, the average number of accidents at a crossing is 1 per 30 days. We assume an exponential distribution with density
   $$\lambda(\theta) = 30\exp(-30\theta), \quad [\text{day}]^{-1}.$$
2. We collect our own data and result in an observation x. We assume it follows Poisson distribution.
   - Three accidents have been recorded after monitoring the roundabout for one year. The likelihood is
   $$p(x \mid \theta) = \frac{(365\theta)^x}{x!}\exp(-365\theta),$$
   where $x = 3$.
3. We use the information in x to update our knowledge on θ.
   - By Bayes' formula
   $$\pi(\theta \mid x) = \frac{p(x \mid \theta)\,\lambda(\theta)}{m(x)} = \frac{p(x \mid \theta)\,\lambda(\theta)}{\int p(x \mid t)\,\lambda(t)\,d\nu(t)}.$$

In a Bayesian model, we will have many distributions

- prior distribution with density $\lambda(\theta)$.
- conditional distribution $X \mid \theta$ (likelihood) with density $p(x \mid \theta)$.
- joint distribution of (θ, X) with density $p(x, \theta) = p(x \mid \theta)\,\lambda(\theta)$.
- posterior distribution with density $\lambda(\theta \mid x)$.
- marginal distribution of X with density $m(x) = \int f(x \mid \theta)\,\lambda(\theta)\,d\nu(\theta)$.

Example 1. Find the posterior distribution.

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}(\theta)$, $i = 1, ..., n$. The prior is $\theta \sim \text{Beta}(a_0, b_0)$. The posterior satisfies

$$\begin{aligned} \lambda(\theta \mid x) &\propto \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{n-x_i} \cdot \frac{1}{B(a_0, b_0)} \theta^{a_0-1} (1-\theta)^{b_0-1} \\ &\propto \theta^{a_0+\sum_{i=1}^n x_i - 1} (1-\theta)^{b_0+n-\sum_{i=1}^n x_i - 1} \end{aligned}$$

   The posterior is $\text{Beta}(a_0 + \sum_{i=1}^n x_i, b_0 + n - \sum_{i=1}^n x_i)$.

2. Suppose that we have an iid sample $X_i \mid \mu \sim N(\mu, \sigma^2)$, $i = 1, ..., n$, where σ^2 is known. The prior is $\mu \sim N(\mu_0, \sigma_0^2)$. The posterior satisfies

$$\begin{aligned} \lambda(\theta \mid x) &\propto \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\}\right] \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\ &\propto \exp\left\{-\frac{-2\theta\sum_{i=1}^n x_i + n\theta^2}{2\sigma^2} - \frac{\theta^2 - 2\mu_0\theta}{2\sigma_0^2}\right\} \\ &= \exp\left\{-\frac{\left(n\sigma_0^2+\sigma^2\right)\theta^2 - 2\left(\sigma_0^2\sum_{i=1}^n x_i + \sigma^2\mu_0\right)\theta}{2\sigma^2\sigma_0^2}\right\} \\ &\propto \exp\left\{-\frac{\left(\theta - \frac{\sigma_0^2\sum_{i=1}^n x_i + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}\right)^2}{2\sigma^2\sigma_0^2/\left(n\sigma_0^2+\sigma^2\right)}\right\} \\ &\sim N\left(\frac{\sigma_0^2\sum_{i=1}^n x_i + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \quad \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\right). \end{aligned}$$

Claim 1 (Bayesian Inference Principle). Information on the underlying parameter θ is entirely contained in the posterior distribution $\Lambda(\theta \mid x)$. That is, all statistical inference are based on the posterior distribution $\Lambda(\theta \mid x)$.

Some examples are

1. posterior mean: $\text{E}[\theta \mid x]$.

2. posterior mode (MAP): θ that maximizes $\lambda(\theta \mid x)$.

3. predictive distribution of a new observation:

$$p(y \mid x) = \int p(y \mid x, \theta)\, \lambda(\theta \mid x)\, d\nu(\theta).$$

5.2 Prior Distribution

We generally have two types of priors

1. A subjective prior incorporates our prior knowledge, such as expert advise or similar experiences

2. An objective prior fulfills some desired (theoretical) properties.

5.2.1 Conjugate Prior

Definition 1 (Conjugate Prior)**.** Let $\mathcal{F}$ be a family of probability distributions on Θ. If $\Lambda(\cdot) \in \mathcal{F}$ and $\Lambda(\cdot \mid x) \in \mathcal{F}$ for every x, then the family of distributions $\mathcal{F}$ is conjugate. The prior distribution that is an element in a conjugate family is called a conjugate prior.

Main benefit of a conjugate prior: tractability (we only need to update the hyperparameters without changing the family of distributions).

Example 2. Conjugate prior example:

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}(\theta)$. When the prior is $\theta \sim \text{Beta}(a_0, b_0)$, the posterior is $\text{Beta}(a_0 + \sum_{i=1}^n x_i, b_0 + n - \sum_{i=1}^n x_i)$.

2. Suppose that we have an iid sample $X_i \mid \mu \sim N(\mu, \sigma^2)$, $i = 1, ..., n$, where σ^2 is known. When the prior is $\mu \sim N(\mu_0, \sigma_0^2)$, the posterior is $N\left(\frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2}, \quad \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)$.

5.2.2 Uniform Prior

Definition 2 (Uniform Prior or Flat Prior)**.** The Laplace prior is $\lambda(\theta)$ is a constant over Θ.

Disambiguation: The name Laplace prior is often referred to as that θ follows a Laplace distribution

$$\pi(\theta) \quad = \quad \frac{1}{2b_0} \exp\left(-\frac{|\theta - a_0|}{b_0}\right), \quad -\infty < \theta < \infty.$$

The ambiguation is caused by that the uniform prior is proposed by Laplace.

- Intuitively speaking, a constant $\lambda(\theta)$ means that we treat all θ equally. Hence, it is often viewed as a non-informative prior.

- The uniform prior often yields

$$\int_\Theta \pi(\theta)\, d\nu(\theta) \quad = \quad \infty.$$

  Such prior distribution is said to be an improper prior. As long as the posterior distribution is proper, we can still use the improper prior. But the posterior, even exists, may not follow the rules of probability, e.g., marginalization paradox says that the marginal distribution cannot be recovered from the joint distribution and the conditional distribution.

- Another issue of the uniform prior is that it is not invariant against reparametrization.

  – Suppose that we choose a prior for $\theta \in \Theta$.

  – Now we reparameterize to $\eta = \eta(\theta)$ such that $\theta = h(\eta)$. Then,

$$\lambda_\eta(\eta) \quad = \quad \lambda_\theta(h(\eta)) \left| \det\left[\frac{\partial h(\eta)}{\partial \eta^T}\right] \right|,$$

    which is not a constant.

  – A constant prior on θ does not always yield a constant prior on $\eta(\theta)$, even though η is a strictly monotone transformation.

5.2.3 Jeffreys Prior

Definition 3 (Jeffreys Prior)**.** Consider a statistical model with density $f(x \mid \theta)$ and Fisher information matrix $\mathcal{I}(\theta)$. The Jeffreys prior is

$$\lambda(\theta) \quad \propto \quad [\det(\mathcal{I}(\theta))]^{1/2}.$$

Proposition 1. *The Jeffreys prior is invariant to reparametrization under smooth monotone transformation, that is*

$$\lambda_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| \quad = \quad \lambda_\eta(\eta).$$

Proof. With reparametrization, we have the relation between the Fisher information matrices

$$\begin{aligned} \mathcal{I}_\eta(\eta) &= \left[\frac{d\theta}{d\eta^T} \right]^T \mathcal{I}_\theta(\theta) \frac{dh(\eta)}{d\eta^T} \\ \mathcal{I}_\theta(\theta) &= \left[\frac{d\eta}{d\theta^T} \right]^T \mathcal{I}_\eta(\eta) \frac{d\eta}{d\theta^T}. \end{aligned}$$

Invariance means that the prior that we obtained from applying the procedure to η is the same as the prior that we obtained from applying the procedure to θ and make the change of variables. Let the procedure be $\lambda(\theta) \propto [\det(\mathcal{I}(\theta))]^{1/2}$. Then,

$$\begin{aligned} \lambda_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| &= [\det(\mathcal{I}_\theta(\theta))]^{1/2} \left| \det\left[\frac{\partial h(\eta)}{\partial \eta^T} \right] \right| \\ &= \left[\det\left(\left[\frac{d\eta}{d\theta^T} \right]^T \mathcal{I}_\eta(\eta) \frac{d\eta}{d\theta^T} \right) \right]^{1/2} \left| \det\left[\frac{\partial h(\eta)}{\partial \eta^T} \right] \right| \\ &= [\det(\mathcal{I}_\eta(\eta))]^{1/2}, \\ \lambda_\eta(\eta) &= [\det(\mathcal{I}_\eta(\eta))]^{1/2} \text{ the procedure yields this.} \end{aligned}$$

They are the same. Hence, Jeffreys prior is invariant. □

Example 3. Suppose that $X_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$, $i = 1, ..., n$. The likelihood is

$$p(x \mid \theta) \quad = \quad \exp\left\{ -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 \right\}.$$

Then,

$$\begin{aligned} \frac{\partial \log p(x \mid \theta)}{\partial \theta} &= \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2 \end{bmatrix}, \\ \frac{\partial^2 \log p(x \mid \theta)}{\partial \theta \partial \theta^T} &= \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{(\sigma^2)^2}\sum_{i=1}^n (x_i - \mu) \\ -\frac{1}{(\sigma^2)^2}\sum_{i=1}^n (x_i - \mu) & \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^n (x_i - \mu)^2 \end{bmatrix} \\ &\Downarrow \\ \mathcal{I}(\theta) &= -\mathrm{E}\left[\frac{\partial^2 \log p(x \mid \theta)}{\partial \theta \partial \theta^T} \mid \theta \right] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}. \end{aligned}$$

Hence, the Jeffreys prior is

$$\lambda(\theta) \quad \propto \quad [\det(\mathcal{I}(\theta))]^{1/2} = \left[\frac{n^2}{2\sigma^6} \right]^{1/2} \propto \frac{1}{\sigma^3}$$

still an improper prior.

When we have multiple parameters, it is also common to use the independent Jeffreys prior:

1. Obtain the Jeffeys prior for each parameter separately by fixing the others.
2. Multiple the single parameter Jeffreys prior together.

Example 4. Suppose that $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Treating σ^2 as known, we obtain

$$\begin{aligned} \frac{\partial^2 \log p\left(x \mid \theta\right)}{\partial \mu^2} &= -\frac{n}{\sigma^2}, \\ \mathcal{I}\left(\mu\right) &= -\mathrm{E}\left[\frac{\partial^2 \log p\left(x \mid \theta\right)}{\partial \mu^2} \mid \theta\right] = \frac{n}{\sigma^2} \propto 1, \\ \lambda\left(\mu\right) &= 1. \end{aligned}$$

Treating μ as known, we obtain

$$\begin{aligned} \frac{\partial^2 \log p\left(x \mid \theta\right)}{\partial \left(\sigma^2\right)^2} &= \frac{n}{2\left(\sigma^2\right)^2} - \frac{1}{\left(\sigma^2\right)^3} \sum_{i=1}^{n} \left(x_i - \mu\right)^2, \\ \mathcal{I}\left(\sigma^2\right) &= -\mathrm{E}\left[\frac{\partial^2 \log p\left(x \mid \theta\right)}{\partial \left(\sigma^2\right)^2} \mid \theta\right] = \frac{n}{2\sigma^4}, \\ \lambda\left(\sigma^2\right) &= \sqrt{\frac{n}{2\sigma^4}} \propto \sigma^{-2}. \end{aligned}$$

Hence, the independent Jeffreys prior is

$$\lambda\left(\theta\right) \propto \lambda\left(\mu\right)\lambda\left(\sigma^2\right) \propto \frac{1}{\sigma^2},$$

still an improper prior.

5.2.4 Reference Prior

Suppose that there are two probability measure P and Q with respective densities p and q. The Kullback-Leibler divergence is

$$\mathrm{KL}\left(p, q\right) = \mathrm{E}\left[\log\left(\frac{p}{q}\right)\right],$$

where the expectation is taken under P-probability.The Kullback-Leibler divergence between $\lambda\left(\theta \mid x\right)$ and $\lambda\left(\theta\right)$ is

$$\begin{aligned} \mathrm{KL}\left(\lambda\left(\theta \mid x\right), \lambda\left(\theta\right)\right) &= \int \lambda\left(\theta \mid x\right) \log\left(\frac{\lambda\left(\theta \mid x\right)}{\lambda\left(\theta\right)}\right) d\nu\left(\theta\right) \\ &= \int \lambda\left(\theta \mid x\right) \log\left(\frac{p\left(x, \theta\right)}{\lambda\left(\theta\right) m\left(x\right)}\right) d\nu\left(\theta\right), \end{aligned}$$

which is a function of x. In probability theory, the mutual information of two random variables X and Y is defined as

$$\mathrm{MI}\left(X, Y\right) = \mathrm{E}\left[\log\left(\frac{p\left(x, y\right)}{p\left(x\right)p\left(y\right)}\right)\right] = \mathrm{KL}\left(p\left(x, y\right), p\left(x\right)p\left(y\right)\right).$$

It is a measure to quantity the information in $p\left(x, y\right)$ instead of $p\left(x\right)p\left(y\right)$.

- The expected Kullback-Leibler divergence between $\lambda\left(\theta \mid x\right)$ and $\lambda\left(\theta\right)$, denoted by $\mathrm{E}\left[\mathrm{KL}\left(\lambda\left(\theta \mid x\right), \lambda\left(\theta\right)\right)\right]$, is the mutual information of $\theta \mid x$ and θ.

Definition 4. The reference prior maximizes the mutual information of the prior and posterior.

Suppose that some regularity conditions are satisfied, including Θ is a compact set, the Fisher information equals to the negative expected Hessian, among others. It has been proved that, asymptotically ($n \to \infty$), the joint Jeffreys prior maximizes the mutual information.

5.3 Bayesian Estimation

5.3.1 MAP

Definition 5. The maximum a posteriori (MAP) estimator is the mode of the posterior $\lambda(\theta \mid x)$ as

$$\hat{\theta}_{\text{MAP}}(x) = \arg\sup_{\theta} \lambda(\theta \mid x).$$

The MLE satisfies

$$\hat{\theta}_{\text{MLE}}(x) = \arg\sup_{\theta} p(x \mid \theta).$$

Hence, MAP is a penalized MLE since

$$\begin{aligned}\hat{\theta}_{\text{MAP}}(x) &= \arg\sup_{\theta} \log \lambda(\theta \mid x) = \arg\sup_{\theta} \left[\log p(x \mid \theta) + \log \lambda(\theta)\right], \\ \hat{\theta}_{\text{MLE}}(x) &= \arg\sup_{\theta} \log p(x \mid \theta).\end{aligned}$$

Note that $\lambda(\theta \mid x) \propto p(x \mid \theta)\lambda(\theta)$ and $m(x)$ does not include any θ.

- The MAP estimator only requires the kernel of $p(x \mid \theta)\lambda(\theta)$.
- We can skip the integration step to obtain $m(x)$.

Example 5. Let $X_1, ..., X_n$ be iid $N(0, \sigma^2)$. The parameter of interest is $\theta = \sigma^{-2}$. We assume that the prior of θ is Gamma(a, b). The posterior is

$$\begin{aligned}\lambda(\theta \mid x) &\propto \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta^{-1}}} \exp\left\{-\frac{x_i^2}{2\theta^{-1}}\right\}\right] \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) \\ &\propto \theta^{a+n/2-1} \exp\left\{-\left(b + \frac{\sum_{i=1}^{n} x_i^2}{2}\right)\theta\right\}.\end{aligned}$$

We only need to maximize the kernel as

$$\log \lambda(\theta \mid x) = \text{constant} + \left(a + \frac{n}{2} - 1\right)\log\theta - \left(b + \frac{\sum_{i=1}^{n} x_i^2}{2}\right)\theta.$$

Hence, the MAP estimator is

$$\theta = \frac{a + 2^{-1}n - 1}{b + 2^{-1}\sum_{i=1}^{n} x_i^2} = \frac{2a + n - 2}{2b + \sum_{i=1}^{n} x_i^2}.$$

Remark 1. Suppose that the posterior is $\lambda(\theta, \tau \mid x)$, where θ is the parameter of interest. The mode of $\lambda(\theta, \tau \mid x)$ may not equal the marginal posterior mode, the mode of $\lambda(\theta \mid x)$.

Example 6. Consider the normal-inverse-gamma model, where the posterior is

$$\mu \mid \sigma^2, x \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \qquad \sigma^2 \mid x \sim \text{InvGamma}(a_n, b_n),$$

where μ_n, a_n, and b_n are known constants. The posterior is

$$\pi(\theta \mid x) \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \cdot \text{InvGamma}(a_n, b_n),$$

where

$$\begin{aligned}
\mu_n &= \frac{\mu_0\lambda_0 + \sum_{i=1}^n x_i}{\lambda_0 + n}, \\
a_n &= a_0 + \frac{n}{2}, \\
b_n &= b_0 + \frac{1}{2}\left[\sum_{i=1}^n x_i^2 + \lambda_0\mu_0^2 - \frac{\left(\mu_0\lambda_0 + \sum_{i=1}^n x_i\right)^2}{\lambda_0 + n}\right].
\end{aligned}$$

Then,

$$\log \pi(\theta \mid x) = \text{Constant} - \frac{1}{2}\log\left(\sigma^2\right) - \frac{(\lambda_0 + n)(\mu - \mu_n)^2}{2\sigma^2} - (a_n + 1)\log\left(\sigma^2\right) - \frac{b_n}{\sigma^2}.$$

1. The joint mode is obtained from

$$\frac{\partial \log \pi(\theta \mid x)}{\partial \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix}} = \begin{bmatrix} -\frac{(\lambda_0+n)(\mu-\mu_n)}{\sigma^2} \\ -\frac{2a_n+3}{2\sigma^2} + \frac{(\lambda_0+n)(\mu-\mu_n)^2+2b_n}{2(\sigma^2)^2}\end{bmatrix}.$$

   Then,

$$\mu = \mu_n \quad \text{and} \quad \sigma^2 = \frac{2b_n}{2a_n + 3}.$$

2. The marginal of σ^2 is InvGamma(a_n, b_n). Hence, the univariate mode is

$$\hat{\sigma}^2 = \frac{b_n}{a_n + 1}$$

   not the same as the joint mode.

Recall that MLE is invariant with respect to reparametrization. However, MAP is not invariant with respect to reparametrization, because of the Jacobian.

Example 7. Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}(n, \theta)$. Let the prior be $\theta \sim \text{Beta}(a_0, b_0)$, where $a_0 > 1$ and $b_0 > 1$. The posterior is Beta(a_n, b_n), where $a_n = a_0 + x$ and $b_n = b_0 + n - x$. We are interested in the odds $\theta/(1-\theta)$.

1. The mode of this Beta distribution is

$$\frac{a_n - 1}{a_n + b_n - 2}$$

   If we plug it into the odds, we obtain

$$\frac{\hat{\theta}}{1-\hat{\theta}} = \frac{\frac{a_n-1}{a_n+b_n-2}}{1 - \frac{a_n-1}{a_n+b_n-2}} = \frac{a_n - 1}{b_n - 1}.$$

2. When we consider the odds directly, we do a change of variables, then

$$\begin{aligned}
\lambda(\eta \mid x) &\propto \theta^{a_n-1}(1-\theta)^{b_n-1}\left|\frac{d\theta}{d\eta}\right| \\
&= \left(\frac{\eta}{1+\eta}\right)^{a_n-1}\left(\frac{1}{1+\eta}\right)^{b_n-1}\frac{1}{(1+\eta)^2} \\
&= \exp\left\{(a_n - 1)\log\eta - (a_n + b_n)\log(1+\eta)\right\}.
\end{aligned}$$

   The MAP is

$$\hat{\eta} = \frac{a_n - 1}{b_n + 1}.$$

5.3.2 Posterior Mean

An alternative to MAP is the posterior mean

$$\hat{\theta}_{\text{Mean}} = \mathrm{E}\left[\theta \mid x\right].$$

Consider the weighted L_2 loss between θ and an estimator d:

$$L_W\left(\theta, d\right) = \left(\theta - d\right)^T W \left(\theta - d\right),$$

where W is a $p \times p$ positive definite matrix and d is an estimator of θ using x.

Theorem 1. *Suppose that there exists an estimator d such that $E\left[L_W\left(\theta, d\right) \mid x\right] < \infty$. Then, the posterior mean minimizes $E\left[L_W\left(\theta, d\right) \mid x\right]$, where W does not depend on θ but can depend on x.*

Proof. Let $\delta_B = \mathrm{E}\left[\theta \mid X = x\right]$. Then, for any estimator $d = d\left(x\right)$, we have

$$\begin{aligned}
\mathrm{E}\left[\left(\theta - d\right)^T W \left(\theta - d\right) \mid x\right] &= \mathrm{E}\left[\left(\theta - \delta_B + \delta_B - d\right)^T W \left(\theta - \delta_B + \delta_B - d\right) \mid x\right] \\
&= \mathrm{E}\left[\left(\theta - \delta_B\right)^T W \left(\theta - \delta_B\right) \mid x\right] + 2\left(\delta_B - d\right)^T W \underbrace{\mathrm{E}\left[\left(\theta - \delta_B\right) \mid x\right]}_{=0} \\
&\quad + \mathrm{E}\left[\underbrace{\left(\delta_B - d\right)^T W \left(\delta_B - d\right)}_{\geq 0} \mid x\right] \\
&\geq \mathrm{E}\left[\left(\theta - \delta_B\right)^T W \left(\theta - \delta_B\right) \mid x\right].
\end{aligned}$$

□

Remark 2. Comparison of MAP and posterior mean

1. The marginal posterior mean is the same as the joint posterior mean. But marginal MAP is not necessarily the same as joint MAP.
2. To obtain the closed form expression of $\mathrm{E}\left[\theta \mid x\right]$, we need the normalizing constant of $\pi\left(\theta \mid x\right)$. MAP does not require the normalizing constant.
3. The posterior mean is not invariant with respect to reparametrization either.
4. It can even happen that the posterior mean does not exist, even though the posterior is proper.

Example 8 (Non-Existence of Posterior Mean)**.** Let X_1, ..., X_n be iid from a two parameter Weibull distribution

$$f\left(x \mid \theta, \beta\right) = \frac{\beta x^{\beta-1}}{\theta^{\beta}} \exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\}, \quad x > 0,\ \theta > 0,\ \beta > 0.$$

Consider the proper priors

$$\begin{aligned}
\pi\left(\theta \mid \beta\right) &= \frac{\beta b_0^{a_0}}{\Gamma\left(a_0\right)} \frac{1}{\theta^{a_0\beta+1}} \exp\left(-\frac{b_0}{\theta^{\beta}}\right), \text{ "InvGamma" prior} \\
\pi\left(\beta\right) &= \frac{d_0^{c_0}}{\Gamma\left(c_0\right)} \beta^{c_0-1} \exp\left(-d_0\beta\right). \text{ Gamma prior}
\end{aligned}$$

With probability 1, the posterior mean of θ^k does not exist for any $k > 0$.

5.3.3 Predictive Distribution

In frequentist statistics, the prediction of a new observation z after observing x is

$$\hat{z}(x) = \int zp\left(z \mid x, \hat{\theta}\right) d\mu(z).$$

In Bayesian statistics, the predictive distribution of a new observation z after observing x is

$$p(z \mid x) = \int p(z \mid x, \theta) \lambda(\theta \mid x) d\nu(\theta).$$

A point predictor can be the expectation of the predictive distribution.

Example 9. Suppose that we have an iid sample $X_i \mid \mu \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$, where σ^2 is known. The prior is $\mu \sim N\left(\mu_0, \sigma_0^2\right)$. We have shown that the posterior is $\mu \mid x \sim N\left(\mu_n \quad \sigma_n^2\right)$, where

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}, \qquad \mu_n = \frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2}.$$

The predictive distribution is

$$f(x_0 \mid x) = \int p(x_0 \mid x, \mu) \lambda(\mu \mid x) d\nu(\mu).$$

Under the iid assumption, $p(x_0 \mid x, \mu) = p(x_0 \mid \mu)$ is normal density $N\left(\mu, \sigma^2\right)$. Hence,

$$f(x_0 \mid x) = \int \underbrace{p(x_0 \mid x, \mu)}_{N(\mu, \sigma^2)} \underbrace{\lambda(\mu \mid x)}_{N(\mu_n, \sigma_n^2)} d\mu \sim N\left(x_0 \mu_n, \ \sigma^2 + \sigma_n^2 x_0^2\right).$$

5.4 Bayesian Test

We can formulate the hypothesis testing as a decision problem. Suppose that the loss of the wrong decision is

	Truth	
Decision	H_0	H_1
H_0	0	a_1
H_1	a_0	0

such that $a_0 + a_1 > 0$.

Theorem 2. *The optimal Bayes test that minimizes the posterior expected loss $E[L \mid x]$ and the expected loss $E[L]$ is*

$$\phi(x) = \begin{cases} 1, & \text{if } P(H_0 \mid x) < \frac{a_1}{a_0 + a_1}, \\ 0, & \text{if } P(H_0 \mid x) \geq \frac{a_1}{a_0 + a_1}. \end{cases}$$

Proof. For any test ϕ^*, the posterior risk for a given x is

$$\begin{aligned} \mathrm{E}\left[L\left(\phi^*\right) \mid x\right] &= \underbrace{a_0 1\left(\phi^* = 1\right) \mathrm{P}\left(H_0 \mid x\right)}_{\text{truth is } H_0} + a_1 \underbrace{1\left(\phi^* = 0\right)}_{=1-1(\phi^*=1)} \mathrm{P}\left(H_1 \mid x\right) \\ \text{simplify} &= \left[a_0 \mathrm{P}\left(H_0 \mid x\right) - a_1 \mathrm{P}\left(H_1 \mid x\right)\right] 1\left(\phi^* = 1\right) + a_1 \mathrm{P}\left(H_1 \mid x\right). \end{aligned}$$

The difference

$$\begin{aligned}
\mathrm{E}\left[L\left(\phi^{*}\right) \mid x\right]-\mathrm{E}\left[L\left(\phi\right) \mid x\right] &= \left[a_{0}\mathrm{P}\left(H_{0} \mid x\right)-a_{1}\mathrm{P}\left(H_{1} \mid x\right)\right] 1\left(\phi^{*}=1\right) \\
&\quad -\left[a_{0}\mathrm{P}\left(H_{0} \mid x\right)-a_{1}\mathrm{P}\left(H_{1} \mid x\right)\right] 1\left(\mathrm{P}\left(H_{0} \mid x\right)<\frac{a_{1}}{a_{0}+a_{1}}\right) \\
&= \left[a_{0}\mathrm{P}\left(H_{0} \mid x\right)-a_{1}\underbrace{\mathrm{P}\left(H_{1} \mid x\right)}_{=1-\mathrm{P}\left(H_{0} \mid x\right)}\right]\left[1\left(\phi^{*}=1\right)-1\left(\phi=1\right)\right] \\
&= \left(a_{0}+a_{1}\right)\left[\mathrm{P}\left(H_{0} \mid x\right)-\frac{a_{1}}{a_{0}+a_{1}}\right]\left[1\left(\phi^{*}=1\right)-1\left(\phi=1\right)\right].
\end{aligned}$$

1. If $\phi=1$, we must have $\mathrm{P}\left(H_{0} \mid x\right)<\frac{a_{1}}{a_{0}+a_{1}}$ and $1\left(\phi^{*}=1\right)-1\left(\phi=1\right)=1\left(\phi^{*}=1\right)-1$, which means that

$$\mathrm{E}\left[L\left(\phi^{*}\right) \mid x\right]-\mathrm{E}\left[L\left(\phi\right) \mid x\right] = \frac{1}{a_{0}+a_{1}}\left[\underbrace{\mathrm{P}\left(H_{0} \mid x\right)-\frac{a_{1}}{a_{0}+a_{1}}}_{<0}\right]\left[\underbrace{1\left(\phi^{*}=1\right)-1\left(\phi=1\right)}_{\leq 0}\right] \geq 0$$

2. If $\phi=0$, we must have $\mathrm{P}\left(H_{0} \mid x\right) \geq \frac{a_{1}}{a_{0}+a_{1}}$ and $1\left(\phi^{*}=1\right)-1\left(\phi=1\right)=1\left(\phi^{*}=1\right)$, which means that

$$\mathrm{E}\left[L\left(\phi^{*}\right) \mid x\right]-\mathrm{E}\left[L\left(\phi\right) \mid x\right] = \frac{1}{a_{0}+a_{1}}\left[\underbrace{\mathrm{P}\left(H_{0} \mid x\right)-\frac{a_{1}}{a_{0}+a_{1}}}_{\geq 0}\right]\left[\underbrace{1\left(\phi^{*}=1\right)-1\left(\phi=1\right)}_{\geq 0}\right] \geq 0$$

Above all, we obtain $\mathrm{E}\left[L\left(\phi^{*}\right) \mid x\right]-\mathrm{E}\left[L\left(\phi\right) \mid x\right] \geq 0$ for any x. Thus, $\mathrm{E}\left[L\left(\phi^{*}\right)\right]-\mathrm{E}\left[L\left(\phi\right)\right] \geq 0$ if we integrate out x using its marginal distribution $m\left(x\right)$.

Example 10. Suppose that independent $X_{i} \mid \theta \sim N\left(\theta, \sigma^{2}\right)$ for $i=1, ..., n$, where σ^{2} is known. The prior is $\theta \sim N\left(\mu_{0}, \sigma_{0}^{2}\right)$. We are interested in testing

$$H_{0}: \theta \leq 0, \quad \text{versus} \quad H_{1}: \theta>0.$$

The posterior is

$$\theta \mid x \sim N\left(\frac{\sigma_{0}^{2}\sum_{i=1}^{n} x_{i}+\sigma^{2}\mu_{0}}{\sigma_{0}^{2}+\sigma^{2}}, \quad \frac{\sigma^{2}\sigma_{0}^{2}}{\sigma_{0}^{2}+\sigma^{2}}\right).$$

The posterior probability is

$$\mathrm{P}\left(H_{0} \mid x\right) = \mathrm{P}\left(\theta \leq 0 \mid x\right)=\mathrm{P}\left(N\left(\frac{\sigma_{0}^{2}\sum_{i=1}^{n} x_{i}+\sigma^{2}\mu_{0}}{\sigma_{0}^{2}+\sigma^{2}}, \quad \frac{\sigma^{2}\sigma_{0}^{2}}{\sigma_{0}^{2}+\sigma^{2}}\right) \leq 0 \mid x\right)=\Phi\left(-\frac{\frac{\sigma_{0}^{2}\sum_{i=1}^{n} x_{i}+\sigma^{2}\mu_{0}}{\sigma_{0}^{2}+\sigma^{2}}}{\sqrt{\frac{\sigma^{2}\sigma_{0}^{2}}{\sigma_{0}^{2}+\sigma^{2}}}}\right).$$

□

Definition 6. Consider testing $H_{0}: \theta \in \Theta_{0}$ versus $H_{1}: \theta \in \Theta_{1}$. The Bayes factor is defined to be

$$B_{01}\left(x\right) = \frac{\mathrm{P}\left(H_{0} \mid x\right) / \mathrm{P}\left(H_{1} \mid x\right)}{\mathrm{P}\left(H_{0}\right) / \mathrm{P}\left(H_{1}\right)}.$$

A large $B_{01}\left(x\right)$ indicates that the marginal likelihood under H_{0} is higher than that under H_{1}. A rule-of-thumb to interpret the value of Bayes factor B_{10} (instead of B_{01}) is as follows.

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

A side note is that the marginal likelihood $m(x)$ is the omitted normalizing constant when we use $\lambda(\theta \mid x) \propto p(x \mid \theta)\lambda(\theta)$. We have to keep track all constants now!

Definition 7. A set $C(x)$ is a α-credible set if the posterior distribution satisfies

$$\mathrm{P}(\theta \in C(x) \mid x) \quad \geq \quad 1-\alpha, \quad \alpha \in [0,1].$$

1. It is highest posterior density (HPD) if it can be written as

$$\{\theta : \lambda(\theta \mid x) > k_\alpha\} \subseteq C(x) \subseteq \{\theta : \lambda(\theta \mid x) \geq k_\alpha\},$$

where k_α is the largest bound such that

$$\mathrm{P}(\theta \in C(x) \mid x) \quad \geq \quad 1-\alpha.$$

2. It is an equal tailed credible interval if the lower and upper bounds satisfy

$$\mathrm{P}(\theta \leq L(x) \mid x) \quad = \quad \mathrm{P}(\theta \geq U(x) \mid x) = \alpha/2.$$

Example 11. Let $X_1, ..., X_n$ be iid $N(0, \sigma^2)$. We assume that the prior of σ^2 is InvGamma(a_0, b_0). The posterior is

$$\sigma^2 \mid x \quad \sim \quad \text{InvGamma}\left(a_0 + \frac{n}{2},\, b_0 + \sum_{i=1}^{n} x_i^2\right).$$

The credible set is obtained from the posterior.