18. Bayesian Estimation

ECE 830, Spring 2014

The Bayesian Paradigm

Given a parameter θ, we assume observations are generated according to $p(x|\theta)$. In our work so far, we have treated the parameter θ like a fixed and deterministic quantity while the observation x is the realization of a random process.

It is tempting to interpret the likelihood as a measure of how likely different values of θ are given the data, but this is not always possible; for example, often

$$\int p(x|\theta) d\theta \to \infty$$

Another problematic issue is the mathematical formalization of statements like: "Based on the measurements of x, I am 95% confident that θ falls in a certain range."

Example: Unfair coin

Suppose you toss a coin 10 times and each time it comes up "heads." It might be reasonable to say that we are 99% sure that the coin is unfair, biased towards heads.

Formally, we can think about this in a hypothesis testing framework:

$$H_0 : \text{prob heads } \equiv \theta > 0.5$$

$$\text{Let } k := \sum_{n=1}^{10} x_n$$

$$p(x|\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k} \text{ binomial likelihood}$$

$$p(\theta > 0.5|x) = ?$$

Example: (cont.)

The problem with this is that

$$p(\theta \in H_0|x)$$

implies that θ is a random, not deterministic, quantity.

So, while "confidence" statements are very reasonable and in fact a normal part of "everyday thinking," this idea can not be supported from the classical perspective.

All of these "deficiencies" can be circumvented by a change in how we view the parameter θ.

Bayes Rule

If we view θ as the realization of a random variable with density $p(\theta)$, then Bayes Rule (Bayes, 1763) shows that

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\widetilde{\theta})\, p(\widetilde{\theta})\, d\widetilde{\theta}}$$

Thus, from this perspective we obtain a well-defined inversion:

Given x, the parameter θ is distributed according to $p(\theta|x)$.

From here, confidence measures such as $p(\theta \in H_0|x)$ are perfectly legitimate quantities to ask for.

Bayesian statistical models

Definition: Bayesian statistical model

A Bayesian statistical model is composed of a *data generation model*, $p(x|\theta)$, and a *prior* distribution on the parameters, $p(\theta)$.

The prior distribution (or "prior" for short) models the uncertainty in the parameter. More specifically, $p(\theta)$ models our knowledge - or a lack thereof - prior to collecting data.

Notice that

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} \propto p(x|\theta)\, p(\theta)$$

Hence, $p(\theta|x)$ is proportional to the likelihood function multiplied by the prior.

The Bayes Advantage

Bayesian analysis has some significant advantages over classical statistical analysis:

1. Properly inverts the relationship between causes and effects
2. Permits meaningful assessments in confidence regions
3. Enables the incorporation of prior knowledge into the analysis (which could come from previous experiments, for example)
4. Leads to more accurate estimators (provided the prior knowledge is accurate)

Example: DC level in AWGN

$$x_n = A + w_n\ , \quad n = 1, \cdots, N$$
$$w_n \sim \mathcal{N}(0, \sigma^2)\ \ iid$$
$$\widehat{A} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{MVUE and MLE estimator}$$

Now suppose that we have prior knowledge that $-A_0 \leq A \leq A_0$. We might incorporate this by forming a new estimator

$$\widetilde{A} = \left\{ \begin{array}{l} \\ \\ \\ \end{array} \right.$$

This is called a truncated sample mean estimator of A.

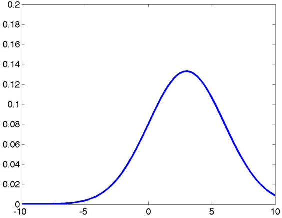
Example: (cont.)

Is $\widetilde{A}$ a better estimator of A than the sample mean $\widehat{A}$? Let $p(a)$ denote the density of $\widehat{A}$. Since $\widehat{A} = \frac{1}{N}\sum x_n$,

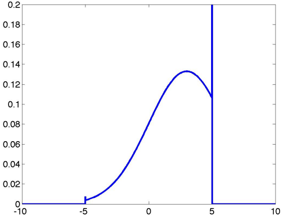
$$p(a) =$$

The density of $\widetilde{A}$ is given by

$$\widetilde{p}(a) \quad =$$



$p(a)$



$\widetilde{p}(a)$

Now consider the MSE of the sample mean $\widehat{A}$:

$$\begin{aligned}
\mathsf{MSE}(\widehat{A}) &= \int_{-\infty}^{\infty} (a-A)^2\, p(a)\, da \\
&= \int_{-\infty}^{-A_0} (a-A)^2\, p(a)\, da + \int_{-A_0}^{A_0} (a-A)^2\, p(a)\, da \\
&\quad + \int_{A_0}^{\infty} (a-A)^2\, p(a)\, da \\
&> \int_{-\infty}^{-A_0} (-A_0-A)^2\, p(a)\, da + \int_{-A_0}^{A_0} (a-A)^2\, p(a)\, da \\
&\quad + \int_{A_0}^{\infty} (A_0-A)^2\, p(a)\, da \\
&= (-A_0-A)^2\, \mathbb{P}(\widehat{A} \le -A_0) + \int_{-A_0}^{A_0} (a-A)^2\, p(a)\, da \\
&\quad + (A-A_0)^2\, \mathbb{P}(\widehat{A} \ge A_0) \\
&= \mathsf{MSE}(\widetilde{A})
\end{aligned}$$

$$\mathsf{MSE}(\widehat{A}) > \mathsf{MSE}(\widetilde{A})$$

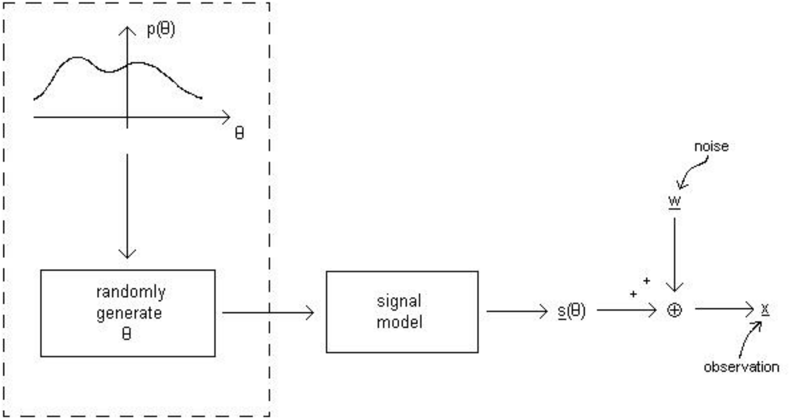
Note

1. $\widetilde{A}$ is biased
2. Although $\widehat{A}$ is MVUE, $\widetilde{A}$ is better in the MSE sense
3. Prior information is aptly described by regarding A as a random variable with a prior distribution.

$$\text{uniform}(-A_0, A_0)$$

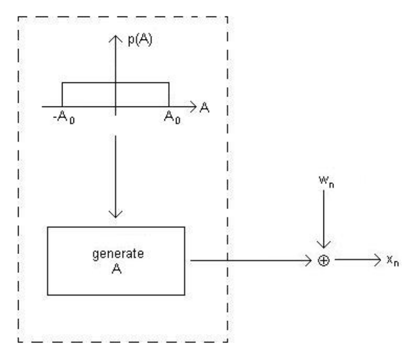
$\Rightarrow$ We know $-A_0 \leq A \leq A_0$, but otherwise A is arbitrary.

The Bayesian Approach to Statistical Modeling



Example:

$x_n = A + w_n \ , \ \ n = 1, \cdots, N$



The prior distribution allows us to incorporate prior information regarding unknown parameter - probable values of parameter are supported by prior. Basically, the prior reflects what we believe "nature" will probably throw at us.

Elements of Bayesian Analysis

(a) Joint distribution

$$p(x, \theta) =$$

(b) Marginal distributions

$$p(x) = \int$$

$$p(\theta) = \int$$

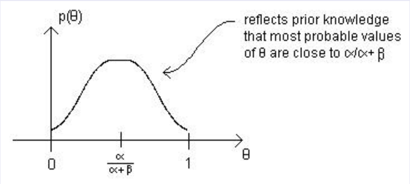
(c) Posterior distribution

$$p(\theta|x) =$$

Example: Binomial + Beta

$$
\begin{aligned}
p(x|\theta) &= \binom{n}{x}\theta^{x}(1-\theta)^{n-x}, 0 \le \theta \le 1 \\
&= \text{binomial likelihood} \\
p(\theta) &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\
&= \text{Beta prior distribution}
\end{aligned}
$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ is the Gamma function



Example: (cont.)

- Joint Density

$$p(x, \theta) =$$

- Marginal Density

$$p(x) =$$

- Posterior Density

$$p(\theta|x) =$$

Bayesian Estimation

We are interested in estimating θ given the observation x within a Bayesian framework. Naturally, then, any estimation strategy will be based on the posterior distribution $p(\theta|x)$.

However, we need a criterion for assessing the quality of potential estimators.

Loss

Definition: Loss

The quality of an *estimate* $\widehat{\theta}$ is measured by a real-valued *loss (or cost) function*

$$L(\theta, \widehat{\theta}).$$

For example, *squared error* or quadratic loss is simply

$$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta}(x))^\top (\theta - \widehat{\theta}(x)) = \|\theta - \widehat{\theta}(x)\|^2.$$

Definition: Bayes Risk

The quality of an estimator is measured by the expected loss, known as the Bayes risk:

$$R(\widehat{\theta}) := \mathbb{E}_{x,\theta}\left[L(\theta, \widehat{\theta})\right].$$

Note that the expectation is with respect to both x and θ.

For example, if x and θ are jointly continuous, then

$$
\begin{aligned}
R(\widehat{\theta}) &= \iint L(\theta, \widehat{\theta}(x)) p(\theta, x) dx d\theta \\
&= \iint L(\theta, \widehat{\theta}(x)) p(x|\theta) p(\theta) dx d\theta \\
&=
\end{aligned}
$$

In general, Bayesian estimation seeks the estimator

$$\begin{aligned}
\widehat{\theta} &= \arg\min_{\widetilde{\theta}} R(\widetilde{\theta}) \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{x,\theta} \left\{ L\left(\theta, \widetilde{\theta}(x)\right) \right\} \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{x} \left\{ \mathbb{E}_{\theta|x} \left\{ L\left(\theta, \widetilde{\theta}(x)\right) |x = x \right\} \right\}
\end{aligned}$$

minimizing the Bayes risk. Thus, given the data x, the "best" or optimal estimator under a given loss function is given by

This is called the "posterior expected loss"; it depends only on the loss function and the posterior distribution.

Bayes Minimum MSE

Measure the loss as $L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|^2$. In classical estimation, we tried to minimize

$$\mathbb{E}_x \left[L(\theta, \widehat{\theta}(x)) \right],$$

but didn't get a practical estimator. (Recall why.) In the Bayesian setting, we try to minimize

$$\mathrm{BMSE}(\widehat{\theta}) := \mathbb{E}_{x,\theta} \left[L(\theta, \widehat{\theta}(x)) \right]$$

and get a very different result.

Definition

The estimator that minimizes the $\mathrm{BMSE}(\widehat{\theta})$ is called the *minimum mean squared error (MMSE)* estimator.

Now note

$$
\begin{aligned}
&\mathbb{E}_{\theta|x=x}\left[\left(\theta-\widehat{\theta}(x)\right)^{\top}\left(\theta-\widehat{\theta}(x)\right)|x\right] \\
&= \mathbb{E}_{\theta|x}\left[\left(\theta-\mathbb{E}[\theta|x]+\mathbb{E}[\theta|x]-\widehat{\theta}(x)\right)^{\top}\right. \\
&\qquad \left.\times\left(\theta-\mathbb{E}[\theta|x]+\mathbb{E}[\theta|x]-\widehat{\theta}(x)\right)|x\right] \\
&= \mathbb{E}_{\theta|x}\left[(\theta-\mathbb{E}[\theta|x])^{\top}(\theta-\mathbb{E}[\theta|x])|x\right] \\
&\quad +2\mathbb{E}_{\theta|x}\left[(\theta-\mathbb{E}[\theta|x])^{\top}\left(\mathbb{E}[\theta|x]-\widehat{\theta}(x)\right)|x\right] \\
&\quad +\mathbb{E}_{\theta|x}\left[\left(\mathbb{E}[\theta|x]-\widehat{\theta}(x)\right)^{\top}\left(\mathbb{E}[\theta|x]-\widehat{\theta}(x)\right)|x\right]
\end{aligned}
$$

The first term is independent of $\widehat{\theta}(x)$ and the second term is 0. The third term can be minimized by taking

$$\widehat{\theta}_{\mathrm{MMSE}}(x)=\mathbb{E}[\theta|x]=\int \theta p(\theta|x)d\theta$$

which is the .

Example: DC Level in AWGN

$$x_n = A + w_n$$

$n = 1, \cdots, N$, $w_n \sim \mathcal{N}(0, \sigma^2)$. Prior for unknown parameter A:

$$p(a) = \mathsf{Unif}[-A_0, A_0]$$

$$p(x|A) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - A)^2\right\}$$

$$p(A|x) = \begin{cases} \dfrac{\frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - A)^2\right\}}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - a)^2\right\} da} & \text{if } |A| \le A_0 \\ 0 & \text{if } |A| > A_0 \end{cases}$$

Example: (cont.)

Bayes Minimum MSE Estimator:

$$\begin{aligned}\widehat{A} &= \mathbb{E}\left[A|x\right] = \int_{-\infty}^{\infty} Ap(A|x)dA \\ &= \frac{\int_{-A_0}^{A_0} A \cdot \frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - A)^2\right\} dA}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - a)^2\right\} da}\end{aligned}$$

Notes:

1. No closed-form estimator
2. As $A_0 \to \infty$,
3. For smaller A_0, truncated integral produces an $\widehat{A}$ that is a function of x, σ^2, and A_0
4. As N increases σ^2/N decreases and posterior $p(A|x)$ becomes tightly clustered about $\frac{1}{N}\sum x_n$

   (the data "swamps out" the prior)

Other Common Loss Functions

Absolute Error Loss (Laplace, 1773):

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_1 \equiv \sum_{i=1}^{p} |\theta_i - \widehat{\theta}_i|$$

Scalar case:

$$\begin{aligned}
\mathbb{E}\left[L(\theta, \widehat{\theta})|x\right] &= \int_{-\infty}^{\infty} |\theta - \widehat{\theta}| p(\theta|x) d\theta \\
&= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) p(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) p(\theta|x) d\theta
\end{aligned}$$

The optimal estimator under this loss is referred to the "minimum mean absolute error" (MMAE) estimator.

To see what estimator minimises this loss, we differentiate $\mathbb{E}\left[L(\theta,\widehat{\theta})|x\right]$ with respect to $\widehat{\theta}$ (using Leibnitz's rule) to get

$$\frac{\partial}{\partial\widehat{\theta}}\mathbb{E}\left[L(\theta,\widehat{\theta})|x\right] = P(\widehat{\theta}(x)|x) - (1 - P(\widehat{\theta}(x)|x)),$$

where $P(\theta|x)$ is the posterior cumulative distribution function of θ given x. Setting this equal to zero, this implies $P(\widehat{\theta}(x)|x) = 1/2$ or

$$\mathbb{P}(\theta < \widehat{\theta}|x) = \mathbb{P}(\theta > \widehat{\theta}|x).$$

The optimal $\widehat{\theta}$ under absolute error loss is

Uniform Loss:

$$L(\theta, \widehat{\theta}) = I_{\{\|\widehat{\theta}-\theta\|>\epsilon\}} = \begin{cases} 1 & \text{if } \|\theta - \widehat{\theta}\| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

where $\epsilon > 0$ is small. The posterior expected loss is

$$\mathbb{E}\left[L(\theta, \widehat{\theta})|x\right] = \mathbb{E}\left[I_{\{\|\widehat{\theta}-\theta\|>\epsilon\}}|x\right] = \mathbb{P}(\|\widehat{\theta} - \theta\| > \epsilon|x)$$

which is the posterior probability that θ deviates from $\widehat{\theta}(x)$ by more then ϵ. To minimize this uniform loss we must choose $\widehat{\theta}$ to be the value of θ with highest posterior probability.

The optimal estimator $\widehat{\theta}$ under uniform loss is the

Taking the limit as $\epsilon \longrightarrow 0$ gives:

Definition

Maximum A Posteriori (MAP) estimator - the value of θ where $p(\theta|x)$ is maximized:

$$\widehat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\widetilde{\theta}} p(\widetilde{\theta}|x) = \arg\max_{\widetilde{\theta}} p(x|\widetilde{\theta})p(\widetilde{\theta})$$

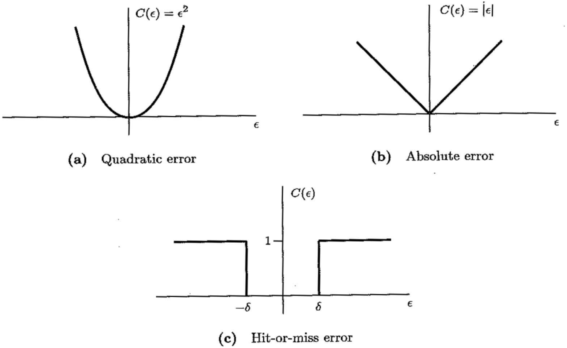
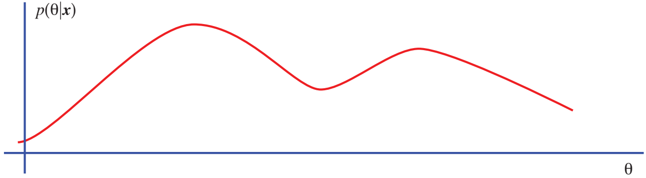
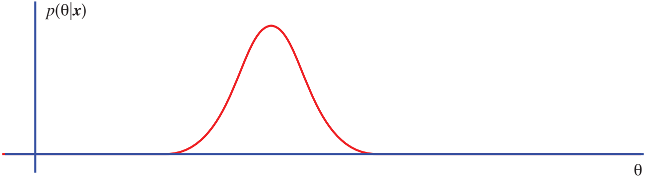


Figure 11.1 Examples of cost function



(a) General posterior PDF



(b) Gaussian posterior PDF

If the posterior is symmetric and unimodal, then

Computation

Both $\widehat{\theta}_{\mathrm{MMSE}}$ and $\widehat{\theta}_{\mathrm{MMAE}}$ require integrating with respect to $p(\theta|x)$. Often this calculation will be intractable. How can we approximate these estimators numerically?
One common approach: if we can simulate $\theta_1, \ldots, \theta_M$ from $p(\theta|x)$, then we can apply the following Monte Carlo estimates:

$$\widehat{\theta}_{\mathrm{MMSE}}(x) \approx$$

$$\widehat{\theta}_{\mathrm{MMAE}}(x) \approx$$

If the posterior mode cannot be determined analytically, then many numerical approaches for MLE can be applied.
Which of the three loss functions is used is often dictated by computational considerations.