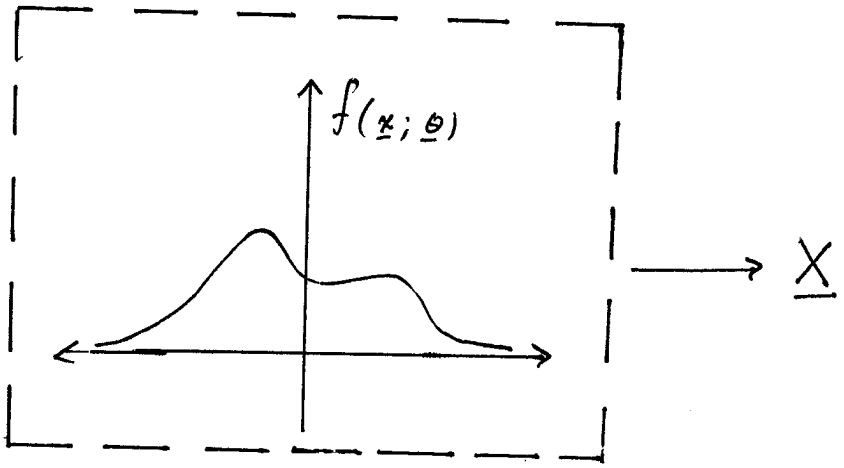
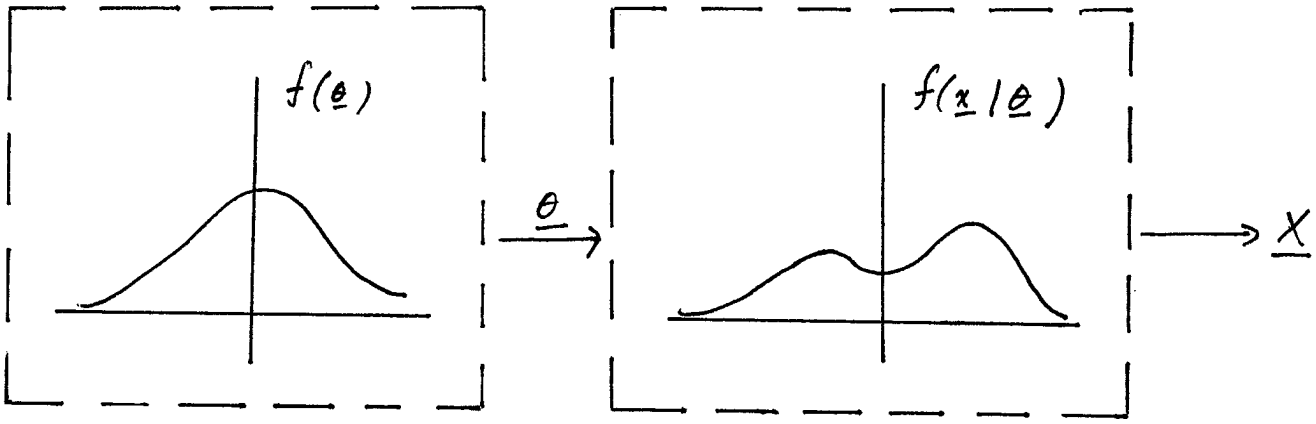
BAYESIAN ESTIMATION

Bayesian Statistical Modeling

In classical estimation, the unknown parameter $\underline{\theta}$ is viewed as nonrandom, and a statistical model is specified entirely by a data model (likelihood) $f(\underline{x}; \underline{\theta})$



In Bayesian estimation, the unknown parameter is viewed as random. A statistical model is specified in terms of a conditional pdf/pmf $f(\underline{x} \mid \underline{\theta})$ and a prior distribution $f(\underline{\theta})$ on $\underline{\theta}$.



The prior $f(\underline{\theta})$ is specified by the investigator and reflects "prior knowledge" about the uncertainty in $\underline{\theta}$.

Note that we now write $f(\underline{x} \mid \underline{\theta})$ instead of $f(\underline{x}; \underline{\theta})$ to reflect that $\underline{\theta}$ is random.

By Bayes' rule, we may express the <u>posterior</u> distribution of $\underline{\theta}$ given $\underline{x}$ as

$$f(\underline{\theta} \mid \underline{x}) = \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x})}$$

$$= \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{\int f(\underline{x} \mid \underline{\theta}') \, f(\underline{\theta}') \, d\underline{\theta}'}$$

Whereas the prior reflects our uncertainty in θ before x is observed, the posterior represents our uncertainty in θ after x is observed.

Example 1 Suppose we have a coin whose probability θ of turning up heads is unknown. We toss the coin N times and observe X heads.

The natural model for X given θ is binomial:

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

One possible prior for θ is the beta distribution:

$$f(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \le \theta \le 1$$

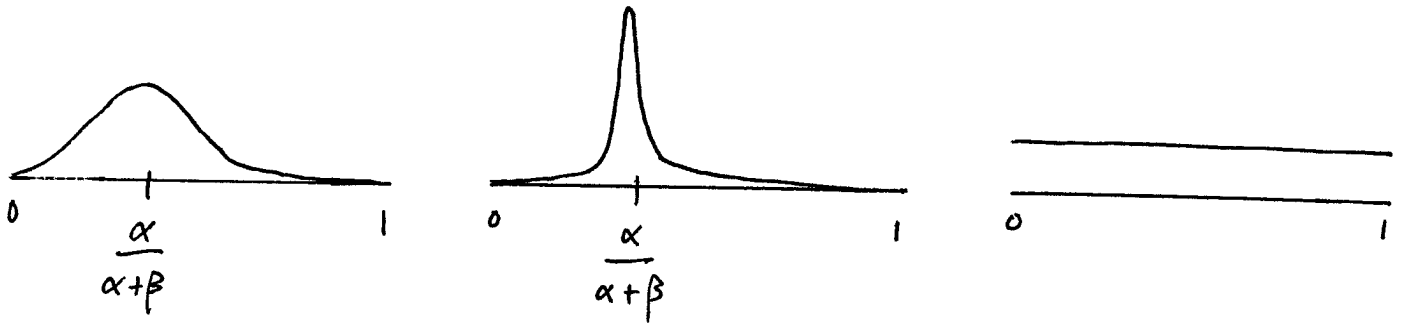
where

$$B(\alpha,\beta) := \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

and $\alpha, \beta \ge 1$.

$$E\{\theta\} = \frac{\alpha}{\alpha+\beta}, \quad Var\{\theta\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



α, β small $\qquad$ α, β large $\qquad$ $\alpha = \beta = 1$ (uniform)

The parameters α, β must be set by the user to reflect prior knowledge

- $\alpha = \beta = 1 \implies \theta$ could be anywhere
- $\alpha = \beta = 2 \implies \theta$ is probably fair or close to fair, but not really sure
- $\alpha = \beta = 10 \implies \theta$ is almost certainly fair

Let's see how our belief about θ changes once x is observed.

Viewing x as a constant, we have

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

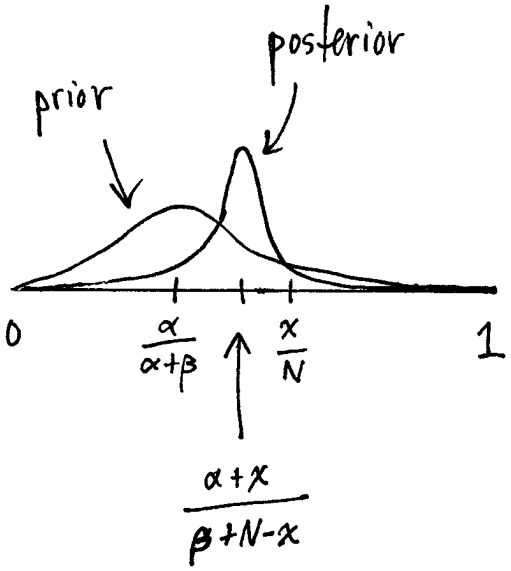
$$\propto p(x|\theta)\, f(x)$$

$$= \binom{N}{x} \theta^x (1-\theta)^{N-x} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

$$\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}$$

Since $f(\theta|x)$ is a density it must integrate to 1 and we recognize $\theta|x \sim \text{Beta}(\alpha+x, \beta+N-x)$ and

$$f(\theta|x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}}{B(\alpha+x, \beta+N-x)}, \quad 0 \le \theta \le 1$$



The posterior is shifted toward the observed frequency and is more concentrated (reflecting greater certainty).

The beta prior is said to be conjugate to the binomial data model because the prior and posterior belong to the same family.

Confidence Statements

One advantage of Bayesian over classical estimation is that in Bayesian inference, confidence statements are more natural.

For example, if we toss a coin 10 times and observe 10 heads, it makes perfect sense to assert "it is highly probable that the coin is unfair and biased towards heads."

Formally, let

$$\mathcal{H} = \left(\tfrac{1}{2}, 1\right]$$

$$X \sim \binom{N}{x} \theta^x (1-\theta)^{N-x}.$$

What is

$$\text{Prob}\left(\theta \in \mathcal{H} \mid X = x\right)?$$

Asking such a question already suggests that θ is random. In the Bayesian framework, the answer to the question is

$$\text{Prob}\left(\theta \in \textcircled{A} \mid \underline{x}\right) = \int_{\textcircled{A}} f(\theta \mid x)\, d\theta$$

Even though such confidence statements are a normal part of "everyday thinking," they are less natural in the classical setting

Likelihood Principle

All Bayesian methods for estimation are based on the posterior. As a consequence, Bayesian estimation conforms to the likelihood principle. To see this, suppose $\underline{x}_1$ and $\underline{x}_2$ are such that

$$f(\underline{x}_1 \mid \underline{\theta}) = c \cdot f(\underline{x}_2 \mid \underline{\theta}) \qquad \forall \underline{\theta}$$

for some constant c. Then

$$f(\underline{\theta} \mid \underline{x}_1) = \frac{f(\underline{x}_1 \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x}_1)}$$

$$= \left[\frac{c f(\underline{x}_2)}{f(\underline{x}_1)}\right] \cdot \frac{f(\underline{x}_2 \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x}_2)}$$

$$\propto f(\underline{\theta} \mid \underline{x}_2)$$

$$\Rightarrow f(\underline{\theta} \mid \underline{x}_1) = f(\underline{\theta} \mid \underline{x}_2).$$

Sufficiency Principle

Bayesian inference also obeys the sufficiency principle, which states that if $\tau(\underline{x}_1) = \tau(\underline{x}_2)$, where $\underline{T} = \tau(\underline{X})$ is a sufficient statistic, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same inference about $\underline{\theta}$.

To see this, note

$$f(\underline{\theta} \mid \underline{x}) = \frac{f(\underline{x} \mid \underline{\theta})\, f(\underline{\theta})}{f(\underline{x})}$$

$$= \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{\int f(\underline{x} \mid \underline{\theta}')\, f(\underline{\theta}')\, d\underline{\theta}'}$$

$$= \frac{g(\underline{t} \mid \underline{\theta})\, h(\underline{x})\, f(\underline{\theta})}{\int g(\underline{t} \mid \underline{\theta}')\, h(\underline{x})\, f(\underline{\theta}')\, d\underline{\theta}'}$$

$$= \frac{g(\underline{t} \mid \underline{\theta})\, f(\underline{\theta})}{\int g(\underline{t} \mid \underline{\theta}')\, f(\underline{\theta}')\, d\underline{\theta}'}$$

which depends on $\underline{x}$ only through $\underline{t} = \tau(\underline{x})$.

Pros of Bayesian Inference

- allows incorporation of prior information
- leads to better estimates if prior information is accurate
- provides for valid confidence statements
- satisfies the likelihood and sufficiency principles

Cons of Bayesian Inference

- prior knowledge can be difficult to specify
- can lead to worse estimates (relative to classical methods) if prior knowledge is inaccurate
- in practice, the choice of prior can be dictated by tractability considerations (e.g., conjugate priors are usually preferred regardless of how appropriate they are)

Bayesian Estimation

The goal of Bayesian estimation is the same as classical estimation: given an observation $\underline{x}$, estimate a specific value $\underline{\theta}$.

Unfortunately, the convention (which I will adhere to) is to refer to the random parameter *and* its realizations as $\underline{\theta}$.

So, given $\underline{x}$, we want to estimate the specific realization $\underline{\theta}$ that describes the model $f(\underline{x} \mid \underline{\theta})$.

Loss functions and Risk

The quality of an *estimate* is measured by a *loss* (or *cost*) function

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x}))$$

For example, the quadratic loss (squared error) is

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) = (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))^T (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))$$

$$= \| \underline{\theta} - \hat{\underline{\theta}}(\underline{x}) \|^2$$

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

$$R(\hat{\underline{\theta}}) = E_{\underline{X},\underline{\theta}}\left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X})) \right\}$$

Note: The expectation is with respect to both $\underline{X}$ and $\underline{\theta}$. For example, if $\underline{X}$ & $\underline{\theta}$ are jointly continuous, then

$$R(\hat{\underline{\theta}}) = \int\int L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) f(\underline{x}, \underline{\theta}) \, d\underline{x} \, d\underline{\theta}$$

$$= \int\int L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) f(\underline{x}|\underline{\theta}) f(\underline{\theta}) \, d\underline{x} \, d\underline{\theta}$$

In general, Bayesian estimation seeks the estimator

$$\hat{\underline{\theta}} = \arg\max_{\underline{\phi}} R(\phi)$$

minimizing the Bayes risk. The optimal estimator will depend on the statistical model and the loss.

In fact, the optimal estimator may be expressed solely in terms of the loss $L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x}))$ and the posterior $f(\underline{\theta} \mid \underline{x})$. To see this, note

$$R(\hat{\underline{\theta}}) = E_{\underline{X},\underline{\theta}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X})) \right\}$$

$$= E_{\underline{X}} \left\{ E_{\underline{\theta}|\underline{X}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{X} = \underline{x} \right\} \right\}$$

Thus, to minimize the risk, $\hat{\underline{\theta}}(\underline{x})$ must minimize

$$E_{\underline{\theta}|\underline{X}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{X} = \underline{x} \right\}$$

for each $\underline{x}$.

Posterior expected loss: depends only on loss & posterior

Said another way, the optimal estimator is

$$\hat{\underline{\theta}}(\underline{x}) =$$

(a)

Let's apply this result to some specific loss functions.

Minimum Mean Squared Error

In classical estimation, where only $\underline{X}$ is random, the MSE criterion

$$E_{\underline{X}}\left\{ (\underline{\theta} - \hat{\underline{\theta}})^T (\underline{\theta} - \hat{\underline{\theta}}) \right\}$$

did not lead to a practical estimator. In the Bayesian setup, the situation is different.

We define the <u>Bayesian MSE</u> to be the Bayes risk when the loss function is the squared error:

$$\text{BMSE}(\hat{\underline{\theta}}) := E_{\underline{X},\underline{\theta}}\left\{ (\underline{\theta} - \hat{\underline{\theta}})^T (\underline{\theta} - \hat{\underline{\theta}}) \right\}$$

The estimator that minimizes $\text{BMSE}(\hat{\underline{\theta}})$ is called the <u>minimum mean squared error</u> (MMSE) estimator

As we just saw, the MMSE estimator must minimize

$$E_{\underline{\theta}|\underline{X}}\left\{ (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))^T (\underline{\theta} - \hat{\underline{\theta}}(\underline{x})) \,\middle|\, \underline{X} = \underline{x} \right\}$$

for each $\underline{x}$.

Observe

$$E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-\hat{\underline{\theta}}(\underline{x}))^T(\underline{\theta}-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$= E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}]+E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T(\underline{\theta}-E[\underline{\theta}|\underline{x}]+E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$= E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])^T(\underline{\theta}-E[\underline{\theta}|\underline{x}])\,|\,\underline{x}\right\}$$

$$+ 2E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])^T(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$+ E_{\underline{\theta}|\underline{x}}\left\{(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

First term: independent of $\hat{\underline{\theta}}(\underline{x})$

Second term: $2(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T \cdot \underbrace{E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])\,|\,\underline{x}\right\}}_{=\underline{0}}$

Third term: minimized by taking

$$\hat{\underline{\theta}}(\underline{x}) = E[\underline{\theta}|\underline{x}]$$

$$= \int \underline{\theta}\, f(\underline{\theta}|\underline{x})\, d\underline{\theta}$$

(b)

$$=$$

Exercise | Suppose $X \sim \text{Bin}(N, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. Find the MMSE estimator of θ. Prove the formula for the mean of a Beta random variable.

Solution | Earlier we saw

$$f(\theta|x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}}{B(\alpha+x, \beta+N-x)}$$

$$\Rightarrow \hat{\theta}(x) = E[\underline{\theta}|\underline{x}] = \int \theta\, f(\theta|x)\, d\theta$$

$$= \frac{1}{B(\alpha+x,\beta+N-x)} \int \theta^{\alpha+x}(1-\theta)^{\beta+N-x-1}\, d\theta$$

$$= \frac{B(\alpha+x+1, \beta+N-x)}{B(\alpha+x, \beta+N-x)}$$

$$= \frac{\Gamma(\alpha+x+1)\,\Gamma(\beta+N-x)}{\Gamma(\alpha+\beta+N+1)} \cdot \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+x)\cdot\Gamma(\beta+N-x)}$$

$$= \frac{\alpha+x}{\alpha+\beta+N}$$

$$\boxed{\frac{\Gamma(z+1)}{\Gamma(z)} = z}$$

Minimum Mean Absolute Error

For a scalar parameter θ define the absolute error loss

$$L(\theta, \hat{\theta}(\underline{x})) = |\theta - \hat{\theta}(\underline{x})|.$$

The posterior expected loss may be written

$$E_{\theta|\underline{x}}\left[L(\theta, \hat{\theta}(\underline{x})) \mid \underline{x} \right]$$

$$= \int_{-\infty}^{\infty} |\theta - \hat{\theta}(\underline{x})| \, f(\theta \mid \underline{x}) \, d\theta$$

integration by parts

$$= \int_{-\infty}^{\hat{\theta}(\underline{x})} (\hat{\theta}(\underline{x}) - \theta) f(\theta|\underline{x}) d\theta + \int_{\hat{\theta}(\underline{x})}^{\infty} (\theta - \hat{\theta}(\underline{x})) f(\theta|\underline{x}) d\theta$$

$$= \int_{-\infty}^{\hat{\theta}(\underline{x})} F(\theta \mid \underline{x}) \, d\theta + \int_{\hat{\theta}(\underline{x})}^{\infty} (1 - F(\theta \mid \underline{x})) \, d\theta$$

where $F(\theta \mid \underline{x})$ is the posterior cumulative distribution function of θ given $\underline{x}$.

To minimize this expression let's take the derivative w.r.t $\hat{\theta}(\underline{x})$:

$$F(\hat{\theta}(\underline{x}) \mid \underline{x}) - (1 - F(\hat{\theta}(\underline{x}) \mid \underline{x})) = 0$$

$$\Rightarrow F(\hat{\theta}(\underline{x}) \mid \underline{x}) = \frac{1}{2}$$

$\Rightarrow \hat{\theta}(\underline{x})$ is the _____

Minimum Mean Uniform Error Estimation

The <u>uniform error loss</u> is defined to be

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) = I_{\{\|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon\}}$$

$$= \begin{cases} 1 & \text{if } \|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

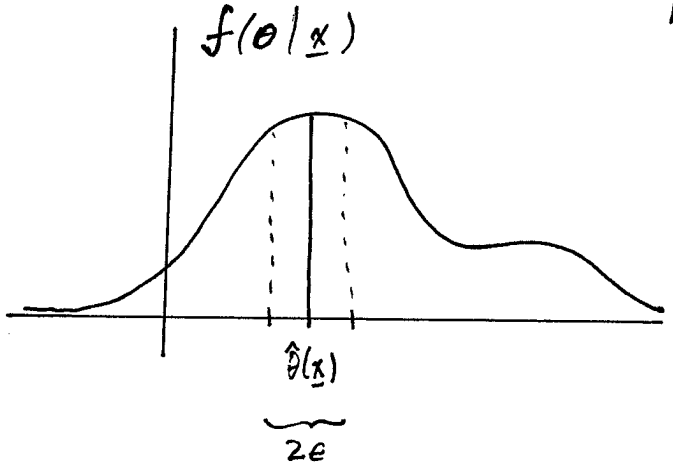
where $\epsilon > 0$ is small.

The posterior expected loss is

$$E\{L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{x}\} = P(\|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon \mid \underline{x}),$$

the posterior probability that $\underline{\theta}$ deviates from $\hat{\underline{\theta}}(\underline{x})$ by more than ϵ. This probability is minimized, for ϵ sufficiently small, when $\hat{\underline{\theta}}(\underline{x})$ is near the _____.

(d)



Taking the limit as $\epsilon \to 0$ gives rise to the

maximum a posteriori (MAP) estimator:

$$\begin{aligned}
\hat{\underline{\theta}}(\underline{x}) &= \arg\max_{\underline{\theta}} \; f(\underline{\theta} \mid \underline{x}) \\
&= \arg\max_{\underline{\theta}} \; \frac{f(\underline{x} \mid \underline{\theta})\, f(\underline{\theta})}{f(\underline{x})} \\
&= \arg\max_{\underline{\theta}} \; f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta}).
\end{aligned}$$

This last expression is often easiest to compute: It avoids having to determine $f(\underline{x})$ or $f(\underline{\theta} \mid \underline{x})$.

Exercise | Suppose $X \sim \text{Bin}(N, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. Find the MAP estimator of θ.

Solution | Recall

$$f(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}$$

$$\Rightarrow \log f(\theta|x) = (\alpha+x-1)\log\theta + (\beta+N-x-1)\log(1-\theta) + C$$

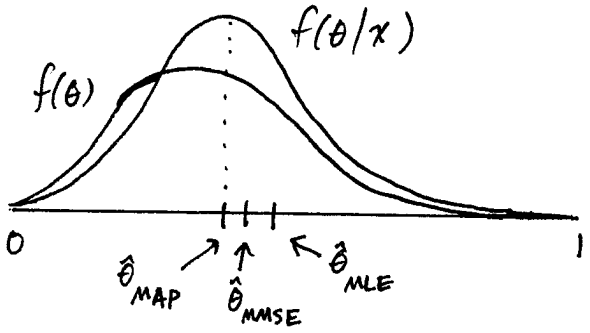
$$\Rightarrow \frac{\partial \log f(\theta|x)}{\partial\theta} = \frac{\alpha+x-1}{\theta} - \frac{\beta+N-x-1}{1-\theta} = 0$$

$$\Rightarrow \hat{\theta}_{MAP}(x) = \frac{\alpha+x-1}{\alpha+\beta+N-2}$$

Compare:

$$\hat{\theta}_{MMSE}(x) = \frac{\alpha+x}{\alpha+\beta+N}$$

$$\hat{\theta}_{MLE}(x) = \frac{x}{N}$$

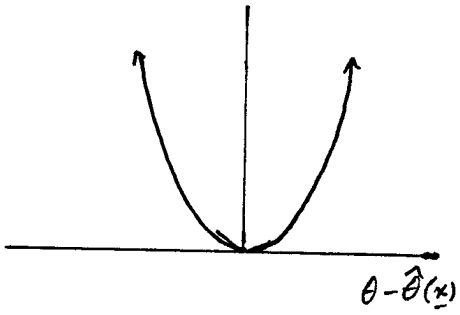


Note that both Bayesian estimators tend to the classical estimator as $N \to \infty$; a flood of data can overwhelm any amount of prior knowledge.

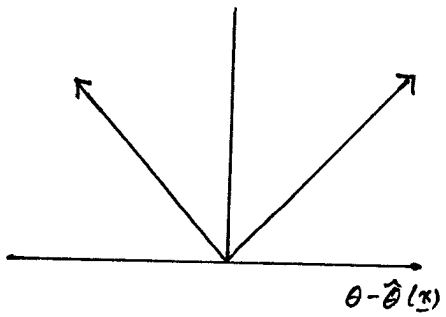
Note: the posterior median requires a closed form expression for the posterior CDF which is not available.

Discussion and Summary

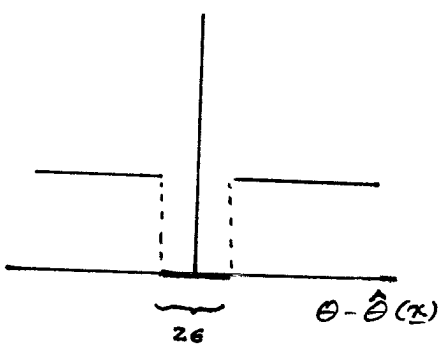
squared error



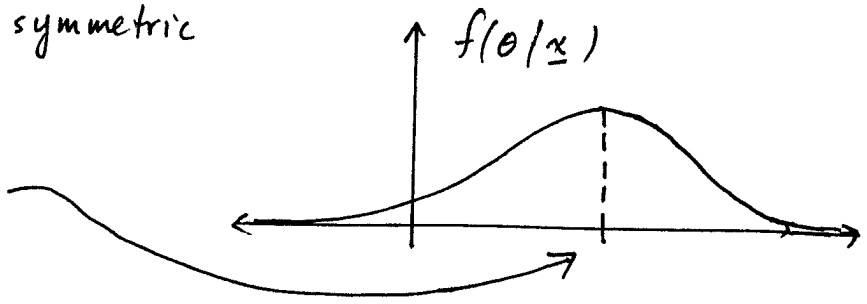
absolute error



uniform error



- These are the primary three loss functions used
- Estimate depends on $\underline{x}$ through the posterior:
  posterior mean, posterior median, and posterior mode
- If the posterior is symmetric and unimodal, then
  $\hat{\theta}_{MMSE} = \hat{\theta}_{MMAE} = \hat{\theta}_{MAP}$



- Limiting cases:
  - $N \to \infty \implies$ Bayesian $\to$ classical
  - prior $\to$ least informative $\not\Rightarrow$ Bayesian $\to$ classical
- $\hat{\theta}_{MMAE}$ does not generalize well to vector $\underline{\theta}$.
- Both $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MMAE}$ require integrating w.r.t. $f(\underline{\theta}|\underline{x})$. Often this calculation will be intractable. How can we approximate these estimators numerically?

If we can simulate $\underline{\theta}_1, \ldots, \underline{\theta}_M$ from $f(\underline{\theta} | \underline{x})$, then we can apply the following Monte Carlo estimates:

$$\hat{\theta}_{MMSE}(\underline{x}) \approx \frac{1}{M} \sum_{i=1}^{M} \theta_i$$

$$\hat{\theta}_{MMAE}(\underline{x}) \approx \text{median}\{\theta_1, \ldots, \theta_M\}$$

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

Key

a. $\hat{\underline{\theta}}(\underline{x}) = \arg\min_{\underline{\phi}} E_{\underline{\theta}|\underline{X}}\{L(\underline{\theta}, \underline{\phi}) \mid \underline{X} = \underline{x}\}$

b. posterior mean

c. posterior median

d. posterior mode