14 Bayesian Estimation

Thomas Bayes (18th-century mathematician and statistician)

Sir Harold Jeffreys (famous 20th-century mathematician and statistician) wrote that Bayes' theorem "is to the theory of probability what Pythagoras's theorem is to geometry"

14.1 Review: Properties of ML Estimator

Data: i.i.d. sample of size n drawn from $P(X|\theta)$

Consistency: the sequence of MLE estimates $\widehat{\theta}$ converges in probability to the true parameter value θ

Asymptotic Normality: as the sample size increases, the distribution of the MLE tends to the Gaussian distribution with mean θ (and covariance matrix equal to the inverse of the Fisher information matrix)

Efficiency: No consistent estimator has lower asymptotic mean squared error than the ML estimator (ML estimator achieves the Cramer-Rao lower bound when the sample size tends to infinity)

14.2 Bayes' Rule / Theorem

For events A and B, $P(A|B) = P(B|A)P(A)/P(B)$

Proof follows from our definition of conditional probability, i.e., $P(X|Y) := P(X \cap Y)/P(Y)$

14.3 Example (Coin Flip)

Consider that we don't know if a coin is fair / unfair

We have 2 possibilities in our mind:
(1) Coin fair, i.e., $P(\text{head}) = p = 0.5$
(2) Coin biased towards heads with $P(\text{head}) = q = 0.7$

We have a belief (**prior** to observing data) that $P(\text{CoinFair}) = 0.8$

Now we experiment with the coin, collect data, and recompute the probability that the coin is fair

$$P(\text{CoinFair}|\text{Data}) = P(\text{Data}|\text{CoinFair})P(\text{CoinFair})/P(\text{Data})$$

Given: We have data = n observations with r heads and $(n-r)$ tails. What does the data do to our belief ?

$$\begin{aligned}
P(\text{Data}|\text{CoinFair}) &= C_r^n 0.5^r 0.5^{n-r} \\
P(\text{Data}|\text{CoinUnfair}) &= C_r^n 0.7^r 0.3^{n-r} \\
P(\text{Data}) &= P(\text{Data}|\text{CoinFair})P(\text{CoinFair}) + P(\text{Data}|\text{CoinUnfair})P(\text{CoinUnfair}) \\
P(\text{CoinFair}|\text{Data}) &= \frac{0.5^r 0.5^{n-r} \times 0.8}{0.5^r 0.5^{n-r} \times 0.8 + 0.7^r 0.3^{n-r} \times 0.2}
\end{aligned}$$

Case 1: If $n = 20, r = 11$, then $P(\text{CoinFair}|\text{Data}) = 0.9074$ which is more than 0.8. So the data has strengthened our belief !!
Why has this happened ? Because 11 heads out of 20 is more like the fair coin.

Case 2: If $n = 20, r = 13$, then $P(\text{CoinFair}|\text{Data}) = 0.6429$ which is less than 0.8. So the data has weakened our belief !!
Why has this happened ? Because 13 heads out of 20 is more like the unfair coin.

Case 3: If $n = 20, r = 12$, then $P(\text{CoinFair}|\text{Data}) = 0.8077$ which is close to 0.8.

14.4 Example (Box)

There are two boxes:
(i) one with 4 black balls and 1 white ball
(ii) another with 1 black ball and 3 white balls

You pick one box at random (*prior* probability of picking any box is 0.5).

Then select a ball from the box. It turns out to be white (*data*).

Given that the ball is white, what is the probability that you picked the 1st box ?

Solution: $P(Box1|W) = P(W|Box1)P(Box1)/P(W)$ where,
using total probability, $P(W) = P(W|Box1)P(Box1) + P(W|Box2)P(Box2)$

$P(Box1|W)$ comes out to 0.2105
Prior probability for $P(Box1)$ was 0.5

14.5 Example: Gaussian (Unknown mean, Known variance)

Given: Data $\{x_i\}_{i=1}^N$ derived from a Gaussian distribution with known variance σ^2, but unknown mean μ

Treat mean μ as a random variable

Prior belief on μ is that it is derived from a Gaussian with mean μ_0 and variance σ_0^2

Associated Generative Model here: first draw μ from prior, then draw data given μ. Draw a picture

Goal: Estimate μ, given prior and data

What if we ignore the prior ? (ML estimation seen before)
What if we ignore the likelihood / data ? ($\mu = \mu_0$)
A possible solution: Maximize posterior w.r.t. μ

Posterior: $P(\mu|x_1, \cdots, x_N) = P(x_1, \cdots, x_N|\mu)P(\mu)/P(x_1, \cdots, x_N)$

Assume sample mean = $\bar{x}$

Then MAP estimate for the mean is :

$$\mu = \frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/N}{\sigma_0^2 + \sigma^2/N}$$

What if $N = 1$?
What if $N \to \infty$? (data dominates the prior)
What if $\sigma_0 \to \infty$? (weak prior: ignore the prior)
What if $\sigma_0 \to 0$? (strong prior: ignore the data)

14.6 Posterior Mean Estimate to Minimize MSE

Given data: $\{x_i\}_{i=1}^n$ drawn from $P(X|\theta)$

We have a prior $P(\theta)$ on RV θ

Posterior = conditional density $P(\theta|x_1, \cdots, x_n) = \frac{P(x_1,\cdots,x_n|\theta)P(\theta)}{\int_\theta P(x_1,\cdots,x_n,\theta)d\theta}$

Question: Given a PDF $P(\theta|x_1, \cdots, x_n)$ on the true parameter θ, what is the best estimate $\widehat{\theta}^*$ to minimize mean squared error $E_{P(\theta|x_1,\cdots,x_n)}[(\widehat{\theta}-\theta)^2]$?

Answer: The PDF mean $E_{P(\theta|x_1,\cdots,x_n)}[\theta]$. This is also a Bayes estimate.

14.7 Loss functions and Risk functions

Loss function $L(\widehat{\theta}|\theta)$:= loss incurred in obtaining the estimate as $\widehat{\theta}$, when the true value was θ.
We know that, given the data, the true value θ is distributed as per the posterior PDF $P(\theta|x_1, \cdots, x_n)$

Risk function $R(\widehat{\theta})$:= expected loss := expectation of the loss function $L(\widehat{\theta}|\theta)$ under the posterior PDF $P(\theta|x_1, \cdots, x_n)$

Goal: Choose $\widehat{\theta}$ to minimize risk

Example 1: Squared-error loss function: $L(\widehat{\theta}) = (\widehat{\theta}-\theta)^2$

Risk function $= E_{P(\theta|x_1,\cdots,x_n)}[(\widehat{\theta}-\theta)^2]$ = mean squared error

Let risk minimizer = θ^*

Then, $\frac{\partial}{\partial\widehat{\theta}} E_{P(\theta|x_1,\cdots,x_n)}[(\widehat{\theta}-\theta)^2]\Big|_{\widehat{\theta}=\theta^*} = 0$

Thus, $\theta^* = E_{P(\theta|x_1,\cdots,x_n)}[\theta] =$ Posterior mean

Example 2.1: Zero-one loss function (case of discrete RV θ): $L(\widehat{\theta}) = I(\widehat{\theta} \neq \theta)$

Risk function $= R(\widehat{\theta}) = E_{P(\theta|x_1,\cdots,x_n)}[I(\widehat{\theta} \neq \theta)]$

$= \sum_{\theta \neq \widehat{\theta}} P(\theta|x_1, \cdots, x_n)$
$= 1 - P(\theta = \widehat{\theta}|x_1, \cdots, x_n)$

Thus, the risk function is minimized when $\widehat{\theta} = \arg\max_\theta P(\theta|x_1, \cdots, x_n)$ = MAP estimate

Example 2.2: Zero-one loss function (case of continuous RV θ)

Assume that the loss function is an *inverted* rectangular pulse —_— with height 1 and an infinitesimally small width $\epsilon > 0$ (we do NOT make $\epsilon = 0$), with center of the pulse at the true parameter value θ. i.e.,
$L(\widehat{\theta}|\theta) = 0$; if $\widehat{\theta} \in (\theta - \epsilon/2, \theta + \epsilon/2)$
$L(\widehat{\theta}|\theta) = 1$; otherwise

For such a loss function, the risk function $1 - \int_{\widehat{\theta}-\epsilon/2}^{\widehat{\theta}+\epsilon/2} P(\theta|x_1, \cdots, x_n)d\theta$ is minimized when the pulse center is placed at the mode of the PDF.

Take the limit, as $\epsilon \to 0$, of $\arg\max_{\widehat{\theta}} \int_{\widehat{\theta}-\epsilon/2}^{\widehat{\theta}+\epsilon/2} P(\theta|x_1, \cdots, x_n)d\theta$

Draw a picture. Bimodal PDF. One peak is wide. Another peak is narrow.

Example 3: Absolute-error loss function $L(\widehat{\theta}) = |\widehat{\theta}-\theta|$

Risk function $= E_{P(\theta|x)}[|\widehat{\theta}-\theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta}-\theta| P(\theta|x)d\theta$
$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta}-\theta)P(\theta|x)d\theta + \int_{\widehat{\theta}}^{\infty} (\theta-\widehat{\theta})P(\theta|x)d\theta$

The risk function is minimized when its derivative is zero.

How to take the derivative of an integral where the limits are also a function of the variable of interest ?
Leibniz's Integral Rule (draw picture):
$\frac{\partial}{\partial a}\int_{l(a)}^{u(a)} f(z,a)dz = \int_{l(a)}^{u(a)} \frac{\partial f}{\partial a}dz + f(z=u(a),a)\frac{\partial u}{\partial a} - f(z=l(a),a)\frac{\partial l}{\partial a}$

In our case, $f(z \equiv \theta, a \equiv \widehat{\theta}) \propto (\widehat{\theta} - \theta)P(\theta|x)$
In our case, for the 1st integral: $f(z = u(a), a) = 0$ and the lower-limit term doesn't arise
In our case, for the 2nd integral: $f(z = l(a), a) = 0$ and the upper-limit term doesn't arise
Thus, the derivative of our risk function w.r.t. $\widehat{\theta}$ is:
$= \int_{-\infty}^{\widehat{\theta}} (+1)P(\theta|x)d\theta + \int_{\widehat{\theta}}^{\infty} (-1)P(\theta|x)d\theta$
$= \int_{-\infty}^{\widehat{\theta}} P(\theta|x)d\theta - \int_{\widehat{\theta}}^{\infty} P(\theta|x)d\theta$
This is zero when $\widehat{\theta}$ = median of $P(\theta|x)$

The median will be a minimizer if the 2nd derivative is positive. Is that so ?

In this case, for both integrals, $\frac{\partial f}{\partial a} = 0$
In this case, for 1st integral, the lower-limit term doesn't arise
In this case, for 2nd integral, the upper-limit term doesn't arise
Thus, the 2nd derivative of our risk function w.r.t. $\widehat{\theta}$, evaluated at $\widehat{\theta}$ = median of $P(\theta|x)$, is:
$= P(\widehat{\theta}|x) + P(\widehat{\theta}|x) \geq 0$

Note: the median $\widehat{\theta}$ isn't unique if $P(\widehat{\theta}|x) = 0$

14.8 Example: i.i.d. Bernoulli

Given: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter θ and PDF $P(x = 1|\theta) = \theta, P(x = 0|\theta) = 1 - \theta$

Data: $x_1, \cdots, x_n$

Estimate $\theta \in (0, 1)$

Prior $P(\theta) = 1, \forall\theta \in (0, 1)$

Answer:

Rewrite PDF as $P(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0, 1\}$

$P(\theta|x_1, \cdots, x_n) = P(x_1, \cdots, x_n|\theta)/P(x_1, \cdots, x_n)$
where
Numerator $= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$

If we want the posterior mean, then we need to care about the denominator as well
Denominator $= \int_0^1 \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}d\theta$
To handle the integral in the denominator, we exploit the result / trick: $\int_0^1 \theta^m(1-\theta)^r d\theta = m!r!/(m + r + 1)!$

Let $x = \sum_i x_i$
Then, $P(\theta|x_1, \cdots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$

Thus, $E_{P(\theta|x_1,\cdots,x_n)}[\theta] = \int_0^1 \theta \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}d\theta = \frac{x+1}{n+2}$

Thus, Bayes posterior-mean estimator $= \frac{\sum_i X_i + 1}{n+2}$

Note: ML estimator $= \max_\theta \log\left(\theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i}\right)$

$= \max_\theta X \log\theta + (n-X)\log(1-\theta)$, where $X := \sum_i X_i$
$= X/n$
$= \sum_i X_i/n$

Check that the 2nd derivative is negative (Use the facts: $X \geq 0$ and $n - X \geq 0$ and $0 < \theta < 1$)

Note: In this case, ML estimator $\equiv$ MAP estimator; because $P(\theta) = 1$

Note: When $n = 0$, Bayes estimate $= 0.5$, the mid-point of the interval $(0, 1)$. This is what we get when we solely rely on the prior

Note: Asymptotically, i.e., as $n \to \infty$, the Bayes estimator tends to the ML estimator

What happens to the Bayes estimate and ML estimate when true $\theta = 0$ or true $\theta = 1$? Assume n is large.

14.9 Example: i.i.d. Gaussian

Given: $X_1, \cdots, X_n$ i.i.d. $G(\theta, \sigma_0^2)$. Unknown mean. Known variance.
Prior: $P(\theta) := G(\theta; \mu; \sigma^2)$
Bayes posterior-mean estimate = ?

Answer:

Property 1: Product of 2 Gaussians is another Gaussian: $G(z;\mu_1,\sigma_1^2)G(z;\mu_2,\sigma_2^2) \propto G(z;\mu_3,\sigma_3^2)$

Numerator exponent $= \frac{(z-\mu_1)^2}{2\sigma_1^2} + \frac{(z-\mu_2)^2}{2\sigma_2^2}$
$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_2^2\right)$
$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z\right) + c$, where $c =$ constant independent of z
$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - \frac{2\mu_1\sigma_2^2+2\mu_2\sigma_1^2}{\sigma_2^2+\sigma_1^2}z\right) + c$, where $c =$ constant independent of z
$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - 2\mu_3 z + \mu_3^2\right) + c'$, where $c' =$ constant independent of z and where $\mu_3 = \frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}$
$= \frac{1}{2\sigma_3^2}(z-\mu_3)^2 + c'$, where $c' =$ constant independent of z where $\sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$

In our case, we have two PDFs on θ, i.e.,

Prior $P(\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp((\theta-\mu)^2/(2\sigma^2)) = G(\theta;\mu,\sigma^2)$

Likelihood $P(x_1,\cdots,x_n|\theta) = \frac{1}{(2\pi)^{n/2}\sigma_0^n}\exp(-\sum_i(x_i-\theta)^2/(2\sigma_0^2)) = G(\theta;x_1,\sigma_0^2)\cdots G(\theta;x_n,\sigma_0^2)$

The negative exponent here can be written as:
$(n\theta^2 - 2(\sum_i x_i)\theta)/(2\sigma_0^2) + c$, where $c =$ constant independent of θ
$= (\theta^2 - 2(\sum_i x_i/n)\theta)/(2\sigma_0^2/n) + c$
$\propto G(\theta; \sum_i x_i/n, \sigma_0^2/n)$

Let $x = \sum_i x_i/n$

Thus, the (normalized) product of the prior and the likelihood gives a Gaussian $G(\theta;\mu^*,\sigma^{*2})$, where
$\mu^* = \frac{\mu\sigma_0^2/n + x\sigma^2}{\sigma^2+\sigma_0^2/n}, \sigma^{*2} = \frac{\sigma^2\sigma_0^2/n}{\sigma^2+\sigma_0^2/n}$

Bayes estimate = mean of posterior = μ^*, which also happens to be the Gaussian posterior's mode = MAP estimate

Note: As the data sample size $n \to \infty$, the mean $\mu^* \to x$ and variance $\sigma^{*2} \to 0$.
Thus, the posterior becomes a delta function at $\theta = x =$ sample mean
In this case, the Bayes estimate converges to the ML estimate = sample mean

14.10 MAP Estimation and ML Estimation

Consider the likelihood function $P(x_1, \cdots, x_n|\theta)$

Consider prior $P(\theta) = 1/(b-a)$ for $\theta \in (a,b)$, i.e., a uniform distribution over (a,b)

Then, posterior PDF $= \frac{P(x_1,\cdots,x_n|\theta)P(\theta)}{\int_a^b P(x_1,\cdots,x_n|\theta)P(\theta)d\theta}$, for $\theta \in (a,b)$
$= \frac{P(x_1,\cdots,x_n|\theta)}{\int_a^b P(x_1,\cdots,x_n|\theta)d\theta}$, for $\theta \in (a,b)$

Maximum of the posterior within (a,b) = maximum of $P(x_1, \cdots, x_n|\theta)$ within (a,b)
If the mode of the likelihood function lied within (a,b), then the mode of the posterior $\equiv$ ML estimate

14.11 Bayes Interval Estimate

Previous analysis gives a point estimate for the parameter θ
How do we get an interval estimate for the parameter θ ?
We can do this by finding a, b such that $\int_a^b P(\theta|x_1, \cdots, x_n)d\theta = 1-\alpha$, where probability α is given.
We can get such information in some special cases, relatively easily

14.11.1 Example: Gaussian

Question: Suppose signal of value s is sent from A to B.
Because of the noisy communication channel, signal received at B has a Gaussian PDF with mean s and variance 60.
A priori, it is known that the signal s being sent is selected from a Gaussian PDF with mean 50 and variance 100.
Given: Value received at B is 40.
Find an interval (a,b) s.t. the probability of the signal being in that interval is 0.9

Answer:
Using formulas derived before for the posterior $P(s|x_1 = 40)$ of parameter s given data x_1,
Posterior mean = $\frac{50*60+40*100}{60+100} = 43.75$
Posterior variance = $\frac{60*100}{60+100} = 37.5$

We know that the posterior PDF is Gaussian
Thus, $Z := \frac{S-43.75}{\sqrt{37.5}}$ has a standard Normal PDF
For a standard Normal PDF, we know that the probability mass within $Z \in (-1.645, +1.645)$ is 0.9

Thus, we want to find S s.t. $P(-1.645 < Z < 1.645|\text{data}) = 0.9$
i.e., $P(-1.645 < \frac{S-43.75}{\sqrt{37.5}} < 1.645|\text{data}) = 0.9$
i.e., $P(33.68 < S < 53.83|\text{data}) = 0.9$

Thus, the desired interval is $(a = 33.68, b = 53.83)$

14.12 Conjugate Priors

If the posterior PDFs $P(\theta|x)$ are in the same family as the prior PDF $P(\theta)$, then:
(i) the prior and posterior are called *conjugate* PDFs, and
(ii) the prior is called the conjugate prior for the likelihood function

Advantage of conjugate priors: The posterior has a closed-form expression because the denominator / normalizing constant has a closed-form expression
$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$
Otherwise, a difficult numerical integration may be required to approximate the normalization factor

Example: Binomial Likelihood and Beta prior

1) Likelihood of s successes in n tries: $P(s,n|\theta) = ^n C_s \theta^s (1-\theta)^{n-s}$, where $n \in \mathbb{N}$, $s \in \mathbb{I}_{\geq 0}$
2) Prior: $P(\theta) =$ beta$(\theta; a \in \mathbb{R}^+, b \in \mathbb{R}^+) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$, Note: $a > 0, b > 0$
3) Posterior $\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1} \equiv$ beta$(\theta; a+s, b+n-s)$

• We know that the **mean** of the beta PDF beta$(\theta; a, b)$ is $a/(a+b)$

Thus, Bayes estimate = posterior mean = $(a+s)/(a+b+n)$
$= w(a/(a+b)) + (1-w)(s/n)$, where weight $w = (a+b)/(a+b+n)$

Note: When the sample size $n = 0$, the posterior mean $= a/(a+b) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

If prior $P(\theta) = 1$ is uniform over $\theta \in (0,1)$, i.e., beta$(\theta, 1, 1)$
In that case, the likelihood determines the posterior

• We know that the **mode** of the beta PDF beta$(\theta; a, b)$ is $(a-1)/(a+b-2)$ for $a, b > 1$

So, posterior mode $= (a+s-1)/(a+b+n-2)$
$= w((a-1)/(a+b-2)) + (1-w)(s/n)$, where weight $w = (a+b-2)/(a+b+n-2)$

Note: When the sample size $n = 0$, the posterior mode $= (a-1)/(a+b-2) =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate

Example: Gaussian (known mean μ, unknown variance θ) and Inverse Gamma

1) Likelihood: $P(x_1, \cdots, x_n|\mu, \theta) \propto \prod_{i=1}^n \theta^{-0.5} \exp(-0.5(x_i - \mu)^2/\theta)$
2) Prior = Inverse Gamma PDF: $P(\theta; a, b) \propto \theta^{-a-1} \exp(-b/\theta)$, where $a > 0, b > 0$
3) Posterior = Inverse Gamma PDF: $P(\theta; a + n/2, b + \sum_i (x_i - \mu)^2/2)$

• **Mean** of the inverse Gamma $P(\theta; a, b) = b/(a-1)$, for $a > 1$

Thus, Bayes estimate = posterior mean = $(b + \sum_i (x_i - \mu)^2/2)/(a + n/2 - 1)$
$= (2b + \sum_i (x_i - \mu)^2)/(2a + n - 2)$
$= w(b/(a-1)) + (1-w)\sum_i (x_i - \mu)^2/n$, where weight $w = (2a-2)/(2a+n-2)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= b/(a-1) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

• **Mode** of the inverse Gamma $P(\theta; a, b) = b/(a+1)$

So, posterior mode = $(b + \sum_i (x_i - \mu)^2/2)/(a + n/2 + 1)$
$= (2b + \sum_i (x_i - \mu)^2)/(2a + n + 2)$
$= w(b/(a+1)) + (1-w)\sum_i (x_i - \mu)^2/n$, where weight $w = (2a+2)/(2a+n+2)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mode $= b/(a+1) =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate

An *"uninformative"* (misnomer) prior for the Gaussian *mean* is the (improper) uniform PDF

Why improper ? Because it doesn't integrate to a finite number

Why uninformative ? Because:
i) posterior PDF driven by the likelihood function alone
ii) the prior on θ is invariant to any change in the true θ, which could cause translation of the data x_i (Duda-Hart-Stork).
Note: translation of data also implies that the MLE estimate of the mean also gets translated.

Uninformative priors express "objective" (impersonal; unaffected by personal beliefs) information such as "the variable is positive" or "the variable is less than some limit".
Uninformative priors yield results *close to* what we would get with non-Bayesian (e.g., ML) analysis

An *"uninformative"* (and improper) prior for the Gaussian *standard deviation* σ is $P(\sigma) = 1/\sigma$

Why uninformative ? Because of scale invariance, as follows.
Assume data x comes from a Gaussian with mean zero. Consider the RVs $\log(X)$ and $\log(\sigma)$. If the data x get scaled (which implies that the MLE for the standard deviation σ also gets scaled) in the original domain by factor a, then a term $\log(a)$ gets added in the log domain. Scale-invariant prior on $\sigma \to$ translation-invariant prior on $\log(\sigma) \to$ uniform PDF on $\log(\sigma)$.
Transform the RV $U := \log(\Sigma)$ with $P(U) = c$, to get the RV $V := \exp(U)$. Transformation of variables implies that $P(v) = c/v$.

Same analysis applies to the Gaussian variance.
The uninformative prior for the Gaussian variance θ is the inverse Gamma PDF with parameters $a = b \to 0$, which implies $P(\theta) \propto 1/\theta$ where $\theta = \sigma^2$. This is an improper PDF.

Example: Poisson PDF and Gamma prior

Use this example to motivate the general result for exponential families later
1) Likelihood: $P(k_1, \cdots, k_n|\lambda) = \prod_i \lambda^{k_i} \exp(-\lambda)/k_i!$, where $\lambda \in \mathbb{R}^+, k_i \in \mathbb{I}^+$
2) Prior: $P(\theta) =$ Gamma$(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$, where $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+, \lambda \in \mathbb{R}^+$
3) Posterior: $\propto \lambda^{\sum_i k_i + \alpha - 1} \exp(-n\lambda - \beta\lambda) \equiv$ Gamma$(\lambda; \sum_i k_i + \alpha, n + \beta)$

• For a Gamma distribution Gamma$(\lambda|\alpha, \beta)$, we know that the **mean** is α/β

Thus, the Bayes estimate = posterior mean = $(\sum_i k_i + \alpha)/(n + \beta)$
$= w(\alpha/\beta) + (1 - w) \sum_i k_i/n$, where weight $w = \beta/(\beta + n)$
$= w(\alpha/\beta) + (1 - w)\widehat{\lambda}_{\text{MLE}}$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= \alpha/\beta =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

• For a Gamma distribution Gamma$(\lambda|\alpha, \beta)$, we know that the **mode** is $(\alpha - 1)/\beta$ when $\alpha \geq 1$. When $\alpha < 1$, the case is tricky.

Then, posterior mode = $(\sum_i k_i + \alpha - 1)/(n + \beta)$
$= w((\alpha - 1)/\beta) + (1 - w) \sum_i k_i/n$, where weight $w = \beta/(\beta + n)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mode $= (\alpha - 1)/\beta =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate

14.13 Exponential Family of PDFs

Interesting result: PDFs in the exponential family (typically) have conjugate priors.

Definition: A **single-parameter** exponential family is a set of PDFs where each PDF can be expressed in the form:

$P(x|\theta) = \exp\left[\eta(\theta)T(x) - A(\theta) + B(x)\right] = g(\theta)h(x)\exp[\eta(\theta)T(x)]$
where $T(x), B(x), \eta(\theta), A(\theta)$ are known functions
and
the support of the distribution cannot depend on θ.

So, uniform distribution isn't in this family.

Interpretation: The parameters θ and observation variables x must *factorize* either directly or within either part of an exponential operation

Consider the *canonical form* of the exponential family where $\eta(\theta) := \theta$, i.e., $\eta(\cdot)$ is identity
Note: It is always possible to convert an exponential family to canonical form, by defining a transformed parameter $\theta' = \eta(\theta)$

Example: Bernoulli

$P(X = x; \theta) = \theta^x(1-\theta)^{1-x} = \exp(x \log\theta + (1-x)\log(1-\theta)) = \exp(x\log(\theta/(1-\theta)) + \log(1-\theta))$

$\eta = \log(\theta/(1-\theta)$
$T(x) = x$
$g(\eta) = \exp(\log(1-\theta)) = (1-\theta)$
$h(x) = 1$

Example: Poisson

$P(X = x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!} = \exp(-\lambda)(1/x!)\exp[x\log\lambda]$

$\eta = \log\lambda$
$T(x) = x$
$g(\eta) = \exp(-\lambda)$
$h(x) = 1/x!$

Definition: A **multi-parameter** exponential family is a set of PDFs where each PDF can be expressed in the form:
$P(x|\eta) = \exp\left[\eta^\top T(x) - A(\eta) + B(x)\right]$
where $T(x), B(x), A(\eta)$ are known functions.

Example: Gaussian

$P(X = x; \mu, \sigma^2) = (1/\sigma)(1/\sqrt{2\pi})\exp(-0.5x^2/\sigma^2 + \mu x/\sigma^2 - 0.5\mu^2/\sigma^2)$

$\eta = [-0.5/\sigma^2, \mu/\sigma^2]^\top$
$T(x) = [x^2, x]^\top$
$g(\eta) = (1/\sigma)\exp(-0.5\mu^2/\sigma^2)$
$h(x) = (1/\sqrt{2\pi})$

Some Properties:

(1) The random variable $T(x)$ is sufficient for parameter θ

$T(X)$ is a function of data only; not any parameter.

Sufficient Statistic: Statistic $T(X)$ is sufficient for parameter θ if there isn't any information in X regarding θ beyond that in $T(X)$.
If our goal is to estimate θ, all we need is $T(X)$ and X can be discarded.

(2) If **i.i.d.** RVs $\{X_i\}$ are from the one-parameter exponential family, then their joint PDF is also from the one-parameter exponential family (with sufficient statistic $\sum_i T(X_i)$).

The joint PDF is $P(x_1, x_2, \cdots, x_N|\theta) = \left(\prod_{n=1}^N h(x_n)\right) \exp\left(\eta^\top \sum_{n=1}^N T(x_i) - NA(\eta)\right)$

For i.i.d. observations from (i) Bernoulli PMF or (ii) Poisson PDF, sufficient statistic for parameter is the sum $\sum_n x_n$

For i.i.d. observations from (i) Gaussian PDF, sufficient statistic for parameter is the vector sum $[\sum_n x_n^2, \sum_n x_n]$

What other PDFs aren't in the exponential family ?

$P(x|\theta) = [f(x)g(\theta)]^{h(x)j(\theta)} = \exp([h(x)\log f(x)]j(\theta) + h(x)[j(\theta)\log g(\theta)])$

Laplace / Double-Exponential PDF: $P(x|\theta) := 0.5\exp(-|x-\theta|)$ (Proof is non-trivial)

• How do we go about guessing what the conjugate prior is ?

Step (1) For the exponential family, the likelihood function for data $\{x_i\}_{i=1}^N$ is:
$L(\theta|x_1, \cdots, x_N) = (\Pi_i \exp(B(x_i)))\exp(\theta(\sum_i T(x_i)) - NA(\theta))$

Step (2) Consider the prior $P(\theta|\alpha,\beta) = H(\alpha,\beta)\exp(\alpha\theta - \beta A(\theta))$

Diaconis and Ylvisaker 1979 gave conditions on the hyper-parameters α, β under which this PDF is integrable (i.e., proper)

Step (3) The posterior PDF $\propto \exp(\theta(\alpha + \sum_i T(x_i)) - (\beta+N)A(\theta))$ that belongs to the exponential family w.r.t. variable θ and has the same form as the prior
The conversion from the prior to the posterior simply replaces $\alpha \to \alpha + \sum_i T(x_i)$ and $\beta \to \beta + N$

Because the prior can be normalized, so can the posterior

14.14 Kullback-Leibler Divergence / Dissimilarity

Continuous RVs: $D(P(X|\theta_1), Q(X|\theta_2)) := \int_x P(x|\theta_1)\log\frac{P(x|\theta_1)}{Q(x|\theta_2)}dx$

Discrete RVs: $D(P(X|\theta_1), Q(X|\theta_2)) := \sum_x P(x|\theta_1)\log\frac{P(x|\theta_1)}{Q(x|\theta_2)}$

Defined only under the following condition: $Q(x) = 0$ implies $P(x) = 0$

When $P(x) \to 0$ and $Q(x) > 0$, the contribution of the x-th term is zero because $\lim_{P(x)\to 0} P(x)\log P(x) = 0$

When $P(x) \to 0$ and $Q(x) \to 0$, we use the convention / *interpretation* that $0\log\frac{0}{0} = 0$; Cover and Thomas (2nd Ed.). Basically, ignore such cases. Can see this as an outcome of regularization: (i) Bayesian prior or (ii) convex combination of each of the given PDFs $P(X)$ and $Q(X)$ with uniform PDF $U(X)$).

Properties:
1) When PMFs / PDFs $P(X)$ and $Q(X)$ are identical (almost everywhere; in the continuous case), then $D(P,Q) = 0$
2) $D(P,Q) \geq 0$, for all P, Q

For discrete PMFs, this inequality is known as the Gibbs' inequality
Proof (discrete case):
We know that $\log x \leq x - 1$
So, $-\log x \geq -(x-1)$
$\sum_{x|P(x)>0} P(x)\log\frac{P(x)}{Q(x)}$
$= -\sum_{x|P(x)>0} P(x)\log\frac{Q(x)}{P(x)}$

$\geq -\sum_{x|P(x)>0} P(x)(\frac{Q(x)}{P(x)} - 1)$
$= -\sum_{x|P(x)>0} Q(x) + \sum_{x|P(x)>0} P(x)$
$= -\sum_{x|P(x)>0} Q(x) + 1$
≥ 0

So, $\sum_{x|P(x)>0} P(x) \log P(x) \geq \sum_{x|P(x)>0} P(x) \log Q(x)$

If we extend the summation to all remaining x', then the LHS stays the same (because $\lim_{P(x')\to 0} P(x') \log P(x') = 0$) and the RHS also stays the same (because $P(x') = 0$)
Thus, $\sum_x P(x) \log P(x) \geq \sum_x P(x) \log Q(x)$

Thus, $D(P||Q) \geq 0$

When is $D(P||Q) = 0$?
For this to happen, Condition 1: $P(x) = Q(x), \forall x : P(x) > 0$, i.e., when $\log \frac{P(x)}{Q(x)} = 0 = \frac{P(x)}{Q(x)} - 1$ making the first inequality as an equality
The second inequality becomes an inequality when $\sum_{x:P(x)>0} Q(x) = 1$
Alternatively, because $\sum_{x:P(x)>0} P(x) = 1$, and $P(x) = Q(x)$ on this domain, we have $\sum_{x:P(x)>0} Q(x)$ also $= 1$
Thus, for all $x : P(x) = 0$, we have $Q(x)$ also $= 0$
Thus, $P(x) = Q(x), \forall x$

For continuous PMFs, the proof uses Jensen's inequality.

Jensen's inequality: If $f(\cdot)$ is a convex function and X is a random variable, then $E[f(X)] \geq f(E[X])$

Proof of Jensen's inequality:
Let $\mu := E[X]$
Draw a line tangent to the convex function $f(X)$, touching it at $(\mu, f(\mu))$
The tangent, say, $aX + b$ lies below the function $f(X), \forall X$
LHS = $E[f(X)] \geq E[aX + b] = a\mu + b = f(\mu) =$ RHS

Another variant of Jensen's Inequality:
$E_{P(X)}[f(g(X))] \geq f(E_{P(X)}[g(X)])$, when $f(\cdot)$ is convex and $g(\cdot)$ can be any function.

Proof: LHS
$= \sum_{i=1}^n P(x_i) f(g(x_i)) = P(x_n) f(g(x_n)) + (1 - P(x_n)) \sum_{i=1}^{n-1} P'(x_i) f(g(x_i))$, where $P'(x_i) := P(x_i)/(1 - P(x_n))$
$\geq P(x_n) f(g(x_n)) + (1 - P(x_n)) f(\sum_{i=1}^{n-1} P'(x_i) g(x_i))$ (because of the induction hypothesis)
$\geq f\left(P(x_n) g(x_n) + (1 - P(x_n)) \sum_{i=1}^{n-1} P'(x_i) g(x_i)\right)$ (because of the definition of convexity of $f(\cdot)$)
$= f\left(\sum_{i=1}^n P(x_i) g(x_i)\right)$
This proof extends to the continuous case.

Proof of KL Divergence being non-negative (continuous case):
$D(P||Q) = E_{P(X)}[\log(P(X)/Q(X))] = E_{P(X)}[-\log(Q(X)/P(X))]$
Take $f(\cdot) := -\log(\cdot)$ as the convex function
Take $g(X) := Q(X)/P(X)$
Then, $D(P||Q) \geq -\log E_{P(X)}[Q(X)/P(X)] = -\log 1 = 0$

KL-Divergence Property: $D(\cdot, \cdot)$ is asymmetric. Not a "distance metric".

14.15 KL Divergence and MLE

Empirical Estimate of PMF / PDF of data: $\widehat{P}(X = x) := \frac{1}{N} \sum_{n=1}^N \delta(x; x_n)$

Discrete RV: $\delta(x; x_n)$ is the Kronecker delta function

Continuous RV: $\delta(x; x_n)$ is the Dirac delta function(al)

For Discrete RV, KL divergence between empirical PDF and actual PDF:
$D(\widehat{P}(X), P(X|\theta))$
$= \sum_x \widehat{P}(x) \log \widehat{P}(x) - \sum_x \widehat{P}(x) \log P(x|\theta)$
$= \sum_x \widehat{P}(x) \log \widehat{P}(x) - \sum_x (1/N) \sum_n \delta(x; x_n) \log P(x|\theta)$
$= \sum_x \widehat{P}(x) \log \widehat{P}(x) - (1/N) \sum_n \sum_x \delta(x; x_n) \log P(x|\theta)$
$= \sum_x \widehat{P}(x) \log \widehat{P}(x) - (1/N) \sum_n \log P(x_n|\theta)$
where the second term is the average log-likelihood function

Thus, minimizing this KL divergence is the same as maximizing the likelihood function

For Continuous RV, KL divergence between empirical PDF and actual PDF:
$D(\widehat{P}(X), P(X|\theta))$
$= \int_x \widehat{P}(x) \log \widehat{P}(x) dx - \int_x \widehat{P}(x) \log P(x|\theta) dx$
$= \int_x \widehat{P}(x) \log \widehat{P}(x) - \int_x (1/N) \sum_n \delta(x; x_n) \log P(x|\theta) dx$
$= \int_x \widehat{P}(x) \log \widehat{P}(x) - (1/N) \sum_n \int_x \delta(x; x_n) \log P(x|\theta) dx$
$= \int_x \widehat{P}(x) \log \widehat{P}(x) - (1/N) \sum_n \log P(x_n|\theta)$
where the second term is the average log-likelihood function

Thus, minimizing this KL divergence is the same as maximizing the likelihood function

14.16 Fisher Information

Key Question: How much information can a sample of data provide about the unknown parameter ?

(1) If likelihood function $P(\text{data}|\theta)$ is sharply peaked w.r.t. Δ changes in θ around $\theta = \theta_{\text{true}}$, it is easy to estimate θ_{true} from the given data sample of size N.

Example 1: Bernoulli RV with θ close (equal) to 0 or 1

Example 2: Estimating Gaussian mean $\theta := \mu$ in two cases: (i) when variance σ^2 (known) is huge, (ii) when σ^2 is tiny.
Data drawn from $G(x; \mu, \sigma^2)$ in 2nd case has a smaller spread.
Likelihood in 2nd case more peaked.
For a small sample of size N (say, N = 5), mean estimate (sample mean; always unbiased = always high accuracy) is much more precise (= much lower variance) in 2nd case

(2) If likelihood function $P(\text{data}|\theta)$ has a large spread w.r.t. changes in θ around θ_{true}, it will take very many N-sized data samples to get the ML estimate of θ to be at / close to θ_{true}

First, consider the average (expected) derivative of the log-likelihood function:
$E_{P(X|\theta_{\text{true}})}[\frac{\partial}{\partial\theta} \log P(X|\theta)\big|_{\theta=\theta_{\text{true}}}]$
$= \int_x P(x|\theta) \frac{\partial P(x|\theta)}{\partial\theta} / P(x|\theta) dx$
$= \int_x \frac{\partial}{\partial\theta} P(x|\theta) dx$
$= \frac{\partial}{\partial\theta} \int_x P(x|\theta) dx$
$= \frac{\partial}{\partial\theta} 1$
$= 0$
The expectation / integral isn't over θ, but over different instances of observed data $x \sim P(X|\theta_{\text{true}})$
The expectation is zero for all θ_{true}
Now, consider the expected squared slope (slope variance) of the log-likelihood function $\log P(X|\theta)$, evaluated at $\theta = \theta_{\text{true}}$, i.e.,
$I(\theta_{\text{true}}) := E_{P(X|\theta_{\text{true}})}[\Big(\frac{\partial}{\partial\theta} \log P(X|\theta)\big|_{\theta_{\text{true}}}\Big)^2]$

The Fisher information $I(\theta_{\text{true}}) \geq 0$

If $\log P(X|\theta)$ didn't contain θ, then the derivative would be 0, and the data wouldn't contain any information about θ

There is another way to look at Fisher information.

$$\text{Consider } \frac{\partial^2}{\partial \theta^2} \log P(X|\theta) = \frac{\frac{\partial^2 P(X|\theta)}{\partial \theta^2}}{P(X|\theta)} - \left(\frac{\frac{\partial P(X|\theta)}{\partial \theta}}{P(X|\theta)} \right)^2 = \frac{\frac{\partial^2 P(X|\theta)}{\partial \theta^2}}{P(X|\theta)} - \left(\frac{\partial \log P(X|\theta)}{\partial \theta} \right)^2 \tag{4}$$

Now, (i) evaluate LHS and RHS at $\theta := \theta_{\text{true}}$ and (ii) take expectation w.r.t. $P(X|\theta_{\text{true}})$:

$$E_{P(X|\theta_{\text{true}})} \left[\frac{\partial^2}{\partial \theta^2} \log P(X|\theta) \Big|_{\theta=\theta_{\text{true}}} \right] = E_{P(X|\theta_{\text{true}})} \left[\frac{\frac{\partial^2 P(X|\theta)}{\partial \theta^2}}{P(X|\theta)} \Big|_{\theta=\theta_{\text{true}}} \right] - I(\theta) = -I(\theta), \text{ because} \tag{5}$$

$$E_{P(X|\theta_{\text{true}})} \left[\frac{\frac{\partial^2 P(X|\theta)}{\partial \theta^2}}{P(X|\theta)} \Big|_{\theta=\theta_{\text{true}}} \right] = \int_x \frac{\partial^2 P(x|\theta)}{\partial \theta^2} dx = \frac{\partial^2}{\partial \theta^2} \int_x P(X|\theta) dx = 0 \tag{6}$$

So, Fisher information is the expectation (over $x \sim P(X|\theta_{\text{true}})$) of the negative 2nd-derivative (curvature) of the log-likelihood function $\log P(x|\theta)$ evaluated at $\theta = \theta_{\text{true}}$
So, larger Fisher information means the log-likelihood function $\log P(x|\theta)$ is expected to be more concave and more curved at $\theta = \theta_{\text{true}}$

Example: Bernoulli RV

$\log P(x|\theta) = x \log \theta + (1-x) \log(1-\theta)$
$\frac{\partial}{\partial \theta} \log P(x|\theta) = x/\theta - (1-x)/(1-\theta)$
$\frac{\partial^2}{\partial \theta^2} \log P(x|\theta) = -x/\theta^2 - (1-x)/(1-\theta)^2$
$I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} \log P(x|\theta)] = \theta/\theta^2 + (1-\theta)/(1-\theta)^2 = 1/(\theta(1-\theta))$
So, $I(\theta)$ is large when θ close to 0 or 1

For a dataset of size N, $I_N(\theta) = N/(\theta(1-\theta))$

Example: Gaussian RV

Unknown mean parameter $\theta = \mu$. Known variance σ^2.
$\frac{\partial}{\partial \mu} \log P(x|\mu) = (x-\mu)/\sigma^2$
$\frac{\partial^2}{\partial \mu^2} \log P(x|\mu) = -1/\sigma^2$
$I(\mu) = 1/\sigma^2$
Here, $I(\mu)$ is independent of μ, but rather depends on the other parameter σ^2

For a dataset of size N, $I_N(\mu) = N/\sigma^2$

14.17 Cramer-Rao Lower Bound

Let RV X model a dataset.

Assumption: Consider an **unbiased** estimator $\widehat{\theta}(X)$

Then, $E_{P(X|\theta_{\text{true}})}[\widehat{\theta}(X) - \theta_{\text{true}}] = 0 = \left(\int_x P(x|\theta)[\widehat{\theta}(x) - \theta] dx \right) \Big|_{\theta=\theta_{\text{true}}}$

This holds for all θ_{true}.
That is, $\int_x P(x|\theta')[\widehat{\theta}(x) - \theta'] dx$ is a function of θ' that is identically zero. So, its derivative is also identically zero.

Thus, $0 = \frac{\partial}{\partial \theta} \left(\int_x P(x|\theta)[\widehat{\theta}(x) - \theta] dx \right) \Big|_{\theta=\theta_{\text{true}}}$

For convenience, lets call θ_{true} as θ

Thus, $\int_x [\widehat{\theta}(x) - \theta] \frac{\partial}{\partial \theta} P(x|\theta) dx = \int_x P(x|\theta) dx = 1$

Thus, $1 = \int_x [\widehat{\theta}(x) - \theta] P(x|\theta) \frac{\partial}{\partial \theta} \log P(x|\theta) dx$
Thus, $1 = \int_x \left([\widehat{\theta}(x) - \theta] \sqrt{P(x|\theta)} \right) \left(\sqrt{P(x|\theta)} \frac{\partial}{\partial \theta} \log P(x|\theta) \right) dx$
Thus, $1 = \left[\int_x \left([\widehat{\theta}(x) - \theta] \sqrt{P(x|\theta)} \right) \left(\sqrt{P(x|\theta)} \frac{\partial}{\partial \theta} \log P(x|\theta) \right) dx \right]^2$

Using Cauchy-Schwarz inequality, $1 \leq \int_x [\widehat{\theta}(x) - \theta]^2 P(x|\theta) dx \cdot \int_x P(x|\theta) \left(\frac{\partial}{\partial \theta} \log P(x|\theta) \right)^2 dx$

Thus, Var$(\widehat{\theta}(X)) \geq I(\theta)^{-1}$

For i.i.d. Gaussian RVs, any estimator of the unknown mean (known variance) will have variance $\geq \sigma^2/n$.
We know that the ML estimator's variance $= \sigma^2/n$.
Thus, this ML estimator is an *efficient* estimator / *minimum variance unbiased estimator*.

Bayesian estimation can lead to lower mean squared error, for finite data, at the cost of introudcing a bias in the estimator (vis-a-vis unbiased ML estimator).

Let $X \sim$ Binomial(n, θ), i.e., each try is Bernoulli with probability of success θ
* MLE estimator (unbiased): $\widehat{\theta}_{\text{MLE}}(\theta) := X/n$
* MLE estimator's variance: $=$ Var$(X/n) = \theta(1-\theta)/n$
Consider prior Beta$(a = 1, b = 1)$ on θ, as before.
* Bayes mean estimator: $\widehat{\theta}_{\text{Bayes}}(\theta) := (X+1)/(n+2) = w(X/n) + (1-w)0.5$
* Bias of Bayes mean estimator: $(n\theta + 1)/(n+2) - \theta = (1-w)(0.5 - \theta)$
* Variance of Bayes estimator: $=$ Var$(X)/(n+2)^2 = (\theta(1-\theta)/n) * (1/(n+2)^2) = w^2 \theta(1-\theta)/n$
MSE = Bias^2 + Variance
MSE of MLE estimator is mostly (i.e., for most values of $\theta \in (0,1)$) greater than the MSE of Bayes estimator. Plot.

14.18 Bayesian Cramer-Rao Lower Bound

Applications of the van Trees Inequality: A Bayesian Cramer-Rao Bound
Bernoulli 1995, `https://www.jstor.org/stable/3318681`

Let X model a dataset.
Consider likelihood $P(X|\theta)$ with "parameter" / RV θ
Consider a prior PDF $Q(\theta|\alpha)$ on "parameter" / RV θ with hyper-parameter α

$E_{Q(\theta|\alpha)}[E_{P(X|\theta)}[\widehat{\theta}(X) - \theta]^2]$

$\geq$

$(E_{Q(\theta|\alpha)}[I_P(\theta)] + J_Q(\theta))^{-1}$

where
$I_P(\theta)$ is the Fisher information of the likelihood associated with PDF / model $P(X|\theta)$, and
$J_Q(\theta)$ is the "prior information" of the prior PDF / model $Q(\theta|\alpha)$

Unlike the CRLB, the Bayesian-CRLB gives us a lower bound for all (biased and unbiased both) estimators.

Assumption: Consider the prior θ defined on (compact) interval (a, b) such that:
$Q(\theta|\alpha) \to 0$ as $\theta \to a$ and as $\theta \to b$

Then, similar to our strategy in proving CRLB, lets consider

$\int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right) \frac{\partial}{\partial\theta} \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$
$= \int_x \int_\theta \widehat{\theta}(x) \frac{\partial}{\partial\theta} \left(P(x|\theta)Q(\theta|\alpha)\right) d\theta dx - \int_x \int_\theta \theta \frac{\partial}{\partial\theta} \left(P(x|\theta)Q(\theta|\alpha)\right) d\theta dx$

1st term includes the inner integral:

$\int_\theta \widehat{\theta}(x) \frac{\partial}{\partial\theta} [P(x|\theta)Q(\theta|\alpha)] d\theta$
$= \widehat{\theta}(x) \int_\theta \frac{\partial}{\partial\theta} [P(x|\theta)Q(\theta|\alpha)] d\theta$
$= \widehat{\theta}(x) [P(x|\theta)Q(\theta|\alpha)]_a^b$
$= 0$, because the prior $Q(\theta|\alpha)$ goes to zero at the boundary points a and b
So, the 1st term reduces to zero

2nd term (without the negative sign) includes an inner integral:

$\int_\theta \theta \frac{\partial}{\partial\theta} \left[P(x|\theta)Q(\theta|\alpha)\right] d\theta = [\theta P(x|\theta)Q(\theta|\alpha)]_a^b - \int_\theta P(x|\theta)Q(\theta|\alpha) d\theta$
$= 0 - \int_\theta P(x|\theta)Q(\theta|\alpha) d\theta$

So, 2nd term (with the negative sign) equals:
$\int_x \int_\theta P(x|\theta)Q(\theta|\alpha) d\theta dx$
$= \int_\theta Q(\theta|\alpha) \left(\int_x P(x|\theta) dx\right) d\theta$
$= 1$

So, our original term equals 1:
$1 = \int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right) \frac{\partial}{\partial\theta} \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$
$= \int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right) P(x|\theta)Q(\theta|\alpha) \frac{1}{P(x|\theta)Q(\theta|\alpha)} \frac{\partial}{\partial\theta} \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$
$= \int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right) \sqrt{P(x|\theta)Q(\theta|\alpha)} \sqrt{P(x|\theta)Q(\theta|\alpha)} \frac{\partial}{\partial\theta} \log\left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$

Now, we apply the Cauchy-Schwarz inequality:

$1 \leq \int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right)^2 P(x|\theta)Q(\theta|\alpha) dx d\theta \cdot \int_{\theta=a}^{b} \int_x P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)Q(\theta|\alpha)\right]^2 dx d\theta$

where

1st integral $=$ expected squared error (NOT variance; because bias of estimator $\widehat{\theta}(x)$ may be non-zero)

2nd integral:

$= \int_{\theta=a}^{b} \int_x P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)\right]^2 dx d\theta + \int_{\theta=a}^{b} \int_x P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log Q(\theta|\alpha)\right]^2 dx d\theta$

$+2 \int_{\theta=a}^{b} \int_x P(x|\theta)Q(\theta|\alpha) \frac{\partial}{\partial\theta} \log P(x|\theta) \frac{\partial}{\partial\theta} \log Q(\theta|\alpha) dx d\theta$

where

1st term $= \int_{\theta=a}^{b} Q(\theta|\alpha) \left(\int_x P(x|\theta) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)\right]^2 dx\right) d\theta = E_{Q(\theta|\alpha)}[I_P(\theta)]$

2nd term $= \int_x P(x|\theta) dx \cdot \int_{\theta=a}^{b} Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log Q(\theta|\alpha)\right]^2 d\theta = J_Q(\theta)$

3rd term $= 2 \int_{\theta=a}^{b} \frac{\partial}{\partial\theta} Q(\theta|\alpha) \cdot \int_x \frac{\partial}{\partial\theta} P(x|\theta) dx \cdot d\theta = 0$, because the inner integral is zero.

Q.E.D.