LECTURE 9B: BAYESIAN INFERENCE

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 5, 2025

1. Bayesian Methods

Bayesian method is a different approach to estimating parameters of a model.

From the Bayes' Theorem: let A and B be two events, and $P(B) \neq 0$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

$$= \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Let $X_1, \ldots, X_n$ be the data-generating process with the pdf $f(x_1, \ldots, x_n|\theta)$, where θ is an unknown parameter. Let $\pi(\theta)$ denote the researcher's prior belief about θ. Now $\pi(\theta)$ is a pdf.

Then, given the realization $(x_1, \ldots, x_n)$ from the joint likelihood $f(x_1, \ldots, x_n|\theta)$, we *update* our prior according to the Bayes' rule:

$$\pi(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{f(x_1, \ldots, x_n)} \tag{3}$$

Where $f(x)$ is the marginal distribution of x, i.e. $f(x) = \int f(x|\theta)\pi(\theta)d\theta$.

$\pi(\theta|x_1, \ldots, x_n)$ is called the *Posterior distribution* of θ. Our *Bayes estimator* is $\pi(\theta|x_1, \ldots, x_n)$, which is an entire probability distribution, not a single point estimate. To report a single point estimate from this distribution, we usually report the mean of this posterior distribution, called the posterior mean, $\mathbb{E}_{\pi(\theta|x_1,\ldots,x_n)}[\theta] = \int \theta\pi(\theta|x_1, \ldots, x_n)d\theta$.

To quantify the uncertainty around the posterior mean, we can report the posterior variance, which is: $\text{Var}_{\pi(\theta|x_1,\ldots,x_n)}(\theta)$. Alternatively, we can also report the mode, the median, or other summary statistics of the posterior distribution.

The marginal distribution $f(x)$ does not depend on θ, it is just a constant. As such, we can express the posterior distribution as:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \tag{4}$$

The constant of proportionality can be found by computing the constant C such that $Cf(x|\theta)\pi(\theta)$ integrates to one (with respect to θ), ensuring that $\pi(\theta|x)$ is a valid probability distribution.

1.1. Frequentist vs Bayesian estimators

All the estimators we encountered prior to this lecture have been Frequentist estimators. A *Frequentist* estimator of θ is a function of only the data $(x_1, \ldots, x_n)$. For instance, the sample mean and the sample variance are Frequentist estimators. The Maximum Likelihood estimator (MLE), $\hat{\theta}(x_1, \ldots, x_n) = \text{argmax}_\theta f(x_1, \ldots, x_n|\theta)$, is also a Frequentist estimator.

Frequentist estimators	Bayesian estimators
θ is a constant (there is a ground truth).	θ is not a constant, there is no ground truth. Fundamentally, θ is a random variable.
Requires the sampling model $f(x_1, \ldots, x_n\|\theta)$	Requires $f(x_1, \ldots, x_n\|\theta)$ and a prior distribution $\pi(\theta)$.
Given the dataset $(x_1, \ldots, x_n)$ drawn from $f(x_1, \ldots, x_n\|\theta)$, estimate θ as a function of $(x_1, \ldots, x_n)$. Typically involved optimization.	Given the dataset $(x_1, \ldots, x_n)$, compute $\pi(\theta\|x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n\|\theta)\pi(\theta)$.
Uncertainty about the estimate is given by the sampling distribution.	Uncertainty about the estimate is given by the posterior distribution.
Variation in the sampling distribution is entirely due to the sampling variation in $f(x_1, \ldots, x_n\|\theta)$.	Variation in the posterior distribution is a combination of sampling distribution and the prior distribution.

1.2. Prior distribution

What is the prior distribution $\pi(\theta)$ and how do we specify it?

(i) A prior distribution can be entirely subjective. A researcher's subjective belief about θ.

(ii) A prior distribution can be derived from other models and previous studies. As such, the posterior distribution reflects an updating of the prior $\pi(\theta)$ to the posterior $f(\theta|x)$ when confronted with a new source of data x. In practice, Bayesian methods perform well because it is a form of model-averaging or data-combination.

(iii) A prior distribution itself can be dependent on the current data, or estimated from the current data x. This is the Empirical Bayes approach.

(iv) A prior distribution reflects *model uncertainty*. For instance, you are not sure that $f(x|\theta, \sigma) \sim \mathcal{N}(\theta, \sigma)$ is the right model, so you let $\theta \sim \mathcal{N}(\mu, \tau)$ to be the prior distribution, essentially considering many Normal distributions with random locations.[1]

2. Example

2.1. Normal distributions

Let $X_1, \ldots, X_n$ are iid $\sim \mathcal{N}(\theta, \sigma^2)$, and suppose that the prior distribution is $\pi(\theta) = \mathcal{N}(\mu, \tau^2)$. Just for this example, assume that τ, μ, σ are known.

The likelihood:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\theta)^2/2\sigma^2}$$

The prior:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-(\theta-\mu)^2/2\tau^2}$$

The posterior:

$$f(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{\int f(x_1, \ldots, x_n|\theta)\pi(\theta)d\theta}$$

Observe that $\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$, where $\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$ is the sample mean.

[1] We can set μ to be the sample mean, in the spirit of Empirical Bayes. A reasonable prior would then be $\theta \sim \mathcal{N}(\frac{1}{n}\sum_{i=1}^{n} x_i, 1)$. Our Bayes estimate of σ would be robust to model misspecification.

$$f(\theta|x_1,\ldots,x_n) \propto \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\right)\exp\left(\frac{-(\theta-\mu)^2}{2\tau^2}\right)$$
$$\propto \exp\left(\frac{-(\theta-\tilde{\mu})^2}{2\tilde{\tau}^2}\right)$$

Where:

$$\tilde{\mu} = \tilde{\tau}^2\left(\frac{n}{\sigma^2}\bar{x}+\frac{1}{\tau^2}\mu\right)$$

$$\tilde{\tau}^2 = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

Therefore, the posterior distribution of θ given $x_1,\ldots,x_n$ is:

$$\theta|x_1,\ldots,x_n \sim \mathcal{N}\left(\tilde{\mu},\tilde{\tau}^2\right) \tag{5}$$

Since the variance of the sample mean is σ^2/n, and the variance of the prior distribution is τ^2, then the variance of the posterior distribution is just a harmonic mean between the two variances, σ^2/n and τ^2.

The posterior mean is a weighted sum of the prior mean μ and the sample mean $\bar{x}$ with weights that reflect the precision of the sample mean (given by the reciprocal of σ^2/n) and the precision of the prior (given by the reciprocal of τ^2).

As the sample size n increases, the posterior mean becomes more similar to the sample mean, which is the Frequentist estimator. This means that information from the sample dominates the prior, and the variance of the posterior reflects mostly sampling uncertainty.

As a numerical example, suppose the prior is $\pi(\theta) = \mathcal{N}(5,3)$. Suppose $\bar{x} = 10$, $\sigma^2 = 100$, and $n = 50$. The sampling distribution of the sample mean is $\mathcal{N}(10,2)$. The posterior distribution is $\mathcal{N}(8,1.2)$ according to (5). We can see from below Figure 1, that the posterior distribution lies in between the prior distribution and the sampling distribution.

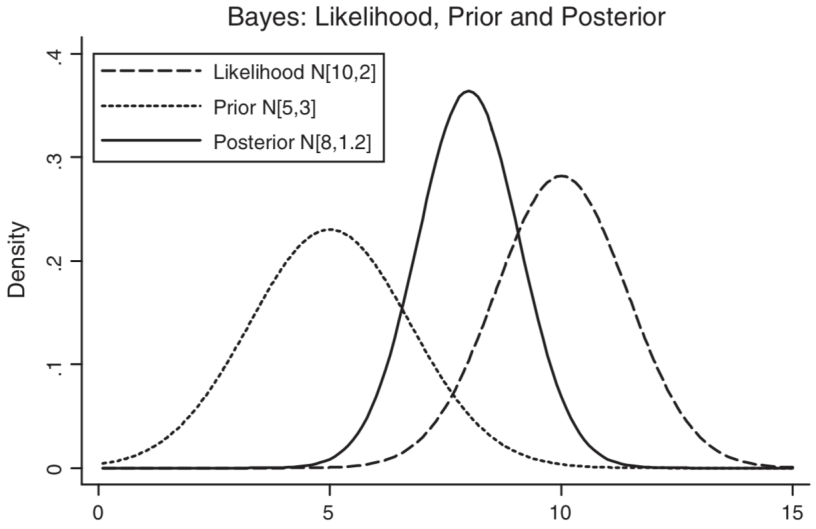

FIGURE 1. The posterior distribution lies in between the prior distribution and the sampling distribution. (Cameron and Trivedi's "Microeconometrics: Methods and Applications")

3. Uninformative prior

Given the likelihood function $f(x|\theta)$, suppose the parameter space of θ is bounded. Let the prior be a Uniform density over the range of θ, i.e. $\pi(\theta) = \frac{1}{C}$, for some number C.

The posterior distribution is:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$
$$\pi(\theta|x) \propto f(x|\theta)$$

Recall that MLE is $\text{argmax}_\theta f(x|\theta)$. Therefore, MLE coincides with the *mode of the posterior distribution* when the prior is uninformative.

When the space of θ is unbounded, a Uniform prior distribution is not proper or well-defined. In practice, we often impose a diffuse prior or a flat prior $\mathcal{N}(0, 100)$, which is a very flat distribution. It is approximately $\pi(\theta) \approx 1/(2\pi \times 100)$, which is a constant.

4. Conjugate prior

The prior distribution $\pi(\theta)$ is a *Conjugate Prior* for the likelihood function $f(x|\theta)$ if the resulting posterior distribution $f(\theta|x)$ belongs to the same probability distribution family as the prior.

In the example before, the conjugate prior for a Normal distribution is a Normal distribution, since the posterior distribution is also a Normal distribution.

In the next example, the conjugate prior for Binomial(n, p) with p as the unknown is the Beta distribution.

4.1. Example: Binomial Bayes Estimator

Suppose $X \sim \text{Binomial}(n, p)$, where p is unknown, that is, $f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$. Let the prior distribution be $\pi(p) \sim \text{Beta}(\alpha, \beta)$. Then it turns out, the posterior distribution given the realization $X = x$ is $p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$.

The mean of the prior distribution Beta(α, β), is $\frac{\alpha}{\alpha+\beta}$. The posterior mean is $\frac{x+\alpha}{n+\alpha+\beta}$, which can be decomposed into a weighted average between the sample information and the prior information:

$$\frac{x+\alpha}{n+\alpha+\beta} = \frac{n}{\alpha+\beta+n}\left(\frac{x}{n}\right) + \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)$$

Here, x/n is the frequentist estimator for p.

4.2. Empirical Bayes Estimator

In a pure Bayesian approach, the Mean Square Error is undefined because there is no true underlying value of the parameter. In an Empirical Bayes Estimator approach however, the prior distribution can be derived or estimated from the existing data, by assuming there is a true value for the underlying parameter.

Let $\hat{p}_B = \frac{X+\alpha}{n+\alpha+\beta}$ be the Bayes estimator of the Binomial parameter p. Suppose the Mean Square Error of this Bayes estimator is:

$$\begin{aligned}
\mathbb{E}[(\hat{p}_B - p)^2] &= \text{Var}(\hat{p}_B) + (\mathbb{E}[\hat{p}_B] - p)^2 \\
&= \text{Var}\left(\frac{X+\alpha}{n+\alpha+\beta}\right) + \left(\mathbb{E}\left[\frac{X+\alpha}{n+\alpha+\beta}\right] - p\right)^2 \\
&= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \left(\frac{np+\alpha}{n+\alpha+\beta} - p\right)^2
\end{aligned}$$

We could try to tune α and β to minimize the MSE. At the choice of $\alpha = \beta = \sqrt{n/4}$, the MSE is minimized and does not depend on p.

The MSE of the resulting Empirical Bayes estimator $\hat{p}_B = \frac{X+\sqrt{n/4}}{n+\sqrt{n}}$ then becomes:

$$\mathbb{E}[(\hat{p}_B - p)^2] = \frac{n}{4(n+\sqrt{n})^2}$$

The frequentist (MLE) estimator of p is $\hat{p} = X/n$. The MSE is:

$$\mathbb{E}[(X/n - p)^2] = \mathrm{Var}(X/n) + (\mathbb{E}[X/n] - p)^2 = \frac{p(1-p)}{n} + 0$$

Comparing the two MSEs, the Bayes estimator does better for an intermediate range of p and when n is small.

Is this surprising? The Bayes estimator is more flexible, and has more parameters that one can tune to optimize the MSE. Essentially, we are considering a distribution over distributions, and in this sense, the estimator is more robust. The Bayes estimator $\hat{p}_B$ doesn't just estimate Binomial(n, p), but Binomial(n, p) over $p \sim \pi(\alpha, \beta)$.

4.3. Markov Chain Monte Carlo (MCMC)

Often the posterior density has no closed form. If the posterior density is univariate, then we can use probability integral transform to sample from this posterior density. More generally, the posterior density will be multivariate. The most common way to sample from a multivariate density is to use the MCMC (Markov Chain Monte Carlo) method.

Essentially, the heavy-lifting in Bayesian analysis is sampling, while the heavy-lifting in frequentist analysis is optimization.

5. Variational Bayesian Inference (VI)

When exact posteriors are intractable and MCMC is impractical for large datasets or real-time applications, *variational inference* replaces sampling with optimization. The posterior $p(\theta \mid x)$ is approximated by a tractable distribution $q(\theta)$ in a chosen family $\mathcal{Q}$.

5.1. Evidence Lower Bound (ELBO

Let $q(\theta) \in \mathcal{Q}$ be a candidate approximation. The quality of q is measured by the Kullback–Leibler divergence.

Definition (Kullback–Leibler divergence). For densities q and p on the same support,

$$\mathrm{KL}(q \,\|\, p) \;=\; \int q(\theta)\, \log \frac{q(\theta)}{p(\theta)}\, d\theta \;=\; \mathbb{E}_q\big[\log q(\theta) - \log p(\theta)\big], \tag{6}$$

with the convention $\mathrm{KL}(q \,\|\, p) = +\infty$ if q is not absolutely continuous with respect to p.

Write the joint density as

$$p(x, \theta) \;=\; f(x \mid \theta)\, p(\theta) \tag{7}$$

and recall Bayes' rule

$$p(\theta \mid x) \;=\; \frac{f(x \mid \theta)\, p(\theta)}{p(x)}, \qquad p(x) = \int f(x \mid \theta)\, p(\theta)\, d\theta. \tag{8}$$

Start from the KL between q and the posterior, substitute (8), and expand:

$$\mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x)) = \mathbb{E}_q[\log q(\theta) - \log p(\theta \mid x)] \tag{9}$$

$$= \mathbb{E}_q[\log q(\theta) - \log f(x \mid \theta) - \log p(\theta)] + \log p(x). \tag{10}$$

The evidence $\log p(x)$ is constant with respect to q. Therefore, minimizing the KL is *equivalent* to maximizing the following functional of q:

$$\mathcal{L}(q) \;:=\; \mathbb{E}_q\big[\log f(x \mid \theta)\big] \;-\; \mathrm{KL}(q(\theta) \,\|\, p(\theta))\,. \tag{11}$$

Equation (11) is the **evidence lower bound (ELBO)**; it will be taken as the primary optimization objective.

Using (7),

$$\mathcal{L}(q) = \mathbb{E}_q\big[\log f(x \mid \theta)\big] + \mathbb{E}_q\big[\log p(\theta)\big] - \mathbb{E}_q\big[\log q(\theta)\big] \tag{12}$$

$$= \mathbb{E}_q\big[\log p(x, \theta)\big] - \mathbb{E}_q\big[\log q(\theta)\big], \tag{13}$$

i.e., "expected joint minus entropy". Combining (10) and (11) yields the standard evidence decomposition

$$\log p(x) \;=\; \mathcal{L}(q) \;+\; \mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x))\,. \tag{14}$$

Hence $\mathcal{L}(q) \le \log p(x)$ with equality iff $q(\theta) = p(\theta \mid x)$ almost everywhere.

5.2. Mean-field variational family

To obtain a tractable approximation while allowing high-dimensional θ, a common restriction is the mean-field family. Partition $\theta = (\theta_1, \ldots, \theta_J)$ and posit

$$q(\theta) \;=\; \prod_{j=1}^{J} q_j(\theta_j). \tag{15}$$

This assumption ignores posterior dependencies across the parameters but if the true posterior lies within this family, then the approximation becomes exact.

5.3. Mean-field Gaussian

A frequently used concrete instance of (15) is the fully factorized Gaussian:

$$q(\theta) \;=\; \prod_{j=1}^{d} \mathcal{N}(\theta_j;\, \mu_j, \sigma_j^2). \tag{16}$$

The entropy is available in closed form,

$$\mathsf{H}(q) \;=\; -\,\mathbb{E}_q[\log q(\theta)] \;=\; \tfrac{1}{2}\sum_{j=1}^{d}\big(1 + \log 2\pi + \log \sigma_j^2\big), \tag{17}$$

so maximizing (11) over the mean-field Gaussian family reduces to optimization in the parameters $\{\mu_j, \sigma_j^2\}_{j=1}^{d}$.[2]

As a note, when the prior $p(\theta)$ is also Gaussian, the KL divergence between two gaussians is available in closed form. For $q = \mathcal{N}(\mu, \Sigma)$ and $p = \mathcal{N}(m, S)$ in $\mathbb{R}^d$,

$$\mathrm{KL}(q \,\|\, p) \;=\; \tfrac{1}{2}\Big(\mathrm{tr}(S^{-1}\Sigma) + (m-\mu)^{\top} S^{-1}(m-\mu) - d + \log \tfrac{|S|}{|\Sigma|}\Big). \tag{18}$$

[2]When expectations $\mathbb{E}_q[\log f(x \mid \theta)]$ are not available in closed form, unbiased Monte Carlo estimates (e.g., via the reparameterization $\theta = \mu + \sigma \odot \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$) provide gradients for first-order optimization.