Bayesian Modeling Problems

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

PART A: CONCEPTUAL QUESTIONS

A1. What is the difference between a 95% frequentist *confidence interval* and a 95% Bayesian *credible interval* for a parameter?

Answer: A 95% confidence interval is constructed through repeated-sampling theory: if we repeatedly sample data and build intervals in the same way, 95% of those intervals would contain the true parameter. It *does not* allow a probability statement about the specific interval from one experiment. In contrast, a 95% Bayesian credible interval represents a direct probability statement about the parameter given the observed data: there is a 95% posterior probability that the parameter lies within the credible interval. The confidence interval depends only on the data (and an assumed sampling distribution) and has no prior, whereas the credible interval incorporates prior information and the observed data to give a probability distribution for the parameter.

A2. Name and briefly describe the two types of uncertainty often characterized in Bayesian modeling.

Answer: In Bayesian modeling we distinguish: - *Aleatoric uncertainty*: the inherent randomness or noise in the outcome, due to natural variability. Even with infinite data, this uncertainty remains (e.g. coin flips have 50% chance each time). - *Epistemic uncertainty*: the uncertainty due to lack of knowledge or limited data. This uncertainty can be reduced with more information or data. Bayesian methods capture epistemic uncertainty via the spread of the posterior distribution over parameters (which shrinks as data grow), while aleatoric uncertainty is captured in the likelihood (and reflected in the predictive distribution's irreducible variance).

A3. What is a *conjugate prior*? Give an example of a conjugate prior-likelihood pair.

Answer: A conjugate prior is a prior distribution that, when combined with a particular likelihood, yields a posterior of the same functional form as the prior. In other words, prior and posterior are in the same family of distributions. For example, a Beta prior for a Bernoulli/Binomial likelihood is conjugate, since the posterior will also be a Beta distribution. Specifically, if $p \sim \text{Beta}(\alpha, \beta)$ and the data likelihood is $\text{Binomial}(N, p)$ with X successes, then the posterior is $\text{Beta}(\alpha + X,\ \beta + N - X)$.

A4. Explain the distinction between Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) estimation for a model parameter.

Answer: MLE finds the parameter value that maximizes the likelihood (data fit) alone, essentially assuming a uniform (uninformative) prior. It produces a single point estimate that best explains the observed data, but provides no direct measure of uncertainty. MAP estimation finds the parameter that maximizes the *posterior* distribution, i.e. it maximizes (likelihood × prior). Thus MAP incorporates prior beliefs in addition to the data. In cases with a reasonably uninformative prior, MAP and MLE may coincide or be very close. However, with informative priors or limited data, MAP estimates are pulled toward the prior mean (regularization effect), whereas MLE can be more extreme or overfit. MAP also provides just a point estimate.

PART B: MATHEMATICAL DERIVATIONS

B1. Consider a Binomial likelihood with k successes out of N trials for an unknown probability θ, and a $\text{Beta}(\alpha, \beta)$ prior for θ.

(a) Derive the posterior distribution for θ (show that the Beta prior is conjugate to the Binomial likelihood).

(b) Find the MAP (Maximum A Posteriori) estimate of θ from this posterior, and compare it to the MLE (Maximum Likelihood Estimate).

Answer:

(a) The likelihood is $P(k \mid \theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$. The prior is $\mathrm{Beta}(\alpha, \beta)$ with density proportional to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. The unnormalized posterior is:

$$p(\theta \mid k) \propto \underbrace{\theta^x(1-\theta)^{N-x}}_{\text{likelihood}} \times \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} = \theta^{\,x+\alpha-1}\,(1-\theta)^{\,N-x+\beta-1}\,.$$

This is the kernel of a $\mathrm{Beta}(x+\alpha,\ N-x+\beta)$ distribution. Thus, the posterior is

$$\theta \mid k \sim \mathrm{Beta}(\alpha + k,\ \beta + N - k)\,.$$

(b) The MAP estimate is the mode of the Beta posterior. For $\mathrm{Beta}(a, b)$ with $a > 1, b > 1$, the mode is $\frac{a-1}{a+b-2}$. Here $a = \alpha + k$ and $b = \beta + N - k$, so

$$\theta_{\mathrm{MAP}} = \frac{\alpha + k - 1}{\alpha + \beta + N - 2}\,.$$

On the other hand, the MLE ignores the prior and maximizes $\theta^k(1-\theta)^{N-k}$, which gives $p_{\mathrm{MLE}} = \frac{k}{N}$. We can see that if the prior is uniform ($\alpha = \beta = 1$), then p_{MAP} reduces to k/N (so MAP = MLE). In general with $\alpha, \beta > 1$, p_{MAP} will be a weighted estimate that pulls k/N toward the prior mean $\frac{\alpha-1}{\alpha+\beta-2}$. For example, with very small sample N, a heavily informative prior will make p_{MAP} closer to the prior belief than to the MLE. As N grows large, the likelihood dominates and $p_{\mathrm{MAP}} \approx p_{\mathrm{MLE}}$.

PART C: PARAMETRIC ANALYSIS (What if we change parameters?)

C1. You consider two Bayesian models for a probability θ (e.g. the fraction of defective items in manufacturing) with different prior strengths:

- Model A: $\theta \sim \mathrm{Beta}(1, 1)$ (a very weak, uniform prior).
- Model B: $\theta \sim \mathrm{Beta}(10, 10)$ (a stronger prior peaked around 0.5).

Both models are updated on the same dataset. Answer the following:

(a) Which model's posterior distribution for θ will have the larger variance? Explain why.

(b) Which model is likely to yield more extreme posterior *predictive* probabilities for new observations (i.e. predictions closer to 0 or 1)? Why?

(c) In a scenario where θ is actually very small (a rare event case), which prior (Model A or Model B) would be more appropriate to use, and why?

Answer:

(a) Model A (Beta(1,1) prior) will have a higher posterior variance. The Beta(1,1) prior contributes essentially no information (it's flat), so the posterior is driven only by the data, which if limited, leads to a broad (high-variance) posterior. Model B's stronger prior (Beta(10,10)) adds "pseudo-observations" that concentrate the posterior, resulting in lower variance. Intuitively, the weak prior model is less confident (more spread-out posterior) while the strong prior model is more confident (tighter posterior) about θ after seeing the same data.

(b) Model B (strong prior) tends to yield more extreme predictive probabilities (closer to 0 or 1) for new observations. This is because the posterior in Model B is more concentrated—after updating, Model B is more confident about the value of θ and will make predictions with less hesitation. Model A's posterior, being broader, effectively "hedges" more, yielding predictions closer to moderate values (e.g. around 0.5 for a Bernoulli outcome) when averaging over uncertainty. In other words, Model B's confidence leads to predictions that, for a given dataset, are closer to a deterministic outcome (high or low probability), whereas Model A often predicts with more caution (closer to 50% when uncertainty is high).

(c) If θ is believed to be very small (a rare event), a stronger informative prior that reflects this (like a Beta prior heavily skewed toward 0) is preferable. Using a weak or uniform prior (Model A) in a rare-event scenario means the posterior would be highly influenced by even a few observed successes, potentially overestimating θ. A strong prior (Model B, or even one skewed to low θ) incorporates prior knowledge of rarity, preventing the posterior from reacting too extremely to limited data. It provides necessary regularization: for example, if we observe one or two failures in a small sample, a strong low-θ prior will keep the posterior's estimate of θ moderate, avoiding the conclusion that θ is large when the data are sparse.

PART D: IMPLEMENTATION AND PRACTICAL QUESTIONS

D1. In a Beta-Binomial model, you have updated the posterior for θ given observed data. How can you compute the probability that the next trial will be a success? Provide:

(a) an analytic expression for this posterior predictive probability in terms of the posterior Beta parameters, and

(b) a brief description of how you could approximate this probability via Monte Carlo simulation.

Answer:

(a) If the posterior after observing data is $\theta \mid \text{data} \sim \text{Beta}(\alpha_{\text{post}}, \beta_{\text{post}})$, then the predictive probability that the next Bernoulli trial is a success is the posterior mean $E[\theta \mid \text{data}]$. This is given by

$$P(\text{success on next trial} \mid \text{data}) = \frac{\alpha_{\text{post}}}{\alpha_{\text{post}} + \beta_{\text{post}}}.$$

Equivalently, one can derive this by the integral $\int_0^1 \theta\, f(\theta \mid \text{data})\, d\theta$, which for a Beta distribution yields $\alpha_{\text{post}}/(\alpha_{\text{post}} + \beta_{\text{post}})$.

(b) To approximate this via Monte Carlo, we can sample many θ values from the posterior and average their outcomes. For example:

- Draw $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ i.i.d. from $\text{Beta}(\alpha_{\text{post}}, \beta_{\text{post}})$.
- For each sampled $\theta^{(i)}$, simulate a Bernoulli trial (success with probability $\theta^{(i)}$).
- Estimate the predictive success probability as the fraction of these M simulated trials that resulted in success.

In practice, this Monte Carlo estimate will converge to the true analytic value $\frac{\alpha_{\text{post}}}{\alpha_{\text{post}}+\beta_{\text{post}}}$ as M becomes large. This approach is useful if we want to simulate more complex outcomes or if an analytic formula were not readily available.

PART E: Other problems

E1. What is an *equal-tailed* 95% credible interval, and how does it differ from a 95% *highest posterior density* (HPD) credible interval? In what situations would these two types of intervals give notably different results?

Answer: An equal-tailed 95% credible interval is defined by the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles of the posterior distribution. It excludes 2.5% of the posterior probability on each tail. In contrast, a 95% highest posterior density (HPD) interval is the narrowest interval containing 95% of the posterior mass; every point inside an HPD interval has higher posterior density than any point outside the interval.

For a symmetric posterior distribution (e.g. a Gaussian), the equal-tailed and HPD intervals coincide. However, in skewed or multimodal distributions, they can differ significantly. The HPD interval will generally be shorter and anchored around the mode(s) of the distribution, whereas an equal-tailed interval might include more low-density region in one tail to achieve the 2.5%–97.5% cutoffs. For example, if the posterior of a rate parameter θ is highly skewed (say heavy tail toward 1), the equal-tailed 95% interval might run from $\theta = 0.01$ to $\theta = 0.30$, even if most of the probability is tightly concentrated below 0.20. The 95% HPD might be something like $[0.01, 0.22]$, excluding the upper low-density tail beyond 0.22. In general, the HPD interval is more informative for skewed distributions, since it always captures the region of highest belief (it will include the posterior mode), while the equal-tailed interval is easier to compute but may not be the most compact credible interval.

E2. **Problem: Manufacturing Defect Detection**. A manufacturing line produces electronic components, and the historical defect rate is around 5%. You decide to use Bayesian modeling to monitor the defect rate in production. You assume a Beta prior for the defect probability θ. In a random sample of $N = 10$ components from a new batch, you observe $X = 2$ defective units.

(a) Why might a Bayesian approach be preferable to a frequentist approach (e.g. using just the sample proportion) for estimating θ in this scenario?

(b) Assume a prior $\theta \sim \text{Beta}(2, 38)$ reflecting the historical belief (mean $\approx$ 5%). Update this prior with the data. Give the posterior distribution for θ, and calculate the posterior mean and a 95% credible interval for θ.

(c) Suppose any defect rate above 10% is deemed unacceptable. If the cost of missing an unacceptably high defect rate (a “false negative” of not acting when $\theta > 0.10$) is 5 times the cost of a false alarm (stopping production when it was actually fine), should you halt production to investigate? Base your decision on the posterior probability that $\theta > 0.10$ and the given cost ratio.

Answer:

(a) A Bayesian approach offers several advantages in this scenario:

1. *Incorporation of prior knowledge:* We have a historical expectation of 5% defects. Bayesian analysis allows us to encode this prior belief (via a Beta prior) and update it with the new data. A frequentist approach would treat the new sample in isolation, potentially overreacting to a small sample result.
2. *Better performance with limited data:* With only $N = 10$ samples, the sample proportion (MLE) is a very noisy estimate (20% in this sample). The Bayesian posterior combines prior information with this limited data, yielding a more stable estimate (shrinking towards 5%). This regularization helps prevent extreme conclusions from small samples.

3. *Uncertainty quantification:* The Bayesian result is a full posterior distribution for θ, from which we can derive credible intervals. For example, we can say there's a high probability that θ lies in a reasonable range (rather than just giving a point estimate 0.2). The frequentist approach might give a confidence interval, but its interpretation is indirect and it could be wide due to the small N.

4. *Decision-making under risk:* Bayesian analysis directly provides the probability of scenarios of interest (e.g. $P(\theta > 0.10 \mid \text{data})$) which is useful for making decisions. In contrast, a classical test might give a θ-value for "$\theta = 0.10$ vs not," but it won't directly tell us the probability that the defect rate exceeds 10%. For cost-sensitive decisions, the Bayesian framework is more natural.

(b) The prior is Beta(2, 38). Given $X = 2$ defects out of $N = 10$, the posterior for θ is:

$$\theta \mid X = 2 \;\sim\; \text{Beta}(2+2,\; 38+8) \;=\; \text{Beta}(4,\; 46)\,.$$

The posterior mean is

$$E[\theta \mid \text{data}] = \frac{4}{4+46} = 0.08\,,$$

i.e. about an 8% defect rate (higher than the prior 5%, but much lower than the raw sample 20% thanks to the prior).

For the 95% credible interval, we find the 2.5% and 97.5% quantiles of Beta(4, 46). (This can be done via a software routine - not during the exam.) The 95% equal-tailed credible interval is approximately

$$\theta \in [0.023,\; 0.169]\,,$$

i.e. about [2.3%, 16.9%].