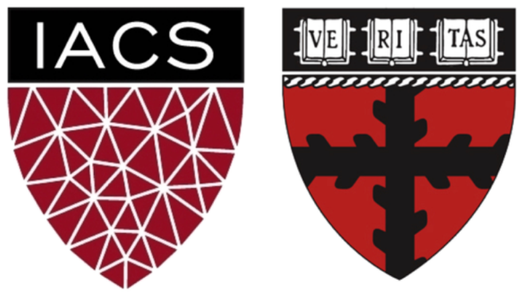
Lecture #4: Bayesian versus Frequentist Inference

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020


IACS
VE RI TAS

Outline

1. Review of Bayesian Modeling
2. Examples of Conjugate and Non-Conjugate Models
3. Connections to Frequentist Inference

Review of Bayesian Modeling

The Bayesian Modeling Process

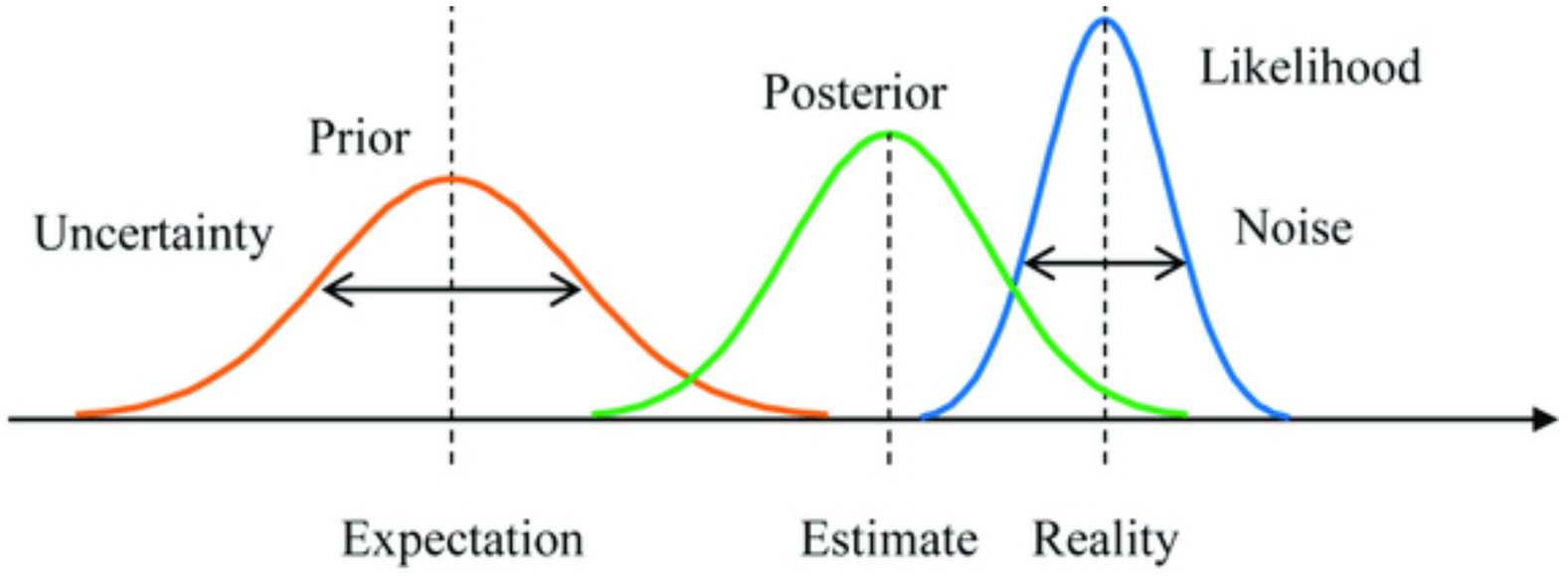
In order to make statements about Y, the outcome, and θ, parameters of the distribution generating the data, we form the joint distribution over both variables and use the various marginals/conditional distributions to reason about Y and θ.

1. we form the ***joint distribution*** over both variables $p(Y, \theta) = p(Y|\theta)p(\theta)$.

2. we can condition on the observed outcome to make inferences about θ,
$$p(\theta|Y) = \frac{p(Y, \theta)}{p(Y)}$$
where $p(\theta|Y)$ is called the ***posterior distribution*** and $p(Y)$ is called the ***evidence***.

3. before any data is observed, we can simulate data by using our prior
$$p(Y^*) = \int_\Theta p(Y^*, \theta)d\theta = \int_\Theta p(Y^*|\theta)p(\theta)d\theta$$
where Y^* represents new data and $p(Y^*)$ is called the ***prior predictive***.
4. after observing data, we can simulate new data simliar to the observed data by using our posterior
$$p(Y^*|Y) = \int_\Theta p(Y^*, \theta|Y)d\theta = \int_\Theta p(Y^*|\theta)p(\theta|Y)d\theta$$
where Y^* represents new data and $p(Y^*|Y)$ is called the ***posterior predictive***.

Bayesian Update



Model Evaluation

How do we know that our Bayesian model is a good fit for the data?

1. **(Log-likelihood)** We can compute the marginal log-likelihood of the data under our posterior. That is, give a set of test data $\{y_1^*, \dots, y_M^*\}$, compute

$$\log \prod_{m=1}^{M} p(\mathbf{y}_m^* | \text{Data}) = \sum_{m=1}^{M} \log p(\mathbf{y}_m^* | \text{Data}) = \sum_{m=1}^{M} \log \int_{\Theta} p(\mathbf{y}_m^* | \theta) p(\theta | \text{Data}) d\theta$$

   This is simply called the **log-likelihood** of the data under the Bayesian model.

2. **(Posterior Predictive Check)** We can also compare the synthetic data generated from our posterior predictive:

   A. sample from the posterior $\theta_s \sim p(\theta | Y)$
   B. plug the posterior samples into the likelihood, and sample synthetic data from the likelihood $Y_s \sim p(Y | \theta_s)$.

   We can then compare the synthetic data from the posterior predictive to the observed data.

Components of Bayesian Inference

We see that in order to evaluate Bayesian models we need to be able to perform two tasks:

1. integration over the posterior (required by log-likelihood)
2. sampling from the posterior (required by the posterior predictive check).

Both requirements becomes easier if know the closed form expression for the posterior, e.g. if the prior is conjugate to the likelihood.

Examples of Conjugate and Non-Conjugate Models

Bayesian Model for (Univariate) Gaussian Likelihood with Known Variance

The Bayesian Model

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known. We place a normal prior on μ, $\mu \sim \mathcal{N}(m, s^2)$.

Question: is our choice of prior appropriate?

Bayesian Model for (Univariate) Gaussian Likelihood with Known Variance

Inference: The posterior $p(\mu|Y)$ is then:

$$p(\mu|Y) = \frac{p(Y|\mu)p(\mu)}{p(Y)} = \frac{\overbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(Y-\mu)^2}{2\sigma^2}\right\}}^{\text{likelihood}}\overbrace{\frac{1}{\sqrt{2\pi s^2}}\exp\left\{-\frac{(m-\mu)^2}{2s^2}\right\}}^{\text{prior}}}{p(Y)}$$

We can simplify the posterior as:

$$\begin{aligned} p(\mu|Y) &= const * \frac{\exp\left\{-\frac{s^2(Y-\mu)^2+\sigma^2(m-\mu)^2}{2s^2\sigma^2}\right\}}{p(Y)} \\ &= const * \exp\left\{\frac{s^2Y^2+\sigma^2m^2}{\sigma^2s^2}\right\}\exp\left\{-\frac{(s^2+\sigma^2)\mu^2-2(s^2Y+\sigma^2m)\mu}{2s^2\sigma^2}\right\} \\ &= const * \exp\left\{-\frac{\left(\mu-\frac{s^2Y+\sigma^2m}{s^2+\sigma^2}\right)^2}{2s^2\sigma^2}\right\} \quad \text{(Completing the square)} \end{aligned}$$

Thus, we see that the posterior is a normal distribution, $\mathcal{N}\left(\frac{s^2Y+\sigma^2m}{s^2+\sigma^2}, s^2\sigma^2\right)$.

Bayesian Model for (Univariate) Gaussian Likelihood with Known Mean

The Bayesian Model

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$, with μ known. We place an inverse-gamma prior on σ^2, $\sigma^2 \sim IG(\alpha, \beta)$.

Question: is our choice of prior appropriate?

Bayesian Model for (Univariate) Gaussian Likelihood with Known Mean

Inference: The posterior $p(\sigma^2|Y)$ is then:

$$p(\sigma^2|Y) = \frac{p(Y|\sigma^2)p(\sigma^2)}{p(Y)} = \frac{\overbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(Y-\mu)^2}{2\sigma^2}\right\}}^{\text{likelihood}}\overbrace{\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma^2\right)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma^2}\right\}}^{\text{prior}}}{p(Y)}$$

We can simplify the posterior as:

$$p(\sigma^2|Y) = const * \left(\sigma^2\right)^{-(\alpha+0.5)-1}\exp\left\{-\frac{\frac{(Y-\mu)^2}{2}+\beta}{\sigma^2}\right\}$$

Thus, we see that the posterior is an inverse gamma distribution,
$IG\left(\alpha+0.5, \frac{(Y-\mu)^2}{2}+\beta\right)$.

Bayesian Model for (Univariate) Gaussian Likelihood with Unknown Mean and Variance

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$, with both parameters unknown. We place a normal prior on μ, $\mu \sim \mathcal{N}(m, s^2)$, and an inverse-gamma prior on σ^2, $\sigma^2 \sim IG(\alpha, \beta)$.

The posterior $p(\sigma^2|Y)$ is then:

$$p(\mu, \sigma^2|Y) = \frac{\overbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(Y-\mu)^2}{2\sigma^2}\right\}}^{\text{likelihood}}\overbrace{\frac{1}{\sqrt{2\pi s^2}}\exp\left\{-\frac{(m-\mu)^2}{2s^2}\right\}}^{\text{prior on } \mu}\overbrace{\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma^2\right)^{-\alpha-1}\exp\left\{-\frac{\beta}{\sigma^2}\right\}}^{\text{prior on } \sigma^2}}{p(Y)}$$

Can the posterior be simplified so that we recognize the form of the distribution?

Non-Conjugate Models

We know that conjugate priors yield closed-form expressions for the posterior. In all our examples, these posteriors have been both easy to sample from and easy to integrate over. That is, we can evaluate our Bayesian models: can compute the log-likelihood of the data and perform posterior predictive checks.

So why would we ever consider non-conjugate priors?

Suppose that $Y \sim \mathcal{N}(\mu, 2)$ represent the height of a person randomly selected from a population. Would the conjugate prior $\mu \sim \mathcal{N}(5.7, 1)$ be appropriate for this application?

If we wanted to use the prior $\mu \sim Ga(5.7, 1)$, would we be able to derive a closed form expression for the posterior?

Connections with Frequentist Inference

Point Estimates from the Posterior

If you absolutely wanted to derive a point estimate for the parameters θ in the likelihood from your Bayesian model, there are two common ways to do it:

1. ***the posterior mean***

$$\theta_{\text{post mean}} = \mathbb{E}_{\theta \sim p(\theta|Y)}\left[\theta|Y\right] = \int \theta p(\theta|Y) d\theta$$

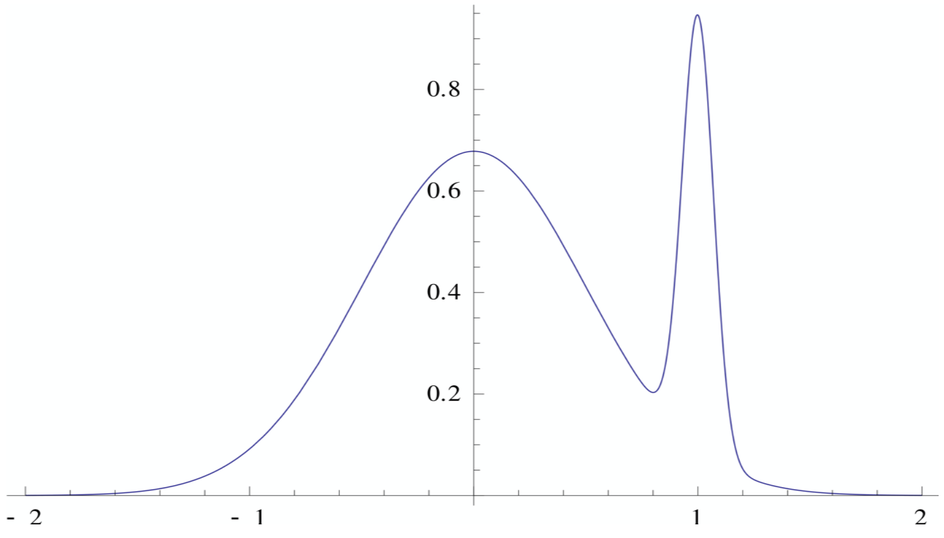
2. ***the posterior mode*** or ***maximum a posterior (MAP)*** estimate

$$\theta_{\text{MAP}} = \text{argmax}_{\theta} p(\theta|Y)$$

Question: is it better to summarize the entire posterior using a point estimate? I.e. why should we keep the posterior distribution around?

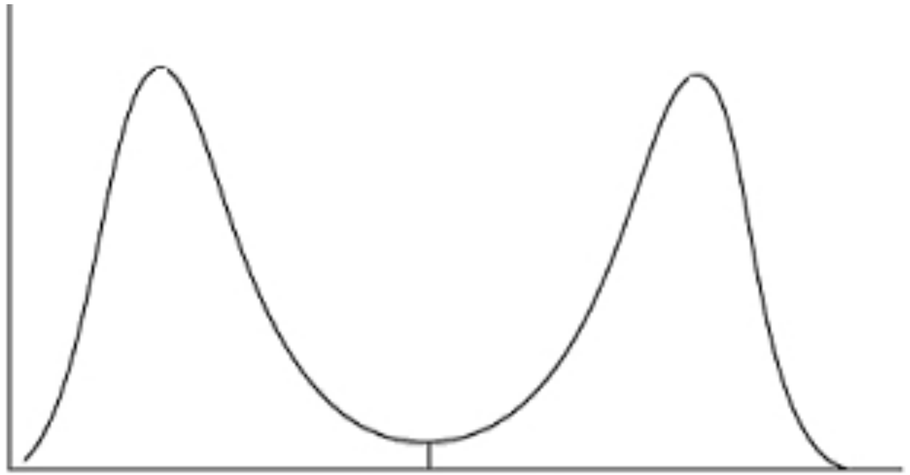
Point Estimates Can Be Misleading

The posterior mode can be an atypical point:



Point Estimates Can Be Misleading

The posterior mean can be an unlikely point:



Comparison of Posterior Point Estimates and MLE

Beta-Binomial Model for Coin Flips

- Likelihood: $Bin(N, \theta)$
- Prior: $Beta(\alpha, \beta)$
- MLE: $\frac{Y}{N}$
- MAP: $\frac{Y+\alpha-1}{N+\alpha+\beta-2}$
- Posterior Mean: $\frac{Y+\alpha}{N+\alpha+\beta}$

Question: What is the effect of the prior on the posterior point estimates? Imagine if $Y = 10, N = 11, \alpha = 100, \beta = 300$. What if $Y = 1,000, N = 11,000, \alpha = 1, \beta = 3$?

The Coin Toss Example: Revisited Yet Again

Recall that one way to prevent the MLE from overfitting is to add ***regularization terms***:

$$\theta_{\text{MLE Reg}} = \frac{Y + \alpha}{N + \beta}.$$

This is very similar to the MAP and posterior mean estimates:

- MAP: $\frac{Y+\alpha-1}{N+\alpha+\beta-2}$
- Posterior Mean: $\frac{Y+\alpha}{N+\alpha+\beta}$

In fact, we have see that one effect of adding a prior is that it **regularizes** our inference about θ.

Question: What happens to the MAP and posterior mean estimates as N (and hence Y) becomes very large?

$$\lim_{N\to\infty} \frac{Y + \alpha - 1}{N + \alpha + \beta - 2} = ?$$

Law of Large Numbers for Bayesian Inference

In general, in Bayesian inference we are **less interested asymptotic behavior**. But the properties of the asymptotic distribution of the posterior can be useful.

Theorem: (Berstein-von Mises)

"Under some conditions, as $N \rightarrow \infty$ the posterior distribution converges to a Gaussian distribution centred at the MLE with covariance matrix given by a function of the Fisher information matrix at the true population parameter value."

Consequences

1. The posterior point estimates approach the MLE, with large samples sizes.

2. It may be valid to approximate the posterior with a Gaussian, with large samples sizes. This will become a very important idea during the second half of the course!

Computational Comparisons

1. **Computation of the MLE is an optimization problem.** Although difficult, there are many established methods for performing optimization (even when the objective function is not convex -- i.e. many local optima).

   More importantly, there are algorithms to perform general, automatic optimization (e.g. gradient descent) on a large class of functions -- that is, we do not need to artisanally solve an optimization problem for each statistical model.

2. **Computation of the posterior (thus far) is an process of choosing the right priors and noting that the posterior distribution is of the same type as the prior.** The derivation is simple so long as we use conjugate priors. But many intuitively appropriate priors (like the inverse gamma and normal priors for a univariate gaussian) are not conjugate. In those cases, it becomes intractable to

   - compute posterior ~~mode~~ or mean
   - simulate samples from the posterior (and hence simulate samples from the posterior predictive)