MVE550 2022 Lecture 2
Compendium Chapter 1:
Basics of Bayesian inference. Conjugacy.
Prediction. Discretization.

Petter Mostad

Chalmers University

November 3, 2022

Outline for today

- Idea of Bayesian inference: Predicting from conditional stochastic models.
- Tossing a coin: The Beta Binomial conjugacy.
- The Poisson Gamma conjugacy.
- Computations of predictive distributions.
- Bayesian inference using discretization or numerical integration.

Example: Throwing a dice

- If you are trowing a fair six-sided dice, your stochastic model would be that each outcome has probability $1/6$.
- New observations would be independent of old observations: To make predictions, you don't need data.
- Assume instead the dice may be biased in some way, but you don't know exactly how.
- A way to make predictions would be to first acquire data, i.e., record approximately how often each outcome occurs, and use that information when predicting. Outcomes would be *dependent*.
- Thus you use a more complex stochastic model that reasonably models the dependency.
- Given a sequence 1, 5, 6, 1, 3, 1, 1, 2, 1, 5, the probability for 1 in the next throw is then computed as

$$\Pr(1 \mid 1,5,6,1,3,1,1,2,1,5) = \frac{\Pr(1,5,6,1,3,1,1,2,1,5,1)}{\Pr(1,5,6,1,3,1,1,2,1,5)}$$

Biased coin example

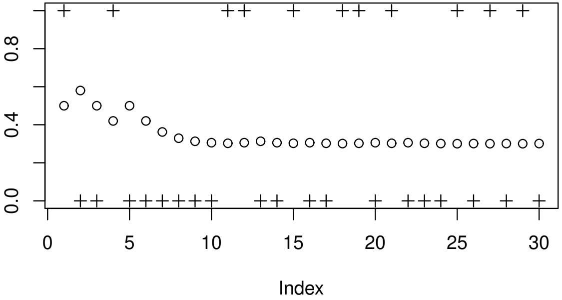

Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is that θ, the probability of heads, is either 0.7 or 0.5, with $\Pr(\theta = 0.7) = \Pr(\theta = 0.3) = 0.5$.

Reformulation using the underlying parameter θ

- A more common approach: Define the model in terms of a parameter θ, so that all observations are independent *given* θ.
- In our case: θ is a discrete random variable, possible values 0.7 and 0.3:
$$\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5.$$
.
- If y is count of heads in n first throws, and y_{new} is count of heads in the next throw:
$$y \mid \theta \sim \text{Binomial}(n, \theta) \qquad \text{and} \qquad y_{new} \mid \theta \sim \text{Binomial}(1, \theta)$$
- We can use
$$\pi(y_{new} \mid y) = \sum_{\theta} \pi(y_{new} \mid \theta)\pi(\theta \mid y) \;\text{ and }\; \pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}$$
- For example, $\pi(\theta = 0.3 \mid y) = \frac{\pi(y|\theta=0.3)\pi(\theta=0.3)}{\pi(y|\theta=0.7)\pi(\theta=0.7)+\pi(y|\theta=0.3)\pi(\theta=0.3)}$.
- We get exactly the same results as above. (Prove!)

General terminology

- **The probability distribution for θ, $\pi(\theta)$, is called the prior.**
- **The probability distribution for the data y given θ, $\pi(y \mid \theta)$ is called the likelihood, when it is viewed as a function of θ.**
- **The probability distribution for θ given the value of the data y, $\pi(\theta \mid y)$ is called the posterior.**

Finding the posterior for θ using a uniform prior

- The conditional model $\pi(\theta \mid y)$ (the posterior for θ) can be computed with Bayes formula. We get

$$\begin{aligned}\pi(\theta \mid y) &= \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} = \frac{\pi(y \mid \theta)\pi(\theta)}{\int_0^1 \pi(y \mid \theta)\pi(\theta)\, d\theta} \\ &= \frac{\text{Binomial}(y; n, \theta)}{\int_0^1 \text{Binomial}(y; n, \theta)\, d\theta} = \frac{\theta^y (1-\theta)^{n-y}}{\int_0^1 \theta^y (1-\theta)^{n-y}\, d\theta}.\end{aligned}$$

- Before we continue with computing the integral, we review the definition of the Beta distribution.

Review of definition: The Beta distribution

θ has a Beta distribution on $[0, 1]$, with parameters α and β, if its density has the form

$$\pi(\theta \mid \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta *function* defined by

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

where $\Gamma(t)$ is the *Gamma function* defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx.$$

Recall that for positive integers, $\Gamma(n) = (n-1)! = 1 \cdot \cdots \cdot (n-1)$. See for example Wikipedia for more properties of the Beta distribution, and the Beta and Gamma functions. We write $\pi(\theta \mid \alpha, \beta) = \mathrm{Beta}(\theta; \alpha, \beta)$ for the Beta density; we then also write $\theta \sim \mathrm{Beta}(\alpha, \beta)$.

Biased coin example

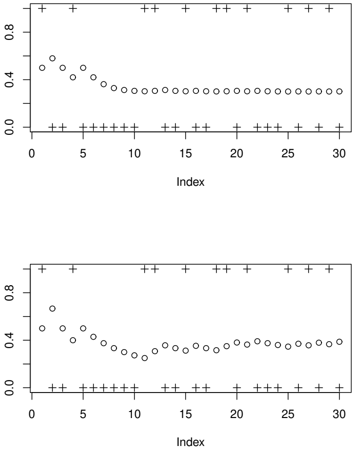

Figure: The probability of heads at each point in a sequence of observations, or the probability of "success", conditioning on the previous observations. The priors used are $\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5$ (left) and $\theta \sim \text{Uniform}(0, 1)$

Using a Beta distribution as prior

- Assume the prior is $\theta \sim \text{Beta}(\alpha, \beta)$.
- The posterior becomes (prove!)

$$\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$$

- The prediction becomes (prove!)

$$\pi(y_{new} = 1 \mid y) = \text{E}\left(\theta \mid y\right) = \frac{y + \alpha}{n + \alpha + \beta}.$$

- **DEFINITION: Given a likelihood model $\pi(x \mid \theta)$. A conjugate family of priors to this likelihood is a parametric family of distributions for θ so that if the prior is in this family, the posterior $\theta \mid x$ is also in the family.**

Biased coin example

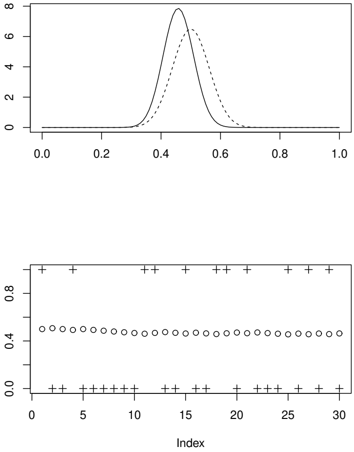

Figure: Left: The prior Beta$(33.4, 33.4)$ and the posterior Beta$(33.4 + 11, 33.4 + 19)$ for θ. Right: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of

Example: The Poisson-Gamma conjugacy

- Assume $\pi(x \mid \theta) = \text{Poisson}(x; \theta)$, i.e., that

$$\pi(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!}$$

- Then $\pi(\theta \mid \alpha, \beta) = \text{Gamma}(\theta; \alpha, \beta)$ where α, β are positive parameters, is a conjugate family. Recall that

$$\text{Gamma}(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$

- Specifically, we have the posterior

$$\pi(\theta \mid x) = \text{Gamma}\left(\theta; \alpha + x, \beta + 1\right).$$

Poisson-Gamma example

- We make repeated observations of a Poisson(θ) distributed variable for some $\theta > 0$. The observed values are $x_1 = 20$, $x_2 = 24$, and $x_3 = 23$. What is the posterior distribution for θ given this data?
- We first must decide on a prior for θ. In this example we use $\pi(\theta) \propto_\theta \frac{1}{\theta}$.
- Note that this is an *improper* prior; it is a "density" that does not integrate to 1! However, using such improper priors is possible in Bayesian statistics.
- We get the posterior after observing x_1:

$$\theta \mid x_1 \sim \text{Gamma}(20, 1)$$

- Using this as prior, we get after also observing x_2:

$$\theta \mid x_1, x_2 \sim \text{Gamma}(20 + 24, 1 + 1)$$

and similar for the last observation x_3.

Poisson-Gamma example

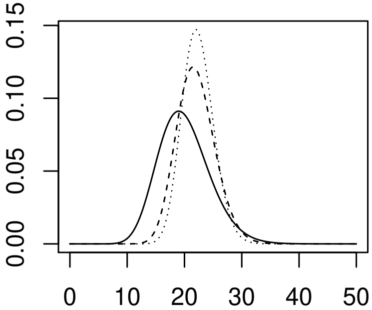

Figure: The posteriors after one, two, and three observations, where $x_1 = 20, x_2 = 24$, and $x_3 = 23$. Note how increasing amounts of data leads to a narrower posterior density.

Predictive distribution for the Poisson Gamma conjugacy

- We have seen: If $k \mid \theta \sim \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$ then $\theta \mid k \sim \text{Gamma}(\alpha + k, \beta + 1)$.
- Direct computation gives the prior predictive distribution

$$\pi(k) = \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(\theta \mid k)} = \frac{\beta^\alpha \Gamma(\alpha + k)}{(\beta + 1)^{\alpha + k}\Gamma(\alpha) k!}$$

- Note that the positive integer x has a Negative Binomial distribution with parameters r and p if its probability mass function is

$$\pi(x \mid r, p) = \binom{x + r - 1}{x} \cdot (1 - p)^x p^r = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)}(1 - p)^x p^r$$

- We get that the prior predictive is $\text{Negative-Binomial}(\alpha, \beta/(1 + \beta))$.
- Note that we can get the posterior predictive by simply replacing the α and β of the prior with the corresponding $\alpha + k$ and $\beta + 1$ of the posterior.

Poisson-Gamma example

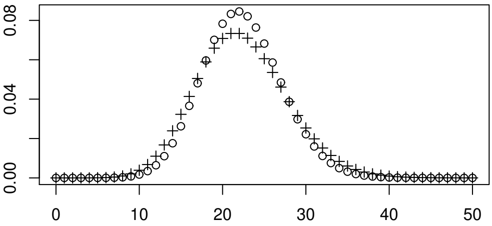

Figure: Two different ways of predicting the values of k_4, given the observations $k_1 = 20, k_2 = 24, k_3 = 23$. The pluses represent the Bayesian predictions using the posterior predictive; the circles represent the Frequentist predictions, using the Poisson distribution with parameter $(20 + 24 + 23)/3 = 22.33$.

Bayesian inference using discretization

If the sample space of θ is finite, Bayesian inference is quite easy:

- The prior distribution $\pi(\theta)$ is represented by a vector.
- The posterior distribution $\pi(\theta \mid y)$ is obtained by termwise multiplication of the vectors $\pi(y \mid \theta)$ and $\pi(\theta)$ and normalizing so the result sums to 1.
- The prediction $\pi(y_{new} \mid y) = \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta$ simplifies to taking the sum of the termwise product of the vectors $\pi(y_{new} \mid \theta)$ and $\pi(\theta \mid y)$.
- USAGE: Approximate a 1D (and 2D) prior $\pi(\theta)$ by finding $\theta_1, \dots, \theta_k$ equally spaced in the definition area for θ, compute $\pi(\theta_i)$ and normalize these values so that they sum to 1.
- Check out the R code in the example of Section 1.5 of the Compendium!

Bayesian inference using numerical integration

- The prediction we want to make can be expressed as a quotient of integrals:

$$
\begin{aligned}
\pi(y_{new} \mid y) &= \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta \\
&= \int_\theta \pi(y_{new} \mid \theta) \frac{\pi(y \mid \theta)\pi(\theta)}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta}\, d\theta \\
&= \frac{\int_\theta \pi(y_{new} \mid \theta)\pi(y \mid \theta)\pi(\theta)\, d\theta}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta}
\end{aligned}
$$

- One idea: Compute these integrals using numerical integration.
- Can work well as long as the dimension of θ is low (max 2 or 3?) and the functions are well-behaved.
- Check out the R code in the example of Section 1.6 of the Compendium!