MSA101/MVE187 2022 Lecture 2

Petter Mostad

Chalmers University

August 31, 2022

What do I expect from you in this course?

- Formal expectations:
  - Three individual obligatory assignments
  - A final written exam, determining the grade
- In addition, my actual expectations:
  - Get familiar with the information on the Canvas course page.
  - Read up on literature BEFORE lectures.
    Student: "I think maybe students should be encouraged to skim through the relevant book chapters before the lectures because it really helped me when doing so."
  - Be active in connection with lectures. Ask questions!
  - Take responsibility for learning assumed background knowledge, such as running R and basic probability. But also ask me for help!
  - Make sure you do exercises that help YOU learn. Take advantage of the exercise sessions.

What can you expect from the course?

- A Canvas course page, also used for handing in assignments.
- Two lectures each week. Three lectures will be with Umberto Picchini.
- One exercise session each week.
- Outside lectures and exercise sessions I will answer mail (and Canvas messages) when I have time.

Required knowledge

- in **basic probability theory**:
  - Basic knowledge of distributions, densities, conditional distributions, expectations ...
  - Some familiarity with standard distributions such as Binomial, Poisson, Gamma (but no need to memorize; check out old exam appendices!).
  - Consult your previous statistics/probability textbooks!
- in **classical statistics**:
  - ...not much, you have mostly seen this in the first lecture.
- in **computation**:
  - We use R. Learn R now!
  - ...in fact, no advanced programming is needed to get through this course.

Overview for today

- Definition and examples of conjugacy. How to compute in practice.
- Predictive distributions when using conjugate families.
- The exponential family of distributions.

Review from last lecture: Bayesian framework

- Prediction variable Y_{pred}, data Y_{data}, parameter θ.
- Specify a complete model by specifying prior $\pi(\theta)$, likelihood $\pi(Y_{data} \mid \theta)$, and prediction distribution $\pi(Y_{pred} \mid \theta)$.
- Derive the posterior $\pi(\theta \mid Y_{data})$.
- Make predictions using

$$\pi(Y_{pred} \mid Y_{data}) = \int \pi(Y_{pred} \mid \theta)\pi(\theta \mid Y_{data})\, d\theta$$

Review from last lecture: Notation

- For standard distributions, we use similar but different notation for a random variable itself, and its density (or probability mass function).
- Example: We write

$$Y \sim \text{Binomial}(N, p) \qquad \text{and} \qquad \pi(y) = \text{Binomial}(y; N, p)$$

- so we have

$$\text{Binomial}(y; N, p) = \binom{N}{y} p^y (1-p)^{N-y}.$$

- We define

$$\text{expression 1} \qquad \propto_\theta \qquad \text{expression 2}$$

  to mean that the second expression is equal to the first expression except for a factor that does not contain the variable θ.
- We say that expression 2 is proportional to expression 1 as a function of θ.
- For example

$$\binom{N}{y} \theta^y (1-\theta)^{N-y} \propto_\theta \theta^y (1-\theta)^{N-y}$$

Review from last time: The biased coin

- $Y_{pred} = 1$ or 0 (heads or tails). Y_{data}: Number of heads in N previous throws. θ: prob. of heads.
- We use $Y_{data} = y \sim \text{Binomial}(N, \theta)$ and $Y_{pred} \sim \text{Binomial}(1, \theta)$.
- We first used a prior with two possible values for θ: 0.7 and 0.3, with equal probabilities.
- We now compute the posterior when the prior is $\theta \sim \text{Uniform}(0, 1)$.

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 $\theta \sim \mathrm{Beta}(\alpha, \beta)$.

The biased coin, continued

- We get from the definition of Beta density that $\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta = B(\alpha,\beta)$.
- Show that the posterior becomes

$$\pi(\theta \mid y) = \frac{\theta^y(1-\theta)^{N-y}}{B(y+1, N-y+1)}.$$

- We see that

$$\theta \mid y \sim \mathrm{Beta}(y+1, N-y+1)$$

- NOTE: Computations can be made simpler, by not keeping track of factors not containing y!

Using a Beta distribution as prior

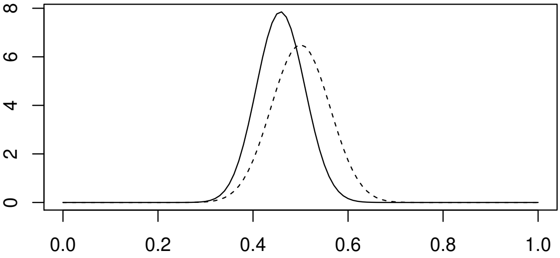
- Assume the prior is $\theta \sim \text{Beta}(\alpha, \beta)$. Compute the posterior!
- The posterior becomes

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

- DEFINITION: Given a likelihood model $\pi(y \mid \theta)$. A *conjugate family of priors* to this likelihood is a parametric family of distributions so that if the prior for θ is in this family, the posterior $\theta \mid y$ is also in the family.
- So the Beta family is conjugate to the Binomial likelihood: *The Beta-Binomial conjugacy*.
- NOTE: $\text{Uniform}(0, 1) = \text{Beta}(1, 1)$, so our previous example is part of this example.

Biased coin example, continued

- The prior $\pi(\theta) = 1$ may not be the most realistic.
- Better: $\pi(\theta) = \text{Beta}(\theta; 33.4, 33.4)$: Has 90% of its probability in the interval $[0.4, 0.6]$.



-
- The figure includes the posterior density $\text{Beta}(\theta; 33.4 + 11, 33.4 + 19)$.

Biased coin example, continued

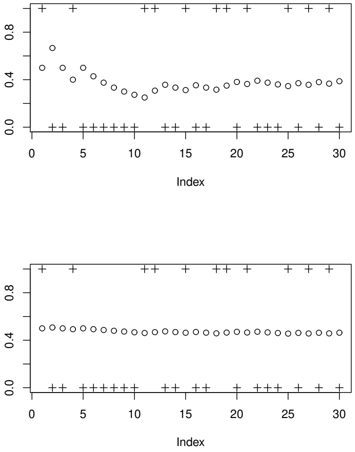


Figure: The probability of heads at each point in a sequence of observations, or the probability of "success", conditioning on the previous observations. The priors used are $\theta \sim \text{Uniform}(0,1)$ (left) and $\theta \sim \text{Beta}(33.4, 33.4)$ (right).

Example: The Poisson-Gamma conjugacy

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

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

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

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

- Compute the posterior!
- We get

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

- See Albert Section 3.3 for a computational example.

Example: The Normal-Gamma conjugacy

- Assume the likelihood is $\pi(y \mid \tau) = \text{Normal}(y; \mu, 1/\tau)$, so that y is normally distributed with known mean μ and unknown *precision* τ. The likelihood becomes

$$\pi(y \mid \tau) = \frac{1}{\sqrt{2\pi 1/\tau}} \exp\left(-\frac{1}{2/\tau}(y-\mu)^2\right) \propto_\tau \tau^{1/2} \exp\left(-\frac{1}{2}(y-\mu)^2 \tau\right)$$

- Prove: $\pi(\tau \mid \alpha, \beta) = \text{Gamma}(\tau; \alpha, \beta)$ is a conjugate family, where

$$\pi(\tau \mid \alpha, \beta) \propto_\tau \tau^{\alpha-1} \exp(-\beta\tau).$$

- Specifically, we get the posterior below:

$$\pi(\tau \mid y) = \text{Gamma}\left(\tau; \alpha + \frac{1}{2}, \beta + \frac{1}{2}(y-\mu)^2\right).$$

- We can also describe this conjugacy using the variance σ^2 and an inverse Gamma (or inverse Chi-squared) distribution.

Example: The Normal-Normal conjugacy

- Assume the likelihood is $\pi(y \mid \theta) = \text{Normal}(y; \theta, 1/\tau_0)$, where τ_0 is a known and fixed precision.
- Then $\pi(\theta \mid \mu, \tau) = \text{Normal}(\theta; \mu, 1/\tau)$, where τ is positive and μ has any real value, is a conjugate family.
- Specifically, we have the posterior

$$\pi(\theta \mid y) = \text{Normal}\left(\theta; \frac{\tau_0 y + \tau\mu}{\tau_0 + \tau}, \frac{1}{\tau_0 + \tau}\right)$$

- PROOF: Use completion of squares.

PROOF

$$
\begin{aligned}
\pi(\theta \mid y) &\propto_\theta \pi(y \mid \theta)\pi(\theta) \\
&\propto_\theta \exp\left(-\frac{\tau_0}{2}(y-\theta)^2\right)\exp\left(-\frac{\tau}{2}(\theta-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\tau_0 y^2 - 2\tau_0 y\theta + \tau_0\theta^2 + \tau\theta^2 - 2\tau\theta\mu + \tau\mu^2\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}\left[(\tau_0+\tau)\theta^2 - 2(\tau_0 y + \tau\mu)\theta\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}(\tau_0+\tau)\left(\theta - \frac{\tau_0 y + \tau\mu}{\tau_0+\tau}\right)^2\right) \\
&\propto_\theta \text{Normal}\left(\theta; \frac{\tau_0 y + \tau\mu}{\tau_0+\tau}, \frac{1}{\tau_0+\tau}\right)
\end{aligned}
$$

Conditionally independent data

- Assume $Y_{data} = (y_1, y_2)$, where y_1 and y_2 are conditionally independent given θ, i.e.,

$$\pi(y_1 \mid \theta, y_2) = \pi(y_1 \mid \theta).$$

- Then

$$\pi(\theta \mid y_1, y_2) \propto_\theta \pi(y_1, y_2 \mid \theta)\pi(\theta) = \pi(y_1 \mid \theta)\pi(y_2 \mid \theta)\pi(\theta)$$

- NOTE: We may first find the posterior given y_2, then use this posterior as the prior when finding the posterior given y_1: The result will be the posterior given y_1 and y_2.
- NOTE: We may **update** the prior on θ **sequentially** with data $y_1, y_2, \ldots, y_n$, as long as all the y_i are conditionally independent given θ.

Example: Normal distribution with fixed variance 1

- Assume $Y_{data} = (y_1, y_2, \ldots, y_n)$ where, independently given θ,

$$y_1, y_2, \ldots, y_n \sim \text{Normal}(\theta, 1)$$

- If the prior is $\theta \sim \text{Normal}(\mu, 1/\tau)$, we get

$$\theta \mid y_1 \sim \text{Normal}\left(\frac{y_1 + \tau\mu}{1+\tau}, \frac{1}{1+\tau}\right)$$

- Repeated updates give (writing $\overline{y} = (y_1 + \cdots + y_n)/n$)

$$\theta \mid y_1, \ldots, y_n \sim \text{Normal}\left(\frac{n\overline{y} + \tau\mu}{n+\tau}, \frac{1}{n+\tau}\right).$$

- We see that, using the *improper prior* $\pi(\theta) \propto_\theta 1$, or setting $\tau = 0$, gives the posterior $\text{Normal}(\overline{y}, 1/n)$.

Predictive distributions

- If $\pi(y \mid \theta)$ is a likelihood and $\pi(\theta)$ is some density, then

$$\pi(y) = \int \pi(y \mid \theta)\pi(\theta)\, d\theta$$

  is called a *predictive distribution*.

- If $y \mid \theta \sim \text{Binomial}(N, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$, show that

$$\begin{aligned} \pi(y) &= \int \text{Binomial}(y; N, \theta)\, \text{Beta}(\theta; \alpha, \beta)\, d\theta \\ &= \binom{N}{y} \frac{B(\alpha + y, \beta + N - y)}{B(\alpha, \beta)} \end{aligned}$$

- This is called a Beta-Binomial distribution:

$$\pi(y) = \text{Beta-Binomial}(y; N, \alpha, \beta).$$

Predictive distributions when you have conjugacy

- When $\pi(\theta)$ is in a conjugate family to $\pi(y \mid \theta)$, we can always analytically compute the integral defining the predictive distribution!
- In fact, we can always compute the predictive distribution without any integration at all! Use

$$\pi(y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(\theta \mid y)}$$

- Example: Compute the Beta-Binomial result above without considering integration.

Prior predictive / posterior predictive

- If $\pi(\theta)$ is considered a prior we call $\pi(y) = \int \pi(y \mid \theta)\pi(\theta)\, d\theta$ a *prior predictive*.
- If we condition on (conditionally independent) Y_{data}, we get

$$\pi(Y_{\text{pred}} \mid Y_{\text{data}}) = \int \pi(Y_{\text{pred}} \mid \theta)\pi(\theta \mid Y_{\text{data}})\, d\theta.$$

  It is the same type of formula, but $\pi(Y_{\text{pred}} \mid Y_{\text{data}})$ is now called the *posterior predictive*.
- NOTE: What can be considered a prior in one perspective can be considered a posterior in another perspective.

Predictive distribution for the Poisson-Gamma conjugacy

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

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

- 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 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 parameters after the update with data.

Poisson-Gamma example

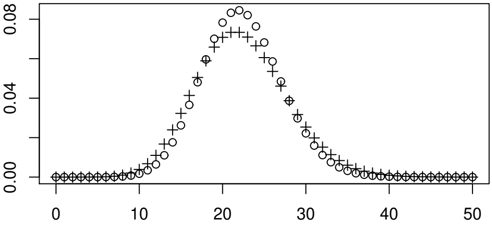


Figure: Two different ways of predicting the values of k_4, given the observations $k_1 = 20, k_2 = 24, k_3 = 23$ when $k_i \mid \theta \sim \text{Poisson}(\theta)$ and an improper $\text{Gamma}(0, 0)$ prior. The pluses represent the Bayesian predictions using the posterior predictive; the circles represent the Frequentist predictions, using the Poisson distribution with parameter $(20 + 24 + 23)/3 = 22.33$.

Example: Predictive distribution for the Normal-Normal conjugacy

- Assume $\pi(y \mid \theta) = \text{Normal}(y; \theta, 1/\tau_0)$ and $\pi(\theta) = \text{Normal}(\mu, 1/\tau)$.
- Instead of using the type of computations above, the following is simpler:
  - We know from general theory of the normal distribution that $\pi(y)$ is normal.
  - $E(y) = E(E(y \mid \theta)) = E(\theta) = \mu$.
  - $Var(y) = Var(E(y \mid \theta)) + E(Var(y \mid \theta)) = Var(\theta) + E(1/\tau_0) = 1/\tau + 1/\tau_0$.
- So for the prior predictive we get

$$\pi(y) = \text{Normal}(y; \mu; 1/\tau + 1/\tau_0)$$

Exponential distribution families

- Many parametric families of distributions can be written in a particular form:

  $$\pi(x \mid \eta) = h(x)g(\eta)\exp\left(\eta \cdot u(x)\right)$$

  where η and $u(x)$ are vectors, $\eta \cdot u(x)$ is their dot product, and η is called the "natural parameters" of the family.
- Some examples of exponential families of distributions, corresponding to particular choices of g, h, and u:
  - Normal distributions.
  - Beta distributions.
  - Poisson distributions.
  - Gamma distributions.
  - Bernoulli distributions and Binomial distributions for a fixed N.
  - Multinomial distributions for a fixed N.
  - ....and many more.
- Exponential families of distributions share many properties and can be studied together.

Conjugacies and exponential families

- If $\pi(x \mid \eta) = h(x)g(\eta) \exp(\eta \cdot u(x))$, then a conjugate family of priors for η is given as
$$\pi(\eta \mid \nu, \beta) \propto_\eta g(\eta)^\nu \exp(\eta \cdot \beta) .$$
The posterior becomes
$$\pi(\eta \mid x) \propto_\eta g(\eta)^{\nu+1} \exp(\eta \cdot (\beta + u(x))) .$$
- Essentially all examples of conjugacy fit into the framework above, so the above describes conjugacy in general.
- Note that the conjugate family of priors is also an exponential family.

Some properties

Assume $\pi(x \mid \eta) = h(x)g(\eta)\exp(\eta \cdot u(x))$.

- The expectation (and further moments) of $u(x)$ can be expressed with a differentiation of $g(\eta)$:

$$\mathrm{E}_{x|\eta}[u(x)] = -\nabla_\eta \log g(\eta).$$

- Given data $x_1, x_2, \ldots, x_N$ and a prior $\pi(\eta \mid \nu, \beta) \propto_\eta g(\eta)^\nu \exp(\eta \cdot \beta)$ the posterior becomes

$$\pi(\eta \mid x_1, \ldots, x_N) \propto_\eta g(\eta)^{\nu+N} \exp\left(\eta \cdot \left(\beta + \sum_{i=1}^{N} u(x_i)\right)\right).$$

- With for example a flat prior ($\mu = 0$, $\beta = 0$), the posterior is $\propto_\eta g(\eta)^N \exp\left(\eta \cdot \sum_{i=1}^N u(x_i)\right)$ and
  - The posterior (i.e., likelihood) depends only on $\sum_i u(x_i)$.
  - The maximum posterior (i.e., maximum likelihood) is the $\hat{\eta}$ satisfying

$$-\nabla_\eta \log g(\hat{\eta}) = \frac{1}{N} \sum_{i=1}^{N} u(x_i).$$