MVE550 2019 Lecture 2

Petter Mostad

Chalmers University

November 7, 2019

Basics of Bayesian inference

Outline:

- 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.

Inference

- We want to use *stochastic models* to make probabilistic predictions about future observations based on previous observations (i.e., data).
- Simple example: Predict the range of an electric car as normally distributed with *parameters* μ (expectation) and σ (standard deviation).
- Adapting a model to data is called *inference*.
- The *classical* or *frequentist* inference *paradigm*: Define a model in terms of *unknown parameters*, *estimate* these parameters using the data, and *predict* from the model with the estimated parameters plugged in.
- The *Bayesian* inference paradigm: Build a stochastic model (a probability distribution) with variables representing both observed data and the future data one would like to predict. Use for prediction the conditional distribution with data variables fixed to their observed values.

A biased coin example

You believe a coin is *biased*, and that the chance for heads is *either* 0.7 *or* 0.3. The probability for each of these possibilities is 0.5.

- Objective: Learn in which direction the bias goes by observing repeated throws of the coin.
- The probability of observing k heads in n throws is

$$\Pr(k) = 0.5 \cdot \text{Binomial}(k; n, 0.7) + 0.5 \cdot \text{Binomial}(k; n, 0.3).$$

- The probability of observing a *specific sequence* with k heads in n throws is

$$\Pr(k) = 0.5 \cdot 0.7^k 0.3^{n-k} + 0.5 \cdot 0.3^k 0.7^{n-k}.$$

- One can compute the probability of observing any *sequence*. The prediction for observing H after observing for example $HTTHTTT$ can be computed as

$$\frac{\Pr(HTTHTTTH)}{\Pr(HTTHTTT)}.$$

Biased coin example

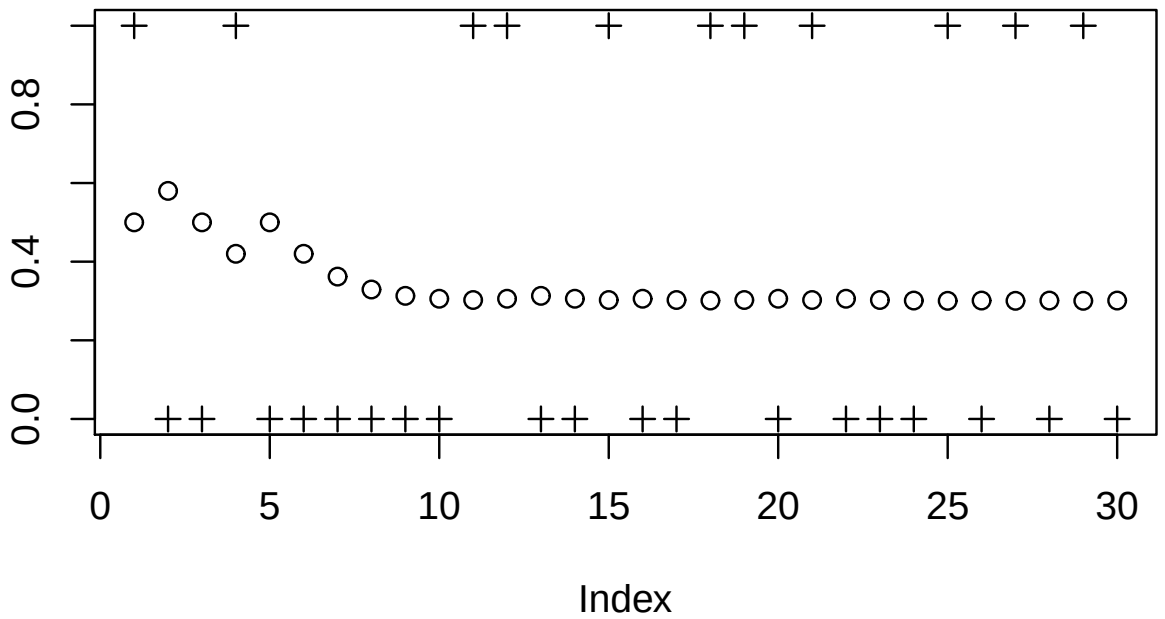

Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is $\theta = 0.7^x \cdot 0.3^{1-x}$ where $x \sim \text{Bernoulli}(0.5)$.

Example: Learning about a proportion

- An experiment is performed n times. We assume there is a probability θ for "success" each time, and that the outcomes are independent. Let X be the observed number of successes. We get $X \sim \text{Binomial}(n, \theta)$. Given $X = x$, what do we know about θ?
- For a Bayesian analysis, we need a joint probability density (or mass function) $\pi(X, \theta)$. We have defined $\pi(X \mid \theta)$ (the *likelihood*). We need to define $\pi(\theta)$ (the *prior*).
- Let us first try with the prior $\theta \sim \text{Uniform}[0, 1]$.

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

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

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

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

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, $\Gamma(n) = (n-1)! = 0 \cdot 1 \cdot \cdots \cdot (n-1)$. See for example Wikipedia for more properties of the Beta distribution, and the Beta and Gamma functions. We write $\pi(\theta \mid \alpha, \beta) = \mathsf{Beta}(\theta; \alpha, \beta)$ for the Beta density; we then also write $\theta \sim \mathsf{Beta}(\alpha, \beta)$.

Example: Learning about a proportion, continued

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

$$\theta \mid (X = x) \sim \text{Beta}\,(x + 1, n - x + 1)\,.$$

- The prediction we want can be found as an expecation of a Beta distribution:

$$\pi(y = 1 \mid k) = \mathsf{E}\,(\theta) = \frac{k+1}{k+2}.$$

- The conditional model $\pi(\theta \mid X = x)$ can be computed most easily using proportionality computaions. We get

$$\pi(\theta \mid X = x) \propto_\theta \theta^x (1-\theta)^{n-x}.$$

- We can then recognize this as a Beta distribution:
$\theta \mid X = x \sim \text{Beta}(x + 1, n - x + 1)$

Biased coin example

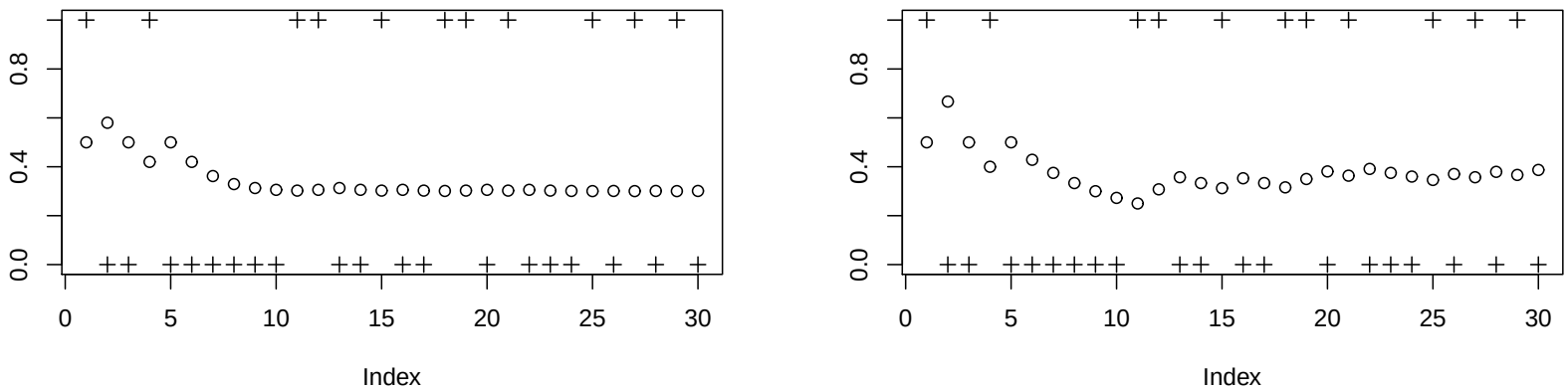

Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The priors used are $\theta = 0.7^x \cdot 0.3^{1-x}$ where $x \sim \text{Bernoulli}(0.5)$ (left), and $\theta \sim \text{Uniform}(0, 1)$ (right).

Using a Beta distribution as prior

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

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

- The prediction becomes

$$\pi(y = 1 \mid k) = \text{E}(\theta) = \frac{k + \alpha}{k + \alpha + \beta}.$$

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

Biased coin example

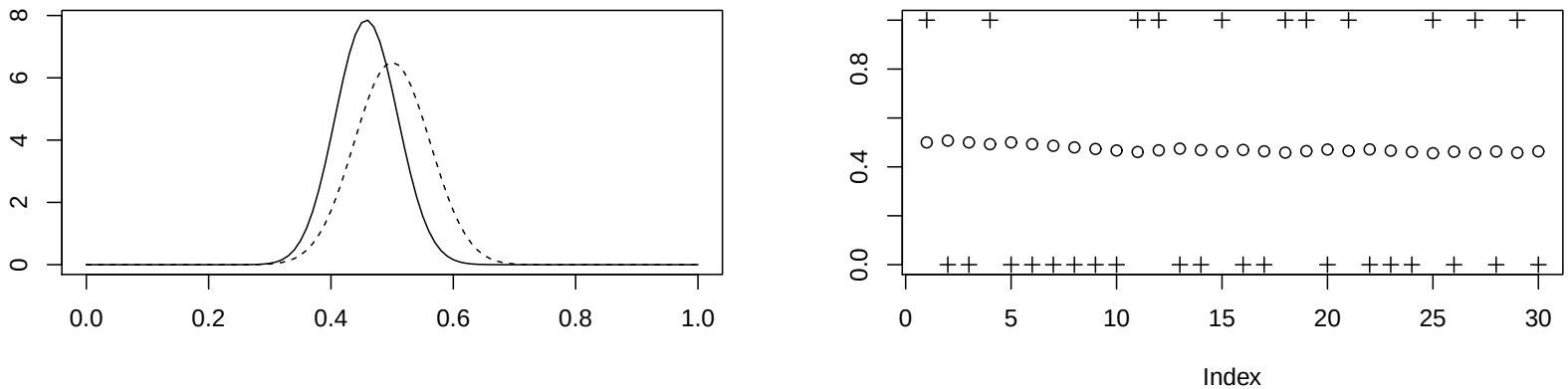

Figure: Left: The prior Beta$(33.4, 33.4)$ and the posterior Beta$(33.4 + 11, 33.4 + 19)$ for θ. Right: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails, using the shown prior.

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

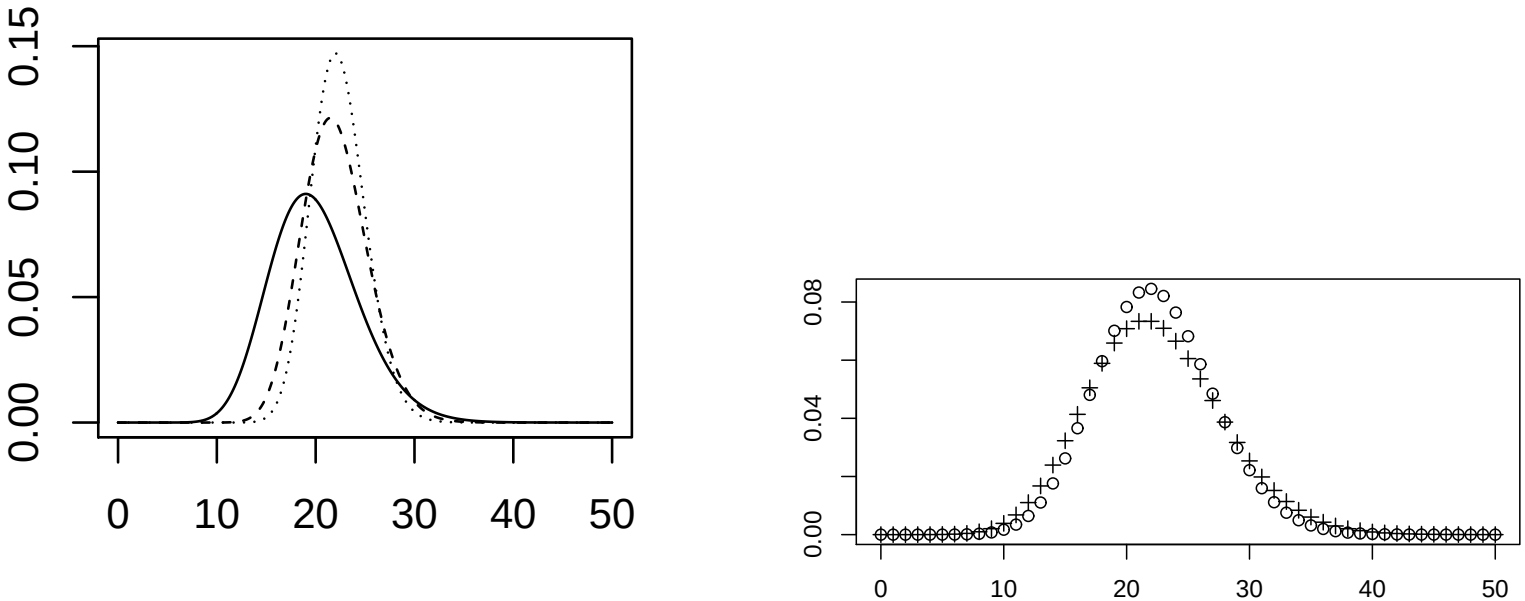

Figure: Left: The posteriors after one, two, and three observations, where $k_1 = 20, k_2 = 24$, and $k_3 = 23$. right: Two different ways of predicting the values of k_4, given the observations of k_1, k_2, k_3. The pluses represent the Bayesian predictions; the circles represent the Frequentist predictions, using the Poisson distribution with parameter $(20 + 24 + 23)/3 = 22.33$.

Prediction

The Bayesian paradigm implies:

- The usefulness of a model lies in its ability to predict.
- We create a joint probability model for the parameters θ, the observed data x, and data we would like to predict x_{new}. Often on the form $\pi(\theta, x, x_{new}) = \pi(\theta)\pi(x \mid \theta)\pi(x_{new} \mid \theta)$.
- The distribution for x_{new} is given by conditioning on the observed x and marginalizing out θ:

$$\begin{aligned} \pi(x_{new} \mid x) &= \int_\theta \pi(\theta, x_{new} \mid x)\, d\theta = \int_\theta \pi(x_{new} \mid \theta, x)\pi(\theta \mid x)\, d\theta \\ &= \int_\theta \pi(x_{new} \mid \theta)\pi(\theta \mid x)\, d\theta \end{aligned}$$

  This is called the *posterior predictive distribution*.
- It is also possible to look at the predictive distribution for x before it has been observed. This is called the *prior predictive distribution*:

$$\pi(x) = \int_\theta \pi(x, \theta)\, d\theta = \int_\theta \pi(x \mid \theta)\pi(\theta)\, d\theta$$

Predictive distributions when using conjugate priors

- When using a conjugate prior, not only do we have an analytic expression for the posterior density for θ, we also have analytic expressions for the prior predictive density and the posterior predictive density.
- To see this for the prior predictive density, use this formula derived from Bayes formula:

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

  The prior predictive density is on the left and all expressions on the right have analytic formulas.
- Note that, when using the right hand side for computing, θ will necessarily eventually disappear.
- As the posterior predictive distribution is on the same form as the prior predictive, we also get an analytic formula for it. Specifically, we can write

$$\pi(x_{new} \mid x) = \frac{\pi(x_{new} \mid \theta)\pi(\theta \mid x)}{\pi(\theta \mid x_{new}, x)}.$$

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 computaion gives the 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 if its probability mass function is

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

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

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 $\theta_1, \ldots, \theta_k$ equally spaced in the definition area for θ, compute $\pi(\theta_i)$ and normalize these values so that they sum to 1.

Bayesian inference using numerical integration

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

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

- 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.