Bayesian Theory and Computation

Lecture 2: Bayesian Inference for Simple Models

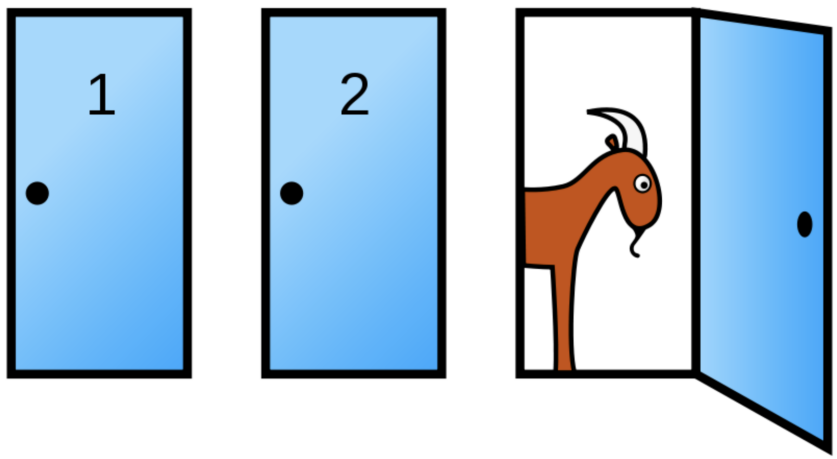
Cheng Zhang

School of Mathematical Sciences, Peking University

Feb 19, 2025

Monty Hall Problem

The problem is based on a TV game show hosted by Monty Hall. Suppose now your are on this show, and you are given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



- At the beginning, the car can be behind any of the three doors with equal probability. That is

$$p(D_1) = p(D_2) = p(D_3) = \frac{1}{3}$$

- Now you pick door No. 1, and Monty open door No. 3. The conditional probability of openning, OD_3, given the three possibilities (i.e., D_1, D_2 and D_3) are

$$p(OD_3|D_1) = \frac{1}{2}$$
$$p(OD_3|D_2) = 1$$
$$p(OD_3|D_3) = 0$$

- Now using the law of total probability we can find the marginal probability for opening door No. 3

$$p(OD_3) = \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2}$$

- Using Bayes' theorem, we have

$$P(D_2|OD_3) = \frac{p(D_2)p(OD_3|D_2)}{p(OD_3)} = \frac{\frac{1}{3} \times 1}{\frac{1}{2}} = \frac{2}{3}$$

$$p(D_1|OD_3) = \frac{p(D_1)p(OD_3|D_1)}{p(OD_3)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}$$

- Therefore, probability of winning doubles if you switch.

- Consider a set of observations $x = (x_1, \ldots, x_n)$. In constructing the joint distribution of these observations, we might believe that the indices are uninformative.
- For example, we toss an old-fashioned thumbtack on a soft surface and keep track of whether the sharp point is up or down. After n tosses, we believe the joint distribution remains the same regardless of which order we consider.
- In this experiment, we do not expect those tosses close together in time to be more similar to each other compared to other tosses.
- We also believe that the above comments are true for any subsets of tosses. That is, if $n = 100$, (x_4, x_{17}) has the same joint distribution as (x_{81}, x_{22}), and (x_{30}, x_{16}, x_{92}) has the same joint distribution of (x_{18}, x_{10}, x_{99}) and so forth

- Such *symmetry* or *similarity* could be expressed as

$$p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n})$$

  where π represent all permutations on $\{1, 2, \ldots, n\}$.
- We call a sequence to be *exchangeable* if this property holds for any finite subset of it.
- For a sequence of coin tossing, let's denote head as 1 and tail as 0. Then exchangeability means the joint probability of any fixed set of 0's and 1's does not change when we permute them.
- For example, if $n = 3$, then $p(100) = p(010) = p(001)$, i.e., there is nothing special about the location of 1. Also $p(110) = p(101) = p(011)$.

Exchangeability

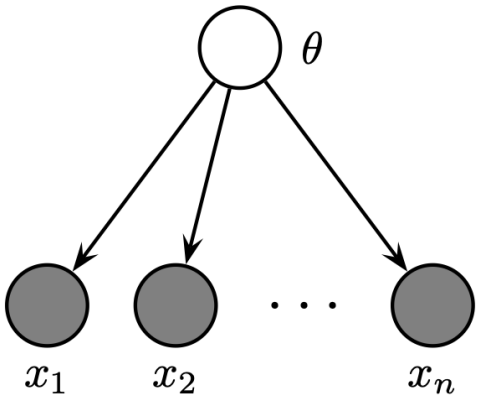
- We should be careful about our judgement of exchangeability.
- Consider the age of students in this class. Assume that all we know about students is their names.
- We might regard their age as exchangeable, which means students' names are not informative in defining the joint distribution.
- However, what if we also know whether a student is in a master's program or a PhD program?
- We know that in general, master students tend to be younger.
- Therefore, it would be more appropriate to assume exchangeability only within each group (i.e., master's and PhD).

- **De Finetti's Theorem** (1930s) A sequence of random variables $(x_1, x_2, \ldots)$ is infinitely exchangeable iff, for all n,

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta) \tag{1}$$

  for some measure P on θ.
- We can replace $P(d\theta)$ with $p(\theta)d\theta$ if the distribution on θ has a density.
- Clearly, the product $\prod_{i=1}^{n} p(x_i|\theta)$ is permutation invariant. Therefore, any sequence distribution that can be written as (1) for all n must be exchangeable.

De Finetti's Representation Theorem



By De Finetti's theorem, if we have exchangeable data

- There must exist a parameter θ.
- There must exist a likelihood $p(x|\theta)$.
- There must exist a distribution P on θ.
- The above quantities must exist so as to render the data conditionally independent.

This provide another evidence for the Bayesian formulation!

- To perform Bayesian inference, we need to specify the model $p(x|\theta)$ for the observed data and the prior $p(\theta)$ for the parameter of the model.
- The next step is to make probabilistic conclusions regarding the unobserved quantity θ given the observed data x, which is called *posterior* distribution.

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

- This simple formula is the essential part of Bayesian analysis. It is used not only for expressing updated belief about model parameters, but also for making decisions (e.g., accepting or rejecting a hypothesis) and predictions.

- Next, we will discuss some simple models commonly used for typical random variables.
- These models are based on our assumption for the underlying mechanism that generates the observed data.
- The focus in this model is on one single parameter, which represents the population mean.
- If there are other parameters in the model, we would regard them as nuisance parameters.
- Later, we will discuss multi-parameter models.

Binomial Model

- Consider a sequence of independent binary random variables, $x_1, x_2, \ldots$, such as "head/tails", "cancer/non-cancer", or "win/loss", such that $x_i \in \{0, 1\}$.
- Denote the probability of observing 1 as θ, then

$$x_i|\theta \sim \text{Bernoulli}(\theta), \quad p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- If $x_1, x_2, \ldots, x_n$ are iid (hence exchangeable) binary random variables with Bernoulli distribution, the sum $y = \sum_i x_i$ (i.e., number of 1's in the sequence) has a Binomial(n, θ) distribution

$$p(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Assuming the prior $p(\theta)$, the marginal distribution of y can be obtained as

$$p(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} p(\theta) d\theta$$

- Let's say we are quite ignorant about possible value of θ. That is to say, we think θ is uniformly distributed in $[0, 1]$, i.e., $p(\theta) = 1, 0 \leq \theta \leq 1$.
- Then we have

$$\begin{aligned} p(y) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \frac{n!}{y!(n-y)!} \text{Beta}(y+1, n-y+1) \\ &= \frac{n!}{y!(n-y)!} \cdot \frac{y!(n-y)!}{(n+1)!} = \frac{1}{n+1} \end{aligned}$$

- The posterior then can be evaluated easily

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{(n+1)!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \end{aligned}$$

- This is a Beta$(y+1, n-y+1)$ distribution with expectation

$$\mathbb{E}(\theta|y) = \frac{y+1}{n+2}$$

- Note that the posterior mean is a compromise between the prior mean and the sample proportion. As the sample size increases, the effect of prior diminishes.

Posterior Predictive Distribution

- Sometimes our objective is to use the posterior to predict future observations.
- That is, after observing some data, $y = \sum_i x_i$, we want to predict the next observation, x_{n+1}, which we denote as $\tilde{x}$.
- We can sum (or integrate) over posterior distribution of θ, i.e., $p(\theta|y)$, to form the *posterior predictive distribution*

$$p(\tilde{x}|y) = \int_0^1 p(\tilde{x}|\theta, y)p(\theta|y)d\theta$$

Since $\tilde{x}$ is independent of y given θ, we have

$$p(\tilde{x}|y) = \int_0^1 p(\tilde{x}|\theta)p(\theta|y)d\theta$$

- For the above binomial model, since $p(x = 1|\theta) = \theta$, the posterior predictive distribution can be obtained as

$$p(\tilde{x} = 1|y) = \int_0^1 \theta p(\theta|y) d\theta$$

- This is just the posterior mean of θ, which we computed before

$$p(\tilde{x} = 1|y) = \frac{y+1}{n+2}$$

Predicting The Election Result

- We want to predict which one of two candidates, A or B, will win the election.
- Let's denote the probability that A wins as θ, and we assume *a priori* the probability of winning for candidate A has uniform distribution.
- We ask 10 people which candidate they would choose in this election. Of 10 people surveyed, 3 people said they are going to vote for A.
- Our updated belief in A's winning has now a Beta$(4, 8)$ distribution.
- The posterior expectation of A's winning is $\frac{4}{12} \approx 0.33$, which is also the probability that the next person we survey votes for A.
- Note that this is almost the same as the maximum likelihood estimation $\frac{3}{10} = 0.3$

- In the above example, the derivation of posterior distribution was quite simple since it had a closed form.
- This was due to our choice of prior, i.e., uniform distribution. Note that uniform prior on $[0, 1]$ is in fact Beta$(1, 1)$ distribution.
- Therefore, for the above binomial model, both prior and posterior are Beta distributions.
- This is called "conjugacy" and the prior is called a "conjugate" prior.
- **Conjugacy** is informally defined as a situation where the prior distribution $p(\theta)$ and the corresponding posterior distribution, $p(\theta|y)$ belong to the same distribution family.
- Using conjugate priors makes sampling and Bayesian inference much easier compared to non-conjugate priors.

- Many widely used distributions, e.g., Normal, Bernoulli, Poisson, belong to a large class of distributions, called *exponential family*, that takes the following form

$$p(y_i|\theta) = h(y_i)g(\theta)\exp(\phi(\theta)^T s(y_i))$$

- $\phi(\theta)$ is called the *natural parameter* of the family
- The joint distribution for a set of conditionally independent observations, $y = (y_1, y_2, \ldots, y_n)$ is

$$p(y|\theta) = \prod_i h(y_i)g(\theta)^n \exp\left(\phi(\theta)^T \sum_i s(y_i)\right)$$

- $t(y) = \sum_i s(y_i)$ is a *sufficient statistic* for θ.

- For an exponential family, we have an natural choice of conjugate priors

$$p(\theta) \propto g(\theta)^{\eta} \exp\left(\phi(\theta)^T \nu\right)$$

- Easy to check the posterior would have a similar form

$$p(\theta|y) \propto g(\theta)^{\eta+n} \exp\left(\phi(\theta)^T (\nu + t(y))\right)$$

- In this case, $p(\theta)$ is a conjugate prior.

- Let's look at the binomial model again:

$$\begin{aligned} p(y|\theta, n) &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \\ &= \binom{n}{y} \exp\left[y \log\left(\frac{\theta}{1-\theta} \right) + n \log(1-\theta) \right] \\ &= \binom{n}{y} (1-\theta)^n \exp\left[y \log\left(\frac{\theta}{1-\theta} \right) \right] \end{aligned}$$

- Therefore,

$$g(\theta) = 1-\theta, \quad \phi(\theta) = \log\left(\frac{\theta}{1-\theta} \right) = \text{logit}(\theta)$$

- Recall that a conjugate prior is proportional to

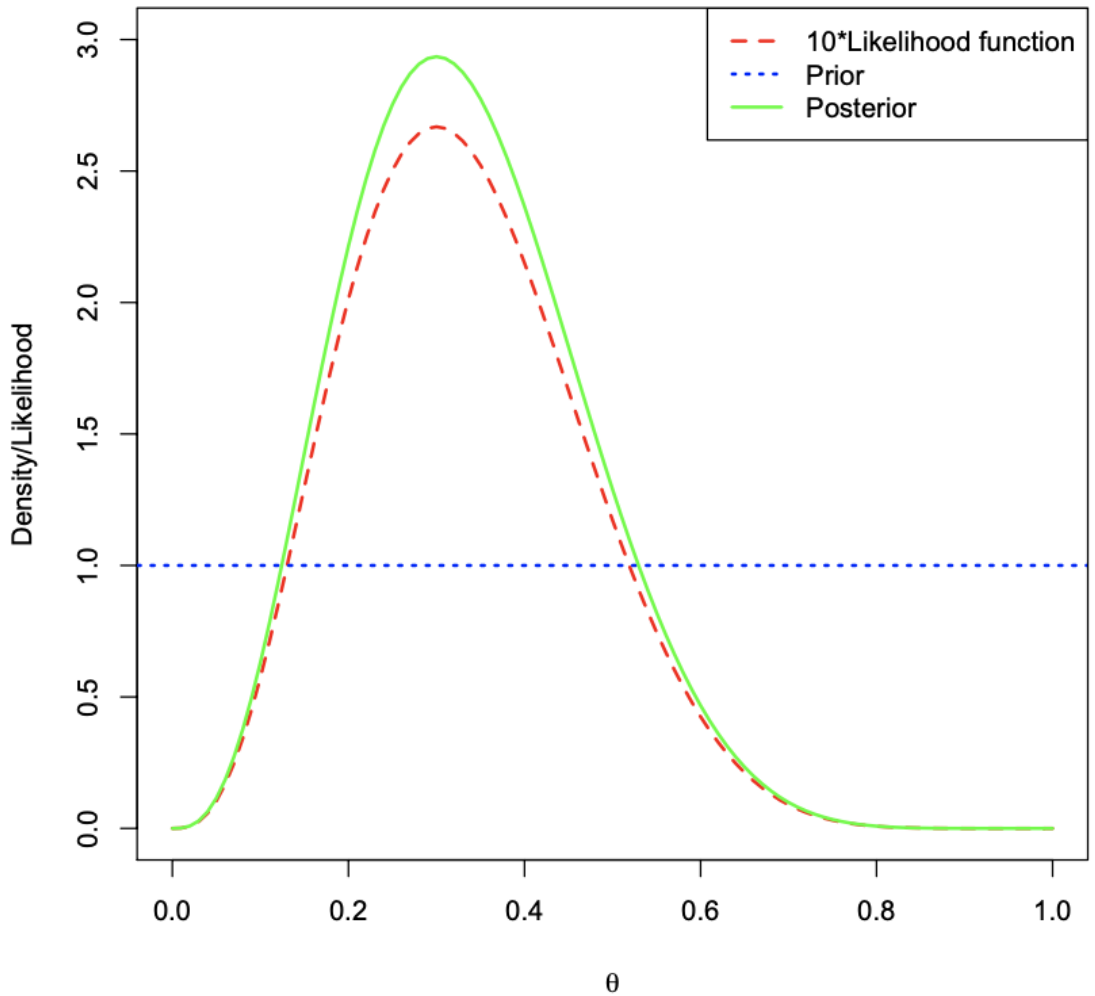
$$p(\theta) \propto g(\theta)^{\eta} \exp\left(\phi(\theta)^{T} \nu\right)$$

- Therefore, the conjugate prior for the Binomial model has the following form:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- This is a Beta(α, β) distribution.
- We can interpret this prior as observing $\alpha - 1$ prior success and $\beta - 1$ prior failure. That is, the prior acts as additional data.
- And the posterior distribution is Beta$(\alpha + y, \beta + n - y)$.
- Note that the uniform distribution we used before is in fact a special case of Beta distribution where $\alpha = \beta = 1$.

The Election Example

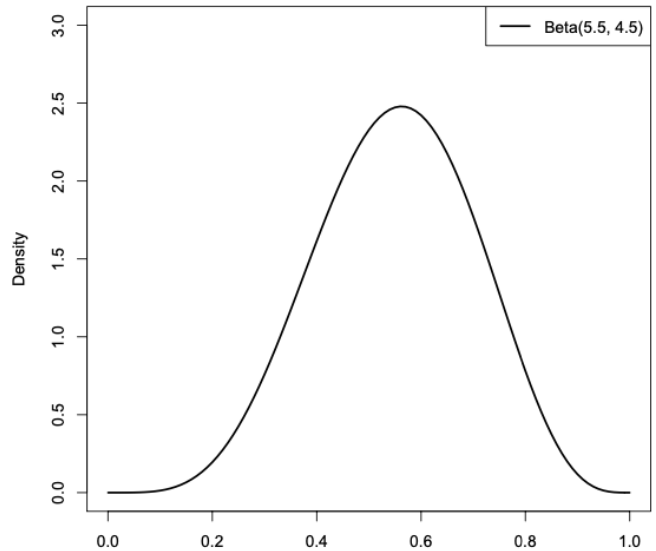


Election Example: Informative Prior

- As before, assume that we have surveyed 10 people and 3 of them are going to vote for candidate A.
- This time, however, we know that candidate A belongs to the party that in the previous elections won about 55% of votes.
- Instead of a uniform prior, we could use a more informative Beta prior which reflects such prior information.
- For example, we could choose a Beta prior whose mean is $\frac{\alpha}{\alpha+\beta} = 0.55$, and it is broad enough to reflect the extent of our uncertainty.
- We should always use a reasonably broad prior.

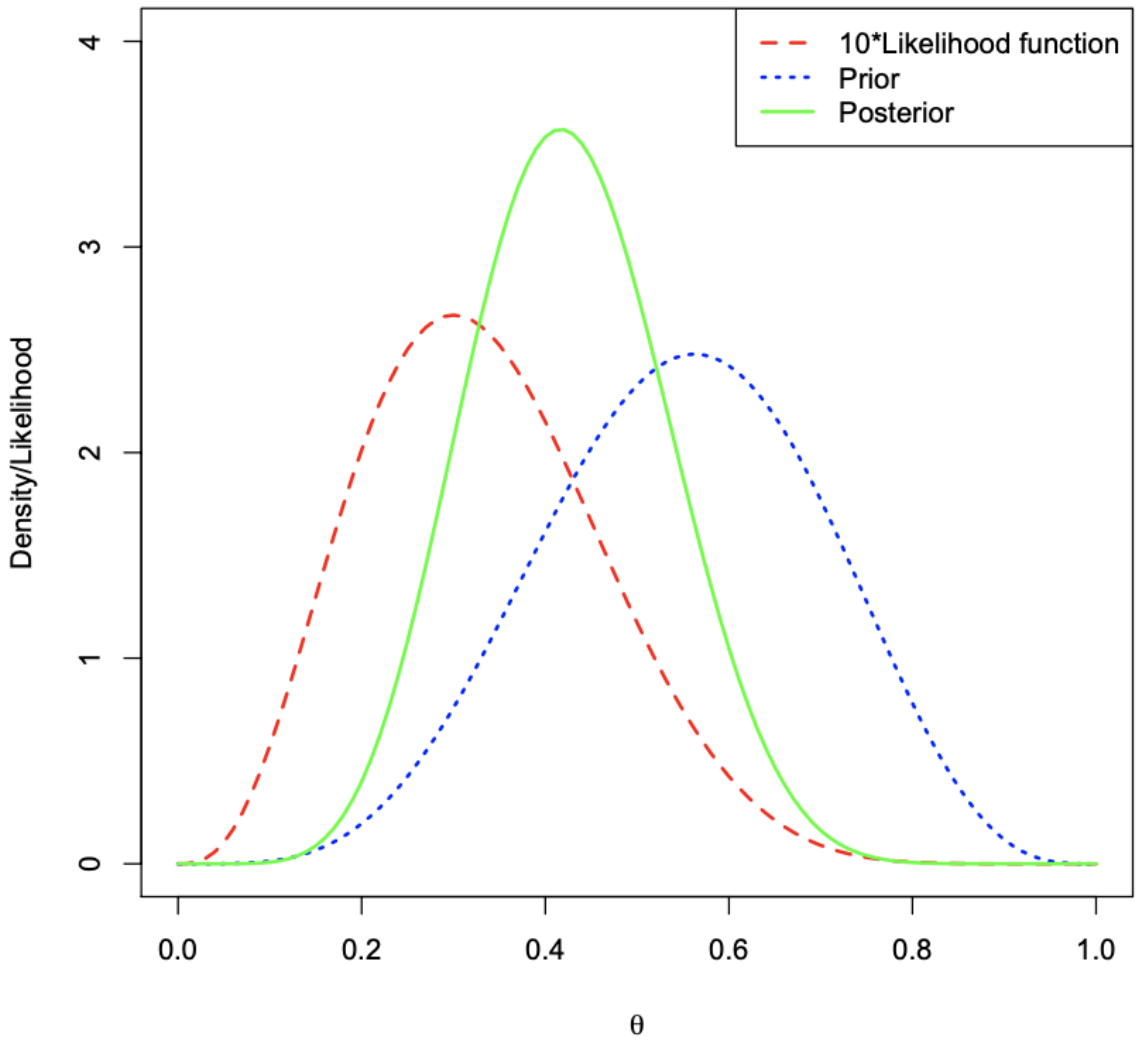
Election Example: Informative Prior

- We choose a $p(\theta) = \text{Beta}(5.5, 4.5)$ as our prior
- Plot you prior distribution or generate samples from it to make sure it is a good representation of your opinion.



- The posterior distribution now is $\text{Beta}(8.5, 11.5)$. So while the MLE is 0.3, the posterior expectation now is 0.425, which is a compromise between the observed data and the prior.

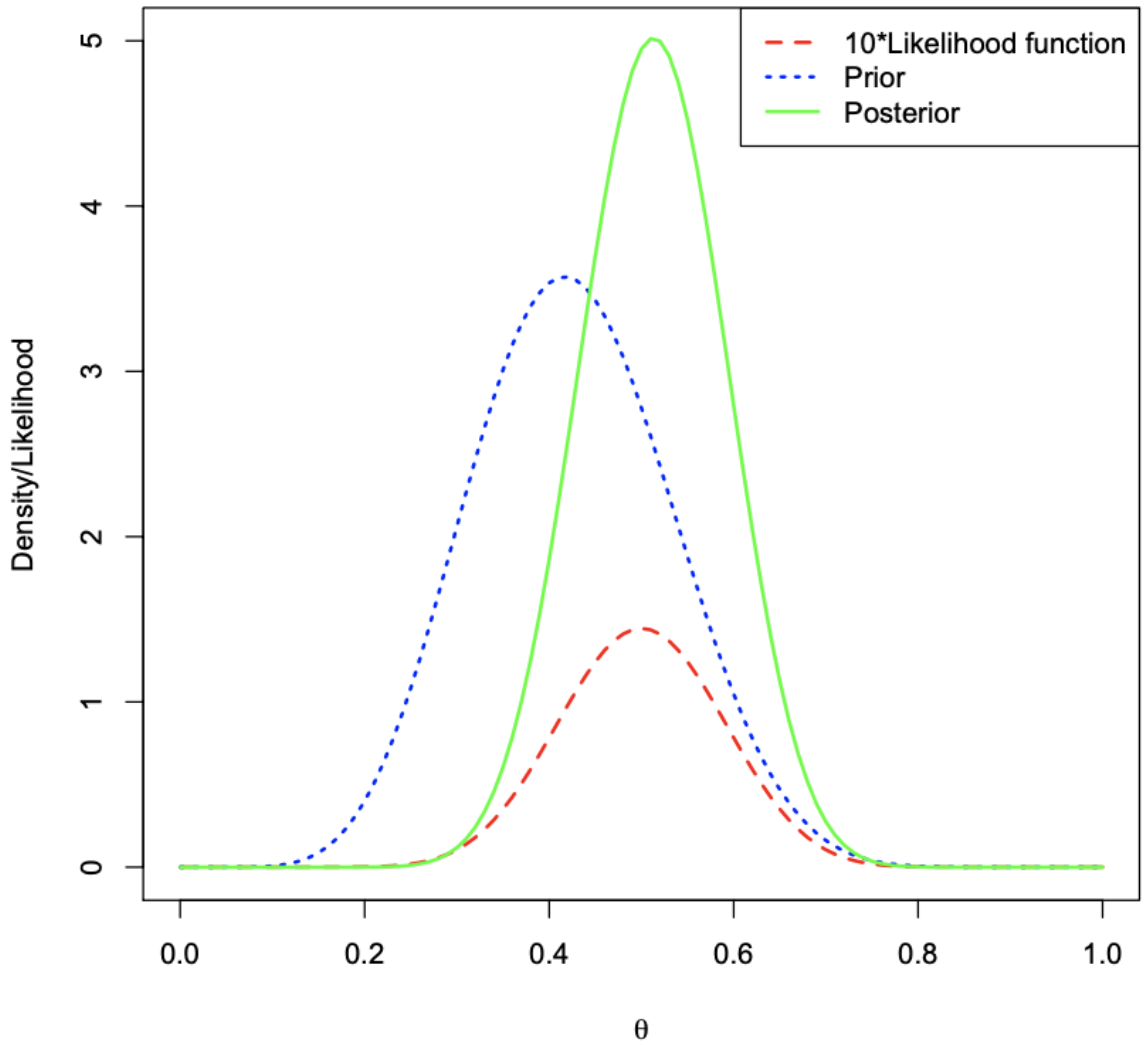
Election Example: Informative Prior



Election Example: More Data

- Now assume that we have obtained additional bugdet to survey 20 more people. The result shows that 12 out of 20 are going to vote for candidate A.
- It make sense to update our opinion based on this new information. It is also reasonable not to ignore the previous data.
- However, we do not need to start our analysis from the beginning. We can use the previous posterior distribution, $p(\theta) = \text{Beta}(8.5, 11.5)$, as our new prior and obtain a new posterior based on the more recent data.
- Our new posterior is therefore $p(\theta|y) = \text{Beta}(20.5, 19.5)$.
- The posterior expectation $\frac{20.5}{20.5+19.5} = 0.51$ and the MLE $15/30 = 0.5$ are now getting closer as the amount of data increases.

Election Example: More Data



Poisson Model

- Poisson model is another member of exponential family and is commonly used for count data.
- Assume we have observed $y = (y_1, y_2, \ldots, y_n)$:

$$p(y|\theta) = \prod_i \frac{\theta^{y_i} \exp(-\theta)}{y_i!} \propto \exp(-n\theta) \exp\left(\log\theta \sum_i y_i\right)$$

- The conjugate prior would have the following form:

$$p(\theta) \propto (\exp(-\theta))^\eta \exp(\nu \log\theta) \propto \exp(-\eta\theta)\theta^\nu$$

- Using $p(\theta) \propto \exp(-\beta\theta)\theta^{\alpha-1}$, which is a Gamma$(\alpha, \beta)$ distribution, as our prior, we have the following posterior

$$\theta|y \sim \text{Gamma}(\alpha + \sum_i y_i, \beta + n)$$

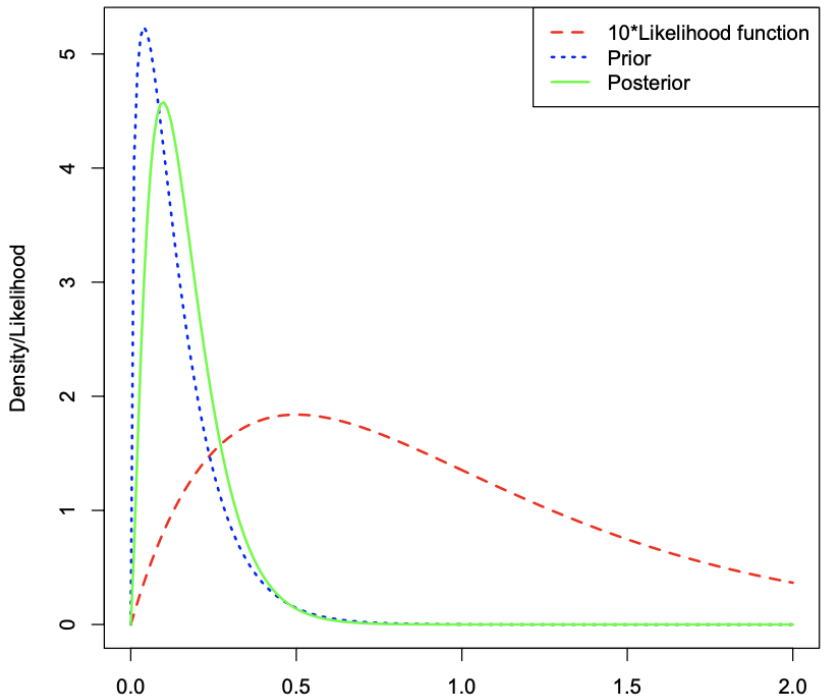
- When David Beckham joined LA Galaxy, he scored one goal in his first two MLS games.
- Assume that after the manager of LA Galaxy wanted to predict the number of goals Beckham would score in the remaining games.
- We model the number of goals, y_i, he scores in a game using a Poisson model with parameter θ.
- The maximum likelihood estimate is $\hat{\theta} = 0.5$.
- Now let's use a Gamma(α, β) prior for θ.
- Since we don't have a clue for α and β, we should use a noninformative prior that reflects our lack of information.
- Alternatively, we might want to use Beckham's history in Real Madrid to build a prior opinion.

- When in Madrid, Beckham scored 3 goals in 22 games (i.e., $3/22 \approx 0.14$ on average) during 06-07 season.
- We could choose a Gamma prior with mean around 0.14, for example, making sure it is broad enough to reflect our uncertainty.
- Be careful when working with Gamma distribution since there are two different ways of parameterizing it. Here, we use the form

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(a)} x^{\alpha-1} \exp(-\beta x).$$

- The mean of Gamma(α, β) is α/β.
- For our example, we could use the conjugate Gamma$(1.4, 10)$ prior with mean $1.4/10 = 0.14$.

- Since Gamma is a conjugate prior for the parameter of Poisson model, the posterior also has a Gamma distribution, which in this case is a Gamma$(1.4+1, 10+2)$ distribution.
- The expected number of goals is therefore $2.4/12 = 0.2$



- Posterior is again a compromise between the prior and the data (likelihood).
- In this example, as shown in the graph, the posterior is more similar to the prior than the likelihood.
- This is due to the fact that the amount of data is small.
- As the amount of data increases the influence of prior on posterior decreases while the effect of likelihood increases.
- In 2008-2009, Beckham played 25 games and scored 5 goals. This is a 0.2 average, which is much closer to our estimate compared to the MLE, which is 0.5

- The normal distribution is also a member of exponential families.
- We first consider a situation where there is only one observation and the variance is known.

$$y \sim \mathcal{N}(\theta, \sigma^2), \quad p(y|\theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$$

- So the general form of a conjugate prior is $p(\theta) \propto \exp(a\theta^2 + b\theta)$, which can be parameterized as

$$p(\theta) \propto \exp\left(-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right)$$

a $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution.

- As a result, the posterior distribution would be

$$p(\theta|\sigma, y) \propto \exp\left(-\frac{(y-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\tau_0^2}\right)$$

- When you complete the square, the posterior would also become a normal distribution:

$$p(\theta|\sigma, y) \propto \exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right)$$

which is a $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution with

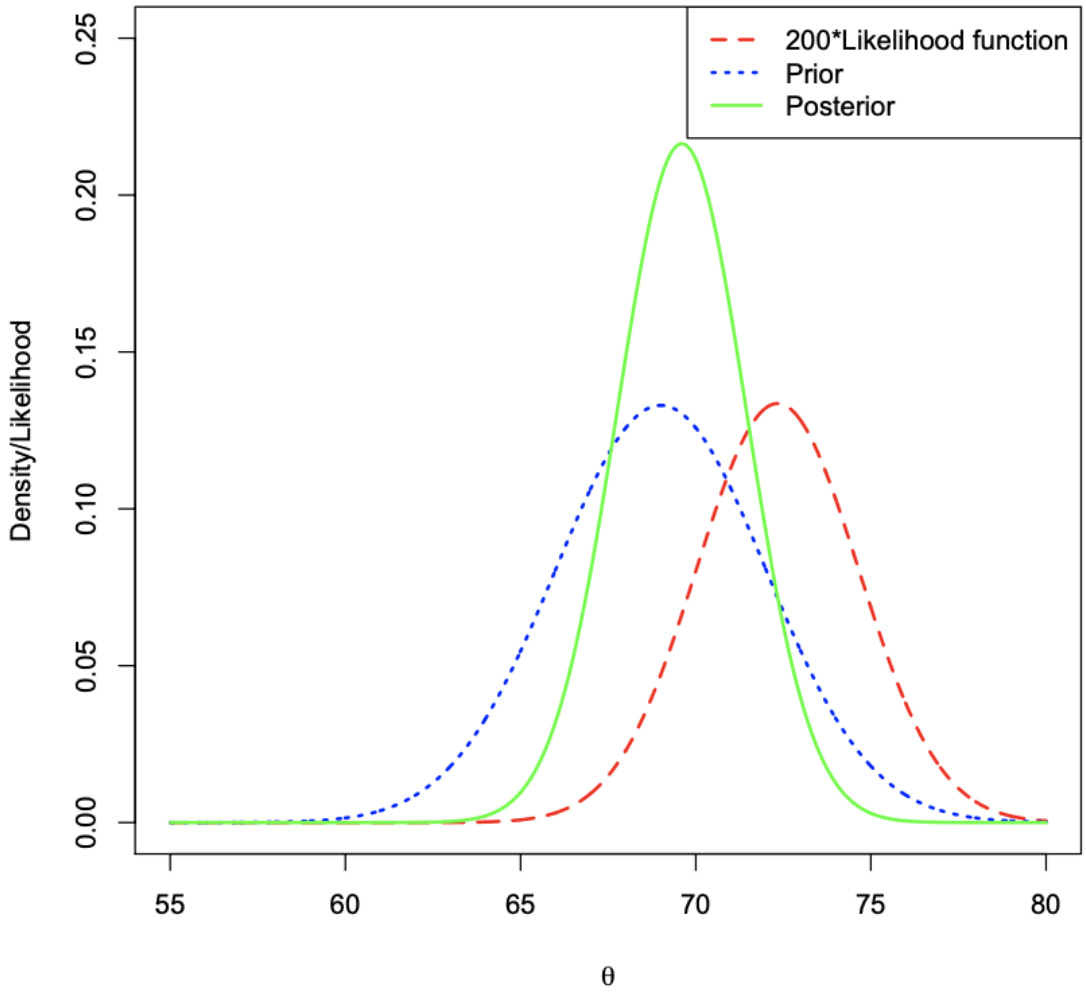
$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- Similarly, for n observations, we update the likelihood with $\bar{y}$ and the posterior is a $\mathcal{N}(\mu_n, \sigma_n^2)$ distribution with

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- Let's assume that the height (in inch) of students in this class follows a normal distribution $\mathcal{N}(\theta, 16)$. We use a $\theta \sim \mathcal{N}(65, 9)$ prior. We measure the height of three students: $y_1 = 72, y_2 = 75, y_3 = 70$.
- The posterior is $\theta|y \sim \mathcal{N}(69.6, 3.4)$.
- The role of prior is substantial here due to the small sample size. The prior modifies the likelihood based estimate (i.e., $\bar{y} = 72.3$), which could have been misleading since the data happened to be from tall people.

Again, the posterior distribution could be interpreted as a compromise between the prior and the likelihood



- In the above example, what would be our prediction for the height of next person we observe?
- Denote our prediction as $\tilde{y}$, and the corresponding distribution as $p(\tilde{y}|y)$, i.e., the posterior predictive probability. As before,

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

- By integrating out θ, the conditional distribution of $\tilde{y}$ given y is normal with the following mean and variance:

$$\begin{aligned}
\mathbb{E}(\tilde{y}|y) &= \mathbb{E}(\mathbb{E}(\tilde{y}|\theta, y)|y) = \mathbb{E}(\theta|y) = \mu_n \\
\mathbb{V}\text{ar}(\tilde{y}|y) &= \mathbb{E}(\mathbb{V}\text{ar}(\tilde{y}|\theta, y)|y) + \mathbb{V}\text{ar}(\mathbb{E}(\tilde{y}|\theta, y)|y) \\
&= \mathbb{E}(\sigma^2|y) + \mathbb{V}\text{ar}(\theta|y) = \sigma^2 + \tau_n^2
\end{aligned}$$

- We could use μ_n, the posterior expectation of θ, as our single point estimate for $\tilde{y}$.
- The variation around this estimate (i.e., our uncertainty) comes from two different sources: σ^2, the sampling variation (which is assumed fixed here) of data according to the model, and τ_n^2, the posterior variation of the model parameter θ, given the observed data.
- In the height example, our guess for the height of the fourth student can be expressed by a $\mathcal{N}(69.6, 19.4)$ distribution.

- For situations where the mean is fixed and the variance, σ^2, is the parameter of interest, we use the following scaled inverse-χ^2 prior:

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \quad p(\sigma^2) \propto \frac{1}{\sigma^{2+\nu_0}} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

  where ν_0 is the degree of freedom and σ_0 is the scale parameter.
- The posterior would also be scaled inverse-χ^2 with $\nu_0 + n$ degrees of freedom and scale equal to $\frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (y_i - \mu)^2}{\nu_0 + n}$.
- Recall that scaled inverse-χ^2 can be reviewed as Gamma distribution with a different parameterization (for precision $\gamma^2 = 1/\sigma^2$). Therefore, we can also use a Gamma prior as conjugate prior.

- When both μ and σ^2 are unknown, the only way to make the priors conjugate is to make the prior for μ dependent on σ^2 as follows:

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/k_0)$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

- In general, we do not recommend this prior
- As we will see later, we don't have to specify our prior this way.

- This is a multiparameter generalization of binomial distribution
- For example, in the election problem, we might have more than two candidates.
- If y has a multinomial distribution with J groups, the sampling distribution would have the following form

$$p(y|\theta) \propto \prod_{j=1}^{J} \theta_j^{y_j}$$

where $\sum_j \theta_j = 1$ and $\theta_j \geq 0, j = 1, \ldots, J$.

- The conjugate prior for this model is the Dirichlet distribution

$$p(\theta|\alpha) \propto \prod_{j=1}^{J} \theta_j^{\alpha_j - 1}, \quad \theta_j \geq 0, \sum_{j=1}^{J} \theta_j = 1, \alpha_j > 0$$

  which is a multivariate generalization of the Beta distribution
- For this distribution, $\mathbb{E}(\theta_j) = \alpha_j / \sum_{j'} \alpha'_j$.
- A special case is the *symmetric Dirichlet distribution*, where $\alpha_j = \alpha, j = 1, \ldots, J$. This is useful when no prior preference is available. The scalar parameter α is called the concentration parameter.

- The posterior is also a Dirichlet distribution

$$p(\theta|\alpha, y) \propto \prod_{j=1}^{J} \theta_j^{y_j+\alpha_j-1}$$

- Consider the election example. Let's assume another candidate, C, enters the race and a new poll shows that out of 100 people surveyed 24 people vote for A, 45 for B, and 31 for C.
- Let's denote the probability of winning by θ_j, where $j \in \{A, B, C\} \equiv \{1, 2, 3\}$. Assume a symmetric Dirichlet prior with $\alpha = 1$.
- The posterior distribution of θ has a Dirichlet$(25, 46, 32)$. The probability of winning (i.e., $\mathbb{E}(\theta_j)$) for candidate A, B, C becomes $25/103, 46/103$, and $32/103$ respectively.

- For multivariate normal distribution with known covariance, Σ, we assume

$$x \sim \mathcal{N}(\mu, \Sigma)$$
$$\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$$

The posterior distribution of μ given n observations is also a multivariate normal distribution,

$$\mu | x \sim \mathcal{N}(\mu_n, \Sigma_n)$$

where

$$\mu_n = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})$$
$$\Sigma_n = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$$

- For multivariate normal distribution with known mean, μ, we assume

$$x \sim \mathcal{N}(\mu, \Sigma)$$
$$\Sigma \sim \text{Inv-Wishart}(\nu_0, \Lambda_0)$$

- The posterior distribution of Σ given n observations is also an inverse Wishart distribution,

$$\Sigma | x \sim \text{Inv-Wishart}(\nu_n, \Lambda_n)$$

where

$$\nu_n = \nu_0 + n$$
$$\Lambda_n = \Lambda_0 + \sum_i (x_i - \mu)(x_i - \mu)^T$$

References

- M. J. Schervish. Theory of Statistics. Springer. 1995.