STA 630: Bayesian Inference - Lecture 2

Zeya Wang

University of Kentucky

Spring 2025

Lecture 3: Conjugate Models - Outline

- Conjugate priors
- Exponential family
- Poisson model
- Monte Carlo estimates

Conjugate Priors

- Example:

  Likelihood $x \mid \theta \sim \text{Bin}(n, \theta)$ and prior $\theta \sim \text{Beta}(\alpha, \beta)$, then posterior is $\theta \mid x \sim \text{Beta}(\alpha + x, \beta + n - x)$

  The chosen prior results in a posterior which takes the same parametric form (typically same functional form of the likelihood)

- Definition:

  $\mathcal{F}$ is a family of distributions $p(x \mid \theta)$ for the data

  $\mathcal{P}$ is a family of prior distributions $p(\theta)$

  Then, $\mathcal{P}$ is conjugate with respect to $\mathcal{F}$ if $p(\theta \mid x)$ is in $\mathcal{P}$ for all $p(x \mid \theta)$ in $\mathcal{F}$ and $p(\theta)$ in $\mathcal{P}$.

- Advantage: Easy to derive the posterior distribution. Disadvantage: Expert opinion may not conform to conjugate prior.

Conjugate prior for exponential family

Suppose we have a random sample $x = (x_1, \ldots, x_n)$ from an exponential family distribution

$$p(x \mid \theta) = f(x)g(\theta)e^{\phi(\theta)^T u(x)} \longrightarrow L(\theta) \propto \prod_{i=1}^{n} p(x_i \mid \theta)$$

[Binomial: $\phi(\theta) = \text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right), g(\theta) = \left(1 + e^{\phi(\theta)}\right)^{-1}, u(x) = x$]

The conjugate prior for θ is $p(\theta) \propto [g(\theta)]^{\eta} e^{\phi(\theta)^T \nu}$, with posterior

$$p(\theta \mid x) \propto [g(\theta)]^{\eta+n} e^{\phi(\theta)^T (\nu + t)}$$

with $t = t(x) = \sum_i u(x_i)$ the sufficient statistics for θ (i.e., it contains all info about θ available in the data; $L(\theta)$ depends on the data only via t).
Prior inputs η and ν are hyperparameters
Examples: Binomial (n. successes in n trials); Poisson (n. occurrences in a given interval); Exponential (waiting times); Normal distribution

Conjugate prior for exponential family

Sampling distr	Conjugate prior
$\text{Bin}(n, \theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$
$\text{Poisson}(\theta)$	$\theta \sim Ga(\alpha, \beta) \propto e^{-\beta\theta}\theta^{\alpha-1}$
$\text{Exp}(\theta)$	$\theta \sim Ga(\alpha, \beta)$
$\text{Gamma}(\mu, \theta), \mu$ known	$\theta \sim Ga(\alpha, \beta)$
$\text{Beta}(\theta, \beta), \beta$ known	$\theta \sim Ga(\alpha, \gamma)$
$N(\theta, \sigma_0^2), \sigma_0$ known	$\theta \sim N(\mu, \tau^2)$
$N(\mu, \sigma^2), \mu$ known	$\tau = 1/\sigma^2 \sim Ga(\alpha, \beta)$
	$\sigma^2 \sim \text{Inv} - Ga(\alpha, \beta) \propto e^{-\beta/\sigma^2}/\sigma^{2(\alpha-1)}$

Example of non-conjugate prior: $N(\theta, \sigma_0^2)$ with $\theta > 0$. Possible priors: Gamma or truncated normal $\theta \sim N(\mu, \tau^2)\, I(\theta > 0)$. Analytical derivation of posterior difficult or impossible, and one typically needs to numerically compute posterior summaries.

Poisson Model

Random sample $x_1, \ldots, x_n$ from a Poisson (λ). Pdf of X is

$$p(x \mid \lambda) \propto \frac{\exp(-\lambda)\lambda^x}{x!}, \quad L(\lambda) \propto \prod_{i=1}^{n} p\left(x_i \mid \lambda\right) \propto e^{-n\lambda}\lambda^t$$

with $t = \sum x_i$.
Conjugate prior for λ is $\text{Ga}(\alpha, \beta), \pi(\lambda) \propto e^{-\beta\lambda}\lambda^{\alpha-1}$
Posterior distribution is $p(\lambda \mid x) \propto L(\lambda)\pi(\lambda) \propto e^{-(n+\beta)\lambda}\lambda^{t+\alpha-1}$, that is

$$\lambda \mid x \sim \text{Gamma}(\alpha + t, \beta + n)$$

Point estimates: $E[\lambda \mid x] = \frac{\alpha+t}{\beta+n}, \text{var}[\lambda \mid x] = \frac{\alpha+t}{(\beta+n)^2}$

$\left[E[\lambda \mid x] = \frac{\beta}{\beta+n}\frac{\alpha}{\beta} + \frac{n}{\beta+n}\frac{t}{n} = w\times \text{ prior mean } +(1-w)\times \text{ MLE, } w = \beta/(\beta+n)\right]$
For large n and fixed $\beta, w \approx 0$. For fixed n and $\beta \approx 0$ (non-inf prior), $w \approx 0$ (point estimate $\approx$ MLE)
95% credible interval for $\lambda : (a, b)$ with 2.5 th and 97.5 th percentiles

Practice2: Example on birth rates

Data gathered by the General Social Survey during the 1990s on the educational level and number of children of 155 women who were 40 years old at the time of the survey. Aim is to compare the women with college degrees to those without. Let $Y_{1,1}, \ldots, Y_{n_1,1}$ be the number of kids of the n_1 women without college degrees and $Y_{1,2}, \ldots, Y_{n_2,2}$ be the number of kids of the n_2 women with college degrees.
Assume

$$Y_{1,1}, \ldots, Y_{n_1,1} \mid \theta_1 \sim \text{ i.i.d.Poisson } (\theta_1), Y_{1,2}, \ldots, Y_{n_2,2} \mid \theta_2 \sim \text{ i.i.d.Poisson } (\theta_2)$$

Data:
$n_1 = 111, \sum_i Y_{i,1} = 217, \bar{Y}_1 = 1.95, \quad n_2 = 44, \sum_i Y_{i,2} = 66, \bar{Y}_2 = 1.50$
Prior: Choose $\theta_i \sim Ga(2,1), i = 1,2$
Posterior: $\theta_1 \,|\, Y_1 \sim Ga(217+2, 111+1), \quad \theta_2 |\, Y_2 \sim Ga(66+2, 44+1)$
See plot of posteriors and code for posterior means and credible intervals.
Conclude that $P(\theta_1 > \theta_2 \mid Y_1, Y_2) = .97$.

Practice2: Example on birth rates

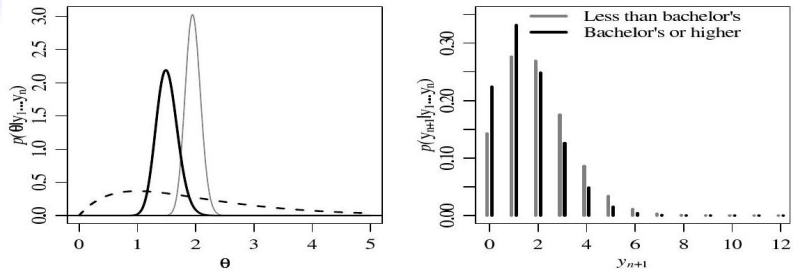


Figure: Prior (dashed line) and posterior distr; predictive distr for n . kids

```
a<-2; b<-1; n1<-111; sy1<-217; n2<-44; sy2<-66
(a+sy1)/(b+n1), (a+sy2)/(b+n2) # posterior means
[1] 1.955357, 1.511111
qgamma(c(.025, .975), a+sy1, b+n1) # credible interval
[1] 1.704943 2.222679
qgamma(c(.025, .975), a+sy2, b+n2)
[1] 1.173437 1.890836
```

Predictive Distribution $p(z \mid x), z = 0, 1, \ldots$ for a future z

$$
\begin{aligned}
p(z \mid x) &= \int p(z \mid \lambda) p(\lambda \mid x) d\lambda \\
&= \int_0^\infty \frac{\exp(-\lambda)\lambda^z}{z!} \frac{(\beta+n)^{t+\alpha}}{\Gamma(t+\alpha)} e^{-(n+\beta)\lambda} \lambda^{t+\alpha-1} d\lambda \\
&= \frac{(\beta+n)^{t+\alpha}}{z!\Gamma(t+\alpha)} \int_0^\infty \exp(-\lambda(\beta+n+1)) \lambda^{z+t+\alpha-1} d\lambda
\end{aligned}
$$

We know that is $\lambda \sim Ga(\alpha, \beta)$ then $\int \frac{\beta^\alpha e^{-\beta\lambda}\lambda^{\alpha-1}}{\Gamma(\alpha)} d\lambda = 1$, that is $\frac{\Gamma(\alpha)}{\beta^\alpha} = \int e^{-\beta\lambda}\lambda^{\alpha-1} d\lambda$. Substitute $\alpha \leftarrow \alpha + t + z$ and $\beta \leftarrow \beta + n + 1$

$$
\begin{aligned}
p(z \mid x) &= \frac{(\beta+n)^{t+\alpha}}{z!\Gamma(t+\alpha)} \frac{\Gamma(\alpha+t+z)}{(\beta+n+1)^{\alpha+t+z}} \\
&= \binom{z+t+\alpha-1}{z} \left(\frac{\beta+n}{\beta+n+1}\right)^{t+\alpha} \left(\frac{1}{\beta+n+1}\right)^z
\end{aligned}
$$

Predictive Distribution $p(z \mid x), z = 0, 1, \ldots$ for a future z

That is $z \mid x \sim Neg - \text{binomial}(\alpha + t, \beta + n)$

$$z \mid x \sim \text{Neg} - \text{binomial}(\alpha + t, \beta + n)$$

[General result: $\text{Neg} - \text{Bin}(x \mid \alpha, \beta) = \int \text{Poisson}(x \mid \lambda) Ga(\lambda \mid \alpha, \beta) d\lambda$]
$E[z \mid x] = \frac{\alpha+t}{\beta+n}, \quad \text{var}[z \mid x] = \frac{\alpha+t}{\beta+n} \frac{\beta+n+1}{\beta+n}$
Back to the example: See plots of predictive densities. Conclude that
$P(z_1 > z_2 \mid Y_1, Y_2) = .48$ and $P(z_1 = z_2 \mid Y_1, Y_2) = .22$ [strong evidence of difference between populations does not mean that the actual difference is large]

```
y<- 0:10
dnbinom(y, size=(a+sy1), mu=(a+sy1)/(n+n1))
dnbinom(y, size=(a+sy2), mu=(a+sy2)/(n+n2))
```

Other model: The exponential distribution (for waiting times).
$p(x \mid \theta) = \theta e^{-\theta x}$ [same as $Ga(1, \theta)$]
$\theta \sim Ga(\alpha, \beta)$
$\theta \mid x \sim Ga(\alpha + 1, \beta + x)$

Monte Carlo (MC) approximation to posterior distributions

Calculate exact posterior quantities can be difficult or impossible at time. Suppose we can take a sample of m values from the posterior distribution of θ, that is $\theta_1, \ldots, \theta_m \sim p(\theta \mid x)$ i.i.d. for large m. Then posterior quantities can be approximated by Monte Carlo estimates.
By the law of large numbers:

- Empirical distribution of the samples θ_i 's approximates $p(\theta \mid x)$ - use histogram or kernel density estimator (use density() in R)
- $\frac{1}{m}\sum_i \theta_i$ approximates $E[\theta \mid x]$ and $\frac{1}{m}\sum_i g(\theta_i)$ approximates $E[g(\theta) \mid x]$
- Cumulative ordered values approximate cdf $F(\theta \mid x)$
- Probability of $(g(\theta) > c)$ approximated by proportion of samples where event $(g(\theta_i) > c)$ occurs
- Sample moments/quantiles/functions approximate true moments/quantiles/functions

Approximation improves with increasing m. Method extends easily to higher dimensional parameters.

Example: Practice2 on birth rates revisited

Consider the number of kids of the $n_2 = 44$ women with college degrees,
$Y_{1,2}, \ldots, Y_{n_2,2} \mid \theta_2 \sim$ i.i.d.Poisson (θ_2), $\quad \sum_i Y_{i,2} = 66$
$\theta_2 \sim Ga(2,1), \quad \theta_2 \mid Y_2 \sim Ga(66+2, 44+1)$

```
a<-2; b<-1; n<-44; sy<-66
theta.mc<-rgamma(1000,a+sy,b+n)
mean(theta.mc)
(a+sy)/(b+n)
[1] 1.501015
mean(theta.mc<1.75)
[1] 1.51111
pgamma(1.75,a+sy,b+n)
[1] 0.899
[1] 0.8998
quantile(theta.mc,c(.025,.975)) qgamma(c(.025,.975),
a+s y, b+n)
[1] 1.1801941 .892473
[1] 1.1734371 .890836
```

Also, $\theta_1 \mid Y_1 \sim Ga(217+2, 111+1)$ for women without college degrees. Conclude that $P(z_1 > z_2 \mid Y_1, Y_2) = .48$

Example: Practice2 on birth rates revisited

We have $p(z \mid x) = \int p(z \mid \theta) p(\theta \mid x) d\theta$, that is a posterior expectation of $p(z \mid \theta)$. Therefore MC estimates can be obtained via

$$\text{sample } \theta^{(1)} \sim p(\theta \mid x), \text{ sample } z^{(1)} \sim p\left(z \mid \theta^{(1)}\right)$$
$$\text{sample } \theta^{(2)} \sim p(\theta \mid x), \text{ sample } z^{(2)} \sim p\left(z \mid \theta^{(2)}\right)$$

then $z^{(1)}, z^{(2)}, \ldots$ is a sample from the posterior predictive distribution

```
a<-2; b<-1; n1<- 111; sy1<- 217; n2<-44; sy2<-66
theta1.mc<-rgamma(10000,a+sy1,b+n1)
theta2.mc<-rgamma(10000,a+sy2,b+n2)
z1.mc<-rpois(10000,theta1.mc)
z2.mc<-rpois(10000,theta2.mc)
mean(z1.mc>z2.mc)
[1] 0.4823
```

Lecture 4: Prior Types - Outline

- Subjective and objective priors
- Non-informative priors
- Jeffreys' priors
- Diffuse priors

Choice of Prior Distribution

What is the interpretation of probability?

- Objective Probability

  As normative and objective representations of what is rational to believe about a parameter, usually in a state of ignorance (dice, coin, etc.)

- Frequentist probability

  Long run frequency when a process is repeated

- Subjective Probability

  Personal judgment about an event. But, one must be "coherent".

Two general types of Priors:

- Subjective Priors

  Priors chosen to reflect expert opinion or personal beliefs

- Objective Priors

  Priors chosen to let the data (i.e., likelihood) dominate the posterior distribution, and hence inference. These are generally determined based on the sampling model in use.

Subjective Priors

Ways of Specifying

- For single events:

  Relative probability of A to "not A" (as for lotteries/betting)

- Continuous quantities: prob. distribution

  divide range into intervals, and assign probabilities for each (as for events, above) or specify percentiles.

  Then, fit a smooth probability density.

- Examples: $(i)\theta \sim \text{Beta}(\alpha, \beta)$. From expert: mean $= .3$ and s.d. $= 0.1$. Then solve for α and β and find $\text{Beta}(9.2, 13.8)$. (ii) $\theta \sim N(\mu, s)$. From expert: an educated guess of 25 and a 95% confidence that θ is between 10 and 40. Then, $\theta \sim N(25, 30/4)$.

Subjective priors may be classified into two classes:

- Conjugate Priors
- Non-conjugate priors

Non-informative or "Objective" Priors

Goal: Choose priors which reflect a state of "no-information" about θ. Central problem: specifying a prior distribution for a parameter about which nothing is known.

For example, if θ can only have a finite set of values, it seems natural to assume all values equally likely a priori. Such specifications fall into the general class of noninformative priors. Roughly speaking a prior is non-informative if it has minimal impact on the posterior, i.e. it does not change much over the region where the likelihood is appreciable.

This notion was first used by Bayes(1763), and Laplace(1812) in Astronomy. A formal development is given in Box and Tiao (1973) and in Kass (1989, Statistical Science). Still the notion is not adequately well-defined.
Two types:

- Reference or "default" priors
- Diffuse priors

Reference Priors

Example: $x = (x_1, \ldots, x_n)$ from $N(\theta, \sigma^2)$, θ unknown, σ known.
What is the prior that reflects a state of no information about θ and lets the likelihood dominate the posterior?
The "flat" prior: $\pi(\theta) = 1$ for $-\infty < \theta < \infty$
Some features of the flat prior:

- Non-informative: All values of θ are equally likely. No information.
- Likelihood dominates posterior: $p(\theta \mid x) \sim L(\theta)\pi(\theta) = L(\theta)$
- May not be a probability distribution, i.e. $\int \pi(\theta) d\theta = \infty$ ("improper prior").
  Not a major concern if posterior is proper.
  Ex: $X \sim \text{Poisson}(\lambda)$, $0 < \lambda < \infty$, $\lambda \sim Ga(\alpha, \beta)$. Then $\alpha = 1, \beta \to 0$ gives $\pi(\lambda) \sim 1$ but with a proper posterior $\lambda \mid x \sim Ga(\alpha + t, n)$ (also, example above).
  In complex models, however, posterior may also be improper and one cannot make inference.
- Other "default" priors: $\pi(\lambda) \sim \frac{1}{\lambda^{1/2}} [= Ga(1/2, \beta \to 0)]$ and $\pi(\lambda) \sim \frac{1}{\lambda} [= Ga(0, \beta \to 0)]$ (both with proper posterior).

Change of variable

Another concern with reference priors: The "constant/flat rule" prior is not transformation invariant.

If $\tau = g(\theta)$ is a monotone function of θ with inverse $g^{-1}(\tau) = \theta$ then the density of τ is obtained from the density of θ as

$$p(\tau) = p\left(g^{-1}(\tau)\right)\left|\frac{\partial g^{-1}(\tau)}{\partial \tau}\right|, \text{ with Jacobian } J = \frac{\partial g^{-1}(\tau)}{\partial \tau}$$

Ex: $\pi(\theta) = 1$ and $\tau = e^{\theta}$. Implied $\pi(\tau) = 1/\tau$, no-longer "flat".

Binomial model with $\theta \sim \text{Beta}(1, 1)$ uniform for $0 \leq \theta \leq 1$

$\theta \sim \text{Beta}(0, 0)$ uniform for log odds, $\tau = \log[\theta/(1-\theta)]$

$\theta \sim \text{Beta}(1, -1)$ uniform for odds, $\tau = \theta/(1-\theta)$

A density cannot be simultaneously uniform in all transformations. Problem with uniform distribution as "non-informative" (no information on a transformation of θ does not imply no information on θ).

Remedy: Jeffreys' priors.

Jeffreys' Priors

Let $p(x \mid \theta)$ be a (single-parameter) probability model. The Fisher information is defined as

$$I(\theta) = -E_{x|\theta}\left[\frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2}\right]$$

and the Jeffreys' prior is $\pi(\theta) \sim \sqrt{I(\theta)}$. Properties:

- Locally uniform, i.e., flat in the region of the likelihood.
- Transformation invariant, ie. all parameterizations lead to same prior

  $$\pi(\theta) = \sqrt{I(\theta)} \text{ and } \pi(\tau) = \sqrt{I(\tau)} \text{ with } \tau = g(\theta)$$

  same prior obtained whether (i) apply Jeffreys rule to get $\pi(\theta)$ and then transform to get $\pi(\tau)$ or (ii) transform to τ and then apply Jeffreys rule to get $\pi(\tau)$.
- Caution: Posterior may not be proper.
- Does not work well for multi-parameter models, $\theta = (\theta_1, \ldots, \theta_k)$, $I(\theta)$ is a matrix of partial derivatives and $\pi(\theta) \sim \sqrt{|I(\theta)|}$.

Examples

Binomial: $x \sim \text{Binomial}(n, \theta)$,
$\log(p(x \mid \theta)) = x \log(\theta) + (n - x) \log(1 - \theta), \quad E[x \mid \theta] = n\theta,$

$$\frac{\partial^2 \log(p(x \mid \theta))}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}, \quad I(\theta) = \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Jeffreys' prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, i.e., $\theta \sim \text{Beta}(1/2, 1/2)$ (conjugate, proper)
Poisson: $x \sim \text{Poisson}(\lambda)$,
$\log(p(x \mid \lambda)) = \lambda + x \log(\lambda) - \log(x!)$ (single obs), $\quad E[x \mid \lambda] = \lambda$

$$\frac{\partial^2 \log(p(x \mid \lambda))}{\partial \lambda^2} = -\frac{x}{\lambda^2}, \quad I(\lambda) = \frac{1}{\lambda}$$

Jeffreys' prior is $\pi(\lambda) \propto \lambda^{-1/2}$, i.e., $\theta \sim Ga(1/2, \beta \to 0)$ (conjugate, improper with proper posterior)

Diffuse Priors

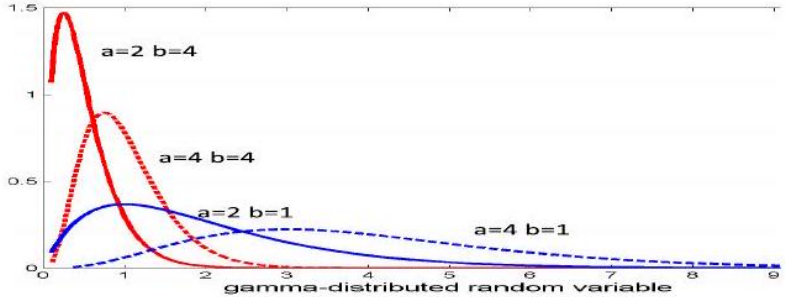
Motivation is to use a proper prior with "little" info about the parameter.

- Conjugate priors are proper
- The spread of a prior is a measure of the amount of uncertainty about the parameter expressed by the prior.

So, choose a conjugate prior with large standard deviation as a diffuse or weakly informative prior.

Diffuse Priors

Examples: $(i) X \sim \text{Poisson}(\lambda), \quad 0 < \lambda < \infty, \quad \lambda \sim Ga(\alpha, \beta)$. A diffuse prior is $Ga(\alpha, 1)$ with α the prior expectation.



(ii) $X \sim N(\theta, \sigma^2)$ with σ^2 known. A diffuse prior for θ is $N(\mu_0, 1000)$.