MSA101/MVE187 2017 Lecture 3

Petter Mostad

Chalmers University

September 5, 2017

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

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

- Prove this yourself!
- See Albert Section 3.3 for a computational example.

Example: The Normal-Gamma conjugacy

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

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

- Then $\pi(\tau \mid \alpha, \beta) = \text{Gamma}(\tau; \alpha, \beta)$ is a conjugate family, so that

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

- Specifically, we get the posterior below. (Mention noninformative)

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

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

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

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

Example: Predictive distribution for the Beta-Binomial conjugacy

- Assume $\pi(x \mid \theta) = \text{Binomial}(x; n, \theta)$ and $\pi(\theta) = \text{Beta}(\theta; \alpha, \beta)$.
- We get for the prior predictive

$$\begin{aligned}
\pi(x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)} \\
&= \frac{\text{Binomial}(x; n, \theta)\,\text{Beta}(\theta; \alpha, \beta)}{\text{Beta}(\theta; \alpha + x, \beta + n - x)} \\
&= \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} / \,\text{B}(\alpha, \beta)}{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} / \,\text{B}(\alpha + x, \beta + n - x)} \\
&= \binom{n}{x} \frac{\text{B}(\alpha + x, \beta + n - x)}{\text{B}(\alpha, \beta)}
\end{aligned}$$

- This is the Beta-Binomial distribution with parameters n, α, and β.

Example: Predictive distribution for the Normal-Normal conjugacy

- Assume $\pi(x \mid \theta) = \text{Normal}(x; \theta, 1/\tau_0)$ 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 $\pi(x)$ is normal.
  - $E(x) = E(E(x \mid \theta)) = E(\theta) = \mu$.
  - $Var(x) = Var(E(x \mid \theta)) + E(Var(x \mid \theta)) = Var(\theta) + E(1/\tau_0) = 1/\tau + 1/\tau_0$.
- So for the prior predictive we get

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

Example: Conjugacy for normal normal likelihood, no parameters known

- Assume $X \sim \text{Normal}(\mu, 1/\tau)$, with both μ and τ uncertain. The likelihood becomes

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

- Then the Normal-Gamma family is conjugate: The pair (μ, τ) has a Normal-Gamma distribution with parameters $\mu_0, \lambda > 0, \alpha > 0, \beta > 0$ if the density has the form

$$\pi(\mu, \tau \mid \mu_0, \lambda, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - 1/2} \exp\left(-\beta\tau - \frac{\lambda\tau}{2}(\mu - \mu_0)^2\right)$$

- Note: If (μ, τ) has the Normal-Gamma distribution above, we have $\tau \sim \text{Gamma}(\alpha, \beta)$ and $\mu \mid \tau \sim \text{Normal}(\mu_0, 1/(\lambda\tau))$.

Example: Multinomial-Dirichlet conjugacy

- Assume $x = (x_1, \ldots, x_n) \sim \text{Multinomial}(m, \theta_1, \theta_2, \ldots, \theta_n)$, with $\theta_1 + \cdots + \theta_n = 1$, so that x_i counts the number of results of type i in m independent trials, if results of type i have probability θ_i. The probability mass function is

$$\pi(x \mid \theta_1, \ldots, \theta_n) = \frac{m!}{x_1! \ldots x_k!} \theta_1^{x_1} \ldots \theta_n^{x_n}$$

- $(\theta_1, \ldots, \theta_n)$ has a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_n$ if the density can be written as

$$\pi(\theta_1, \ldots, \theta_n \mid \alpha_1, \ldots, \alpha_n) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_n)} \theta_1^{\alpha_1 - 1} \ldots \theta_n^{\alpha_n - 1}$$

- Prove that the Dirichlet family is a conjugate family to the Multinomial likelhiood!

Mixtures of conjugate distributions

- Assume we have a model $\pi(x \mid \theta)$ and a conjugate family of priors with densities $g(\theta; \gamma)$, where $\gamma \in Q$. For a fixed integer $k > 1$ define a new family of prior densities as consisting of all sums

$$\sum_{i=1}^{k} \alpha_i g(\theta; \gamma_i)$$

  where $\alpha_i > 0$, $\sum_{i=1}^{k} \alpha_i = 1$, and $\gamma_i \in Q$. Then, the new family is also a conjugate family.

- To assemble a proof: First, write $f_i(x)$ for the prior predictive density when using the prior $g(\theta; \gamma_i)$. We have shown above that it has an analytic form. Also, we know that, when using this prior, the posterior for θ has the form $g(\theta; \gamma_i')$ for some $\gamma_i' \in Q$. So we can write $\pi(x \mid \theta) g(\theta; \gamma_i) = f_i(x) g(\theta; \gamma_i')$.

Mixtures of conjugate distributions, cont.

We can compute the prior predictive as

$$\begin{aligned}\pi(x) &= \int \pi(x \mid \theta) \left[\sum_{i=1}^{k} \alpha_i g(\theta; \gamma_i)\right] d\theta \\ &= \sum_{i=1}^{k} \alpha_i \int \pi(x \mid \theta) g(\theta; \gamma_i)\, d\theta = \sum_{i=1}^{k} \alpha_i f_i(x)\end{aligned}$$

We get the posterior distribution

$$\pi(\theta \mid x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} = \sum_{i=1}^{k} \frac{\alpha_i}{\pi(x)} \pi(x \mid \theta) g(\theta; \gamma_i) = \sum_{i=1}^{k} \frac{\alpha_i f_i(x)}{\pi(x)} g(\theta; \gamma_i')$$

Thus the posterior has the same forrm as the prior: We have conjugacy.

Stepwise Bayesian updating

- Assume $x = (x_1, \ldots, x_m)$ is a random sample, so that

$$\pi(x \mid \theta) = \prod_{i=1}^{m} \pi(x_i \mid \theta)$$

- Using a prior $\pi(\theta)$ the posterior becomes

$$\pi(\theta \mid x) \propto_\theta \prod_{i=1}^{m} \pi(x_i \mid \theta)\pi(\theta)$$

- If we first update only with the observations $x_1, \ldots, x_k$, we get the posterior

$$\pi(\theta \mid x_1, \ldots, x_k) \propto_\theta \prod_{i=1}^{k} \pi(x_i \mid \theta)\pi(\theta)$$

- We see that if we use this as the prior and update with the remaining data $(x_{k+1}, \ldots, x_m)$, we get the same result as before.
- In Bayesian statistics, we may subdivide the data into data subsets and update the model stepwise with the data, as long as all the data sets are mutually independent given the model parameter θ.

Model choice / choosing priors

- Model choice is often divided into choosing a likelihood and choosing a prior distribution for the parameters θ.
- One possibility for setting a prior is to choose one that reflects "no knowledge". Unfortunately, not mathematically clear what this means; experts are disagreeing. "Non-informative priors". An example: The flat prior in the Normal distribution example above.
- In most examples, there is contextual information and it is reasonable to use an "informative prior". How to determine it?
  - *Elicit* parameters of a prior from "experts" or yourself, asking what seems to reflect existing knowledge. (E.g., `beta.select` in LearnBayes package)
  - Use a posterior given data from another previous source; previous knowledge is assumed based on those data.
- There are advanced methods for "model choice" in general (outside scope of course).

"Model choice" using weighed models and Bayes Factors

- NOTE: Instead of mixtures of conjugate distributions, one can use mixtures of *any* set of priors $g_i(\theta)$. The prior predictives $f_i(x)$ and the posteriors $g_i'(\theta)$ exist, even if they may be difficult to compute.
- We get that the posterior is a mixture of the corresponding posteriors, with weights updated using the prior predictive values $f_i(x)$ for the data.
- If we have only $k = 2$ priors, with weights α_1 and $\alpha_2 = 1 - \alpha_1$, and if we denote the posterior weights α_1' and $\alpha_2' = 1 - \alpha_1'$, we get

$$\frac{\alpha_1'}{1-\alpha_1'} = \frac{f_1(x)}{f_2(x)} \cdot \frac{\alpha_1}{1-\alpha_1}$$

  i.e., the posterior odds $\alpha_1'/(1-\alpha_1')$ is equal to the likelihood ratio $f_1(x)/f_2(x)$ times the prior odds $\alpha_1/(1-\alpha_1)$.
- $f_1(x)/f_2(x)$ is called the *Bayes factor*.

Robustness

- Another approach to the choice of prior: Check if switching between different choices matters for the final result.
- NOTE: For any posterior, there exists a prior that will give this posterior (assuming nonzero densities).
- Revised question: Do *reasonable* changes in the prior affect the result much?
- If not, the prior is called *robust* for this likelihood.
- Example: See Albert 3.4