STAT740: Informative Prior Distributions

Alyson Wilson

Department of Statistics
North Carolina State University

August 28, 2013

Conditional Independence

Suppose $Y_1, \ldots, Y_n$ are random variables and that θ is a parameter. We say that $Y_1, \ldots, Y_n$ are *conditionally independent* given θ if for every collection of n sets $\{A_1, \ldots, A_n\}$,

$$Pr(Y_1 \in A_1, \ldots, Y_n \in A_n \,|\, \theta) = Pr(Y_1 \in A_1 \,|\, \theta) \times \cdots \times Pr(Y_n \in A_n \,|\, \theta)$$

The joint density is given by

$$p(y_1, \ldots, y_n \,|\, \theta) = p_{Y_1}(y_1 \,|\, \theta) \times \cdots \times p_{Y_n}(y_n \,|\, \theta)$$

If the marginal densities are all the same, we say the $Y_1, \ldots, Y_n$ are *conditionally independent and identically distributed* given theta, written

$$Y_1, \ldots, Y_n \,|\, \theta \sim \text{i.i.d. } p(y \,|\, \theta)$$

Exchangeability

Let $p(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations π of $\{1, \ldots, n\}$, then $Y_1, \ldots, Y_n$ are *exchangeable*.

Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels convey no information about the outcomes.

Claim: If $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.

de Finetti's Theorem

Let $\{Y_1, Y_2, \ldots\}$ be a potentially infinite sequence of random variables all having a common sample space $\mathscr{Y}$. Suppose that for any n we believe that $Y_1, \ldots, Y_n$ is exchangeable. Then we can write

$$p(y_1, \ldots, y_n) = \int \left(\prod_{i=1}^{n} p(y_i \mid \theta) \right) p(\theta) d\theta$$

for some parameter θ, some prior distribution on θ, and some sampling distribution $p(y \mid \theta)$. The prior and sampling distribution depend on the form of $p(y_1, \ldots, y_n)$.

Prior Distributions

Science and Art

The prior distribution on our parameter summarizes our “pre-data” information. Specifying a prior distribution is both a science and an art.

When specifying a “classical” statistical model (a sampling distribution for the data), we make choices about which features of the data are important to capture (symmetry, unimodality), and which are not.

When specifying a Bayesian statistical model (a sampling distribution for the data and a prior distribution for the parameters), we similarly make choices about which features of the data and information about the parameter are important to capture, and which are not.

Prior Distributions

We will consider two different approaches to specifying prior distributions:

- Informative priors
- Noninformative priors

Informative Prior Distributions

What is an informative prior distribution?

We usually define it using its negative: any prior distribution trying to capture something besides "I have no idea."

The process of gathering information to use for a prior distribution is called *elicitation*.

Many Bayesians subscribe to *subjective probability*, where probability is interpreted to reflect personal belief in the chance of occurrence. (It is a valid question to ask whose personal belief we are modeling.)

Informative Prior Distributions

Approaches

- Histogram approach: Bin the parameter space and decide on the amount of probability (or the relative "likelihood") in each bin. Normalize if required. Smooth if desired.
- Relative likelihood approach: Identify the "most likely" and "least likely" points in the parameter space. Determine the relative "likelihood" for several more points. Interpolate (usually linearly). Normalize.

Informative Prior Distributions

Approaches

- Match a given functional form: Choose a favorite parametric family of distributions. For example,
  - for a parameter constrained to be between 0 and 1, use a beta distribution.
  - for a parameter constrained to be greater than 0, use a gamma distribution.
  - for a parameter defined on the real line, use a normal distribution.

  Specify enough additional information about the parameter to determine *hyperparameter* values for the parametric family. Use means, variances, quantiles, medians—whatever can be elicited.

  Note: You don't have to use a familiar probability distribution—you can use any function that integrates to 1.

Informative Prior Distributions

Conjugate Priors

For some sampling distributions, there are particularly convenient functional forms to use a prior distributions. These are called *conjugate priors*.

Informally, a conjugate prior is one where the prior distribution and the posterior distribution have the same form.

Conjugate Priors

We've already seen an example of a conjugate prior. In our t-shirt example, the prior distribution we used was Beta(1.08,14.96), our sampling distribution was Binomial(120,p), and our posterior distribution was Beta(15.08,120.96). The prior distribution and the posterior distribution have the same form.

More formally,

$$\begin{aligned} Y \mid p &\sim \text{Binomial}(n, p) \\ p &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

$$\begin{aligned} \pi(p \mid y) &\propto p^{y}(1-p)^{n-y} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha+y-1}(1-p)^{\beta+n-y-1} \end{aligned}$$

Therefore, $p \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$. We say that the beta distribution is the conjugate prior for the binomial sampling distribution.

Beta-Binomial

Notice that the posterior mean is

$$\begin{aligned}\frac{\alpha + y}{\alpha + \beta + n} &= (\frac{\alpha+\beta}{\alpha+\beta+n})(\frac{\alpha}{\alpha+\beta}) + (1 - \frac{\alpha+\beta}{\alpha+\beta+n})(\frac{y}{n}) \\ &= (\frac{\alpha+\beta}{\alpha+\beta+n})(\frac{\alpha}{\alpha+\beta}) + (\frac{n}{\alpha+\beta+n})(\frac{y}{n})\end{aligned}$$

a linear combination of the prior mean and the sample mean (maximum likelihood estimate).

With no observations, the posterior mean is the prior mean. As the number of observations becomes large (for fixed α and β), the posterior mean $\approx \frac{y}{n}$.

Conjugate Priors

Example: Poisson Distribution

Assume that we have data $y \in (0, 1, \ldots)$ that represent the number of events in a fixed time period. If the expected number of events in the time period is $\lambda > 0$, then we use a Poisson distribution as the sampling distribution.

$$f(y_i \mid \lambda) = \frac{\lambda^{y_i} \exp(\lambda)}{y_i!}$$

If we assume that the y_i are conditionally independent given λ, then

$$\begin{aligned} f(y|\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \\ &= \frac{\lambda^{n\bar{y}} \exp(-n\lambda)}{\prod_{i=1}^{n} y_i!} \end{aligned}$$

Conjugate Prior

Example: Poisson Distribution

Consider Gamma(α, β) as prior for λ:

$$\pi(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

The posterior distribution is also Gamma:

$$\begin{aligned} \pi(\lambda \,|\, y) &\propto \lambda^{n\bar{y}} \exp(-n\lambda)\lambda^{\alpha-1} \exp(-\beta\lambda) \\ &\propto \lambda^{n\bar{y}+\alpha-1} \exp(-n\lambda - \beta\lambda) \end{aligned}$$

For $\eta = n\bar{y} + \alpha$ and $\nu = n + \beta$, the posterior distribution is Gamma(η, ν).

Poisson-Gamma

Notice that the posterior mean is

$$\begin{aligned}\frac{n\bar{y}+\alpha}{n+\beta} &= (\frac{\beta}{n+\beta})(\frac{\alpha}{\beta})+(1-\frac{\beta}{n+\beta})(\frac{n\bar{y}}{n}) \\ &= (\frac{\beta}{n+\beta})(\frac{\alpha}{\beta})+(\frac{n}{n+\beta})(\bar{y})\end{aligned}$$

a linear combination of the prior mean and the sample mean (maximum likelihood estimate).

With no observations, the posterior mean is the prior mean. As the number of observations becomes large (for fixed α and β), the posterior mean $\approx \bar{y}$.

See Diaconis and Ylvisaker (1979) for a more detailed treatment of conjugate priors for exponential families.

Prior Distributions

Given Functional Form

Once you have chosen a given functional form, use the knowledge from Tversky and Kahneman (1974) to elicit, for example, a measure of center and quantiles. Keep in mind when eliciting quantiles that experts are often conservative.

Choose parameters that match these elicited quantities.

Check your results with the expert. You may do this using the prior distribution itself or using the prior predictive distribution. Draw a picture and try out some probability statements like "there's only a 5% chance that the lifetime exceeds."

Thoughts on informative prior distributions

- Sometimes experts find it difficult to talk about parameters and are more comfortable talking about predictions. In this case, you need to do your elicitation on the *prior predictive distribution*. For example, your expert may be more able to talk about how many t-shirts he expects to sell to 20 alumni rather than the proportion of alumni who will buy shirts. (Chaloner and Duncan 1983; Gavaskar 1988)
- Experts may find it easier to talk about a reparameterized version of the model. For example, if you model a part's lifetime using an Exponential(λ) distribution, the experts may not be able to talk about the failure rate (λ), but perhaps the mean time to failure ($1/\lambda$) or the reliability/probability of working at time t ($\exp(-\lambda t)$). (Martz and Waller 1982)
- Always make some predictions of something observable using your prior to do a sanity check.
- Remember, any place that has probability zero in the prior has probability zero in the posterior.