Bayesian data analysis – reading instructions 2

Aki Vehtari

Chapter 2 – outline

Outline of the chapter 2

- 2.1 Binomial model (e.g. biased coin flipping)
- 2.2 Posterior as compromise between data and prior information
- 2.3 Posterior summaries
- 2.4 Informative prior distributions (skip exponential families and sufficient statistics)
- 2.5 Gaussian model with known variance
- 2.6 Other single parameter models
    - in this course the normal distribution with known mean but unknwon variance is the most important
    - glance through Poisson and exponential
- 2.7 glance through this example, which illustrates benefits of prior information, no need to read all the details (it's quite long example)
- 2.8 Noninformative priors
- 2.9 Weakly informative priors

Laplace's approach for approximating integrals is discussed in more detail in Chapter 4.

R and Python demos at https://avehtari.github.io/BDA_course_Aalto/demos.html

- demo2_1: Binomial model and Beta posterior.
- demo2_2: Comparison of posterior distributions with different parameter values for the Beta prior distribution.
- demo2_3: Use samples to plot histogram with quantiles, and the same for a transformed variable.
- demo2_4: Grid sampling using inverse-cdf method.

Chapter 2 – most important terms

Find all the terms and symbols listed below. When reading the chapter, write down questions related to things unclear for you or things you think might be unclear for others. See also the additional comments below.

- binomial model
- Bernoulli trial
- exchangeability
- Bin, $\binom{n}{y}$
- Laplace's law of succession
- think which expectations in eqs. 2.7-2.8
- summarizing posterior inference
- mode, mean, median, standard deviation, variance, quantile

- central posterior interval
- highest posterior density interval / region
- uninformative / informative prior distribution
- principle of insufficient reason
- hyperparameter
- conjugacy, conjugate family, conjugate prior distribution, natural conjugate prior
- nonconjugate prior
- normal distribution
- conjugate prior for mean of normal distribution with known variance
- posterior for mean of normal distribution with known variance
- precision
- posterior predictive distribution
- normal model with known mean but unknown variance
- proper and improper prior
- unnormalized density
- Jeffreys' invariance principle
- note non-uniqueness of noniformative priors for the binomial parameter
- difficulties with noninformative priors
- weakly informative priors

Integration over Beta distribution

Chapter 2 has an example of analysing the ratio of girls born in Paris 1745–1770. Laplace used binomial model and uniform prior which produces Beta distribution as posterior distribution. Laplace wanted to calculate $p(\theta \geq 0.5)$, which is obtained as

$$\begin{aligned} p(\theta \geq 0.5) &= \int_{0.5}^{1} p(\theta|y,n,M)d\theta \\ &= \frac{493473!}{241945!251527!} \int_{0.5}^{1} \theta^{y}(1-\theta)^{n-y} d\theta \end{aligned}$$

Note that $\Gamma(n) = (n-1)!$. Integral has a form which is called *incomplete Beta function*. Bayes and Laplace had difficulties in computing this, but nowadays there are several series and continued fraction expressions. Furthermore usually the normalisation term is computed by computing $\log(\Gamma(\cdot))$ directly without explicitly computing $\Gamma(\cdot)$. Bayes was able to solve integral given small n and y. In case of large n and y, Laplace used Gaussian approximation of the posterior (more in chapter 4). In this specific case, R pbeta gives the same results as Laplace's result with at least 3 digit accuracy.

Numerical accuracy

Laplace calculated

$$p(\theta \geq 0.5|y,n,M) \approx 1.15 \times 10^{-42}.$$

Correspondingly Laplace could have calculated

$$p(\theta \geq 0.5|y,n,M) = 1 - p(\theta \leq 0.5|y,n,M),$$

which in theory could be computed in R with `1-pbeta(0.5,y+1,n-y+1)`. In practice this fails, due to the limitation in the floating point representation used by the computers. In R the largest floating point number which is smaller than 1 is about 1-eps/4, where eps is about 2.22×10^{-16} (the smallest floating point number larger than 1 is 1+eps). Thus the result from `pbeta(0.5,y+1,n-y+1)` will be rounded to 1 and $1 - 1 = 0 \neq 1.15 \times 10^{-42}$. We can compute $p(\theta \geq 0.5|y, n, M)$ in R with `pbeta(0.5, y+1, n-y+1, lower.tail=FALSE)`.

Highest Posterior Density interval

HPD interval is not invariant to reparametrization. Here's an illustrative example (using R and package `HDInterval`):

```
> r <- exp(rnorm(1000))
> quantile(log(r),c(.05, .95))
       5%       95%
-1.532931  1.655137
> log(quantile(r,c(.05, .95)))
       5%       95%
-1.532925  1.655139
> hdi(log(r), credMass = 0.9)
    lower     upper
-1.449125  1.739169
attr(,"credMass")
[1] 0.9
> log(hdi(r, credMass = 0.9))
    lower     upper
-2.607574  1.318569
attr(,"credMass")
[1] 0.9
```

Gaussian distribution in more complex models and methods

Gaussian distribution is commonly used in mixture models, hierarchical models, hierarchical prior structures, scale mixture distributions, Gaussian latent variable models, Gaussian processes, Gaussian random Markov fields, Kalman filters, proposal distribution in Monte Carlo methods, etc.

Predictive distribution

Often the predictive distribution is more interesting than the posterior distribution. The posterior distribution describes the uncertainty in the parameters (like the proportion of red chips in the bag), but the predictive distribution describes also the uncertainty about the future event (like which color is picked next). This difference is important, for example, if we want to what could happen if some treatment is given to a patient.

In case of Gaussian distribution with known variance σ^2 the model is

$$y \sim \mathrm{N}(\theta, \sigma^2),$$

where σ^2 describes aleatoric uncertainty. Using uniform prior the posterior is

$$p(\theta|y) \sim \mathrm{N}(\theta|\bar{y}, \sigma^2/n),$$

where σ^2/n described epistemic uncertainty related to θ. Using uniform prior the posterior predictive distribution for new $\tilde{y}$ is

$$p(\tilde{y}|y) \sim \mathrm{N}(\tilde{y}|\bar{y}, \sigma^2 + \sigma^2/n),$$

where the uncertainty is sum of epistemic (σ^2/n) and aleatoric uncertainty (σ^2).

Non-informative and weakly informative priors

Our thinking has advanced since sections 2.8 and 2.9 were written. We're even more strongly in favor weakly informative priors, and in favor of more information in the priors. Non-informative priors are likely to produce more unstable estimates (higher variance), and the lectures include also examples of how seemingly non-informative priors can actually be informative on some aspect. See further discussion and example in the Prior Choice Wiki (https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. Thus Prior Choice Wiki will see also some further updates (we're doing research and learning more all the time).

Should we worry about rigged priors?

Andrew Gelman's blog post answering worries that data analyst would choose a too optimistic prior http://andrewgelman.com/2017/10/04/worry-rigged-priors/.

Exchangeability

You don't need to understand or use the term exchangeability before Chapter 5 and Lecture 7. At this point and until Chapter 5 and Lecture 7, it is sufficient that you know that 1) independence is stronger condition than exchangeability, 2) independence implies exchangeability, 3) exchangeability does not imply independence, 4) exchangeability is related to what information is available instead of the properties of unknown underlying data generating mechanism. If you want to know more about exchangeability right now, then read BDA Section 5.2 and BDA_notes_ch5.