STATS 700 Bayesian Inference and Computation **Fall 2017**

Lecture 6: Oct 9,11 Hierarchical Model Example, Importance Sampling

Lecturer: CHEN Yang *Scribe: Ran Bi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Hierarchical Models: Examples and Applications

Bayes: prior + likelihood → posterior
↓
Simple Parameter / Multiple Parameter: inference, analytical posterior
↓
Hierarchical Models

6.1.1 Eight Schools Example

A study analyzes the effects of special coaching programs on test scores. Separate randomized experiments were performed to estimate the effects of coaching programs for the SAT in each of eight high schools.[1]
casual inference problem: $Y_i(1) - -treatment; Y_i(0) - -control$

$$causal\ effect = \frac{1}{N}\sum(Y_i(1) - Y_i(0)) \approx \frac{\sum(Y_i(1))}{M_1} - \frac{\sum(Y_j(0))}{M_2}$$

• separate estimates:it is difficult statistically to distinguish between any of the experiments. Treating each experiment separately and applying the simple normal analysis in each yields 95% posterior intervals that all overlap substantially[1].
• pooled estimates: by treating the data as eight normally distributed observations with known variances, estimate the common effect to be 7.7 points with standard error equal to 4.1[1].

6.1.2 Gaussian Model(continue with the schools example)

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$$

θ_j: true treatment effect. When $\theta_1 = \theta_2 = ... = \theta$, it becomes single parameter model.

$$y_j|\theta_j, \sigma_j \sim N(\theta_j, \sigma_j^2)$$

y_i: point estimate of the effect θ_j, σ_j:the corresponding estimated standard error, which is assumed known[1].

•hierarchical model:

$$p(\theta_1, ..., \theta_z) = \prod N(\theta_j; \mu, \tau^2)$$

$$\overline{y_j} \sim N(\mu, \frac{\sigma^2}{n_j} + \tau^2)$$

$$p(\tau|y) = \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)}$$

$$p(\theta_j|y_1...y_z) \rightarrow random\ samples: \theta_j^{(1)}\theta_j^{(2)}...\theta_j^{(M)} \rightarrow random\ samples\ from\ joint: P(\mu,\tau^2,\theta 1...\theta z|y_1...y_z)$$

if we have $p(\mu,\tau^2|observations)$:

$$p(\theta_i|\mu,\tau,\overline{y_j},\sigma^2) \sim N(\frac{\frac{\mu}{\tau^2}+\frac{\overline{y_j}}{\sigma_i^2}}{\frac{1}{\tau^2}+\frac{1}{\sigma_i^2}}, \frac{1}{\frac{1}{\tau^2}+\frac{1}{\sigma_i^2}})$$

6.1.3 Weakly informative priors for hierarchical variance parameters

- Uniform prior distribution
- Inverse Gamma prior distribution:
  an attempt at noninformativeness within the conditionally conjugate family[1]
  improper densities inverse-gamma(ϵ,ϵ) on τ^2 as $\epsilon \rightarrow 0$

6.2 Importance Sampling

Importance sampling is a method that is used for computing expectations using a random sample drawn from an approximation to the target distribution[1].

6.2.1 Quantity of interest

$$\mu = E_\pi(h(x)) = \int h(x)\pi(x)dx = \int h(x)p(x|y)dx$$

6.2.2 Importance sampling algorithm

We cannot generate random draws of x from $\pi(x)$, but g(.) is a probability density from which we can generate random draws.

- draw $x^{(1)},...,x^{(m)}$ from g(.)
- calculate the importance weight: $w^{(j)} = \frac{\pi(x^{(j)})}{g(x^{(j)})}$, $j = 1,..,m$

- approximate $\hat{\mu}$:

$$\hat{\mu} = \frac{w^{(1)}h(x^{(1)}) + ... + w^{(1)}h(x^{(1)})}{w^{(1)} + ... + w^{(m)}}$$

6.2.3 Effective sample size (ESS)

ESS measures how different the trial distrition is from the target distribution.

$$ESS = \frac{m}{1 + var_g[w(x)]}$$

prove $\frac{var_\pi h(y)}{var_g(h(x)w(x))} \approx \frac{1}{1+var_g(w(x))}$:

$$\hat{\mu} = \frac{\frac{1}{m}\sum h(x^{(j)})w(x^{(j)})}{\frac{1}{m}\sum w(x^{(j)})} = \frac{\tilde{\mu}}{\overline{W}}$$

$Z = h(\mathbf{x})w(\mathbf{x}),\ W = w(\mathbf{x})$

$$E_g(\hat{\mu}) \approx E_g\overline{Z}[1 - (\overline{W} - 1) + (\overline{W} - 1)^2 + ... \approx \mu - \frac{cov_g(W, Z)}{m} + \frac{\mu var_g W}{m}$$

$$var_g(\hat{\mu}) \approx \frac{1}{m}[\mu^2 var_g(W) + var_g(Z) - 2\mu cov_g(W, Z)]$$

$$MSE(\tilde{\mu}) = E_g(\tilde{\mu} - \mu)^2 = \frac{var_g(Z)}{m}$$

$$MSE(\hat{\mu}) = \frac{1}{m}MSE(\tilde{\mu}) + \frac{1}{m}[\mu^2 var_g(W) - 2\mu cov_g(W, Z)] + O(m^{-2})$$

$H = h(\mathbf{x}),\ then\ Z = WH,\ \mu = E_g(WH)$

$$cov_g(W, Z) = cov_\pi(W, H) + \mu E_\pi(W) - \mu$$

$$var_g(Z) = E_\pi(WH^2) - \mu^2 \approx E_\pi(W)E_\pi^2(H) + var_\pi(H)E_\pi(W) + 2\mu cov_\pi(W, H) - \mu^2$$

$$var_g(\tilde{\mu}) \approx var_\pi(H)\frac{1 + var_p(W)}{m}$$

if $\hat{\mu}$ were estimated by $\sum h(\mathbf{y}^{(j)})/m$, then the efficiency of $\hat{\mu}$ is:

$$\frac{var_\pi h(y)}{var_g(h(x)w(x))} \approx \frac{1}{1 + var_g(w(x))}$$

6.2.4 Marginalization

$$var_g[\frac{f(z_1, z_2)}{g(z_1, z_2)}] \geq var_g[\frac{f_1(z_1)}{g_1(z_1)}] = var_g[\frac{\int f(z_1, z_2)dz_2}{\int g(z_1, z_2)dz_2}]$$

$f(z_1, z_2), g(z_1, z_2)$: probability density
$f_1(z_1), g_1(z_1)$: marginal density

References

[1] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis.* CRC press Boca Raton, FL, 2014.