STA 601: Fall 2019, Midterm

October 24, 2019

Community Standard

To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

I have adhered to the Duke Community Standard in completing this exam.

Name: Yifei Wang

NetID: yw323

Signature: Yifei Wang

Please write your name at the top of every page!

Common distributions

Normal with mean θ and variance σ^2: $p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(y-\theta)^2)$ for $y \in \mathbb{R}$

Multivariate (p-dimensional) normal with mean vector θ and covariance matrix Σ: $p(y|\theta,\Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp(-\frac{1}{2}(y-\theta)^t\Sigma^{-1}(y-\theta))$ for $y \in \mathbb{R}^p$

Exponential with mean λ and variance λ^2: $p(y|\lambda) = \frac{1}{\lambda}\exp(-\frac{y}{\lambda})$ for $y > 0$

Gamma with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$: $p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-\beta y)$ for $y > 0$

Inverse Gamma with mean $\frac{\beta}{\alpha-1}$ and variance $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$: $p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{-(\alpha+1)}\exp(-\frac{\beta}{y})$ for $y > 0$

Uniform distribution on $(0,1)$: $p(y) = 1$ for $y \in [0,1]$

Poisson distribution with mean θ: $p(y|\theta) = \frac{\exp(-\theta)\theta^y}{y!}$ for y a non-negative integer

Σ is an inverse Wishart distribution with parameters (ν_0, S_0^{-1}):

$$p(\Sigma) \propto |\Sigma|^{-(\nu_0+p+1)/2}\exp(-\mathrm{tr}(S_0\Sigma^{-1})/2) \text{ and } E[\Sigma] = \frac{1}{\nu_0-p-1}S_0$$

Beta distribution with mean $\frac{a}{a+b}$: $p(y|a,b) = \frac{1}{B(a,b)}y^{a-1}(1-y)^{b-1}$ for $y \in [0,1]$ and $a > 0, b > 0$

33/48

Some figures in this exam are from Bayesian Data Analysis, 2nd Edition.

7

1. (2 points each) State/Define the following:

   (a) State Bayes' theorem.

   $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$, what's A and B?

   -1

   (b) Describe Geweke's z statistic.

   It describe the convergence of the MCMC ~~chains~~.

   -1

   (c) TRUE or FALSE: Let $X_1 \ldots, X_n|\theta \sim \text{Poisson}(\theta)$ and consider the proper conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$. The posterior mean is unbiased for θ. FALSE.

   $\alpha + \Sigma y,\ \beta + n.$

   $\frac{\alpha + \Sigma y}{\beta + n}$

   (d) TRUE or FALSE: The proper way to deal with Missing Not at Random data is by performing a complete case analysis. TRUE

   (e) TRUE or FALSE: The conditional distributions used in the MICE procedure necessarily define a proper joint distribution. FALSE

$\begin{pmatrix} 1-\rho^2 & \rho^2 \\ \rho & 1-\rho^2 \end{pmatrix}$ $\frac{1}{1-\rho^2}$ $\frac{-\rho}{1-\rho^2}$ $\frac{1}{1-\rho^2}$ (5) $a+b\rho=1$ $a-a\rho^2=1$

$a\rho+b=0$ $a(1-\rho^2)=1$

2. (10 points) Different samplers for normal distributions. Consider a single observation (y_1, y_2) from a bivariate normal distribution with unknown mean (θ_1, θ_2) and known and fixed covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Consider a uniform prior on $\theta = (\theta_1, \theta_2)$: $p(\theta_1, \theta_2) \propto 1$.

(a) Derive the joint posterior for (θ_1, θ_2) conditional on the data and the known covariance matrix. Describe a direct sampler from this distribution. 2

(b) Derive the Gibbs sampler for the above posterior (that is, what are the distributions of $\theta_1|\theta_2, y_1, y_2, \Sigma$ and $\theta_2|\theta_1, y_1, y_2, \Sigma$?) 0

(c) If you have 1000 samples from the direct sampler in part (a) and 1000 samples from the Gibbs sampler in part (b), which one has a bigger effective sample size? Which one likely provides a better approximation to the posterior with this many samples? 3

$\Sigma_1^{-1} = \begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}$

(a) $Y \sim N\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \Sigma_1\right)$

$p\left(y=\begin{pmatrix}y_1\\ y_2\end{pmatrix}\right) = \frac{1}{2\pi} \cdot |\Sigma_1|^{-1} \exp\left(-\frac{1}{2}(y-\theta)^T \Sigma_1^{-1}(y-\theta)\right)$

$P(\theta|y, \Sigma_1) \propto P(y|\theta, \Sigma_1)\, p(\theta)$

$\propto \exp\left(-\frac{1}{2}(y-\theta)^T \Sigma_1^{-1}(y-\theta)\right)$

$\cancel{\propto \exp\left(-\frac{1}{2}\left[(y_1-\theta_1)^2+(y_2-\theta_2)^2+2\rho(y_1-\theta_1)(y_2-\theta_2)\right]\right)}$

$\propto \exp\left(-\frac{1}{2(1-\rho^2)}\left[(y_1-\theta_1)^2+(y_2-\theta_2)^2-2\rho(y_1-\theta_1)(y_2-\theta_2)\right]\right)$

$\sim N\left(\begin{pmatrix}y_1\\ y_2\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$

This is a multivariate normal distribution, and we could directly sample from it.

how?

(b). $P(\theta_1|\theta_2, y_1, y_2, \Sigma_1)$

$\propto P(y_1|\theta_1, \theta_2, y_2, \Sigma_1)\, P(\theta_1|\theta_2, y_2, \Sigma_1)$

~~We know that~~ $\cancel{P(\theta_1|\theta_2, y_2, \Sigma_1) \propto 1}$

$\propto P(y_1|\theta_1, \theta_2, y_2, \Sigma_1)\, P(y_2|\theta_1, \theta_2, \Sigma_1)\, P(\theta_1|\theta_2, \Sigma_1)$

We know that $P(\theta_1|\theta_2, \Sigma_1) \propto 1$

$P(y_2|\theta_1, \theta_2, \Sigma_1) = \int_{y_1} P(y_1, y_2|\theta_1, \theta_2, \Sigma_1)\, dy_1$

$\sim N(\theta_2, 1)$

$P(y_1|\theta_1, \theta_2, y_2, \Sigma_1)$

$\sim N(\theta_1$

(c). (a) has a bigger effective sample size.

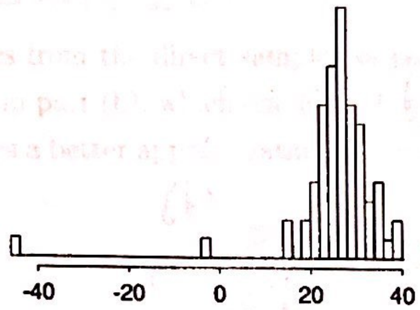
(a) is more likely providing a better approximation to the posterior.

because these samples are independent.

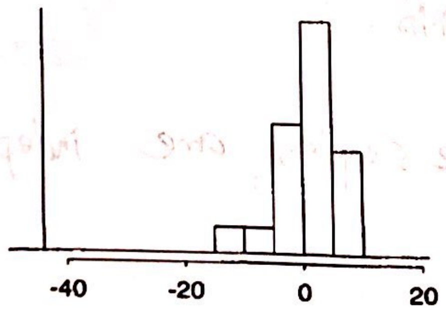
9/10

3. (10 points) Posterior predictive checks: Simon Newcomb made 66 measurements of the speed of light in 1882 by measuring the amount of time required for light to travel 7442 meters. Those are presented in the first plot below. The data are recorded as deviations from 24800 nanoseconds (and so can be positive and negative). We model these as normally distributed with mean μ and variance σ^2 with a prior $p(\mu, \sigma^2) \propto 1$.

(a) Before considering any analysis, give one argument for and one argument against using the normal distribution for modeling these data.



(b) The smallest observation in Newcomb's dataset is -44. Below we plot the smallest observation from the posterior predictive distribution across 20 hypothetical replications. The vertical line is Newcomb's original value.



What does this picture tell you about the models? Does the normal model capture everything in the data well?

(c) Write out the scheme for getting the data to draw the posterior predictive distribution above. For each step in the scheme, state which distributions you are sampling from.

(a). ① ~~This~~ The data seems to have a bell shape and symmetric about some ~~axis~~ line, so it's reasonable to use normal

② Some data are really far from the high density region ~~of~~, which is very rare in the normal distribution, so normal might not be great here.

(b). This ~~I~~ shows that the normal model could not capture well on everything. The data have the feature that the smallest observation is −44. However, none of the synthetic data have, or even close to, this feature. The normal model failed in capturing this data feature. But it might be great at capturing the mean of the data.

(c). for s in 1...20:

① sample one ~~[illegible]~~ $\sigma^{2(s)}$ from $P(\sigma^2 | Y)$ What are these dist.? (−1)

② sample one $\theta^{(s)}$ from $P(\theta | Y, \sigma^{2(s)})$

③ sample 66. $\tilde{Y}_i^{(s)}$ from $P(\tilde{Y} | \theta^{(s)}, \sigma^{2(s)})$, which is $N(\theta^{(s)}, \sigma^{2(s)})$ for i in 1...66.

④ Compute and record the smallest value in $Y_i^{(s)}$ for i in 1...66. as $t_{min}^{(s)}$

plot a histogram among all $t_{min}^{(s)}$ for s in 1...20

5

4. (10 points) Let $\theta_A = \mu + \delta$ and $\theta_B = \mu - \delta$. Let the prior on μ be $N(\mu_0, \sigma_0^2)$ and δ be $N(\delta_0, \tau_0^2)$. What is the induced joint prior on θ_A and θ_B?

$$\mu \sim N(\mu_0, \sigma_0^2)$$

$$\delta \sim N(\delta_0, \tau_0^2)$$

Thus we have,

$$\theta_A \sim N(\mu_0 + \delta_0, \sigma_0^2 + \tau_0^2)$$

should be $\sigma_0^2 + \tau_0^2$ also

$$\theta_B \sim N(\mu_0 - \delta_0, \sigma_0^2 - \tau_0^2)$$

Joint prior is.

$$P(\theta_A, \theta_B \mid \mu_0, \sigma_0^2, \delta_0, \tau_0^2)$$

$$= P(\theta_A \mid \theta_B, \mu_0, \sigma_0^2, \delta_0, \tau_0^2)\, P(\theta_B \mid \mu_0, \sigma_0^2, \delta_0, \tau_0^2)$$

$$P(\theta_B \mid \mu_0, \delta_0, \sigma_0^2, \tau_0^2)$$

$$= \frac{1}{\sqrt{2\pi(\sigma_0^2 - \tau_0^2)}} \exp\left(-\frac{1}{2(\sigma_0^2 - \tau_0^2)} (\theta_B - \mu_0 + \delta_0)^2\right)$$

We also have. $\theta_A = \theta_B + 2\delta$, this shows.

$$P(\theta_A \mid \theta_B, \mu_0, \delta_0, \sigma_0^2, \tau_0^2)$$

$$= P(\theta_A \mid \theta_B, \delta_0, \tau_0^2)$$

Thus.

$$\theta_A \mid \theta_B \sim N(\theta_B + 2\delta_0, \tau_0^2)$$

Combining all these we have.

$$P(\theta_A, \theta_B \mid \mu_0, \sigma_0^2, \delta_0, \tau_0^2)$$

$$= \frac{1}{\sqrt{2\pi(\sigma_0^2 - \tau_0^2)}} \exp\left(-\frac{(\theta_B - \mu_0 + \delta_0)^2}{2(\sigma_0^2 - \tau_0^2)}\right) \times$$

$$\frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{(\theta_A - \theta_B - 2\delta_0)^2}{2\tau_0^2}\right)$$

$$= N(\mu_0 - \delta_0, \sigma_0^2 - \tau_0^2) \times N(\theta_B + 2\delta_0, \tau_0^2)$$

need the joint prior, not factored version

$$\theta_A, \theta_B \sim N_2\left(\begin{bmatrix} \mu_0 + \delta_0 \\ \mu_0 - \delta_0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 + \tau_0^2 & \sigma_0^2 - \tau_0^2 \\ \sigma_0^2 - \tau_0^2 & \sigma_0^2 + \tau_0^2 \end{bmatrix}\right)$$

~~α~~ Gamma $\frac{\alpha}{\beta}$ $\frac{\alpha}{\beta^2}$ $\frac{\exp(-\theta)\theta^y}{y!}$ $\alpha = 7\beta$ $\alpha = \frac{8}{3}\beta^2$ $\frac{8}{3}\beta = 7$ $\beta = \frac{21}{8}$ $\alpha = \frac{147}{8}$

7

5. (10 points) Setting prior parameters. You are studying email behavior of individuals in a university. For a random sample of 10 students you observe the count of emails they send in one day ($S_1, \ldots, S_{10}$) and for a random sample of 20 professors you observe the count of emails they send in 10 days: ($P_1, \ldots, P_{20}$).

(a) What is a reasonable sampling model for the counts of student and faculty emails?

(b) Imagine that the per-day email rate (θ) for students and faculty is the same, use your choice of sampling model in (a) to write down $S_1, \ldots, S_{10}|\theta \sim$? and $P_1, \ldots, P_{20}|\theta \sim$?

$\frac{1280 \times 8 + 147}{8}$

(c) What is the conjugate prior for θ in the sampling model above?

$\frac{240 + 21}{8}$

(d) You read the literature and find three studies where the per-day email rate on university campuses are 5, 7 and 9. Compute the mean and the variance of those three numbers—use those to help specify the parameters in the conjugate prior above. 80 1200

(e) You observe $\frac{1}{10}\sum_{i=1}^{10} s_i = 8$ and $\frac{1}{20}\sum_{i=1}^{20} p_i = 60$, report the posterior mean for θ given the observed data.

(a). Different Poison models with differet parameter (per-day email rate) [we could have]

(b). $p(S_1 \ldots S_{10}|\theta) = \prod_{i=1}^{10} \frac{\exp(-\theta)\theta^{S_i}}{S_i!} \propto \theta^{\sum_{i=1}^{10} S_i} \exp(-10\theta)$ ✓

$p(P_1 \ldots P_{20}|\theta) = \prod_{i=1}^{20} \frac{\exp(-\theta)\theta^{P_i}}{P_i!} \propto \theta^{\sum_{i=1}^{20} P_i} \exp(-20\theta)$ ✗

(c). Gamma distribution with parameters α and β. ✓

(d). $\text{mean}(5,7,9) = 7$ ✓ $\text{var}(5,7,9) = \frac{8}{3}$ ✓

If in the literature we found that these three studies have similar sample sizes., and we want to have a relatively weak prior, we can set prior mean $\frac{\alpha}{\beta} = 7$ and prior variance $\frac{\alpha}{\beta^2} = \frac{8}{3}$. This gives us $\begin{cases} \alpha = \frac{147}{8} \\ \beta = \frac{21}{8} \end{cases}$ ✓

(e). $P(\theta|S,P) \propto P(S|\theta) P(P|\theta) P(\theta)$ ✓ 9

$\propto \theta^{1280} \exp(-30\theta) \cdot \theta^{\frac{147}{8}-1} \cdot \exp(-\frac{21}{8}\theta) \sim \text{Gamma}(1280 + \frac{147}{8}, 30 + \frac{21}{8})$ ✓ ✗

$E[\theta|S,P] = (1280 + \frac{147}{8})/(30 + \frac{21}{8})$