MATH 341 / 650.3 Spring 2020 Homework #8

Tuesday 19th May, 2020

Problem 1

These are questions about McGrayne's book, chapter 17 and the Epilogue.

(a) [easy] What do the computer scientists who adopted Bayesian methods care most about and whose view do they subscribe to? (p233)

Computer scientist cared only about results and now they subscribe to John Tukey's view.

(b) [easy] How was "Stanley" able to cross the Nevada desert?

Stanley was moving in an average speed, its camera took images of the route and the computer estimated the probability of various obstacles. As the robot navigated sharp turns and cliffs and generally stayed on course, and it slowed down to avoid even unlikely catastrophes.

(c) [easy] What two factors are leading to the "crumbling of the Tower of Babel?"

- The ability to accumulate evidence is an optimal survival strategy.
- Dependence on Bayes' theorem for calculating each individual has mastered a topic and is ready for a new challenge.

(d) [harder] Does the brain work through iterative Bayesian modeling?

Yes.

(e) [easy] According to Geman, what is the most powerful argument for Bayesian Statistics?

"...there is no more powerful argument for Bayes than its recognition of the brain's inner structures and prior expectations."

Distribution of r.v.	Quantile Function	PMF / PDF function	CDF function	Sampling Function
beta	`qbeta`$(p,\ \alpha,\ \beta)$	`d-`$(x,\ \alpha,\ \beta)$	`p-`$(x,\ \alpha,\ \beta)$	`r-`$(\alpha,\ \beta)$
betabinomial	`qbetabinom`$(p,\ n,\ \alpha,\ \beta)$	`d-`$(x,\ n,\ \alpha,\ \beta)$	`p-`$(x,\ n,\ \alpha,\ \beta)$	`r-`$(n,\ \alpha,\ \beta)$
binomial	`qbinom`$(p,\ n,\ \theta)$	`d-`$(x,\ n,\ \theta)$	`p-`$(x,\ n,\ \theta)$	`r-`$(n,\ \theta)$
exponential	`qexp`$(p,\ \theta)$	`d-`$(x,\ \theta)$	`p-`$(x,\ \theta)$	`r-`(θ)
gamma	`qgamma`$(p,\ \alpha,\ \beta)$	`d-`$(x,\ \alpha,\ \beta)$	`p-`$(x,\ \alpha,\ \beta)$	`r-`$(\alpha,\ \beta)$
inversegamma	`qinvgamma`$(p,\ \alpha,\ \beta)$	`d-`$(x,\ \alpha,\ \beta)$	`p-`$(x,\ \alpha,\ \beta)$	`r-`$(\alpha,\ \beta)$
negative-binomial	`qnbinom`$(p,\ r,\ \theta)$	`d-`$(x,\ r,\ \theta)$	`p-`$(x,\ r,\ \theta)$	`r-`$(r,\ \theta)$
normal (univariate)	`qnorm`$(p,\ \theta,\ \sigma)$	`d-`$(x,\ \theta,\ \sigma)$	`p-`$(x,\ \theta,\ \sigma)$	`r-`$(\theta,\ \sigma)$
poisson	`qpois`$(p,\ \theta)$	`d-`$(x,\ \theta)$	`p-`$(x,\ \theta)$	`r-`(θ)
T (standard)	`qt`$(p,\ \nu)$	`d-`$(x,\ \nu)$	`p-`$(x,\ \nu)$	`r-`(ν)
T (nonstandard)	`qt.scaled`$(p,\ \nu,\ \mu,\ \sigma)$	`d-`$(x,\ \nu,\ \mu,\ \sigma)$	`p-`$(x,\ \nu,\ \mu,\ \sigma)$	`r-`$(\nu,\ \mu,\ \sigma)$
uniform	`qunif`$(p,\ a,\ b)$	`d-`$(x,\ a,\ b)$	`p-`$(x,\ a,\ b)$	`r-`$(a,\ b)$

Table 1: Functions from `R` (in alphabetical order) that can be used on this assignment and exams. The hyphen in colums 3, 4 and 5 is shorthand notation for the full text of the r.v. which can be found in column 2.

Problem 2

This problem is about the normal-normal model using a "semi-conjugate" prior. Assume $X_1, \ldots, X_n \mid \theta, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\theta,\ \sigma^2)$ throughout.

(a) [easy] If θ and σ^2 are assumed to be independent, how can $\mathbb{P}(\theta,\ \sigma^2)$ be factored?

$$\mathbb{P}\left(\theta,\ \sigma^2\right) = \mathbb{P}(\theta)\,\mathbb{P}\left(\ \sigma^2\right)$$

(b) [easy] If $\mathbb{P}(\theta) = \mathcal{N}(\mu_0,\ \tau^2)$ and $\mathbb{P}(\sigma^2) \sim \text{InvGamma}\left(\frac{n_0}{2},\ \frac{n_0\sigma_0^2}{2}\right)$, find the kernel of the joint posterior, $\mathbb{P}(\theta,\ \sigma^2 \mid X)$.

$$\begin{aligned}
\mathbb{P}\left(\theta,\ \sigma^2 \mid X\right) &= \mathbb{P}\left(X \mid \theta,\ \sigma^2\right)\mathbb{P}(\theta)\,\mathbb{P}\left(\ \sigma^2\right) \\
&= \Pi \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \theta)^2\right) \exp\left(\frac{1}{2\tau^2}(\theta - \mu_0)^2\right) \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{x}-\theta)^2\right)} e^{-\frac{1}{2\tau^2}(\theta-\mu_0)^2} \left(\sigma^2\right)^{-\frac{n_0}{2}-1}
\end{aligned}$$

(c) [difficult] Show that this kernel can be factored into the kernel of a normal where the leftover is *not* the kernel of an inverse gamma. This is in the lecture notes.

$$\begin{aligned}
&\left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{x}-\theta)^2\right)} e^{-\frac{1}{2\tau^2}(\theta-\mu_0)^2} \left(\sigma^2\right)^{-\frac{n_0}{2}-1} \\
&\quad \propto \left(\sigma^2\right)^{-\frac{n}{2}-\frac{n_0}{2}-1} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n_0\sigma_0^2 + n\bar{x}^2\right)} e^{\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\theta - \left(\frac{n}{2\sigma^2}+\frac{1}{\tau^2}\right)\theta^2}
\end{aligned}$$

$$= (\sigma^2)^{-\frac{n}{2}-\frac{n_0}{2}-1} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n_0\sigma_0^2+n\bar{x}^2\right)} \sqrt{\frac{\pi}{b}} e^{-\frac{a^2}{4b}} \mathcal{N}\left(\frac{a}{2b}, \frac{1}{2b}\right) \mathbb{P}\left(X,\ \sigma^2 \mid \theta\right)$$

$$\propto (\sigma^2)^{-\frac{n}{2}-\frac{n_0}{2}-1} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n_0\sigma_0^2+n\bar{x}^2\right)} \left(\frac{n}{2\sigma^2}+\frac{1}{\pi}\right)^{-\frac{1}{2}} e^{\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2/4\left(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2}\right)}$$

(d) [difficult] [MA] Find the posterior mode of σ^2 using $k(\sigma^2 \mid X)$.

(e) [difficult] Describe how you would sample from $k(\sigma^2 \mid X)$. Make all steps explicit and use the notation from Table 1.

(a) Select $\sigma^2_{min}, \sigma^2_{max}\Delta$ for: $< \sigma^2_{min}, \sigma^2_{min}+\Delta, \cdots, \sigma^2_{max} >$

(b) Compute $C \approx \frac{1}{\sum_{\sigma^2 \in G} k(\sigma^2 \mid X)}$ for all $\sigma_0^2 \in G$ compute $F(\sigma_0^2 \mid X) = \mathbb{P}\left(\sigma^2 \leq \sigma_0^2 \mid X\right)$

(c) Draw u from runif(0,1) and ship $X_{samp} = avgmin_{\sigma^2 \in G}\{F(\sigma^2 \mid X) \geq u\}$

(d) Repeat c

(f) [difficult] Describe how you would sample from $\mathbb{P}\left(\theta,\ \sigma^2 \mid X\right)$. Make use of the sampling algorithm in the previous question. Make all steps explicit and use the notation from Table 1.

(a) Draw σ^2_{samp} from $k(\sigma^2 \mid X)$

(b) Draw θ_{samp} from $\mathcal{N}\left(\theta,\ \sigma^2\right)$

(c) Ship $< \theta_{samp}, \sigma^2_{samp} >$ as one sample

(d) Repeat a-c s times

(g) [difficult] What are the two main disadvantages of grid sampling?

(a) Selecting the min, max and Δ are needed.

(b) if Δ is small and $dim[\theta]$ is large, grid sampling fails.

(h) [difficult] Why do you think the prior $\mathbb{P}(\theta) = \mathcal{N}\left(\mu_0,\ \tau^2\right)$ and $\mathbb{P}\left(\sigma^2\right) \sim \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$ is called "semi-conjugate"?

It combines the normal and the inverse gamma to make a normal inverse gamma distribution.

(i) [E.C.] [MA] Find the MMSE of σ^2

Problem 3

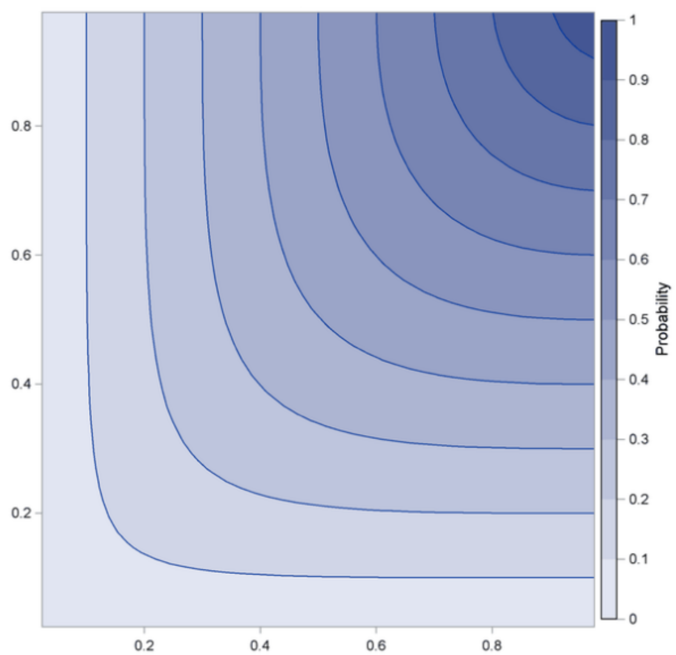
These are questions which introduce Gibbs Sampling.

(a) [easy] Outline the systematic sweep Gibbs Sampler algorithm below (in your notes).

   (a) Begin with a reasonable value for $\theta (\theta_0 = \theta)$
   (b) Draw σ_1^2 from $\mathbb{P}(\sigma^2 \mid X, \theta)$ where $\theta_0 = \theta$
   (c) Draw θ_1 from $\mathbb{P}(\theta \mid X, \sigma^2)$ where $\sigma^2 = \sigma_1^2$
   (d) Draw σ_2^2 from $\mathbb{P}(\sigma^2 \mid X, \theta)$ where $\theta = \theta_1$
   (e) Draw θ_2 from $\mathbb{P}(\theta \mid X, \sigma^2)$ where $\sigma^2 = \sigma_2^2$
   (f) Repeat s times

(b) [E.C.] Under what conditions does this algorithm converge?

(c) [easy] Pretend you are estimating $\mathbb{P}(\theta_1,\ \theta_2 \mid X)$ and the joint posterior looks like the picture below where the x axis is θ_1 and the y axis is θ_2 and darker colors indicate higher probability. Begin at $[\theta_1, \theta_2] = [0.5, 0.5]$ and simulate 5 iterations of the systematic sweep Gibbs sampling algorithm by drawing new points on the plot (just as we did in class).



Problem 4

These are questions about the change point model and the Gibbs sampler to draw inference for its parameters. You will have to use R to do this question. If you do not have it installed on your computer, you can use R online without installing anything by using a site like jupyter. You copy code into the black box and click the "run" button atop. Then you enter more code into the next box and click "run" again, etc.

(a) [easy] Consider the change point Poisson model we looked at in class. We have m exchangeable Poisson r.v.'s with parameter λ_1 followed by $n-m$ exchangeable Poisson r.v.'s with parameter λ_2. Both rate parameters and the value of m are unknown so the parameter space is 3-dimensional. Write the likelihood below.

$$\begin{aligned}\mathbb{P}(\lambda_1, \lambda_2, m \mid X_1, \cdots, X_n) &\propto \mathbb{P}(X_1, \cdots, X_n \mid \lambda_1, \lambda_2, m)\, \mathbb{P}(lambda_1, \lambda_2, m) \\ &= \left(\prod_{t=1}^{m} \frac{e^{-\lambda_1}\lambda_1^{X_t}}{X_t!}\right)\left(\prod_{t=m+1}^{n} \frac{e^{-\lambda_2}\lambda_2^{X_t}}{X_t!}\right)\left(\frac{1}{n-1}\right)\end{aligned}$$

(b) [easy] Consider the model in (a) where $\lambda_1 = 2$ and $\lambda_2 = 4$ and $m = 10$ and $n = 30$. Run the code on lines 1–14 of the code at the link here by copying them from the website and pasting them into an R console. This will plot a realization of the data with those parameters. Can you identify the change point visually?

The change point is approximately 12.

(c) [easy] Consider the model in (a) but we are blinded to the true values of the parameters given in (b) and we wish to estimate them via a Gibbs sampler. Run the code on lines 16–78 of the code at the link here which will run 10,000 iterations. What iteration number do you think the sampler converged?

It looks like 10.

(d) [easy] Now we wish to assess autocorrelation among the chains from the Gibbs sampler run in (d). Run the code on lines 79–89 of the code at the link here. What do we mod our chains by to thin them out so the chains represent independent samples?

8

(e) [easy] Run the code on lines 91–121 of the code at the link here which will first burn and thin the chains. Explain these three plots. What distributions do these frequency histograms approximate? You must have $\mathbb{P}$(something) in your answer. What are the blue lines? What are the red lines? What are the grey lines? Read the code if you have to for the answers.

- The histograms approximate $\mathbb{P}(\lambda_1 \mid X_1, \cdots, X_n, \lambda_2, m)$, $\mathbb{P}(\lambda_2 \mid X_1, \cdots, X_n, \lambda_1, m)$ and $\mathbb{P}(m \mid X_1, \cdots, X_n, \lambda_1, \lambda_2)$.
- The blue line are the means ($\lambda_1 = 1.6, \lambda_2 = 3.25, m = 9.4$).
- The red lines are the true values ($\lambda_1 = 2, \lambda_2 = 4, m = 10$).
- The grey lines are the 95% credible regions ($\lambda_1 : [0.7, 2.7], \lambda_2 : [2.4, 4.2], m : [2, 19]$).

(f) [difficult] Test the following hypothesis: $H_0 : m \leq 15$ by approximating the p-value from one of the plots in (e).

$$\begin{aligned} P_{val} &= \mathbb{P}\left(m \leq 15 \mid X\right) \\ &= \frac{1}{30}\Sigma \mathbb{1} m \leq 15 \\ &= 0.9597598 > (\alpha = 0.05) \end{aligned}$$

(g) [difficult] [M.A.] Explain a procedure to test $H_0 : \lambda_1 = \lambda_2$. You can use the plots if you wish, but you do not have to.

(h) [difficult] What exactly would come from $\mathbb{P}\left(X^* \mid X\right)$ in the context of this problem? Assume X^* is the same dimension of X (in our toy example, $n = 30$). Explain in full detail. Be careful!

It would predict the 31st data point given our estimates $(\lambda_1, \lambda_2, m)$ and the previous data. $\mathbb{P}\left(X^* \mid X\right) = \int\int\int \mathbb{P}\left(X^* \mid \lambda_1, \lambda_2, m\right) \mathbb{P}\left(\lambda_1, \lambda_2, m \mid X\right) d\lambda_1 d\lambda_2 dm$

(i) [E.C.] Explain how you would estimate $\mathbb{C}\text{ov}\left[\lambda_1, \lambda_2\right]$ and what do you think this estimate will be close to?

Problem 5

These are questions about the mixture-of-two-normals model and the Gibbs sampler to draw inference for its parameters. You will have to use R to do this question. If you do not have it installed on your computer you can use this website which will give you provide you with a workable R console.

(a) [easy] Consider the mixture-of-two-normals model we looked at in class. Write the likelihood below.

$$\mathbb{P}\left(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2,\ \rho\right) = \prod_{i=1}^{n}(\rho \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}\left(x_i - \theta_1\right)^2\right) + (1-\rho)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}\left(x_i - \theta_2\right)^2\right))$$

(b) [easy] Consider the model in (a) with $\theta_1 = 0,\ \theta_2 = 4,\ \sigma_1^2 = 2,\ \sigma_2^2 = 1$ and $\rho = 2$. These are the targets of inference so pretend you don't know their values! Run the code on lines 1–16 of the code at the link here by copying them from the website and pasting them into an R console. This will plot a realization of the data with those parameters. Can you identify that it's a mixture of two normals visually?

Yes since there are two humps. Some will samples belong to the first and some to the second one.

(c) [easy] Consider the model in (a) but we are blinded to the true values of the parameters given in (b) and we wish to estimate them via a Gibbs sampler. Run the code on lines 19–92 of the code at the link here which will run 10,000 iterations. What iteration number do you think the sampler converged?

Iteration number of about 50

(d) [easy] Now we wish to assess autocorrelation among the chains from the Gibbs sampler run in (d). Run the code on lines 96–103 of the code at the link here. What do we mod our chains by to thin them out so the chains represent independent samples?

We can modify the thinning, include more samples, or more iterations.

(e) [easy] Run the code on lines 120–152 of the code at the link here which will first burn and thin the chains. Explain these five plots. What distributions do these frequency histograms approximate?

The blue line represents the best guess for each of the five parameters. These distograms approximate the normal distribution.

(f) [difficult] Provide and approximate $CR_{\rho,95\%}$. Does it capture the true value of ρ?

$$\mathrm{CR}_{\rho,95\%} = [0.235, 0.375], \text{ It does not capture } \rho$$

(g) [difficult] Explain carefully how you would approximate $\mathbb{P}(X^* \mid X)$.

(h) [difficult] Explain carefully how you would approximate the probability that the 17th observation belonged to the $\mathcal{N}(\theta_1,\, \sigma_1^2)$ distribution.

Take th expectation of the best guess. With that you can see the probability if the 17th observation belongs to $\mathcal{N}(\theta_1,\, \sigma_1^2)$

(i) [easy] If one of the θ's did not have a known conditional distribution, which algorithm could you use? Would this algorithm take longer or shorter to converge than the Gibbs sampler you've seen here?

You could use grid sampling. This algorithm would take longer to converge.

(j) [E.C.] Explain carefully how you would test if $\theta_1 \neq \theta_2$.

Problem 6

This question is about a famous extra sensory perception (ESP) experiment.

According to quantum mechanics, an event should happen with exactly probability 50%. So if someone comes with a claim that they have ESP and can affect this event with their mind, then the event no longer has probability 50% (this is what we are trying to prove).

An experiment was run with someone claiming they have ESP. They tried to affect the event with their mind. The data is as follows: of $n = 104,490,000$ observations, the number of events was $x = 52,263,970$. We will now test to see if the person has ESP a number of ways and try to reconcile the differences.

(a) [easy] What is the MLE for the probability of the event?

$$\hat{\theta}_{\text{MMSE}} = \bar{x} = \frac{1}{n}\sum x_i = 0.5001$$

(b) [easy] Run a two-sided frequentist test at $\alpha = 5\%$ and report the decision (i.e. retain or reject) and the p value.

$$\begin{aligned}
\text{RR} &= [0.5 \ \pm \ 2\sqrt{\frac{0.5(1-0.5)}{104,490,000}}] = [0.49, 0.50] \\
\text{H}_0 &: \theta = 0.5 \\
\text{H}_a &: \theta \neq 0.5 \\
\theta &\in \text{RR}_{\theta,95\%} \text{ Retain } H_0 \\
P_{val} &= 0.003 < \alpha = 0.5
\end{aligned}$$

(c) [harder] Run a two-sided Bayesian test assuming the principle of indifference. Use the equivalence region approach with $\delta = 0.01$. Report the decision (i.e. retain or reject) and the p value.

$$\begin{aligned}
\text{H}_0 &: \theta = 0.5 \\
\text{H}_a &: \theta \neq 0.5 \\
\text{CR}_{\theta,95\%} &= [\text{qbeta}(0.025, x+1, n-x+1), \text{qbeta}(0.975, x+1, n-x+1)] \\
&= [0.50008567, 0.5002774]
\end{aligned}$$

Ratain H_0 since $\theta \in \text{CR}_{\theta,95\%}$

(d) [difficult] Calculate the Bayes Factor B where the alternative hypothesis of $\theta \neq 0.5$ can be summed up with $\theta \sim \mathrm{U}\,(0,\,1)$.

$$
\begin{aligned}
\mathrm{H}_0 &: \theta = 0.5 \\
\mathrm{H}_a &: \theta \neq 0.5 \\
\alpha &= 5\% \\
\mathrm{B} &= \frac{\mathrm{B}(52,263,971,\ \ 52,226,031)}{(0.5)^{104,490,000}} \\
&\approx 0.33
\end{aligned}
$$

Retain H_0

(e) [difficult] Try to reconcile parts (b), (c), (d), (e). Why are there different answers? What is going on??

With the frequentist pval, bayesian pval, and Bayes Factor B we access "statistical significance". We could also messure "clinical significance", how important the actual effect of theta is. Bayes factors puts both of these ideas together.