MATH 341 / 650.3 Spring 2021 Homework #7

Professor Adam Kapelner

NOT DUE

(this document last updated Thursday 13th May, 2021 at 4:52pm)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read about the normal-inverse-gamma conjugate model with Jeffrey's priors, marginal posteriors in the normal-inverse-gamma conjugate model with Jeffrey's priors, how a non-conjugate model can break our inferential paradigm, how this can be solved with grid sampling, disadvantages to grid sampling, the systematic sweep Gibbs sampler, and Bayesian inference in the change point model.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to attempt the *difficult* problems.

Problems marked "[MA]" are for the masters students only (those enrolled in the 650.3 course). For those in 341, doing these questions will count as extra credit.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 10 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

These are questions about McGrayne's book, chapter 16.

(a) [easy] What was the main problem facing Bayesian Statistics in the early 1980's?

(b) [harder] What is the "curse of dimensionality?"

(c) [easy] How did Bayesian Statistics help sociologists?

(d) [easy] How did Gibbs sampling come to be?

(e) [easy] Were the Geman brothers the first to discover the Gibbs sampler?

(f) [easy] Who officially discovered the expectation-maximization (EM) algorithm? And who *really* discovered it?

(g) [harder] How did Bayesians "break" the curse of dimensionality?

(h) [harder] Consider the integrals we use in class to find expectations or to approximate PDF's / PMF's — how can they be replaced?

(i) [easy] What did physicists call "Markov Chain Monte Carlo" (MCMC)? (p222)

(j) [easy] Why is sampling called "Monte Carlo" and who named it that?

(k) [easy] The Metropolis-Hastings (MH) Algorithm is world famous and used in myriad applications. Why didn't Hastings get any credit?

(l) [easy] The combination of Bayesian Statistics + MCMC has been called ... (p224)

(m) [E.C.] p225 talks about Thomas Kuhn's ideas of "paradigm shifts." What is a "paradigm shift" and does Bayesian Statistics + MCMC qualify?

(n) [easy] How did the BUGS software change the world?

(o) [easy] Lindley said that Bayesian Statistics would win out over Frequentist Statistics because it was more logical. What in reality was the reason for the eventual victory of Bayes?

(p) [E.C.] One of my PhD advisors, Ed George at Wharton told me that "Bayesian Statistics is really 'knowledge engineering."' Is this true? Explain.

(q) [E.C.] Take a look at the software Stan. What kind of potential does it have to change the world? Note: I had an opportunity to work on Stan as a postdoc (right after I finished his PhD) but chose to come to QC instead.

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

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

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

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

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

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

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

Now we will move to the Bayesian normal-normal model for estimating both the mean and variance and demonstrate similarities with the classical results.

(a) [harder] If $X_1, \ldots, X_n \mid \theta, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\theta,\ \sigma^2)$ and X represents all $X_1, \ldots, X_n$, Find the kernel of $\mathbb{P}(\theta,\ \sigma^2 \mid X)$ if $\mathbb{P}(\theta,\ \sigma^2) \propto \frac{1}{\sigma^2}$. Use the substitution that we made in class:

$$\sum_{i=1}^{n}(x_i - \theta)^2 = (n-1)s^2 + n(\bar{x} - \theta)^2$$

where $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. We do this here because this substitution is important for what comes next.

(b) [harder] Using Bayes Rule, break up $\mathbb{P}(\theta,\ \sigma^2 \mid X)$ into two pieces. How are those two pieces distributed?

(c) [harder] Using your answer from (b), explain in English how you can create samples from the distribution $\mathbb{P}(\theta,\ \sigma^2 \mid X)$ that look like $\{[\theta_1, \sigma_1^2], [\theta_2, \sigma_2^2], \ldots, [\theta_S, \sigma_S^2]\}$.

(d) [difficult] Using these samples, how would you estimate $\mathbb{E}[\theta \mid X]$ and $\mathbb{E}[\sigma^2 \mid X]$? Why is $\mathbb{E}[\theta \mid X]$ of paramount importance?

(e) [difficult] Using these samples, how would you estimate a 95% CR for θ?

(f) [difficult] Using these samples, how would you obtain a p-val for testing if $\sigma^2 > 1.364$?

(g) [difficult] [MA] Using these samples, how would you estimate $\text{Corr}\left[\theta \mid X,\ \sigma^2 \mid X\right]$ i.e. the correlation between the posterior distributions of the two parameters?

(h) [easy] Find $\mathbb{P}\left(\theta \mid X,\ \sigma^2\right)$ by using the full posterior kernel from (a) and then conditioning on σ^2. You should get the same answer as we did before the midterm.

(i) [easy] Find $\mathbb{P}(\sigma^2 \mid X,\ \theta)$ by using the full posterior kernel from (a) and then conditioning on θ. You should get the same answer as we did before the midterm.

(j) [difficult] Show that $\mathbb{P}(\theta \mid X)$ is a non-standard T distribution and find its parameters. Assume the prior $\mathbb{P}(\theta,\ \sigma^2) \propto \frac{1}{\sigma^2}$. The answer is in the notes, but try to do it yourself.

(k) [difficult] Show that $\mathbb{P}(\sigma^2 \mid X)$ is an inverse gamma and find its parameters. Assume the prior $\mathbb{P}(\theta,\ \sigma^2) \propto \frac{1}{\sigma^2}$. The answer is in the notes, but try to do it yourself.

(l) [easy] Write down the distribution of $\mathbb{P}(X^* \mid X)$ assuming the prior $\mathbb{P}(\theta,\ \sigma^2) \propto \frac{1}{\sigma^2}$. This is in the notes.

(m) [E.C.] Prove what you wrote in the previous question: $\mathbb{P}(X^* \mid X)$ is the non-standard T distribution and find its parameters.

(n) [harder] Explain how to sample from the distribution of $\mathbb{P}(X^* \mid X)$. Also in the notes.

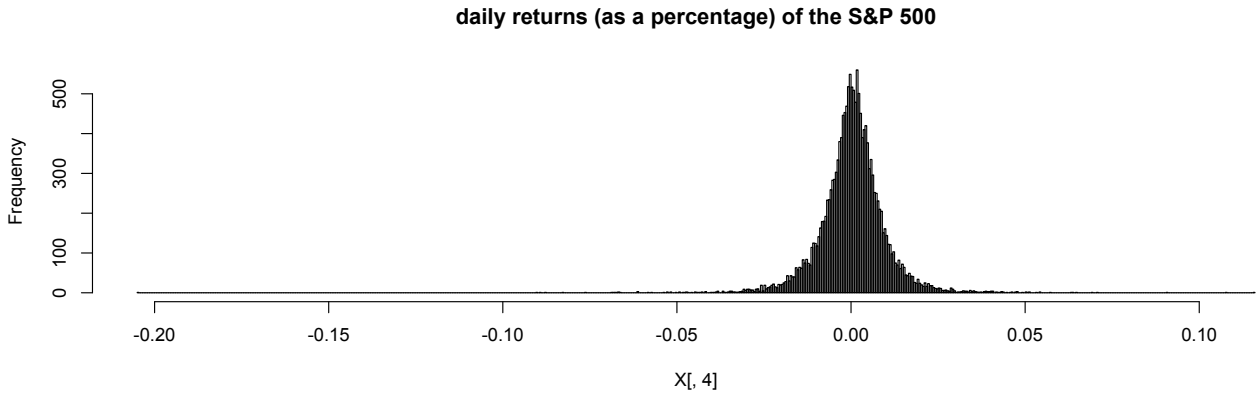
(o) [harder] Now consider the informative conjugate prior of $\mathbb{P}(\theta,\ \sigma^2) = \mathbb{P}(\theta \mid \sigma^2)\,\mathbb{P}(\sigma^2)$ where $\mathbb{P}(\theta \mid \sigma^2) = \mathcal{N}\left(\mu_0,\ \frac{\sigma^2}{m}\right)$ and $\mathbb{P}(\sigma^2) = \text{InvGamma}\left(\frac{n_0}{2},\ \frac{n_0\sigma_0^2}{2}\right)$ i.e. the general normal-inverse-gamma. What is its kernel? Collect common terms and be neat.

(p) [difficult] [MA] If $X_1, \ldots, X_n \mid \theta, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\theta,\ \sigma^2)$ and given the general prior above, find the posterior and demonstrate it that the normal-inverse gamma is conjugate for the normal likelihood with both mean and variance unknown. This is what I did *not* do in class.

Problem 3

We model the returns of S&P 500 here.

(a) [easy] Below are the 16,428 daily returns (as a percentage) of the S&P 500 dating back to January 4, 1950 and the code used to generate it. Does the data look normal? Yes/no



(b) [harder] Do you think the data is $\overset{iid}{\sim}$? Explain.

(c) [harder] Assume $\overset{iid}{\sim}$ normal data regardless of what you wrote in (a) and (b). The sample average is $\bar{x} = 0.0003415$ and the sample standard deviation is $s = 0.0096$. Under an objective prior, give a 95% credible region for the true mean daily return.

(d) [difficult] Give a 95% credible region for *tomorrow's* return using functions in Table 1.

Problem 4

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?

(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, $\mathbb{P}(\theta,\ \sigma^2 \mid X)$.

(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.

(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.

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

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

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

(i) [E.C.] [MA] Find the MMSE of σ^2

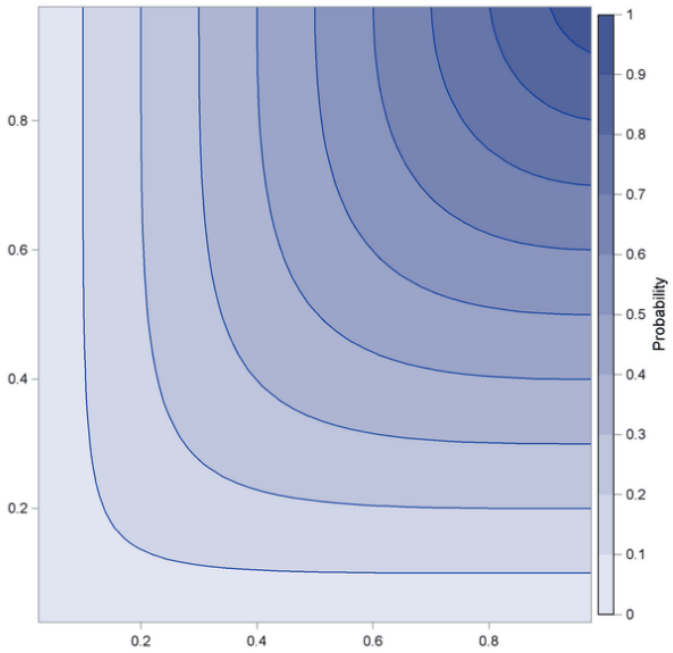
Problem 5

These are questions which introduce Gibbs Sampling.

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

(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 6

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.

(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?

(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?

(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?

(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 $\mathbb{P}(\text{something})$ 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.

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

(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 $\mathbb{P}(X^* \mid X)$ 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!

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