MATH 343 / 643 Homework #1

Osman Khan

Monday 4th March, 2024

Problem 1

These are general questions about Gibbs Sampling and Metropolis-within-Gibbs Sampling.

(a) [easy] Let dim $[\boldsymbol{\theta}] = p$ and assume a prior $f(\boldsymbol{\theta})$ to be continuous. Describe the steps of the systematic sweep Gibbs Sampler algorithm below that will converge to $f(\boldsymbol{\theta} \mid \boldsymbol{X})$. Label the steps that are necessary for the p dimensions separately e.g. Step 2.1, Step 2.2, ..., Step 2.p. You need to reference these step numbers later on in the problem.

Step 0. Initialize $\vec{\theta_0} = [\theta_{0,1}, \theta_{0,2}, \dots \theta_{0,p}]$

Step 1.0 The steps below will give us $\vec{\theta_1}$

Step 1.1 Draw $\theta_{1,1}$ from $f(\theta_1 | \theta_{0,2}, \theta_{0,3}, \dots \theta_{0,p}, \vec{X})$

Step 1.2. Draw $\theta_{1,2}$ from $f(\theta_2 | \theta_{1,1}, \theta_{0,3}, \dots \theta_{0,p}, \vec{X})$

$\vdots$

Step 1.K. Draw $\theta_{1,p}$ from $f(\theta_k | \theta_{1,1}, \theta_{1,2}, \dots \theta_{1,p-1}, \vec{X})$

Step 2.0. Repeat Steps 1-K at each integer step (including this) until 'convergence', which is $f(\theta_1, \theta_2, \dots \theta_p | \vec{X})$. See the difference between $\vec{\theta_0}$&$\vec{\theta_1}$

(b) [easy] What are all the items you need to know in order to write code for that implements a Gibbs Sampler?

We need i) some initial state of the variables (can possibly be chosen randomly), ii) number for burning, iii) number for thinning, iv) the ability to sample one parameter given everything else.

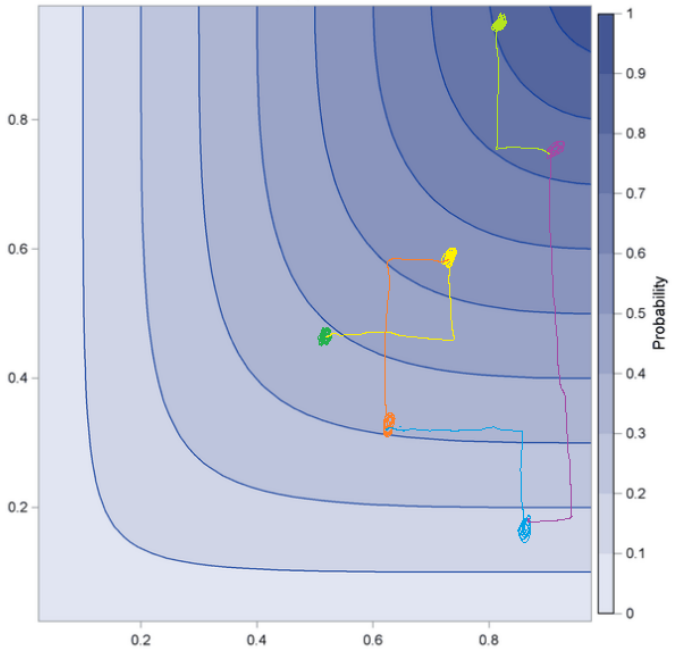
(c) [easy] Explain what burning of the chain is and why it is necessary.

Burning is the discarding of the samples at the start as they are not useful and are not as representative of the target posterior distribution.

(d) [easy] Explain what thinning of the chain is and why it is necessary.

Gibbs sampling can suffer from autocorrelation. Thinning helps in preventing that, as picking samples at some step helps in reducing dependence between the samples picked.

(e) [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.



Color of iterations respectively:Yellow, Orange, Azure Blue, Purple Lime Green.

(f) [easy] Consider the need to implement a Metropolis Hastings step within the Sampler for θ_j. Why would you need to do this? At which step (reference your steps in part a) would you require it?

Metropolis-Hastings is implemented when sampling is difficult. It will be implemented after step 0 for whichsoever step .i's parameter for which sampling is difficult.

(g) [easy] If Supp $[\theta_j] = \mathbb{R}$, propose a default proposal distribution to start with:

$q(\theta_{t,j} \mid \theta_{t-1,j}, \phi) =$

Remember, the mean of proposal distributions should be $\theta_{t-1,j}$ (or close to that value) and ϕ are additional parameters which may or may not be used.

Normal $(\theta_j, \sigma^2) = N(\theta_{t-1,j}, 1^2)$

(h) [harder] How do you know if this proposal distribution is a good choice or not?

This is a good choice as it is defined for all reals. We also don't know whether or not the tails are heavy, so Normal is a safe bet. In addition, the mean is $\theta_{t-1,j}$.

(i) [difficult] If Supp $[\theta_j] = (0, \infty)$, propose a proposal distribution

$q(\theta_{t,j} \,|\, \theta_{t-1,j}, \phi) =$

A Gamma Distribution = Gamma $(\theta_{t-1,j}, 1)$. The mean is $\theta_{t-1,j}$.

(j) [difficult] [MA] If Supp $[\theta_j] = [0, 1]$, propose a proposal distribution

$q(\theta_{t,j} \,|\, \theta_{t-1,j}, \phi) =$

A Beta Distribution = Beta $(\theta_{t-1,j} - \frac{1}{2}, \frac{3}{2})$. The mean is $\theta_{t-1,j}$.

Problem 2

Consider a count model that has many zeroes. We choose to fit it with a hurdle model

$$X_1, \ldots, X_n \overset{iid}{\sim} \begin{cases} 0 & \text{w.p. } \theta_1 \\ \text{ShiftedExtNegBinomial}(\theta_2, \theta_3, +1) & \text{w.p. } 1 - \theta_1 \end{cases}$$

where the shifted distribution is just the extended negative binomial distribution so that the probability of realizing a count of one is the probability of realizing a count of zero, the probability of realizing a count of two is the probability of realizing a count of one, etc. i.e.

$$\text{ShiftedExtNegBinomial}(\theta_2, \theta_3, +1) := p(x) = \frac{\Gamma(x_i - 1 + \theta_2)}{(x_i - 1)!\Gamma(\theta_2)}(1 - \theta_3)^{x_i - 1}\theta_3^{\theta_2}.$$

(a) [harder] What is the parameter space for all three parameters of interest? This may require looking at your MATH 340 notes.

θ_3 & $\theta_1 \in [0, 1], \theta_2 \in (0, \infty)$

(b) [harder] Assume a flat prior $f(\theta_1, \theta_2, \theta_3) \propto 1$. Find the kernel of the posterior distribution $f(\theta_1, \theta_2, \theta_3 \,|\, \boldsymbol{x}, n_0, n_+)$ where $\boldsymbol{x} := \{x_1, \ldots, x_n\}$, the observations. Let n_0 be the number of zeroes in the dataset and $n_+ := n - n_0$, the number > 0 in the dataset.

$f(\theta_1, \theta_2, \theta_3 \,|\, \boldsymbol{x}, n_0, n_+) = \Pi_{i=1}^n \theta^{\mathbb{1}_{x_i=0}=n_0}(\frac{\Gamma(x_i-1+\theta_2)}{(x_i-1)!\Gamma(\theta_2)}(1-\theta_3)^{1-x_i}(\theta_3)^{\theta_2})^{\mathbb{1}_{x_i>0}=n_+}$

$\vdots$

$\alpha\theta_1^{n_0}\frac{\Gamma(x_i-1+\theta_2)^{n_+}}{\Gamma(\theta_2)^{n_+}}(1-\theta_3)^{\sum(1-x_i)\mathbb{1}_{x-i>0}}\theta_3^{\theta_2 n_+}$

(c) [harder] Find the log of the kernel of the posterior distribution.

$n_0 \ln(\theta_1) + n_+ \ln(x_i - 1 + \theta_2) - n_+ \ln(\Gamma(\theta_2)) + n_+ \ln(\theta_3) + \sum(1 - x_i)\mathbb{1}_{x_i>0}(\ln(1 - \theta_3))$

(d) [easy] Find the conditional distribution $f(\theta_1 \mid \boldsymbol{x}, n_0, n_+, \theta_2, \theta_3)$ as a brand name rv.

$f(\theta_1 \mid \boldsymbol{x}, n_0, n_+, \theta_2, \theta_3) = \theta_1^{\sum(\mathbb{1}_{x_i=0})} = \theta_1^{n_0} = \theta_1\theta_1^{n_0-1}(1-\theta_1)^{1-1} = \theta_1\theta_1^{n_0-1}(1-\theta_1)^{1-1}\frac{(\theta_1-1)!}{(\theta_1-1)!} = \theta_1^{n_0-1}(1-\theta_1)^{1-1}\frac{\Gamma(\theta_1+1)}{\Gamma(\theta_1)\Gamma(1)} \sim Beta(\theta_1, 1)$

(e) [easy] Find the kernel of the conditional distribution $f(\theta_2 \mid \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3)$.

$f(\theta_2 \mid \boldsymbol{x}, n_0, n_+, \theta_2, \theta_3) = \theta_1^{\sum(\mathbb{1}_{x_i=0})} = (\frac{\Gamma(x_i-1+\theta_2)}{\Gamma(\theta_2)}(\theta_3)^{\theta_2})^{\mathbb{1}_{x_i>0}=n_+}$

(f) [easy] Is the conditional distribution $f(\theta_2 \mid \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3)$ a brand name rv?

No

(g) [easy] Given your answer in (a), the Supp $[\theta_2]$ and your answer from problem 1(k) which was marked difficult, provide a proposal distribution

$q(\theta_{t,2} \mid \theta_{t-1,2}, \phi) =$

A Gamma Distribution = Gamma$(\theta_{t,2}, 1)$

(h) [easy] Find the conditional distribution $f(\theta_3 \mid \boldsymbol{x}, n_0, n_+, \theta_1, \theta_2)$ as a brand name rv.

Problem 3

These are general questions about Permutation Testing.

(a) [easy] What are the null and alternative hypotheses for a two-sample permutation test?

$H_0 = DGP_1 = DGP_2 = DGP; H_a = DGP_1 \neq DGP_2$

(b) [easy] Let n_1 and n_2 be the sample sizes from population one and population two respectively. How many possible sample "permutations" are there? I put permutations in quotes because it's not truly a "permutation" in the sense that you were taught in MATH 241.

The possible sample permutations are:
$\sum_{i=1}^{n} \binom{n}{n_1}$

(c) [easy] Give three examples of a test statistic to employ within the body of the loop of a permutation test.

The three examples of a test statistic are:
a) $\hat{\hat{\theta}}_b = \bar{x}_{b1} - \bar{x}_{b2}$,
b) $\hat{\hat{\theta}}_b = S_{b1}^2 - S_{b2}^2$,
c) $\hat{\hat{\theta}}_b = \frac{S_{b1}^2}{S_{b2}^2}$

(d) [difficult] Explain how you would calculate a p-value in a permutation test.

I will choose a test statistic, and will compute it over all possible permutations (considering n is a reasonable number). Then, I will take the average of the test statistic computed over all possible permutations. This is different from what we covered in class, which was: $\min\{2P(\hat{\theta}_b > \hat{\hat{\theta}}), 2P(\hat{\theta}_b < \hat{\hat{\theta}})\}$, which was about expanding the retainment region so that it contains the test statistic.

Problem 4

These are general questions about the Bootstrap. Assume $X_1, \ldots, X_n \overset{iid}{\sim}$ some DGP.

(a) [easy] Describe the steps in the bootstrap procedure for the estimate $\hat{\hat{\theta}} := w(x_1, \ldots, x_n)$ which estimates θ.

Step 0) Specify a B. That will be the number of times resampling will be done from the sample: $\{x_1, x_2, \ldots, x_n\}$. Choose a test statistic of interest: ϕ

Step 1) Draw $\{x_{b,1}, x_{b,2}, \ldots, x_{b,n}\}$ with replacement from the sample, and compute $\hat{\hat{\phi}}$ from the resampling.

Step 2) Repeat Step 1) B number of times.

Step 3) The collection of tests statistics you get, i.e $\{\hat{\hat{\phi}}_1, \hat{\hat{\phi}}_2, \ldots, \hat{\hat{\phi}}_B\}$, can be tought of as samples from $\hat{\phi}_{boot_n}$which is approximately ϕ.

(b) [easy] In what situations should the bootstrap be employed instead of other inferential procedures you learned about?

Bootstrap should be employed when repeating the experiment is expensive. It should also be employed as we get more computing power. Most importantly, it should be used when we are interested in a (parametric) test statistic for which our regular methods (i.e. wald test, t test, etc) do not have reasonable power (for reasons such as not enough sample size, etc).

(c) [difficult] Explain in what situations the bootstrap fails.

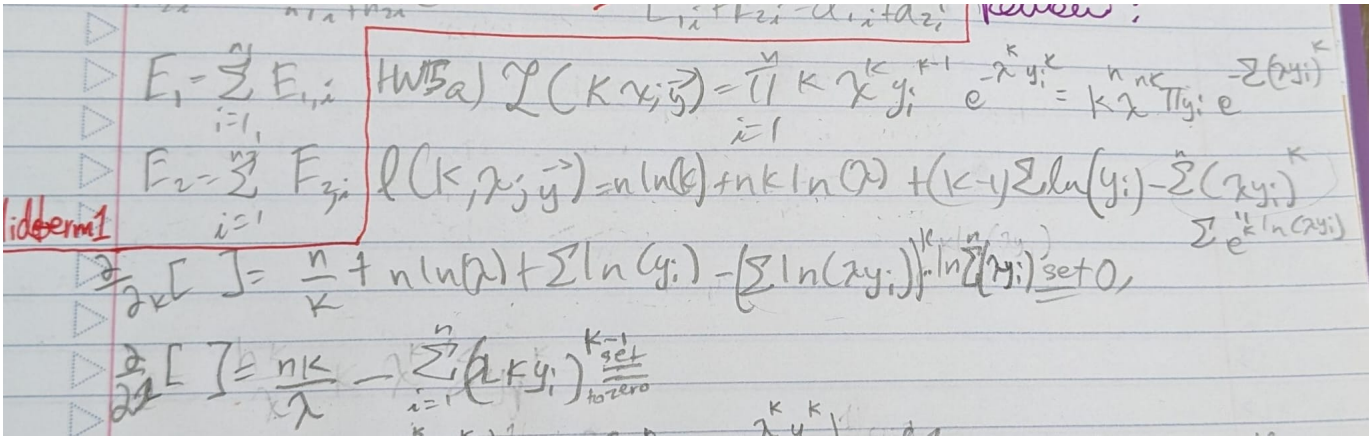
Bootstrap fails when we choose non-representative samples or when we don't choose an appropriate test statistic.

Problem 5

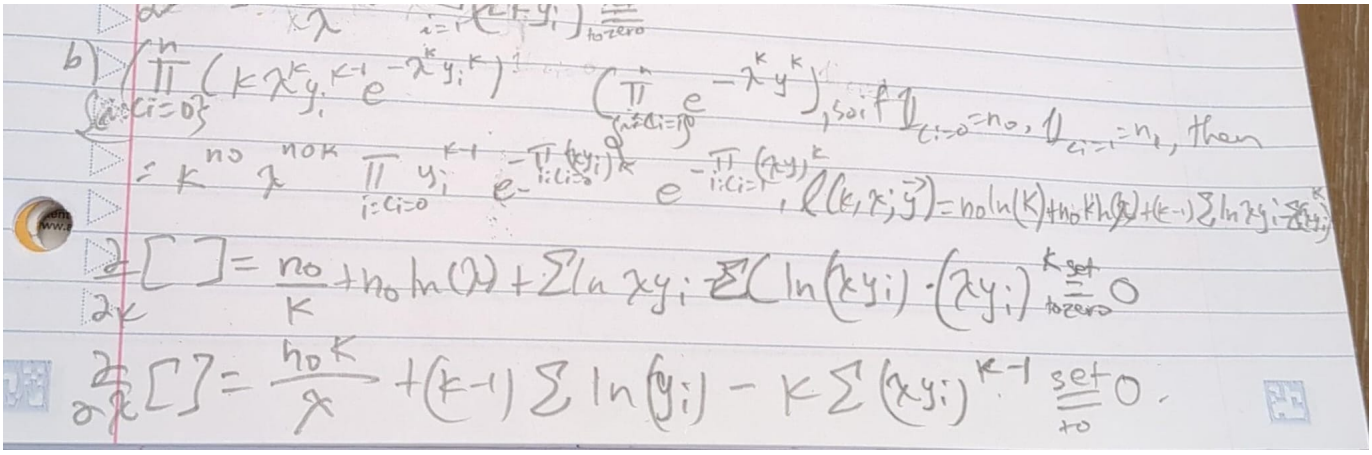
These are questions about parametric survival using the Weibull model i.e.

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Weibull}(k, \lambda) := f(y) = k\lambda^k y^{k-1} e^{-\lambda^k y^k} \mathbb{1}_{y>0}, \quad F(y) = 1 - e^{-\lambda^k y^k}, \quad S(y) = e^{-\lambda^k y^k}$$

(a) [difficult] Assume no censoring in the data. Find closed form expressions and/or equations for the MLEs of k and λ



(b) [difficult] Assume censoring in the data so that $\mathbf{c}$ is the binary vector that is one when censored and zero if measured. Let $\boldsymbol{y}$ be the vector of measurements or censored values if not measured. Find $\ell(k, \lambda; \boldsymbol{y}, \boldsymbol{c})$.



(c) [harder] In class we proved that $\mathbb{E}[Y] = \frac{1}{\lambda}\Gamma\left(1+\frac{1}{k}\right)$. Use this result to find $\mathbb{E}[Y \mid Y > a]$ where $a > 0$. You should first find the density of the truncated distribution. Then the expectation of this distribution will basically follow the same steps as found in lecture when we derived the expectation.

(d) [harder] Describe the steps in an EM algorithm to find the maximum likelihood estimates of k and λ.

Step 0. Have guesses for the missing $y_{i's}$ i.e. t_f

Step 1. Compute $\hat{\hat{K}}^{MLE}$&$\hat{\hat{\lambda}}^{MLE}$ (M-step)

Step 2. Compute better guesses for the missing y_i's $E[Y_i|Y_i > t_f]$ which is the same for all i's censored. (E-step)

Step 3. Repeat Steps 1 & 2 until convergence i.e. your k, λ estimates don't change too much between iterations.

Problem 6

These are questions about nonparametric survival inference.

(a) [harder] Explain how the Kaplan-Meier estimator differs from the empirical survival function if there is censoring at all different times before and after the maximum measured survival. There is only one difference!

The censored data is considered in the death that occurs before the censoring.

(b) [easy] Consider the dataset $y = \{79, 81, 92+, 95, 105+, 107, 122\}$ where the "+" signs indicate censored values. Draw the Kaplan-Meier estimate of $S(y)$. Try to make it to scale as best as possible.

(c) [easy] Write the hypotheses for the log-rank test.

$H_0 : DGP_1 = DGP_2$

$H_a : DGP_1 \neq DGP_2$

(d) [easy] Write the formula for the test statistic in the log-rank test.

$\hat{\theta} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \dot{\sim} \chi^2_1$, where:

$O_1 = \#$ of events in Sample 1 $\Rightarrow O_1 = \sum \mathbb{1}_{c_{1i}} = 0$,

$O_2 = \#$ of events in Sample 2 $\Rightarrow O_1 = \sum \mathbb{1}_{c_{2i}} = 0$

$E_{1i} := (d_{1i} + d_{2i}) \frac{n_{1i}}{n_{1i} + n_{2i}}$

$E_{2i} := (d_{1i} + d_{2i}) \frac{n_{2i}}{n_{1i} + n_{2i}}$

$E_{1i} + E_{2i} = d_{1i} + d_{2i}$

$E_1 = \sum_{i=1}^{n_1} E_{1i}$

$E_2 = \sum_{i=1}^{n_2} E_{2i}$