MATH 343 / 643 Homework #1

Natasha Watson

Sunday 2nd March, 2025

Problem 1

These are general questions about 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 1.0: intialize $\vec{\theta} = [\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,p}]$

Step 1.1: = draw iteration 1 $\theta_{1,1}$ from $f(\theta_1 \mid x; \theta_2 = \theta_{0,2}, \ldots, \theta_p = \theta_{0,p})$

Step 1.2: = draw $\theta_{1,2}$ from $f(\theta_2 \mid x; \theta_1 = \theta_{1,1}, \theta_3 = \theta_{0,3}, \ldots, \theta_p = \theta_{0,p})$

Step 1.3: = draw $\theta_{1,3}$ from $f(\theta_3 \mid x; \theta_1 = \theta_{1,1}, \theta_2 = \theta_{2,1}, \theta_4 = \theta_{0,4} \ldots, \theta_p = \theta_{0,p})$

.

.

.

Step 1.p: = draw $\theta_{1,p}$ from $f(\theta_p \mid x; \theta_1 = \theta_{1,1}, \ldots, \theta_{p-1} = \theta_{1,p-1})$

Step 2.1: = repeat steps 1 - p using $\vec{\theta_0} = \vec{\theta_1}$ from steps 1 - p.

.

.

.

Continue until convergence or until a preset number of iterations

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

You need the number of parameters, the jdf and the conditional densities

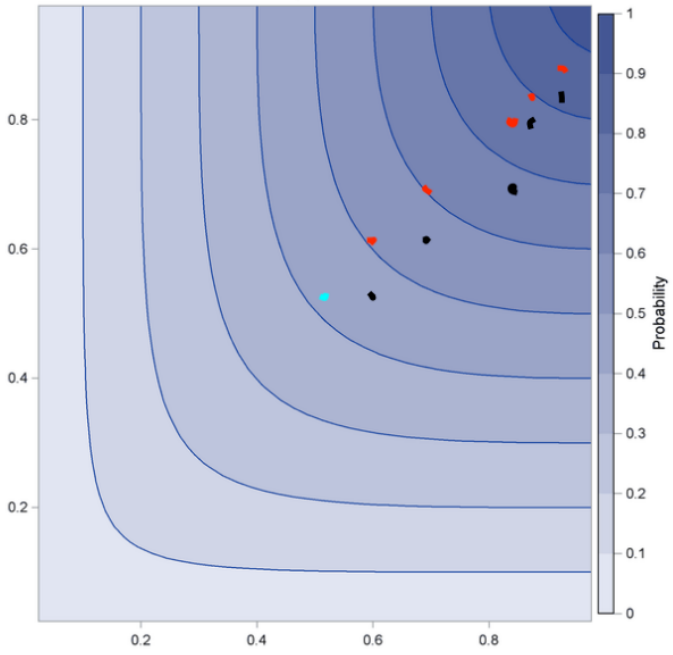
(c) [easy] Explain what burning of the Gibbs sample chain is and why it is necessary.

The burning is the initial phase where samples are discarded. This is necessary because the sampler starts from an arbitrarily chosen initial state, thus, samples drawn during the early iterations are heavily influenced by this initial state rather than the true distribution.

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

Thinning of the Gibbs Chain: For some T, $\vec{\theta}_t$ is virtually independent of $\vec{\theta}_{t+T}$, where T is called the "thinning value" i.e. it takes thirteen iterations to achieve independence. We do this because ultimately we want iid samples from the posterior.

(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.



The blue point denotes our initial point (0.5, 0.5). Each black point is our horizontal move and the corresponding black point is the vertical move. We move into the darker contours as they are the regions of higher probability.

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.

$\theta_1 \in (0, 1)$
$\theta_2 \in (0, \infty)$
$\theta_3 \in (0, 1)$

(b) [harder] Assume a flat prior $f(\theta_1, \theta_2, \theta_3) \propto 1$. Find the kernel of the posterior distribution $f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{x}, n_0, n_+)$ 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.

Because the prior is $\propto 1$, we know the kernel of the posterior will be $\propto$ the likelihood of the distribution.

Because we have a piecewise function given to us, it is best to think of the probability in two cases:

$$P(X = x \mid \theta_1, \theta_2, \theta_3) = \begin{cases} \theta_1 & x = 0 \\ (1 - \theta_1) \cdot p(x \mid \theta_2, \theta_3) & x \geq 1 \end{cases}$$

where $p(x \mid \theta_2, \theta_3) = \frac{\Gamma(x_i - 1 + \theta_2)}{(x_i - 1)!\Gamma(\theta_2)}(1-\theta_3)^{x_i - 1}\theta_3^{\theta_2}$

Denote counts of zero as:

$$\theta_1^{n_0}(1 - \theta_1)^{n_+}$$

Denote positive counts as:

$$\prod_{i=1}^{n_+} \frac{\Gamma(x_i - 1 + \theta_2)}{(x_i - 1)!\Gamma(\theta_2)}(1-\theta_3)^{x_i - 1}\theta_3^{\theta_2}$$

Posterior is:

$$\propto \theta_1^{n_0}(1 - \theta_1)^{n_+}(1 - \theta_3)^{\sum x_i - 1}\theta_3^{\theta_2 n_+} \cdot \prod_{i=1}^{n_+} \frac{\Gamma(x_i - 1 + \theta_2)}{(x_i - 1)!\Gamma(\theta_2)}$$

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

$$f(\theta_1 \mid \boldsymbol{x}, n_0, n_+, \theta_2, \theta_3) \propto \theta_1^{n_0 + 1 - 1}(1 - \theta_1)^{n_+ + 1 - 1}$$

$$\text{Beta}\big(n_0 + 1, n_+ + 1\big)$$

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

$$f(\theta_2 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3) \propto \theta_3^{\theta_2 n_+} \cdot \prod_{i=1}^{n_+} \frac{\Gamma(x_i - 1 + \theta_2)}{\Gamma(\theta_2)}$$

(e) [easy] Is the conditional distribution $f(\theta_2 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3)$ a brand name rv? Yes/no

No.

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

$$f(\theta_3 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_2) \propto (1-\theta_3)^{\sum x_i - 1 + 1 - 1} \theta_3^{\theta_2 n_+ + 1 - 1} \propto \text{Beta}\Big(\sum_{i=1}^{n_+}(x_i - 1) + 1, \theta_2 n_+ + 1\Big)$$

(g) [easy] Is it possible to get inference for this model using a Gibbs Sampler? Why or why not?

Because we don't have a closed form conditional probability function for θ_2, we likely cannot use the Gibb's sampler. The whole point of Gibb's is that for a whole posterior that is not in closed form, to use the closed form conditional distributions to create the model.

Problem 3

Consider the change point model

$$X_1, X_2, \ldots, X_{\theta_3} \overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2) \quad \text{independent of} \quad X_{\theta_3+1}, X_{\theta_3+2}, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$$

(a) [harder] What is the parameter space for all five parameters of interest?

Recall a change point model of two indpt random variables. In this context of two normals (ind), a time series is assumed to follow a normal distribution with one set of parameters (θ_1, σ_1^2) before a critical parameter, the "change point" (θ_3) which is the specific time point where the distribution switches from the first normal to the second, with a distinct set of parameters after that point (θ_2, σ_2^2). θ_3 is a discrete parameter because it indexes the position in the ordered sequence of observations, and it must come after the first observation and before the last observation (n)

$$\sigma_1^2 \in 0, \infty)$$
$$\sigma_2^2 \in 0, \infty)$$
$$\theta_1 \in \mathbb{R}$$
$$\theta_2 \in \mathbb{R}$$
$$\theta_3 \in \{1, 2, \ldots, n-1\}$$

(b) [harder] Assume a flat prior $\theta_1, \theta_2, \theta_3$ and Jeffrey's prior for σ_1^2, σ_2^2 which are assumed a priori independent of one another. Find the kernel of the posterior distribution.

$f(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \theta_3 \mid \vec{x}) \propto f(\vec{x} \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \theta_3) \cdot f(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \theta_3)$

We will use Jeffrie's prior for variance, and the Laplace prior for the mean. Recall that to find the posterior of a change point model, we are multiplying two processes together.

$$\prod_{i=1}^{\theta_3} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2sigsq_1}(x_i-\theta_1)^2} \cdot \prod_{i=\theta_3+1}^{n} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2sigsq_2}(x_i-\theta_2)^2}$$

$$\propto \left(\sigma_1^2\right)^{-\frac{\theta_3}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-\frac{n-\theta_3}{2}} e^{-\frac{1}{2\sigma_2^2}\sum_{i=\theta_3+1}^{n}(x_i-\theta_2)^2}$$

Multiply by Jeffries and Laplace Priors:

$$f(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \theta_3) \propto 1 \cdot 1 \cdot 1 \cdot \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}$$

$$\Rightarrow \propto \left(\sigma_1^2\right)^{-\frac{\theta_3}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-\frac{n-\theta_3}{2}} e^{-\frac{1}{2\sigma_2^2}\sum_{i=\theta_3+1}^{n}(x_i-\theta_2)^2} \cdot \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}$$

$$\Rightarrow \propto \left(\sigma_1^2\right)^{-(\frac{\theta_3}{2})-1} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-(\frac{n-\theta_3}{2})-1} e^{-\frac{1}{2\sigma_2^2}\sum_{i=\theta_3+1}^{n}(x_i-\theta_2)^2}$$

(c) [harder] Find the kernels of all five conditional distributions. If they are proportional to a known distribution, name it.

Starting with θ_1, we use the portion of the posterior that is proportional to θ_1:

$$\propto e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2}$$

Foil and complete the square:

$$\propto e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i^2-2x_i\theta_1+\theta_1^2)}$$

$$\propto e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(\sum_{i=1}^{\theta_3}x_i^2-\sum_{i=1}^{\theta_3}2x_i\theta_1+\theta_3\theta_1^2)}$$

$$\propto e^{\frac{\sum_{i=1}^{\theta_3}x_i}{\sigma_1^2}\cdot\theta_1-\frac{\theta_2}{2\sigma_1^2}\cdot\theta_1^2)}$$

Recall from math 340 that this resembles $e^{a\theta-b\theta^2}$ which is $\propto \mathcal{N}\left(\frac{a}{2b}, \frac{1}{2b}\right)$; where a $= \frac{\sum_{i=1}^{\theta_3}x_i}{\sigma_1^2}$ and b $= \frac{\theta_2}{2\sigma_1^2}$

$$2b = \frac{\theta_3}{\sigma_1^2}$$

and

$$\frac{a}{2b} = \frac{\sum_{i=1}^{\theta_3} x_i}{\cancel{\sigma_1^2}} \cdot \frac{\cancel{\sigma_1^2}}{\theta_3} = \frac{\sum_{i=1}^{\theta_3} x_i}{\theta_3}$$

Therefore,

$$f(\theta_1 \mid -) \propto \mathcal{N}\left(\bar{x}_{1:\theta_3}, \frac{\sigma_1^2}{\theta_3}\right)$$

Similarly for θ_2,

$$f(\theta_2 \mid -) \propto \mathcal{N}\left(\bar{x}_{\theta_3+1:n}, \frac{\sigma_2^2}{n-\theta_3}\right)$$

For σ_1^2 and σ_2^2:

$$f(\sigma_1^2 \mid -) \propto \left(\sigma_1^2\right)^{-(\frac{\theta_3}{2})-1} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2}$$

Recall Inverse Gamma:

$$\text{InvGamma}(\alpha, \beta) \coloneqq \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$$

Therefore, we can see if we let $\alpha = \frac{\theta_3}{2}$ and $\beta = \frac{\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2}{2}$, then:

$$f(\sigma_1^2 \mid -) \propto \text{InvGamma}\left(\frac{\theta_3}{2}, \frac{\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2}{2}\right)$$

Similarly,

$$f(\sigma_2^2 \mid -) \propto \text{InvGamma}\left(\frac{n-\theta_3}{2}, \frac{\sum_{i=\theta_3}^{n}(x_i-\theta_2)^2}{2}\right)$$

for θ_3:

$$\propto (\sigma_2^2)^{-\frac{\theta_3}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-(\frac{n-\theta_3}{2})-1} e^{-\frac{1}{2\sigma_2^2}\sum_{i=\theta_3+1}^{n}(x_i-\theta_2)^2}$$

(d) [harder] Find the conditional PMF of θ_3.

$$P(\theta_3 \mid -) = \frac{(\sigma_2^2)^{-\frac{\theta_3}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\theta_3}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-(\frac{n-\theta_3}{2})-1} e^{-\frac{1}{2\sigma_2^2}\sum_{i=\theta_3+1}^{n}(x_i-\theta_2)^2}}{\sum_{k=1}^{n-1}(\sigma_2^2)^{-\frac{k}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{k}(x_i-\theta_1)^2} \left(\sigma_2^2\right)^{-(\frac{n-k}{2})-1} e^{-\frac{1}{2\sigma_2^2}\sum_{i=k+1}^{n}(x_i-\theta_2)^2}}$$

(e) [easy] Is it possible to get inference for this model using a Gibbs Sampler? Why or why not?

Yes, because we have all the needed conditional densities and can grid sample for θ_3, which will converge becasue the resolution is picked for you

Problem 4

Consider the discrete mixture model:

$$X_1, \ldots, X_n \overset{iid}{\sim} \begin{cases} \text{Poisson}\,(\theta_0) & \text{w.p. } \rho \\ \text{Poisson}\,(\theta_1) & \text{w.p. } 1-\rho \end{cases}$$

(a) [harder] What is the parameter space for all three parameters of interest?

$$\rho \in (0,1)$$
$$\theta_0 \in (0,\infty)$$
$$\theta_1 \in (0,\infty)$$

(b) [harder] Assume a flat prior one all parameters. Find the kernel of the posterior distribution.

Recall the formula for discrete mixture distribution:

$$p_Y(y) = \sum_{all\ x} p_x(x) p_{y|x}(y,x)$$

Therefore we have

$$\Rightarrow \left[\rho\left(\frac{\theta_0^x e^{-\theta_0}}{x!}\right)\right] + \left[1-\rho\left(\frac{\theta_1^x e^{-\theta_1}}{x!}\right)\right]$$

The posterior would be the likelihood:

$$p(\theta_0,\theta_1,\rho \mid x) = \prod_{i=1}^n \left[\rho\left(\frac{\theta_0^x e^{-\theta_0}}{x!}\right)\right] + \left[1-\rho\left(\frac{\theta_1^x e^{-\theta_1}}{x!}\right)\right] \propto \prod_{i=1}^n \left[\rho\left(\theta_0^x e^{-\theta_0}\right)\right] + \left[1-\rho\left(\theta_1^x e^{-\theta_1}\right)\right]$$

(c) [easy] Is this proportional to any known distribution?

No.

(d) [harder] Is it possible to make a Gibbs Sampler to get inference here? Why or why not.

Not really, there is no closed form of the posterior, or of the conditional distributions.

(e) [harder] Let's use data augmentation. Add $I_1, \ldots, I_n$ as parameters whose parameter space is $\{0,1\}$ where $I_i = 1$ denotes that the ith observation has membership in the Poisson (θ_0) distribution and $I_i = 0$ denotes that the ith observation has membership in the Poisson (θ_1) distribution. Now find the kernel of the posterior distribution.

We use the data augmentation to get a manageable posterior that we can sample from via Gibb's:

$$\Rightarrow p(\theta_0, \theta_1, \rho, I_1, \ldots, I_n \mid \vec{x}) \propto p(\vec{x} \mid \theta_0, \theta_1, \rho, I_1, \ldots, I_n) \cdot p(\theta_0, \theta_1, \rho, I_1, \ldots, I_n)$$

Where we now have latent indicators $I_1, \ldots I_n$ with

$$I_i = \begin{cases} 1, & \text{if } X_i \sim \text{Poisson}\,(\theta_0) \\ 0, & \text{if } X_i \sim \text{Poisson}\,(\theta_1) \end{cases}$$

Assume the previous priors and a Laplace prior for $I_1, \ldots, I_n$. We now have:

$$\propto \prod_{i=1}^{n} \left[\rho\left(\theta_0^{x_i} e^{-\theta_0}\right)\right]^{I_i} \left[1 - \rho\left(\theta_1^{x_i} e^{-\theta_1}\right)\right]^{1-I_i}$$

Define $n_0 = \sum_{i=1}^{n} I_i$ and $n_1 = \sum_{i=1}^{n} 1 - I_i$

We now have:

$$\propto \rho^{n_0}(1-\rho)^{n_1} e^{-\theta_0 n_0} e^{-\theta_1 n_1} \theta_0^{\sum I_i x_i} \theta_1^{\sum (1-I_i) x_i}$$

(f) [harder] Find the kernels of all four conditional distributions (for $\theta_0, \theta_1, \rho, I_i$). If they are proportional to a known distribution, name it.

$$p(\theta_0 \mid -) \propto e^{-\theta_0 n_0} \theta_0^{\sum I_i x_i}$$

Recall $\text{Gamma}\,(\alpha,\, \beta) \coloneqq \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ We can see:

$$p(\theta_0 \mid -) \propto \text{Gamma}\left(\left(\sum_{i=1}^{n} I_i x_i\right) + 1,\, n_0\right)$$

And we can deduce that:

$$p(\theta_1 \mid -) \propto \text{Gamma}\left(\left(\sum_{i=1}^{n} (1 - I_i) x_i\right) + 1,\, n_1\right)$$

For ρ:

$$p(\rho \mid -) \propto \rho^{n_0}(1-\rho)^{n_1}$$

Recall the Beta distribtuion $\coloneqq \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

Subbing in values for *alpha* and *beta* we have the brand name conditional distribution,

$$p(\rho \mid -) \propto \text{Beta}\,(n_0 + 1,\, n_1 + 1)$$

Now for I

For each observation x_i the contribution to the joint kernel is $\left[\rho\left(\theta_0^{x_i}e^{-\theta_0}\right)\right]^{I_i}\left[1-\rho\left(\theta_1^{x_i}e^{-\theta_1}\right)\right]^{1-I_i}$.

For $I_i = 1$

$$P(I_i = 1 \mid -) \propto \rho e^{-\theta_0}\theta_0^{x_i}$$

for $I_i = 0$

$$P(I_i = 1 \mid -) \propto (1-\rho)e^{-\theta_1}\theta_1^{x_i}$$

Each I_i is a Bernouilli RV with success probability

$$p_i = \frac{\rho e^{-\theta_0}\theta_0^{x_i}}{\rho e^{-\theta_0}\theta_0^{x_i} + (1-\rho)e^{-\theta_1}\theta_1^{x_i}}$$

Therefore,

$$I_i \mid - \sim \text{Bernouilli}\left(\frac{\rho e^{-\theta_0}\theta_0^{x_i}}{\rho e^{-\theta_0}\theta_0^{x_i} + (1-\rho)e^{-\theta_1}\theta_1^{x_i}}\right)$$

(g) [easy] Is it possible to get inference for this model using a Gibbs Sampler after data augmentation? Why or why not?

Yes, because we now have known conditional distributions; however, it will be much slower and if you pick the wrong bounds then you mess up the sample

Problem 5

These are general questions about Permutation Testing.

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

$$H_0 : DGP_1 = DGP_2$$
$$H_a : DGP_1 \neq DGP_2$$

(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.

Recall, in Fisher's permutation test, we are randomly permuting the indices of the original two populations

Therefore, the number of "permutations" are the number of possible ways to assign n_1 observations from population 1 and n_2 observations from population 2 to a combined sample of size $n_2 + n_2$

This is given by $\binom{n_1+n_2}{n_1}$

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

1: $\hat{\hat{\theta}}_b = \bar{x}_{b,2} - \bar{x}_{b,1}$

2: $\hat{\hat{\theta}}_b = s^2_{b_1} - s^2_{b_2}$

3: $\hat{\hat{\theta}}_b = d_{b_1} - d_{b_2}$, where d denotes the sup difference in ECDF's

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

First you would calculate a test stat (observed). The under the null, shuffle the group labels among all $n_1 + n_2$ observations. For each permutation, reassign observations to two groups of sizes n_1 and n_2 and then calculate the test statistic for this (perm). Repeat many times or for all possible permutations if possible to build the distribution under the null.

You then determine the extremity of the permuted test stats tot he observed statistic. The p-value is then the proportion of permuted test statistics that are at least as extreme as the observed statistic

Ex: $p_{\text{val}} = min\{2P(\hat{\theta}_b > \hat{\hat{\theta}}), 2P(\hat{\theta}_b < \hat{\hat{\theta}})\}$

Problem 6

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.

The bootstrap is a resampling method that uses your observed data to approximate the sampling distribution of your estimator.
Let $\phi = g(\theta)$. Define a bootstrap sample as a sample of n values with replacement from a set of size n. This will mean there are some duplicates as well as some missing samples. (e.g. Consider the dataset $x_1, \ldots, x_n$. A bootstrap sample will be $x_{b_1}, \ldots x_{b_n}$, and you will likely have 2/3 of the unique original values and 1/3 missing. This can duplicate many times.)

Then consider a large number (B) of bootstrap samples from the dataset. For each sample b, compute an estimate $\hat{\theta}^{*(b)} = w(x^*_1, \ldots x^*_n)$ which will give you a collection of bootstrap replicates. The empirical distribution of the $\hat{\theta}^{*(b)\prime}s$ serves as an approximation to the sampling distribution of $\hat{\hat{\theta}}$. Note that this procedure lets you approximate the variability of your estimator without relying on strong parametric assumptions about the underlying distribution.

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

Bootstrap method is useful when

- The sampling distribution is complex or unknown
- Parametric assumptions are unavailable
- You have a dataset with small to moderate sample size but complex estimator
- Standard inferential methods are not easily applicable

(c) [difficult] Explain in what situations the bootstrap fails. Read online about this.

Bootstrap can give misleading results when the statistic is a non-smooth function of the data such as a trimmed mean or other "irregular" estimators. It also can fail when the sampling distribution has discontinuities. Also, if the parameter is "on the edge" i.e. it lies on the boundary of the parameter space, the bootstrap may fail. When the distribution is heavy-tailed or the variance is infinite is another instance of when bootstrap can fail. Also, bootstrap approximation will be poor if being used on a small sample size.

Problem 7

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

Recall to find the MLE, we first take the likelihood, then the log likelihood and then the score function and set it equal to zero.

$$\mathcal{L}(k, \lambda) = \Pi_{i=1}^n k\lambda^k y_i^{k-1} e^{-\lambda^k y_i^k} = k^n \lambda^{nk} \cdot \sum_{i=1}^n y_i^{k-1} e^{-\sum_{i=1}^n y_i^k \lambda^k}$$

$$\ell(k, \lambda) = \ln\left(k^n \lambda^{nk} \cdot \sum_{i=1}^n y_i^{k-1} e^{-\sum_{i=1}^n y_i^k \lambda^k} \right)$$

$$= n \ln(k) + nk \ln(\lambda) + (k-1) \sum_{i=1}^n \ln(y_i) - \sum_{i=1}^n y_i^k \lambda^k$$

$$\frac{d}{dy} = \frac{nk}{\lambda} - k\lambda^{k-1}\sum_{i=1}^{n} y_i^k \overset{set}{=} 0$$

$$\hat{\hat{\lambda}}^{\text{MLE}} = \left(\frac{n}{\sum_{i=1}^{n} y_i^k}\right)^{\frac{1}{k}}$$

For k:

$$\frac{d}{dk}\frac{n}{k} + n\ln(\lambda) + \sum_{i=1}^{n}\ln(y_i) - \sum_{i=1}^{n}\lambda^k y_i^k \ln(\lambda y_i)$$

After setting to 0, we do not get a closed form; we have:

$$\frac{1}{k} = \frac{\sum_{i=1}^{n} y_i^k \ln(y_i)}{\sum_{i=1}^{n} y_i^k} - \frac{1}{n}\sum_{i=1}^{n}\ln(y_i)$$

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

$$\mathcal{L}(k, \lambda; \vec{y}, \vec{c}) = \Pi_{i:c_i=1} f(y_i; k, \lambda) \Pi_{i:c_i=0} P(Y > y_i)$$

$$\Rightarrow \Pi_{i:c_i=1}^{n} \lambda^k k y_i^{k-1} e^{-\lambda^k y_i^k} \Pi_{i:c_i=0}^{n} e^{-\lambda^k \sum_{i:c_i=0}^{n} y_i^k}$$

let $n_1 = \sum \mathbb{1}_{c_i=1}$ and let $n_0 = \sum \mathbb{1}_{c_i=0}$

$$\Rightarrow (\lambda^k k)^{n_1} \Pi_{i:c_i=1}^{n} y_i^{k-1} e^{-\lambda^k \sum_{i:c_i=1}^{n} y_i^k} e^{-\lambda^k \sum_{i:c_i=0}^{n} y_i^k}$$

$$\ell(k, \lambda; \vec{y}, \vec{c}) = n_1 k \ln(\lambda) + n_1 \ln(k) + (k-1)\sum \ln(y_i) - \lambda^k \sum y_i^k$$

Problem 8

These are questions about nonparametric survival inference.

(a) [easy] Show that the empirical survival function is equal to the product limit estimator form with no censoring. Make sure to define what your notation means.

The product estimator with no censoring (KM estimator):

$$\hat{\hat{S}}(t_k) = \Pi_{i=1}^{k}\left(1 - \frac{1}{n-(i-1)}\right)$$

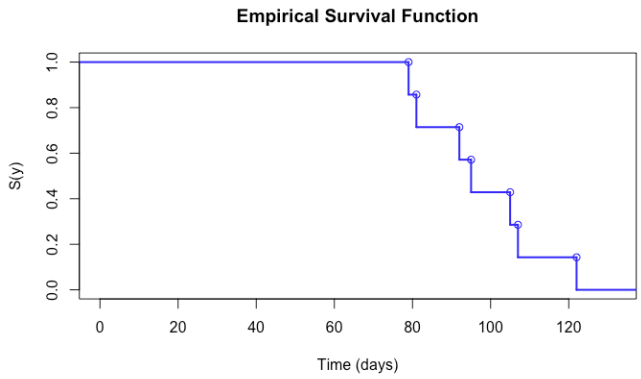
$$\Rightarrow \frac{n-k}{n}$$

Empirical Survival:

$$\frac{1}{n}\sum_{\mathbb{1}=1}^{n} y_i > y$$

$$\Rightarrow \frac{n-k}{n}$$

(b) [easy] Consider the dataset $y = \{79, 81, 92, 95, 105, 107, 122\}$ measured in days. Draw the estimate of $S(y)$.



(c) [harder] Let your parameter of interest θ be survival past 106 days. Compute a 95% CI for θ.

$$\text{CI}_{\hat{\hat{S}},95\%} = \left[\hat{\hat{S}} \pm 1.96\sqrt{\frac{\hat{\hat{S}}(106)(1-\hat{\hat{S}}(106))}{7}}\right] = [-0.048, 0.620]$$

(d) [harder] Test $H_a : \theta > 0.5$.

$$\text{RET}_{\theta,5\%} = \left[0, \theta_0 + 1.96\sqrt{\frac{\theta_0(1-\theta_0)}{7}}\right] = [0, 0.870]$$

$\frac{2}{7} \in$ the retainment region, therefore, retain the null

(e) [easy] Explain how you would use the bootstrap to find a CI for the median. Explain why the bootstrap won't be so accurate in this example.

(1) Sample with replacement many times from the learning set

(2) Compute the median for each sample

(3) Compose the collection $\{m_1, m_2, \ldots, m_B\}$ which is the bootstrap approximation

(4) Construct a confidence interval by finding a 2.5 percent quantile and a 97.5 percent quantile.

In our sample, n is very small and the median is not a smooth test statistic, so bootstrap may not be very accurate.

(f) [harder] Rederive the Kaplan-Meier estimator for the survival function.

Survival can be written as a product of the conditional survival probabilities over intervals defined by event times

$$S(t) = \mathcal{P}(T > t) = \Pi^n_{t_{(j)} \leq t} \mathcal{P}(T > t_{(j)} \mid T > t_j)$$

At each event time $t_{(j)}$, an empriical estimate for the conditional probability of surviving past $t_{(j)}$ is:

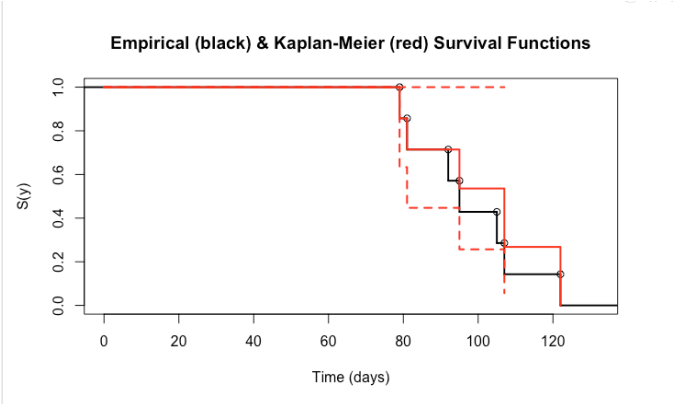
$$\mathcal{P}(T > t_{(j)} \mid T > t_j) \approx 1 - \frac{d_j}{n_j}$$

Where $\frac{d_j}{n_j}$ is the observed probability at time t_j of those at risk of death.
Plugging back in:

$$\Pi^n_{t_j \leq t}(1 - \frac{d_j}{n_j})$$

When there is no censoring, $n_j = n - j + 1$ and every observation will be an event so $d_j = 1$

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



(h) [harder] Explain how you would use the bootstrap to find a CI for the median.

(1) Sample in pairs with replacement many times from the learning set (y_i, δ_i)

(2) Compute the median for each sample

(3) Compose the collection $\{m_1, m_2, \ldots, m_B\}$ which is the bootstrap approximation

(4) Construct a confidence interval by finding a 2.5 percent quantile and a 97.5 percent quantile.

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

$$H_a : \text{MED}[Y_1] \neq \text{MED}[Y_2]$$
$$H_0 : \text{MED}[Y_1] = \text{MED}[Y_2]$$

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

$$\hat{\theta} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \xrightarrow{d} \chi^2_1$$