Assignment #3 STA355H1S

due Friday March 20, 2020

Instructions: Solutions to problems 1 and 2 are to be submitted on Quercus (PDF files only) – the deadline is 11:59pm on March 20. You are strongly encouraged to do problems 3 through 6 but these are **not** to be submitted for grading.

Problems to hand in:

1. Suppose that $X_1, \cdots, X_n$ are independent Gamma random variables with pdf

$$f(x; \lambda, \alpha) = \frac{\lambda^\alpha x^{\alpha-1} \exp(-\lambda x)}{\Gamma(\alpha)} \quad \text{for } x > 0$$

where $\lambda > 0$ and $\alpha > 0$ are unknown parameters. Given $X_1 = x_1, \cdots, X_n = x_n$, the likelihood function is

$$\mathcal{L}(\lambda, \alpha) = \frac{\lambda^{n\alpha} \left\{ \prod_{i=1}^n x_i^{\alpha-1} \right\} \exp\left(-\lambda \sum_{i=1}^n x_i \right)}{[\Gamma(\alpha)]^n}$$

(a) Assume the following prior distribution for (λ, α):

$$\pi(\lambda, \alpha) = \frac{1}{10000} \exp(-\lambda/100) \exp(-\alpha/100) \quad \text{for } \lambda, \alpha > 0$$

Given $X_1 = x_1, \cdots, X_n = x_n$, show that the posterior density of α is

$$\pi(\alpha | x_1, \cdots, x_n) = K(x_1, \cdots, x_n) \frac{\Gamma(n\alpha + 1)}{[\Gamma(\alpha)]^n} \exp\left(\alpha \sum_{i=1}^n \ln(x_i) - \frac{\alpha}{100} \right) \left(\frac{1}{100} + \sum_{i=1}^n x_i \right)^{-(n\alpha+1)}$$

(b) Data on intervals (in hours) between failures of air conditioning units on ten Boeing aircraft are given (and analyzed) in Example T of Cox & Snell[1] (1981). These data (199 observations) are provided in a file on Quercus. Using the prior for (λ, α) in part (a), compute the posterior distribution for α.

Note: To compute the posterior density, you will need to compute the normalizing constant – on Quercus, I will give some suggestions on how to do this. A simple estimate of α is $\widehat{\alpha} = \bar{x}^2/s^2$ where $\bar{x}$ and s^2 are the sample mean and variance, respectively. We would expect the posterior to be concentrated around this estimate.

[1] Cox, D.R. and Snell, E.J. (1981) *Applied Statistics: Principles and Examples.* Chapman and Hall, New York.

(c) [Bonus] We may be interested in "testing" whether an Exponential model is an appropriate model for the data. One approach to doing this is to put a prior probability θ on the Exponential model (i.e. a Gamma model with $\alpha = 1$) and prior probability $1-\theta$ on the more general Gamma model. This leads to a prior distribution on (λ, α) having a point mass θ (which is a hyperparameter) at $\alpha = 1$ so that

$$\pi(\lambda, 1) = \frac{1}{100}\theta \exp(-\lambda/100) \quad \text{for } \lambda > 0$$

and

$$\pi(\lambda, \alpha) = \frac{1}{10000}(1-\theta)\exp(-\lambda/100)\exp(-\alpha/100) \quad \text{otherwise}$$

(This type of prior is an example of a "spike-and-slab" prior used in Bayesian model selection.) We then have

$$P(\alpha = 1|x_1, \cdots, x_n) = \frac{\displaystyle\int_0^\infty \pi(\lambda, 1)\mathcal{L}(\lambda, 1)\, d\lambda}{\displaystyle\int_0^\infty \pi(\lambda, 1)\mathcal{L}(\lambda, 1)\, d\lambda + \int_0^\infty \int_0^\infty \pi(\lambda, \alpha)\mathcal{L}(\lambda, \alpha)\, d\lambda\, d\alpha}.$$

For $\theta = 0.1, 0.2, 0.3, \cdots, 0.9$, evaluate $P(\alpha = 1|x_1, \cdots, x_n)$. (This is easier than it looks as much of the work has been done in part (b).)

2. Suppose that F is a continuous distribution with density f. The mode of the distribution here is defined to be the global maximizer of the density function. (In other applications, it is useful to think of modes as local maxima of the density function.)

In some applications (for example, when the data are "contaminated" with outliers), the centre of the distribution is better described by the mode than by the mean or median; however, unlike the mean and median, the mode turns out to be a difficult parameter to estimate. There is a document on Quercus that discusses a few of the various methods for estimating the mode.

The τ-shorth is the shortest interval that contains at least a fraction τ of the observations $X_1, \cdots, X_n$; it will have the form $[X_{(a)}, X_{(b)}]$ where $b - a + 1 \geq \tau n$. A number of mode estimators are based on the observations lying in $[X_{(a)}, X_{(b)}]$; for example, we can the sample mean of these observation or the sample median. (Note that taking the sample mean of the observations in $[X_{(a)}, X_{(b)}]$ is like a trimmed mean although the trimming here is typically asymmetrical.) In this problem, we will consider estimating the mode using the midpoint of the interval, that is, $\widehat{\mu} = (X_{(a)} + X_{(b)})/2$. This estimator is called a Venter estimator. An R function to compute this estimator for a given value of τ is available on Quercus in a file `venter.txt`.

(a) For the Hidalgo stamp thickness data considered in Assignment #2, compute Venter estimates for various values of τ. How small does τ need to be in order that the estimate

"makes sense"? (Recall from Assignment #2 that the density seemed to have a number of local maxima but one clear global maximum.)

(b) Suppose that the underlying density f is asymmetric and unimodal. The choice of τ for the Venter estimator involves a bias-variance tradeoff: If τ is small then we should expect the estimator to have a small bias but larger variance with the bias increasing and the variance decreasing as τ increases.

Suppose that $X_1, \cdots, X_n$ are independent Gamma random variables with density

$$f(x;\alpha) = \frac{x^{\alpha-1}\exp(-x)}{\Gamma(\alpha)}$$

where we will assume that $\alpha > 1$; in this case, the mode is $\alpha - 1$.

Using Monte Carlo simulation, estimate the MSE of the Venter estimator for $\tau = 0.5$ and $\tau = 0.1$, sample sizes $n = 100$ and $n = 1000$, and shape parameters $\alpha = 2$ and $\alpha = 10$. (8 simulations in total.) For $\alpha = 2$ and $\alpha = 10$, which Venter estimator seems to be better (on the basis of MSE)?

(c) The density of Venter estimator $\widehat{\mu}$ under the Gamma model in part (b) can be estimated using the fact that $\widehat{\mu}/(X_1 + \cdots + X_n)$ and $X_1 + \cdots + X_n$ are independent. (This fact follows from Basu's Theorem. which you may encounter in STA452/453.)

We can exploit this independence as follows: Define $T = X_1 + \cdots + X_n$; T has a Gamma distribution with shape parameter n. Then

$$\begin{aligned}
P(\widehat{\mu} \le x) &= P\left(\frac{\widehat{\mu}}{T} \le \frac{x}{T}\right) \\
&= \int_0^\infty P\left(\frac{\widehat{\mu}}{T} \le \frac{x}{T}\Big|T = t\right)\frac{t^{n\alpha-1}\exp(-t)}{\Gamma(n\alpha)}\,dt \\
&= \int_0^\infty P\left(\frac{\widehat{\mu}}{T} \le \frac{x}{t}\Big|T = t\right)\frac{t^{n\alpha-1}\exp(-t)}{\Gamma(n\alpha)}\,dt \\
&= \int_0^\infty P\left(\frac{\widehat{\mu}}{T} \le \frac{x}{t}\right)\frac{t^{n\alpha-1}\exp(-t)}{\Gamma(n\alpha)}\,dt
\end{aligned}$$

Thus given $(\widehat{\mu}_1, T_1), \cdots, (\widehat{\mu}_N, T_N)$, we can estimate

$$P\left(\frac{\widehat{\mu}}{T} \le \frac{x}{t}\right)$$

by

$$\frac{1}{N}\sum_{i=1}^N I\left(\frac{\widehat{\mu}_i}{T_i} \le \frac{x}{t}\right) = \frac{1}{N}\sum_{i=1}^N I\left(t \le \frac{T_i}{\widehat{\mu}_i}x\right)$$

and so we can estimate $P(\widehat{\mu} \le x)$ by

$$\frac{1}{N}\sum_{i=1}^N \int_0^{xT_i/\widehat{\mu}_i} \frac{t^{n\alpha-1}\exp(-t)}{\Gamma(n\alpha)}\,dt$$

and differentiating, we obtain the density estimate

$$\widehat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\widehat{\mu}_i}\left(\frac{T_i}{\widehat{\mu}_i}x\right)^{n\alpha-1} \frac{\exp(-xT_i/\widehat{\mu}_i)}{\Gamma(n\alpha)}$$

Using the simulation data in part (b) for $n = 100$, $\alpha = 2$ and $\tau = 0.5$, compute $\widehat{f}(x)$.

Supplemental problems (not to hand in):

3. Suppose that $X_1, \cdots, X_n$ are independent random variables with density or mass function $f(x;\theta)$ and suppose that we estimate θ using the maximum likelihood estimator $\widehat{\theta}$; we estimate its standard error using the observed Fisher information estimator

$$\widehat{\text{se}}(\widehat{\theta}) = \left\{-\sum_{i=1}^{n} \ell''(X_i;\widehat{\theta})\right\}^{-1/2}$$

where $\ell'(x;\theta), \ell''(x;\theta)$ are the first two partial derivatives of $\ln f(x;\theta)$ with respect to θ. Alternatively, we could use the jackknife to estimate the standard error of $\widehat{\theta}$; if our model is correct then we would expect (hope) that the two estimates are similar. In order to investigate this, we need to be able to get a good approximation to the "leave-one-out" estimators $\{\widehat{\theta}_{-i}\}$.
(a) Show that $\widehat{\theta}_{-i}$ satisfies the equation

$$\ell'(X_i;\widehat{\theta}_{-i}) = \sum_{j=1}^{n} \ell'(X_j;\widehat{\theta}_{-i}).$$

(b) Expand the right hand side in (a), in a Taylor series around $\widehat{\theta}$ to show that

$$\widehat{\theta}_{-i} - \widehat{\theta} \approx \frac{\ell'(X_i;\widehat{\theta}_{-i})}{\sum_{j=1}^{n} \ell''(X_j;\widehat{\theta})} \approx \frac{\ell'(X_i;\widehat{\theta})}{\sum_{j=1}^{n} \ell''(X_j;\widehat{\theta})}$$

and so

$$\widehat{\theta}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{-i} \approx \widehat{\theta}.$$

(You should try to think about the magnitude of the approximation error but a rigorous proof is not required.)
(c) Use the results of part (b) to derive an approximation for the jackknife estimator of the standard error. Comment on the differences between the two estimators - in particular, why is there a difference? (Hint: What type of model – parametric or non-parametric – are we assuming for the two standard error estimators?)

(d) For the air conditioning data considered in Assignment #1, compute the two standard error estimates for the parameter λ in the Exponential model ($f(x;\lambda) = \lambda \exp(-\lambda x)$ for $x \geq 0$). Do these two estimates tell you anything about how well the Exponential model fits the data?

4. Suppose that $X_1, \cdots, X_n$ are independent continuous random variables with density $f(x;\theta)$ where θ is real-valued. We are often not able to observe the X_i's exactly rather only if they belong to some region B_k ($k = 1, \cdots, m$); an example of this is *interval censoring* in survival analysis where we are unable to observe an exact failure time but know that the failure occurs in some finite time interval. Intuitively, we should be able to estimate θ more efficiently with the actual values of $\{X_i\}$; in this problem, we will show that this is true (at least) for MLEs.

Assume that $B_1, \cdots, B_m$ are disjoint sets such that $P(X_i \in \cup_{k=1}^m B_k) = 1$. Define independent discrete random variables $Y_1, \cdots, Y_n$ where $Y_i = k$ if $X_i \in B_k$; the probability mass function of Y_i is

$$p(k;\theta) = P_\theta(X_i \in B_k) = \int_{x \in B_k} f(x;\theta)\, dx \quad \text{for } k = 1, \cdots, m.$$

Also define

$$\begin{aligned} I_X(\theta) &= \mathrm{Var}_\theta \left[\frac{\partial}{\partial \theta} \ln f(X_i;\theta) \right] \\ &= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta} \ln f(x;\theta) \right]^2 f(x;\theta)\, dx \\ I_Y(\theta) &= \mathrm{Var}_\theta \left[\frac{\partial}{\partial \theta} \ln p(Y_i;\theta) \right] \\ &= \sum_{k=1}^{m} \left[\frac{\partial}{\partial \theta} \ln p(k;\theta) \right]^2 p(k;\theta). \end{aligned}$$

Under the standard MLE regularly conditions, the MLE of θ based on $X_1, \cdots, X_n$ will have variance approximately $1/\{nI_X(\theta)\}$ while the MLE based on $Y_1, \cdots, Y_n$ will have variance approximately $1/\{nI_Y(\theta)\}$.

(a) Assume the usual regularity conditions for $f(x;\theta)$, in particular, that $f(x;\theta)$ can be differentiated with respect to θ inside integral signs with impunity! Show that $I_X(\theta) \geq I_Y(\theta)$ and indicate under what conditions there will be strict inequality.

Hints: (i) $f(x;\theta)/p(k;\theta)$ is a density function on B_k.
(ii) For any function g,

$$\int_{-\infty}^{\infty} g(x) f(x;\theta)\, dx = \sum_{k=1}^{m} p(k;\theta) \int_{x \in B_k} g(x) \frac{f(x;\theta)}{p(k;\theta)}\, dx.$$

(iii) For any random variable U, $E(U^2) \geq [E(U)]^2$ with strict inequality unless U is constant.

(b) Under what conditions on $B_1, \cdots, B_m$ will $I_X(\theta) \approx I_Y(\theta)$?

5. In seismology, the Gutenberg-Richter law states that, in a given region, the number of earthquakes N greater than a certain magnitude m satisfies the relationship

$$\log_{10}(N) = a - b \times m$$

for some constants a and b; the parameter b is called the b-value and characterizes the seismic activity in a region. The Gutenberg-Richter law can be used to predict the probability of large earthquakes although this is a very crude instrument. On Quercus, there is a file containing earthquakes magnitudes for 433 earthquakes in California of magnitude (rounded to the nearest tenth) of 5.0 and greater from 1932–1992.
(a) If we have earthquakes of (exact) magnitudes $M_1, \cdots, M_n$ greater than some known m_0, the Gutenberg-Richter law suggests that $M_1, \cdots, M_n$ can be modeled as independent random variables with a shifted Exponential density

$$f(x; \beta) = \beta \exp(-\beta(x - m_0)) \quad \text{for } x \geq m_0.$$

where $\beta = b \times \ln(10)$ and m_0 is assumed known. However, if the magnitudes are rounded to the nearest δ then they are effectively discrete random variables taking values $x_k = m_0 + \delta/2 + k\delta$ for $k = 0, 1, 2, \cdots$ with probability mass function

$$\begin{aligned} p(x_k; \beta) &= P(m_0 + k\delta \leq M < m_0 + (k+1)\delta) \\ &= \exp(-\beta k\delta) - \exp(-\beta(k+1)\delta) \\ &= \exp(-\beta(x_k - m_0 - \delta/2)) \{1 - \exp(-\beta\delta)\} \quad \text{for } k = 0, 1, 2, \cdots. \end{aligned}$$

If $X_1, \cdots, X_n$ are the rounded magnitudes, find the MLE of β. (There is a closed-form expression for the MLE in terms of the sample mean of $X_1, \cdots, X_n$.)
(b) Compute the MLE of β for the earthquake data (using $m_0 = 4.95$ and $\delta = 0.1$) as well as estimates of its standard error using (i) the Fisher information and (ii) the jackknife. Use these to construct approximate 95% confidence intervals for β. How similar are the intervals?

6. In genetics, the Hardy-Weinberg equilibrium model characterizes the distributions of genotype frequencies in populations that are not evolving and is a fundamental model of population genetics. In particular, for a genetic locus with two alleles A and a, the frequencies of the genotypes AA, Aa, and aa are

$$P_\theta(\text{genotype} = AA) = \theta^2, \; P_\theta(\text{genotype} = Aa) = 2\theta(1-\theta), \text{ and } P_\theta(\text{genotype} = aa) = (1-\theta)^2$$

where θ, the frequency of the allele A in the population, is unknown.
In a sample of n individuals, suppose we observe $X_{AA} = x_1$, $X_{Aa} = x_2$, and $X_{aa} = x_3$ individuals with genotypes AA, Aa, and aa, respectively. Then the likelihood function is

$$\mathcal{L}(\theta) = \{\theta^2\}^{x_1}\{2\theta(1-\theta)\}^{x_2}\{(1-\theta)^2\}^{x_3}.$$

(The likelihood function follows from the fact that we can write

$$P_\theta(\text{genotype} = g) = \{\theta^2\}^{I(g=AA)}\{2\theta(1-\theta)\}^{I(g=Aa)}\{(1-\theta)^2\}^{I(g=aa)};$$

multiplying these together over the n individuals in the sample gives the likelihood function.)
(a) Find the MLE of θ and give an estimator of its standard error using the observed Fisher information.
(b) A useful family of prior distributions for θ is the Beta family:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 \le \theta \le 1$$

where $\alpha > 0$ and $\beta > 0$ are hyperparameters. What is the posterior distribution of θ given $X_{AA} = x_1$, $X_{Aa} = x_2$, and $X_{aa} = x_3$?