Department of Statistical Science Duke University

FIRST YEAR EXAM - SPRING 2014

Monday, May 5th, 2014

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Put the problem number and your assigned code on the top of **each page**.
3. Write only on **one side** of the page (solutions on the reverse side of the page will be ignored).
4. Start each problem on a new page.
5. It is to your advantage to show your work and explain your answers.
   Do not erase anything– just draw a line through work you do not want graded.
6. You have 3 hours to finish the written exam: Questions 1-6 inclusive. Attempt all questions; note that credit is not necessarily equally allocated across questions.
7. This is a closed book exam. No notes are permitted.

Problem 1. A random quantity $x_0 > 0$ is exponentially distributed conditional on the exponential parameter $\theta > 0$, i.e., $x_0|\theta \sim$ *Ex*(θ) with $E(x_0|\theta) = 1/\theta$. Also, θ has a gamma prior, $\theta \sim$ *Ga*(a, b) with mean a/b and variance a/b^2 for some $a > 0, b > 0$.

1. What is $E(x_0)$?
2. What is the marginal p.d.f. $p(x_0)$?

A second random quantity x_1 has the same conditional distribution $x_1|\theta \sim$ *Ex*(θ) given θ. Also, x_0, x_1 are conditionally independent given θ.

3. What is $p(x_1|x_0)$? Given the complete expression for this conditional density.
4. Sketch one directed graph that represents the joint distribution of θ, x_0, x_1.
5. Sketch the implied undirected graph that represents all conditional dependencies in the distribution of θ, x_0, x_1.

Relabeling x_0, x_1 as x_{t-1}, x_t, the conditional distribution derived in part 3 above is chosen as the transition distribution for a first-order Markov process over time $t = 0, 1, 2, \ldots$.

6. Is the resulting process $x_t, t = 0, 1, 2, \ldots,$ a stationary process? Why or why not?
7. Is it reversible? Why or why not?

Problem 2. Suppose $Y|\mu \sim N(\mu, 1)$, and the prior on μ is $\pi(\mu) = 0.5 \cdot \mathbb{1}(\mu = -1) + 0.5 \cdot \mathbb{1}(\mu = 1)$ where $\mathbb{1}(\cdot)$ is the indicator function. That is, under this prior, μ is supported on {-1,1} and takes either value with 50/50 chance.

1. Show that the posterior $\pi(\mu|y) \propto e^{-y}\mathbb{1}(\mu = -1) + e^{y}\mathbb{1}(\mu = 1)$.
2. Give a simple expression for the Bayes estimator δ under the absolute error loss $L(\mu, a) = |\mu - a|$ and show that its risk function equals $R(\mu, \delta) = 2\Phi(-1)$ for $\mu \in \{-1, 1\}$. What is the associated Bayes risk?
3. Is δ above a minimax estimator under the absolute error loss? Justify.

Problem 3. Suppose $Z \sim N(\mu, 1)$. We can write $Z = Y + K$, where $K = \lfloor Z \rfloor \in \{0, \pm 1, \pm 2, \ldots\}$, i.e. the largest integer $\leq Z$, and $Y = Z - K \in [0, 1)$.

1. Find $p(y|\mu)$, the marginal density of Y. (It is okay to leave the expression in terms of the standard normal pdf.)

2. If we change the distribution of Z to $Z \sim N(\mu + 1, 1)$ how does this affect the marginal distribution of Y? If $Z \sim N(\mu + .5, 1)$ how does this affect the marginal distribution of Y?

Problem 4. Consider an experiment to estimate four unknown parameters $\boldsymbol{\theta} = (\mu, \alpha, \beta, \gamma)^T$ with four observations

$$\begin{aligned} Y_{11} &= \mu + \alpha + \beta + \gamma + \epsilon_{11} \\ Y_{12} &= \mu + \alpha - \beta - \gamma + \epsilon_{12} \\ Y_{21} &= \mu - \alpha + \beta - \gamma + \epsilon_{21} \\ Y_{22} &= \mu - \alpha - \beta + \gamma + \epsilon_{22} \end{aligned}$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

1. Write this as a linear model using matrix notation, and define the design matrix.
2. Are the maximum likelihood estimates of $\boldsymbol{\theta}$ unique? Justify.
3. If possible, compute $\mathsf{Cov}(\hat{\boldsymbol{\theta}})$ and simplify or explain why it does not exist.
4. Suppose that you use the following prior for $\boldsymbol{\theta}, \phi$

   $$p(\boldsymbol{\theta}, \phi) \propto 1/\phi$$

   where $\phi = 1/\sigma^2$

   i) Find the full conditional for ϕ given $\boldsymbol{\theta}$. Simplify as much as possible. Provide the name of the distribution and the values of the corresponding parameters if possible.

   ii) Is the joint posterior distribution of $\boldsymbol{\theta}$ and ϕ a proper distribution? Justify.

Problem 5. In a clinical trial, a group of n patients with a certain tumor type were treated with a standard surgical procedure and were followed up to measure time until relapse. Let $X_1, \cdots, X_n$ denote these relapse times (in months). Another group of m patients with the same tumor type were given a new treatment and their relapse times were recorded in $Y_1, \cdots, Y_m$. Consider the model:

$$X_i \overset{\text{iid}}{\sim} \textit{Ex}(\lambda), \quad Y_j \overset{\text{iid}}{\sim} \textit{Ex}(\lambda\theta), \quad \lambda > 0, \theta > 0$$

with the two groups assumed conditionally independent given the model parameters. Give brief answers to the following questions on inference about λ, the simplest answers do not require any optimization, differentiation or integration.

1. Explain why maximum likelihood inference (testing or interval) on λ will not involve any Y_j.
2. Explain why Bayesian inference on λ will not involve any Y_j if the conditional prior on θ given λ is

   $$\pi(\theta|\lambda) = \lambda g(\lambda\theta), \quad \theta > 0, \lambda > 0$$

   for some density function (proper or improper) g on $(0, \infty)$.
3. Does the same hold if one uses a product prior on (θ, λ), i.e., $\pi(\theta, \lambda) = g(\theta)h(\lambda)$ for some densities g and h on $(0, \infty)$? Explain.

Problem 6. A random variable $X \sim \mathsf{Fr}(\alpha, \beta)$ with the Fréchet distribution with parameters $\alpha, \beta > 0$ has CDF

$$\Pr[X \le x] = e^{-\beta x^{-\alpha}}, \qquad x > 0$$

and hence pdf

$$f(x) = \alpha\beta x^{-\alpha-1} e^{-\beta x^{-\alpha}} \mathbb{1}(x > 0).$$

A random variable $Y \sim \mathsf{Ca}(0,1)$ with the standard Cauchy distribution has CDF

$$\Pr[Y \le y] = \frac{1}{2} + \frac{1}{\pi} \arctan(y)$$

and pdf

$$g(y) = \frac{1/\pi}{1 + y^2}.$$

In the following, $\{X_n\} \overset{\text{iid}}{\sim} \mathsf{Fr}(\alpha, \beta)$ and $Y \sim \mathsf{Ca}(0,1)$; also $X_n^* := \max\{X_j :\ 1 \le j \le n\}$.

1. Find the probability distribution for $X_n^* := \max\{X_j :\ 1 \le j \le n\}$. Give either its name (including the values of any parameters) or a formula for either its pdf or CDF.
2. Find α and β s.t. $\Pr[X_n > u] = \Pr[Y > u] + o(u^{-1})$ as $u \to \infty$.
3. Find $\Pr[X_{n+1} = X_{n+1}^*]$ (or, equivalently, $\Pr[X_{n+1} > X_n^*]$).
4. For $c \ge 1$ find $\Pr[X_{n+1} > c\, X_n^*]$, the probability that X_{n+1} is at least c times larger than any of its predecessors.

Note your answer to (4) with $c = 1$ should reduce to (3). Also, neither of (3) and (4) depends on the scale factor $\beta > 0$, so you may take $\beta = 1$ without any loss of generality for (3) and (4).

FYE 2014
May 5th to 7th 2014

Department of Statistical Science
Duke University

Take Home Data Analysis Problem

Tree growth provides essential information about forest ecology. One common method to estimate tree growth is based on repeated tape measurements of the diameter of the same tree, and the diameter increment is the difference between the current and previous measurement.

The dataset "`diamdata.txt`" (**http://stat.duke.edu/∼lm186/data/diamdata.txt**) contains diameter measurement data for 88 trees obtained from a mapped stand in Duke Forest. The stand was established in 1991 for the purpose of studying forest dynamics. The measurements are made at breast height marked by a nail that holds a tag indicating the identifying tree number. Diameter censuses were conducted at intervals of one to four years starting in 1993. Each year, some trees died and were removed from the census, and some new trees were planted and added to the census, resulting in different numbers of trees measured in each census and different numbers of measurements for each tree. Each tree is indexed by a unique ID (`ID`), and the diameter of a given year (`year`) is the variable `cm`. For each year, three variables of weather information are also available: annual precipitation (`annualprec`), average summer (Jun. - Sep.) Palmer Drought Severity Index (PDSI) (`summerpdsi`), and average winter (Jan. - Mar.) temperature (`wintertemp`).

Explore and analyze the data to *infer about the pattern of tree growth over time*. In this regard, we may be interested in learning about both stand (population) level growth and individual level growth. Write a report of up to three pages regarding tree growth based on your analysis that is understandable and useful to ecologists. Details of key statistical methods or models should be given.

READ THE SIGNATURE SHEET CAREFULLY. Submit your report electronically to Karen Herndon by email (`karen@stat.duke.edu`). Your report file should be named `fye14_codename.pdf`. Your report should not contain your name or any other identifier. It MUST include your assigned code name.

Take-home Applied Exam Signature Sheet

- Keep your answer concise and to the point.
- Present your results in a three page (maximum) report addressing the primary questions posed.
- Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results.
- You may include code and other plots in a supplemental appendix; BUT, you should not assume that graders will read beyond the main report; all relevant material should be within the three page limit.
- You may use all notes, books, software etc from courses and studies to date, and build on your cumulated experience in applied modeling and data analysis.
- BUT– you are also bound by this honor pledge and must sign below to confirm this:
  - I confirm that this Take-home Exam submission is my work alone.
  - I have not consulted at all with any other students, whether they are taking the exam or not.
  - I have not copied nor adapted the work of others, nor provided help or advice to others on this exam.
  - I have not sought out or used any external sources (past student projects, publications, web sites, etc) that explicitly address any aspects of the specific data set and applied problem here. In particular, I have not used web searches to find previous references to the data and earlier analyses of this specific data set and problem, of any kind.
- Sign below and hand this in with your solution.

<u>Name:</u>

<u>Signature:</u>

<u>Date:</u> May 7th 2014

Distribution	Notation	$f(x) =$ **pdf (pmf)**	Support	Mean	Variance	
Beta	$Be(a,b)$	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$x \in (0,1)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	
Bernoulli	$Bern(p)$	$f(x) = p^x q^{(1-x)}$	$x \in \{0,1\}$	p	$p\,q$	$(q = 1-p)$
Binomial	$Bin(n,p)$	$f(x) = \binom{n}{x} p^x q^{(n-x)}$	$x \in \{0,\cdots,n\}$	$n\,p$	$n\,p\,q$	$(q = 1-p)$
Chi-square	$\chi^2(\nu)$	$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$	$x \in \mathbb{R}_+$	ν	2ν	
Exponential	$Ex(\lambda)$	$f(x) = \lambda\, e^{-\lambda x}$	$x \in \mathbb{R}_+$	$1/\lambda$	$1/\lambda^2$	
Gamma	$Ga(\nu,\lambda)$	$f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$	$x \in \mathbb{R}_+$	ν/λ	ν/λ^2	
Geometric	$Geo(p)$	$f(x) = p\,q^x$	$x \in \mathbb{Z}_+$	q/p	q/p^2	$(q = 1-p)$
		$f(y) = p\,q^{y-1}$	$y \in \{1,...\}$	$1/p$	q/p^2	$(y = x+1)$
HyperGeo.	$HG(n,M,N)$	$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$x \in 0,\cdots,n$	$n\,p$	$n\,p\,(1-p)\frac{N-n}{N-1}$	$(p = \frac{M}{N})$
Logistic	$Lo(\mu,\beta)$	$f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$	$x \in \mathbb{R}$	μ	$\pi^2\beta^2/3$	
Log Normal	$LN(\mu,\sigma^2)$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}_+$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$	
Neg. Binom.	$NB(\alpha,p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha q^x$	$x \in \mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$	$(q = 1-p)$
		$f(y) = \binom{y-1}{y-\alpha} p^\alpha q^{y-\alpha}$	$y \in \{\alpha,...\}$	α/p	$\alpha q/p^2$	$(y = x+\alpha)$
Normal	$N(\mu,\sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}$	μ	σ^2	
Pareto	$Pa(\alpha,\epsilon)$	$f(x) = \alpha\,\epsilon^\alpha / x^{\alpha+1}$	$x \in (\epsilon,\infty)$	$\frac{\epsilon\,\alpha}{\alpha-1}$	$\frac{\epsilon^2\alpha}{(\alpha-1)^2(\alpha-2)}$	
Poisson	$Poi(\lambda)$	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{Z}_+$	λ	λ	
Snedecor F	$F(\nu_1,\nu_2)$	$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \times$ $x^{\frac{\nu_1-2}{2}} \left[1+\frac{\nu_1}{\nu_2}x\right]^{-\frac{\nu_1+\nu_2}{2}}$	$x \in \mathbb{R}_+$	$\frac{\nu_2}{\nu_2-2}$	$\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}$	
Student t	$t(\nu)$	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} [1+x^2/\nu]^{-(\nu+1)/2}$	$x \in \mathbb{R}$	0	$\nu/(\nu-2)$	
Uniform	$U(a,b)$	$f(x) = \frac{1}{b-a}$	$x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Weibull	$Wei(\alpha,\beta)$	$f(x) = \alpha\beta\, x^{\alpha-1} e^{-\beta\, x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	