Department of Statistical Science Duke University

FIRST YEAR EXAM - SPRING 2015

Monday 4th May 2015

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Please use **black pen/ink** (no pencils) to complete your final solutions.
3. Put the problem number and your assigned code on the top of **each page**.
4. Write only on **one side** of the page (solutions on the reverse side of the page will be ignored).
5. Start each problem on a new page.
6. 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.
7. You have 4 hours to finish the written exam.
8. **You may choose to complete five out of the six questions. In case you attempt all six questions, only five will be graded and please clearly indicate which five you choose to be graded.**
9. All five graded questions will carry *equal* weight.
10. This is a closed book exam. No notes are permitted.

1. A first-order Markov process with real-valued states x_t is generated from a transition pdf $p(x_t|x_{t-1})$ given by

$$x_t \sim \begin{cases} N(x_t|\phi x_{t-1}, v), & \text{with probability } \alpha, \\ N(x_t|0, s), & \text{otherwise,} \end{cases}$$

where $|\phi| < 1$, $s = v/(1-\phi^2)$ and for some probability α. That is, the $1-$step transition distribution is the $2-$component normal mixture, $p(x_t|x_{t-1}) = \alpha N(x_t|\phi x_{t-1}, v) + (1-\alpha)N(x_t|0, s)$.

   (a) What is the conditional mean $E(x_t|x_{t-1})$ of this state transition distribution?

   (b) Show that the conditional variance $V(x_t|x_{t-1})$ depends quadratically on x_{t-1} and give the explicit formula for this conditional variance.

   (c) Show that the Markov process is ergodic.

   (d) Identify the unique stationary marginal distribution $\pi(x_t)$ for all t.

   (e) Is the process reversible? Either prove or disprove.

2. Counts of clicks on I similar/related ads on a web page on each of T business days are modeled as conditionally independent over days and between ads, with

$$(x_{it}|\theta_i) \sim Po(\theta_i), \qquad \text{independently}, \qquad (i = 1 : I, t = 1 : T),$$

where x_{it} is the number of clicks on ad i on day t. Variation between ads is reflected in the hierarchical model

$$(\theta_i|\mu) \sim Ga(a, a\mu),$$

with $a > 1$ and with the x_{it} conditionally independent of μ given θ_i. Write $X_i = \{x_{it},\ t = 1 : T\}$ for the data for ad i on all days, and $X = \{X_i,\ i = 1 : I\}$ for the full data set.

(a) Show that $(\theta_i|X, \mu) \sim Ga(a + Ty_i, a\mu + T)$, independently, where y_i is a function of the data X. Identify y_i.

(b) Show that the marginal likelihood for μ is $p(X|\mu) \propto \mu^{aI}(T + a\mu)^{-q}$, where $q = I(a + T\bar{x})$ and $\bar{x}$ is the overall sample mean.

(c) Assuming the reference prior $p(\mu) \propto 1/\mu$, show that the posterior for μ is such that $\mu = \phi/\bar{x}$ where $\phi \sim F_{k,h}$ with $k = 2Ia$ and $h = 2IT\bar{x}$.

(d) Give an expression for the posterior harmonic mean of μ, $E(\mu^{-1}|X)$, in terms of $\bar{x}$, a, and I.

Some facts related to the F distribution. If $\phi \sim F_{k,h}$ then:

- $p(\phi) \propto \phi^{k/2-1}/(h + k\phi)^{(k+h)/2}$,
- $\phi^{-1} \sim F_{h,k}$,
- $E(\phi) = h/(h - 2)$ if $h > 2$, and
- $E((h + k\phi)^{-1}) = (h + k)^{-1}$.

3. The random variables $X \sim \textit{Ex}(\theta)$ and $Y \sim \textit{Po}(\lambda)$ are independent, with pdf and pmf

$$f(x \mid \theta) = \theta e^{-\theta x}, \quad x > 0 \qquad\qquad p(y \mid \lambda) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y \in \{0, 1, 2, \dots\}$$

respectively. Their means are $\mathsf{E}[X] = 1/\theta$, $\mathsf{E}[Y] = \lambda$ and their variances are $\mathsf{Var}[X] = 1/\theta^2$, $\mathsf{Var}[Y] = \lambda$.

(a) What does it *mean* mathematically for an absolutely continuously distributed random variable like X and a discrete random variable like Y to be *independent*?

(b) As a function of $\theta > 0$ and $\lambda > 0$, find: $\mathsf{P}[X > Y] =$

(c) If $\{X_i\} \overset{\text{iid}}{\sim} \textit{Ex}(\theta)$ and $\{Y_i\} \overset{\text{iid}}{\sim} \textit{Po}(\lambda)$ are all independent, and if $\theta = \lambda = 2$ and $n = 100$, find the approximate probability

$$\mathsf{P}\Big[\bar{Y}_n - \bar{X}_n > 1.8\Big] \approx$$

where $\bar{X}_n := \frac{1}{n}\sum_{i \le n} X_i$ and $\bar{Y}_n := \frac{1}{n}\sum_{i \le n} Y_i$ denote the sample means.

(d) If $\{X_i\} \overset{\text{iid}}{\sim} \textit{Ex}(\theta)$ and $\{Y_i\} \overset{\text{iid}}{\sim} \textit{Po}(\lambda)$ are all independent, and if $\theta = \lambda = 2$, does the sequence $\bar{X}_n \bar{Y}_n$ converge almost-surely? If so, to what limit and why? If not, why?

$$\lim_{n \to \infty} \bar{X}_n \bar{Y}_n =$$

and justify your answer.

4. Let $X_1, X_2, \ldots$ be an infinite sequence of $i.i.d$ Poisson(λ) random variables, and N is a geometric random variable, independent of the X's, with success probability $p \in (0, 1)$. Specifically, $P(N = k) = p(1-p)^{k-1}$ for $k = 1, 2, \ldots$. An experiment is carried out in which N and only $X_1, X_2, \ldots, X_N$ are observed.

   (a) If λ is *unknown* but p is *known*, find a sufficient statistic that is as simple as possible. Is it a complete statistic?

   (b) For this and all following parts, assume that both λ and p are *unknown*. What is a complete sufficient statistic?

   (c) Find a UMVU estimator for $e^{-\lambda}$.

   (d) Find a level α UMPU test for $H : \lambda = 1$ vs $\lambda > 1$. (Just give the form of the test. You don't have to find the exact expressions of the constants involved.)

5. An experimenter exploring dose response adopts the model for the responses

$$y_i = \beta_0 + \beta_1 d_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2 d_i), \qquad i = 1 : n, \tag{1}$$

where the d_i are positive dose levels; note that error variances are proportional to these dose levels, with σ assumed known.

(a) The experimenter then realizes that she is *absolutely* convinced that: if the dose is zero, then the *expected* response *must* be zero. She adopts the improper prior $p(\beta_1) \propto$ constant for the free slope parameter, after fixing this constraint. What is the resulting posterior distribution for $(\beta_1 \mid y_{1:n})$ (in terms of summary statistics)?

(b) The experimenter then faces a design question: she can choose each $d_i \in \{1, 2, 3\}$ only, and faces the decision of choosing the numbers n_1 of $d_i = 1$, n_2 of $d_i = 2$ and n_3 of $d_i = 3$, constrained by a fixed total available sample size n.

Using the model and assumptions of (5a) above, what are the optimal sample sizes n_1, n_2, n_3 such that the posterior variance of $\beta_1 | y_{1:n}$ is minimized?

(c) As a colleague, you are concerned that the mean function may not go through the origin. Does the above design allow you to check this assumption? Explain.

(d) You encourage a reanalysis that does not impose the zero intercept constraint, and uses the reference prior $p(\beta_0, \beta_1) \propto$ constant. What now is the posterior variance of β_1? Express the result in terms of n_1, n_2, n_3 and their total n.

(e) Under this more general model, can the dose-sample allocation in (5b) above be the optimal design?

6. A random quantity x has a Pareto (power-law) distribution $x \sim Pa(\alpha, c)$ with p.d.f.

$$p(x|\alpha, c) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}(x > c) \equiv \begin{cases} \alpha c^\alpha / x^{\alpha+1}, & \text{for } x > c, \\ 0, & \text{otherwise.} \end{cases}$$

for some parameters $\alpha > 0, c > 0$.

(a) Suppose $x \sim Pa(\alpha, c_0)$ for some $\alpha > 0, c_0 > 0$ and you now learn that $x > c$ for some other constant $c > c_0$. What is the resulting distribution of $x \mid x > c$?

(b) For this and the following parts, suppose that $x_1, \ldots, x_n$ are the sizes of the populations of the n *largest* cities in North Carolina, and a demographer assumes the model

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} Pa(\alpha, c).$$

There is a potential issue with using an i.i.d. model for this data, due to the fact that we are only seeing the n *largest* cities; briefly comment on why an issue arises if we are interested in inference on c.

(c) How does the answer to the first question (6a) above alleviate this issue when c is known and we are interested only in inference for α?

(d) Using the prior $\pi(\alpha, c) \propto 1/\alpha$ on $\alpha > 0$ and $c > 0$, derive the full conditionals for Gibbs sampling from the posterior $p(\alpha, c|x_1, \ldots, x_n)$. Identify the full conditional for α as a well-known distribution.

(e) How would you simulate from the full conditional of c?

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	$Po(\lambda)$	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{Z}_+$	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	$Wei(\alpha,\beta)$	$f(x) = \alpha\beta\, x^{\alpha-1}\, e^{-\beta\, x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	

FYE 2015 Department of Statistical Science
April 20th, 12noon to April 23rd, 12noon, 2015 Duke University

Take Home Data Analysis Problem

The dataset "`IrishElectricity.txt`" (`https://stat.duke.edu/~fl35/data/IrishElectricity.txt`) provides daily total electricity consumption (in kilowatt-hour (kWh)) for 151 households in Ireland for the four month period between November 15, 2009 to March 15, 2010 (column 7-127). Each row represents a household. The dataset can be read into R using command `read.table`.

For each household, there are also six covariates (column 1-6) on demographics:

`Age`:	Age of the head of the household, ordinal (1 to 6: 1=youngest, 6=oldest)
`Attitude-Reduce Bill`:	How strongly the person feel about reducing bill, ordinal (1 to 5: 1= strongest)
`Attitude-Environment`:	How strongly the person feel about the environment, ordinal (1 to 5: 1= strongest)
`Education`:	Education level of the the head of the household, ordinal (1 to 5: 1= no edu, 2=elementary, 3=middle school, 4=high school, 5= college or above)
`Bedroom`:	Number of bedrooms
`Resident`:	Number of residents in the household

Analyze the data to address the following two questions:

1. Describe the patterns of household electricity consumption. In this regard, we are interested in learning about both population level usage as well as household level usage.

2. On January 1, 2010, there was a major change in the tariff structure on electricity consumption. Is there a difference in energy consumption pattern before versus after the policy change? During the 2.5 months after the change does the pattern of energy consumption at the household level tend to return to the pattern before the policy change?

Write a report (maximum 3 pages) based on your analysis describing your findings. Be thorough in your exploratory analyses and exploratory use of models; applied work that overly emphasizes complicated modeling to begin is often less valuable than careful, incisive evaluation of data through simpler, exploratory models– at least to begin.

Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results.

Take-home Applied Exam

- Present your results in a three page (maximum) report addressing the primary questions posed. Keep your answers concise and to the point.
- 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.
- You may freely use other resources– code, literature, etc– from whatever source you like, so long as you do not violate the condition 4 below.
- **To Confirm**– you are also bound by this honor pledge and must sign below to confirm this:
    1. I confirm that this Take-home Exam submission is my work alone.
    2. I have not consulted at all with any other students, whether they are taking the exam or not.
    3. I have not copied nor adapted the work of others, nor provided help or advice to others on this exam.
    4. 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 before or at 12noon, April 23 (Thursday) 2015 to Lori Rauch at Room 214, Old Chem.

Name:

Signature:

Date: April 23rd 2015