

# FIRST YEAR EXAM

Monday May 4, 2009; 9:00 – 12:00am

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. No problem is associated with any particular course, so you should **attempt to work as many parts as feasible**.
7. You have 3 hours to finish.
8. This is a closed book exam. No notes are permitted.  
A page with common p.d.f. and p.m.f. formulas is attached.
9. The Take-Home practical will be available from Karen Herndon in 214 Old Chemistry immediately after dropping of this written exam and is required to be handed in by 5:00 PM on Tuesday May 5th to Karen Herndon in 214 Old Chemistry.

1. Suppose  $Y_i$  follows a Poisson distribution with mean  $\beta x_i$ , where  $x_i$  is a fixed, known, positive constant.

$$\Pr(Y_i = y_i) = (\beta x_i)^{y_i} e^{-\beta x_i} / y_i! \quad y_i = 0, 1, 2, \dots$$

- (a) Find a minimal sufficient statistic for this family of distributions.
- (b) Find the Cramer-Rao bound for the variance of an unbiased estimator of  $\beta$ .
- (c) Find the MLE for  $\beta$ .
- (d) Find the variance of the MLE for  $\beta$ .
- (e) Give an approximate 95% confidence interval for  $\beta$  when  $n$  is large. Carefully describe any theorem that you use.

2. Let  $X_1, \dots, X_n$  be independent identically distributed Normal random variables with mean  $\mu$  and standard deviation  $\mu$  with  $\mu > 0$ .
- (a) Is this a one-parameter exponential family? (show yes or no)
  - (b) Let  $\phi = 1/\mu$ . Suppose your prior distribution for  $\phi$  is a  $G(a, b)$  with rate  $b$ <sup>1</sup>. Find the posterior distribution up to any normalizing constant.
  - (c) Give an expression for a conjugate prior density for  $\phi$  (up to any normalizing constant).
  - (d) What is Jeffreys' prior for  $\phi$ ?
  - (e) What is Jeffreys' prior for  $\mu$ ?
  - (f) Based on a sample of size  $n = 10$  the posterior for  $\phi$  under the Jeffreys' prior for  $\phi$  leads to a 95% highest posterior density (HPD) interval for  $\phi$  of (98,116). Find a 95% posterior credible interval for  $\mu$ .
  - (g) Based on this information, can you identify a 95% HPD interval for  $\mu$ ?

---

<sup>1</sup>Recall  $\phi \sim G(a, b)$  with shape  $a > 0$  and rate  $b > 0$  ( $1/b$  is also known as the scale) has density

$$f(\phi) = \frac{1}{\Gamma(a)} b^a \phi^{a-1} e^{-b\phi}, \quad \phi > 0.$$

3. (a) Two real-valued random variables  $X$  and  $Y$  have a joint density function of the form  $f(x, y) \propto \exp(-Q(x, y)/2)$  where

$$Q(x, y) = x^2 + y^2 - 2axy$$

for some constant  $a$  such that  $|a| < 1$ . Derive answers to the following questions and state the results using standard notation for normal distributions (i.e,  $Z \sim N(m, v)$  and so forth).

- i. What is the conditional distribution of  $X|Y$ ?
  - ii. What is the conditional distribution of  $Y|X$ ?
  - iii. *Without using direct integration*, what is the marginal distribution of  $X$ ?
- (b) In another context, it is known that both  $X$  and  $Y$  are positive so the density function is modified to  $f(x, y) \propto \exp(-Q(x, y)/2)$  for  $x > 0, y > 0$  with  $f(x, y) = 0$  otherwise. Derive the explicit form of the density function  $f(x|y)$  under this modification.

4. Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be the open unit interval  $\Omega = (0, 1]$  with the Borel sets  $\mathcal{F} = \mathcal{B}(\Omega)$  and Lebesgue measure  $\mathbb{P}(d\omega) = d\omega$ . Set  $X_n(\omega) := n 1_{(0, 1/n^2]}(\omega)$  for each integer  $n \in \mathbb{N}$ .

(a) Does  $X_n(\omega)$  converge at each point  $\omega \in \Omega$ ? If so, find the limit  $X(\omega) \equiv \lim_{n \rightarrow \infty} X_n(\omega)$ ; if not, tell why.

Circle one:      **Yes**      **No**       $X(\omega) =$

Reasoning:

(b) Does  $\int_{\Omega} |X_n - X| d\mathbb{P} \rightarrow 0$ ?    **Y**   **N**    Show why (or why not).

(c) Does  $\int_{\Omega} |X_n - X|^2 d\mathbb{P} \rightarrow 0$ ?    **Y**   **N**    Show why (or why not).

(d) Is  $\{X_n\}$  uniformly bounded by a positive integrable random variable  $Y$ ?

If so, find a suitable  $Y \geq |X_n|$  and verify  $Y \in L_1$ ; if not, explain.    **Y**   **N**    .

(e) Is  $\{X_n^2\}$  uniformly bounded by a positive integrable random variable  $Y$ ?

If so, find a suitable  $Y \geq X_n^2$  and verify  $Y \in L_1$ ; if not, explain.

**Y**   **N**

5. Consider the linear model

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N(0, \sigma^2 I_n)$  and  $I_n$  is the  $n \times n$  identity matrix and  $X$  is a  $n \times p$  matrix of rank  $p$ . Let  $U\Lambda V^T$  be the singular value decomposition of  $X$ , where  $\Lambda$  is diagonal with positive diagonal elements. Consider transforming the model above,

$$Y_* \equiv U^T Y = U^T X\beta + U^T \epsilon.$$

(a) Show that

$$Y_* \sim N(\Lambda\alpha, \sigma^2 I_p)$$

where  $\alpha = V^T \beta$ .

- (b) Show that the Maximum Likelihood Estimate (MLE) of  $\alpha_j$  is  $y_{*j}/\lambda_j$  for  $j = 1, \dots, p$  where  $y_{*j}$  is the  $j$ th element of  $Y_*$  and  $\lambda_j$  is the  $j$ th diagonal element of  $\Lambda$ .
- (c) Suppose the prior distribution for  $\alpha$  is  $N(0, I_p/k)$ . Find the posterior mean of  $\alpha$  as a function of the MLE of  $\alpha$ .
- (d) Find the posterior mean of  $\beta$ .
- (e) Now treating the posterior mean as an estimator of  $\alpha$ , find the expected mean squared error:  $E[(\tilde{\alpha} - \alpha)^T(\tilde{\alpha} - \alpha)]$  where  $\tilde{\alpha}$  is the posterior mean of  $\alpha$  and the expectation is taken with respect to the conditional distribution  $f(Y_* | \alpha, \sigma^2)$ .

6. The first-order autoregression model is commonly utilized in analysis of time series data and longitudinal data. This model assumes:

$$X_{t+1} = \theta X_t + \epsilon_t \quad t = 0, \dots, n$$

where  $X_t$  are observed random variables with  $X_0 = 0$ ;  $\epsilon_t$  are i.i.d  $N(0, \sigma^2)$  with known  $\sigma^2$ ; and the unknown parameter  $\theta$  controls the correlation between two successive observations  $X_t, X_{t+1}$ . This is a model for the observed data  $\mathbf{X} = (X_0, \dots, X_{n+1})$  in the sense that  $\epsilon_t$  are unobserved except for  $t = 0$ .

- (a) Show that for  $n \geq 1$ ,

$$X_n = \sum_{t=0}^{n-1} \epsilon_t \theta^{n-t-1}.$$

- (b) Find an expression for the score for  $\theta$ :  $S(\mathbf{X}; \theta) = \frac{\partial \log f(\mathbf{X}|\theta)}{\partial \theta}$ .  
(c) Find the Fisher information for  $\theta$  based on observed data  $\mathbf{X}$ .

## 7. FYE'09 Take-Home problem

Turn in solution to Karen Herndon in Room 214 Old Chemistry by 5pm on Tuesday, May 5.

A chemical engineering experiment was run to study heat transfer in a shallow fluidized bed. Data is collected on the following four candidate regressors -  $X_1$ : fluidizing gas flow rate in pounds per hour,  $X_2$ : supernatant gas flow rate in pounds per hour,  $X_3$ : supernatant gas inlet nozzle opening in millimeters,  $X_4$ : supernatant gas inlet temperature, °F.

The measured responses are  $Y_1$ : heat transfer coefficient,  $Y_2$ : thermal efficiency. Twenty observations were gathered with the data below.

Obs	$Y_1$	$Y_2$	$X_1$	$X_2$	$X_3$	$X_4$
1	41.85	38.75	69.69	170.83	45	219.74
2	155.32	51.87	113.46	230.06	25	181.22
3	99.62	53.79	113.54	228.19	65	179.05
4	49.41	53.84	118.75	117.73	65	281.30
5	72.96	49.17	119.72	117.60	25	282.30
6	107.70	47.61	168.38	173.46	45	216.14
7	97.23	64.19	169.35	169.85	45	223.88
8	105.85	52.73	169.85	170.86	45	222.80
9	95.34	51.00	170.89	173.92	80	218.34
10	111.90	47.37	171.31	173.34	25	218.12
11	100.01	43.18	171.43	171.43	45	219.20
12	175.38	71.23	171.59	263.49	45	168.62
13	117.30	49.30	171.63	171.63	45	217.58
14	217.41	50.87	171.93	170.91	10	219.92
15	41.73	54.44	173.92	71.73	45	296.60
16	151.40	47.93	221.44	217.39	65	189.14
17	220.63	42.91	222.74	221.73	25	86.08
18	131.66	68.60	228.90	114.40	25	285.80
19	80.54	64.94	231.19	113.52	65	286.34
20	152.96	43.18	236.84	167.77	45	221.72

### Questions.

- Explore linear normal regression models for predicting heat transfer coefficient,  $Y_1$  using the  $X_j, j = 1, 2, 3, 4$ . Investigate relevant transformations of  $Y_1$  and consider quadratic terms and product terms. Check usual model assumptions.
- Suppose for a new run of the experiment, we are given only that  $x_1 = 116.9$  and  $x_2 = 172.1$ . Estimate the associated heat transfer coefficient. Estimate the probability that this coefficient exceeds 100.



- (c) Suppose we only observe whether or not  $Y_1 > 100$ . Adopting a suitable model choice criterion, obtain a good model using the  $X_j, j = 1, 2, 3, 4$  for explaining the probability that the heat transfer coefficient exceeds 100.
- (d) Does the regression model that you selected for predicting  $Y_1$  in part (a) do a good job of predicting  $Y_2$ ?

Please be rigorous in providing a full justification for each of your answers, including all relevant statistical details, calculations and results. Report the results in a manner interpretable by a chemical engineer interested in the study conclusions.

The data can be found at: <http://www.stat.duke.edu/programs/grad/FYE09takehome.xls>