CSC591: Foundations of Data Science
HW5: Bayesian Inference, Missing Data Analysis

Released: 11/25/15
Due: **12/04/15 (23:55pm);** (One day late: -25%; -100% after that).

Student Name:
Student ID:

Notes

- Submit single zip file containing: (1) all solutions as single pdf file (Filename: Lastname_StudentID.pdf); (2) separate R code file for R question with appropriately named readme files.
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth 5% of total grade + **Bonus questions account for 4% of grade**
- You can discuss with your friends, but solution should be yours.
- Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
- All submissions must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- No makeups; for regarding policies, refer to syllabus and 1st day lecture slides.

Q#	Max Points	Your Score
1	30	
2	35	
3	10	
4	25	
R	R project (Bonus: 4% of Grade)	

Note: For each question (and subparts) please also list how much time it took to solve it.

Q1. Bayesian Inferencing (30 points)

(a) Following table give prior distribution for the proportion of defective parts produced by a machine

p (proportion of defects)	0.2	0.3
f(p) : posterior probability	0.7	0.3

Let x denote the number of defectives among a random sample of size 2.

(1) Find the posterior probability distribution of p, given that x is observed. (**10 points)**

(2) Estimate the proportion of defectives being produced by the machine if the random sample of size 2 yields 2 defects. **(10 points)**

(b) One of the standard measures (say "acceleration") reported for cars is the time (in seconds) required to reach 0-60 mph. A company determines that the acceleration for their new car is a normal r.v. with a s.d of 0.8 sec. Assume a normal prior distribution, N(8, 0.2). If 10 of the production cars are tested and determined that the average acceleration is 9 sec, then find the 95% Bayesian interval for μ . **(10 points)**

Q2. The following table summarizes two exam scores. Left half of the table gives complete scores and right half gives an example of missing data. (35 points)

Complete Data		Missing Data	
mt1	mt2	mt1	mt2
74	66	74	66
70	58	70	58
66	74	66	74
55	47	55	47
52	61	52	61
47	38	47	38
45	32	45	
38	46	38	
33	41	33	41
28	44	28	

Answer the following (using data given in the above table):

(1) Based on the missing data, determine missing data pattern and justify your answer (5 points)

(2) Compute Mean and Standard Error (SE) for (i) complete data, and (ii) missing data using list-wise deletion. (10 points)

(3) Comment on bias of the estimates of (2.ii) as compared to estimates from complete data (2.i). (5 points)

(4) Impute missing data using simple regression (see slide 22 from w15-c1-missing-data). (10 points)

(5) Compute Mean and SE on imputed data, and comment on bias of the estimates. (5 points)

Q3. (**10 points**)

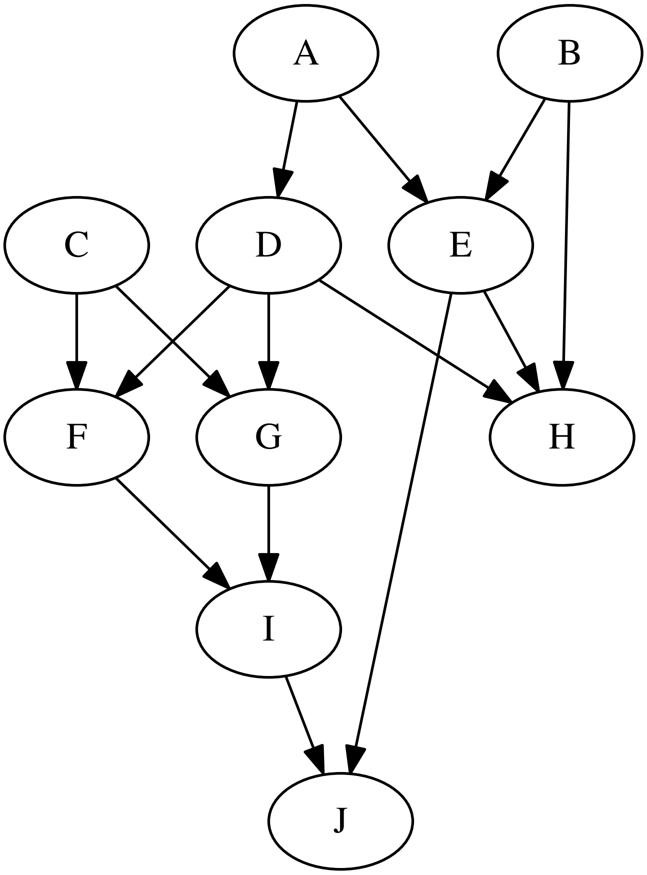
(a) Describe missing data patterns and missing data mechanisms (**5 points**)

(b) Describe various traditional methods for dealing with missing data, highlight advantage and disadvantages of each method (**5 points**)

Q4. Bayes Networks (25 points)

(a)

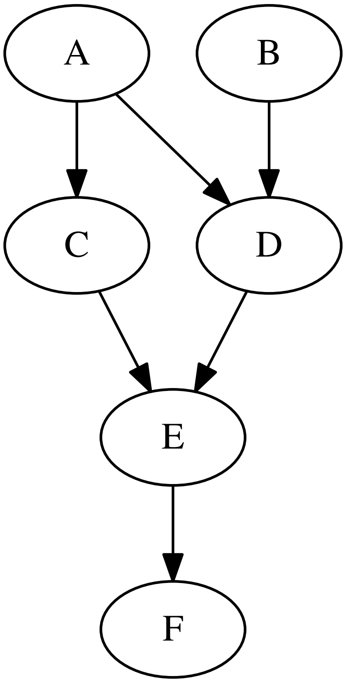
For each independence statement, state whether the independence is implied by the given Bayes net. If there are any active trails between the given variables, name one.



- $A \perp C$?
- $A \perp C \mid E$?
- $A \perp C \mid I$?
- $D \perp I$?
- $D \perp I \mid G$?
- $D \perp I \mid F, G, J$?
- $D \perp I \mid F, G, J, A$?
- $F \perp H$?
- $F \perp H \mid A$?
- $F \perp H \mid D$?
- $F \perp H \mid D, J$?
- $F \perp H \mid D, I, J$?

Part 2

Given the following Bayes net (all variables are binary):



$P(A = 1) = 0.7$

$P(B = 1) = 0.4$

$P(C = 1 \mid A = 0) = 0.2$

$P(C = 1 \mid A = 1) = 0.7$

$P(D = 1 \mid A = 0, B = 0) = 0.3$

$P(D = 1 \mid A = 0, B = 1) = 0.7$

$P(D = 1 \mid A = 1, B = 0) = 0.6$

$P(D = 1 \mid A = 1, B = 1) = 0.2$

$P(E = 1 \mid C = 0, D = 0) = 0.8$

$P(E = 1 \mid C = 0, D = 1) = 0.6$

$P(E = 1 \mid C = 1, D = 0) = 0.1$

$P(E = 1 \mid C = 1, D = 1) = 0.5$

$P(F = 1 \mid E = 0) = 0.9$

$P(F = 1 \mid E = 1) = 0.6$

Compute:

a) $P(A = 1, B = 0, C = 0, D = 1, E = 0, F = 1)$

b) $P(A = 1, E = 0)$
Use the elimination ordering F, D, C, B (F is the innermost sum)

R. Bonus Question (R implementation) (**4% of grade**) (Please note that this is completely optional; use your time wisely as the implementation may take time).

(You can use any 2-d data, real or simulated for implementation; test data will be provided later to answer part b of this question)

(**a**) Implement G-Means (paper is provided under additional resources) (Algorithm 1, listed on page 3). (submit code as separate file; make single zip file)

(**b**) Generate 2-d plots (scatter plots and draw ellipsoids) (data will be provided later), include these plots as part of h/w solution)