Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
MIDTERM EXAM Fall 2016 (Nov. 1, 2016)

60 points; 2 hours; closed book, no notes. *Answer on exam sheets where possible, & put **your name** on them.* No red ink. You can use $\Gamma(a, b, ...) = -a \log a - b \log b - ...$ When constraints are asked about, they must always be linearly independent. If any question seems ambiguous, ask me about it.

1. Basics [22 points]

(a) For the following contingency table, with probabilities a+b+c+d=1, if X and Y are independent, then (circle all that are true; 2 pts)

	Y_1	Y_2
X_1	a	b
X_2	c	d

i. a/b = c/d ii. a/c = b/d iii. a/d = b/c iv. none of the above

(b) A test is given to detect a disease. A 'negative' test result means that this condition is not detected, i.e., the patient is judged to be free of the disease; a 'positive' result means that the condition is detected, i.e., the patient is judged to have the disease. The actual condition of the patient and the test conclusions are summarized in these frequencies: TN = true negatives, FP = false positives, FN = false negatives, TP = true positives. The frequency margins of the actual distribution are N = TN + FP and P = FN + TP.

	Test	(-)	(+)	
Actual	(-)	TN	FP	N
	(+)	FN	TP	P

In the communication situation, where S means message sent and R means message received, H(R|S) is called 'noise' and H(S|R) is called 'equivocation.' In the present case, the Actual situation is the message sent, and the Test results are the message received. Which of the following is true (circle one; 2 pts)?

i. FP = 'noise' and FN = 'equivocation'
ii. FP = 'equivocation' and FN = 'noise'
iii. neither: FP and FN do not map 1:1 onto these conditional entropies

(c) Consider the following data table.

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	0	3/16	1/16	0
A_2	3/16	0	0	9/16

True or False? (circle one; 2 pts): For this ABZ table, T(A:B) = 0, i.e., there is no constraint between A and B.

(d) Consider data ABZ where T(A:Z) > 0 and T_B(A:Z) = 0. Ignore the possible relevance of variables other than A,B, and Z. These two facts can be interpreted as saying (circle all interpretations that *could* be true; 2 pts)

i. A directly predicts B which directly predicts Z, so A predicts Z indirectly
ii. B directly predicts A which directly predicts Z, so B predicts Z indirectly
iii. A directly and separately predicts both B and Z
iv. B directly and separately predicts both A and Z.
v. A and B both directly predict Z via a 3-way interaction effect
vi. A does not predict Z at all, either directly or indirectly
vii. B does not predict Z at all, either directly or indirectly

(e1) If the BN (Bayesian network) model q(ABZ) = p(A) p(B) p(Z|AB) is used to predict Z from AB, it will give the same predictions as which *RA* model? (circle one; 2 pts)

i. ABZ	ii. AB:AZ:BZ	iii. AB:AZ	iv. AB:BZ	v. AZ:BZ
vi. AB:Z	vii. AZ:B	viii. BZ:A	ix. A:B:Z	x. none of these

(e2) For A, B, and Z being binary variables, how many degrees of freedom are needed to specify this BN model? df = _____ (2 pts)

(f) Probability distribution I has values p(x) = .4, .3, .2, .1, for x = 1, 2, 3, 4 respectively, and probability distribution II has values p(x) = .4, .1, .2, .3 for x = 1, 2, 3, 4 respectively. One can calculate variances for these two distributions, where variance is the sum of square distances of values of x from the average x, weighting each square distance by the probability of that value. Entropy is a nominal data analog of variance, and also measures the spread of a distribution. Is the *following* statement True or False? (circle one; 2 pts): The entropy of distribution I = the entropy of distribution II.

(g) True or False? (circle one; 2 pts): Transmission, like correlation, ranges in value from –1 for perfect inverse association to +1 for perfect direct association.

(h) True or False? (circle one; 2 pts): If a triadic ABC relation exhibits non-zero constraint, then there exists at least one dyadic projection (AB, AC, and/or BC) that exhibits non-zero constraint.

(i) Consider the contingency table whose X and Y "margins" are shown below. Write an expression in terms only of the numerical constants, .3, .7, .4, .6 for T_{max}, the maximum value that T(X:Y) could possibly have: T_{max} = (2 pts)

	Y_1	Y_2	
X_1			.3
X_2			.7
	.4	.6	

(j) True or False (circle one; 2 pts):

H(AB:AZ:BZ) = H(AB) + H(AZ) + H(BZ) – H(A) – H(B) – H(Z).

2. Structures [16 points]

(a) In a top-down search of neutral system ABCD, one can descend in one step from ABC:ABD:CD to ABC:ABD or to ABC:AD:BD:CD. All variables have cardinality 2. What is the relationship between df values for the descendant models? (circle one; 2 pts)

i. df(ABC:ABD)<df(ABC:AD:BD:CD) ii. df(ABC:ABD)>df(ABC:AD:BD:CD)
iii. df(ABC:ABD)=df(ABC:AD:BD:CD)

(b) What is the nearest common ancestor of ABC:D and AB:BC:CD:DA? (2 pts)

(c) What is the nearest common descendant of ABC:D and AB:BC:CD:DA? (2 pts)

(d) $T_A(B:Z) = 0$. Circle all models where T(model) = 0. (2 pts)

i. ABZ	ii. AB:AZ:BZ	iii. AB:AZ	iv. AB:BZ	v. AZ:BZ
vi. AB:Z	vii. AZ:B	viii. BZ:A	ix. A:B:Z	x. none of these

(e) Consider a directed system in which A, B, and C are IVs, and Z is a DV. How many *general* structures are there? Exemplify each general structure with one specific structure. Circle all structures which *have* loops. (4 pts)

(f) $\Delta df(ABC:ABD:ACD:BCD \rightarrow AB:AC:AD:BC:BD:CD) = df(ABC:ABD:ACD:BCD) - df(AB:AC:AD:BC:BD:CD)$. Let A, B, C, & D have cardinalities of 2, 3, 3, & 4, respectively. Use the log-linear method and compute this Δdf. (2 pts)

$\Delta df =$

(g) For model AB:BC:CD, where |A| = 2, |B| = 3, |C| = 4, |D| = 5, the number of constraints are _____ (fill in a precise number). (2 pts)

3. Information-theoretic Reconstruction [22 points]

(a) The expression for $I(m_k \rightarrow m_j)$, the information distance between models m_k and m_j, where m_k is above m_j, where q_k means $q(m_k)$ and q_j means $q(m_j)$, is: (circle one; 2 pts)

i. $\Sigma\ p \log [q_k / q_j]$ ii. $\Sigma\ p \log [q_j / q_k]$ iii. $\Sigma\ q_k \log [q_k / q_j]$
iv. $\Sigma\ q_k \log [q_j / q_k]$ v. $\Sigma\ q_j \log [q_k / q_j]$ vi. $\Sigma\ q_j \log [q_j / q_k]$

(b) T(AB:Z) – T(AB:BZ) is the amount of information (circle one; 2 pts):

i. captured in model AB:Z ii. captured in model AB:BZ
iii. lost in model AB:Z iv. lost in model AB:BZ v. none of above

(c) Suppose I go *down* the lattice of structures from ancestor to descendant. After each of these three measures, write whether the measure is MD (monotonically decreasing or staying the same), MI, (monotonically increasing or staying the same), or NM (not monotonic, i.e., could either increase or decrease) (3 pts):

(i) H (ii) T (iii) df

(d) On the left is *observed* data (p) for a directed system, with sample size N and probabilities a..h (all non-zero). On the right is the *calculated* table (q) for AB:AZ:BZ.

p(ABZ)

	Z:	0		1	
	B:	0	1	0	1
A:	0	a	b	c	d
	1	e	f	g	h

$q_{AB:AZ:BZ}(ABZ)$

	Z:	0		1	
	B:	0	1	0	1
A:	0	q_1	q_2	q_3	q_4
	1	q_5	q_6	q_7	q_8

(d1) Give a numerical answer (1 pt): df(AB:AZ:BZ) =

(d2) Fill in entries of the tables for AB, AZ, and BZ that are projected from p and q. Fill in *only* cells on the left and right where equating corresponding cell contents constitute the constraints that the AB:AZ:BZ model imposes on entropy maximization. That is, the number of cells on either side that you fill in (in terms of parameters a…h or unknowns $q_1 \ldots q_8$) should be exactly the number of constraint equations. Do this for the projections *sequentially*. First show the maximum number of *linearly independent* constraint equations for AB; then the maximum number of *additional* linearly independent constraint equations for AZ, given those selected for AB; then the maximum number of *additional* linearly independent constraint equations for BZ, given those selected for AB and AZ. (3 pts)

Projected from p

	B:	0	1
A:	0		
	1		

Projected from q

	B:	0	1
A:	0		
	1		

	Z:	0	1
A:	0		
	1		

	Z:	0	1
A:	0		
	1		

	Z:	0	1
B:	0		
	1		

	Z:	0	1
B:	0		
	1		

(d3) Write the full set of constraints in the form M **q** = M **p**, where p and q are column vectors, by filling in, under 'Matrix, M' *only rows sufficient and necessary* to define this constraint matrix for AB:AZ:BZ. Fill in values *only where the value is 1*; leave the matrix element blank when its value is 0. The same matrix is 'M' is understood to be on the right. (M may or may not require as many rows as shown here.) (3 pts)

Matrix, M

q_1			a
q_2			b
q_3			c
q_4	=	M	d
q_5			e
q_6			f
q_7			g
q_8			h

(d4) In fitting AB:AZ:BZ, what quantity, expressed in terms of unknown q's and/or parameters a…h, is maximized subject to the above constraints? (2 pts)

(d5) The maximum entropy solution for the q's is obtained by (circle one; 2 pts)

i. $q_{AB:AZ:BZ}(ABZ) = p(AB)\ p(AZ)\ p(BZ)$
ii. $q_{AB:AZ:BZ}(ABZ) = p(AB)\ p(AZ)\ p(BZ) / [\ p(A)\ p(B)\ p(Z)\ ]$
iii. iteration (IPF) that imposes the projections until convergence

(e) Given a directed system with variables A, B, C, and Z, the information distance I(ABC:ABZ → ABC:AZ) = (circle all that are true, 4 pts)

i. H(Z|A) – H(Z|AB) ii. H(Z|A) + H(Z|AB)
iii. H(Z|B) – H(Z|AB) iv. H(Z|B) + H(Z|AB)
v. H(Z|AB) – H(Z|A) vi. H(Z|AB) + H(Z|A)
vii. H(Z|AB) – H(Z|B) viii. H(Z|AB) + H(Z|B)
ix. T(A:Z) x. T(B:Z) xi. T_B(A:Z) xii. T_A(B:Z)
xiii. T(AB:Z) xiv. T(ABZ) xv. none of the above