Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
FINAL EXAM Fall 2016 (Dec 6, 2016)

60 points; 2 hours; closed book, closed notes. *Answer questions on exam sheets, where possible, and put **your name** on them.* You can use $\Gamma(a, b, ...) = -a \log a - b \log b - ...$

1. [16 points] H and T

(a) Consider data below for a *neutral* system, with probability values a…h. In the Lattice of Structures presented in class, the bottom model is the independence model, X:Y:Z. An alternative is the uniform distribution which doesn't preserve the X, Y, and Z marginal distributions. There are other models *between* X:Y:Z and the uniform distribution; for example, the X:Y:Φ model, which says that q(XYZ) agrees with the X and Y margins but is uniform (Φ) in Z. Using parameters a through h, what transmission quantity would you use to assess the agreement of this X:Y:Φ model with the data? Express this quantity using $\sum$ p log p/q expression (*not* using entropies) and give only *the first two terms* of the expression. (2 pts)

X	Y	Z	p
0	0	0	a
0	0	1	b
0	1	0	c
0	1	1	d
1	0	0	e
1	0	1	f
1	1	0	g
1	1	1	h

(b) For *directed* system ABCZ, T(ABC:BZ) – T(ABC:ABZ) = (circle one; 2 pts)

i. H(Z) – H(Z|A)
ii. H(Z|A) – H(Z)
iii. H(Z) – H(Z|B)
iv. H(Z|B) – H(Z)
v. H(Z) – H(Z|AB)
vi. H(Z|AB) – H(Z)
vii. H(Z|AB) – H(Z|A)
viii. H(Z|A) – H(Z|AB)
ix. H(Z|AB) – H(Z|B)
x. H(Z|B) – H(Z|AB)
xi. none of these

(c) True or false? (circle one; 2 pts): H(m), the Shannon entropy of model m, does *not* depend on whether the reference is the top or the bottom.

(d) (2 pts) For the contingency table whose X and Y margins are shown below, write an expression in terms of the constants, a, b, c, d for T_{max}, the maximum value that T(X:Y) could possibly have: T_{max} =

	Y_1	Y_2	
X_1			a
X_2			b
	c	d	

(e) Let I(m) = the *normalized* information for model m, which is 1 for data, m_0, and 0 for independence model, m_{ind}. Write an expression for I(m) in terms of some or all of the quantities $H(m_0)$, H(m), $H(m_{ind})$ and any necessary constants. (2 pts)

I(m) =

(f) Data ABZ is fit with the AZ:BZ model. Is the following statement True or False? (circle one; 2 pts): The AB distribution projected from the calculated ABZ for this AZ:BZ model *can* exhibit some non-zero-strength constraint, i.e., it is possible that, for this projected AB distribution, T(A:B) > 0.

(g) True or false? (circle one; 2 pts): For data m_0 and a model m that is fit to the data, if $df(m) = df(m_0)$, then T(m) = 0.

(h) True or false? (circle one; 2 pts): For data m_0 and models m_j and m_k that are fit to the data, if $df(m_j) = df(m_k)$, then $T(m_j) = T(m_k)$.

2. [26 points] SEARCH, PICKING A BEST MODEL

(a) Where the bottom is the reference and one is searching upwards, suppose one uses two approaches to select models. In approach I, one selects the model with maximum ΔAIC or ΔBIC. In approach II, one uses either α relative to the reference ($\alpha_{cumulative}$) or the αs for all steps from the reference to the model under consideration – call this set of values $\alpha_{incremental}$ – and one selects the highest information model that is statistically significant by one or the other α criterion.

(a1) In approach I, (circle one; 2 pts)

i. using ΔAIC to pick a model lets one go higher than using ΔBIC
ii. using ΔBIC to pick a model lets one go higher than using ΔAIC
iii. these two criteria pick the same model

(a2) In approach II, (circle one; 2 pts)

i. using $\alpha_{cumulative}$ to pick a model lets one go higher than using $\alpha_{incremental}$
ii. using $\alpha_{incremental}$ to pick a model lets one go higher than using $\alpha_{cumulative}$
iii. these two criteria pick the same model

(a3) Suppose when the sample size is small, I select a model that has the highest %correct in the training data. This (third) approach is (circle one; 2 pts)

i. smart because it will produce a high %correct in the test data
ii. not smart because it will usually result in overfitting
iii. equivalent to using $\alpha_{cumulative}$
iv. equivalent to using ΔAIC

(b1) If one has many IVs and wants to select a subset of them that have predictive value with calculations that are fast, one should choose the independence model as a reference and starting model and search upwards considering (circle one; 2 pts)

i. only loopless models ii. only disjoint models
iii. all models (including those with loops)

(b2) Once one has this smaller subset of IVs, suppose that the reduction of uncertainty of the DV, with *all* these IVs in one predicting component, is not statistically significant. Suppose one does not wish to discard any of these IVs. What could one do to try to achieve statistical significance, without losing much predictive power? (circle one; 2 pts)

i. one could go up the lattice from this loopless model & try models with loops
ii. one could go down the lattice from this loopless model & try models with loops

(c) An RA analysis of a 3-variable directed system gives the following results, where the reference model for α^* is AB:Z and the reference model for $\alpha^{\#}$ is AB:AZ:BZ.

Structure	ΔDF	ΔLR	α^*	$\alpha\#$	%ΔdH(Z)
ABZ	8	27.00	0.0007	0.0047	9.10
AB:AZ:BZ#	4	11.98		1.0000	4.04
AB:AZ	2	8.22	0.0164	-	2.77
AB:BZ	2	4.20	0.1227	-	1.41
AB:Z*	0	0.00	1.0000	-	0.00

(c1) Using the provided Chi-square table, the range of values for α^* for model AB:AZ:BZ is from _____ to _____ (fill in *numeric* values; 2 pts).

(c2) What can be said about the relative power of A vs. B to predict Z (circle one; 2 pts)

i. A predicts better than B ii. B predicts better than A
iii. A and B predict equally well iv. neither A nor B predicts Z at all

(c3) Is there a statistically significant (p-value ≤ .05) triadic interaction effect between A, B, and Z? (circle one; 2 pts)

i. no ii. yes

(d) True or false?(circle one, 2 pts): For a directed system model to be useful, its %correct(DV|IV) must be greater than the %correct(DV) for the independence model.

(e) Given the following *neutral* system analysis, with reference = top, let consideration#1 in picking a model be to maximize simplicity subject to the constraint that the model retains at least 3/4 of the information in the data. Let consideration#2 in picking a model be that the probability of a Type I error be in the range advocated in the log-linear book.

#		L^2	Δdf	α	Info
12	ABCD	0.00	0	1.00	1.00
11	ABC:ABD:ACD:BCD	0.00	1	1.00	1.00
10	ABC:ABD:BCD	0.00	2	1.00	1.00
9	ADB:BCD:AC	1.33	3	0.73	1.00
8	BCD:AB:AC:AD	2.45	4	0.66	1.00
7	AB:AC:AD:BC:BD:CD	29.09	5	0.00	0.99
6	AB:AC:BC:BD:CD	74.27	6	0.00	0.97
5	AC:AD:BC:BD	193.25	7	0.00	0.91
4	AC:BC:BD	404.83	8	0.00	0.83
3	BC:BD:A	973.57	9	0.00	0.59
2	BD:A:C	1574.54	10	0.00	0.33
1	A:B:C:D	2366.00	11	0.00	0.00

(e1) Looking at the above output, (2 pts),

i. consideration#1 advocates model # _____

ii. consideration#2 advocates model #_____

(e2) Occam outputs ΔAIC = AIC(reference) – AIC(model). Recall that, for the purposes of the ΔAIC calculation, AIC is a sum of an error term, L^2 and a weighted complexity term, where the weighting factor is 2. Calculate *numerical* values for ΔAIC for the models below, and circle the model favored by the ΔAIC selection criterion. (2 pts)

ΔAIC(ADB:BCD:AC) = _____

ΔAIC(BCD:AB:AC:AD) = _____

(f1) For a *directed* system, where the reference is independence, and α = probability of Type I error, one is usually more concerned with preventing (circle one; 2 pts)

i. Type I errors than Type II errors, & one wants α small (e.g., .05)
ii. Type I errors than Type II errors, & one wants α intermediate (e.g., .10 to .35)
iii. Type I errors than Type II errors, & one wants α large (e.g., .75)
iv. Type I errors than Type II errors, & one wants α very large (e.g., near 1.0)
v. Type II errors than Type I errors, & one wants α small (e.g., .05)
vi. Type II errors than Type I errors, & one wants α intermediate (e.g., .10 to .35)
vii. Type II errors than Type I errors, & one wants α large (e.g., .75)
viii. Type II errors than Type I errors, & one wants α very large (e.g., near 1.0)

(f2) Now suppose, for the directed system, one chooses the top as the reference model, and that while one's *primary* objective is to have a model that adequately fit the data, one does also want a simple model (but not so simple that it unmistakably differs from the data). Follow the guidelines about α values in the log-linear text used in the course. One is usually more concerned with preventing (circle one; 2 pts)

i. Type I errors than Type II errors, & one wants α small (e.g., .05)
ii. Type I errors than Type II errors, & one wants α intermediate (e.g., .10 to .35)
iii. Type I errors than Type II errors, & one wants α large (e.g., .75)
iv. Type I errors than Type II errors, & one wants α very large (e.g., near 1.0)
v. Type II errors than Type I errors, & one wants α small (e.g., .05)
vi. Type II errors than Type I errors, & one wants α intermediate (e.g., .10 to .35)
vii. Type II errors than Type I errors, & one wants α large (e.g., .75)
viii. Type II errors than Type I errors, & one wants α very large (e.g., near 1.0)

3. [16] FITTING A MODEL, USING ITS CONDITIONAL PROBABILITIES

(a) A 'disjoint' model in OCCAM is defined differently for neutral and directed systems. For neutral systems, it is defined as a structure in which no variable appears in more than one component, e.g., ABC:DE:FGHI. For directed systems, it is defined as a structure in which no IV appears in more than one *predicting* component, e.g., ABC:ABZ:CZ.

(a1) Algebraic fitting (getting a q distribution) is possible for (circle one; 2 pts)

i. all disjoint *neutral* models
ii. some disjoint *neutral* models
iii. no disjoint *neutral* models

(a2) Algebraic fitting is possible for (circle one; 2 pts)

i. all disjoint *directed* models
ii. some disjoint *directed* models
iii. no disjoint *directed* models

(b) Consider on the left an *observed* (data) probability table (p) for a *directed* system, with sample size N. None of parameters (a...h) is 0. Let the *calculated* table (q) for model AB:BZ be the table on the right.

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	a	b	e	f
A_2	c	d	g	h

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	q_1	q_2	q_3	q_4
A_2	q_5	q_6	q_7	q_8

(b1) Solve for q_7 *algebraically* for model AB:BZ in terms of a…h. (2 pts)

q_7 =

(b2) *Use IPF* to obtain q_6^{AB} and $q_6^{AB:BZ}$ in terms of parameters a…h. $q_6^{initial} = 1/8$.

(b2.1) After imposing AB, we get (1 pt)

$q_6^{AB} = q_6^{initial} * $ _____

(b2.2) Next, after imposing also BZ, we get (3 pts)

$q_6^{AB:BZ} = q_6^{AB} * $ _____

(c) (2 pts) Given the fit table below for a directed system model of medical data, where Z_0 is healthy and Z_1 is diseased, where $p(Z_0)$ and $p(Z_1)$ are conditional probabilities given the IV state, where Ratio = $p(Z_0)/p(Z_1)$, and where the right-most column is the p-value comparing conditional distributions for each IV state with the margins (in the last row).

A	B	C	N	$p(Z_0)$	$p(Z_1)$	Ratio	p-value
0	0	0	20	.52	.48	1.1	.92
0	0	1	19	.16	.84	0.2	.00
1	0	0	30	.52	.48	1.1	.90
1	0	1	18	.16	.84	0.2	.00
0	1	0	24	.52	.48	1.1	.91
0	1	1	13	.52	.48	1.1	.93
1	1	0	38	.73	.27	2.7	.01
1	1	1	14	.73	.27	2.7	.09
			176	.51	.49		

The most favorable ABC state for a patient to be in is _____

The least favorable ABC state for a patient to be in is _____

(d) If a 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

(e) Suppose that for model m = ABC:AZ:BZ, with binary Z, the *frequency* of (A_i,B_j) is too *small* for differences between $q(Z_0|A_iB_j)$ and $q(Z_1|A_iB_j)$ to be statistically significant. A reasonable strategy for this situation is to (circle one; 2 pts)

i. predict using the rule coming from the independence model, AB:Z
ii. choose a *child* of m with fewer IVs, e.g., ABC:AZ & use the rule for A_i
iii. choose a *parent* of m, e.g., ABC:AZ:BZ:CZ and use a rule from its q
iv. obtain the needed rule from the data, ABZ

4. [2 points only!] INTRO TO FOURIER COMPOSITION (REGRESSION-RA)

	Y_0	Y_1	
X_0	.1	.2	.3
X_1	.3	.4	.7
	.4	.6	

For the above distribution, do composition for X:Y by 'imposing the X and Y margins' in a *new way* as follows: set $q_{X:Y}(x,y) = p(x)/|Y| + p(y)/|X| - K$. This distributes a marginal value of X or a Y equally to all cells that contribute to this marginal value and at the end of the process subtracts the same constant from each cell to make the sum of q's equal 1. This equation *looks like* a regression equation, where p(x) and p(y), multiplied by parameters, *add* and not multiply. (In standard RA, $q_{X:Y}(x,y)$ is p(x) *multiplied* by p(y).)

Fill in the table as follows. Start with cells having zero probabilities (not uniform!). Take each probability in the X margin, divide this probability by the number of Y states, and *add* the result to every cell in the row that contributes to this marginal probability (i.e., add .15 to the x_0y_0 and x_0y_1 cells, and add .35 to the x_1y_0 and x_1y_1 cells). Similarly, add the corresponding contributions for the Y margins. Then subtract some constant, K, from every cell, choosing the constant so that the total calculated probability in the table is 1.

(a) Show the calculations and *the resulting values* in the table below. (1 pt)
(b) What is interesting about these results? (1 pt)

	y_0	y_1
x_0		
x_1		