NAME (PRINT): _____
Last/Surname First /Given Name

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO MISSISSAUGA
APRIL 2017 FINAL EXAMINATION
STA431H5S
Structural Equation Models
Jerry Brunner
Duration - 3 hours
Aids: Calculator Model(s): Any calculator is okay. Formula sheet will be supplied.

The University of Toronto Mississauga and you, as a student, share a commitment to academic integrity. You are reminded that you may be charged with an academic offence for possessing any unauthorized aids during the writing of an exam. Clear, sealable, plastic bags have been provided for all electronic devices with storage, including but not limited to: cell phones, SMART devices, tablets, laptops, calculators, and MP3 players. Please turn off all devices, seal them in the bag provided, and place the bag under your desk for the duration of the examination. You will not be able to touch the bag or its contents until the exam is over.

If, during an exam, any of these items are found on your person or in the area of your desk other than in the clear, sealable, plastic bag, you may be charged with an academic offence. A typical penalty for an academic offence may cause you to fail the course.

*Please note, once this exam has begun, you **CANNOT** re-write it.*

Qn. #	Value	Score
1	10	
2	10	
3	10	
4	35	
5	15	
6	20	
Total = 100 Points		

10 points

1. Let the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent observations from a distribution with $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$, $E(Y_i) = \mu_y$, $Var(Y_i) = \sigma_y^2$ and $Cov(X_i, Y_i) = \sigma_{xy}$. Show that $\widehat{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ is a consistent estimator of σ_{xy}. You have more room than you need.

continued on page 3

10 points

2. Let $\mathbf{D}_1, \ldots, \mathbf{D}_n$ be a random sample from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Writing $\mathbf{D}_i = \begin{pmatrix} \mathbf{X}_i \\ \hline \mathbf{Y}_i \end{pmatrix}$, where $\mathbf{X}_i$ is $p \times 1$ and $\mathbf{Y}_i$ is $q \times 1$, we have $cov(\mathbf{D}_i) = \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \hline \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y \end{array}\right)$, and $\widehat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xy} \\ \hline \widehat{\boldsymbol{\Sigma}}_{yx} & \widehat{\boldsymbol{\Sigma}}_y \end{array}\right)$. We want to test whether the vector of observations $\mathbf{X}_i$ is independent of the vector of observations $\mathbf{Y}_i$. Because zero covariance implies independence for the multivariate normal, the null hypothesis is $H_0 : \boldsymbol{\Sigma}_{xy} = \mathbf{0}$.

   (a) Starting from the formula sheet, write down and simplify the likelihood evaluated at the unrestricted MLE.

   (b) Using the fact that zero covariance implies independence for the multivariate normal, give the likelihood evaluated at the restricted MLE — that is, at the MLE restricted by $H_0 : \boldsymbol{\Sigma}_{xy} = \mathbf{0}$. If you think about it you can just write down the answer without any calculation.

continued on page 4

(c) Calculate and simplify the large-sample likelihood ratio statistic G^2 for testing $H_0 : \boldsymbol{\Sigma}_{xy} = \mathbf{0}$, which is equivalent to $\mathbf{X}_i$ and $\mathbf{Y}_i$ independent. Your answer is a formula. **Circle your final answer**.

(d) What are the degrees of freedom of the test?

continued on page 5

10 points

3. Two latent explanatory variables X_1 and X_2 (say motivation and ability) potentially have non-zero covariance with one another. Six observable job performance measures Y_1 through Y_6 are related to X_1 and X_2 as follows:

$$\begin{aligned} Y_1 &= \beta_{11}X_1 + \beta_{12}X_2 + \epsilon_1 \\ Y_2 &= \beta_{21}X_1 + \beta_{22}X_2 + \epsilon_2 \\ Y_3 &= \beta_{31}X_1 + \beta_{32}X_2 + \epsilon_3 \\ Y_4 &= \beta_{41}X_1 + \beta_{42}X_2 + \epsilon_4 \\ Y_5 &= \beta_{51}X_1 + \beta_{52}X_2 + \epsilon_5 \\ Y_6 &= \beta_{61}X_1 + \beta_{62}X_2 + \epsilon_6, \end{aligned}$$

where the β_{ij} quantities are unknown parameters, and ϵ_1 through ϵ_6 are independent of one another and independent of X_1 and X_2. The model has been centered and re-parameterized so that $Var(X_1) = Var(X_2) = 1$, making ϕ_{12} a correlation.

(a) Does this model pass the test of the parameter count rule? Answer Yes or No and give the numbers.

(b) You start to wonder whether the parameters are identifiable, and then you realize that while this is expressed as a regression model, it's really unrestricted factor analysis. The fact that there are exactly two factors and six observable variables just gets in the way. Accordingly, you write $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i$.

i. What is $\boldsymbol{\Sigma} = cov(\mathbf{Y}_i)$? Give your answer in matrix notation, *not* specific to the example above. You have a lot more room than you need.

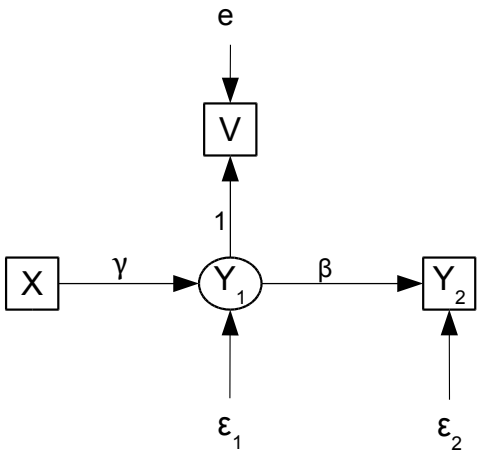
continued on page 6

ii. Give two different parameter sets $(\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ and $(\boldsymbol{\beta}', \boldsymbol{\Phi}', \boldsymbol{\Psi}')$ that yield the same $\boldsymbol{\Sigma}$. An easy way to do this is to let $\boldsymbol{\Phi} = \mathbf{I}$, and let $\mathbf{R}$ be an arbitrary orthogonal (rotation) matrix satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. You have a lot more room than you need.

continued on page 7

35 points

4. In a study of the determinants of healthy food choices, young parents were asked to refrain from eating for a certain number of hours before coming to the study. Then they rated how hungry they were (as part of a long questionnaire) and also put together some menus for small children. The question was how hunger in an adult could affect affect their choice of healthy food for children. In the path diagram below, the symbol on the arrow from X to Y_1 might look like the letter Y, but actually it's the Greek letter γ (gamma).



X is food deprivation time.
Y_1 is true hunger.
V is reported hunger.
Y_2 is how nutritious the chosen menus were.

with $Var(X) = \phi$, $Var(e) = \omega$, $Var(\epsilon_j) = \psi_j$.

(a) Give the centered model equations in scalar form.

(b) List the parameters that appear in the covariance matrix of the observable data.

(c) Does this model pass the test of the parameter count rule? Answer Yes or No and give the numbers.

(d) Why is it reasonable to assume $\gamma > 0$?

continued on page 8

(e) Write $\boldsymbol{\Sigma}$ (the covariance matrix of an observable data vector $\mathbf{D}_i$) in terms of the model parameters.

(f) Give a Method of Moments estimator of β. Don't forget the hats!

(g) Prove that the parameter ω is identifiable by writing it explicitly in terms of σ_{ij} quantities. The answer is a formula. Show a little work and **circle your final answer.**

continued on page 9

(h) You may take it for granted that it s possible to solve uniquely for all the parameters provided $\gamma \neq 0$ and $\beta \neq 0$. You don't have to do the work. How do you know that your Method of Moments estimator of β is also the Maximum Likelihood estimator?

(i) Suppose the null hypothesis $\beta = 0$ is true. Is the parameter β still identifiable? Answer Yes or No and say how you know.

(j) If the null hypothesis $\beta = 0$ is true, how do you know that the remaining parameters cannot all be identifiable (except possibly on a set of volume zero in the restricted parmeter space; that was a hint.)

(k) If you tried to carry out a likelihood ratio test of $H_0 : \beta = 0$, why would you expect numerical trouble?

(l) Even though $H_0 : \beta = 0$ has just one equals sign, I am attracted to a two-degree-of-freedom test. Give my null hypothesis in terms of σ_{ij} quantities. No explanation is needed.

continued on page 10

15 points

5. Volunteers of legal drinking age come to the Alcohol Lab and fill out a health questionnaire, which includes a question about their weight. Then they have a brief physical exam including a measurement of their body weight on a scale. Next, they are given a randomly assigned amount of alcohol to drink. They all drink the same volume of liquid, but the concentration of alcohol varies and it's mixed with a very sweet juice cocktail so presumably it's hard for the subjects to judge how strong it is.

   At this point we have three observable variables: two independent assessments of body weight, and amount of alcohol consumed, which is measured without error.

   Then the subject has two breath tests of blood alcohol level. The subject is seated in a chair and stays there without speaking while two different technicians come in and take measurements. That way they do not see how steady the subject is on his or her feet, and they do not hear how slurred his or her speech might be. Finally, a nurse comes in and takes a blood sample. Now we have three independent measurements of true blood alcohol level.

   Finally, the subject fills out some more questionnaires, including a racism questionnaire and a sexism questionnaire. There are undoubtedly quite a few latent variables operating here including true racism and true sexism (possibly correlated), but we will not model this complicated and interesting process. Instead, we'll run straight arrows directly from blood alcohol level to the questionnaire responses, treating all the complicated processes as unspecified omitted variables that are part of the error terms.

   (a) Make a path diagram. Put coefficients on all the arrows *except* the ones coming from error terms. You will fix some of the factor loadings to one in order to obtain identifiability. Make it reasonable, and in this case write the number 1 on the arrows.

   (b) How do you know that the parameters of the latent variable model are identifiable? Just cite the letter and number of the rule or give its name; no discussion or explanation is necessary.

continued on page 11

20 points

6. In the Arthritis data, rheumatoid arthritis patients were assessed for disease severity, and then for pain and exercise/physical activity at two time points. First comes `proc means` output, largely useful to remind you of what the variables are, and then `proc calis` code and output. The questions are on last page of this exam.

Exercise and Arthritis Pain: STA431s17 Final Exam

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
clinical	Disease severity based on clinical assessment	494	6.0740283	1.2440408	2.2000000	9.8400000
xray	Disease severity based on x-ray	494	10.0627530	2.1791259	4.0000000	16.0000000
blood	Disease severity based on blood test	494	40.6194332	13.3820786	3.0000000	80.0000000
selfpain1	Self-reported pain at time one	494	6.0526316	1.1540840	2.0000000	10.0000000
EEG1	Pain assessed from brain waves at time one	494	50.6989879	9.1847534	24.1000000	75.9000000
selfexer1	Self-reported exercise/physical activity at time one	494	5.9291498	1.2305445	2.0000000	9.0000000
spouseexer1	Spouse report of exercise/physical activity at time one	494	5.9838057	1.2200742	3.0000000	10.0000000
acceler1	Accelerometer (fitness tracker) data at time one	494	14.8943320	4.4283991	0.6000000	30.0000000
selfpain2	Self-reported pain at time two	494	6.0607287	1.1934389	2.0000000	9.0000000
EEG2	Pain assessed from brain waves at time two	494	50.6325911	9.3300237	24.1000000	82.8000000
selfexer2	Self-reported exercise/physical activity at time two	494	5.9858300	1.1838165	3.0000000	9.0000000
spouseexer2	Spouse report of exercise/physical activity at time two	494	5.9352227	1.2033819	3.0000000	9.0000000
acceler2	Accelerometer (fitness tracker) data at time two	494	14.7961538	4.1011606	3.6000000	29.9000000

```
proc calis pshort nostand vardef=n;
     fitindex on(only) = [chisq df probchi];
     var clinical xray blood selfpain1 EEG1 selfexer1 spouseexer1 acceler1
                           selfpain2 EEG2 selfexer2 spouseexer2 acceler2;
     lineqs Fpain1 = gamma*Fseverity + epsilon1,
            Fexer1 = beta1*Fpain1 + epsilon2,
            Fpain2 = beta2*Fexer1 + gamma*Fseverity + epsilon3,
            Fexer2 = beta1*Fpain2 + epsilon4,
            clinical    =         Fseverity + e1,
            xray        = lambda2*Fseverity + e2,
            blood       = lambda3*Fseverity + e3,
            selfpain1   =         Fpain1 + e4,
            EEG1        = lambda5*Fpain1 + e5,
            selfexer1   =         Fexer1 + e6,
            spouseexer1 = lambda7*Fexer1 + e7,
            acceler1    = lambda8*Fexer1 + e8,
            selfpain2   =         Fpain2 + e9,
            EEG2        = lambda5*Fpain2 + e10,
            selfexer2   =         Fexer2 + e11,
            spouseexer2 = lambda7*Fexer2 + e12,
            acceler2    = lambda8*Fexer2 + e13;
     variance Fseverity = phi, epsilon1-epsilon4 = psi1-psi4,
              e1-e13 = omega01-omega13;
     cov epsilon1 epsilon3 = psi13, epsilon2 epsilon4 = psi24;
     bounds phi psi1-psi4 omega01-omega13 > 0;
```

continued on page 12

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Fit Summary	
Chi-Square	71.3955
Chi-Square DF	63
Pr > Chi-Square	0.2189

Effects in Linear Equations

Variable	Predictor	Parameter	Estimate	Standard Error	t Value	Pr > |t|
Fpain1	Fseverity	gamma	0.54990	0.04169	13.1891	<.0001
Fexer1	Fpain1	beta1	-1.03143	0.05901	-17.4781	<.0001
Fpain2	Fexer1	beta2	-0.02037	0.04317	-0.4718	0.6371
Fpain2	Fseverity	gamma	0.54990	0.04169	13.1891	<.0001
Fexer2	Fpain2	beta1	-1.03143	0.05901	-17.4781	<.0001
clinical	Fseverity		1.00000			
xray	Fseverity	lambda2	1.94410	0.08030	24.2101	<.0001
blood	Fseverity	lambda3	12.17585	0.49276	24.7097	<.0001
selfpain1	Fpain1		1.00000			
EEG1	Fpain1	lambda5	7.90009	0.42289	18.6813	<.0001
selfexer1	Fexer1		1.00000			
spouseexer1	Fexer1	lambda7	0.96672	0.03759	25.7162	<.0001
acceler1	Fexer1	lambda8	3.98701	0.12760	31.2465	<.0001
selfpain2	Fpain2		1.00000			
EEG2	Fpain2	lambda5	7.90009	0.42289	18.6813	<.0001
selfexer2	Fexer2		1.00000			
spouseexer2	Fexer2	lambda7	0.96672	0.03759	25.7162	<.0001
acceler2	Fexer2	lambda8	3.98701	0.12760	31.2465	<.0001

continued on page 13

Estimates for Variances of Exogenous Variables						
Variable Type	**Variable**	**Parameter**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Latent	**Fseverity**	phi	1.02412	0.09501	10.7790	<.0001
Disturbance	**epsilon1**	psi1	0.38311	0.04430	8.6476	<.0001
	epsilon2	psi2	0.24060	0.04063	5.9222	<.0001
	epsilon3	psi3	0.26804	0.04385	6.1124	<.0001
	epsilon4	psi4	0.19855	0.03592	5.5272	<.0001
Error	**e1**	omega01	0.52038	0.03952	13.1674	<.0001
	e2	omega02	0.86826	0.09299	9.3367	<.0001
	e3	omega03	26.88985	3.37477	7.9679	<.0001
	e4	omega04	0.68055	0.05306	12.8269	<.0001
	e5	omega05	43.56905	3.37478	12.9102	<.0001
	e6	omega06	0.54776	0.04153	13.1884	<.0001
	e7	omega07	0.62657	0.04583	13.6704	<.0001
	e8	omega08	3.78539	0.39482	9.5877	<.0001
	e9	omega09	0.78136	0.05708	13.6888	<.0001
	e10	omega10	47.20638	3.46789	13.6124	<.0001
	e11	omega11	0.54800	0.04081	13.4295	<.0001
	e12	omega12	0.61019	0.04418	13.8125	<.0001
	e13	omega13	3.73043	0.37532	9.9392	<.0001

Covariances Among Exogenous Variables						
Var1	**Var2**	**Parameter**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
epsilon1	**epsilon3**	psi13	0.26409	0.03748	7.0463	<.0001
epsilon2	**epsilon4**	psi24	0.20628	0.03032	6.8039	<.0001

continued on page 14

(a) We want to know whether pain at Time One affects exercise at Time One.

i.

H_0 in symbols or names from the printout	t Statistic	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

ii. In plain, non-statistical language, what do you conclude?

(b) We want to know whether pain at Time Two affects exercise at Time Two.

i.

H_0 in symbols or names from the printout	t Statistic	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

ii. In plain, non-statistical language, what do you conclude?

(c) We want to know whether exercise at Time One affects pain at Time Two.

i.

H_0 in symbols or names from the printout	t Statistic	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

ii. In plain, non-statistical language, what do you conclude?

(d) What is the estimated reliability of clinical assessment as a measure of disease severity? The answer is a number. Show a little work and **circle your answer**.

(e) What is the estimated reliability of the blood test as a measure of disease severity? The answer is a number. Show a little work and **circle your answer**.