Last Name: *First Name:* *Student Number:*

UNIVERSITY OF TORONTO
Faculty of Arts and Science

DECEMBER EXAMINATIONS 2008

STA 437H1 F (plus STA 1005)

Duration - 3 hours

1	
2	
3	
4	
5	
6	
7	
T	

No books, notes, or calculators are allowed. A sheet of formulas and a table for the F distribution are included at the back of this exam.

Answer all questions in the space provided; if you run out of space, use the back of a page, indicating where the answer continues.

For all questions where the answer is a number, you must provide an actual numerical answer (eg, 1.5 or 3/2), not just a formula that could be evaluated to give this number.

Except where noted, you must explain how you obtained your answer to obtain full credit.

The seven questions are worth equal amounts.

1. Prove each of the following statements.

   a) Prove that if $\mathbf{e}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_A, and $\mathbf{e}$ is also an eigenvector of $\mathbf{B}$ with eigenvalue λ_B, then $\mathbf{e}$ is an eigenvector of $\mathbf{A} + 2\mathbf{B}$. Also, find what eigenvalue $\mathbf{e}$ has as an eigenvector of $\mathbf{A} + 2\mathbf{B}$.

   b) Prove that if $\mathbf{e}_1$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, and $\mathbf{e}_2$ is also an eigenvector of $\mathbf{A}$ with eigenvalue λ, then $\mathbf{e}_1 - \mathbf{e}_2$ is an eigenvector of $\mathbf{A}$. Also, find what eigenvalue is associated with $\mathbf{e}_1 - \mathbf{e}_2$.

   c) Let $\mathbf{B}$ be a $p \times k$ matrix, and c be a scalar. Prove that if $\mathbf{e}$ is an eigenvector of $\mathbf{B}'\mathbf{B}$ with eigenvalue λ, then $\mathbf{Be}$ is an eigenvector of $\mathbf{A} = \mathbf{BB}' + c\mathbf{I}$, where $\mathbf{I}$ is the $p \times p$ identity matrix. Also, find the eigenvalue that $\mathbf{Be}$ has as an eigenvector of $\mathbf{A}$.

2. The distribution of a vector $\mathbf{X}$ of length $p = 4$ is given by a factor analysis model with $k = 1$ common factors, in which the true values of the parameters are mean $\mu = \mathbf{0}$, covariance of specific factors $\psi = 4\,\mathbf{I}$, and loadings matrix $\mathbf{L} = [\,3\ 2\ 1\ 0\,]'$.

   a) Give another value for the loadings matrix, $\mathbf{L}$, that will produce the same distribution for $\mathbf{X}$. No explanation is required.

   b) Find the covariance matrix of $\mathbf{X}$.

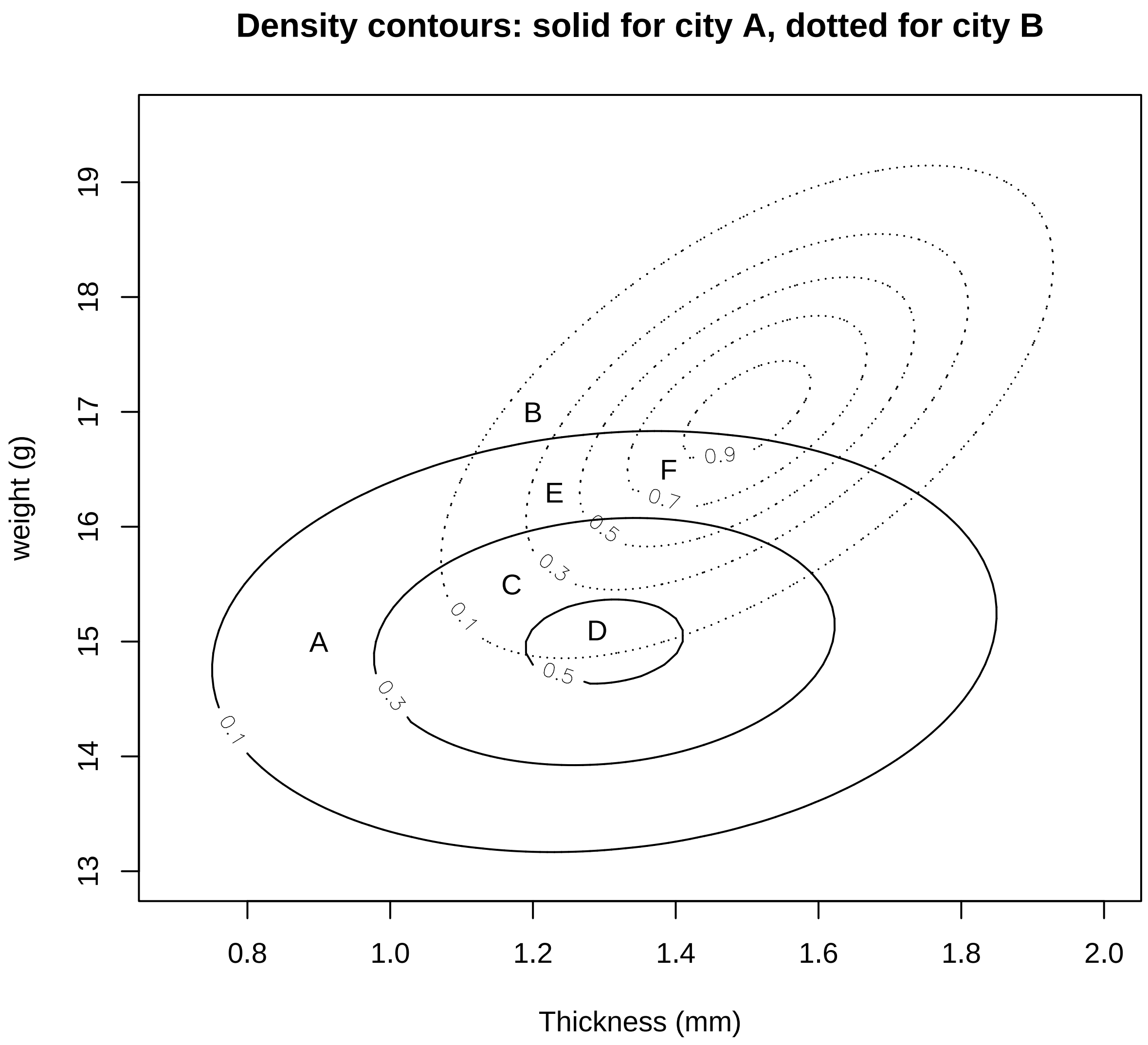
   c) Find the first principal component direction of the covariance matrix for $\mathbf{X}$ (any vector pointing in that direction will do) and the variance in that direction.

d) Find the conditional distribution of the common factor given that we observe that the first component of $\mathbf{X}$ is -2. Assume that the remaining three components of $\mathbf{X}$ are not observed, and that we know the true values of the model parameters.

3. A researcher is studying the use of ancient Roman coins as a way of tracking patterns of trade in the Roman empire. It is known that a certain type of coin was made by mints located in two distant cities, which we'll call A and B. Samples of 200 of these coins that were made in city A and of 200 of these coins that were made in city B were obtained. For each coin in these samples, the thickness and the weight were measured. Using this data, the researcher hopes to classify other coins of this type, as either being made in city A or in city B.

   The data in both samples appears to be normally distributed. The sample mean vectors and sample covariance matrices for the two samples were computed. Using these, contour plots of the estimated probability density function for the thickness and weight for coins from city A and from city B were produced. These plots are shown superimposed on the next page, with the plot for city A using solid lines, and the plot for city B using dotted lines.

   Six points, labelled A to F, are also shown in the plot. You are to decide how a coin whose measured thickness and weight are at these points should be classified (as being from city A or city B). You should classify a coin according to which city it is more likely to have been made in. Costs of misclassifying coins as coming from city A or city B are assumed to be equal. However, depending on where a coin was found, the chance of it coming from city A or city B (before considering the measurements on the coin) may vary, as described in parts (a) and (b) on the next page.



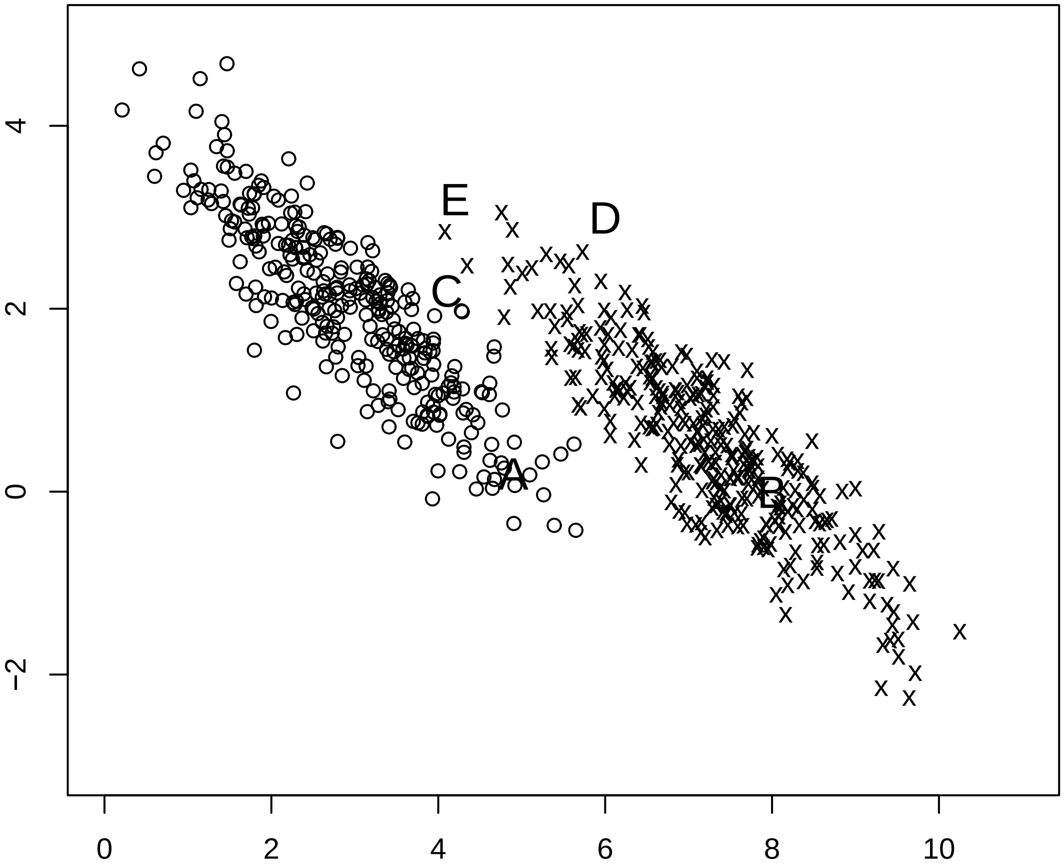
a) How would you classify a coin whose measurements lie at each labelled point if city A and city B are equally likely to be the source of the coin, and mis-classification costs are equal? Write A or B for each point, or write U if you can't tell clearly from the information shown. No explanation is required.

A: B: C: D: E: F:

b) How would you classify a coin whose measurements lie at each labelled point if city A is twice as likely as city B to be the source of the coin, and mis-classification costs are equal? Write A or B for each point, or write U if you can't tell clearly from the information shown. No explanation is required.

A: B: C: D: E: F:

4. The scatterplot below shows the values of two variables for 300 observations from each of two classes (600 total). The observations are marked by class — either, O or X:



a) Five points are marked with the labels A, B, C, D, and E. If the two classes occur equally often, and the cost of misclassification is the same for both classes, how would you classify each of these five points? Write O or X for each point, or write U if you can't tell clearly from the information shown. No explanation is required.

A: B: C: D: E:

b) Suppose that we reduced these observations from two variables to a single variable by projecting them on the first principal component direction, as found using all 600 observations, from both classes. How well would you expect to be able to classify a new observation on the basis of its projection on this first principal component? Would it work just as well as with the original variables? Or almost as well? Or much worse? Or perhaps classification would not be possible at all? Explain.

5. Suppose that we are interested in whether the kind of food that people in Toronto eat on Fridays is different from the kind of food that people in Toronto eat on Sundays. To investigate this, we measure the calories, fat content, and sugar content of everything that 100 randomly-selected people from Toronto eat on one Friday, and we also measure the same things for what 100 people eat on one Sunday. (Of course, having data from more than one Friday and more than one Sunday would be better, but don't worry about that.)

   We want to test the null hypothesis that there is no difference in the mean calories, fat content, and sugar content of food eaten on this Friday and food eaten this Sunday. Two hypothesis tests might be considered for this — the one-sample T^2 test based on the 100 differences between measurements of what is eaten on Friday and what is eaten Sunday, and the two-sample T^2 test using a pooled covariance estimate.

   (a) Suppose that the 100 people whose food we tested on this Friday are the *same* as the 100 people whose food we tested on this Sunday. Which of the two tests would be appropriate in this situation? What assumptions must hold for this to be an exactly valid test? What (if anything) might go wrong if the other test was used instead?

(b) Suppose that the 100 people whose food we tested on this Friday are *different* from the 100 people whose food we tested on this Sunday. Which of the two tests would be appropriate in this situation? What assumptions must hold for this to be an exactly valid test? What (if anything) might go wrong if the other test was used instead?

6. An experiment was conducted to determine whether protein and fibre content for wheat grown with fertilizer 1 is different from that for wheat grown with fertilizer 2. Wheat was grown in 22 plots. On 11 of these plots, fertilizer A was used; on the other 11 plots, fertilizer B was used. The protein and fibre content (in percent) of the wheat from each plot was measured. The sample mean vectors and sample covariance matrices from these measurements were as follows (subscript indicates fertilizer 1 or fertilizer 2):

$$\overline{\mathbf{X}}_1 = [12.1 \;\; 14.3]', \qquad \overline{\mathbf{X}}_2 = [10.1 \;\; 13.3]'$$

$$\mathbf{S}_1 = \begin{bmatrix} 2.2 & -1.1 \\ -1.1 & 0.9 \end{bmatrix}, \qquad \mathbf{S}_2 = \begin{bmatrix} 2.3 & -1.0 \\ -1.0 & 1.1 \end{bmatrix}$$

a) Find the pooled estimate of the covariance matrix for this data.

b) Does it seem appropriate to use the two-sample T^2 test with this pooled covariance estimate to test the null hypothesis that the mean protein and fibre content is the same for both fertilizers? Explain why or why not.

c) Suppose that we decide to use the two-sample T^2 test with pooled covariance estimate. Will we reject the null hypothesis of no diffence at the 0.01 level?

Note: The inverse of the pooled covariance estimate is

$$\begin{bmatrix} 0.87 & 0.92 \\ 0.92 & 1.96 \end{bmatrix}$$

7. Say whether each of the following is true or false. No explanation is required.

(a) If A and B are both $k \times k$ symmetric positive definite matrices, then $A + B$ is also a symmetric positive definite matrix.

(b) When principal component analysis is done, it makes no difference whether the eigenvectors of the covariance matrix or of the correlation matrix are found — the results are essentially the same.

(c) When factor analysis is done, it makes no difference whether the covariance matrix or the correlation matrix is modeled as having the form $\mathbf{LL}' + \psi$ — the results are essentially the same.

(d) When finding a confidence ellipse for the mean of a multivariate normal distribution using the T^2 statistic, it is necessary to use the χ^2 distribution when $n - p$ is large; the F distribution should not be used in this situation.

(e) If random variables X and Y are uncorrelated, and each has a normal distribution, then the joint distribution of (X, Y) must be multivariate normal.

(f) If random variables X and Y are independent, and both have finite variance, their correlation must be zero.

(g) When n is large, the Central Limit Theorem guarantees that the distribution of the sample mean of $X_1, \ldots, X_n$ will be close to normal even if $X_1, \ldots, X_n$ are not independent, provided that the variances of the X_i are finite.

(h) Any non-zero vector of length k is an eigenvector of the $k \times k$ identity matrix.

(i) When testing for equality of all means in a multivariate analysis of variance model, it is necessary to start by standardizing all the variables to have mean zero and standard deviation one, so that the choice of units for the variables will not affect the results.

(j) If you find simultaneous 95% confidence intervals for the 20 means of a multivariate normal distribution for 20 variables, using the Bonferroni correction, the intervals you obtain from a sample of size 1000 will be approximately 10 times larger than the intervals obtained using a sample of size 10000.

Formulas for possible use in this exam

Here are some of the formulas relating to the material we covered. These formulas might or might not be relevant to some of the questions on the final exam. Note that this sheet just has the formulas, not the full statements of theorems that these formulas may be part of.

Sample covariance:

$$\mathbf{S} \;=\; \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j-\overline{\mathbf{X}})\,(\mathbf{X}_j-\overline{\mathbf{X}})'$$

Covariance of a random vector:

$$\mathrm{Cov}(\mathbf{X}) \;=\; E\big[(\mathbf{X}-E(\mathbf{X}))\,(\mathbf{X}-E(\mathbf{X}))'\big]$$

Covariance of transformed random vector:

$$\mathrm{Cov}(\mathbf{CX}) \;=\; \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$$

Probability density function for multivariate normal:

$$f(\mathbf{x}) \;=\; (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp(-(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)/2)$$

Conditional mean and covariance for multivariate normal:

$$\begin{aligned}\text{Mean of } \mathbf{X}_1 \text{ given } \mathbf{X}_2=\mathbf{x}_2 \;&=\; \mu_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\mu_2)\\ \text{Covariance of } \mathbf{X}_1 \text{ given } \mathbf{X}_2=\mathbf{x}_2 \;&=\; \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{aligned}$$

T^2 statistic for one sample:

$$T^2 \;=\; n(\overline{\mathbf{X}}-\mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}}-\mu_0)$$

The distribution of T^2 under the null hypothesis that $\mu=\mu_0$ is $[(n–1)p/(n–p)]F_{p,n-p}$, which is approximately χ^2_p when $n-p$ and n/p are both large.

T^2 statistic for two samples, using pooled covariance estimate:

$$T^2 \;=\; ((\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2)\,-\,\delta_0)'\;[(1/n_1+1/n_2)\,\mathbf{S}_{\text{pooled}}]^{-1}\;((\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2)\,-\,\delta_0)$$

Here, $\mathbf{S}_{\text{pooled}} \;=\; ((n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2)\;/\;(n_1+n_2-2)$. The distribution of T^2 is $[(n_1{+}n_2{–}2)p/(n_1{+}n_2{–}p{–}1)]F_{p,n_1+n_2-p-1}$ under the null hypothesis that $\mu_1-\mu_2=\delta_0$. This distribution is approximately χ^2_p when n_1+n_2-p and $(n_1+n_2)\,/\,p$ are both large.

The factor analysis model:

$$\mathbf{X} \;=\; \mu\;+\;\mathbf{LF}\;+\;\epsilon$$

where $\mathbf{F}\sim N(0,\mathbf{I})$ and $\epsilon\sim N(0,\psi)$, with ψ diagonal. $\mathbf{F}$ and ϵ are independent.