Stat 407 Exam 1 (Fall 2012) Name _____

Be sure to show your working, explain your answers, ask the instructor if you don't understand a question. GOOD LUCK!

1. *(3 pts)* Fill in the blanks.

   _____ _____ involves a collection of methods for studying data where there are dependencies between multiple variables. It includes methods such as principal component analysis, factor analysis, discriminant analysis, cluster analysis, MANOVA.

2. *(9 pts)* For the following data set, answer these questions:

School	Type	SAT	Acceptance	StudentDoll	Top.10.Pct	PctPhD	GradPct
Smith	Lib Arts	1195	57	25271	65	90	87
UCLA	Univ	1142	43	26859	96	100	61
Colgate	Lib Arts	1258	38	17520	61	78	85
Williams	Lib Arts	1336	28	23772	86	90	93
Middlebury	Lib Arts	1255	25	24718	65	89	92
Brown	Univ	1281	24	24201	80	98	90
Claremont McKenna	Lib Arts	1260	36	20377	68	94	74
U of Rochester	Univ	1155	56	38597	52	96	73
Wesleyan (CT)	Lib Arts	1290	35	19948	73	87	91
Washington U (MO)	Univ	1225	54	39883	71	98	76
Occidental	Lib Arts	1170	49	20192	54	93	72
Swarthmore	Lib Arts	1310	24	27487	78	93	88

   (a) *(2)* How many cases are there? _____

   (b) *(2)* What is the **dimension** of the data (how many real-valued variables are there)? _____

   (c) *(3)* Is this possibly the sample mean? Explain you answer.
   $\bar{X} = 5203.85$

   _____ _____

   (d) *(2)* The correlation matrix for the data is below. Which two variables have the strongest correlation, either positive or negative? _____ _____

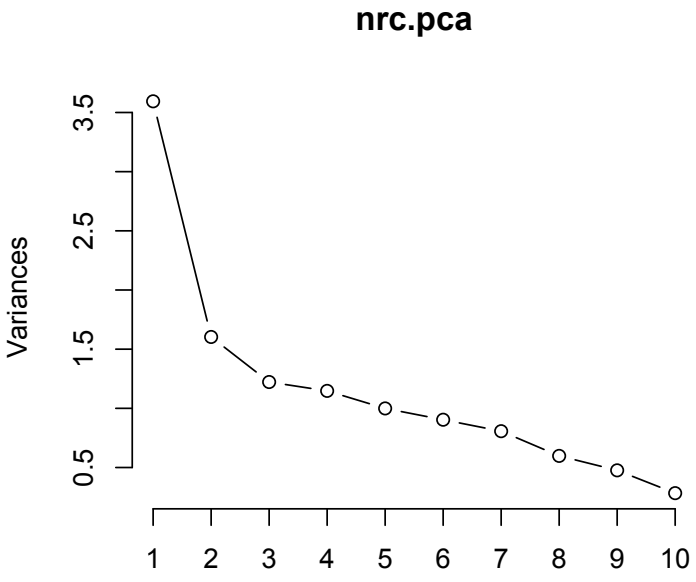
	SAT	Acceptance	StudentDoll	Top.10.Pct	PctPhD	GradPct
SAT	1.00	-0.78	-0.35	0.30	-0.38	0.80
Acceptance	-0.78	1.00	0.47	-0.46	0.18	-0.58
StudentDoll	-0.35	0.47	1.00	-0.06	0.59	-0.30
Top.10.Pct	0.30	-0.46	-0.06	1.00	0.35	0.01
PctPhD	-0.38	0.18	0.59	0.35	1.00	-0.53
GradPct	0.80	-0.58	-0.30	0.01	-0.53	1.00

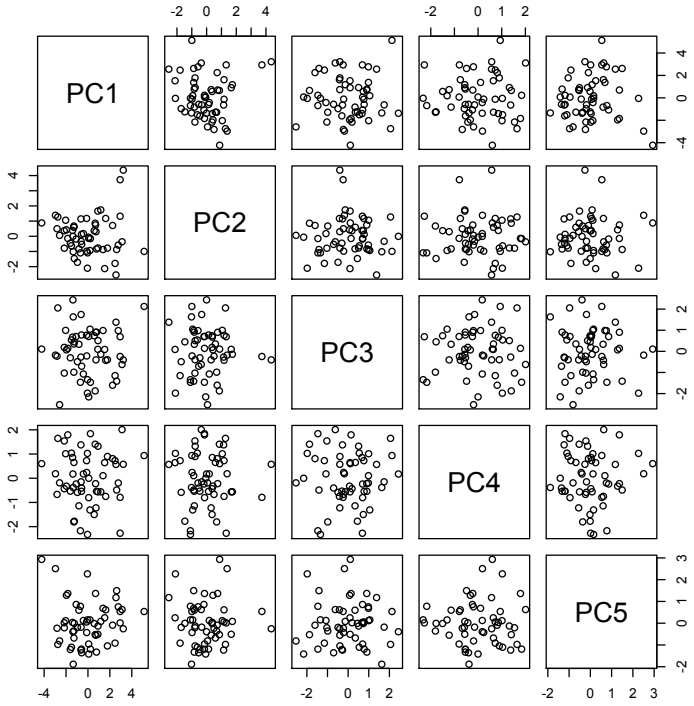
3. *(5 pts)* Match the definitions to the terms:

A. Factor analysis	a. The data multiplied by a vector $\mathbf{a}_{p\times 1}$ that takes the form: $\mathbf{Xa} = [\alpha_1 X_{11} + \alpha_2 X_{21} + \ldots + \alpha_p X_{p1} \quad \ldots \quad \alpha_1 X_{1n} + \alpha_2 X_{2n} + \ldots + \alpha_p X_{pn}]_{n\times 1}$ where $\|\mathbf{a}\| = \sqrt{\alpha_1^2 + \ldots + \alpha_p^2} = 1$.
B. Summary statistics for multivariate data	b. Sample mean vector, correlation matrix, variance-covariance matrix.
C. Data projection	c. For two data points $\mathbf{A}, \mathbf{B}$, $d(\mathbf{A}, \mathbf{B}) = \sqrt{(\mathbf{A}-\mathbf{B})'\mathbf{S}^{-1}(\mathbf{A}-\mathbf{B})}$
D. Mahalanobis distance	d. The direction of maximum variance in the data.
E. First Eigenvector	e. When its not possible to observe the variables of interest directly, measure whats possible and create the variables of interest from the observed data.

4. *(12 pts)* The next question examines principal component analysis for data collected by the National Research Council to compare the performance of statistics graduate programs across the USA. The principal components was computed on the correlation matrix, for 10 variables of the ranking data.

	PC1	PC2	PC3	PC4	PC5
AvNumPubs	0.38	-0.06	0.16	-0.04	0.24
AvCitations	0.38	-0.24	0.17	0.23	0.13
PctFacultywGrants	0.35	-0.08	0.35	0.10	-0.13
NegMedianTimetoDegree	-0.04	0.05	0.25	0.60	-0.52
FemaleFaculty	-0.11	0.25	-0.23	0.23	0.60
FemaleStudents	-0.12	0.42	-0.34	0.18	-0.27
InternatStudents	0.11	-0.20	-0.41	-0.44	-0.45
AvNumPhDs	0.36	0.47	0.05	-0.08	-0.06
PctInterdiscFaculty	0.19	-0.31	-0.52	0.40	0.08
AvGRE	0.32	-0.30	-0.14	-0.15	-0.00
TotalFaculty	0.40	0.17	-0.35	0.25	-0.06
NumStudents	0.35	0.47	0.02	-0.24	-0.01
Variance	3.59	1.60	1.22	1.15	1.00
Cum %					





(a) *(2)* What is the total variance? _____

(b) *(2)* Fill in the last row of the PCA summary.

(c) *(2* **Based on the scree plot only** how many PCs would be suggested? _____

(d) *(2)* Which **three** variables contribute the most to first principal component? _____ _____ _____

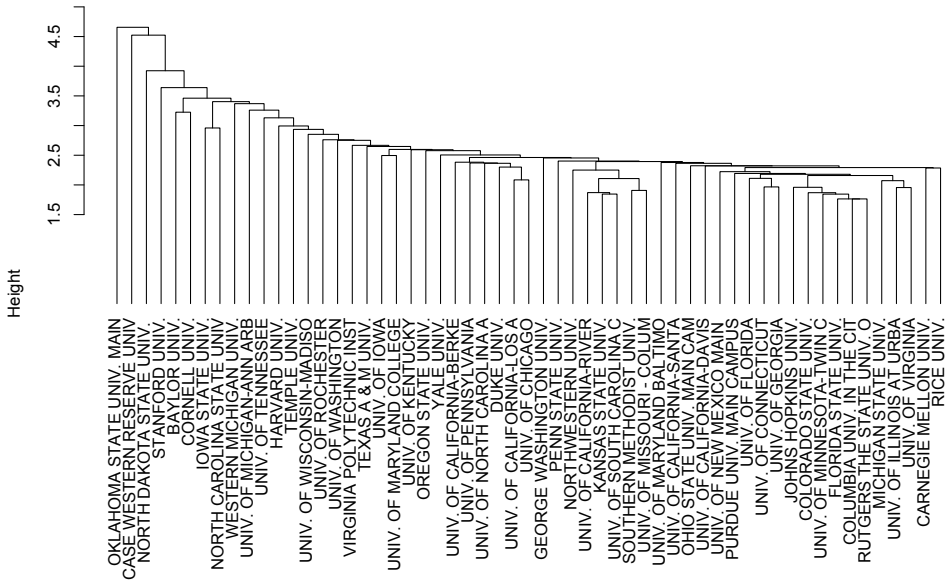
(e) *(2)* Based on the eigenvector coefficients, using your own words, explain how PC1 differs from PC2.

(f) *(2)* On the scatterplot matrix of the first five PCs, mark anything that suggests there were problems with the PCA. If you don't see anything simply write this.

5. *(6 pts)* In this question, we are doing a cluster analysis of the NRC statistics graduate program ratings data, used in the previous question.

   (a) *(2)* There are 57 statistics departments in the data set. What is the size of the interpoint distance matrix? _____

   (b) *(2)* The dendrogram below was produced by hierarchical clustering of the NRC data. What linkage method do you think was used? Explain why you think this.



   _____

   (c) *(2)* These are the results comparing the 4 cluster solutions from two different clusterings. What would be the proportion of cases where the two methods **disagree**?

	Method 2			
Method 1	1	2	3	4
1	4	11	0	0
2	0	23	0	0
3	0	1	11	0
4	0	1	0	6

6. *(10 pts)* For the following data, we will conduct a k-means cluster analysis using Euclidean distance with $k = 2$. The two seed means are $\bar{\mathbf{X}}_1^0 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$, $\bar{\mathbf{X}}_2^0 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$.

Obs Num	Var 1	Var 2	Dist from $\bar{\mathbf{X}}_1^0$	Dist from $\bar{\mathbf{X}}_2^0$	Cluster id
1	0	2	3.2		
2	1	1	2.8		
3	3	2	1.0	2.8	1
4	-1	0		4.5	
5	4	5		3.2	

(a) *(2)* Plot the data, and the two initial means.

(b) *(2)* Now using Euclidean distance compute the distances from each point to each mean, and fill in the blanks of the relevant columns of the table above.

(c) *(2)* Assign the 5 points into one of the two clusters, and write this into the table above.

(d) *(2)* Circle the two clusters on your plot.

(e) *(2)* Compute the two means that will seed the next iteration.

$$\bar{\mathbf{X}}_1^1 = \qquad\qquad \bar{\mathbf{X}}_2^1 =$$

7. *(5 pts)* EXTRA CREDIT: Find the word associated with the phrase in this word search (words can be up, down, right, left or diagonal), and fill in the blanks.

s	r	e	t	s	u	l	c
u	m	v	n	n	r	i	o
p	f	d	c	o	l	n	n
e	t	o	r	i	a	k	f
r	n	r	k	s	m	a	u
v	e	t	m	n	o	g	s
i	t	g	e	e	k	e	i
s	a	r	a	m	u	f	o
e	l	d	n	i	c	v	n
d	o	e	s	d	m	o	p

Intercluster distance is called _____.
Unsupervised classification involves looking for _____ in the data.
_____ (abbreviated) is used to find a layout of points that matches the interpoint distance matrix.
Iteratively re-grouping cases with the closest group mean is called _____ cluster analysis.
We compare the results of two cluster algorithms using a _____ matrix.
_____ classification is used to build a rule on a training sample with known classes for classifying new observations.
Inference for a multivariate mean assume that the data is sampled from a _____ (abbreviated) distribution.
The proportion of cases missclassified by a discriminant rule is called the _____ rate.
The number of variables in a multivariate data set is equal to the number of _____.
An unobservable variable is called a _____ variable in factor analysis.