Written test, date: 11. December 2007

Course no. : 02409

Course name: Multivariate Statistics "Multivariat Statistik".

Aids allowed: All usual ones

"Weighting": The questions are given equal weight.

This exam is answered by:

_____   _____   _____
(name)   (signature)   (study no.)

There is a total of 30 questions for the 8 problems. The answers to the 30 questions must be written into the table below.

Problem	1	1	1	1	1	1	1	1	1	2
Question	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.1
Answer										

Problem	2	2	2	2	2	2	3	4	4	4
Question	2.2	2.3	2.4	2.5	2.6	2.7	3.1	4.1	4.2	4.3
Answer										

Problem	5	6	7	7	8	8	8	8	8	8
Question	5.1	6.1	7.1	7.2	8.1	8.2	8.3	8.4	8.5	8.6
Answer										

The possible answers for each question are numbered from 1 to 6. If you enter a wrong number, you may correct it by crossing the wrong number in the table and writing the correct answer immediately below. If there is any doubt about the meaning of a correction then the question will be considered not answered.

Only the front page must be returned. The front page must be returned even if you do not answer any of the questions or if you leave the exam prematurely. Drafts and/or comments are **not** considered, only the numbers entered above are registered.

A correct answer gives 5 points, a wrong answer gives −1 point. Unanswered questions or a 6 (corresponding to "don't know") gives 0 points. The total number of points, needed for a satisfactorily answered exam is determined at the final evaluation of the exam.

Remember to write your name, signature and study number on the front page.

Please note, that there is one and only one correct answer to each question. Furthermore, some of the possible alternative answers may not make sense. When the text refers to SAS-output the values may be rounded to fewer decimal places than in the output itself. Please check that all pages of the exam paper and the enclosure are present.

Problem 1.

Enclosure A with SAS-program and SAS-output belongs to this problem.

The data are observations from the Landsat 4 Thematic Mapper. Data are from Eastern Greenland, from the north-western part of an island called Ymer Ø.

Each observation consists of values of reflected light from six spectral bands:

Variable	spectral band (in μm)	Description
tm1	0.45-0.52	visible blue
tm2	0.52-0.60	visible green
tm3	0.63-0.69	visible red
tm4	0.76-0.90	near infrared
tm5	1.55-1.75	near infrared
tm7	2.08-2.35	near infrared

We are concerned with observations of a certain type of geology called "Bedrock 15".

Question 1.1.

Consider the variance-covariance matrix. The total variance for the first five variables only (i.e. tm1, tm2, tm3, tm4, tm5) amounts to:

1 □ 549.09

2 □ 6

3 □ 10.42+6.56+7.96+6.64+14.90+8.25

4 □ 478.35+54.73+8.27+4.97+2.11

5 □ 549.09-68.04

6 □ Don't know.

Question 1.2.

Consider the variance-covariance matrix. The number of principal components needed to explain at least 95% of the variation is:

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 4

5 ☐ 5

6 ☐ Don't know.

Question 1.3.

Consider the variance-covariance matrix. A test for the smallest 4 eigenvalues being equal has the following number of degrees of freedom (we will assume it is sensible to perform such a test):

1 ☐ 6

2 ☐ 2

3 ☐ 9

4 ☐ 4

5 ☐ 5

6 ☐ Don't know.

Question 1.4.

Consider the correlation matrix. The first normed eigenvector of the correlation matrix is:

1 ☐ $[0.45\ 0.28\ 0.33\ 0.27\ 0.65\ 0.36]'$

2 ☐ $[\frac{108.53}{10.42}\ \frac{65.33}{6.56}\ \frac{78.16}{7.96}\ \frac{60.63}{6.64}\ \frac{122.65}{14.90}\ \frac{70.82}{8.23}]'$

3 ☐ $[0.94\ 0.98\ 0.98\ 0.92\ 0.98\ 0.98]'$

4 ☐ $[1\ 0\ 0\ 0\ 0\ 0]'$

5 ☐ $\frac{1}{\sqrt{5.25}} \cdot [0.96\ 0.97\ 0.96\ 0.93\ 0.88\ 0.91]'$

6 ☐ Don't know.

Question 1.5.

Consider the correlation matrix. The correlation between tm1 and tm3 equals:

1 □ $\frac{78.16}{10.42 \cdot 7.96}$

2 □ $478.35 \cdot 8.27$

3 □ $\frac{108.53}{10.42 \cdot 7.96}$

4 □ $\frac{78.16}{108.53 \cdot 63.38}$

5 □ $\frac{63.38}{10.42 \cdot 7.96}$

6 □ Don't know.

Question 1.6.

Consider the correlation matrix. In the usual test for the hypothesis that the correlation between tm1 and tm3 is equal to 0 we use the following distribution (We will assume it is sensible to suggest such a test):

1 □ t(389)

2 □ t(386)

3 □ t(6)

4 □ t(4)

5 □ t(387)

6 □ Don't know.

Question 1.7.

Consider the correlation matrix. The variance explained by varimax rotated factor 1 is:

1 □ 5.25

2 □ 0.96

3 □ 0.94

4 □ 3.41

5 □ 0.82

6 □ Don't know.

Question 1.8.

Consider the correlation matrix. The fraction of the variance of variable tm1 which is **not** explained by a two factor solution is:

1 □ 6-5.25-0.52

2 □ 1-0.94

3 □ 0.13

4 □ 1-0.96

5 □ 1-0.82

6 □ Don't know.

Question 1.9.

Which one of the following interpretations is correct:

1 □ Rotated factor 1 is mainly concerned with the spectral bands with wavelength below 1μm. Rotated factor 2 is mainly concerned with the spectral bands with wavelength above 1μm.

2 □ Rotated factor 1 is mainly an average of all spectral bands. Unrotated factor 2 is an average of the visual spectral bands.

3 □ Unrotated factor 1 is mainly an average of all spectral bands. Rotated factor 2 is a contrast between spectral bands with wavelength above and below 1μm.

4 □ Unrotated factor 1 is a contrast between spectral bands with wavelength above and below 1μm. Rotated factor 2 is mainly concerned with the spectral bands with wavelength above 1μm.

5 □ Rotated factor 1 is mainly an average of all spectral bands. Unrotated factor 2 is also mainly an average of all spectral bands.

6 □ Don't know.

Problem 2.

Enclosure B with SAS-program and SAS-output belongs to this problem.

The option "simple" in PROC DISCRIM means that SAS will also provide simple statistics like means and standard deviations. The option "tcov" gives the estimated variance-covariance matrix for the total sample.

The data are observations from the Landsat 4 Thematic Mapper. Data are from Eastern Greenland, from the north-western part of an island called Ymer Ø.

Each observation consists of values of reflected light from six spectral bands:

Variable	spectral band (in μm)	Description
tm1	0.45-0.52	visible blue
tm2	0.52-0.60	visible green
tm3	0.63-0.69	visible red
tm4	0.76-0.90	near infrared
tm5	1.55-1.75	near infrared
tm7	2.08-2.35	near infrared

We are concerned with observations of a certain type of geology called "Bedrock 15" which have been divided by the variable "location" into an "East" group and a "West" group.

In the following the assumptions for performing both linear and quadratic discriminant analyses, where one compares locations, are assumed to be fulfilled.

Question 2.1.

Assuming equal variance-covariance matrices in the two groups, Mahalanobis' distance is estimated at:

1 □ 13.81

2 □ 1.22

3 □ 15.36

4 □ 15.05

5 □ 14.07

6 □ Don't know.

Question 2.2.

Assume equal variance-covariance matrices in the two groups. Under the null-hypothesis the usual test statistic for equal means is distributed as:

1 □ t(387)

2 □ t(389)

3 □ F(6,382)

4 □ F(2,386)

5 □ F(290,97)

6 □ Don't know.

Question 2.3.

Using the quadratic discriminant analysis instead of the linear discriminant analysis reduces the number of erroneously classified observations by:

1 □ 1

2 □ 6

3 □ 26

4 □ 12

5 □ 16

6 □ Don't know.

Question 2.4.

Consider:

$$a = \begin{pmatrix} 54.4 \\ 113.9 \end{pmatrix}, b = \begin{pmatrix} 53.1 \\ 111.8 \end{pmatrix}, c = \begin{pmatrix} 58.4 \\ 120.0 \end{pmatrix}$$

and

$$d = \begin{pmatrix} 38.1 & 62.9 \\ 62.9 & 212.9 \end{pmatrix}, e = \begin{pmatrix} 41.7 & 67.4 \\ 67.4 & 200.4 \end{pmatrix}, f = \begin{pmatrix} 39.0 & 64.0 \\ 64.0 & 209.7 \end{pmatrix}, g = \begin{pmatrix} 44.1 & 72.0 \\ 72.0 & 222.0 \end{pmatrix}$$

Suppose we want to make a quadratic classification rule based only on tm4 and tm5. We will then need the vectors and matrices:

1 □ b, c, f

2 □ c, d, e, g

3 □ a, b, c, d

4 □ b, c, d, e

5 □ b, c, f, g

6 □ Don't know.

Question 2.5.

The linear discriminant rule (in the SAS output) can be written in the form: Classify a new observation as "East" if $x'\delta + \alpha > 0$, where x is the vector of measurements of the (new) observation, δ is a constant vector, and α is a constant. We wish to adjust this rule by prior probabilities proportional to the actual counts of East and West observations in the dataset. The correct classification rule can then be written: Classify a new observation as "East" if $x'\delta + \alpha > k$, where k equals:

1 □ 1

2 □ $\ln \frac{98}{291}$

3 □ $\ln \frac{291}{98}$

4 □ $\exp\left(\frac{0.25}{0.75}\right)$

5 □ $\exp\left(\frac{0.75}{0.25}\right)$

6 □ Don't know.

The problem continues on the next page

Question 2.6.

It is hypothesized that a certain linear projection determined by an hypothesized, fixed value of the vector d is optimal. (The classification rule is: Classify a new observation as "East" if $x'd + \alpha > 0$, where x is the measurements of the (new) observation, d is a constant vector, and α is a constant.) The usual test statistic for this is distributed as:

1 □ F(5,382)

2 □ F(6,383)

3 □ F(5,383)

4 □ F(6,382)

5 □ F(5,384)

6 □ Don't know.

Question 2.7.

The generalised variance based on the pooled variance-covariance matrix is:

1 □ 6

2 □ $94.36 \cdot 36.08 \cdot 54.23 \cdot 38.99 \cdot 209.74 \cdot 62.53$

3 □ 387

4 □ $\det \begin{pmatrix} 0 & 1.22 \\ 1.22 & 0 \end{pmatrix}$

5 □ $e^{14.04}$

6 □ Don't know.

Problem 3.

Consider the following regression models with $i = 1, \ldots, n$ observations:

Model	
A	$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
B	$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
C	$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3}$
D	$E(Y_i) = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}$
E	$E(Y_i) = \beta_0 + \beta_1 x_{i1}$
F	$E(Y_i) = \beta_0 + \beta_2 x_{i2}$
G	$E(Y_i) = \beta_0 + \beta_3 x_{i3}$
H	$E(Y_i) = \beta_0$

Question 3.1.

Which sequence complies with successive testing?

1 □ A, E, G

2 □ A, D, H, G

3 □ A, C, E, H

4 □ A, B, D, F

5 □ B, E, F

6 □ Don't know.

Problem 4.

Consider: $\boldsymbol{X} \in N \left(\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} \right)$

Now let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \boldsymbol{X}$

Question 4.1.

$E(Y_2)$ equals:

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 4

5 ☐ 5

6 ☐ Don't know.

Question 4.2.

$V(Y_2)$ equals:

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 4

5 ☐ 5

6 ☐ Don't know.

Question 4.3.

$\mathrm{Cov}(Y_1, Y_2)$ equals:

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 4

5 ☐ 5

6 ☐ Don't know.

Problem 5.

Consider the following multivariate general linear model (with usual distributional assumptions on the residuals):

$$
E\left(\left(\begin{array}{cc} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ Y_{31} & Y_{32} \\ Y_{41} & Y_{42} \\ Y_{51} & Y_{52} \\ Y_{61} & Y_{62} \end{array}\right)\right) = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right) \left(\begin{array}{cc} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{array}\right)
$$

Question 5.1.

A test with null-hypothesis: $\left(\begin{array}{c} \theta_{11} \\ \theta_{12} \end{array}\right) = \left(\begin{array}{c} \theta_{21} \\ \theta_{22} \end{array}\right)$ is equivalent to:

1 □ Testing equality of group means in a one-sided multidimensional analysis of variance with 3 groups.

2 □ Testing equality of group means in a one-sided multidimensional analysis of variance with 6 groups.

3 □ Hotellings T^2 in the one sample case.

4 □ Hotellings T^2 in the two sample case.

5 □ Testing equality of row means in a two-sided multidimensional analysis of variance with 2 rows and 2 columns.

6 □ Don't know.

Problem 6.

Consider the following matrices:

A: $\left(\begin{array}{ccc} 4 & 3 & 0 \\ 3 & 3 & 0 \\ 0 & 0 & 3 \end{array}\right)$, B: $\left(\begin{array}{cc} 3 & -2 \\ -2 & 2 \end{array}\right)$, C: $\left(\begin{array}{cc} 1 & 3 \\ 3 & 2 \end{array}\right)$, D: $\left(\begin{array}{cc} -3 & 2 \\ 2 & -2 \end{array}\right)$, E: $\left(\begin{array}{cc} 2 & 1 \\ 2 & 2 \end{array}\right)$.

Question 6.1.

Which matrices represent variance-covariance matrices?

1 □ A and B

2 □ B and D

3 □ C and E

4 □ A and D

5 □ D and E

6 □ Don't know.

Problem 7.

Consider $D(\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 6 \end{pmatrix}$

Question 7.1.

The total variance of $\boldsymbol{X}$ is:

1 □ 27

2 □ 11

3 □ 3

4 □ 0

5 □ 24

6 □ Don't know.

Question 7.2.

The generalised variance of $\boldsymbol{X}$ is:

1 □ 27

2 □ 11

3 □ 3

4 □ 0

5 □ 24

6 □ Don't know.

Problem 8.

Enclosure C with SAS-program and SAS-output belongs to this problem.

Data are $n = 15$ observations of (logarithmes of) leafburn time and three measured variables of tobacco leaves.

Variable	Description
Y	logarithm of leafburn time in seconds
X1	percentage of nitrogen (N)
X2	percentage of chlorine (Cl)
X3	percentage of potassium (K)

Question 8.1.

$\sum_{i=1}^{15}(Y_i - \overline{Y})^2$ equals:

1 □ 2.59

2 □ 0.41

3 □ 3.0

4 □ 0.86

5 □ 0.04

6 □ Don't know.

Question 8.2.

The coefficient for X3 (percentage of potassium, K) is significant

1 □ at level 0.1 but not at level 0.05

2 □ at level 0.05 but not at level 0.01

3 □ at level 0.01 but not at level 0.005

4 □ at level 0.005 but not at level 0.001

5 □ at level 0.001 but not at level 0.0005

6 □ Don't know.

The problem continues on the next page

Question 8.3.

The observation most influential on the coefficient for X1 (percentage of nitrogen, N) is observation number:

1 □ 1

2 □ 15

3 □ 3

4 □ 7

5 □ 5

6 □ Don't know.

Question 8.4.

The usual test statistic for the hypothesis that the coefficients for all three independent variables (X1, X2, X3) equal 0 simultaneously has the value:

1 □ 22.90

2 □ 0.8620

3 □ $-6.22 - 4.86 + 3.97$

4 □ $(-6.22)^2 + (-4.86)^2 + (3.97)^2$

5 □ 29.97

6 □ Don't know.

Question 8.5.
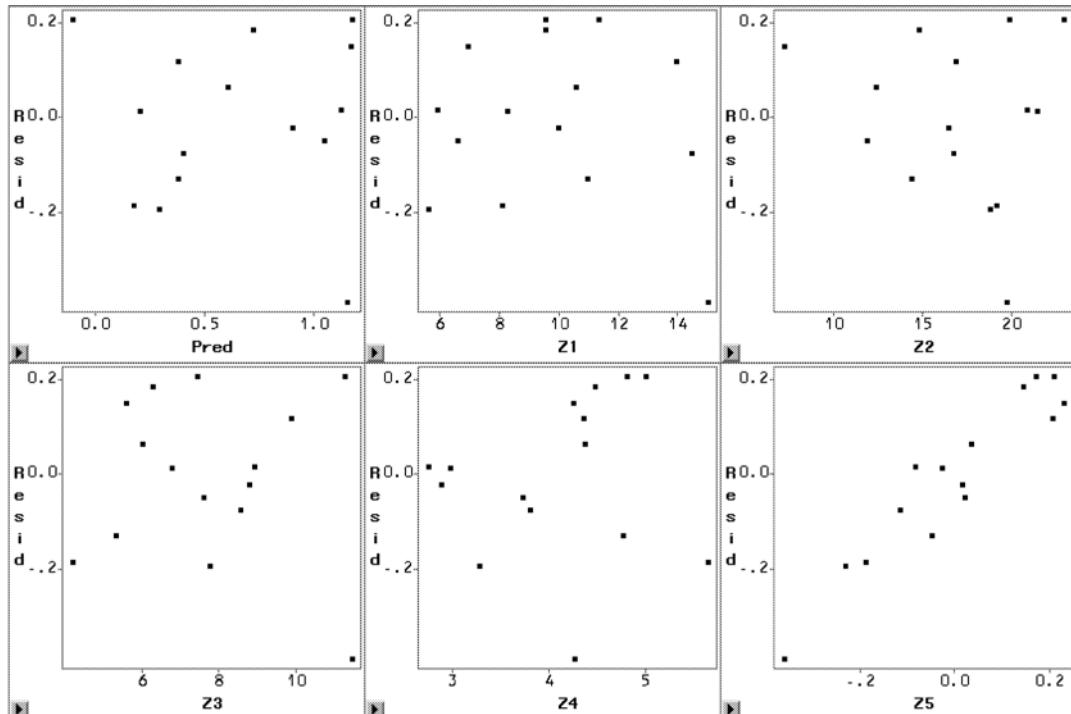
For a new observation the expected value is estimated at 0.7. The variance of the estimate is estimated at 0.01. A 95%-confidence interval for the expected value is then:

1 □ $[0.7 - 2.201 \cdot 0.01; 0.7 + 2.201 \cdot 0.01]$

2 □ $[0.7 - 2.201 \cdot 0.1; 0.7 + 2.201 \cdot 0.1]$

3 □ $[0.7 - 1.796 \cdot 0.01; 0.7 + 1.796 \cdot 0.01]$

4 □ $[0.7 - 1.796 \cdot 0.1; 0.7 + 1.796 \cdot 0.1]$

5 □ $[0.7 - 2.201 \cdot \sqrt{0.1 + 1}; 0.7 + 2.201 \cdot \sqrt{0.1 + 1}]$

6 □ Don't know.

Question 8.6.

Below the residuals from the regression analysis are plotted against: the predicted value (top left plot) and 5 other variables (Z1, Z2, Z3, Z4, Z5) not previously used in the analysis.

Which single variable should one consider to enter into the analysis?

1 □ Z1

2 □ Z2

3 □ Z3

4 □ Z4

5 □ Z5

6 □ Don't know.

```
/* encla.sas   Crtd: 07-11-07 10:44 by BKE. Updt: 25-11-07 12:01 */
/* Purpose:  */

title1 'Enclosure A - Bedrock 15 data';

options obs=20 firstobs=1;

title2 'First 20 observations';
proc print data=stat2.rock;
var tm1 tm2 tm3 tm4 tm5 tm7;
run;

options obs=max firstobs=1;

title2 'Principal component analysis (using all observations)';
proc princomp cov data=stat2.rock;
var tm1 tm2 tm3 tm4 tm5 tm7;
run;

title2 'Factor analysis (using all observations)';
proc factor data=stat2.rock nfactors=2 rotate=varimax;
var tm1 tm2 tm3 tm4 tm5 tm7;
run;
```

Enclosure A - Bedrock 15 data
First 20 observations

Obs	tm1	tm2	tm3	tm4	tm5	tm7
1	118	57	66	60	132	71
2	117	59	68	62	123	68
3	123	63	72	65	131	77
4	119	59	67	62	136	76
5	110	54	62	57	126	68
6	101	47	54	50	117	63
7	110	55	63	57	116	68
8	123	65	74	68	134	79
9	123	65	74	68	134	79
10	119	59	67	62	136	76
11	113	56	65	59	133	75
12	94	46	57	55	110	58
13	94	46	57	55	110	58
14	109	52	60	54	103	58
15	107	51	59	53	116	63
16	97	46	53	50	103	54
17	95	43	51	48	97	55
18	95	43	51	48	97	55
19	95	46	54	51	94	49
20	103	50	58	51	99	54

Enclosure A – Bedrock 15 data
Principal component analysis (using all observations)   3

The PRINCOMP Procedure

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
tm1	0.445133	0.417353	-.622436	-.424607	0.225877	-.094528
tm2	0.276566	0.318448	0.020439	0.266292	-.226518	0.836339
tm3	0.331367	0.411565	0.181402	0.296838	-.573431	-.520546
tm4	0.265632	0.312542	0.716888	-.175186	0.535334	-.023593
tm5	0.645209	-.634746	0.148153	-.308234	-.246364	0.056122
tm7	0.358760	-.233317	-.208259	0.731428	0.470710	-.130107

Enclosure A – Bedrock 15 data
Principal component analysis (using all observations)   2

The PRINCOMP Procedure

Observations   389
Variables   6

Simple Statistics

	tm1	tm2	tm3
Mean	106.0282776	51.27506427	59.50899743
StD	10.4176556	6.56285394	7.96111984

Simple Statistics

	tm4	tm5	tm7
Mean	54.42673522	113.8714653	63.98457584
StD	6.64124141	14.8985896	8.24869678

Covariance Matrix

	tm1	tm2	tm3
tm1	108.5275488	65.3349853	78.1582501
tm2	65.3349853	43.0710519	51.4163332
tm3	78.1582501	51.4163332	63.3794291
tm4	60.6348091	40.2096375	49.3234702
tm5	122.6505512	74.0612196	88.0192007
tm7	70.8226022	44.0300268	51.8532319

Covariance Matrix

	tm4	tm5	tm7
tm1	60.6348091	122.6505512	70.8226022
tm2	40.2096375	74.0612196	44.0300268
tm3	49.3234702	88.0192007	51.8532319
tm4	44.1060875	71.9931360	40.2565990
tm5	71.9931360	221.9679723	117.2067752
tm7	40.2565990	117.2067752	68.0409986

Total Variance   549.09308828

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	478.346907	423.616205	0.8712	0.8712
2	54.730702	46.459887	0.0997	0.9708
3	8.270816	3.304264	0.0151	0.9859
4	4.966552	2.858760	0.0090	0.9949
5	2.107792	1.437473	0.0038	0.9988
6	0.670319		0.0012	1.0000

Page 4

```
                                                            4
            Enclosure A – Bedrock 15 data
       Factor analysis (using all observations)

              The FACTOR Procedure
       Initial Factor Method: Principal Components

         Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6   Average = 1

        Eigenvalue    Difference   Proportion   Cumulative

    1   5.25490316    4.73695530     0.8758       0.8758
    2   0.51794785    0.39152663     0.0863       0.9621
    3   0.12642122    0.06755786     0.0211       0.9832
    4   0.05886336    0.03031389     0.0098       0.9930
    5   0.02854947    0.01523452     0.0048       0.9978
    6   0.01331494                   0.0022       1.0000

    2 factors will be retained by the NFACTOR criterion.

              Factor Pattern

                    Factor1       Factor2

       tm1          0.96176      -0.11123
       tm2          0.97025      -0.19563
       tm3          0.96296      -0.22994
       tm4          0.92717      -0.24870
       tm5          0.88140       0.44795
       tm7          0.90815       0.38977

       Variance Explained by Each Factor

            Factor1       Factor2

          5.2549032     0.5179479

       Final Communality Estimates: Total = 5.772851

    tm1          tm2          tm3          tm4          tm5          tm7

 0.93736279   0.97965497   0.98015889   0.92148444   0.97752324   0.9666668
```

Page 5

```
                                                            5
            Enclosure A – Bedrock 15 data
       Factor analysis (using all observations)

              The FACTOR Procedure
              Rotation Method: Varimax

       Orthogonal Transformation Matrix

                       1           2

             1      0.78129     0.62417
             2     -0.62417     0.78129

              Rotated Factor Pattern

                    Factor1       Factor2

       tm1          0.82084       0.51340
       tm2          0.88015       0.45275
       tm3          0.89587       0.42139
       tm4          0.87961       0.38440
       tm5          0.40904       0.90012
       tm7          0.46625       0.87137

       Variance Explained by Each Factor

            Factor1       Factor2

          3.4094588     2.3633922

       Final Communality Estimates: Total = 5.772851

    tm1          tm2          tm3          tm4          tm5          tm7

 0.93736279   0.97965497   0.98015889   0.92148444   0.97752324   0.9766668
```

```
                    Enclosure B - Bedrock 15 data
First 20 observations showing some West and some East examples            1

Obs    location    tm1    tm2    tm3    tm4    tm5    tm7
 1     West        118     57     66     60    132     71
 2     West        117     59     68     62    123     68
 3     West        123     63     72     65    131     77
 4     West        119     59     67     62    136     76
 5     West        110     54     62     57    126     68
 6     West        101     47     54     50    117     63
 7     West        110     55     63     57    116     68
 8     West        123     65     74     68    134     79
 9     West        123     65     74     68    134     79
10     West        119     59     67     62    136     76
11     West        113     56     65     59    133     75
12     West         94     46     57     55    110     58
13     West         94     46     57     55    110     58
14     East        109     52     60     54    103     58
15     East        107     51     59     53    116     63
16     East         97     46     53     50    103     54
17     East         95     43     51     48     97     55
18     East         95     43     51     48     97     55
19     East         95     46     54     51     94     49
20     East        103     50     58     51     99     54
```

```
/><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>></      2
                    Enclosure B - Bedrock 15 data
           Discriminant analysis (using all observations)

                      The DISCRIM Procedure

           Observations     389        DF Total               388
           Variables          6        DF Within Classes      387
           Classes            2        DF Between Classes        1

                    Class Level Information

                  Variable                                      Prior
location          Name        Frequency      Weight   Proportion   Probability

East              East             291      291.0000    0.748072     0.500000
West              West              98       98.0000    0.251928     0.500000
```

```
/* enclb.sas   Crtd: 07-11-07 11:48 by BKE. Updt: 28-11-07 15:57 */
/* Purpose:  */

title1 'Enclosure B - Bedrock 15 data';

options obs=20 firstobs=1;

title2 'First 20 observations showing some West and some East examples';
proc print data=stat2.rock;
var location tm1 tm2 tm3 tm4 tm5 tm7;
run;

options obs=max firstobs=1;

title2 'Discriminant analysis (using all observations)';
proc discrim data=stat2.rock tcov wcov pcov simple pool=yes;
class location;
var tm1 tm2 tm3 tm4 tm5 tm7;
run;

title2 'Discriminant analysis (using all observations)';
proc discrim data=stat2.rock pool=no;
class location;
var tm1 tm2 tm3 tm4 tm5 tm7;
run;
```

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)   4

The DISCRIM Procedure

Pooled Within-Class Covariance Matrix,   DF = 387

Variable	tm1	tm2	tm3
tm1	94.3605726	55.3706850	66.7586119
tm2	55.3706850	36.0751752	43.4120629
tm3	66.7586119	43.4120629	54.2268541
tm4	52.0995136	34.2171608	42.4710682
tm5	109.3692814	64.7150887	77.3269802
tm7	61.9450924	37.7889357	44.7114212

Pooled Within-Class Covariance Matrix,   DF = 387

Variable	tm4	tm5	tm7
tm1	52.0995136	109.3692814	61.9450924
tm2	34.2171608	64.7150887	37.7889357
tm3	42.4710682	77.3269802	44.7114212
tm4	38.9907163	63.3981010	34.9095569
tm5	63.3981010	209.7426332	108.9816945
tm7	34.9095569	108.9816945	62.5346258

Total-Sample Covariance Matrix,   DF = 388

Variable	tm1	tm2	tm3
tm1	108.5275488	65.3349853	78.1582501
tm2	65.3349853	43.0710519	51.4163332
tm3	78.1582501	51.4163332	63.3794291
tm4	60.6348091	40.2096375	49.3234702
tm5	122.6505512	74.0612196	88.0192007
tm7	70.8226022	44.0300268	51.8532319

Total-Sample Covariance Matrix,   DF = 388

Variable	tm4	tm5	tm7
tm1	60.6348091	122.6505512	70.8226022
tm2	40.2096375	74.0612196	44.0300268
tm3	49.3234702	88.0192007	51.8532319
tm4	44.1060875	71.9931360	40.2565990
tm5	71.9931360	221.9679723	117.2067752
tm7	40.2565990	117.2067752	68.0409986

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)   3

The DISCRIM Procedure
Within-Class Covariance Matrices

location = East,   DF = 290

Variable	tm1	tm2	tm3
tm1	96.4945136	56.6434412	69.2865979
tm2	56.6434412	36.6724337	44.4858514
tm3	69.2865979	44.4858514	55.8954141
tm4	52.8039815	34.1897974	42.5266619
tm5	114.6035668	66.6222538	79.6655950
tm7	66.6111743	39.8695699	47.3322787

location = East,   DF = 290

Variable	tm4	tm5	tm7
tm1	52.8039815	114.6035668	66.6111743
tm2	34.1897974	66.6222538	39.8695699
tm3	42.5266619	79.6655950	47.3322787
tm4	38.0789904	62.8654461	35.6995023
tm5	62.8654461	212.8704823	114.7106174
tm7	35.6995023	114.7106174	67.1984595

location = West,   DF = 97

Variable	tm1	tm2	tm3
tm1	87.9807490	51.5655376	59.2007153
tm2	51.5655376	34.2886598	40.2017673
tm3	59.2007153	40.2017673	49.2383758
tm4	49.9933726	34.2989691	42.3048601
tm5	93.7203871	59.0132548	70.3349463
tm7	47.9949506	31.5684831	36.8758679

location = West,   DF = 97

Variable	tm4	tm5	tm7
tm1	49.9933726	93.7203871	47.9949506
tm2	34.2989691	59.0132548	31.5684831
tm3	42.3048601	70.3349463	36.8758679
tm4	41.7164948	67.3843888	32.5478645
tm5	67.3843888	200.3913318	91.8539870
tm7	32.5478645	91.8539870	48.5912056

6

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to location

From location	East	West
East	0	1.22030
West	1.22030	0

Linear Discriminant Function

$$\text{Constant} = -.5 \, \bar{X}_j' \, COV^{-1} \, \bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1} \, \bar{X}_j$$

Linear Discriminant Function for location

Variable	East	West
Constant	-86.29019	-93.66651
tm1	3.30195	3.25812
tm2	-2.70084	-2.20183
tm3	-2.34413	-2.49623
tm4	1.99085	1.94901
tm5	-0.64291	-0.67579
tm7	1.04745	1.06631

5

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure
Simple Statistics

Total-Sample

Variable	N	Sum	Mean	Variance	Standard Deviation
tm1	389	41245	106.02828	108.52755	10.4177
tm2	389	19946	51.27506	43.07105	6.5629
tm3	389	23149	59.50900	63.37943	7.9611
tm4	389	21172	54.42674	44.10609	6.6412
tm5	389	44296	113.87147	221.96797	14.8986
tm7	389	24890	63.98458	68.04100	8.2487

location = East

Variable	N	Sum	Mean	Variance	Standard Deviation
tm1	291	30214	103.82818	96.49451	9.8232
tm2	291	14472	49.73196	36.67273	6.0558
tm3	291	16803	57.74227	55.89541	7.4763
tm4	291	15453	53.10309	38.07899	6.1708
tm5	291	32534	111.80069	212.87048	14.5901
tm7	291	18218	62.60481	67.19846	8.1975

location = West

Variable	N	Sum	Mean	Variance	Standard Deviation
tm1	98	11031	112.56122	87.98075	9.3798
tm2	98	5474	55.85714	34.28866	5.8557
tm3	98	6346	64.75510	49.23838	7.0170
tm4	98	5719	58.35714	41.71649	6.4588
tm5	98	11762	120.02041	200.39133	14.1560
tm7	98	6672	68.08163	48.59121	6.9707

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
6	14.04138

7

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure

Classification Summary for Calibration Data: STAT2.ROCK
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X - \bar{X}_j)' \, COV^{-1} \, (X - \bar{X}_j)$$

Posterior Probability of Membership in Each location

$$Pr(j|X) = \frac{exp(-.5 \, D_j^2(X))}{SUM_k \, exp(-.5 \, D_k^2(X))}$$

Number of Observations and Percent Classified into location

From location	East	West	Total
East	203 69.76	88 30.24	291 100.00
West	26 26.53	72 73.47	98 100.00
Total	229 58.87	160 41.13	389 100.00
Priors	0.5	0.5	

Error Count Estimates for location

	East	West	Total
Rate	0.3024	0.2653	0.2839
Priors	0.5000	0.5000	

8

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure

Observations	389	DF Total	388
Variables	6	DF Within Classes	387
Classes	2	DF Between Classes	1

Class Level Information

location	Variable Name	Frequency	Weight	Proportion	Prior Probability
East	East	291	291.0000	0.748072	0.500000
West	West	98	98.0000	0.251928	0.500000

Within Covariance Matrix Information

location	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
East	6	13.81192
West	6	14.07423

/><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>><<>></
9

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j) + ln \, |COV_j|$$

Generalized Squared Distance to location

From location	East	West
East	13.81192	15.35945
West	15.04607	14.07423

Right panel

Nov 29 2007 15:02 | **Enclosure C – SAS program** | Page 1

```
/* enclc.sas   Crtd: 09-11-07 10:24 by BKE. Updt: 28-11-07 16:08 */
/* Purpose:  */

title1 'Enclosure C – Tobacco leaf burn data';

title2 'Print of data';
proc print data=stat2.tobacco;
var Y X1 X2 X3;
run;

title2 'Regression analysis';
proc reg data=stat2.tobacco;
model Y=X1 X2 X3 / covb influence r;
run;
```

8

Left panel

Nov 29 2007 15:08 | **Enclosure B – SAS output** | Page 8

10

Enclosure B – Bedrock 15 data
Discriminant analysis (using all observations)

The DISCRIM Procedure
Classification Summary for Calibration Data: STAT2.ROCK
Resubstitution Summary using Quadratic Discriminant Function

Generalized Squared Distance Function

$$D^2_j(X) = (X-\bar{X}_j)' \, COV^{-1}_j \, (X-\bar{X}_j) + \ln |COV_j|$$

Posterior Probability of Membership in Each location

$$Pr(j|X) = \exp(-.5 \, D^2_j(X)) \, / \, SUM \, \exp(-.5 \, D^2_k(X))$$

Number of Observations and Percent Classified into location

From location	East	West	Total
East	219	72	291
	75.26	24.74	100.00
West	16	82	98
	16.33	83.67	100.00
Total	235	154	389
	60.41	39.59	100.00
Priors	0.5	0.5	

Error Count Estimates for location

	East	West	Total
Rate	0.2474	0.1633	0.2053
Priors	0.5000	0.5000	

Enclosure C – Tobacco leaf burn data
Regression analysis

2

```
The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read    15
Number of Observations Used    15
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.58609	0.86203	22.90	<.0001
Error	11	0.41400	0.03764		
Corrected Total	14	3.00009			

Root MSE	0.19400	R-Square	0.8620
Dependent Mean	0.64733	Adj R-Sq	0.8244
Coeff Var	29.96937		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.14539	0.39661	5.41	0.0002
X1	1	-0.60458	0.09726	-6.22	<.0001
X2	1	-0.46762	0.09627	-4.86	0.0005
X3	1	0.19234	0.04841	3.97	0.0022

Covariance of Estimates

Variable	Intercept	X1	X2	X3
Intercept	0.1573021686	-0.032657112	-0.001826202	-0.010113823
X1	-0.032657112	0.0094588453	0.0003932141	0.0002243875
X2	-0.001826202	0.0003932141	0.0092671517	-0.001446839
X3	-0.010113823	0.0002243875	-0.001446839	0.0023431576

Enclosure C – Tobacco leaf burn data
Print of data

1

Obs	Y	X1	X2	X3
1	0.11	4.22	1.35	4.86
2	0.68	3.77	0.23	4.42
3	0.00	3.54	0.76	2.76
4	0.11	3.78	0.39	3.23
5	0.77	3.10	0.64	6.16
6	1.01	2.78	0.64	4.62
7	1.40	2.67	0.90	5.59
8	1.15	3.03	0.97	6.60
9	0.51	4.12	0.62	5.31
10	0.34	3.94	0.45	4.45
11	0.89	2.93	0.25	3.38
12	0.92	3.17	0.20	3.08
13	1.33	2.61	0.20	3.64
14	0.26	3.13	1.48	4.28
15	0.23	2.94	2.22	4.58

Page 3

3

Enclosure C – Tobacco leaf burn data
Regression analysis

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	0.1100	-0.1025	0.1176	0.2125	0.154	1.377
2	0.6800	0.6087	0.0821	0.0713	0.176	0.406
3	0	0.1806	0.0982	-0.1806	0.167	-1.080
4	0.1100	0.2990	0.0875	-0.1890	0.173	-1.091
5	0.7700	1.1567	0.1011	-0.3867	0.166	-2.335
6	1.0100	1.0540	0.0739	-0.0440	0.179	-0.245
7	1.4000	1.1855	0.0935	0.2145	0.170	1.262
8	1.1500	1.1293	0.1129	0.0207	0.158	0.131
9	0.5100	0.3859	0.1050	0.1241	0.163	0.761
10	0.3400	0.4088	0.0830	-0.0688	0.175	-0.393
11	0.8900	0.9072	0.0882	-0.0172	0.173	-0.0994
12	0.9200	0.7278	0.0893	0.1922	0.172	1.116
13	1.3300	1.1740	0.1044	0.1560	0.164	0.954
14	0.2600	0.3842	0.0900	-0.1242	0.172	-0.723
15	0.2300	0.2107	0.1512	0.0193	0.122	0.159

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	-2-1 0 1 2
1	0.275	1.4427	0.3672	1.0859	1.0989	|**
2	0.009	0.3896	0.1792	1.6794	0.1821	|
3	0.101	-1.0890	0.2565	1.2578	-0.6396	**|
4	0.076	-1.1019	0.2034	1.1621	-0.5568	**|
5	0.508	-3.1361	0.2715	0.1298	-1.9145	****|
6	0.003	-0.2344	0.1450	1.6753	-0.0965	|
7	0.120	1.3012	0.3322	1.0201	0.7156	|**
8	0.002	0.1249	0.3386	2.1999	0.0894	|
9	0.060	0.7451	0.2929	1.6681	0.4795	|*
10	0.009	-0.3769	0.1830	1.6938	-0.1784	|
11	0.001	-0.0948	0.2066	1.8387	-0.0484	|
12	0.084	1.1302	0.2118	1.1485	0.5859	|**
13	0.093	0.9495	0.2894	1.4588	0.6059	|*
14	0.036	-0.7060	0.2153	1.5359	-0.3698	*|
15	0.010	0.1514	0.6075	3.6959	0.1884	|

Output Statistics

Obs	Intercept	DFBETAS X1	X2	X3
1	-0.7573	0.8536	0.5159	0.0517
2	-0.0611	0.0931	-0.1061	0.0346
3	-0.2034	-0.1168	-0.1769	0.5315
4	-0.0077	-0.2601	-0.0925	0.2973
5	0.3536	0.3312	0.7061	-1.6001
6	-0.0614	0.0682	0.0202	-0.0111
7	-0.2302	-0.4560	-0.0416	0.3607
8	-0.0203	-0.0174	-0.0098	0.0757
9	-0.3694	-0.3639	-0.1017	0.2223
10	-0.0932	-0.1279	0.0568	-0.0243
11	-0.0373	0.0230	0.0185	0.0215

Page 4

4

Enclosure C – Tobacco leaf burn data
Regression analysis

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Output Statistics

Obs	Intercept	DFBETAS X1	X2	X3
12	0.3683	-0.1284	-0.2169	-0.3355
13	0.5107	-0.4223	-0.2541	-0.1545
14	-0.0942	0.0637	-0.2956	-0.1293
15	-0.0343	-0.0378	0.1718	-0.0498

Sum of Residuals   0
Sum of Squared Residuals   0.41400
Predicted Residual SS (PRESS)   0.77252