

*Written examination, date:* 8. December 2015

*Page 1 of 16 pages Enclosure: 14 pages*

*Course name:* Multivariate Statistics

*Course number:* 02409

*Aids allowed:* All

*Exam duration:* 4 hours

*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 6 problems. The answers to the 30 questions must be written into the table below.

Problem	1	1	1	1	1	1	2	2	2	2
Question	1.1	1.2	1.3	1.4	1.5	1.6	2.1	2.2	2.3	2.4
Answer										

Problem	2	2	2	3	3	3	3	3	3	3
Question	2.5	2.6	2.7	3.1	3.2	3.3	3.4	3.5	3.6	3.7
Answer										

Problem	4	4	4	4	4	4	4	5	5	5
Question	4.1	4.2	4.3	4.4	4.5	4.6	4.7	5.1	5.2	5.3
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”) give 0 points. The total number of points needed for a satisfactorily answered exam is determined at the final evaluation of the exam. Especially note that the grade 10 may be given even if only one answer is wrong or unanswered.  
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. The enclosures do not necessarily contain all the output generated by the given SAS programs. Please check that all pages of the exam paper and the enclosures are present.*

## Problem 1.

Enclosure A with SAS program and SAS output belongs to this problem. The data are taken from “Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado (2015): Modeling and Analysis of Compositional Data, xvii + 247 pp., John Wiley & Sons, Ltd.”. The data give percentages of protein consumption in 25 European countries in 1980-1989 provided by each of nine food categories:

1. Red meats (pork, beef, veal),
2. White meats (poultry( chicken, turkey)),
3. Eggs
4. Milk products
5. Fish (sea and freshwater)
6. Cereals
7. Starch sources (potatoes, rice),
8. Nuts
9. Vegetables (including fruit and orchard products).

The abbreviations used for the 25 countries considered are given below. Furthermore we give a ‘geopolitical’ code where the last letter (E or W) indicates whether the country belonged to the eastern or western block, and the prefix (Nor, Cen, or Sou) indicates the geographical location within Europe.

Country	Abbrev.	GeoPol	Country	Abbrev.	GeoPol
Albania	ALB	SouE	Netherlands	HOL	CenW
Austria	AUS	CenW	Norway	NOR	NorW
Belgium	BEL	CenW	Poland	POL	CenE
Bulgaria	BUL	SouE	Portugal	POR	SouW
Czech Republic	CZE	CenE	Romania	ROM	SouE
Denmark	DEN	NorW	Spain	SPA	SouW
German Fed. Rep.	BRD	CenW	Sweden	SWE	NorW
Finland	FIN	NorW	Switzerland	SCH	CenW
France	FRA	CenW	United Kingdom	UK	NorW
Greece	GRE	NorW	USSR	USSR	CenE
Hungary	HUN	CenE	German Dem. Rep.	DDR	CenE
Ireland	IRE	NorW	Yugoslavia	YUG	SouE
Italy	ITA	SouW			

*The problem continues on the next page*

### **Question 1.1.**

How many principal components should be included in the analysis in order to describe at least 90% of the total variation?

- 1  1
- 2  2
- 3  3
- 4  4
- 5  5
- 6  Don't know

### **Question 1.2.**

The distribution of the usual test statistic for testing the hypothesis that the smallest 5 eigenvalues (in the correlation matrix) are equal against all alternatives is (under the hypothesis)

- 1   $\chi^2(14)$
- 2   $t(14)$
- 3   $\chi^2(18)$
- 4   $F(3,14)$
- 5   $t(9 - 1)$
- 6  Don't know

*The problem continues on the next page*

### Question 1.3.

We now consider a factor analysis with three factors of the data considered above. Consider the following statements on interpretation of an arbitrary factor. The factor basically represents

- I. the mean level of all meat values.
- II. a large correlation with fish, potatoes/rice and vegetables on one side, and a numerically large, negative correlation with cereals on the other side.
- III. the mean level of animal protein sources
- IV. a large correlation with poultry, eggs and potatoes/rice on one side, and a numerically large, negative correlation with nuts and cereals on the other side.
- V. a large correlation with beef/pork/veal, eggs and milk products on one side, and a numerically large, negative correlation with cereals, nuts and vegetables on the other side.

For (VARIMAX rotated factor 1, VARIMAX rotated factor 2) the following characterization is adequate

- 1  (I, II)
- 2  (II, III)
- 3  (III, IV)
- 4  (IV, V)
- 5  (V, I)
- 6  Don't know

### Question 1.4.

The first 2 factor score values for a given country are denoted  $(f_1, f_2)$ . We now consider the northernmost countries in the western block (NorW) and Central European countries in the western block (CenW). The distribution of the factor scores for the two areas satisfy

- 1   $f_1$  has intermediate values for NorW and positive values for CenW  
 $f_2$  has positive and large values for NorW and has intermediate values for CenW.
- 2   $f_1$  has intermediate values for NorW and positive values for CenW  
 $f_2$  has negative values for NorW and has intermediate values for CenW.
- 3   $f_1$  has intermediate values for NorW and negative values for CenW  
 $f_2$  has positive and large values for NorW and has intermediate values for CenW.
- 4   $f_1$  has positive values for NorW and positive values for CenW  
 $f_2$  has positive and large values for NorW and has intermediate values for CenW.
- 5   $f_1$  has intermediate values for NorW and negative values for CenW  
 $f_2$  has positive and large values for NorW and has intermediate values for CenW.
- 6  Don't know

*The problem continues on the next page*

### **Question 1.5.**

The variable that is explained best by the rotated factor model is

- 1  VEG
- 2  STARCH
- 3  FISH
- 4  WMEAT
- 5  CEREAL
- 6  Don't know

### **Question 1.6.**

The reduction in the variance explained by the first factor by going from the unrotated model to the rotated model is

- 1  1.0
- 2  1.5
- 3  2.0
- 4  2.5
- 5  3.0
- 6  Don't know

## **Problem 2.**

We consider a random variable

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix}$$

with expectation vector and dispersion matrix equal to

*The problem continues on the next page*

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1/4 & 1/8 \\ 1 & 4 & 1 & 1 \\ 1/4 & 1 & 1 & 1 \\ 1/8 & 1 & 1 & 4 \end{bmatrix}$$

In the sequel you may find the following expressions useful

$$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}^{-1} = \frac{1}{3} \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\frac{1}{3} \begin{bmatrix} 1/4 & 1/8 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 7/24 & -1/24 \\ 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 7/24 & -1/24 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1/4 & 1 \\ 1/8 & 1 \end{bmatrix} = \begin{bmatrix} 13/192 & 1/4 \\ 1/4 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} - \begin{bmatrix} 13/192 & 1/4 \\ 1/4 & 1 \end{bmatrix} = \begin{bmatrix} 179/192 & 3/4 \\ 3/4 & 3 \end{bmatrix}$$

### Question 2.1.

The squared correlation between  $Y_1$  and  $Y_2$  is

1   $\frac{1}{2}$

2   $\frac{1}{4}$

3   $\frac{1}{8}$

4   $\frac{1}{16}$

5   $\frac{1}{64}$

6  Don't know

*The problem continues on the next page*

**Question 2.2.**

The squared correlation between  $Y_2$  and  $X_1$  is

- 1   $\frac{1}{2}$
- 2   $\frac{1}{4}$
- 3   $\frac{1}{8}$
- 4   $\frac{1}{16}$
- 5   $\frac{1}{64}$
- 6  Don't know

**Question 2.3.**

The squared partial correlation  $\rho_{Y_1 Y_2 | X_1 X_2}^2$  between  $Y_1$  and  $Y_2$  given  $\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$  is

- 1   $\frac{1}{4}$
- 2   $\frac{1}{2}$
- 3   $\frac{6}{7}$
- 4   $\frac{36}{179}$
- 5   $\frac{0.25 - 0.25 \times 0.125}{\sqrt{(1 - 0.25^2)(1 - 0.125^2)}}$
- 6  Don't know

**Question 2.4.**

The conditional mean  $E(Y_2 | \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix})$  is

- 1   $x_1$
- 2   $x_2$
- 3   $\frac{7}{96}x_1 + \frac{1}{4}x_2$
- 4   $\frac{7}{24}x_1 - \frac{1}{24}x_2$
- 5   $\frac{1}{4}x_1 + x_2$
- 6  Don't know

*The problem continues on the next page*

### Question 2.5.

The conditional mean  $E(Y_1 | \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix})$  is

1   $\frac{7}{24}x_1 - \frac{3}{8}x_2$

2   $\frac{7}{24}x_1 - \frac{1}{24}x_2$

3   $\frac{7}{96}x_1 + x_2$

4   $x_2$

5   $x_1$

6  Don't know

### Question 2.6.

The squared multiple correlation  $\rho_{Y_2|X_1X_2}^2$  between  $Y_2$  and  $[X_1 \ X_2]'$  is

1   $\frac{3 \times 192 \times 1}{4 \times 179 \times 3}$

2   $\frac{3}{4}$

3   $\frac{1}{4}$

4   $\frac{1}{3} \times [179/192 \quad 3/4] \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 179/192 \\ 3/4 \end{bmatrix} / 1$

5   $\frac{1 \times 192 \times 1}{16 \times 13 \times 1}$

6  Don't know

### Question 2.7.

The squared multiple correlation  $\rho_{Y_1|X_1X_2}^2$  between  $Y_1$  and  $[X_1 \ X_2]'$  is

1   $\frac{1 \times 192 \times 1}{16 \times 13 \times 1}$

2   $[0.25 \quad 1] \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0.25 \\ 1 \end{bmatrix} / 4$

3   $\frac{9 \times 192}{16 \times 179 \times 3}$

4   $\frac{4 \times (1 - 0.25^2)}{4}$

5   $\frac{13}{192}$

6  Don't know

## Problem 3.

Enclosure B with SAS program and SAS output belongs to this problem. The data are taken from a study reported in “Camilla Himmelstrup Trinderup (2014): Spectral Imaging of Meat Quality - Color and Texture, PHD-201 4-358, DTU Compute, Lyngby”. The problem investigated is the development over time of the diameters of salamis during fermentation. The data considered here comprise diameter measurements (in pixels) on salamis using two different starter cultures for the fermentation (type=1 or type=2). The measurements were taken 2, 3, 9, 14, 21, and 42 days after production.

We are now interested in to which extent the time development depends on the choice of starter culture. We consider the following model and hypotheses.

$$M: \quad E(Y_{tj}) = \alpha_j + \beta_j t + \gamma_j t^2 \quad j = 1,2$$

$$H_1: \quad E(Y_{tj}) = \alpha_j + \beta t + \gamma t^2 \quad j = 1,2$$

$$H_2: \quad E(Y_{tj}) = \alpha + \beta t + \gamma t^2 \quad j = 1,2$$

In all three cases we assume that the errors are independent and normally distributed with the same variance.

### Question 3.1.

What fraction of the total variation is explained by model  $H_2$ ?

- 1  99.60%
- 2  98.56%
- 3  65.36%
- 4  40.51%
- 5  99.51%
- 6  Don't know.

*The problem continues on the next page*

### Question 3.2.

If we assume the model  $M$  and want to test the hypothesis  $H_1$ , the usual test statistic becomes

1   $\frac{(207.74 - 167.56)/2}{167.56/6}$

2  113568

3  173.52

4  359.74

5   $\frac{(207.74 - 167.56)/2}{207.74/8}$

6  Don't know.

### Question 3.3.

The distribution of the test statistic under the hypothesis is?

1   $F(1, 6)$

2   $F(1, 8)$

3   $F(6, 12)$

4   $F(2, 6)$

5   $F(2, 8)$

6  Don't know.

### Question 3.4.

The usual test statistic for testing  $H_2$  assuming that  $H_1$  is true becomes

1  471.85

2   $\frac{(608.33593 - 167.56)/3}{207.74/8}$

3  266.39

4   $\frac{608.33593 - 207.74}{207.74/8}$

5   $\frac{(608.33593 - 167.56)/3}{167.56/6}$

6  Don't know.

*The problem continues on the next page*

### Question 3.5.

We now consider model  $H_2$ . Which observation will cause the largest overall change in the estimated parameter vector if it is omitted from the estimation ?

- 1  4
- 2  6
- 3  8
- 4  10
- 5  12
- 6  Don't know.

### Question 3.6.

For model  $H_2$  the usual 95% confidence interval for the mean value of an observation at day 42 using starter culture type 2 is

- 1   $[1185.835 - 1.96 \times 5.78796, 1185.835 + 1.96 \times 5.78796]$
- 2   $[1185.78 - t(9)_{0.975} \times 5.78796, 1185.78 + t(9)_{0.975} \times 5.78796]$
- 3   $[1185.78 - t(9)_{0.975} \times \sqrt{67.59288}, 1185.78 + t(9)_{0.975} \times \sqrt{67.59288}]$
- 4   $[1185.78 - t(9)_{0.975} \times \sqrt{5.78796/0.49562}, 1185.78 + t(9)_{0.975} \times \sqrt{5.78796/0.49562}]$
- 5   $[1179.67 - t(9)_{0.975} \times 5.78796, 1179.67 + t(9)_{0.975} \times 5.78796]$
- 6  Don't know.

### Question 3.7.

For model  $H_2$  the usual 95% prediction interval for a new observation at day 42 using starter culture type 2 is

- 1   $[1185.835 - t(9)_{0.975} \sqrt{5.78796^2 + 1}, 1185.835 + t(9)_{0.975} \sqrt{5.78796^2 + 1}]$
- 2   $[1185.78 - t(9)_{0.975} \times 5.78796, 1185.78 + t(9)_{0.975} \times 5.78796]$
- 3   $[1185.78 - t(9)_{0.975} \times \sqrt{67.59288 + 1}, 1185.78 + t(9)_{0.975} \times \sqrt{67.59288 + 1}]$
- 4   $[1185.78 - t(9)_{0.975} \times \sqrt{5.78796^2 + 67.59288}, 1185.78 + t(9)_{0.975} \times \sqrt{5.78796^2 + 67.59288}]$
- 5   $[1179.67 - t(9)_{0.975} \times \sqrt{5.78796^2 + 1}, 1179.67 + t(9)_{0.975} \times \sqrt{5.78796^2 + 1}]$
- 6  Don't know.

## Problem 4.

Enclosure C with SAS program and SAS output belongs to this problem. The data are taken from a yet unpublished study at DTU Compute on the possible relation between some specific cell measurements using image analysis and possible heterogeneity in a donor population. The data used here comprise measurements on 20 cells from each of two donors (number 3 and number 4). The variables are

1. *areacell*, the logarithm of the cell area
2. *areanucl*, the logarithm of the nucleus fraction of the cell area
3. *avecytop*, the average cytoplasma intensity for the cell

We are now interested in how suitable the cell measurements are in distinguishing between the donor types represented by the two donors actually investigated.

### Question 4.1.

What is the absolute value  $|t|$  of the usual t-test statistic for assessing whether the mean values for *areacell* are the same for the two donors

- 1   $\sqrt{8.58}$
- 2   $8.58^2$
- 3  8.58
- 4   $\left| \frac{1-8.58^{1/2}}{8.58^{1/2}} \right|$
- 5   $\sqrt[3]{8.58}$
- 6  Don't know.

### Question 4.2.

What is the F-test statistic for assessing whether the mean values for *avecytop* are the same for the two donors

- 1  0.35
- 2  12.82
- 3  9.94
- 4  0.97
- 5  3.99
- 6  Don't know.

*The problem continues on the next page*

### **Question 4.3.**

If we test whether the mean values for [*areacell*, *areanucl*, *avecystop*] are the same for the two donors, then the distribution of the MANOVA test statistic( assuming that the hypothesis is true) becomes

- 1  U(3, 1, 38)
- 2  U(2, 2, 38)
- 3  U(1, 3, 38)
- 4  U(1, 1, 36)
- 5  U(3, 1, 36)
- 6  Don't know.

### **Question 4.4.**

Proc Discrim computes the squared Mahalanobis distance between the two donors. The F-test statistic corresponding to this squared Mahalanobis distance between the two donors using all three variables is

- 1   $0.7262120029 \times 1.0422203745 \times 505.60774572$
- 2  4.14
- 3  1.31103
- 4   $(-3.46803)^2$
- 5   $\frac{1-1.31103^{1/3}}{1.31103^{1/3}}$
- 6  Don't know.

### **Question 4.5.**

If we assume that the prior probabilities are the same, we will classify a cell as belonging to donor (type) 3 if [*areacell*, *areanucl*, *avecystop*][*a b c*]' + *d* > 0, where *a*, *b*, *c* and *d* are constants. The first coefficient *a* is

- 1  -11
- 2  357.55281
- 3  0.1064
- 4  0.9382
- 5  356.61461
- 6  Don't know.

*The problem continues on the next page*

### **Question 4.6.**

The fraction of correctly classified cells from donor 3 using resubstitution is

- 1  60%
- 2  65%
- 3  70%
- 4  75%
- 5  80%
- 6  Don't know.

### **Question 4.7.**

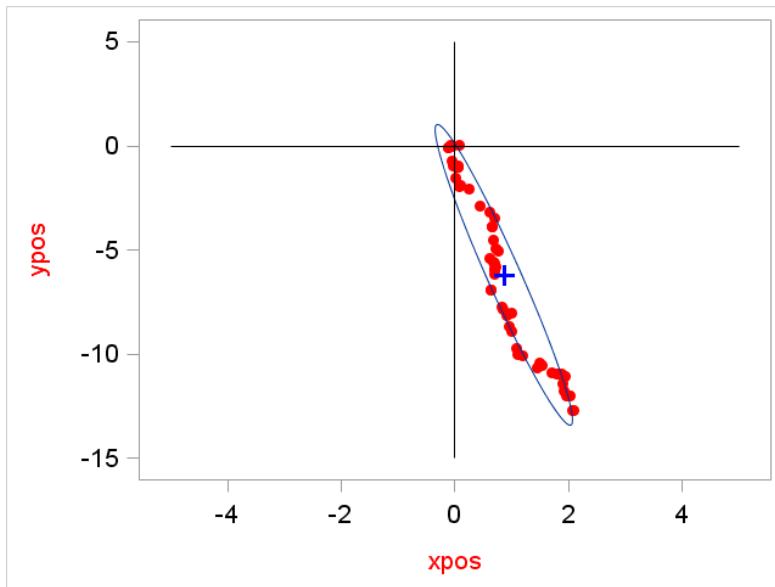
The fraction of correctly classified cells is reduced if we use cross validation instead of resubstitution. The reduction is

- 1  0%
- 2  5%
- 3  10%
- 4  15%
- 5  20%
- 6  Don't know.

### **Problem 5.**

Enclosure D with SAS program and SAS output belongs to this problem. The data are taken from a study on the position of a certain phenomenon in microscopy images of materials samples. In the figure below is shown 58 such positions and the first 5 observations are given in the SAS output in Enclosure D.

*The problem continues on the next page*



### Question 5.1.

What is the value of the first principal component corresponding to observation no 1?

- 1   $-0.1597 \times 0.8598 - 0.9872 \times 6.2359$
- 2   $0.4963 \times 1.2180 - 2.8596 \times 6.4716$
- 3   $0.4963 \times 0.8598 - 0.9872 \times 6.2359$
- 4   $0.9872 \times 2.0779 - 0.1597 \times 12.7075$
- 5   $-0.1597 \times 1.2180 - 0.9872 \times 6.4716$
- 6  Don't know.

### Question 5.2.

The value of the second principal component corresponding to observation no 1 lies in the interval

- 1   $[-2.0, -1.0]$
- 2   $[-1.0, 0.0]$
- 3   $[0.0, 1.0]$
- 4   $[1.0, 2.0]$
- 5   $[2.0, 3.0]$
- 6  Don't know.

*The problem continues on the next page*

### Question 5.3.

If we assume that the data are normally distributed, then the major axis of the contour ellipse for the estimated probability density function lies on the line with the equation

- 1   $0.16xpos + 0.99 ypos + 0.15 = 0$
- 2   $-0.99xpos + 0.16 ypos + 0.15 = 0$
- 3   $-0.16xpos + 0.99 ypos + 0.15 = 0$
- 4   $0.99xpos - 0.16 ypos + 0.15 = 0$
- 5   $0.99xpos + 0.16 ypos + 0.15 = 0$
- 6  Don't know.

## Enclosure A – SAS program

Page 1 of 8

```

/*The data below are taken from "Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado (2015): Modeling and Analysis of Compositional Data, xlii + 247 pp., John Wiley & Sons, Ltd." */

proc print data=sasuser.protein;
var Country GeoPol RMMEAT WMEAT EGG MILK CEREAL STARCH NUTS VEG;
Title 'The Nutrition Data in the Protein Data Set';
ods graphics on;

proc factor data=sasuser.protein plots=(scree initloadings(vector) loadings(vector))
nfactors=3
rotate=varimax
out=scrprotein;
var RMMEAT WMEAT EGG MILK FISH CEREAL STARCH NUTS VEG;
Title "Factor Analysis on the Nutrition Data";
run;
ods graphics on/width=900px height=650px;

/* proc template/ proc sgrender statements produce a plot of the factor scores of the observations */
proc template;
define statgraph facso;
begingraph;
entrytitle 'Factorscores';
entrytitle 'FactorScores';
discreteattrmap name="pro1" / ignorecase=true;
value "SouE" / markerattrs=GraphData1(color=red symbol=circlefilled);
value "SouW" / markerattrs=GraphData2(color=green symbol=trianglefilled);
value "CenE" / markerattrs=GraphData3(color=blue symbol=squarefilled);
value "CenW" / markerattrs=GraphData4(color=nagenta symbol=diamondfilled);
value "NorW" / markerattrs=GraphData5(color=black symbol=triangledownfilled);
enddiscreteattrmap;
discreteattrvar attrvar=GeoPol var=GeoPol attrmap="pro1";
layout overlay;
scatterplot x=factor1 y=factor2 /
group=GeoPol
datalabel=country
name="protein";
discretelegend "protein" / title="GeoPol";
endlayout;
endgraph;
end; run;

proc sgrender data=sasuser.protein template=facsco;
Title "Factor Scores Plot for Nutrition Data";
ods graphics off;

```

## Enclosure A – SAS output

Page 2 of 8

### The Nutrition Data in the Protein Data Set

Obs	Country	GeoPol	RMMEAT	WMEAT	EGG	MILK	FISH	CEREAL	STARCH	NUTS	VEG
1	ALB	SouE	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	AUS	CenW	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
3	BEL	CenW	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
4	BRD	CenW	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
5	BUL	SouE	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
6	CZE	CenE	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
7	DDR	CenE	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
8	DEN	NorW	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
9	FIN	NorW	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
10	FRA	CenW	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
11	GRE	SouW	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
12	HOL	CenW	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
13	HUN	CenE	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
14	IRE	NorW	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
15	ITA	SouW	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
16	NOR	NorW	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
17	POL	CenE	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
18	POR	SouW	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
19	ROM	SouE	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
20	SCH	CenW	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
21	SPA	SouW	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
22	SWE	NorW	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
23	UK	NorW	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
24	USSR	CenE	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
25	YUG	SouE	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

## Enclosure A – SAS Output

## Factor Analysis on the Nutrition Data

## The FACTOR Procedure

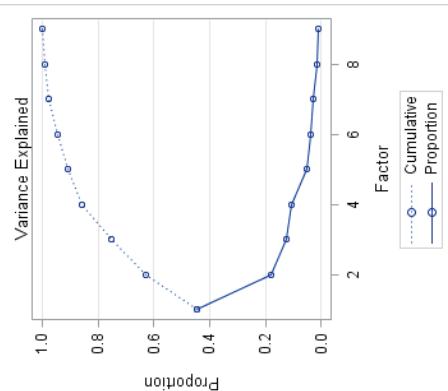
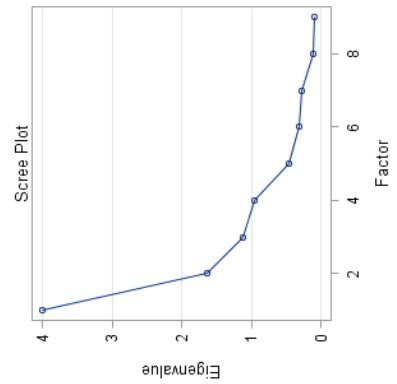
## Initial Factor Method: Principal Components

Prior Community Estimates: ONE

Eigenvalues of the Correlation Matrix: Total  
= 9 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.00643757	2.37143813	0.4452	0.4452
2	1.63499945	0.50707994	0.1817	0.6268
3	1.12791950	0.17325554	0.1253	0.7522
4	0.95466396	0.49082557	0.1061	0.8582
5	0.4633840	0.13870742	0.0515	0.9098
6	0.32513097	0.05352464	0.0361	0.9459
7	0.27160633	0.15531443	0.0302	0.9761
8	0.11629190	0.01718000	0.0129	0.9890
9	0.09911190		0.0110	1.0000

3 factors will be retained by the NFACTOR criterion.



## Enclosure A – SAS Output

The FACTOR Procedure  
Initial Factor Method: Principal Components

## Factor Pattern

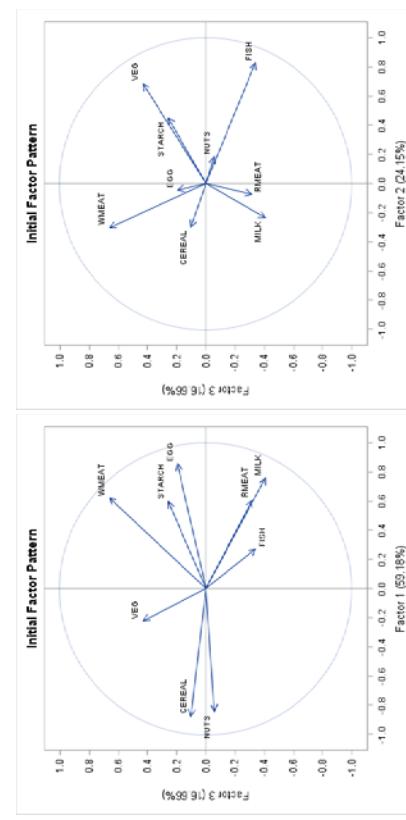
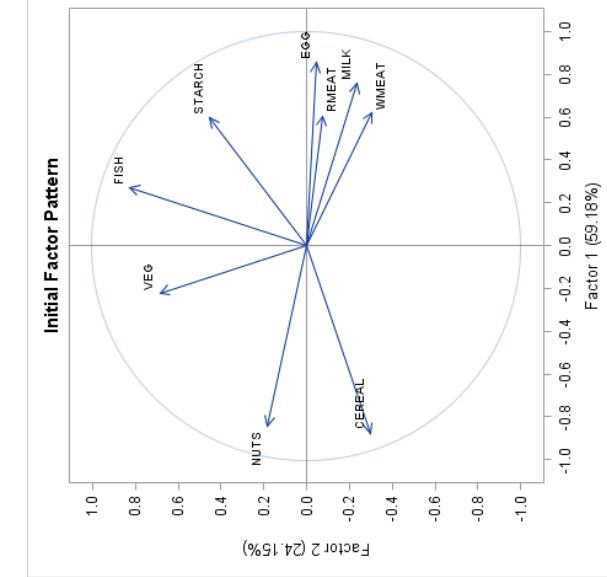
	Factor Pattern		
	Factor1	Factor2	Factor3
RMEAT	0.60571	-0.07193	-0.31604
WMEAT	0.62161	-0.30286	0.66260
EGG	0.85404	-0.04518	0.19279
MILK	0.75606	-0.23603	-0.40958
FISH	0.27152	0.82707	-0.34120
CEREAL	-0.87619	-0.29855	0.10187
STARCH	0.59497	0.45115	0.25805
NUTS	-0.84135	0.18325	-0.05776
VEG	-0.22102	0.68561	0.43284

## Final Communality Estimates: Total = 6.769357

	RMEAT	WMEAT	EGG	MILK	FISH	CEREAL	STARCH	NUTS	VEG
	0.471934	0.917164	0.768599	0.795097	0.874187	0.867220	0.624117	0.744777	0.706260

## Enclosure A – SAS output

## Enclosure A – SAS Output



**The FACTOR Procedure**  
Rotation Method: Varimax

Orthogonal Transformation Matrix			
	1	2	3
RMEAT	0.68520	0.63792	0.35151
WMEAT	-0.11758	-0.37940	0.91773
EGG	0.71880	-0.67016	-0.18496



Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Total
2.4863857	2.3723048	1.9106681	6.769357

Final Communality Estimates: Total = 6.769357					
RMEAT	WMEAT	EGG	MILK	FISH	CEREAL
0.471934	0.917164	0.768599	0.795097	0.874187	0.867220

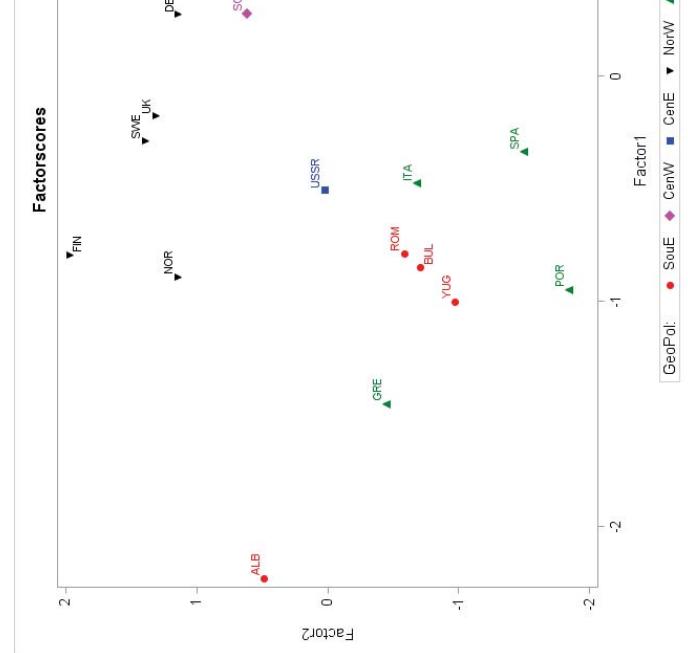
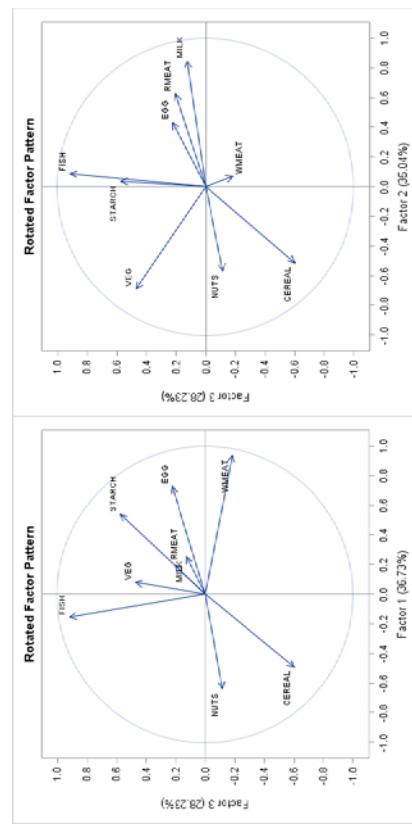
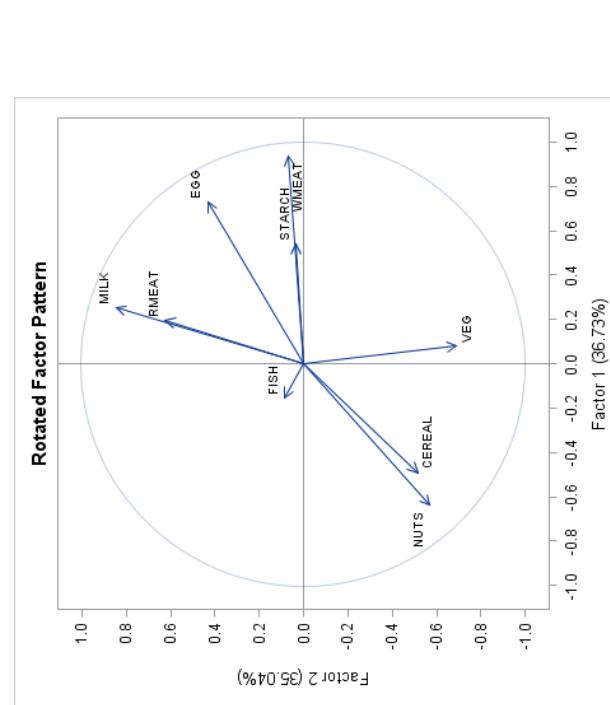
NUTS      VEG

## Scoring Coefficients Estimated by Regression

Standardized Scoring Coefficients			
	Factor1	Factor2	Factor3
RMEAT	-0.09264	0.30091	0.06459
WMEAT	0.55035	-0.22443	-0.22411
EGG	0.27217	0.03192	0.01796
MILK	-0.11474	0.41851	0.00101
FISH	-0.23049	0.05404	0.54401
CEREAL	-0.06346	-0.13076	-0.26116
STARCH	0.23316	-0.16327	0.26312
NUTS	-0.19388	-0.14217	0.03851
VEG	0.18874	-0.45146	0.29447

## Enclosure A – SAS output

## Enclosure A – SAS Output



## Enclosure B – SAS program

**/\*Credit for the data is given in Problem 3.\*/**

```

proc print data=sasuser.salami;
Title 'Salami Data';

proc glm data sasuser.salami;
class type;
model width = type day(type) day*day(type) / noint solution;
Title 'Model: width = type day(type) day*day(type) / noint';
run;

ods graphics on;

proc glm data = sasuser.salami;
class type;
model width = type day day*day / noint solution;
Title 'Model: width = type day day*day / noint';
run;

proc glm data = sasuser.salami;
class type;
model width = day day*day / solution;
output out=ResStat predicted=PREDICTED stdp=StdErrMPV residual=RESIDUAL
student=STUDENTR rstudent=RSTUDENT cookd=CooksD dfits=DFFITS h=HATDiagH;
Title 'Model: width = day day*day with intercept';
run;

proc print data=ResStat; run;

/*Naming of variables in the output dataset ResStat used above:
ResStat: the dataset containing the residuals and influence diagnostics
PREDICTED: the mean predicted value
StdErrMPV: the standard error of the mean predicted value
RESIDUAL: usual naming
STUDENTR: the STUDENT RESIDUAL
RSTUDENT: usual naming
CooksD: usual naming
DFFITS: usual naming
HATDiagH: the leverage, i.e. the i'th diagonal element in the Hat-matrix*/

ods graphics off;

```

## Enclosure B – SAS output

**Salami Data**

obs	day	type	width
1	2	1	1335.00
2	2	2	1341.67
3	3	1	1322.00
4	3	2	1326.00
5	9	1	1265.33
6	9	2	1277.33
7	14	1	1218.67
8	14	2	1246.67
9	21	1	1192.00
10	21	2	1198.33
11	42	1	1179.67
12	42	2	1192.00

## Enclosure B – SAS output

**Model: width = type day(type) day\*day(type)/oint**  
The GLM Procedure

Class Level Information		
Class	Levels	Values
type	2	1 2

Number of Observations Read	
Number of Observations Used	
12	

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	19029406.22	3171567.70	113568	<.0001
Error	6	167.56	27.93		
Uncorrected Total	12	19029573.78			

Dependent Variable: width

Source			DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19029366.04		4757341.51	183201	<.0001	
Error	8	207.74		25.97			
Uncorrected Total	12	19029573.78					

Number of Observations Read		
Number of Observations Used		
12		
12		

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	2	18987814.07	9493907.04	339958	<.0001
day(type)	2	31900.42	15950.21	571.15	<.0001
day*day(type)	2	9691.72	4845.86	173.52	<.0001

R-Square	Coeff Var	Root MSE	width Mean
0.996026	0.420114	5.284572	1257.889

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	2	4796533.417	2398266.708	85877.2	<.0001
day(type)	2	20992.908	10046.454	359.74	<.0001
day*day(type)	2	9691.720	4845.860	173.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	1	1356.682888	4.63762194	292.54	<.0001
type 2	1	1361.405698	4.63762194	293.56	<.0001
day(type) 1	-11.958993	0.60444389	-19.79		
day(type) 2	-10.947553	0.60444389	-18.11		
day*day(type) 1	0.184704	0.01326793	13.92		
day*day(type) 2	0.164246	0.01326793	12.38		

Parameter	Estimate	Standard Error	t Value	Pr >  t
type 1	1353.266515	3.48761964	388.02	<.0001
type 2	1364.822071	3.48761964	391.33	<.0001
day	-11.453273	0.41214472	-27.79	<.0001
day*day	0.174475	0.00904684	19.29	<.0001

## Enclosure B – SAS output

**Model: width = type day(day)\*oint**  
The GLM Procedure

Class Level Information		
Class	Levels	Values
type	2	1 2

Number of Observations Read	
Number of Observations Used	
12	
12	

Dependent Variable: width

Source			DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19029366.04		4757341.51	183201	<.0001	
Error	8	207.74		25.97			
Uncorrected Total	12	19029573.78					

Source	DF	R-Square	Coeff Var	Root MSE	width Mean
Model	4	0.995073	0.405113	5.095873	1257.889

Source	DF	R-Square	Coeff Var	Root MSE	width Mean
Model	4	0.995073	0.405113	5.095873	1257.889

## Enclosure B – SAS output

**Model: width = type day(day)\*oint**  
The GLM Procedure

Class Level Information		
Class	Levels	Values
type	2	1 2

Number of Observations Read	
Number of Observations Used	
12	
12	

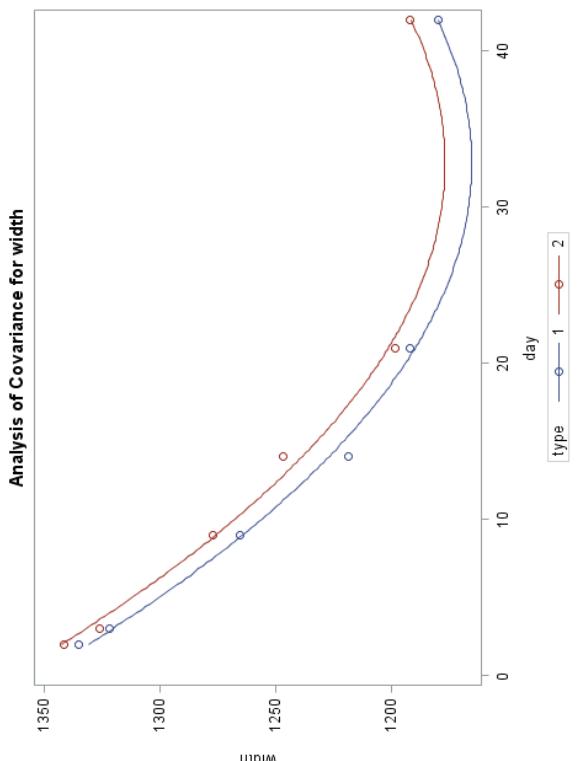
Dependent Variable: width

Source			DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19029366.04		4757341.51	183201	<.0001	
Error	8	207.74		25.97			
Uncorrected Total	12	19029573.78					

Source	DF	R-Square	Coeff Var	Root MSE	width Mean
Model	4	0.995073	0.405113	5.095873	1257.889

Source	DF	R-Square	Coeff Var	Root MSE	width Mean
Model	4	0.995073	0.405113	5.095873	1257.889

## Enclosure B – SAS output

**Analysis of Covariance for width**

**Model: width = day day\*day with intercept**  
The GLM Procedure

Class Level Information	
Class	Level
type	2
	1 2

Number of Observations Read | 12 |  
Number of Observations Used | 12 |

**Dependent Variable: width**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41551.96036	20775.98018	307.37	<.0001
Error	9	608.33593	67.59288		
Corrected Total	11	42160.29629			

R-Square	Coeff Var	Root MSE	width Mean
0.985571	0.653594	8.221489	1257.889

## Enclosure B – SAS output

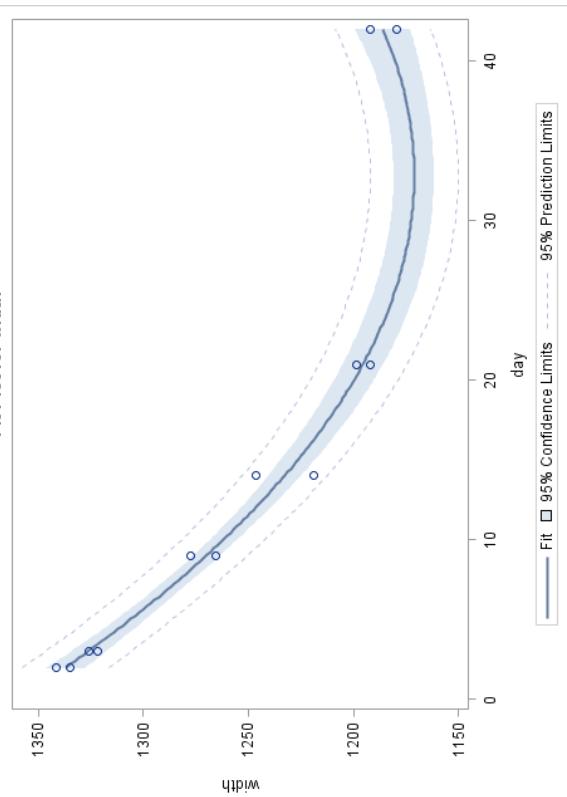
**Model: width = day day\*day with intercept**  
The GLM Procedure

Source	DF	Type I SS	Mean Square	F Value	Pr > F
day	1	31893.44040	31893.44040	471.85	<.0001
day*day	1	9658.51996	9658.51996	142.89	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
day	1	20053.80956	20053.80956	296.69	<.0001
day*day	1	9658.51996	9658.51996	142.89	<.0001

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	1359.044293	5.10177156	266.39	<.0001
day	-11.453273	0.66493877	-17.22	<.0001
day*day	0.1174475	0.01459583	11.95	<.0001

Fit Plot for width



## Enclosure B – SAS output

Page 7 of 7

Model: width = day day\*day with intercept

Obs	day	type	width	PREDICTED	StdErrMPV	RESIDUAL	STUDENT	rSTUDENT	CooksD	DFFITS	HATDiagH
1	2	1	1335.00	1336.84	4.10299	-1.8356	-0.25765	-0.24382	0.00734	-0.14042	0.24906
2	2	2	1341.67	1336.84	4.10299	4.8310	0.67809	0.65629	0.05083	0.37796	0.24906
3	3	1	1322.00	1326.25	3.69001	-4.2547	-0.57912	-0.55647	0.02820	-0.27949	0.20144
4	3	2	1326.00	1326.25	3.69001	-0.2547	-0.03467	-0.03269	0.00010	-0.01642	0.20144
5	9	1	1265.33	1270.10	2.76387	-4.7640	-0.61532	-0.59273	0.01611	-0.21175	0.11318
6	9	2	1277.33	1270.10	2.76387	7.2360	0.93461	0.92731	0.03716	0.33127	0.11318
7	14	1	1218.67	1232.90	3.42423	-14.2289	-1.90367	-2.32224	0.25353	-1.06387	0.17347
8	14	2	1246.67	1232.90	3.42423	13.7711	1.84242	2.20103	0.23748	1.00835	0.17347
9	21	1	1192.00	1195.47	4.25002	-3.4691	-0.49292	-0.47114	0.02954	-0.28451	0.26723
10	21	2	1198.33	1195.47	4.25002	2.8642	0.40698	0.38729	0.02013	0.23388	0.26723
11	42	1	1179.67	1185.78	5.78796	-6.1143	-1.04717	-1.05354	0.35917	-1.04436	0.49562
12	42	2	1192.00	1185.78	5.78796	6.2191	1.06511	1.07418	0.37159	1.06481	0.49562

## Enclosure C – SAS program

Page 1 of 7

*/\*The data in data set sasuser.cells34 are taken from a yet unpublished study at DTU Compute. \*/*

```

proc print data=sasuser.cells34;
Title 'Print of Donor Data';

proc glm data=sasuser.cells34;
class donor;
model areacell areanucl avecytop = donor;
manova h=donor/printn;
Title 'MANOVA on Donor Data'; run;

proc discrim data=sasuser.cells34
method=normal
pool=yes
crossvalidate
class donor;
var areacell areanucl avecytop;
Title 'Discriminant Analysis on Donor Data'; run;

/*Proc g3d produces the three-dimensional scatter plot of the variables.*/
options reset=all border device=activex;

data cells;
set sasuser.cells34;
length color shape $8;
if donor=3 then do; shape="balloon"; color="blue"; end;
if donor=4 then do; shape="balloon"; color="red"; end; run;

proc g3d data=cells;
scatter areacell*areanucl=avecytop/
needle
grid
size=2
color=color
shape=shape
rotate=15 tilt=60;
Title 'Scatterplot of the 3 Variables'; run;

```

## Enclosure C – SAS output

Page 2 of 7

### Print of Donor Data

Obs	donor	areacell	areanucl	avecytop
1	3	9.8575	-1.10239	16.8900
2	3	10.0964	-0.45037	10.8560
3	3	9.7865	-0.94477	14.8680
4	3	10.1149	-0.34893	9.6289
5	3	9.8337	-1.05338	16.4650
6	3	9.7164	-0.29908	5.2071
7	3	9.9957	-0.93055	19.1380
8	3	9.9456	-1.15842	18.7950
9	3	9.7685	-0.45365	8.2803
10	3	10.5485	-0.78963	30.4190
11	3	10.3181	-1.11335	22.0880
12	3	9.9011	-1.00683	17.2810
13	3	9.9581	-1.05485	19.7250
14	3	9.9722	-1.06380	19.4770
15	3	9.9188	-1.17094	18.8260
16	3	9.9865	-1.10721	20.1960
17	3	9.3313	-0.26531	3.1809
18	3	10.1280	-1.19205	25.1210
19	3	9.8315	-0.96812	16.3640
20	3	9.9418	-0.89539	11.5690
21	4	9.5350	-0.16803	2.5070
22	4	9.7966	-0.93315	14.5420
23	4	9.7716	-0.776750	12.3760
24	4	10.0449	-1.01293	18.6900
25	4	9.8601	-0.77070	13.0090
26	4	9.9880	-1.08365	18.3120
27	4	10.0573	-0.40352	8.7063
28	4	9.8320	-0.90553	14.8920
29	4	9.2426	-0.19479	2.0411
30	4	9.1230	-0.26909	2.0096
31	4	9.8361	-0.81242	13.5800
32	4	10.3750	-0.45832	15.1100
33	4	9.6057	-0.24564	4.3744
34	4	9.8459	-0.22116	4.3234
35	4	9.5972	-0.72924	9.9430
36	4	9.6772	-0.74791	11.2510
37	4	9.1902	-0.12910	1.5092
38	4	9.3905	-0.17238	2.2766
39	4	9.6177	-0.71229	10.5370
40	4	9.1747	-0.17497	2.1524

## Enclosure C – SAS output

**MANOVA on Donor Data  
The GLM Procedure**

Class Level Information		
Class	Levels	Values
donor	2	3 4

Number of Observations Read			
Number of Observations Used			

Dependent Variable: areacell

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.72621200	0.72621200	8.58	0.0057
Error	38	3.21492881	0.08460339		
Corrected Total	39	3.94114081			

R-Square	Coeff Var	Root MSE	areacell Mean
0.184264	2.964151	0.290867	9.812815

Source	DF	Type I SS	Mean Square	F Value	Pr > F
donor	1	0.72621200	0.72621200	8.58	0.0057

Source	DF	Type III SS	Mean Square	F Value	Pr > F
donor	1	0.72621200	0.72621200	8.58	0.0057

Similar results for dependent variables "areanucl" and "avecytop" have intentionally been omitted!

## Enclosure C – SAS output

**The GLM Procedure  
Multivariate Analysis of Variance**

E = Error SSCP Matrix			
	areacell	areanucl	avecytop
areacell	3.2149288082	-1.695602673	53.002800162
areanucl	-1.695602673	3.9859687177	-65.35021642
avecytop	53.002800162	-65.35021642	1498.2752357

Partial Correlation Coefficients from the Error SSCP Matrix / Pr >  r			
DF = 38		areacell	areanucl
areacell	1.000000	-0.473665	0.763690
areanucl	-0.473665	1.000000	<.0001
avecytop	0.0023	-0.845638	<.0001

H = Type III SSCP Matrix for donor			
DF		areacell	areanucl
areacell	0.7262120029	-0.869984451	19.161900055
areanucl	-0.869984451	1.0422203745	-22.95549377
avecytop	19.161900055	-22.95549377	505.60774572

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for donor E = Error SSCP Matrix			
S=1 N=17		Characteristic Vector VEV=1	
Characteristic RootPercent	areacell	areanucl	avecytop
0.34500760	100.00	0.13292259	-0.132297260.01507793
0.00000000	0.00	-0.81273507	0.967234430.07471581
0.00000000	0.00	0.56240647	0.469463940.00000000

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall donor Effect					
H = Type III SSCP Matrix for donor					
E = Error SSCP Matrix	S=1 N=17	Value	F Value	Num DF	Den DF
Wilks' Lambda		0.74349022	4.14	3	36
Pillai's Trace		0.25650978	4.14	3	36
Hotelling-Lawley Trace		0.34500760	4.14	3	36
Roy's Greatest Root		0.34500760	4.14	3	36

Pr > F

## Enclosure C – SAS output

**Discriminant Analysis on Donor Data****The DISCRIM Procedure**

Total Sample Size	40	DF Total	39
Variables	3	DF Within Classes	38
Classes	2	DF Between Classes	1

Number of Observations Read	40
Number of Observations Used	40

Class Level Information				
donor	Variable Name	Frequency	Weight	Prior Probability
3	3	20	20.0000	0.500000
4	4	20	20.0000	0.500000

Pooled Covariance Matrix Information			
Natural Log of the Covariance Determinant of the Covariance Matrix Rank			
			-3.46803

Generalized Squared Distance to donor			
From donor	3	4	
3	0	1.31103	
4	1.31103	0	

Linear Discriminant Function for donor			
Variable	3	4	
Constant	-1.695	-1.684	
areacell	357.55281	356.61461	
areanucl	-199.41315	-198.47936	
avecytop	-20.93522	-21.04164	

## Enclosure C – SAS Output

**The DISCRIM Procedure**  
**Classification Summary for Calibration Data: WORK.CELLS**  
**Resubstitution Summary using Linear Discriminant Function**

Number of Observations and Percent Classified into donor			
From donor	3	4	Total
3	15	5	20
	75.00	25.00	100.00
4	7	13	20
	35.00	65.00	100.00
<b>Total</b>	<b>22</b>	<b>18</b>	<b>40</b>
<b>Priors</b>	<b>0.5</b>	<b>0.5</b>	

Error Count Estimates for donor			
Rate	3	4	Total
0.2500	0.3350	0.3000	
0.5000	0.5000		

**The DISCRIM Procedure**  
**Classification Summary for Calibration Data: WORK.CELLS**  
**Cross-validation Summary using Linear Discriminant Function**

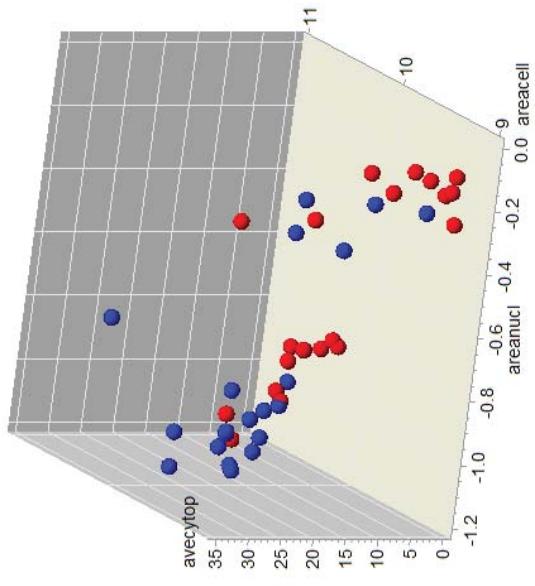
Number of Observations and Percent Classified into donor			
From donor	3	4	Total
3	14	6	20
	70.00	30.00	100.00
4	8	12	20
	40.00	60.00	100.00
<b>Total</b>	<b>22</b>	<b>18</b>	<b>40</b>
<b>Priors</b>	<b>0.5</b>	<b>0.5</b>	

Error Count Estimates for donor			
Rate	3	4	Total
0.3000	0.4000	0.3500	
0.5000	0.5000		

**Enclosure C – SAS output**

Page 7 of 7

**Scatterplot of the 3 Variables**



## Enclosure D – SAS program

```

/* The posdata data set consists of 58 observations of the
position of a given phenomenon in materials science. The data
are a subsample of a data set analyzed at DTU Compute. The
specific nature of the data is not relevant for solving the
present problem */

data posdata;
set sasuser.posdata;
run;

proc print data=posdata(Obs=5);
Title "The 5 First Observations in the Position Data Set";
run;

ods graphics on;

proc princomp data=posdata plots=(scree score) cov;
var xpos ypos;
Title 'Principal Component Analysis of the Position Data Set';
run;

ods graphics off;

```

## Enclosure D – SAS Output

## The 5 First Observations in the Position Data Set

Obs	num	xpos	ypos
1	14	2.07786	-12.7075
2	15	2.07041	-12.6801
3	29	2.01725	-12.0257
4	38	1.97279	-12.0029
5	42	1.96867	-11.9945

Principal Component Analysis of the Position Data Set  
The PRINCOMP Procedure

Observations	58
Variables	2

Simple Statistics		
	xpos	ypos
Mean	0.8398490508	-6.235902582
StD	0.7044660470	4.207708857

Covariance Matrix		
	xpos	ypos
xpos	0.49627241	-2.85962281
ypos	-2.85962281	17.70481382

Total Variance 18.201086236

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	18.1675667	18.1340472	0.9982	0.9982
2	0.0355195		0.0018	1.0000

Eigenvectors		
	Prin1	Prin2
xpos	-.159745	0.987158
ypos	0.987158	0.159745

## Enclosure D – SAS Output

Page 3 of 3

