STAT302/767 Midterm.

19 September 2019, 50 minutes	35 marks total.
FAMILY NAME:	YOUR ID#:
GIVEN NAME:	767/ 302 (circle)

1 Q1

Vineyards maintain spray diaries, where sprays of fertilisers and pesticides are recorded. Among the variables recorded are the number of sprays of each product and the target (a particular insect, fungal disease, or weed species, as well as "additives," which are substances that make the spray disperse better.) Products have been aggregated in to categories by target and chemistry, with hard ("h") chemistry being the most tightly regulated, soft ("s") the least regulated, and an intermediate category ("?"). 18 variables have been created representing the different target-chemistry combinations. We scale these variables to have standard deviation 1 and then perform a principle components analysis.

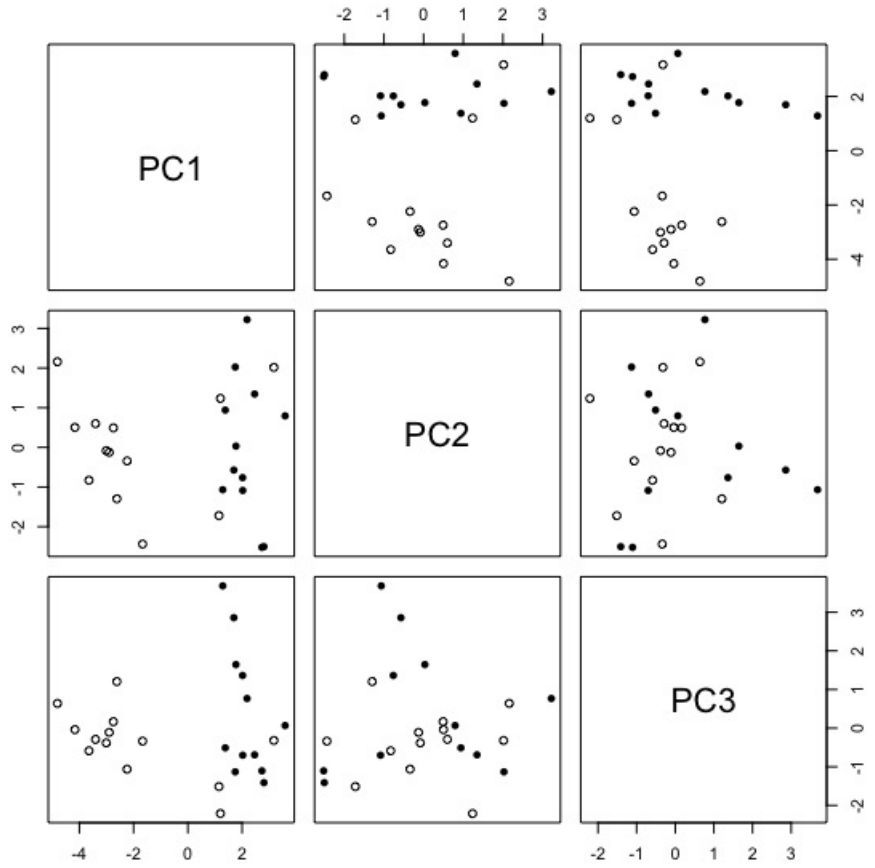
A (2 marks) Discuss the decision to scale the data. In what situations will this be advantageous? In what situations will it be misleading?

B (2 marks)The variances of the resulting principle component scores are given below. Sketch a scree plot.

```
>prcomp(agcount, scale=TRUE)->ag.pr
>round(ag.pr$sdev^2,2)
 [1] 7.39 2.34 1.82 1.22 1.16 0.94 0.79 0.73 0.51 0.33 0.28 0.19
[13] 0.11 0.09 0.05 0.02 0.01 0.00
```

C (3 marks) How many principal components do you suggest using? Explain your reasoning. What proportion of total variability do they account for? (Note that since the variables were initially standardised, the sum of the variances is 18.)

D (1 mark) There are two different management styles, contemporary and future, with future vineyards stating that they strive to eliminate the use of "hard" chemistry. Consider the plot of the first three principal component scores on the following page. Black dots represent vineyards with contemporary management, and open circles represent future management. Identify any component (or combination of components) that separates the two groups reasonably well.


PC1
PC2
PC3
-2 -1 0 1 2 3
-4 -2 0 2
-2 -1 0 1 2 3
-4 -2 0 2
-2 -1 0 1 2 3
-2 -1 0 1 2 3
-2 -1 0 1 2 3

E (2 marks) Below we give the correlations of the original variables and the principle component scores for the first three components. Consider the components identified in (D), and the relative scores of the future and contemporary vineyards. Are the correlations consistent with the stated definitions of future and contemporary management? Explain.

	PC1	PC2	PC3
Additive ?	0.45	0.45	-0.47
Additive h	0.61	-0.18	0.63
Additive s	-0.27	-0.06	-0.07
Botrytis ?	0.77	0.21	0.16
Botrytis h	0.82	-0.04	0.17
Botrytis s	-0.63	0.03	0.02
Downy Mildew ?	-0.56	0.00	0.51
Downy Mildew h	0.76	0.44	-0.05
Downy Mildew s	0.47	-0.48	-0.44
Fertiliser ?	-0.53	0.63	-0.35
Fertiliser s	-0.81	0.13	-0.06
Grasses & Weeds h	0.71	0.39	0.31
Leafroller ?	0.75	0.10	-0.22
Mealy Bug ?	0.49	0.48	-0.22
Mealy Bug h	0.76	-0.32	0.27
Powdery Mildew ?	-0.13	0.59	0.45
Powdery Mildew h	0.92	0.19	-0.08
Powdery Mildew s	-0.56	0.58	0.29

2 Q2

The diet of 215 people is observed, and two sets of variables recorded: 5 macro nutrients (total energy in kJ, carbohydrate, protein, fat, and fibre in grams) and 7 micronutrients (beta-carotene, vitamin C, vitamin A, retinol, vitamin E, vitamin B6 and vitamin B12). A canonical correlation analysis is performed.

A (1 mark) How many pairs of canonical variates will be produced?

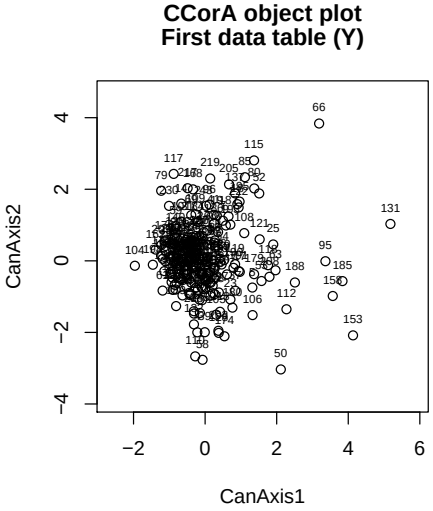
B (4 marks) Two p-values produced by the CCorA function are given below. Explain what null hypothesis they are testing, and the difference between how they are generated. Under what circumstances is each preferred? Do they give the same conclusion in this case? What is that conclusion?

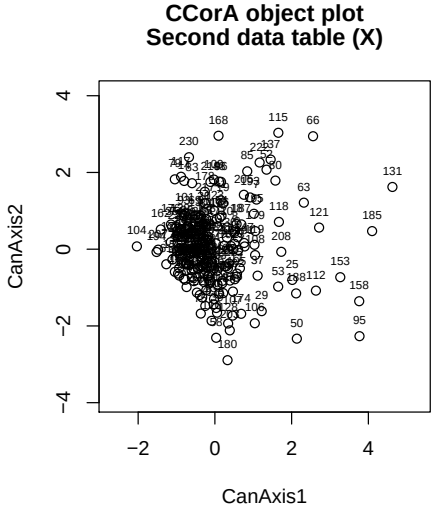
```
> CCorA(micros, macros, permutations=1000)->nutri.CCA
> nutri.CCA$p.perm
[1] 0.000999001
> nutri.CCA$p.Pillai
[1] 4.649273e-123
```

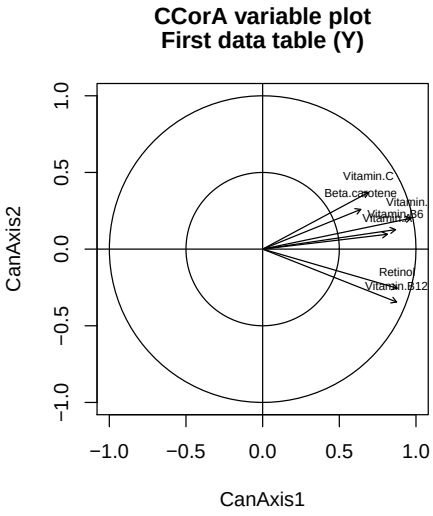
C (2 marks) Examine the RDA-Rsq and RDA-adj-Rsq given below. Comment on the ability to predict someones micro nutrient levels using macronutrient information.

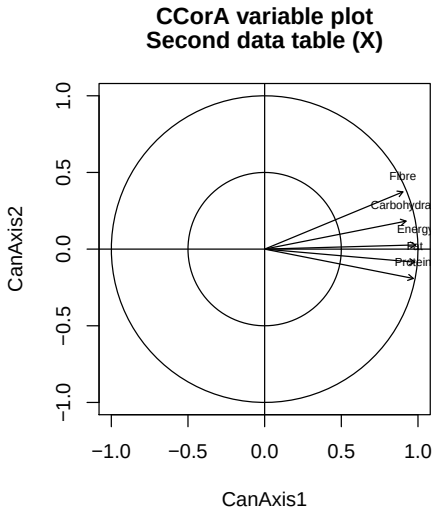
	RDA-R.Sq	RDA-adj-Rsq
micros| macros	0.51	0.50
macros | micros	0.93	0.93

D (2 marks) The biplot is included on the next page. Examine the second row of plots labeled with the variable names. What do the coordinates represent? What does the outer circle represent?


CCorA object plot
First data table (Y)
CanAxis1
CanAxis2


CCorA object plot
Second data table (X)
CanAxis1
CanAxis2


CCorA variable plot
First data table (Y)
Vitamin.C
Beta.carotene
Vitamin.
Vitamin.B6
Retinol
Vitamin.B12
CanAxis1
CanAxis2


CCorA variable plot
Second data table (X)
Fibre
Carbohydra
Energy
Protein
CanAxis1
CanAxis2

E (4 marks) Retinol and vitamin B12 occur only in animal based food sources (milk, meat, eggs). Based on the loadings of the first two canonical variates, which of the macronutrients do you expect to be associated with animal based food sources? Explain your reasoning. In the plot of individuals, where would you expect to find vegan or vegetarian individuals?

```
> round(nutri.CCA$corr.X.Cx,2)
             CanAxis1 CanAxis2 CanAxis3 CanAxis4 CanAxis5
Energy           0.99     0.02     0.16     0.05     0.00
Carbohydrate     0.93     0.17     0.33     0.03     0.05
Protein          0.97    -0.20     0.07    -0.14     0.02
Fat              0.98    -0.08    -0.08     0.15     0.01
Fibre            0.90     0.37     0.04    -0.22     0.00

> round(nutri.CCA$corr.Y.Cy,2)
              CanAxis1 CanAxis2 CanAxis3 CanAxis4 CanAxis5
Beta.carotene     0.63     0.31    -0.17    -0.44    -0.27
Vitamin.C         0.70     0.38     0.16    -0.50    -0.16
Vitamin.A         0.79     0.16    -0.06    -0.34    -0.32
Retinol           0.88    -0.26     0.23     0.03    -0.32
Vitamin.E         0.97     0.21    -0.10    -0.10     0.03
Vitamin.B6        0.87     0.14     0.34    -0.26     0.16
Vitamin.B12       0.88    -0.34     0.09    -0.31     0.04
```

3 Q3

Consider a set of metabolomics data, similar to our first assignment, where the spectral intensity of 333 compounds has been measured on 118 fungal samples. For each compound, a t-test has been performed to examine the difference between two treatment groups (control fungi, and fungi treated with short chain fatty acids). We are interested in discovering compounds whose levels are affected by the treatment.

A (4 marks) Imagine making a histogram of the 333 t-test pvalues. Make two sketches: first, showing what you expect if all the compounds follow the null hypothesis (ie are unaffected by the treatement), and second, what you expect if 20% of the compounds are strongly affected by the treatment (and the other 80% follow the null hypothesis). Put density on the y-axis, put tick marks and labels on both the x- y-axis, and draw roughly to scale.

B (1 mark) If we want to control the family wise type I error rate to be less than 0.05, using the Bonferroni correction, what p-value threshold would we use to declare "discoveries"?

C (1 mark) If we declare tests with a p-value of less than 0.05 to be discoveries, how many discoveries do we expect to find if there are in fact no true differences (ie, the first scenario you sketched above)?

D (1 mark) In fact, 123 p-values are found to be less than 0.05. What is the expected false discovery rate, using the Benjamini and Hochberg method?

E (1 mark) In other situations, we have used linear discriminant analysis, which creates linear combinations of the original variables that maximize an ANOVA f-statistic, to describe the differences between groups (eg treatment and control). What prevents us from using that technique here?

F (4 marks) Suggest an alternate technique that can cope with the problem described in (E). Give the name of the technique, and describe the criteria it optimises.