Lab Assignment No. 8: Answer Key

1) Review your data and report the appropriate univariate statistics. Report the variable correlations. Test whether the correlation matrices for each group have a significantly different structure from a corresponding identity matrix. What can you infer about this dataset? What are you not able to infer from this correlation matrix that a canonical correlation would provide? (**1pt**)

Investigating the dataset there are two sets of variables; one for water oriented indicators and one set specific to soil composition. Both variable sets are focusing on the chemicals contained within each of 96 individual marsh samples taken at different sites.

Before undertaking any specific analyses the variables at hand ought to be reviewed. Table 1 summarizes the univariate moments for the set of 7 variables in the dataset.

Table 1. Variable statistics

Variable	Mean	Median	*SD*	Min	Max	Kurtosis	Skew
Merc_water	80.66	80.67	.19	80.25	81.09	-.57	-.08
Methyl_water	25.99	25.98	.30	25.33	26.63	-.89	.10
Turbidity	.91	.50	.90	.04	5.20	4.18	1.66
Carbon_water	.65	.44	.58	.06	2.86	2.27	1.60
Merc_soil	2.80	1.39	5.06	.33	42.60	41.17	5.77
Sulf_soil	30.44	25.04	33.43	3.72	281.02	35.13	5.27
Phos_soil	198.87	170.34	136.93	5.15	828.62	4.31	-.08

It would appear from Table 1 the average concentration of the different chemicals in both the water and soil samples vary greatly. For instance mercury appears to be much more prevalent in the water samples as opposed to samples taken from the soil.

As with all multivariate tests multivariate normality of the data ought to be investigated. Since there are two distinct groups of variables multivariate normality should be assessed within each variable group (water and soil). In this particular dataset it would seem that multivariate normality did not hold for either group, with the respective Henze-Zirkler tests for water $T = 7.98$, $p < .001$ and for soil $T = 13.00$, $p < .001$. Also, since the covariation between the two variable sets will be analyzed the pooled variable set ought to be tested for multivariate normality. It follows, however, from the single group assessments that no pooled multivariate normality will be achieved ($T = 20.10$ $p < .001$).

The corresponding correlations, summarized in Table 2, for both the individual variable groups and for the cross-group correlations show that the soil sample has all non-significant covariations.

Table 2. Variable correlations

Variable	Merc _water	Methyl _water	Turbidity	Carbon _water	Merc _soil	Sulf _soil	Phos _soil
Merc_water	1.00						
Methyl_water	-.64**	1.00					
Turbidity	-.13	-.09	1.00				
Carbon_water	-.37**	.42**	.46**	1.00			
Merc_soil	-.24*	.33**	.57**	.47**	1.00		
Sulf_soil	-.22*	.22*	.32*	.28*	.64**	1.00	
Phos_soil	.03	-.04	.18	.48**	.06	-.05	1.00

* $p < .05$, ** $p < .001$

It should be pointed out at this point that if we specify *outs* in the *proc corr* statement the corresponding Spearman's rank order correlation will be reported and produced. However, it is the Spearman's product moment correlation that is sought in this example and therefore we ought to specify *outp* instead.

Additionally when conducting a canonical correlation we would want to investigate whether the two groups have approximately equal variance and covariance matrices. However, since it was not demonstrated how to pool the information when unequal variable group are compared we can investigate whether the correlations within each group are equivalent in structure to their corresponding identity matrices.

The water variable group, as compared to a 4 x 4 identity matrix was significantly different, L (10) = 108.23, $p < .001$ (with a critical value of 23.31). Similarly, the soil variable group too varied significantly from the hypothetical structure, L (6) = 50.88 $p < .001$ (with a critical value of 16.82). This indicated that both group variable matrices have correlations significantly different from zero, indicating that treatment of these variables as a group set may be appropriate.

2) Look at the eigenvalues produced from a principal components analysis (use *proc princomp*). What information can you deduce from it? If you were interested in accounting for at least 90% of the data how many components would you retain? (**1pt**)

Conducting a principal components analysis it appears that 78.65% of the data's variance can be accounted for with three components (equivalent here to having used a Kaiser criterion of eigenvalues greater than 1). In order to account for at least 90% we would have to have five components. This suggests that the variables (as a singular set) have common aspects which could be accounted for by a linear combination. The benefit of using canonical correlation here would be to investigate the specific covariation of variable groups between one another.

3) Test the overall null hypothesis that canonical correlations are equal to zero. Give your test statistic, *df*, and *p*-value. (**1pt**)

The omnibus test of at least one canonical correlation being not equal to zero is equivalent to testing whether the first canonical correlation is significantly different from zero. Therefore it follows that the omnibus test can be expressed as:

$$\Lambda = \prod_{i=1}^{k} (1 - R_k^2) ,$$

where k is the number of canonical correlations and R^2 is the squared multiple correlation. Therefore the omnibus test of this example is:

$$\Lambda = (1 - .5258) \times (1 - .2759) \times (1 - .0123)$$
$$\Lambda = .4742 \times .7241 \times .9877 = .3391 \approx .34$$

Wilk's Lambda can be expressed as a *F*-value from a corresponding *F* distribution with 12 and 235.76 degrees of freedom, resulting in F (12, 235.76) = 9.92, p <.001.

4) Test the null hypothesis that the first, second and third canonical correlations are equal to zero. Give your test statistic, d.f., and p-value. (**1pt**)

It follows from question three above that at least one (here the first) canonical correlation is significantly different from zero. The subsequent canonical correlations can be expressed in a similar way, summarized in Table 3.

Table 3. Tests of canonical correlations

R	R_a	R^2	Λ	*F*	df1	df2	*p*
.72	.71	.52	.34	9.92	12	235.76	< .001
.52	.51	.27	.72	5.47	6	180.00	< .001
.11	.07	.01	.99	.57	2	91.00	.57

$$\Lambda_2 = (1-.2759) \times (1-.0123)$$
$$\Lambda_2 = .7241 \times .9877 = .7152 \approx .72$$

and

$$\Lambda_3 = (1-.0123)$$
$$\Lambda_3 = . .9877 \approx .99$$

5) What are the correlations between each variable and their corresponding canonical variables? Discuss. (**1pt**)

The highest correlation for mercury content in the soil samples was for the variable and the first canonical variable (Soil 1). All three soil variables correlated at least moderately with the first canonical variable. Interestingly, the highest correlation for the sulfur content variable was a negative one with the third canonical variable. Much like when analyzing scale consistency it is somewhat peculiar that a correlation would be the highest when negative. At this point it ought to be remembered that multivariate normality was violated and that subsequent analyses may or may not be accurate representations of the true variable interrelations.

Table 4. Soil variable correlations with their canonical variables

Variable	Soil 1	Soil 2	Soil 3
Merc_soil	.92	-.40	-.07
Sulf_soil	.52	-.26	-.81
Phos_soil	.46	.89	.07

Table 5. Water variable correlations with their canonical variables

Variable	Water 1	Water 2	Water 3
Merc_water	-.28	.24	.90
Methyl_water	.39	-.35	-.31
Turbidity	.79	-.19	.02
Carbon_water	.84	.43	-.22

Correlations of the water group variables were not as consistent as those for the soil variables. Here the highest correlation was observed between the mercury content and the third canonical variable. Given that the mercury content in soil samples correlated highest with the first soil canonical variable it would have been feasible to expect a similar pattern in the water variable group. This was not the case. In fact an earlier exploration of the multivariate distribution of the data revealed that there are several multivariate outliers, which could very much account for the lack of consistency in the analysis.

6) What are the correlations between each variable and the canonical variables from the opposite group? Discuss. (**1pt**)

Tables 6 and 7 summarize the cross variable group and canonical variable correlations. The first two soil variables correlated highest with the first water canonical variable. This would suggest a possible major covariation between the pairing of those two variables (merc_soil and sulf_soil) and the turbidity and carbon_water variables (which also correlated highest with the first water and first soil canonical variables. It is plausible that the interrelationship between soil mercury and sulfur contents is somehow co-dependant on soluble carbon (which in return can be reasonably linked to turbidity).

Table 6. Soil variable correlations with the opposite canonical variables

Variable	Water 1	Water 2	Water 3
Merc_soil	.66	-.21	-.01
Sulf_soil	.38	-.14	-.09
Phos_soil	.33	.47	.01

Table 7. Water variable correlations with the opposite canonical variables

Variable	Soil 1	Soil 2	Soil 3
Merc_water	-.20	.13	.10
Methyl_water	.28	-.18	-.03
Turbidity	.58	-.10	.00
Carbon_water	.61	.23	-.02

7) Reproduce in full at least one canonical correlation. (**3pt**)

For this example the second canonical correlation will be reproduced. In order to be able to do so we first need to consider the formula:

$$\rho_2 = \frac{\text{cov}(Soil_2, Water_2)}{\sqrt{\text{var}(Soil_s) \times \text{var}(Water_2)}} \ .$$

In order to solve this proportion we need to first compute the necessary variance and covariance components for the two canonical variables involved. Starting with the second soil canonical variable we need to consult the corresponding raw canonical coefficients (Table 8) as well as the appropriate variance covariances for the involved variables. Even though only a single canonical correlation is being reproduced all the group specific variance and covariances are needed since the canonical variable is a linear composite of the original group variables.

Table 8. Variances and covariances of the observed variables

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Merc_soil	25.6371						
(2) Sulf_soil	107.5528	1117.6360					
(3) Phos_soil	42.9793	-245.6992	18750.6530				
(4) Merc_water	-.2381	-1.4261	.7182	.0376			
(5) Methyl_water	.5163	2.2719	-1.5386	-.0378	.0930		
(6) Turbidity	2.5920	9.7356	21.7117	-.0221	-.0247	.8176	
(7) Carbon_water	1.3861	5.3548	38.1358	-.0418	.0738	.2397	.3352

Table 9. Raw canonical coefficients for soil variables

Variable	Soil 1	Soil 2	Soil 3
Merc_soil	.1818	-.1050	.1496
Sulf_soil	-.0014	.0037	-.0388
Phos_soil	.0029	.0068	-.0003

The variance of the second canonical variable for the soil group can thus be expressed as:

$$\text{var}(Soil_2) = \sum_{s=1}^{3}\sum_{w=1}^{3} a_{2s} \times a_{2w} \times \text{cov}(X_s, X_w),$$

where *a* is the raw canonical coefficient and *X* is the original variable. Since for the first part we shall consider the variance of the second canonical variable there is no consideration of covariates between the different variable groups. Therefore, the above formula can be expressed as:

$$\text{var}(Soil_2) = \sum_{s=1}^{3}\sum_{s=1}^{3} a_{2s} \times a_{2s} \times \text{cov}(X_s, X_s).$$

Using the above formula we can now expand by the first summation sign and compute the sum of the corresponding elements:

$$\begin{aligned}
\text{var}(Soil_2) = \quad & (-.1050) \times (-.1050) \times 25.6371 + \\
& (-.1050) \times .0037 \times 107.5528 + \\
& (-.1050) \times .0068 \times 42.9793 + \\
& .0037 \times (-.1050) \times 107.5528 + \\
& .0037 \times .0037 \times 1117.6360 + \\
& .0037 \times .0068 \times (-245.6992) + \\
& .0068 \times (-.1050) \times 42.9793 + \\
& .0068 \times .0037 \times (-245.6992) + \\
& .0068 \times .0068 \times 18750.6530 \\
\text{var}(Soil_2) = \quad & .2826 - .0418 - \\
& .0307 - .0418 + \\
& .0153 - .0062 - \\
& .0307 - .0062 + \\
& .8670 \\
\text{var}(Soil_2) = \quad & 1.0075
\end{aligned}$$

Next we can express the variance of the second canonical variable for the water group in a similar way,

$$\text{var}(Water_2) = \sum_{s=1}^{4}\sum_{s=1}^{4} a_{2s} \times a_{2s} \times \text{cov}(Y_s, Y_s),$$

using the corresponding appropriate raw canonical coefficients and variance / covariances for the water variable group.

Table 10. Raw canonical coefficients for water variables

Variable	Water 1	Water 2	Water 3
Merc_water	1.1951	-.0643	6.5164
Methyl_water	1.3099	-3.0602	1.8777
Turbidity	.7122	-.9078	.3162
Carbon_water	.8118	2.0682	-.1996

$$\begin{aligned}
\text{var}(Water_2) = {} & (-.0643) \times (-.0643) \times .0376 + \\
& (-.0643) \times (-3.0602) \times (-.0378) + \\
& (-.0643) \times (-.9078) \times (-.0221) + \\
& (-.0643) \times 2.0682 \times (-.0418) + \\
& (-3.0602) \times (-.0643) \times (-.0378) + \\
& (-3.0602) \times (-3.0602) \times (.0930) + \\
& (-3.0602) \times (-.9078) \times (-.0247) + \\
& (-3.0602) \times 2.0682 \times .0738 + \\
& (-.9078) \times (-.0643) \times (-.0221) + \\
& (-.9078) \times (-3.0602) \times (-.0247) + \\
& (-.9078) \times (-.9078) \times .8176 + \\
& (-.9078) \times 2.0682 \times .2397 + \\
& 2.0682 \times (-.0643) \times (-.0418) + \\
& 2.0682 \times (-3.0602) \times .0738 + \\
& 2.0682 \times (-.9078) \times .2397 + \\
& 2.0682 \times 2.0682 \times .3352 \\
\text{var}(Water_2) = {} & .0002 - .0074 - .0013 + .0056 - \\
& .0074 + .8709 - .0686 - .4671 - \\
& .0013 - .0686 + .6738 - .4500 + \\
& .0056 - .4671 - .4500 + 1.4338
\end{aligned}$$

$$\text{var}(Water_2) = 1.0011$$

In order to get the covariance between the two canonical variables we need to revert back to the original formula expressed above. Since now we need to account for variable groups with different numbers of variables the summation subscripts need to reflect not a single but two groups.

$$\text{cov}(Soil_2, Water_2) = \sum_{s=1}^{3}\sum_{w=1}^{4} a_{2s} \times a_{2w} \times \text{cov}(X_s, Y_w)$$

The appropriate covariance is thus expressed as the summation of the weighted covariances across both of the variable groups.

$$\begin{aligned}
\text{cov}(Soil_2, Water_2) = \; & (-.1050) \times (-.0643) \times (-.2381) + \\
& (-.1050) \times (-3.0602) \times .5162 + \\
& (-.1050) \times (-.9078) \times 2.5920 + \\
& (-.1050) \times 2.0682 \times 1.3861 + \\
& .0037 \times (-.0643) \times (-1.4261) + \\
& .0037 \times (-3.0602) \times 2.2719 + \\
& .0037 \times (-.9078) \times 9.7356 + \\
& .0037 \times 2.0682 \times 5.3548 + \\
& .0068 \times (-.0643) \times .7182 + \\
& .0068 \times (-3.0602) \times (-1.5386) + \\
& .0068 \times (-.9078) \times 21.7117 + \\
& .0068 \times 2.0682 \times 38.1358 \\
\text{cov}(Soil_2, Water_2) = \; & -.0016 + .1659 + .2471 - .3010 + \\
& .0003 - .0257 - .0327 + .0410 - \\
& .0003 + .0320 - .1340 + .5363 \\
\text{cov}(Soil_2, Water_2) = \; & .5273
\end{aligned}$$

The canonical correlation of the second canonical variables (*Soil₂* and *Water₂*) can thus be computed as:

$$\rho_2 = \frac{\text{cov}(Soil_2, Water_2)}{\sqrt{\text{var}(Soil_s) \times \text{var}(Water_2)}}$$

$$\rho_2 = \frac{.5273}{\sqrt{1.0075 \times 1.0011}} = \frac{.5273}{\sqrt{1.0086}} = \frac{.5273}{1.0042} = .5250$$

It follows that the second canonical variables (the second canonical pairing) has a correlation of .5250 (the .0002 discrepancy being due to rounding error). This correlation is still significant, see questions 4, and contributes to the understanding of the covariation between the two variable sets.

8) What remarks about the canonical correlations can you make? What inferences can you draw based on your analysis? (**1pt**)

There are two significant canonical correlations cumulatively accounting for about 99.17% of the covariation between the two variable sets. Judging from the cross group variable and canonical variable correlations it would appear that the two sets of variables accounting for the two major sources of covariation between the two groups stem mainly from the interrelationship between merc_soil, turbidity, and carbon_water (driving the correlation for the first canonical pair) and merc_soil, phos_soil and carbon_water driving the second pair (correlation of the second canonical variables).

Based on these observations it would appear that merc_soil, turbidity and carbon_water are key in the interrelationship between these two groups. The variable interconnectivity appears to be accounted for in large by the concentration and measures of these three variables.