STAT505 Assessment #11

1. J&W Exercise 10.18. The data are found in "paper.dat" with columns $y_1 - y_4$ and $z_1 - z_4$ (described in Exercise 7.26).

   Working with the standardized variables $Z_1^{(1)} \dots Z_4^{(1)}$, $Z_1^{(2)} \dots Z_4^{(2)}$. The coefficients from the paper variables for determining the sample canonical variable pairs are

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$
$Z_1^{(1)}$	-1.5054	-3.4956	-5.7015	-5.0848
$Z_2^{(1)}$	-0.2119	-1.5431	3.5252	-0.5867
$Z_3^{(1)}$	1.9984	1.0760	-4.7135	6.0694
$Z_4^{(1)}$	0.6764	3.7679	7.1532	-0.6861

   Similarly, the coefficients from the pulp variables are

	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
$Z_1^{(2)}$	-0.1593	0.6886	-0.5130	2.3330
$Z_2^{(2)}$	0.6325	1.0029	0.0772	-2.1803
$Z_3^{(2)}$	0.3249	0.0050	-1.6631	0.0222
$Z_4^{(2)}$	0.8179	-1.5619	-0.7786	0.0891

   So, for example, the first sample canonical pair is found as follows:

$$\hat{U}_1 = -1.505Z_1^{(1)} - .212Z_2^{(1)} + 1.998Z_3^{(1)} + .676Z_4^{(1)}$$
$$\hat{V}_1 = -.159Z_1^{(2)} + .633Z_2^{(2)} + .325Z_3^{(2)} + .818Z_4^{(2)}$$

   The others are similarly computed. The canonical correlations between pairs are

$$\hat{\rho}_1^* = .9173 \quad \hat{\rho}_2^* = .8169 \quad \hat{\rho}_3^* = .2654 \quad \hat{\rho}_4^* = .0917$$

   The initial test for any nonzero canonical correlation is highly significant (p-value less than .0001); this is also the case when testing the last three correlations. However, the subsequent tests are insignificant. From this, we can conclude that the first two canonical pairs are significantly correlated ($\hat{U}_1$ with $\hat{V}_1$ and $\hat{U}_2$ with $\hat{V}_2$), while the last two pairs are not significantly correlated.

   The first canonical paper variable $\hat{U}_1$ appears to be primarily the difference between (standardized) stress at failure and breaking length, while $\hat{U}_2$ appears to be primarily the difference between burst strength and breaking length. The first two canonical pulp variables appear to be the sum and difference, respectively, of long fiber fraction and zero span tensile.

2. For a sample of 50 sales representatives, measures of sale performance and scores on tests of reasoning and intellectual are recorded. In this analysis we look at how two "sales" variables 1) sales growth and 2) sales profitability relate to three "test" variables 1) mechanical reasoning, 2) abstract reasoning, and 3) mathematics.

   (a) What is the value of the first canonical correlation?

   The first canonical correlation is $\hat{\rho}_1^* = .9876$.

   (b) Write the equations that give the first (unstandardized) pair of canonical variables. You can round the coefficients to three decimal places.

$$\hat{U}_1 = 0.051 * \text{Growth} + 0.063 * \text{Profit}$$
$$\hat{V}_1 = 0.026 * \text{Creative} + 0.072 * \text{Mechanical} + 0.074 * \text{Math}$$

   (c) Suppose that a sales person has data values of sales growth = 93, profitability = 96, mechanical reasoning = 12, abstract reasoning = 10 and math ability = 20. Compute the canonical variable values for the first pair of unstandardized canonical variables. Don't worry about centering.

$$\hat{u}_1 = 0.051(93) + 0.063(96) = 10.791$$
$$\hat{v}_1 = 0.026(10) + 0.072(12) + 0.074(20) = 2.604$$

   (d) What is the result of the test that all canonical correlations equal 0? What is the p-value and conclusion?

   With a Wilk's Lambda test stat of .02369 and p-value less than .0001, we have significant evidence at least one of the canonical correlations (the first one) is nonzero.

   (e) What is the result of the test that the second canonical correlation equals 0? What is the p-value and conclusion?

   The result of the Wilk's lambda test of the second canonical correlation is 0.9637. The p-value is 0.4273, which is insufficient evidence to conclude that the second canonical variate pair is correlated.

   (f) What are the correlations between the sales variables (the x-variables) and the first canonical variable that is constructed from them?

   The correlation between the first canonical variable $\hat{U}_1$ and growth is .9702; its correlation with profit is .9899.

   (g) What are the correlations between the test variables (the y-variables) and the first canonical variable that is constructed from them?

   The correlation between the first canonical variable $\hat{V}_1$ and creative is .5696; its correlation with mechanical and math are .7542 and .9666, respectively.

   (h) Why are there only two canonical correlations in this situation?

   There are two here since the smaller set of variables has only two in it.

   (i) Write a brief conclusion about the relationship between the two sets of variables.

   The first canonical correlation for sales is highly correlated with both Growth (0.97) and Profit (0.9899); it represents the health of the sales. The first canonical correlation for Test is highly correlated with all three tests, especially with Math (0.9666). Overall, these two sets of variables (sales and test) are highly correlated with each other as evidenced by the very strong correlation of the first canonical pair.

SAS code:

```
data paper;
infile 'v:\paper.dat' delimiter='09'x;
input y1 y2 y3 y4 z1 z2 z3 z4;
run;
proc cancorr out=pap vprefix=paper wprefix=pulp;
var y1-y4;
with z1-z4;
run;

data sales;
  infile "v:\sales.dat";
  input growth profit new creative mechanical abs math;
  run;
proc cancorr out = salescan vprefix=sales wprefix=tests;
var growth profit;
with creative mechanical math;
run;
```