Stat501 Solutions and Comments on Assignment 4
Spring 2005

1. (a) $\hat{\lambda}_4 = 4 - (2.855 + 0.721 + 0.253) = 0.171$

(b) Since you are analyzing a correlation matrix, you must use standardized values of individual traits in the formula for the first principal component, i.e.,

$0.55\,Z_{1j} + 0.54\,Z_{2j} + 0.49\,Z_{3j} + 0.42\,Z_{4j}$ where $Z_{ij} = (X_{ij} - \bar{X}_i) / s_i$ is the standardized value of the i-th trait for the j-th tree.

(c) (2.855/4) x 100% = 71.375 %

(d) The component reflects the overall sizes of trees. It is large when a tree has relatively long branches and a relatively large trunk.

(e) The second component is associated with the shape of the tree canopy. It is large for tall trees that have relatively few or short lateral branches, and small for shorter trees with relatively long or many lateral branches.

(f) Since you are analyzing a correlation matrix, the sample correlation between the score for the second principal component and the standardized length of laterals is

$$\hat{e}_{23}\sqrt{\hat{\lambda}_4} = (-.56)\sqrt{0.721)} = -0.4755$$

(g) This is zero from the definition of principal components.

(h) The third and fourth principal components account for only 10.6% of the variation in the four measurements made for the apple trees in this study. These are linear combinations of variables that are nearly constant. The fourth component, for example, indicates that the weight of the tree can be approximately obtained as a linear function of the length of the laterals and the length of the central leader (height of the tree). The third component indicates that the square of the trunk circumference is approximately a linear function of the other three traits. Using these two approximate equations simultaneously leads to a formula for the weight of the tree as a linear function of the square trunk circumference and the length of the central leader , which are relatively easy traits to measure. This may have substantial practical important because a tree must be dug up before it can be weighed. It is much easier to estimate its weight from its height and truck circumference at a specific distance above the ground. You could obtain a formula from the information given in this problem, but notice that you must use standardized values of the traits with this information because we computed the eigenvalues and eigenvectors of the sample correlation matrix. Alternatively, you could get a formula by least squares regression of X1 on X2 and X4 if you had the data.

2. (a).

$$y1 = 0.353\left(\frac{x_1 - u_1}{\sqrt{s_{11}}}\right) + 0.404\left(\frac{x_2 - u_2}{\sqrt{s_{22}}}\right) + 0.396\left(\frac{x_3 - u_3}{\sqrt{s_{33}}}\right) + 0.406\left(\frac{x_4 - u_4}{\sqrt{s_{44}}}\right) +$$
$$0.397\left(\frac{x_5 - u_5}{\sqrt{s_{55}}}\right) + 0.402\left(\frac{x_6 - u_6}{\sqrt{s_{66}}}\right) + 0.265\left(\frac{x_7 - u_7}{\sqrt{s_{77}}}\right)$$

$$y2 = -0.395\left(\frac{x_1 - u_1}{\sqrt{s_{11}}}\right) - 0.270\left(\frac{x_2 - u_2}{\sqrt{s_{22}}}\right) - 0.266\left(\frac{x_3 - u_3}{\sqrt{s_{33}}}\right) - 0.030\left(\frac{x_4 - u_4}{\sqrt{s_{44}}}\right) +$$
$$0.177\left(\frac{x_5 - u_5}{\sqrt{s_{55}}}\right) + 0.216\left(\frac{x_6 - u_6}{\sqrt{s_{66}}}\right) + 0.788\left(\frac{x_7 - u_7}{\sqrt{s_{77}}}\right)$$

The first principal component is the overall measurement of the percentages of mature forest within the seven rings. A large positive score for the first component corresponds to a site with a relatively high percentage of mature forest, and a extreme negative score for the first component corresponds to a site with a low percentage of mature forest.

The second principal component represents a gradient in the percentage of mature forest as one moves from the center to the outer rings. A large positive score for this component corresponds to sites with relatively high percentage of mature forest near the center and relatively low percentages of mature forest in the outer rings.

(b). In the correlation matrix, all correlations are positive, but correlations between the outermost ring and the inner rings is weaker then correlations among the inner rings. This corresponds to the positive loadings of the percentage of mature forest in every ring on the first component. The second component corresponds to the tendency of correlations between percentages of mature forest at a site to become weaker as the distance between the rings increases

(c). 81.01%

(d). By Kaiser's rule or from scree plot, the first two principal components appear to be a good choice. They account for about 80% of the total sample variance of the standardized variables.

(e) and (f) The grand tour and the projection pursuit options in the XGOBI package do not reveal much that cannot be seen by rotating the scores for the first three principal components with the XGOBI package. The first three component scores for the randomly selected sites seem to be more widely scattered, whereas scores for the first three principal components for many of the nesting sites appear to be clustered together at higher positive values of the first component, indicating that percentages of mature forest tend to be higher in rings centered at nesting sites.

(g). Test the null hypothesis: Ho: $\underline{u}_1 = \underline{u}_2$ against the alternative Ha: $\underline{u}_1 \neq \underline{u}_2$.

Hotelling's T^2 test:

$$T^2 = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)^T \left[\left(\frac{2}{n} \right) S_{pooled} \right]^{-1} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2) \sim \frac{(2n-2)p}{(2n-p-1)} F_{p,2n-p-1}$$

$$F = \frac{2n-8}{7(2n-2)} T^2 = 4.81 \sim F_{7,52} \text{ where n=30 and p=7}$$

F=4.81 with p-value=0.0003

Note that Hotelling's T^2 is a monotone function of the likelihood ratio test. You could use a large sample chi-square approximation to –2log(ratio of likelihoods) to test this null hypothesis, but the large sample chi-square approximation would provide a test with inflated type I error levels in small samples.

(h). From the SAS output:

	Univariate t=test	**Wilcoxcon test**
Variables	Pr>|t|	Pr>|Z|
Km91	0.001	0.0015
Km119	0.023	0.0170
Km140	0.0001	0.0001
Km160	0.0039	0.0062
Km177	0.0003	0.0002
Km241	0.0005	0.0006
Km338	0.4964	0.6222

From both the t-tests and the Wilcoxcon tests, we can see that there are significant differences between mean percentage of old growth forest between nesting and random sites within each ring size except the outer most ring.

3. (a) The proportion of the total standardized variance accounted for by the first three principal components is 83%.

(b) The first component is a wealth component. Large positive values correspond to census tracks with relatively high median family income and relatively high percentages of high school graduates and housing units occupied by owners and relatively low percentages of young people not in school, people with Spanish surnames, and housing units with more than one person per room.
The second component is an age and employment component. High values correspond to census tracks with relatively high unemployment rates and relatively high percentages of the population under age 16.

Large positive values of the third component correspond to census tracts where there are relatively high percentages of children, owner occupied homes, and high housing density, but low unemployment. These may be census tracts with high proportions of families with young children living in smaller houses.

(c) It is better to analyze the correlation matrix because median income (X_2) is measured on a different scale than the other variables which are all percentages. The choice of units for (X_2) is arbitrary, but this choice greatly affects the principal components that are produced. If (X_2) is reported in dollars, its variance will be much larger than the variance of the other variables and the first principal component will be dominated by X_2. Conversely, if X_2 is reported in units of thousands of dollars its variance will be closer to that of the other variables and X_2 will not dominate the first principal component.

(d) There is no apparent reason to use more than three principal components. The fourth and fifth components account for only 8% and 5% of the variation in the standardized traits, respectively. The fourth component essentially reflects only the proportion of young people not in school, and the fifth component corresponds to the percentage of owner occupied homes.

(e) Displaying the data with GGOBI does not reveal any distinct groups or clusters of points. There appears to be only one continuous data cloud. Maximizing the central mass index indicates some possible outliers or extreme points. These should be investigated to make sure that the data are valid. (In this case, they are valid data points that should not be deleted.)

4. a)

Source of variation	Df	SS	MS	F	p-value	Conservative df
Training groups	1	18906.25	18906.2	2.56	0.1238	
Subjects(groups)	22	162420.08	7382.73			
Tasks	1	25760.25	25760.2	12.99	0.0016	1,22
Groups*Task	1	3061.78	3061.78	1.54	0.2271	1,22
Errors(b)	22	43622.31	1982.83			
Cues	2	5697.04	2848.52	22.60	<0.0001	1,22
Group*Cues	2	292.63	146.31	1.16	0.3226	1,22
Error(c)	44	5545.00	126.02			
Tasks*Cues	2	345.38	172.69	2.66	0.0815	1,22
Groups*Task*Cues	2	90.51	45.26	0.70	0.5039	1,22
Error(d)	44	2860.78	65.02			
Corrected total	143	268602				

Many students did not report the correct F-tests. By default, the PROC GLM output in SAS incorrectly divides each mean square by the error mean square. Proper F-tests are computed using the TEST statement. These F-tests are printed later in the output.
You can determine appropriate denominator mean squares by looking at expectations of mean squares (formulas for expectations of means squares are obtained from the Q option in the random statement. The notation for the models used in this problem is

$$X_{ijkl} = \mu + \gamma_k + \eta_{kl} + \alpha_i + \alpha\gamma_{ik} + \delta_{ikl} + \beta_j + \beta\gamma_{jk} + \psi_{jkl} + \alpha\beta_{ij} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

where

γ_k is a training type effect $(\gamma_2 = 0)$

α_i is a task effect $(\alpha_2 = 0)$

β_j is a cue condition effect $(\beta_3 = 0)$

$\eta_{kl} \sim \text{NID}(0, \sigma_\eta^2)$ $\quad$ $\psi_{jkl} \sim \text{NID}(0, \sigma_\psi^2)$

$\delta_{ikl} \sim \text{NID}(0, \sigma_\delta^2)$ $\quad$ $\varepsilon_{ijkl} \sim \text{NID}(0, \sigma_\varepsilon^2)$

and the random terms are all mutually independent. Then

$$E(MS_{groups}) = \sigma_\varepsilon^2 + 2\sigma_\psi^2 + 3\sigma_\delta^2 + 6\sigma_\eta^2 + (\text{quadratic form involving fixed effects})$$

and an F-test for testing the null hypothesis that the quadratic form is zero is obtained by dividing MS_{groups} by a mean square with expectation $\sigma_\varepsilon^2 + 2\sigma_\psi^2 + 3\sigma_\delta^2 + 6\sigma_\eta^2$. In this case $E(MS_{\text{subjects within groups}}) = \sigma_\varepsilon^2 + 2\sigma_\psi^2 + 3\sigma_\delta^2 + 6\sigma_\eta^2$ and the F-test is computed as $F = MS_{groups} / MS_{\text{subjects within groups}} = 2.56$ on (1,22) df.

b). There are different levels or correlation for repeated measurements on the same subject depending on whether the repeated measures correspond to the same task or the same cue.

$$\begin{aligned}
\text{Cov}(X_{ijkl}, X_{i'j'kl}) &= \text{Cov}(\eta_{kl} + \delta_{ikl} + \varphi_{jkl} + \varepsilon_{ijkl}, \eta_{kl} + \delta_{i'kl} + \varphi_{j'kl} + \varepsilon_{i'j'kl}) \\
&= \text{Cov}(\eta_{kl}, \eta_{kl}) + \text{Cov}(\delta_{ikl}, \delta_{i'kl}) + \text{Cov}(\varphi_{jkl}, \varphi_{j'kl}) + \text{Cov}(\varepsilon_{ijkl}, \varepsilon_{i'j'kl}) \\
&= \sigma_\eta^2 + \sigma_\delta^2 + \sigma_\varphi^2 + \sigma_\varepsilon^2 = a \qquad \text{if } i=i' \text{ and } j=j' \\
\text{or} &= \sigma_\eta^2 + \sigma_\delta^2 = b \qquad \text{if } i=i' \text{ and } j\neq j' \\
\text{or} &= \sigma_\eta^2 + \sigma_\varphi^2 = c \qquad \text{if } i\neq i \text{ and } j=j' \\
\text{or} &= \sigma_\eta^2 = d \qquad \text{if } i\neq i' \text{ and } j\neq j'
\end{aligned}$$

$$\text{Var}\begin{pmatrix} x_{11kl} \\ x_{12k1} \\ x_{13kl} \\ x_{21kl} \\ x_{22kl} \\ x_{23kl} \end{pmatrix} = \begin{bmatrix} a & b & b & c & d & d \\ b & a & b & d & c & d \\ b & b & a & d & d & c \\ c & d & d & a & b & b \\ d & c & d & b & a & b \\ d & d & c & b & b & a \end{bmatrix}$$

c). The three factor interaction between treatment groups, types of tasks, and cue conditions are not significant. There are no significant two factor interactions. Tasks and cue conditions have significant effects on the mean completion times. Mean completion time was longer for task 1.and mean completion times were longer for the third cue an rather similar for the first two cues. Mean completion times were not significantly different for the two training groups, although the observed completion times tended to be shorter for the enhanced training group. Note that the ifference between mean response for the two training groups is a between subject response and is estimated less accurately then the effects of the within subject factors and their interactions.

5.

$X=A\beta+\varepsilon$ where $\beta=\begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{12} & \mu_{21} & \mu_{22} & \mu_{23} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{21} & \gamma_{22} & \gamma_{23} \end{bmatrix}$

a. $C=\begin{bmatrix}0 & 1\end{bmatrix}_{1x2}$, $M=I_{6x6}$

Wilks Criterion $\hat{\Lambda}=0.627$, F=1.68, df = (6, 17), p-value=0.1859

There is no apparent difference between training groups with respect to the mean completion time of either task under any cue condition.

b. $C=I_{2x2}$, $M=\begin{bmatrix}1 & 1 & 1 & -1 & -1 & -1\end{bmatrix}^T$

Wilks Criterion $\hat{\Lambda}=0.602$, F=7.27, df = (2,22), p-value=0.0038

There is sufficient evidence to indicate that the mean completion times are not the same for the two tasks, averaging across cue conditions for each type of training program.

c. $C=I_{2x2}$, $M=\begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 \end{bmatrix}^T$

Wilks Criterion $\hat{\Lambda}=0.311$, F=8.33, df = (4,42), p-value<0.0001

The mean completion times are significantly different for the three cue conditions, averaging across tasks, for at least one type of training.

d. $C=\begin{bmatrix}0 & 1\end{bmatrix}_{1x2}$, $M=\begin{bmatrix}1 & 1 & 1 & -1 & -1 & -1\end{bmatrix}^T$

Wilks Criterion $\hat{\Lambda}$ =0.934, F=1.54, df = (1, 22), p-value=0.2271

The data do not show that there is interaction between type of training and type of tasks, averaging across cue conditions.

e. $C=\begin{bmatrix}0 & 1\end{bmatrix}_{1x2}$, $M=\begin{bmatrix}1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1\end{bmatrix}^T$

Wilks Criterion: $\hat{\Lambda}$ =0.8737, F=1.517, df = (2,21), p-value=0.2423

The evidence is not significant to reject null hypothesis, the data do not show that there is interaction between types of training and cue conditions, averaging across tasks.

f. $C=\begin{bmatrix}0 & 1\end{bmatrix}_{1x2}$, $M=\begin{bmatrix}1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1\end{bmatrix}^T$

Wilks Criterion $\hat{\Lambda}$ =0.936, F=0.71, df = (2,21), p-value=0.5029

The there factor interaction is not significant.

g. $C=\begin{bmatrix}2 & 1\end{bmatrix}_{1x2}$, $M=\begin{bmatrix}1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1\end{bmatrix}^T$

Wilks Criterion $\hat{\Lambda}$ =0.6787, F=4.97, df = (2,21), p-value=0.0171

There is enough evidence to show that there is significant interaction between type of task and cue condition averaging across training groups.