Multivariate Analysis -Homework No. 5
Due Thursday, February 27, 2003

The data set contains data on a number of variables collected on 50 applicants to a police department of a major metropolitan area. The variables in the first set include.

x_1: applicant's weight in kilograms

x_2: applicant's thigh skinfold thickness in millimeters

x_3: applicant's resting pulse rate

x_4: applicant's diastolic blood pressure

x_5: applicant's total body fat measurement

The variables in the second set include

y_1: the number of chin-ups the applicant was able to complete

y_2: applicant's maximum treadmill speed

y_3: applicant's treadmill endurance time in minutes

1. Compute the correlation matrix and identify $\mathbf{R}_{11}$, $\mathbf{R}_{12}$, $\mathbf{R}_{22}$. Comment briefly on what you see.

2. Carry out a canonical correlation analysis on the standardized data. Obtain and report the values of the canonical correlation coefficients. Obtain the canonical variates. Present these in a table (nothing fancy, I just want to be able to see clearly which variables get which coefficient.

3. Carry out a test of
$$H_o : \mathbf{\Sigma_{12}} = \mathbf{0}$$
vs
$$H_a : \mathbf{\Sigma_{12}} \neq \mathbf{0}.$$
If you use SAS give the results from Wilk's Lambda. If you use R or Splus use the approximate χ^2 test we discussed in class. Give an approximate$p$-value and discuss briefly what the results of the test implies.

4. Attempt an interpretation of the coefficients. This can be difficult but focus on the signs and size of the coefficients, and the corresponding signs of the standardized values of the variables. The canonical correlations are all positive (i.e. high values of U_1 go with high values of V_1 and low values of U_1 with low values o f V_1). In attempting an interpretation focus on what happens as you consider above average values of the variables (positive) and below average values of the variables (negative) along with the corresponding canonical coefficients. That is, focus on the question of determining what combinations of x's and y's are needed to give simultaneously high (and low) values of U_1 and V_1. Let me know if this needs further clarification.

5. Obtain the scores and produce scatterplots of the first two pairs U_1 vs V_1 and U_2 vs V_2. Comment briefly on any unusual patterns you see (outliers, etc.).

Lisa Schwarz
2/28/2003

Stat 437. Homework 5.

1.

	x1	x2	x3	x4	x5	y1	y2	y3
x1	1	0.554	-0.264	-0.052	0.810	-0.576	-0.053	-0.368
x2	0.554	1	-0.006	0.049	0.844	-0.670	-0.208	-0.336
x3	-0.264	-0.006	1	0.234	-0.095	0.155	-0.300	0.007
x4	-0.052	0.049	0.234	1	0.045	0.054	-0.321	0.134
x5	0.810	0.844	-0.095	0.045	1	-0.691	-0.206	-0.405
y1	-0.576	-0.670	0.155	0.054	-0.691	1	0.324	0.213
y2	-0.053	-0.208	-0.300	-0.321	-0.206	0.324	1	-0.022
y3	-0.368	-0.336	0.007	0.134	-0.405	0.213	-0.022	1

The correlation between applicant weight and total body fat is relatively strong (0.810). The correlation between applicant thigh skinfold thickness and total body fat is also relatively high (0.844). The higher the body fat, the higher the skinfold thickness and weight. The correlations between number of chin-ups and the three variables weight, skinfold thickness, and body fat are relatively strong (-0.53, -0.670, and -0.691, respectively). The higher the weight, body fat, and skinfold thickness, the fewer number of chin-ups an applicant was able to complete. There do not seem to be any strong correlations between the three Y variables.

2.

$cor	cor(U1,V1)	cor(U2,V2)	cor(U3,V3)
[1]	0.776524	0.460297	0.186864

$xcoef

	[,1]	[,2]	[,3]	[,4]	[,5]
x1	-0.035	-0.050	0.041	0.143	-0.229
x2	-0.066	-0.030	-0.148	-0.107	-0.217
x3	0.015	0.087	0.103	-0.027	-0.070
x4	0.028	0.075	-0.095	0.081	-0.003
x5	-0.050	0.110	0.111	-0.015	0.386

$ycoef

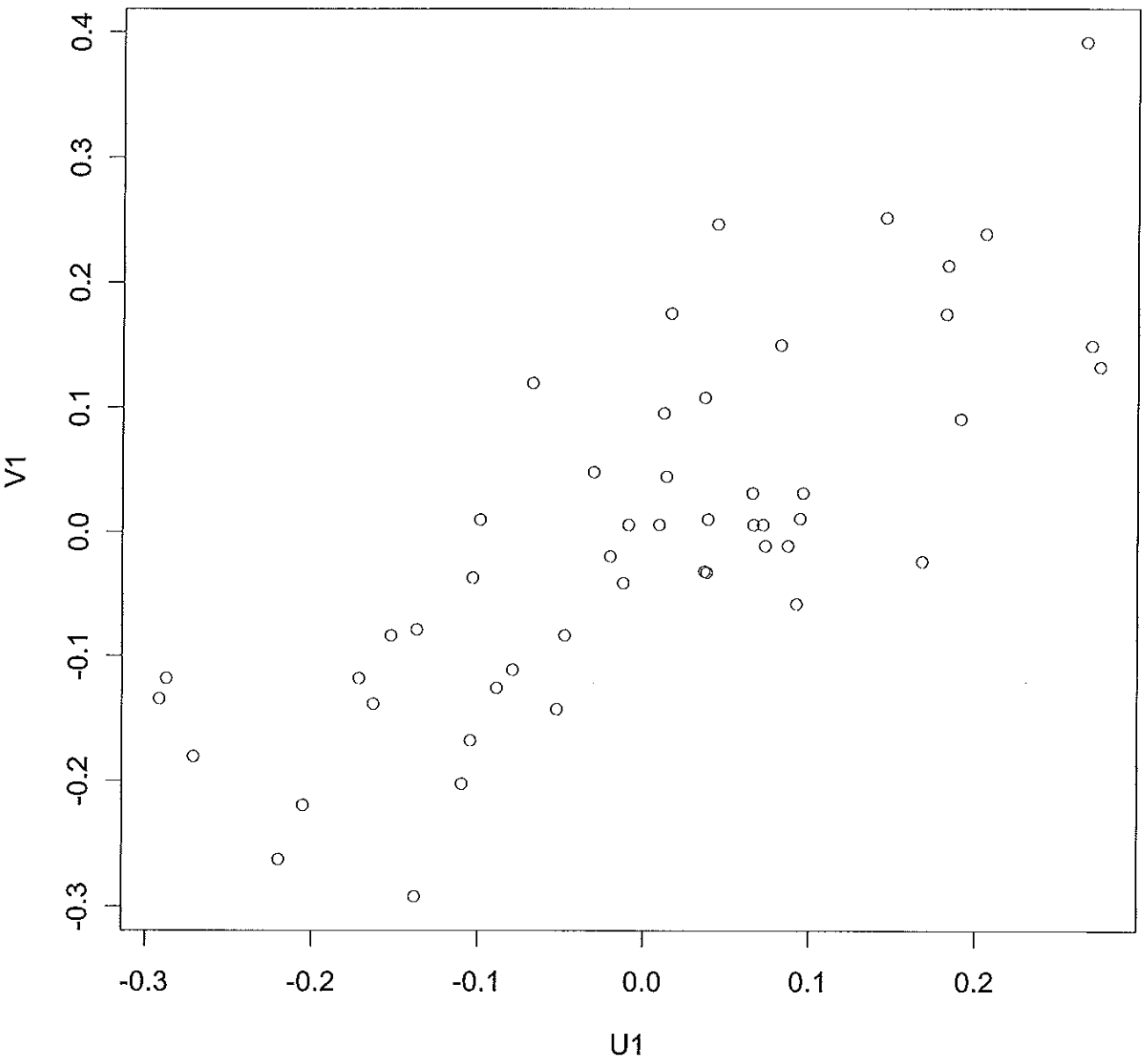
	[,1]	[,2]	[,3]
y1	0.130	0.031	0.079
y2	-0.025	-0.150	0.000
y3	0.048	-0.023	-0.137

$xcenter	x1	x2	x3	x4	x5
	-2.00E-16	1.44E-16	1.34E-16	6.75E-16	4.31E-16

$ycenter	y1	y2	y3
	-7.33E-17	-1.30E-15	9.02E-16

3.
Wilk's lambda indicates we should only consider the first set of canonical variates.

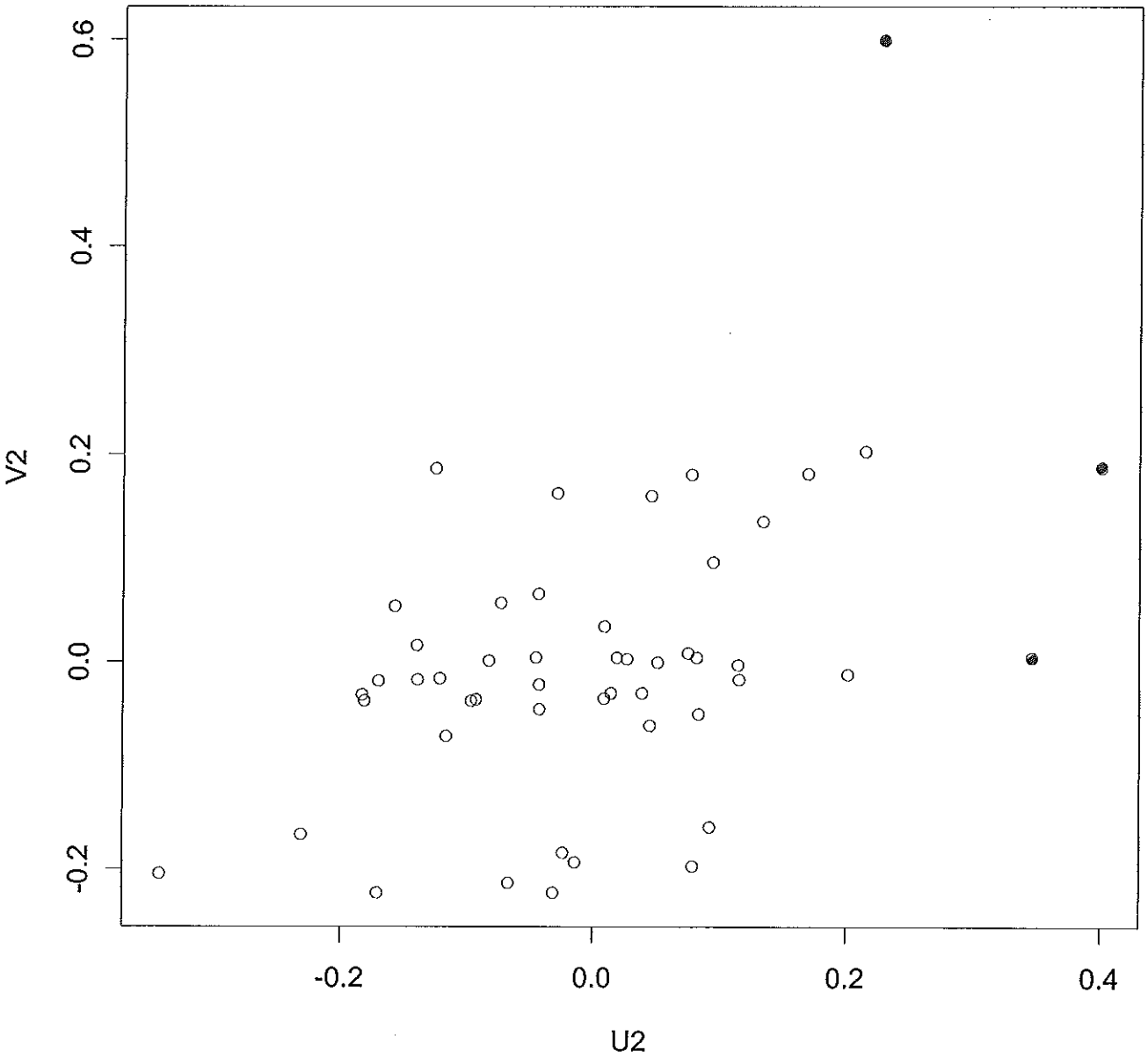
4.
For the first two canonical variables, U_1 is a measure of the difference between two groups. Group 1 is weight, skinfold thickness, and total body fat, while group 2 is pulse rate and blood pressure. V_1 is also a measure of the difference between two groups. Group 1 is number of chin-ups and treadmill endurance time. Group 2 is maximum treadmill speed. Applicants with a large difference between groups 1 and 2 within U_1 also had a large difference between groups 1 and 2 within V_1. In U_1, the coefficient for skinfold thickness is 4.5 times larger than the coefficient for resting pulse rate. The second set of canonical variables is harder to interpret but may show a higher body fat associated with a lower tread mill speed. The third set is just mush.



No obvious outliers. Obviously correlated.

Note: Matlab flips these values on the x and y axes.

Scatter plot of second 2 canonical variables - police data



● = potential outliers. Correlation difficult to see when outliers removed.

18/20