Final Project – Multivariate Statistical Methods

Chi Nguyen, Huong Nguyen, Phuong Dao

05/18/2022

Part I: MANOVA techniques

1. About the Dataset

The heart disease data set is collected from four databases: Cleveland, Hungary, Switzerland, and the VA Long Beach data, publicly available at https://archive.ics.uci.edu/ml/datasets/heart+disease. The dataset dates from 1988 and contains 76 attributes. All the fields are related to heart conditions and can be used to determine whether a patient has heart disease. However, because of the project's scope, we decided not to keep the original data and instead transformed it to meet the project's scope and objectives. The data used in this project is compiled from three data sets: Cleveland, Hungary, and Switzerland, and include six variables:
- Sex
- dataset: Data source
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- oldpeak: ST depression induced by exercise relative to rest
- thalch: maximum heart rate achieved

2. Data Summary

Items	**Values**
Number of rows	342
Number of columns	7
Column Type Frequency	
Categorical	2
Numeric	5
Variable Information	
Level of data source variable	3
Level of sex variable	2
Number of rows in each dataset	114
Number of rows for each sex	171
Number of female patients in each data source	57
Number of male patients in each data source	57

3. Multivariate hypothesis testing

Objective

The objective of multivariate hypothesis testing in this section is to see if there is a significant difference between males and females in heart health indicators (regardless of locations).

Applying technique

Assumption: Two samples of males and females are independent and $\Sigma_1 = \Sigma_2 = \Sigma$

Test hypothesis:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

where μ_1 is the mean vector of Male and μ_2 is the mean vector of Female.

According to **Figure 1** in appendix, we have:

$$\bar{x_1} = (131.05\ 169.62\ 1.00\ 150.36\ )$$

$$\bar{x_2} = (133.34\ 171.69\ 0.68\ 147.29\ )$$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = 13.26$$

Meanwhile, we know that p = 4 (trestbps, chol, oldpeak, thalch), n_1 (number of Male observations) = 171, n_2 (number of Female observations) = 171, $T^2_{.05}(4,340) < T^2_{.05}(4,200) = 9.817\ < 13.26$.

Therefore, we reject H_0 and conclude that the heart health indicators of males and females are significantly different.

4. Multivariate analysis of variance

Objective

The aim of multivariate analysis of variance in this section is to examine if there is a significant difference in heart health indicators between three data sources: Cleveland, Hungary and Switzerland.

Applying technique

Assumptions: 1. The samples are selected independently from 3 populations; 2. The samples are (approximately) Normal; 3. The 3 population variances are equal

Hypothesis test:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: At least one inequality in 3 data sources
(where μ_1 is the mean vector of Cleveland, μ_2 is the mean vector of Hungary and μ_3 is the mean vector of Switzerland)

We have k = 3, p = 4, n = 114. Therefore, $v_H = 3 - 1 = 2$, $v_E = 3(114 - 1) = 339$, $s = min(p, v_H) = 2$, $m = 0.5(|2 - 4| - 1) = 0.5$, $0.5(339 - 4 - 1) = 167$

After running MANOVA test, the result for each test can be found in **Appendix - Figure 2**. In details:

- For Wilks test:
  - $\Lambda = 0.17$.
  - Meanwhile, $\Lambda_{0.05}(4,2,339) > \Lambda_{0.05}(4,2,320) = 0.952 > \Lambda = 0.17$. Thus, we reject H_0.
- For Pillai's trace test:
  - $V^{(s)} = 0.84$.
  - Meanwhile, $V_{0.05}^{(s)}(2,0.5,167) < V_{0.05}^{(s)}(2,0,25) = 0.218 < 0.84$. Thus, we reject H_0.
- For Hotelling-Lawley Trace test:
  - $U^{(s)} = 4.67$. Thus, $\frac{v_E}{v_H} \times U^{(s)} = \frac{339}{2} \times 4.67 = 791.57$
  - Meanwhile, $U_{0.05}^{(s)}(2,4,2 + 339 - 4) = U_{0.05}^{(s)}(2,4,337) < U_{0.05}^{(s)}(2,4,200) = 3.974 < 791.57$. Thus, we reject H_0.
- For Roy's Greatest Root test:
  - $\theta = \frac{4.66}{1+4.66} = 0.82$.
  - Meanwhile,$\theta_{0.05}(2,0.5,167) < \theta_{0.05}(2,0,120) = 0.043 < 0.82$. Thus, we reject H_0.

All four MANOVA tests reject H_0, which means the heart health conditions are significantly different between the three data sources.

Powerful test:

- We have $\frac{\lambda_1}{\Sigma\lambda_i} = 99.65\%$. Therefore, λ_1 is dominant, and μ_1, μ_2, μ_3 lie in one dimension. In that case, $\theta \geq U^{(s)} \geq \Lambda \geq V^{(s)}$, or in other words, Wilk's Test is the most powerful test. Details can be found in **Appendix – Figure 3**.

Part II: Discriminant Analysis and Classification

1. About the Dataset

Dataset used is the same data in part 1.

2. Discriminant Analysis

Objective

In the discriminant analysis for several groups, we are concerned with finding linear combinations of variables that best separate the three groups of multivariate observations (Cleveland, Hungary, and Switzerland).

Applying technique

The eigenvalues of $E^{-1}H$ are $\lambda_1 = 4.6573$ and $\lambda_2 = 0.0161$. (**Appendix – Figure 4**)

According to **Appendix – Figure 5**, the corresponding eigenvectors, which are also the vectors of the discriminant function coefficients are:

$a_1 = (-0.0007, 0.0175, 0.0377, 0.1081\ )$

$a_2 = (0.0477, -0.0007, 0.2091, -0.0032\ )$

The first eigenvalue accounts for a substantial proportion of the total $\lambda_1/(\lambda_1 + \lambda_2) = 0.99$.

Thus, the mean vectors lie largely in one dimension, and one discriminant function suffices to describe most of the separation among the three groups.

The standardized discriminant function coefficients are:

$$a_1^* = (-0.0145\ 0.9969\ 0.0397\ 0.1081\ )$$

$$a_2^* = (0.9319\ \ -0.0386\ 0.2200\ \ -0.0744\ )$$

$\lambda_1/(\lambda_1 + \lambda_2) = 0.99$, we concentrate on a_1^*. From a_1^* (**Appendix – Figure 6**), the second and fourth variables contribute most to separate the groups.

To test the significance of two discriminant functions, we use the test statistics:

$\Lambda_1 = \left(\frac{1}{1+4.6573}\right)\left(\frac{1}{1+0.0161}\right) = 0.17,$

$\Lambda_2 = \left(\frac{1}{1+0.0161}\right) = 0.984$

As $k = 3, p = 4, N - k = 342 - 3 = 339$. The critical value for Λ_1 is $\Lambda_{0.05,4,2,339} > \Lambda_{0.05,4,2,320} = 0.952 >\ \Lambda_1$ and for Λ_2 is $\Lambda_{0.05,3,1,338} < \Lambda_{0.05,3,1,440} = 0.982 <\ \Lambda_2$.

We reject H_o for Λ_1, while fail to reject H_o for Λ_2.

Therefore, the first discriminant function is significant, but the second discriminant function is not. The two procedures agree as to the number of important discriminant function.

Also, from the plot **(Appendix - Figure 10)**, the discriminant function separates groups 1 and 2 from group 3, but the second is ineffective in separating group 1 from group 2.

3. Classification Analysis

Objective

In this section, we examine the allocation of observations to groups, which is predict aspect of discriminant analysis.

Applying technique

Based on **Appendix – Figure 7**, the linear classification functions are:

$$L_1(y) = -53.29 + 0.34y_1 + 0.074y_2 - 0.083y_3 + 0.293y_4$$

$$L_2(y) = -57.59 + 0.35y_1 + 0.083y_2 - 0.0003y_3 + 0.294y_4$$

$$L_3(y) = -42.54 + 0.352y_1 - 0.0007y_2 - 0.208y_3 + 0.272y_4$$

We note that y_1 and y_4 have essentially the same coefficients in all three function and hence do not contribute to classification of y.

To evaluate the performance of a classification procedure to predict group membership, we use the probability of misclassifications, known as the error rate. The proportion of misclassifications resulting from re-substitution is called the apparent error rate.

As we have a large sample, we prefer to use the holdout method for estimating the error rate. The classification table is shown in Appendix – **Figure 8a and 8b**.

The error rate is equal $(39 + 1 + 56)/342 = 0.28$

Besides, using the k-nearest neighbor method of estimating the error rate with k = 5, we have the classification table as in Appendix – **Figure 9a and 9b**.

For each point $y_{ij}, i = 1,2,; j = 1,2, \ldots ,114$, we find the five nearest neighbors classify the point accordingly.

Part III: Principal Component Analysis

1. About the Dataset

The state crime dataset has information on the crime rates and totals across the United States for a wide range of years from 1960 to 2019. The data is available at https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/downloads/download-printable-files. Because the original dataset is large, we took random sample of only 350 observations.

2. Data summary

Items	**Values**
Number of rows	350
Number of columns	19
Column Type Frequency	
Categorical	0
Numeric	19

There are no categorical variables, and 19 variables measure the total numbers of crimes, including burglaries, larcenies, murder, rapes, etc., and number of reported offenses per 100,000.

3. Principal Component Analysis

Objective

Principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. In the sense that p components produce the total system variability, often much of this variability can be accounted for by a small number of k of principal components. Our goal is to seek to find PCs such that as much information in the k components as there is in the original p variables. Once we find such optimal PCs, we can replace the initial p variables consisting of n observations by k principal components.

Applying technique

- The measurements of the variables on the dataset are measured on scales with widely different ranges such as variable X_1 is the population of the state, variable X_{11} to variable X_{19} records the total number of crimes committed. Also, the units of the measurement are not commensurate, for example, variable X_2 to variable X_{10} measure the rate of crime per 100000 population which its unit is very different from the rest of the variables. Thus, we decided to derive the PCs from the correlation matrix. (We could not insert the picture of correlation and covariance due to the large results. Therefore, we attached a pdf file of PCA, which includes correlation and covariance results).

- From the outputs, we can see that if using the **Covariance matrix (Appendix – Figure 11a)**, the first principal component will explain 99.98% of total sample covariance. That is not surprising because the variance of the population variable completely dominates the first principal component determined from the covariance matrix. However, the percent of variance explained by using **Correlation matrix (Appendix – Figure 11b)** is 0.5164, 0.2376, 0.1036, 0.0536, 0.0379, 0.0168, 0.0135, 0.0077, 0.0058, 0.0035, 0.0012, 0.0011, 0.0006, 0.0003, 0.0002, 0.0001, 0.0001. We can see that the first three principal components, collectively explain 85.77% of the total sample variance. In addition, looking at the scree plot **(Appendix – Figure 12)** we also see that three principal components should be the optimal numbers of PCs to retain. Consequently, sample variation is summarized very well by three principal components and a reduction in the data from 350 observations on 19 variables to 350 observations on three principal components.

- From **Appendix – Figure 13**, given the foregoing component coefficients, the first principal component appears to be essentially a similarly weighted sum of the entire total number of crimes including murder, assaults, rape, robbery, burglary, larceny, and lightly weighted by the sum of the rate of crime per 100,000 population.

- However, the second principal component is determined most heavily by the sum of rates violent in all, then by rate violent assaults, rate crime murder, rate violent robbery, rate property all, rate property larceny in order, then rate violent rape, rate property burglary, rate violent motor, and lightly weighted by the difference of the total number of crimes variables and population variables which are essentially contributed to determining the principal 1.

- The third principal component is weighted most heavily by the sum of rate violent motor, rate of property burglary, then the difference with rate property all, rate property larceny, rate violent rape, then the sum of total property motor, rate violent assault. It is lightly weighted by the difference in the total amount of violence except the total crime in property motor.

Part IV: Appendix

1. MANOVA techniques

Figure 1: Multivariate hypothesis testing

Spl 4 rows 4 cols (numeric)

386.47045	204.43829	4.2288579	-4.936687
204.43829	17986.279	20.143118	370.83934
4.2288579	20.143118	1.1011293	-1.383646
-4.936687	370.83934	-1.383646	548.81823

T2 1 row 1 col (numeric)

13.255756

X1BAR 4 rows 1 col (numeric)

131.04678
169.61988
1.0011696
150.35673

X2BAR 4 rows 1 col (numeric)

133.33918
171.69006
0.6789474
147.28655

Figure 2: . Multivariate analysis of variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall dataset Effect
H = Type III SSCP Matrix for dataset
E = Error SSCP Matrix

S=2 M=0.5 N=167

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.17395943	117.40	8	672	<.0001
Pillai's Trace	0.83910320	60.90	8	674	<.0001
Hotelling-Lawley Trace	4.67337661	195.93	8	477.68	<.0001
Roy's Greatest Root	4.65725335	392.37	4	337	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Figure 3: Eigenvalues

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for dataset E = Error SSCP Matrix					
		Characteristic Vector V'EV=1			
Characteristic Root	Percent	trestbps	chol	oldpeak	thalch
4.65725335	99.65	-0.00004031	0.00095369	0.00204844	0.00025258
0.01612325	0.35	0.00259167	-0.00003688	0.01136113	-0.00017374
0.00000000	0.00	0.00022677	-0.00007298	0.00394680	0.00232205
0.00000000	0.00	-0.00111353	-0.00004991	0.05120756	0.00000000

2. Discriminant Analysis and Classification

Figure 4: The eigenvalues of $E^{-1}H$

					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.907323	0.906414	0.009572	0.823236	4.6573	4.6411	0.9965	0.9965	0.17395943	117.40	8	672	<.0001
2	0.125966	0.103316	0.053294	0.015867	0.0161		0.0035	1.0000	0.98413258	1.81	3	337	0.1449

Figure 5: Raw Canonical Coefficients

Raw Canonical Coefficients		
Variable	Can1	Can2
trestbps	-.0007421791	0.0477176892
chol	0.0175593474	-.0006791016
oldpeak	0.0377157324	0.2091805439
thalch	0.0046504090	-.0031988202

Figure 6: Standardized Canonical Coefficients

Pooled Within-Class Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
trestbps	-.0144941614	0.9318880893
chol	0.9968322074	-.0385521342
oldpeak	0.0396663583	0.2199991848
thalch	0.1081487713	-.0743909776

Figure 7: Linear Discriminant Functions

Linear Discriminant Function for dataset			
Variable	**1**	**2**	**3**
Constant	-53.29165	-57.59368	-42.54049
trestbps	0.34047	0.35482	0.35219
chol	0.07487	0.08316	-0.0006949
oldpeak	-0.08317	-0.0003974	-0.20779
thalch	0.29285	0.29412	0.27230

Figure 8a & 8b: Classification table and error rate

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	114	74	39	1
2	114	56	58	0
3	114	0	0	114

Number of Observations and Percent Classified into dataset				
From dataset	1	2	3	Total
1	74 64.91	39 34.21	1 0.88	114 100.00
2	56 49.12	58 50.88	0 0.00	114 100.00
3	0 0.00	0 0.00	114 100.00	114 100.00
Total	130 38.01	97 28.36	115 33.63	342 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for dataset				
	1	2	3	Total
Rate	0.3509	0.4912	0.0000	0.2807
Priors	0.3333	0.3333	0.3333	

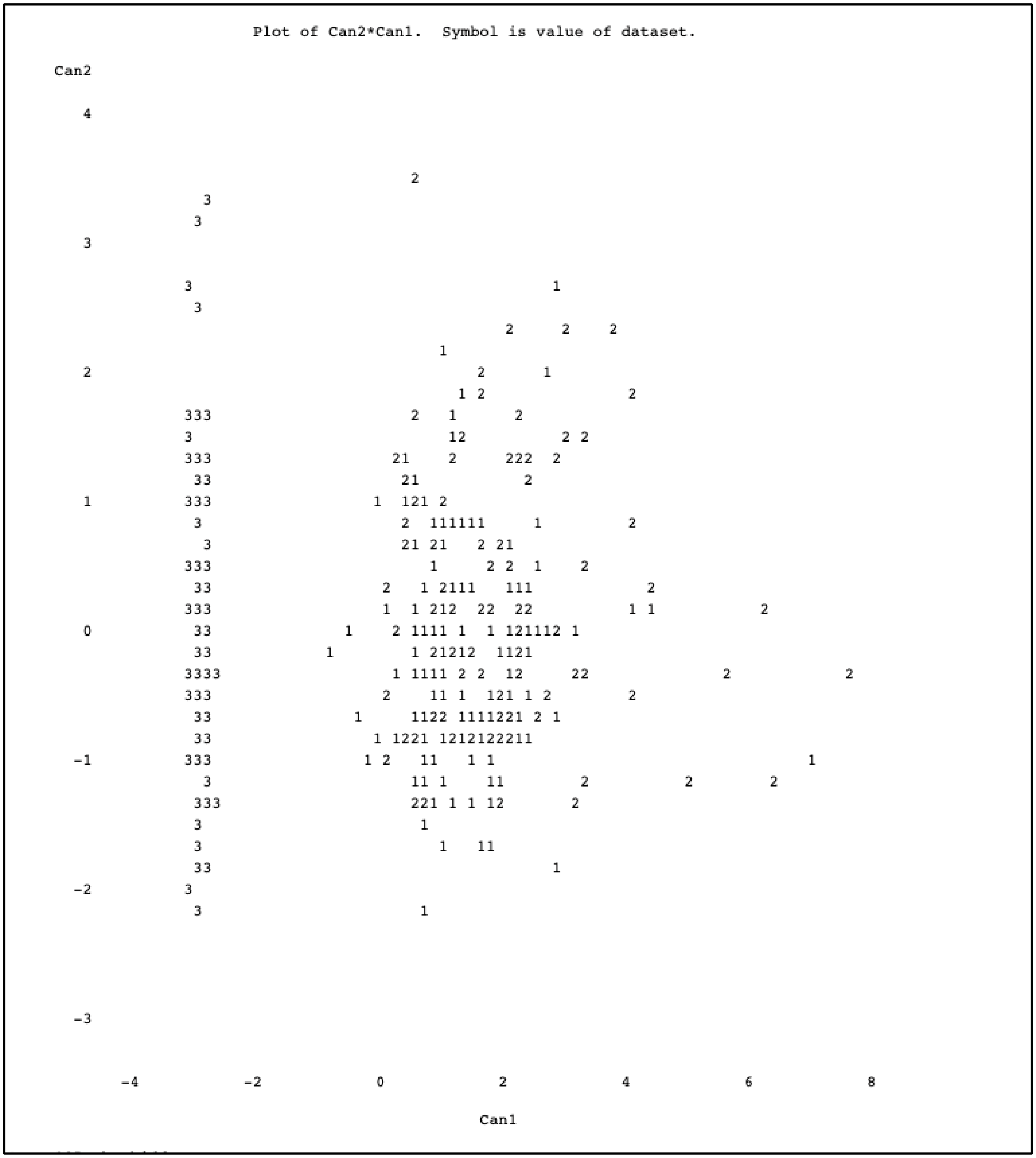
Figure 9a & 9b: Classification table and error rate using the k-nearest neighbor method with k = 5

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	114	54	60	0
2	114	50	64	0
3	114	0	0	114

Number of Observations and Percent Classified into dataset				
From dataset	1	2	3	Total
1	54 47.37	60 52.63	0 0.00	114 100.00
2	50 43.86	64 56.14	0 0.00	114 100.00
3	0 0.00	0 0.00	114 100.00	114 100.00
Total	104 30.41	124 36.26	114 33.33	342 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for dataset				
	1	2	3	Total
Rate	0.5263	0.4386	0.0000	0.3216
Priors	0.3333	0.3333	0.3333	

Figure 10: Plot of 2 Discriminant Functions



3. Principal Component Analysis

Figure 11a & 11b: Eigenvalues of Covariance and Correlation Matrix

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.14042E15	1.14022E15	0.9998	0.9998
2	1.9973E11	1.88772E11	0.0002	1.0000
3	1.09585E10	9777150702	0.0000	1.0000
4	1181372145	845897904	0.0000	1.0000
5	335474240	231800233	0.0000	1.0000
6	103674008	83887075.2	0.0000	1.0000
7	19786932.6	15561495.5	0.0000	1.0000
8	4225437.09	1754016.32	0.0000	1.0000
9	2471420.78	2020970.95	0.0000	1.0000
10	450449.82	180755.444	0.0000	1.0000
11	269694.376	223112.082	0.0000	1.0000
12	46582.2941	10030.1953	0.0000	1.0000
13	36552.0989	14519.474	0.0000	1.0000
14	22032.6249	10871.6798	0.0000	1.0000
15	11160.945	8616.22527	0.0000	1.0000
16	2544.71975	2353.96032	0.0000	1.0000
17	190.759436	190.759436	0.0000	1.0000
18	0	0	0.0000	1.0000
19	0		0.0000	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.81236807	5.29734640	0.5164	0.5164
2	4.51502167	2.54623674	0.2376	0.7541
3	1.96878493	0.95044130	0.1036	0.8577
4	1.01834363	0.29828268	0.0536	0.9113
5	0.72006095	0.40078675	0.0379	0.9492
6	0.31927421	0.06275706	0.0168	0.9660
7	0.25651715	0.11098271	0.0135	0.9795
8	0.14553443	0.03555759	0.0077	0.9872
9	0.10997684	0.04281158	0.0058	0.9929
10	0.06716526	0.04406513	0.0035	0.9965
11	0.02310013	0.00257193	0.0012	0.9977
12	0.02052820	0.00963832	0.0011	0.9988
13	0.01088988	0.00515807	0.0006	0.9993
14	0.00573181	0.00197567	0.0003	0.9996
15	0.00375614	0.00228810	0.0002	0.9998
16	0.00146805	0.00014086	0.0001	0.9999
17	0.00132719	0.00119067	0.0001	1.0000
18	0.00013652	0.00012158	0.0000	1.0000
19	0.00001494		0.0000	1.0000

Figure 12: Scree Plot

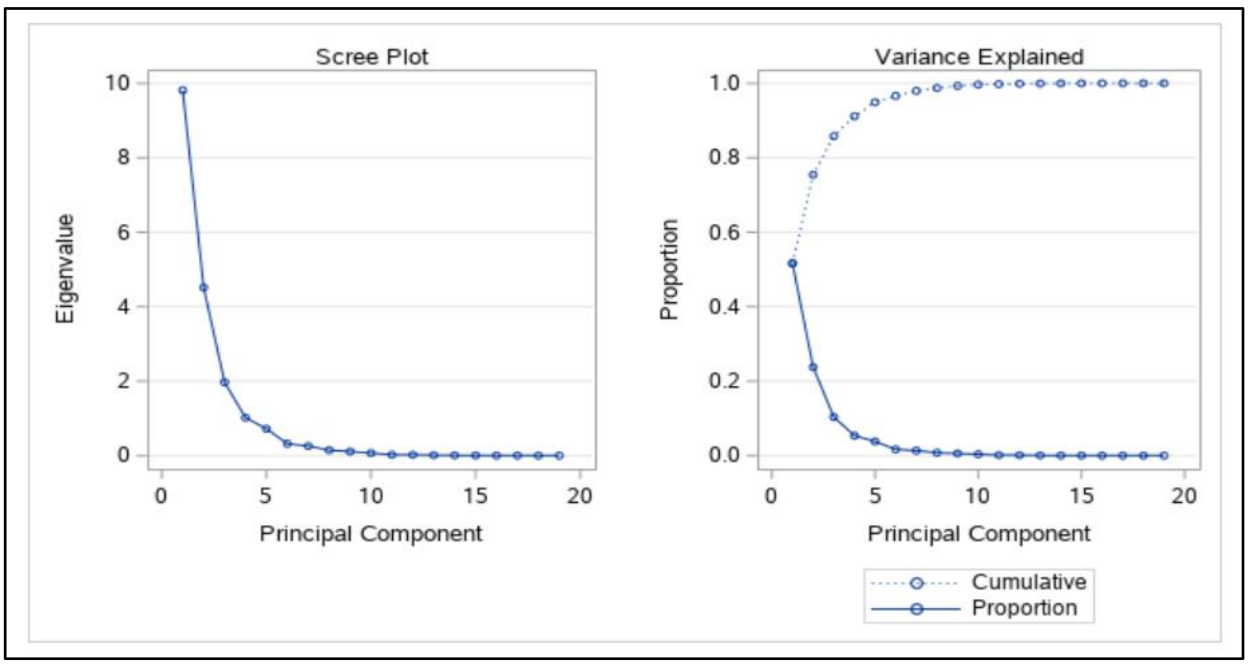


Figure 13: Eigenvectors

Eigenvectors																			
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18	Prin19
x1	0.306334	-.081695	-.019617	-.012993	-.038300	0.071031	-.008956	0.453048	-.044259	0.368398	0.374867	-.113485	0.383332	0.428486	0.059306	-.042756	0.228561	0.088626	-.004699
x2	0.053936	0.363925	-.287115	0.342575	0.332121	0.042733	-.042812	0.025376	0.088892	-.044827	0.107137	0.675005	0.168784	0.060602	-.194506	0.020998	-.037545	-.033894	-.015341
x3	0.030222	0.185361	0.549582	0.255499	0.402492	0.183940	0.197618	0.082896	-.566555	-.093053	-.095532	-.095672	0.021261	-.005646	0.072824	-.005672	0.042552	0.008103	0.005622
x4	0.052210	0.347879	-.283156	0.451032	0.238457	0.010418	-.006251	0.004664	0.323118	0.142991	-.105605	-.601439	-.151782	-.004366	0.116673	-.008356	0.008574	0.031302	0.009506
x5	0.028592	0.100156	0.662601	-.005054	0.099400	-.269853	-.115507	0.180800	0.624260	-.079148	0.104476	0.096052	-.031992	0.003762	0.045673	0.000377	-.044855	-.009710	0.004283
x6	0.086631	0.427748	0.073953	-.184748	-.180468	-.281781	0.026144	0.017867	-.147259	0.065457	-.031764	-.194881	-.028614	0.059659	-.764718	-.000536	0.007638	0.006806	-.052155
x7	0.080980	0.397375	-.106699	-.085211	-.274689	-.441159	0.531413	0.093004	-.100313	-.079579	0.053933	0.118256	-.028938	-.046259	0.465210	-.011982	0.042712	-.009298	0.031995
x8	0.038936	0.334348	0.000411	-.552783	0.127653	0.631475	0.291882	-.053287	0.267079	-.018233	-.021387	-.021373	0.033734	0.010127	0.023852	0.059147	0.011666	0.003554	0.000990
x9	0.042578	0.262816	0.214313	0.385720	-.711932	0.404745	-.177817	-.140489	0.008748	-.008801	0.020716	0.071028	0.041387	0.020790	0.067254	-.006220	-.005928	-.000551	0.003541
x10	0.074737	0.386596	-.028670	-.339626	0.108721	-.104627	-.717194	-.042887	-.235148	0.088431	0.010932	-.006258	0.012549	-.097935	0.335508	-.056828	-.021090	0.003144	0.023349
x11	0.314858	-.047589	-.048141	0.020866	0.018529	0.028916	-.027700	-.050195	0.021597	-.358619	0.233095	-.034208	-.024100	-.357510	-.088469	-.127455	0.301876	0.635441	0.243768
x12	0.307038	-.051498	0.082064	0.023068	0.081690	-.018013	0.100928	-.428362	-.066587	0.292291	0.583576	-.020101	-.219384	-.106437	0.007444	0.247931	-.370907	-.048295	-.058394
x13	0.315258	-.053490	-.044558	0.021187	0.005907	0.032588	-.018240	0.020996	0.028806	-.263153	0.157303	-.121634	0.175840	-.374462	-.049458	-.181566	0.267612	-.694261	-.153819
x14	0.299545	-.058828	0.132475	-.000299	0.057765	-.130969	0.087792	-.534837	0.098591	0.414301	-.424715	0.128423	0.236403	0.011936	0.039101	-.155579	0.330932	0.063575	-.038121
x15	0.316894	-.049040	-.017390	-.006494	-.015659	-.025472	-.010774	0.014975	0.005923	-.114128	-.221300	-.046342	0.184879	0.104447	-.025852	0.170359	-.317832	-.199134	0.784832
x16	0.315563	-.050843	-.037386	-.001442	-.030535	-.016869	0.010919	0.099158	0.018211	-.190529	-.242130	-.086187	0.400442	-.103303	0.038427	-.031591	-.569744	0.233711	-.479426
x17	0.316030	-.049507	-.021775	-.027212	0.008315	0.079247	0.010009	0.052412	-.019285	-.055559	-.061374	0.111572	-.511605	0.320110	-.001088	-.684410	-.179736	-.050990	-.007465
x18	0.307299	-.067091	-.012082	0.012824	-.068373	0.101537	-.014834	0.467579	-.019577	0.356225	-.295936	0.197387	-.386188	-.401057	-.034374	0.308519	0.062673	0.003293	-.053007
x19	0.314620	-.033185	-.037542	-.016861	0.013853	-.043809	-.108808	-.131139	-.005188	-.416904	-.113630	-.005494	-.233364	0.482112	0.059933	0.512066	0.261274	-.025652	-.242234

4. Code

Part 1

```
PROC IMPORT DATAFILE='/home/u61416483/Project/data_official_balanced.csv'
    DBMS=CSV
    OUT=WORK.IMPORT
    REPLACE;
    GETNAMES=YES;
PROC IML;
  USE WORK.IMPORT;
  READ ALL VAR {trestbps chol oldpeak thalch} INTO X;
  X1 = X[1:171,];
  X2 = X[172:342,];
  RESET PRINT;
  N1 = NROW(X1);
  N2 = NROW(X2);
  X1BAR = 1/N1*X1`*J(N1,1);
  X2BAR = 1/N2*X2`*J(N2,1);
  S1 = 1/(N1-1)*X1`*(I(N1)-1/N1*J(N1))*X1;
  S2 = 1/(N2-1)*X2`*(I(N2)-1/N2*J(N2))*X2;
  Spl = 1/(N1+N2-2)*((N1-1)*S1+(N2-1)*S2);
  T2 = N1*N2/(N1+N2)*(X1BAR-X2BAR)`*INV(Spl)*(X1BAR-X2BAR);
RUN;
```

```
TITLE 'MANOVA';
DATA HEART;
PROC IMPORT DATAFILE='/home/u61416483/Project/data_official_balanced.csv'
    DBMS=CSV
    OUT=HEART
    REPLACE;
    GETNAMES=YES;
PROC GLM;
  CLASS dataset;
  MODEL trestbps chol oldpeak thalch = dataset;
  MANOVA H=dataset/PRINTE PRINTH;
RUN;
```

Part 2

```
/* FINAL PROJECT */

DATA UCIHeartDisease;
  INFILE 'UCIsubsample3.dat';
  INPUT dataset trestbps chol oldpeak thalch;
RUN;
PROC FORMAT;
  VALUE dataset 1 = 'Cleveland' 2 = 'Hungary' 3 = 'Switzerland';
RUN;
TITLE 'FinalProject';
PROC CANDISC OUT=CAND;
  CLASS dataset;
RUN;
PROC PRINT DATA=CAND;
RUN;
TITLE 'FinalProject2';
PROC PLOT DATA=CAND;
  PLOT CAN2*CAN1=dataset;
RUN;
```

```
/* FINAL PROJECT - Classification Analysis */

DATA UCIHeartDisease;
  INFILE 'UCIsubsample3.dat';
  INPUT dataset trestbps chol oldpeak thalch;
RUN;
PROC discrim data = UCIHeartDisease oustat=ftstat
method = NORMAL pool = yes list crossvalidate;
class dataset;
var trestbps chol oldpeak thalch;
proc discrim data = UCIHeartDisease oustat = ftstat
method = NORMAL pool = NO list crossvalidate;
class dataset;
var trestbps chol oldpeak thalch;
proc discrim data = UCIHeartDisease oustat = ftstat
method = npar k=5 pool = yes list crossvalidate;
class dataset;
var trestbps chol oldpeak thalch;
```

Part 3

```
PROC IMPORT DATAFILE='/home/u61416483/Project/state_crime_sample.csv'
    DBMS=CSV
    OUT=WORK.IMPORT
    REPLACE;

proc princomp cov;
  var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19;

proc princomp out = crime_component;
  var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19;
run;
```