Multivariate Analysis for Discrimination of Carcinogenesis Stages

STAT764 Multivariate Analysis

Qianqian Yao

Contents

1. Introduction

1.1 Chemometrics and Raman Spectroscopy

Raman micro spectroscopy probes intrinsic molecular vibrations. Thus, the Raman spectrum depends on the chemical composition of the specimen and can therefore be considered as its "molecular fingerprint". As the chemical composition of healthy and tumor tissue differs, Raman micro spectroscopy allows for a discrimination between these states. The aim of the experiment design described in [1] is to define a detection limit for the discrimination of the different stages during carcinogenesis by minimizing the "biological variance" via applying a mouse model.

1.2 Experiment

Two groups of genotypic identical mice with and without tumor suppressor gene p53 are selected. The malignant transformation was stimulated via a regular application of a carcinogen, azoxymethane (AOM). The mice were treated with intraperitoneally once a week for 6 weeks with the carcinogen azoxymethane (AOM). Mini endoscopy was performed on anesthetized mice. For histological and Raman-spectroscopic investigations mice were sacrificed at different time points after the AOM application. The specific loss of gene p53 in the intestine markedly enhances the carcinogen-induced tumor incidence and leads to the development of invasive colorectal tumors beginning about 12 weeks after the first AOM treatment. In total, full preparation of colon and rectum was carried out for 47 individuals; biopsies were taken from 8 individuals.

1.3 Preparation and Biopsy

Biopsies underwent mechanical stress due to the application of forceps while the whole intestine had to reside in the sacrificed animal until preparation, which might influence the constitution of the tissue with respect to blood flow, which stopped which sacrificing the animal. As a consequence of these observations the Raman spectra of the biopsies were discarded for training the classification models only served as test dataset in the evaluation step.

1.4 Raman Spectra

Raman spectra of the 20 µm thick cryosections from colon and rectum of two groups of mice with and without tumor suppressor p53 were recorded with an upright micro-Raman setup. The cells were excited with a 785 nm diode laser. Raman images for the characterization/differentiation of the tumor development in the epithelium were recorded in the scanning mode with a step size of 5 µm and an integration time of 2s per spectrum. As a wavenumber standard 50 Raman spectra of 4-acetamidophenol were measured using the Zeiss 50x objective with an integration time of 1 s per spectrum. A Raman scan will refer to a grid measured pointwise with Raman micro spectroscopy. Thus, a scan consists of a certain number of spectra.

1.5 Labeling Raman Spectra

In a semi-automated process the Raman spectra of epithelial tissue were selected for every scan based on the blinded annotation of a trained pathologist. In this context blinded means that the pathologist made his/her assignments only for the HE stained images and never got in touch with any Raman spectral data to avoid a bias. The annotation of a scan refers to the assignment of predefined entities to the spectra of this scan. The assignment is unambiguously done for every spectrum of the scan. Seven classes were defined for the annotation – four classes describe the states of the adenoma-carcinoma sequence, i.e. normal epithelium, hyperplasia, adenoma and carcinoma. One class, morphology, covers all other tissue types except the epithelium.

1.6 Classification Models

An objective determination of the best model was obtained by a grid-search of the tested models and their respective parameters. These included random forest (RF) classifiers with varying numbers of trees, support vector machines (SVM) with different kernels (linear and radial-base) and cost values, a linear discriminant classifier (LDA) as well as a weighted k-nearest neighbors classifier (KKNN) with varying numbers of neighbors included in the analysis.

1.7 Discussion

[1] discussed the results of statistical analysis and reached some conclusions. A joint use of colon and rectum Raman spectra for the further analysis and model building is allowed. The difference between the two populations can be neglected in the

further modeling building steps and thus, the spectra from mouse individual with and without p53 can be used together in a single dataset.

The results suggest that the difference between normal tissue and hyperplasia are very small as compared to the differences between normal and adenoma tissue. Furthermore, the alterations occurring in the progression from adenoma to carcinoma seem to be – from a Raman spectroscopic point of view – very small as well. This corresponds to the morphological continuum that exists between the adenoma and carcinoma tissue of this mouse model of colon carcinogenesis in which the carcinoma is histologically almost indistinguishable from the adenoma except the fact of an invasive growth pattern. Hence, the most dramatic biochemical changes seem to happen when the polyps develop. This finding is consistent with our expectation and the knowledge about the adenoma-carcinoma sequence.

All results clearly show that hyperplasia is not safely detectable by Raman spectroscopy and thus must be defined as the detection limit. A clinically relevant distinction between normal tissue and the stages, which require further treatment, i.e. adenoma and carcinoma, however, can be obtained with a promising accuracy.

1.8 Crafted data for the project

Experimental unit: molecule of tissue samples

Variables: wavenumber (400, 500, 600,700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600)

Random or not: random

Groups: there are four groups grouping on carcinogenesis staging, which are Adenom, Hyperplasia, Karzinom, Normal.

Group size is 30, total group size is 120.

1.9 Paraphrase of the background of the data

The experiment design is to get the Raman spectra of the tissue samples which are obtained at different carcinogenesis stages of mice. The tissue samples are from different individuals and from different parts of the mice, but the inter-variance and intro-variance can be ignored, and all the Raman spectra can be jointed as one dataset.

The labels of the tissue about which carcinogenesis stages is labeled by professional pathologists for the HE stained image, which is a blinded process meaning that the pathologist made his/her decision only for the HE stained images and never got in touch with any Raman spectral data to avoid a bias.

Each scan on the tissue sample can generate multiple spectra. The annotation of the spectra refers to the labels of the scanned tissue samples. The assignment is unambiguously done for every spectrum of the scan.

In the downloaded dataset there are many columns, and each corresponds to a wavenumber from 400 to 1700, such as 401, 402, and increase by 1. In order to simplify the dataset, only the wavenumber at hundreds are selected, total 13 variables (400, 500, 600,700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600). 30 molecular spectra are selected for each group, thus, total 120 molecular spectra.

By the conclusion of the paper that described the experiment, it is showed that classification models couldn't distinguish the group of Hyperplasia well from the normal. Intuitively understanding is that Hyperplaisa is the initial stage of tumor so the chemical composition didn't change much than the normal stage. The classification models includes random forest (RF) classifiers with varying numbers of trees, support vector machines (SVM) with different kernels (linear and radial-base) and cost values, a linear discriminant classifier (LDA) as well as a weighted k-nearest neighbors classifier (KKNN) with varying numbers of neighbors included in the analysis.

Thus, the motivation of this project is to apply multivariate analysis to explore the group variance, to find the natural groupings of the variables to see how wavenumber correlated with each other.

1.10 Research Questions

The models in the paper couldn't detect hyperplasia from carcinogenesis staging of adenoma-carcinoma sequence, although they can distinguish normal and nonnormal. Thus, by this project we want to use methodologies in multivariate analysis to explore more on the group variances and to find some patterns of variables.

Research Question1 in Section3:

How each group differs from each other by MANOVA?

Research Question2 in Section4:

Classification by discriminant function

Research Question3 in Section5:

How do Principal Components perform on MANOVA analysis?

Research Question4 in Section6:

Clustering Analysis: we already know the group of all observations in dataset, if we ignore the labels first and do cluster analysis, will the observations from one group be clustered into same cluster?

Research Question5 in Section7:

Factor Analysis: are there any patterns among the variables? How the wavenumber relate to each other?

NOTE: some sentences about the background including experiment design and data analysis (1.1~1.7) are picked directly from the paper without paraphrasing.

2. Data Description and Visualization

2.1 Data Summary

```
str(chemometric)
```

```
## 'data.frame':    120 obs. of  14 variables:
##  $ Group: chr  "Adenom" "Adenom" "Adenom" "Adenom" ...
##  $ X400 : num  30.4 19.4 26.9 35.5 27.4 ...
##  $ X500 : num  38.4 32.9 41.3 34.8 32.1 ...
##  $ X600 : num  23.4 14.9 18.3 18.7 17.4 ...
##  $ X700 : num  10.2 11.6 15.4 15.9 13.9 ...
##  $ X800 : num  28.7 14.3 23.2 26.9 20.7 ...
##  $ X900 : num  42.8 37.8 68 57.8 46.5 ...
##  $ X1000: num  111.9 81.8 146.9 135.4 105.8 ...
##  $ X1100: num  78 62.6 98.1 91.9 69.9 ...
##  $ X1200: num  26.8 21.9 27.4 31.3 25 ...
##  $ X1300: num  123 95 150 139 107 ...
##  $ X1400: num  38.2 26.2 38.3 36.6 29.6 ...
##  $ X1500: num  2.7 2.23 2.08 3.26 2.41 ...
##  $ X1600: num  30.7 24.9 33.2 35.1 28.9 ...
```

```
summary(chemometric)
```

```
##     Group                X400             X500             X600        
##  Length:120         Min.   : 12.79   Min.   : -3.317   Min.   : 3.562  
##  Class :character   1st Qu.: 28.01   1st Qu.: 32.769   1st Qu.:14.745  
##  Mode  :character   Median : 33.44   Median : 43.368   Median :18.799  
##                     Mean   : 40.40   Mean   : 46.614   Mean   :22.105  
##                     3rd Qu.: 49.17   3rd Qu.: 55.177   3rd Qu.:25.809  
##                     Max.   :117.14   Max.   :128.407   Max.   :74.922  
##       X700              X800             X900            X1000       
##  Min.   :  9.815   Min.   :-9.145   Min.   : 21.09   Min.   : 42.84  
##  1st Qu.: 15.195   1st Qu.:12.584   1st Qu.: 48.89   1st Qu.:104.19  
##  Median : 19.563   Median :17.119   Median : 67.51   Median :134.61  
##  Mean   : 25.542   Mean   :18.920   Mean   : 74.34   Mean   :155.24  
##  3rd Qu.: 34.122   3rd Qu.:24.883   3rd Qu.: 96.00   3rd Qu.:201.95  
##  Max.   :133.907   Max.   :78.817   Max.   :198.15   Max.   :430.32  
##      X1100            X1200             X1300            X1400       
##  Min.   : 30.15   Min.   : -8.085   Min.   : 47.83   Min.   : 12.47  
##  1st Qu.: 76.60   1st Qu.: 23.243   1st Qu.:104.25   1st Qu.: 34.09  
##  Median : 97.78   Median : 30.706   Median :137.93   Median : 43.28  
##  Mean   :111.10   Mean   : 34.688   Mean   :155.17   Mean   : 49.01  
##  3rd Qu.:143.30   3rd Qu.: 39.337   3rd Qu.:199.39   3rd Qu.: 59.29  
##  Max.   :346.57   Max.   :132.024   Max.   :417.79   Max.   :142.22  
##      X1500              X1600        
##  Min.   :-25.7129   Min.   : 16.11  
##  1st Qu.:  0.9549   1st Qu.: 28.16  
##  Median :  2.7519   Median : 35.86  
##  Mean   :  2.8168   Mean   : 39.50  
##  3rd Qu.:  4.5267   3rd Qu.: 45.00  
##  Max.   : 48.9075   Max.   :130.13
```

Group1: Adenom

Group2: Hp (hyperplasia)

Group3: Karzinom

Group4: Normal

2.2 Correlation

The CORR Procedure

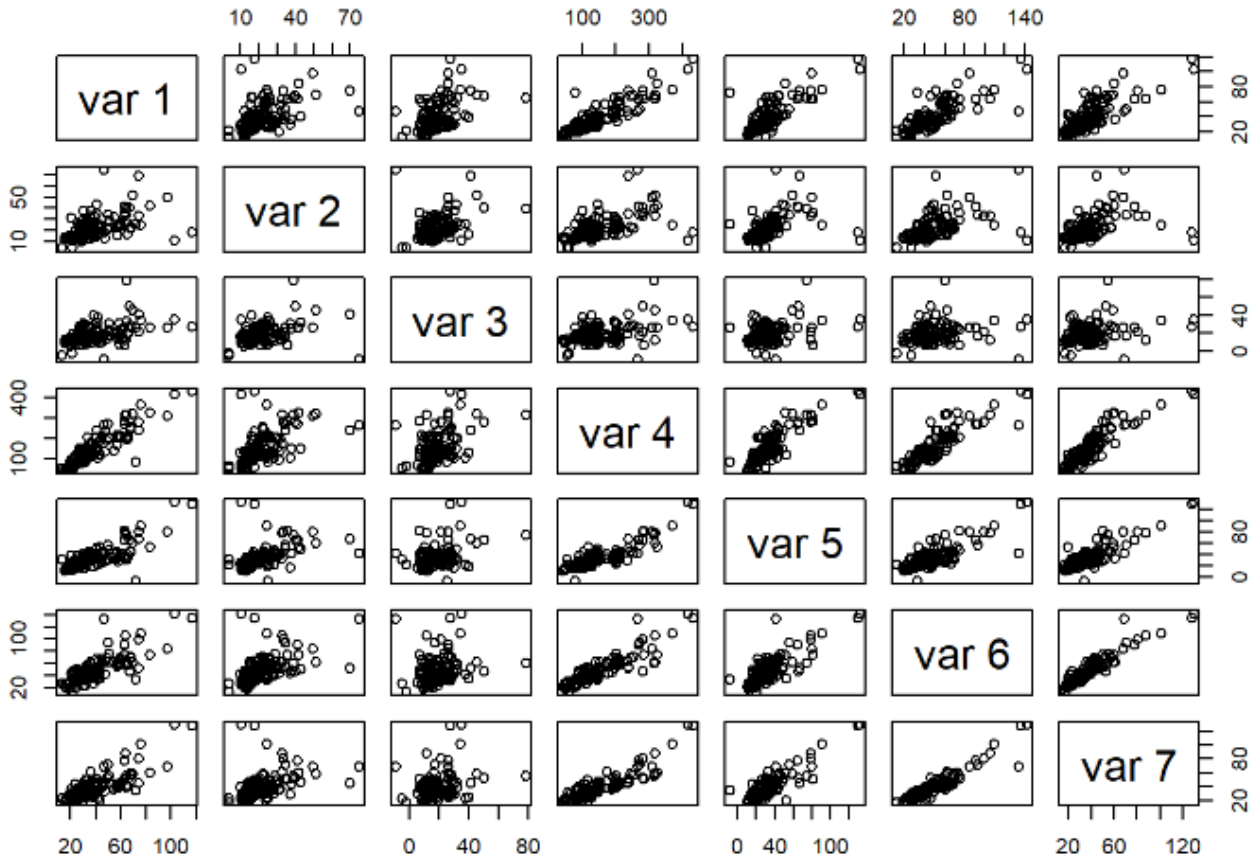
13 Variables:	wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
wv400	120	40.40059	18.96267	4848	12.78675	117.13801
wv500	120	46.61359	21.36755	5594	-3.31698	128.40727
wv600	120	22.10545	11.05560	2653	3.56222	74.92229
wv700	120	25.54176	15.72171	3065	9.81505	133.90667
wv800	120	18.92012	10.84760	2270	-9.14458	78.81740
wv900	120	74.33610	35.07047	8920	21.09073	198.14557
wv1000	120	155.23569	77.81595	18628	42.83851	430.32445
wv1100	120	111.10230	56.29731	13332	30.15437	346.57105
wv1200	120	34.68777	20.38144	4163	-8.08450	132.02398
wv1300	120	155.17171	71.87308	18621	47.83266	417.78657
wv1400	120	49.00679	22.92337	5881	12.46589	142.22443
wv1500	120	2.81684	7.20838	338.02106	-25.71287	48.90754
wv1600	120	39.50316	18.82880	4740	16.10972	130.13334

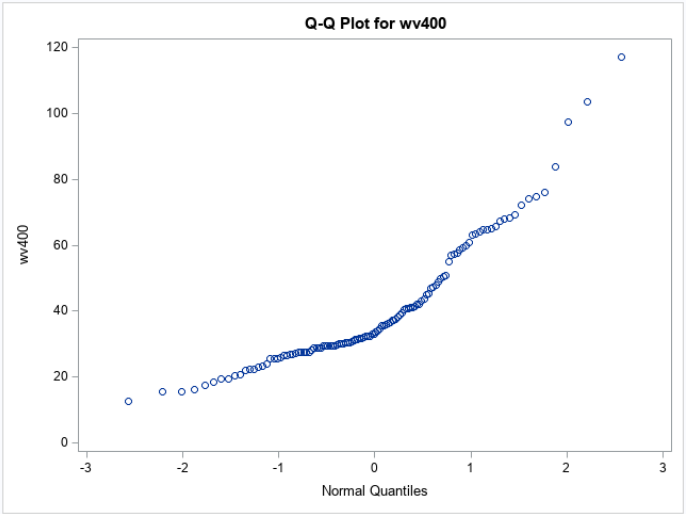
Pearson Correlation Coefficients, N = 120
Prob > |r| under H0: Rho=0

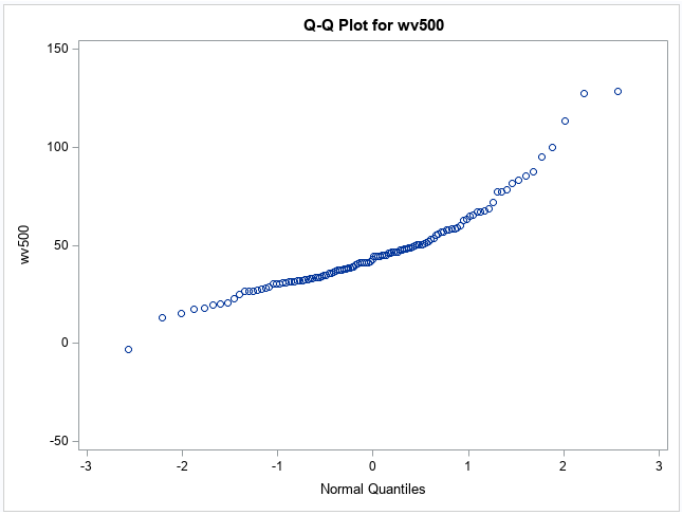
	wv400	wv500	wv600	wv700	wv800	wv900	wv1000	wv1100	wv1200	wv1300	wv1400	wv1500	wv1600
wv400	1.00000	0.86710 <.0001	0.46085 <.0001	0.57373 <.0001	0.46406 <.0001	0.78664 <.0001	0.88377 <.0001	0.88377 <.0001	0.77176 <.0001	0.83758 <.0001	0.78002 <.0001	-0.06030 0.5130	0.79119 <.0001
wv500	0.86710 <.0001	1.00000	0.58553 <.0001	0.65537 <.0001	0.43924 <.0001	0.84482 <.0001	0.88035 <.0001	0.89981 <.0001	0.84865 <.0001	0.85184 <.0001	0.80821 <.0001	-0.02479 0.7881	0.81103 <.0001
wv600	0.46085 <.0001	0.58553 <.0001	1.00000	0.73240 <.0001	0.35280 <.0001	0.68333 <.0001	0.54281 <.0001	0.47862 <.0001	0.42936 <.0001	0.52231 <.0001	0.49929 <.0001	0.45773 <.0001	0.41551 <.0001
wv700	0.57373 <.0001	0.65537 <.0001	0.73240 <.0001	1.00000	0.03504 0.7040	0.77713 <.0001	0.63734 <.0001	0.59725 <.0001	0.50823 <.0001	0.66006 <.0001	0.71354 <.0001	0.37746 <.0001	0.56357 <.0001
wv800	0.46406 <.0001	0.43924 <.0001	0.35280 <.0001	0.03504 0.7040	1.00000	0.29900 0.0009	0.43939 <.0001	0.41886 <.0001	0.36856 <.0001	0.38196 <.0001	0.19228 0.0354	-0.21569 0.0180	0.28980 0.0013
wv900	0.78664 <.0001	0.84482 <.0001	0.68333 <.0001	0.77713 <.0001	0.29900 0.0009	1.00000	0.90847 <.0001	0.85158 <.0001	0.74045 <.0001	0.87298 <.0001	0.80683 <.0001	0.17526 0.0555	0.72893 <.0001
wv1000	0.88377 <.0001	0.88035 <.0001	0.54281 <.0001	0.63734 <.0001	0.43939 <.0001	0.90847 <.0001	1.00000	0.97228 <.0001	0.86294 <.0001	0.96926 <.0001	0.87983 <.0001	-0.02583 0.7795	0.89165 <.0001
wv1100	0.88377 <.0001	0.89981 <.0001	0.47862 <.0001	0.59725 <.0001	0.41886 <.0001	0.85158 <.0001	0.97228 <.0001	1.00000	0.88747 <.0001	0.97573 <.0001	0.91001 <.0001	-0.08744 0.3423	0.93067 <.0001
wv1200	0.77176 <.0001	0.84865 <.0001	0.42936 <.0001	0.50823 <.0001	0.36856 <.0001	0.74045 <.0001	0.86294 <.0001	0.88747 <.0001	1.00000	0.84214 <.0001	0.78929 <.0001	-0.11280 0.2199	0.86896 <.0001
wv1300	0.83758 <.0001	0.85184 <.0001	0.52231 <.0001	0.66006 <.0001	0.38196 <.0001	0.87298 <.0001	0.96926 <.0001	0.97573 <.0001	0.84214 <.0001	1.00000	0.92226 <.0001	-0.03180 0.7302	0.92641 <.0001
wv1400	0.78002 <.0001	0.80821 <.0001	0.49929 <.0001	0.71354 <.0001	0.19228 0.0354	0.80683 <.0001	0.87983 <.0001	0.91001 <.0001	0.78929 <.0001	0.92226 <.0001	1.00000	0.09856 0.2842	0.93290 <.0001
wv1500	-0.06030 0.5130	-0.02479 0.7881	0.45773 <.0001	0.37746 <.0001	-0.21569 0.0180	0.17526 0.0555	-0.02583 0.7795	-0.08744 0.3423	-0.11280 0.2199	-0.03180 0.7302	0.09856 0.2842	1.00000	-0.11190 0.2237
wv1600	0.79119 <.0001	0.81103 <.0001	0.41551 <.0001	0.56357 <.0001	0.28980 0.0013	0.72893 <.0001	0.89165 <.0001	0.93067 <.0001	0.86896 <.0001	0.92641 <.0001	0.93290 <.0001	-0.11190 0.2237	1.00000

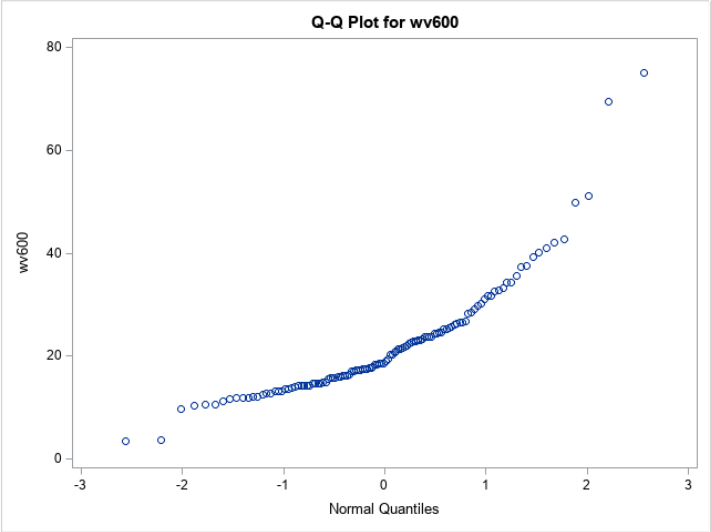


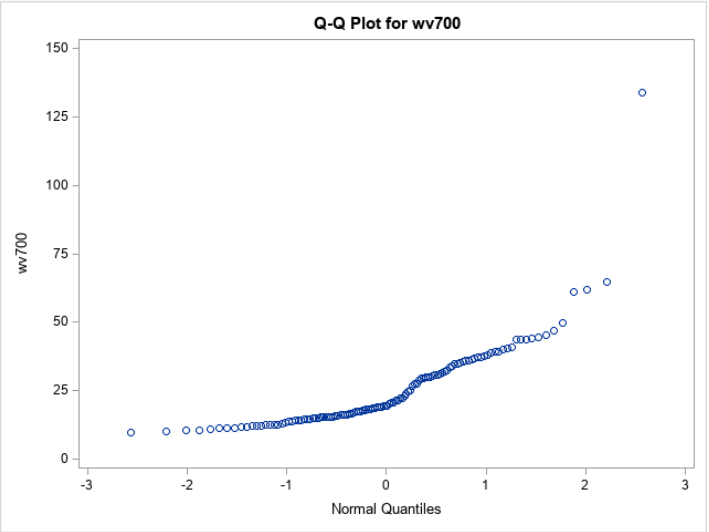
2.3 Normality

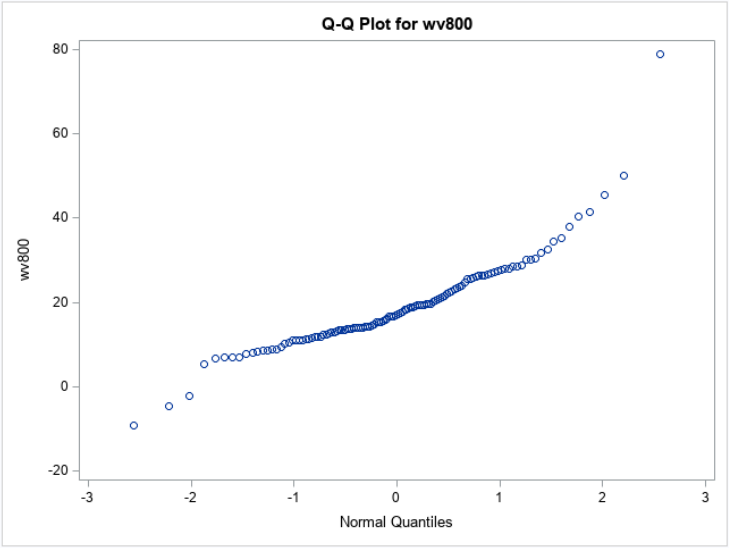
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.855765	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.171965	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.827653	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.957744	Pr > A-Sq	<0.0050

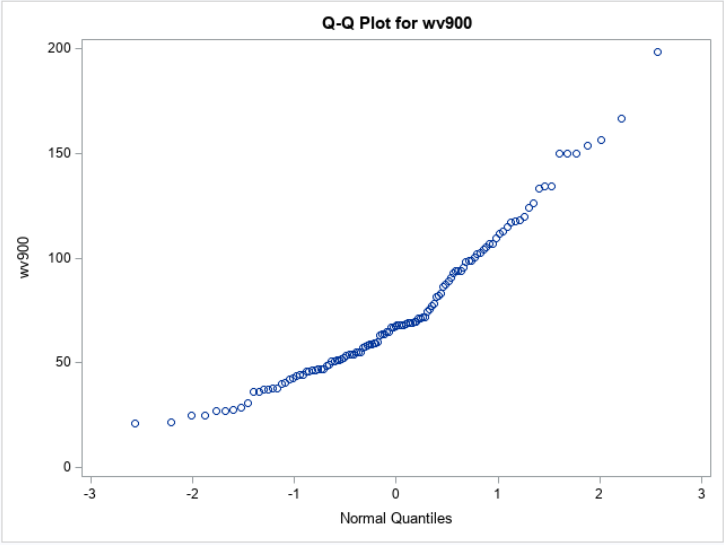


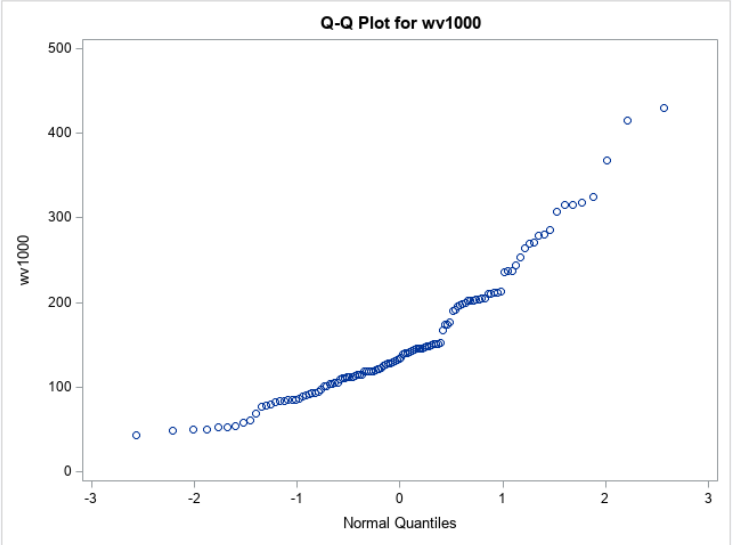


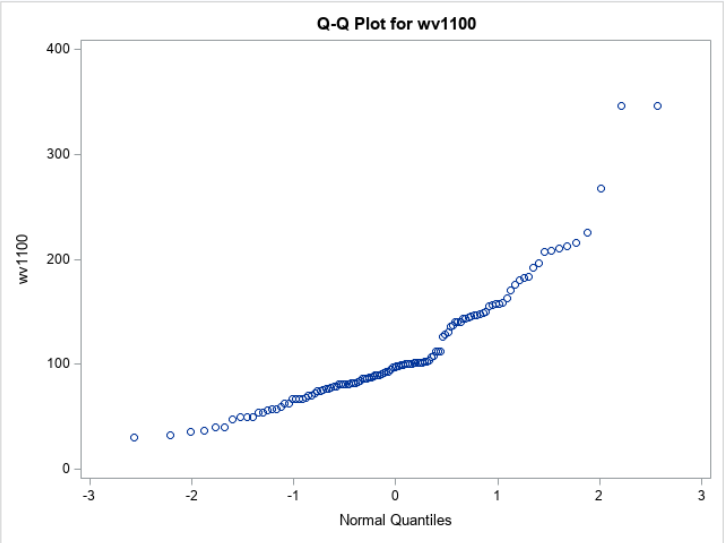

Q-Q Plot for wv600
wv600
Normal Quantiles

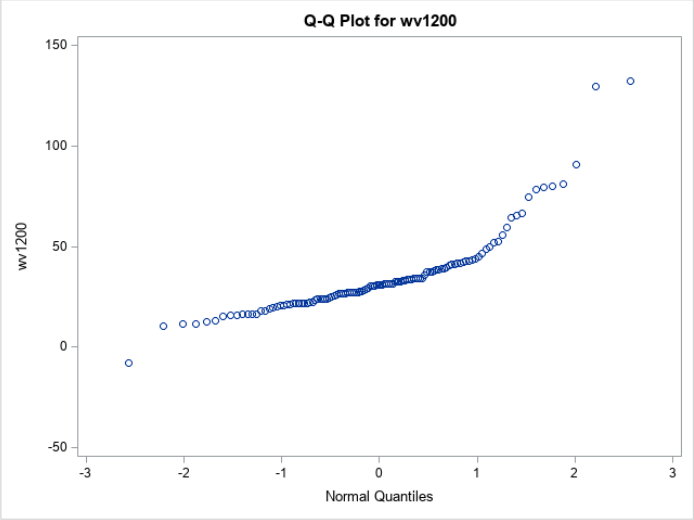

Q-Q Plot for wv700
wv700
Normal Quantiles

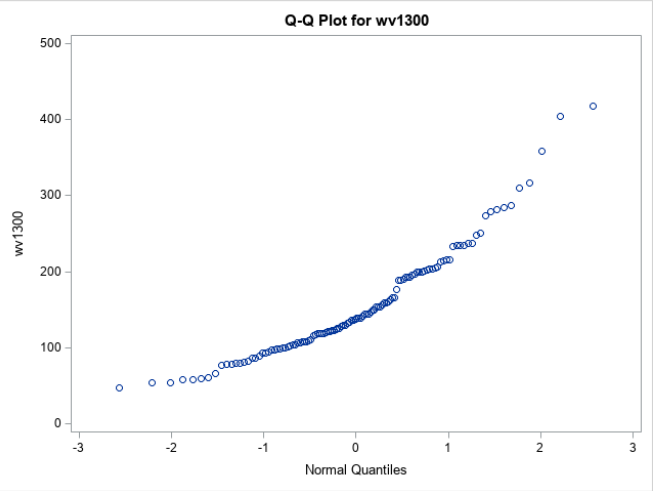

Q-Q Plot for wv800
wv800
Normal Quantiles

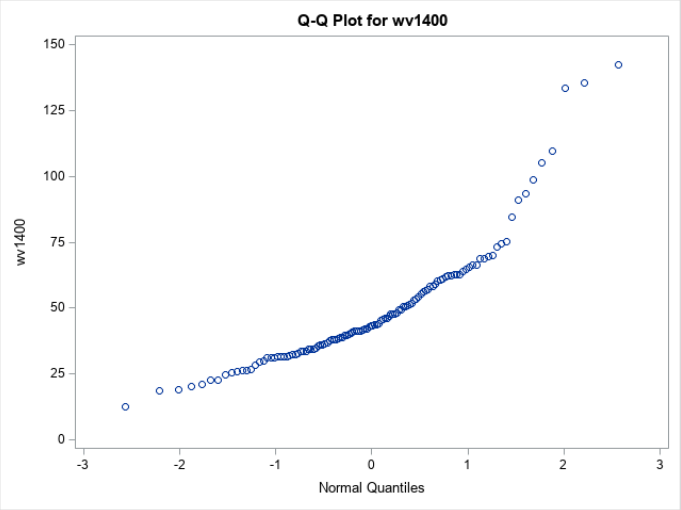

Q-Q Plot for wv900
wv900
Normal Quantiles

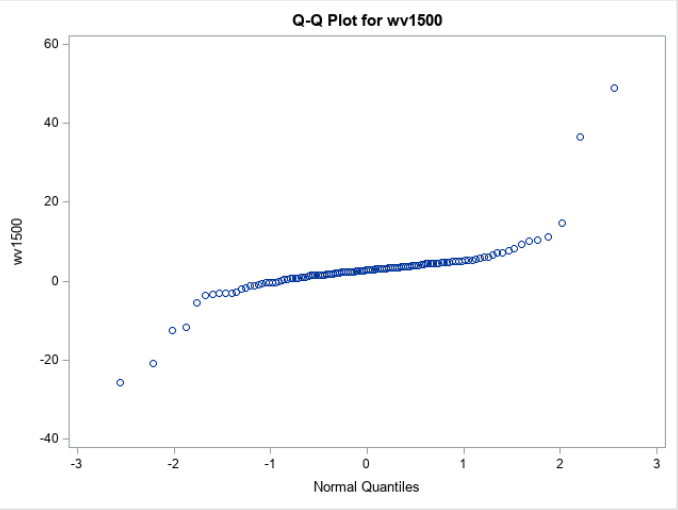

Q-Q Plot for wv1000
wv1000
Normal Quantiles

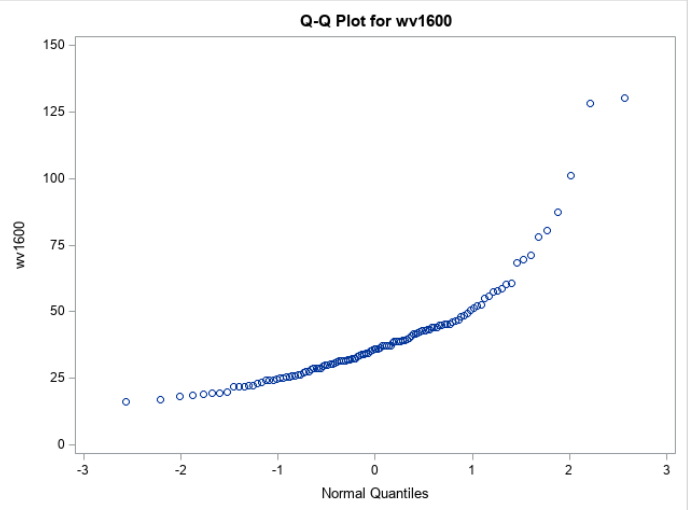

Q-Q Plot for wv1100
wv1100
Normal Quantiles


Q-Q Plot for wv1200
wv1200
150
100
50
0
-50
-3
-2
-1
0
1
2
3
Normal Quantiles


Q-Q Plot for wv1300
wv1300
500
400
300
200
100
0
-3
-2
-1
0
1
2
3
Normal Quantiles


Q-Q Plot for wv1400
wv1400
150
125
100
75
50
25
0
-3
-2
-1
0
1
2
3
Normal Quantiles


Q-Q Plot for wv1500
wv1500
60
40
20
0
-20
-40
-3
-2
-1
0
1
2
3
Normal Quantiles


Q-Q Plot for wv1600
wv1600
150
125
100
75
50
25
0
-3
-2
-1
0
1
2
3
Normal Quantiles

2.4 Interpret and Discuss

Some variables are correlated with each other, which means there is possible multi collinearity and redundancy. Based on the Q-Q plot and the normality test, we can say that the data follow multivariate normal distribution.

3. Multivariate Analysis of Variance

3.1 Motivation

Based on the paper that described the experiment to collect the data, the classification methods of LDA, SVM, KKNN, RF couldn't distinguish hyperplasia from the other carcinogenesis stages, but can distinct normal and non-normal In order to further analyze how each sample of each carcinogenesis stage differ from the other, multivariate analysis for test sample means and sample covariance.

3.2 Method Description, Assumption, and SAS Code

3.2.1 MANOVA

Assumptions of One-way multivariate analysis of variance model are that 1) the four samples of Raman spectra are randomly selected and mutually independent; 2) the observation vectors of four carcinogenesis stages have multivariate normal distribution; 3) the population covariance matrix of observation vectors, $\sum$, is the same for each group of Raman spectra.

The factor effects model for this experiment is as below.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$i = 1, 2, 3, 4 \; ; j = 1, 2, 3, .., 30$$

y_{ij} is the 13×1 observation vector of the four carcinogenesis stages from jth molecule of tissue samples in the ith carcinogenesis stage. μ is the 13×1 overall population mean vector of the four carcinogenesis stages. α_i is the 13×1 effect vector of the ith carcinogenesis stage. $\epsilon_{ij} \sim N_4(\mu, \Sigma)$ is the 13×1 random error vector associated with the from jth molecule of tissue samples in the ith carcinogenesis stage.

The hypothesis to test for a significant difference between at least two mean vectors is as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: at\ least\ two\ population\ mean\ vectors\ differ$$

```
PROC GLM;
CLASS group;
MODEL wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600 =
group;
MANOVA H = group/PRINTE PRINTE MSTAT=EXACT;
RUN;
```

3.2.2 Hypothesis of Equality of Covariance Matrices

In One-way multivariate analysis of variance, we are assuming the equality of the population covariance matrices. However, we don't have sufficient evidence to claim this assumption is correct or wrong. Thus, we need to test the hypothesis of equality of covariance matrices.

The assumptions for this test are that 1) the four samples of Raman spectra are randomly selected and mutually independent; 2) the observation vectors of four carcinogenesis stages have multivariate normal distribution.

The hypothesis to test for a significant difference between at least two covariance matrices is as below:

$$H_0: \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$$

$$H_a: at\ least\ two\ populatin\ covariance\ matrices\ differ$$

```
PROC DISCRIM DATA=chemometric POOL=TEST;
CLASS group;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
run;

PROC STEPDISC STEPWISE SIMPLE STDMEAN TCORR WCORR;
CLASS GROUP;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
TITLE 'STEPWISE SELECTION';
RUN;
```

3.3 Results and Analysis

The results for all the methods for one-way multivariate analysis of variance are listed below. At significant level of 0.05, there is sufficient evidence to reject H_0 and claim that at least two population mean vectors are different.

MANOVA Tests for the Hypothesis of No Overall group Effect
H = Type III SSCP Matrix for group
E = Error SSCP Matrix

S=3 M=4.5 N=51

Statistic	Value	P-Value
Wilks' Lambda	0.25128938	<.0001
Pillai's Trace	1.06440083	<.0001
Hotelling-Lawley Trace	1.86063485	<.0001
Roy's Greatest Root	1.05181080	<.0001

The result for hypothesis of equality of covariance matrices show that at significance level of 0.05, there is sufficient evidence to reject H_0 and claim that at least two population covariance matrices are different.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
988.867647	273	<.0001

The DISCRIM Procedure

Total Sample Size	120	DF Total	119
Variables	13	DF Within Classes	116
Classes	4	DF Between Classes	3

Number of Observations Read	120
Number of Observations Used	120

Class Level Information

group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	30	30.0000	0.250000	0.250000
2	_2	30	30.0000	0.250000	0.250000
3	_3	30	30.0000	0.250000	0.250000
4	_4	30	30.0000	0.250000	0.250000

Within Covariance Matrix Information

group	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	13	37.73176
2	13	54.68957
3	13	54.04841
4	13	41.21567
Pooled	13	57.48403

3.4 Interpret and Discuss

The null hypothesis for equality of sample mean vectors is rejected, which means that there is sufficient evidence to claim at least two population mean vectors differ. Thus, there are variances among groups of four carcinogenesis stages.

The null hypothesis for equality of sample covariance matrices is rejected, which means that there is sufficient evidence to claim that at least two population covariance matrices differ. Thus, the assumption of equal population covariance matrices is violated. And for the following discriminant analysis we will use quadratic classification functions for unequal population covariance matrices.

4. Quadratic Classification for Discriminant Analysis

4.1 Motivation

In section3 we already verified that there are unequal population mean vectors and unequal population covariance matrices. Now we want to allocate observations to groups by quadratic classification functions.

4.2 Method Description and Assumption

The distance function for unequal population covariance matrices is defined as below. Assign y to the group for which $D_i^2(y)$ is a minimum.

$$D_i^2(y) = (y - \bar{y}_i)' S_i^{-1} (y - \bar{y}_i)$$

If we further assume that $f(y|G_i) = N_p(\mu_i, \Sigma_i)$ and incorporate prior probabilities p_1, p_2, p_3, p_4 into classification analysis, the optimal classification rule is as below. Assign y to the group for which $Q_i^2(y)$ is a maximum.

$$Q_i^2(y) = \ln(p_i) - \frac{1}{2}\ln(|S_i|) - \frac{1}{2}(y - \bar{y}_i)' S_i^{-1} (y - \bar{y}_i)$$

```
proc DISCRIM DATA=chemometric LIST CROSSVALIDATE POOL=TEST;
CLASS GROUP;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
RUN;
```

4.3 Results and Analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CHEMOMETRIC
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into group					
From group	1	2	3	4	Total
1	30 100.00	0 0.00	0 0.00	0 0.00	30 100.00
2	1 3.33	26 86.67	0 0.00	3 10.00	30 100.00
3	1 3.33	0 0.00	29 96.67	0 0.00	30 100.00
4	1 3.33	0 0.00	0 0.00	29 96.67	30 100.00
Total	33 27.50	26 21.67	29 24.17	32 26.67	120 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for group					
	1	2	3	4	Total
Rate	0.0000	0.1333	0.0333	0.0333	0.0500
Priors	0.2500	0.2500	0.2500	0.2500	

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CHEMOMETRIC
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into group					
From group	1	2	3	4	Total
1	21 70.00	4 13.33	4 13.33	1 3.33	30 100.00
2	1 3.33	22 73.33	2 6.67	5 16.67	30 100.00
3	1 3.33	6 20.00	22 73.33	1 3.33	30 100.00
4	1 3.33	6 20.00	2 6.67	21 70.00	30 100.00
Total	24 20.00	38 31.67	30 25.00	28 23.33	120 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for group					
	1	2	3	4	Total
Rate	0.3000	0.2667	0.2667	0.3000	0.2833
Priors	0.2500	0.2500	0.2500	0.2500	

From the result we can see that all the classifications in group 1 are correct, and only one observation in group 3 and group4 are wrongly allocated, 2 observations in group 2 are wrongly allocated in group 4.

4.4 Interpret and Discuss

Group 2, which is hyperplasia has higher error rate than the other groups and wrongly allocated in group 4 (normal), which is align with the conclusions in [1]. Intuitively understanding is that Hyperplasia is an initial stage of tumor during which period the chemical composition haven't changed much.

5. Effect of Principal Component Analysis on Group Variances

5.1 Motivation

PCA is a method to abstract information from dataset and to reduce redundancy. It is applied mainly for the purpose of data visualization. But how do PCA perform on the MANOVA? Does it make sense if we apply MANOVA analysis and discriminate analysis on principal components? If we use principal components of the data to perform hypothesis testing of equal population mean vectors and equal population covariance matrices, are the results different than the ones on the original dataset?

5.2 Method Description and Assumption

Principle components are the linear combination of variables based on eigenvectors of sample covariance matrix. We first apply principal components and then perform the methods in section3 and section4 on the principle components.

$$Z_i = Ay_i$$

$$S = CDC'$$

Thus,

$$S_z = ASA' = C'SC = D = Diag(\lambda_1, \lambda_2, \dots, \lambda_p)$$

Principal components:

$$Z_1 = a_1' y$$

$$Z_2 = a_2' y$$

$$\dots$$

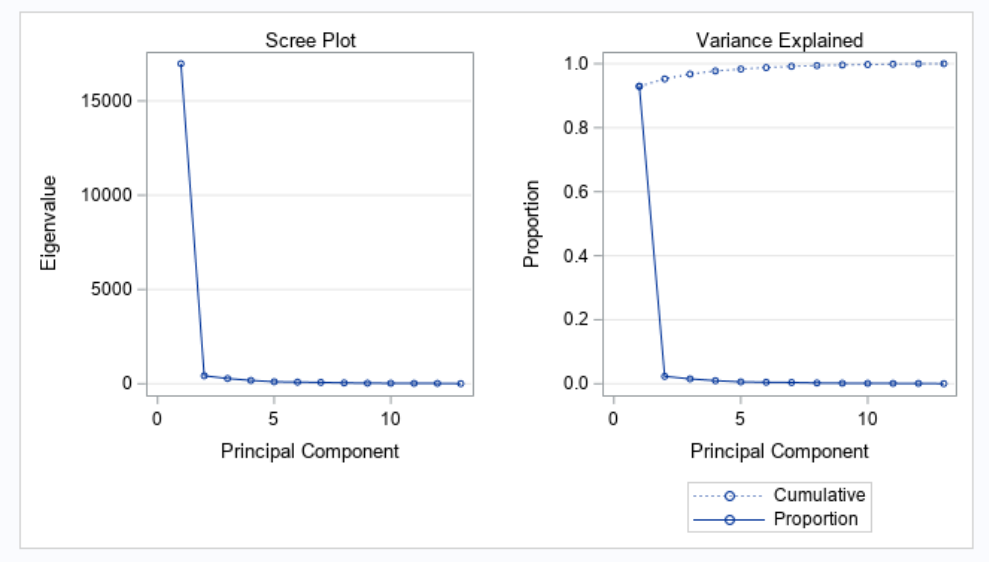
```
PROC PRINCOMP DATA=chemometric COV OUT=RESULTS;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
RUN;
PROC PRINT DATA=RESULTS;
RUN;
PROC PLOT DATA=RESULTS;
PLOT PRIN2*PRIN1;
RUN;
```

5.3 Results and Analysis

5.3.1 Principal Components

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	16976.8621	16553.9048	0.9292	0.9292
2	422.9573	144.7209	0.0231	0.9523
3	278.2365	101.4612	0.0152	0.9675
4	176.7753	69.8474	0.0097	0.9772
5	106.9279	21.1899	0.0059	0.9831
6	85.7379	14.1060	0.0047	0.9878
7	71.6320	27.2002	0.0039	0.9917
8	44.4318	12.6176	0.0024	0.9941
9	31.8142	3.9095	0.0017	0.9959
10	27.9046	6.4778	0.0015	0.9974
11	21.4268	2.1104	0.0012	0.9986
12	19.3164	12.3593	0.0011	0.9996
13	6.9571		0.0004	1.0000

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
wv400	0.128258	-.004701	-.256543	0.245542	0.137837	-.636126	0.107227	-.317034	0.407491	0.346878	-.153832	-.085257	-.066727
wv500	0.146919	0.089091	-.195136	0.548280	0.212643	0.013486	-.286524	0.018201	-.412764	-.136881	-.382057	0.406144	-.062701
wv600	0.046523	0.311615	0.027272	0.109434	0.412361	0.271295	0.305453	0.210470	0.080115	-.070961	-.328958	-.538671	-.311347
wv700	0.080390	0.415367	0.323304	0.246928	0.144843	-.085162	0.214783	-.493469	-.103922	-.356883	0.402661	-.014948	0.192633
wv800	0.034129	-.075850	-.360814	-.104338	0.650147	0.162946	0.128751	0.120535	-.091032	0.294166	0.456571	0.177218	0.192292
wv900	0.243717	0.680129	-.057106	-.042831	-.202305	0.076151	-.435587	0.149583	0.043750	0.399889	0.107462	-.111265	0.168344
wv1000	0.592478	0.101542	-.424195	-.315929	-.276344	-.055079	0.417765	-.020480	-.220493	-.203204	-.045674	0.080693	-.077799
wv1100	0.426659	-.315930	-.057076	0.270275	0.033045	-.112467	-.310493	0.334720	0.207478	-.386423	0.289111	-.349784	0.146142
wv1200	0.136689	-.147122	-.145029	0.416801	-.284833	0.629832	0.128681	-.273683	0.339242	0.181427	0.136602	0.094278	-.141726
wv1300	0.545317	-.174336	0.487413	-.320598	0.310544	0.132762	-.210074	-.244769	0.136180	0.120149	-.205173	0.176065	-.087754
wv1400	0.161164	-.029378	0.407673	0.257675	-.123324	-.220898	0.268998	0.404407	-.204155	0.322058	0.302925	0.160657	-.421527
wv1500	-.001219	0.210519	0.121976	0.039841	0.024204	-.006293	0.277388	0.403952	0.529942	-.190213	-.179304	0.482665	0.341075
wv1600	0.132955	-.216187	0.185773	0.189945	-.101606	0.055911	0.282368	0.021225	-.297230	0.322838	-.263957	-.257056	0.665614



The first eigenvalue explain 92.92% variances of variables.

5.3.2 MANOVA by Principal Components

MANOVA Tests for the Hypothesis of No Overall group Effect
H = Type III SSCP Matrix for group
E = Error SSCP Matrix

S=3 M=4.5 N=51

Statistic	Value	P-Value
Wilks' Lambda	0.25128938	<.0001
Pillai's Trace	1.06440083	<.0001
Hotelling-Lawley Trace	1.86063485	<.0001
Roy's Greatest Root	1.05181080	<.0001

The hypothesis testing results are same as the results for original data, and all reach to the conclusion to reject the null hypothesis.

5.3.3 Hypothesis Testing on Equal Covariance Matrices by Principal Components

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
988.867647	273	<.0001

The hypothesis testing results are same as the results for original data, and all reach to the conclusion to reject the null hypothesis.

The DISCRIM Procedure

Total Sample Size	120	DF Total	119
Variables	13	DF Within Classes	116
Classes	4	DF Between Classes	3

Number of Observations Read	120
Number of Observations Used	120

Class Level Information

group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	30	30.0000	0.250000	0.250000
2	_2	30	30.0000	0.250000	0.250000
3	_3	30	30.0000	0.250000	0.250000
4	_4	30	30.0000	0.250000	0.250000

Within Covariance Matrix Information

group	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	13	37.73176
2	13	54.68957
3	13	54.04841
4	13	41.21567
Pooled	13	57.48403

5.3.4 Quadratic Classification by Principal Components

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.RESULTS
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into group

From group	1	2	3	4	Total
1	30 100.00	0 0.00	0 0.00	0 0.00	30 100.00
2	1 3.33	26 86.67	0 0.00	3 10.00	30 100.00
3	1 3.33	0 0.00	29 96.67	0 0.00	30 100.00
4	1 3.33	0 0.00	0 0.00	29 96.67	30 100.00
Total	33 27.50	26 21.67	29 24.17	32 26.67	120 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for group

	1	2	3	4	Total
Rate	0.0000	0.1333	0.0333	0.0333	0.0500
Priors	0.2500	0.2500	0.2500	0.2500	

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.RESULTS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into group					
From group	1	2	3	4	Total
1	21 70.00	4 13.33	4 13.33	1 3.33	30 100.00
2	1 3.33	22 73.33	2 6.67	5 16.67	30 100.00
3	1 3.33	6 20.00	22 73.33	1 3.33	30 100.00
4	1 3.33	6 20.00	2 6.67	21 70.00	30 100.00
Total	24 20.00	38 31.67	30 25.00	28 23.33	120 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for group					
	1	2	3	4	Total
Rate	0.3000	0.2667	0.2667	0.3000	0.2833
Priors	0.2500	0.2500	0.2500	0.2500	

From the result we can see that all the classifications in group 1 are correct, and only one observation in group 3 and group4 are wrongly allocated, 2 observations in group 2 are wrongly allocated in group 4.

5.4 Interpret and Discuss

The first principal component can explain 92.92% variances in the variables. After we apply MANOVA and discriminant analysis, we can see that the results from principal components are identical as the results from the original data.

Is it necessary if we want to ask “why”?

6. Cluster Analysis: Natural Grouping of Sample Units

6.1 Motivation

We know the label of each observation, if we do natural grouping by clustering analysis, will the observations from one group be clustered as one cluster?

6.2 Method Description and Assumption

K-means clustering: 1) select g items to serve as seeds, i.e., the initial cluster centroids; 2) assign each observation in the data set to the cluster with the nearest seed/centroid based on the Euclidean distance; 3) recalculate the centroid for each cluster. The centroid for a given vector is the mean vector of all observation in that cluster; 4) repeat 2) and 3) until no observations move to different clusters.

```
proc standard data=chemometric out=chemometric0 mean=0 std=1;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
run;

proc print data=chemometric0;
run;

proc fastclus data=chemometric0 radius=3 maxc=4 replace=full maxiter=10 out=Clus_OUT;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
id molecule;
run;
proc sort data=Clus_OUT;
by cluster distance;
run;
proc print data=Clus_OUT;
var molecule cluster distance;
run;
```

6.3 Results and Analysis

6.3.1 Clusters on original data

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	78	0.4683	3.5705	> Radius	3	4.4579
2	3	0.9048	3.6263	> Radius	3	8.5748
3	38	0.7781	6.0740	> Radius	1	4.4579
4	1	.	0		3	10.9191

Table 1 Clusters on original data

Obs	molecule	CLUSTER	DISTANCE
1	**99**	1	0.61536
2	**92**	1	0.66845
3	**5**	1	0.68181

Obs	molecule	CLUSTER	DISTANCE
4	**70**	1	0.69603
5	**9**	1	0.69622
6	**6**	1	0.70940
7	**67**	1	0.74834
8	**14**	1	0.78682
9	**77**	1	0.80533
10	**95**	1	0.81263
11	**78**	1	0.88595
12	**11**	1	0.90507
13	**75**	1	0.90673
14	**102**	1	0.91803
15	**91**	1	0.92448
16	**68**	1	0.96178
17	**100**	1	0.97801
18	**16**	1	1.04229
19	**19**	1	1.09802
20	**97**	1	1.10655
21	**73**	1	1.12006
22	**3**	1	1.13000
23	**65**	1	1.15726
24	**71**	1	1.16748
25	**118**	1	1.17566
26	**69**	1	1.17998
27	**76**	1	1.18799
28	**26**	1	1.20348
29	**4**	1	1.20946
30	**20**	1	1.21142
31	**2**	1	1.21322
32	**80**	1	1.21726
33	**38**	1	1.22349
34	**24**	1	1.23435
35	**98**	1	1.23990
36	**74**	1	1.24742
37	**93**	1	1.26508
38	**63**	1	1.31500
39	**119**	1	1.32924
40	**7**	1	1.37969

Obs	molecule	CLUSTER	DISTANCE
41	**1**	1	1.38542
42	**113**	1	1.47375
43	**117**	1	1.54196
44	**116**	1	1.55651
45	**79**	1	1.55671
46	**18**	1	1.56308
47	**112**	1	1.56735
48	**12**	1	1.57098
49	**96**	1	1.58623
50	**25**	1	1.61591
51	**10**	1	1.62098
52	**83**	1	1.63020
53	**13**	1	1.63100
54	**21**	1	1.67788
55	**8**	1	1.73690
56	**84**	1	1.75212
57	**103**	1	1.77763
58	**17**	1	1.79626
59	**81**	1	1.81544
60	**15**	1	1.84369
61	**23**	1	1.94241
62	**44**	1	2.05881
63	**94**	1	2.21486
64	**62**	1	2.21531
65	**85**	1	2.29224
66	**60**	1	2.36860
67	**22**	1	2.40634
68	**37**	1	2.47654
69	**109**	1	2.48150
70	**31**	1	2.49683
71	**61**	1	2.52556
72	**59**	1	2.52654
73	**101**	1	2.53638
74	**32**	1	2.66325
75	**82**	1	3.04429
76	**72**	1	3.13051
77	**66**	1	3.18831

Obs	molecule	CLUSTER	DISTANCE
78	**64**	1	3.57053
79	**87**	2	1.93870
80	**88**	2	2.09154
81	**86**	2	3.62630
82	**108**	3	0.85945
83	**56**	3	1.00477
84	**50**	3	1.22303
85	**47**	3	1.33021
86	**34**	3	1.39006
87	**58**	3	1.43301
88	**27**	3	1.58613
89	**52**	3	1.63243
90	**30**	3	1.65044
91	**111**	3	1.70571
92	**114**	3	1.75280
93	**48**	3	1.78156
94	**28**	3	1.78497
95	**54**	3	1.80734
96	**55**	3	1.84057
97	**57**	3	1.90494
98	**49**	3	1.91139
99	**40**	3	2.00309
100	**110**	3	2.21270
101	**53**	3	2.26853
102	**115**	3	2.28074
103	**36**	3	2.28433
104	**120**	3	2.31963
105	**29**	3	2.34304
106	**51**	3	2.35826
107	**107**	3	2.44679
108	**42**	3	2.82145
109	**39**	3	2.85589
110	**104**	3	2.89211
111	**90**	3	3.14965
112	**46**	3	3.28353
113	**33**	3	3.50975
114	**45**	3	3.66051

Obs	molecule	CLUSTER	DISTANCE
115	**106**	3	4.14542
116	**43**	3	4.78290
117	**105**	3	4.95895
118	**89**	3	5.37658
119	**41**	3	6.07398
120	**35**	4	0.00000

We know that molecules at ID 1-30 are in group1, 31-60 are in group2, 61-90 are in group3, 91-120 are in group4. But the molecules in each cluster are not as expected.

Based on Table 1, we can see that molecule 86, 87, 88 are clustered as one group, molecule 35 as one group. Most sample units are clustered into the other two groups.

6.3.2 Clusters on Principal Components

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	1	.	0		3	9.4704
2	3	1.4677	4.3833	> Radius	4	5.2902
3	49	1.0029	8.3800	> Radius	4	1.6374
4	67	0.8481	6.9165	> Radius	3	1.6374

Table2 Clusters on Principal Components

Obs	molecule	CLUSTER	DISTANCE
1	**35**	1	0.00000
2	**86**	2	4.28621
3	**90**	2	4.29194
4	**89**	2	4.38334
5	**5**	3	1.49771
6	**62**	3	1.58709
7	**59**	3	1.64709
8	**60**	3	1.64785
9	**63**	3	1.65377
10	**94**	3	1.68683
11	**9**	3	1.77601
12	**44**	3	1.88128
13	**69**	3	1.89822

Obs	molecule	CLUSTER	DISTANCE
14	**19**	3	1.90875
15	**71**	3	1.92507
16	**4**	3	1.95911
17	**116**	3	1.98846
18	**115**	3	2.04658
19	**98**	3	2.12099
20	**61**	3	2.14484
21	**20**	3	2.16439
22	**21**	3	2.23990
23	**23**	3	2.25682
24	**113**	3	2.27005
25	**112**	3	2.30956
26	**26**	3	2.55683
27	**42**	3	2.61597
28	**25**	3	2.63144
29	**40**	3	2.69006
30	**114**	3	2.73651
31	**29**	3	2.90002
32	**103**	3	2.91057
33	**1**	3	2.99911
34	**27**	3	3.10074
35	**30**	3	3.23399
36	**47**	3	3.27861
37	**83**	3	3.32945
38	**81**	3	3.45411
39	**22**	3	3.47392
40	**31**	3	3.49855
41	**39**	3	3.54335
42	**109**	3	3.97287
43	**104**	3	4.13975
44	**82**	3	4.42354
45	**120**	3	4.66021
46	**106**	3	4.66447
47	**24**	3	4.79969
48	**105**	3	5.10324
49	**87**	3	6.28403
50	**43**	3	6.48872

Obs	molecule	CLUSTER	DISTANCE
51	**88**	3	6.85146
52	**66**	3	7.28672
53	**64**	3	8.37999
54	**95**	4	1.20281
55	**91**	4	1.35045
56	**100**	4	1.46021
57	**6**	4	1.46049
58	**97**	4	1.48255
59	**99**	4	1.55221
60	**10**	4	1.58064
61	**92**	4	1.66126
62	**67**	4	1.70023
63	**14**	4	1.74874
64	**58**	4	1.79460
65	**2**	4	1.87260
66	**119**	4	1.89970
67	**13**	4	1.90689
68	**102**	4	1.90840
69	**3**	4	1.93319
70	**93**	4	1.96453
71	**68**	4	1.98357
72	**7**	4	2.05712
73	**65**	4	2.07177
74	**70**	4	2.10036
75	**56**	4	2.12232
76	**76**	4	2.20520
77	**11**	4	2.20894
78	**77**	4	2.27670
79	**38**	4	2.29421
80	**96**	4	2.30979
81	**101**	4	2.32825
82	**18**	4	2.33350
83	**75**	4	2.34966
84	**34**	4	2.39501
85	**78**	4	2.41142
86	**16**	4	2.42092
87	**12**	4	2.48357

Obs	molecule	CLUSTER	DISTANCE
88	**118**	4	2.49372
89	**74**	4	2.59546
90	**117**	4	2.59735
91	**57**	4	2.68104
92	**49**	4	2.83235
93	**73**	4	2.86610
94	**15**	4	2.87971
95	**52**	4	2.88124
96	**51**	4	2.99249
97	**28**	4	3.05484
98	**17**	4	3.09099
99	**54**	4	3.09440
100	**55**	4	3.10866
101	**48**	4	3.22316
102	**32**	4	3.24212
103	**53**	4	3.30180
104	**8**	4	3.31866
105	**107**	4	3.36334
106	**79**	4	3.37919
107	**108**	4	3.44471
108	**110**	4	3.56877
109	**50**	4	3.64648
110	**111**	4	3.71889
111	**37**	4	3.89357
112	**80**	4	3.97949
113	**84**	4	4.13262
114	**72**	4	4.23171
115	**45**	4	4.85995
116	**85**	4	4.87614
117	**46**	4	5.05227
118	**36**	4	5.13635
119	**33**	4	6.71147
120	**41**	4	6.91653

We know that molecule 1-30 are in group1, 31-60 are in group2, 61-90 are in group3, 91-120 are in group4. But the molecules in each cluster are not as expected.

Based on Table 1, we can see that molecule 86, 89, 90 are clustered as one group, molecule 35 as one group. Most sample units are clustered into the other two groups.

6.4 Interpret and Discuss

The clusters are different based on original data and based on principal components. But based on the clusters in section 6.3.1 and section 6.3.2, we can see that molecule 35 is clustered as one group and molecule 86, 87, 88, 89, 90 are clustered as one group. The other majority molecules are in other two groups.

By K-means clustering analysis, we can identify that molecule 35, 86, 87, 88, 89, 90 are outliers.

7. Factor Analysis: Natural Grouping of Variables

7.1 Motivation

The variables are wavenumber of 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600. Are there any patterns among these variables? How these variables of wavenumbers relate to each other?

7.2 Method Description and Assumption

Principal Component Method

$$S = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$$

Step1:

$$S = \hat{\Lambda}\hat{\Lambda}'$$

$$S = CDC' = CD^{\frac{1}{2}}(CD^{\frac{1}{2}})'$$

Thus,

$$\hat{\Lambda} = CD^{\frac{1}{2}} = (\sqrt{\theta_1}C_1, \quad \sqrt{\theta_2}C_2, \quad \ldots \sqrt{\theta_m}C_m)$$

Step2:

$$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$$

$$\Psi_i = S_{ii} - h_i^2$$

Step3:

$$E = S - (\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})$$

```
proc factor data=chemometric method=prin nfactors=2 corr scree ev residuals plot;
VAR wv400 wv500 wv600 wv700 wv800 wv900 wv1000 wv1100 wv1200 wv1300 wv1400 wv1500 wv1600;
run;
```

7.3 Results and Analysis

Overall 4 groups

Factor Pattern		
	Factor1	**Factor2**
wv400	0.89541	-0.14698
wv500	0.93197	-0.05612
wv600	0.63239	0.55501
wv700	0.73174	0.53742
wv800	0.41847	-0.37225
wv900	0.91565	0.20087
wv1000	0.97396	-0.09291
wv1100	0.97131	-0.16834
wv1200	0.88365	-0.20365
wv1300	0.96662	-0.08229
wv1400	0.92199	0.05736
wv1500	0.03881	0.89711
wv1600	0.91262	-0.17826

Overall 4 groups after PCA

Factor Pattern		
	Factor1	**Factor2**
Prin1	1.00000	0.00000
Prin2	0.00000	0.00000
Prin3	0.00000	0.00000
Prin4	0.00000	0.00000
Prin5	0.00000	0.00000
Prin6	0.00000	0.47536
Prin7	0.00000	0.00000
Prin8	0.00000	0.00000
Prin9	0.00000	0.00000
Prin10	0.00000	0.00000
Prin11	0.00000	0.00000
Prin12	0.00000	0.52383
Prin13	0.00000	-0.70685

Normal Group

Factor Pattern		
	Factor1	**Factor2**
wv400	0.91983	-0.01850
wv500	0.96451	-0.09423
wv600	0.83257	-0.07548
wv700	0.92192	-0.01326
wv800	0.53299	-0.03963
wv900	0.97022	-0.15312
wv1000	0.98427	-0.07872
wv1100	0.98555	0.02289
wv1200	0.95869	-0.00484
wv1300	0.97446	0.06973
wv1400	0.95417	0.12709
wv1500	0.06687	0.98779
wv1600	0.93277	0.16826

Hy group

Factor Pattern		
	Factor1	**Factor2**
wv400	0.77505	-0.36793
wv500	0.88487	-0.19908
wv600	0.70309	0.23017
wv700	0.72181	0.63707
wv800	0.21920	-0.79052
wv900	0.95580	0.09501
wv1000	0.93136	-0.23849
wv1100	0.91425	-0.29928
wv1200	0.78649	-0.34665
wv1300	0.93326	-0.10482
wv1400	0.79208	0.50784
wv1500	0.39801	0.83677
wv1600	0.92391	0.05740

Karzinom Group

Factor Pattern		
	Factor1	**Factor2**
wv400	0.92032	0.06650
wv500	0.92815	-0.06610
wv600	0.37445	0.84400
wv700	0.87205	-0.09280
wv800	0.60883	0.09327
wv900	0.94402	0.17267
wv1000	0.98361	0.05442
wv1100	0.99078	-0.03020
wv1200	0.94204	-0.08925
wv1300	0.98726	-0.04152
wv1400	0.97906	-0.01155
wv1500	-0.31306	0.84523
wv1600	0.97921	-0.07901

Adenom Group

Factor Pattern		
	Factor1	**Factor2**
wv400	0.92629	0.03218
wv500	0.89735	0.05270
wv600	0.66553	0.52184
wv700	0.74605	-0.42443
wv800	0.45266	0.62868
wv900	0.80256	0.24472
wv1000	0.96900	0.11748
wv1100	0.98771	-0.03549
wv1200	0.80625	0.35329
wv1300	0.97781	-0.03401
wv1400	0.86699	-0.36318
wv1500	-0.29568	0.76629
wv1600	0.84006	-0.44956

7.4 Interpret and Discuss

wv600, wv700, wv800 don’t clearly associate with any factor; wv400, wv500, wv900, wv1000, wv1100, wv1200, wv1300, wv1400, wv1600 associate with factor1; wv1500 associate with factor2.

8. Conclusion

We performed hypothesis testing on equal population mean vectors and equal population covariance matrices in section 3, the null hypothesis of both equal population mean vectors and equal population covariance matrices are rejected, which mean that there are at least two population mean vectors are different with each other and there are at least two population covariance matrices are different with each other. In section 4 we applied quadratic classification of discriminant function to allocate an observation into a group. The result showed that group 2 have higher error rate and this result align with the conclusion from [1], which is that the classification models couldn't distinguish hyperplasia (group2) from other groups.

In section5 we performed principal components analysis, and applied the methods in section 3 and section 4 on the principal components to see whether the results are same as in senction3 and section4, which we can further verify the performance of principal component analysis. It turned out that the results are same, which means that principal component don't exert any effect on these analysis. As it is commonly noted, principal components are mainly for the purpose of data visualization.

We also applied K-means clustering analysis in section6. Because we know the label of each unit, we can verify whether the sample units with same labels are clustered as one group. Based on the results, we know that all the sample units are clustered into groups regardless of their labels. Intuitively understanding is that K-means is based on the closeness of distance, but the labels are annotated by some professional pathologists based on tissue images, thus, the distance of the Raman spectra doesn't reflect the pathology.

Section 7 is about finding natural groupings among the variables, in which way we can analyze how different wavenumber of Raman spectra relate to each other. Based on the results, we know that wv600, wv700, wv800 don't clearly associate with any factor; wv400, wv500, wv900, wv1000, wv1100, wv1200, wv1300, wv1400, wv1600 associate with factor1; wv1500 associate with factor2.

Reference

[1] Nadine Vogler, et al., Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics, 2015

[2] Shuxia Guo, et al., Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling, 2021

[3] Shuxia Guo, et al., Modified PCA and PLS: Towards a better classification in Raman spectroscopy-based biological applications, 2019

Appendix

SAS code

Output file