DIMENSION REDUCTION TECHNIQUES FOR FUNCTIONAL DATA: AN ILLUSTRATION USING A CANCER SCREENING MEDICAL DEVICE

by

Dalong Cao

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Statistics

Summer 2013

DIMENSION REDUCTION TECHNIQUES

FOR FUNCTIONAL DATA:

AN ILLUSTRATION USING

A CANCER SCREENING MEDICAL DEVICE

by

Dalong Cao

Approved: _____
Jong Soo Lee, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Titus O. Awokuse, Ph.D.
Chair of the Department of Food and Resource Economics

Approved: _____
Mark Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

It takes one and a half years to write the thesis whereby many people have accompanied, supported and helped me improve my work. At first, I would like to especially thank my advisor Jong Soo Lee, for his helpful suggestions, patient guidance, continuous correctness and all the help of my study in the Master's program in Agricultural and Resource Economics. Meanwhile, I have learned a lot and gained plenty of experience from Jong Soo Lee in both academic life and academic research. Also, I would like to show my sincere gratitude to Paul P. Eggermont and Dennis D. Cox for serving as my thesis committee members, helping me improving my results of the thesis all the time. At last, I would like to thank Vincent Lariccia for providing me with many related books and materials which really improve my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Cervical cancer has been one of the most common cancers among women, especially in the developing countries. Our research group has built a new medical device which uses fluorescence spectroscopy for early detection of cervical cancer. The output of the device belongs to the functional data. My role in the project is to determine if the factor of room light is statistically significant using the data setting from the experiment. It is impossible to use traditional statistical method for large functional data because the number of dimensions is much larger than the number of observations. The thesis comes up with some ideas of dimension reduction techniques including PCA and EDA.

Based upon the thoughts of Adaptively Truncated Hotelling T-Square Test, this thesis extends the method to more than two groups which we call it Adaptively Truncated MANOVA. Simulations are made on the three test statistics of Adaptively Truncated MANOVA. However, whether we can apply the Adaptively Truncated MANOVA to the real data still needs more work in the future.

Chapter 1

INTRODUCTION

Chapter 1 shows a brief introduction of the research. The first part shows the background on current diagnose of cervix cancer and the history of cervix cancer. The second part introduces the information about my work in the project and the existing issues. The third part outlines the organization of the whole thesis.

1.1 Background of Cervical Cancer

Cervical cancer is the most common form of cancer in the developing countries and the second most common form of cancer in the worldwide. Every year, 9 of 100,000 women died from cervical cancer in the world. For example, 473,000 persons were detected for cervical cancer and nearly 253,500 died in the year of 2008. In the United States, about 3,870 women died with 11,000 new cases. A large number of deaths occur among the Hispanic women compared to the general population. It was not until 1940s that people found that the HPV (Human Papillomavirus) was the main cause of cervical cancer and HPV-DNA was detected in 1963. Despite HPV, some other viruses and smoking may cause cervical cancer.

In most of the times, there may be no obvious symptoms until the cervical neoplasm has developed into an advanced stage. Therefore, it is quite useful and effective if the cancer is detected in the early time. With the successful diagnose of the cervical cancer, surgeries are able to dramatically reduce the incidences and mortality of cervical cancer. The vaccines of cervical cancer are effective only before the infection happens. The widespread cervical screening by the Papanicolaou test every 3-5 years with fellow-up can reduce the number of the deaths of cervical cancer by 80%. However, even in the developed countries where cervical screening programs have reduced the incidences and mortality from cervical cancer, continued measurements are important to prevent the reappearance of the disease (Miller, 1992).

Using organized cancer screening device with fluorescence spectroscopy which our team has been working on has shown promise for early detection of cancer. Such fluorescence spectroscopy device illuminates the tissue at different excitation and emission wavelengths. Then, experienced person will record the corresponding intensity at a number of excitation and emission wavelengths. The outcomes of the device have an extraordinary number of features, such as 5400 features. Therefore, it is very difficult to use traditional statistical methods to make the quantitative analysis (Lee, et al., 2005).

1.2 Overview of my Research Project

One of the big issues of our device with fluorescence spectroscopy is the repeatability of measurements across devices and time (Lee, et al., 2005). For examples, the amount of sunlight in the room, the relative humidity, the temperature and the voltage may influence the results by the devices with fluorescence spectroscopy. Despite of all the sources of variability above, the probes and the standards of the device are also easy to change in the experiment.

My role in the research focuses on the analyses of sunlight in the room which is measured in LUX. LUX is defined as the strength of the sunlight in the room. In the experiment, the data is directly obtained from a "zoey" instead of real patients. A "zoey" is a hand-made model which can function like a real cervical. After obtaining the data from "zoey", we are expected to use statistical methods to analyze the data. However, the conditions for traditional methods usually fail for functional data setting. In this thesis, we report the design of the experiment, analyses of summary, measurements of unprocessed data, and significance of the factor of room light in explaining variability (Lee, 2005).

Another aim of my project is to introduce the Adaptively Truncated Hotelling T-Square Test (Lee, Cox, & Follen, 2010) and conduct the simulations of the Adaptively Truncated MANOVA. As we have mentioned in the previous paragraph, it is not possible to use traditional methods of multivariate analysis (Rencher, 2002). We need to reduce the dimensions of the original data setting. At first, we select some "important" dimensions from the original data setting. After

leaving out the “unimportant” dimensions from the original data setting, we can apply traditional statistical methods including MANOVA and PCA (Principle Components Analyses). Secondly, we introduce two new test statistics which may solve the high-dimensional data. The first one is called Adaptively Truncated Hotelling T-Square Test. The second one is the Adaptively Truncated MAOVA Test which shares the similar thought of the Adaptively Truncated Hotelling T-square Test. The difference is the number of the groups of the data setting.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 illustrates the description of our new device, the details about the design of the experiment, and an introduction of the statistical methods that are used in the analyses. Chapter 3 contains two major parts. The first part is the results of the traditional statistical methods after reducing the dimensions of the original data. The results include the Exploratory Data Analyses, classical MANOVA, and Principle Components Analyses. The second part is the introduction of Adaptively Truncated Hotelling T-square and Adaptively Truncated MANOVA. Chapter 4 is the conclusion of the whole thesis. References are after Chapter 4.

Chapter 2

MATERIALS AND METHODS

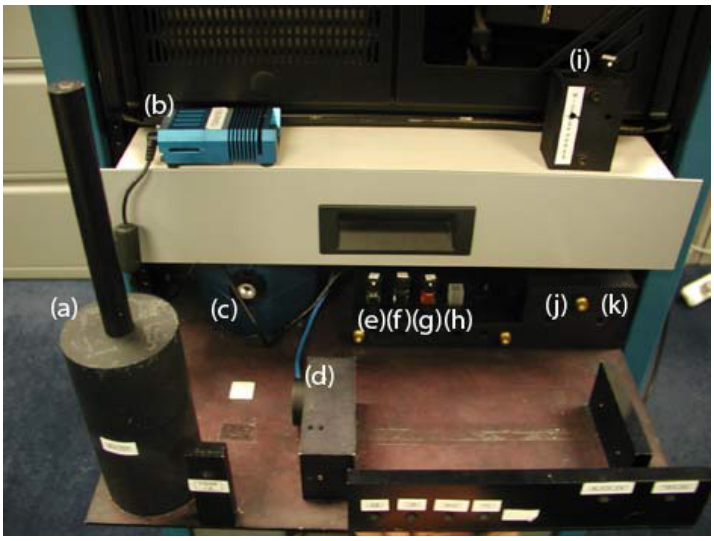
2.1 Description of Device

Fluorescence spectroscopy is used to putting light into the tissue and collecting different wavelengths that are known as the emission wavelength of fluorescence coming back to the devices. The excitation light is filtered so that the excitation wavelengths only change from 300 nm to 530 nm in increments of 10 nm. The range of the emission wavelengths varies with the number of excitation wavelengths. The intensity of fluoresced light is recorded at each excitation-emission wavelength combination. The storage for the intensity of Fluoresced light is called excitation/emission matrix (EEM). The measured fluorescence intensity illustrates absorption, particularly from hemoglobin (Yamal, et al., 2012).

In our experiment, the spectroscopy device is named the Fast Excitation Emission Matrix (FastEEM) system. There exist two kinds of FastEEM devices which are FastEEM2 in Houston,Texas, and FastEEM3 in Vancouver BC, Canada. The FastEEM system works with a xenon arc lamp coupled to a filter wheel to provide excitation light, coupled to a fiberoptic probe to intimate the excitation light to the tissue and the emission light back to a second filter wheel (Pikkula, Shuhatovich,

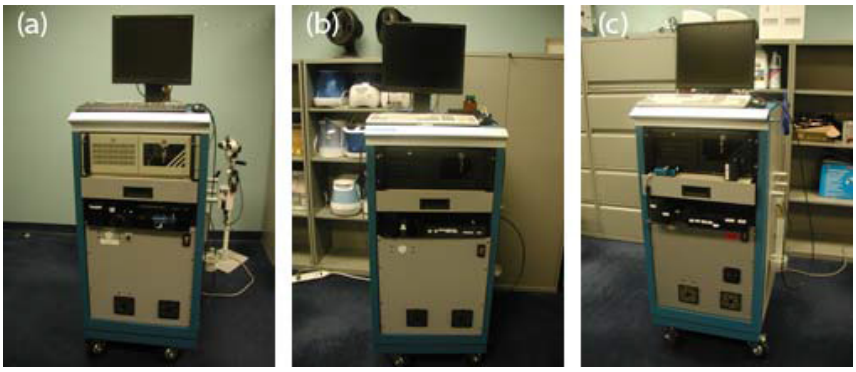
Price, & Serachitopol, 2007). The emission light passes from the second wheel to an imaging cooled charge-coupled device (CCD) camera to record fluorescence intensity as a function of emission wavelength. All components of the FastEEM system and the measurement procedures are controlled by the computer built into the instrument (Lee, et al., 2007).

Except for the FastEEM system, the device has the standard tray which contains different fluorescence and reflectance standards that are integrated into the system and used throughout the experiment. Positive fluorescence standards include Exalite, coumarin, and rhodamine. Negative fluorescence standards are deionized ultra filtered water and the frosted quartz cuvette. Positive reflectance standards are microspheres, Teflon, and an integrating sphere of 99% reflective Spectralon. There is only one negative reflectance standard which is a black 2% reflective Spectralon sample. The transfer buttons of different standards are shown in Figure 2.1 (Cox & Lee, 2008). For the purpose of our experiment, we only study on the standards of frosted cuvette and rhodamine. Meanwhile, three kinds of FastEEM system, MDA FastEEM box, LBJ FastEEM box and MDACC FastEEM box, are used. Different FasteEEM boxes are shown in Figure 2.2. For more specific information of these devices or the data processing, see Lee et al.



Note: water (a), Tungsten lamp (b), mercury lamp and integrating sphere (c), power meter (d), exalite (e), coumarin (f), rhodamine (g), frosted cuvette (h), microspheres, shown in the microspheres holder (i), black 2% spectralon (j), and Teflon (k).

Figure 2.1: Full Set of Standards



Note: (a) BCCA FastEEM box. (b) LBJ FastEEM box. (c) MDACC FastEEM box.

Figure 2.2: Three Different EEM Boxes

2.2 The Design of Experiment

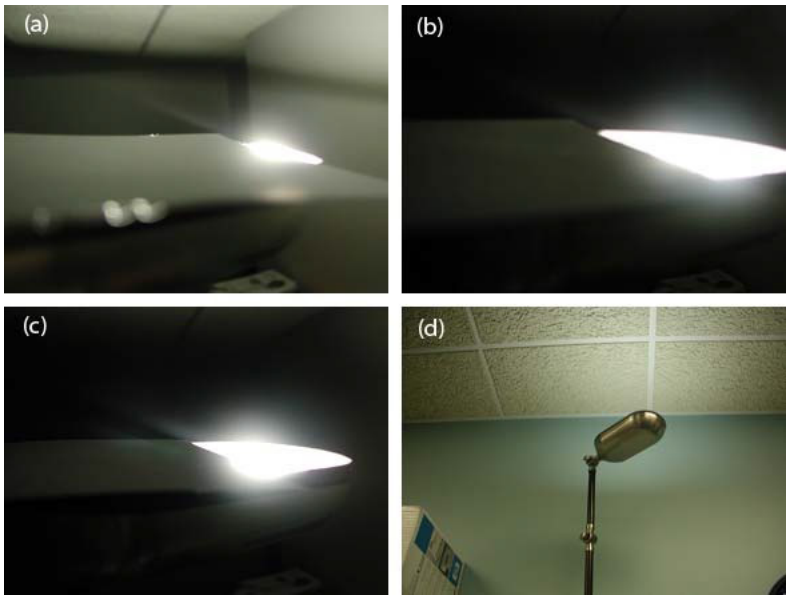
The objective of this experiment is to determine the effects, if any, of room light from a sunlight lamp on FastEEM measurements of rhodamine and frosted cuvette. As mentioned above, the BCCA, LBJ and MDACC FastEEM boxes with full set of standards are used in this experiment. Besides the boxes and standards, a sunlight lamp, a piece of plastic to use as lamp shutter and an environment meter which is used to record the amount of LUX in the room are also hardware in the experiment (Lee and Cox, 2008).

It is necessary to indicate the processes of the experiment as follows: (1) turning FastEEM boxes on and allow one hour for warm up before beginning the experiment; (2) measuring a full set of standards every two hours; (3) turning all room lights off, recording luminosity at a central point in the room and at the FastEEM box using the environment meter; (4) placing dry probe on the rhodamine standard in the standards tray, acquiring five consecutive measurements; (5) repeating the fore steps to get a total of fifteen measurements; (6) exchanging the rhodamine standards for frosted cuvette standards and repeating procedures (4) and (5); (7) consecutively changing the room light to low light, medium low light, medium high light and high light using the piece of plastic and repeating procedures (3)-(6) in each situation. The total number of observations for each FastEEM box combined with different standard is 85. Because we have three FastEEM boxes and two different standards in the experiment, there are 510 observations in total. The Table 2.1 provides more information about the processes of the experiment. The sunlight lamp is used here

because of its ability to replicate the environmental conditions. Figure 2.3 shows more details about how the sunlight lamp is used in each group.

Table 2.1: The Design Table of Experiment

Light in Center of the room (In Lux)	Light at the Box	MDA Rho	MDA Fc	Light at the Box	BCCA Rho	BCCA Fc	Light at the Box	LBJ Rho	LBJ Fc
(No light) 0.22	0.01	1-25	1-25	0.01	1-25	1-25	0.01	1-25	1-25
7.14	0.04	26-40	26-40	0.08	26-40	26-40	0.08	26-40	26-40
10.48	0.08	41-55	41-55	0.12	41-55	41-55	0.12	41-55	41-55
18.39	0.13	56-70	56-70	0.11	56-70	56-70	0.14	56-70	56-70
(All Lights) 847	4.36	71-85	71-85	2.89	71-85	71-85	3.15	71-85	71-85



Note: (a) Low light (Setting 1). (b) Medium low light (Setting 2). (c) Medium high light (Setting 3). (d) High light (Setting 4). Not shown: all lights off (Setting 0). (a), (b), and (c) are shown from a top-down perspective.

Figure 2.3: Sunlight Lamp

2.3 Statistical Methods

Obviously, the data contains a large number of dimensions. It is not proper to use traditional multivariate analysis methods with the data setting from the experiment. Because traditional multivariate analysis methods including MANOVA and Hotelling T-Square Test require the number of dimensions to be much less than the number of observations, we need to reduce the dimension of the original data

setting until the assumption of traditional multivariate analysis methods satisfies (Rencher, 2002).

One of the statistical methods is to use the data with the emission wavelength fixed at a specified value. Meanwhile, the data with the specified emission wavelength should have the largest variance, which means it separates the groups mostly. Because we only have six distinct excitation wavelengths in our experiment, we are able to reduce the dimension to six. Therefore, traditional multivariate analysis methods are able to use. Even though it seems that we may lose much information of the original data setting, it is still a remarkable try under the technical methods of dimension reductions.

Another statistical method is the Principal Components Analysis. Principal Components Analysis is used to reducing the dimensions of the data, which satisfies our objectives. Principle Components Analysis selects some linear combinations of the depending variables. Each linear combination is called a component. We can choose the number of components depending on the total variation percentage. Having decided the number of components, we can transfer the original data setting to a new one using the related eigenvectors of the selected principal components. As a result, we could make quantitative analysis on the new data with lower dimensions sharing the same results of the original data.

Unlike the previous statistical methods, the Principal Components Analysis utilizes all the information of the original data. Also, it transfers the data using linear combinations for dimension reductions. Even though the data with

emission wavelength fixed at a specified value has lower dimensions, it may not provide sufficient information of the original data. In another word, the results of the premier method may not be stationary if some extreme points exist.

2.4 Chapter Summary

This chapter emphasizes the description of the device in the experiment, the design of the experiment and how traditional statistical methods are able to be applied. The first two parts are the backgrounds of the experiment. The last part of this chapter introduces the primary thoughts of dimension reduction techniques. Even though the PCA method transfers the data from higher dimension to a lower dimension without losing much information, the methods with the fixed wavelength of the emission may take the most useful parts of the original data. In a summary, this chapter introduces both the backgrounds of the experiment and the traditional statistical methods of the adjusted dataset using remarkable dimension reduction techniques.

Chapter 3

RESULTS AND ANALYSES OF THE EXPERIMENT

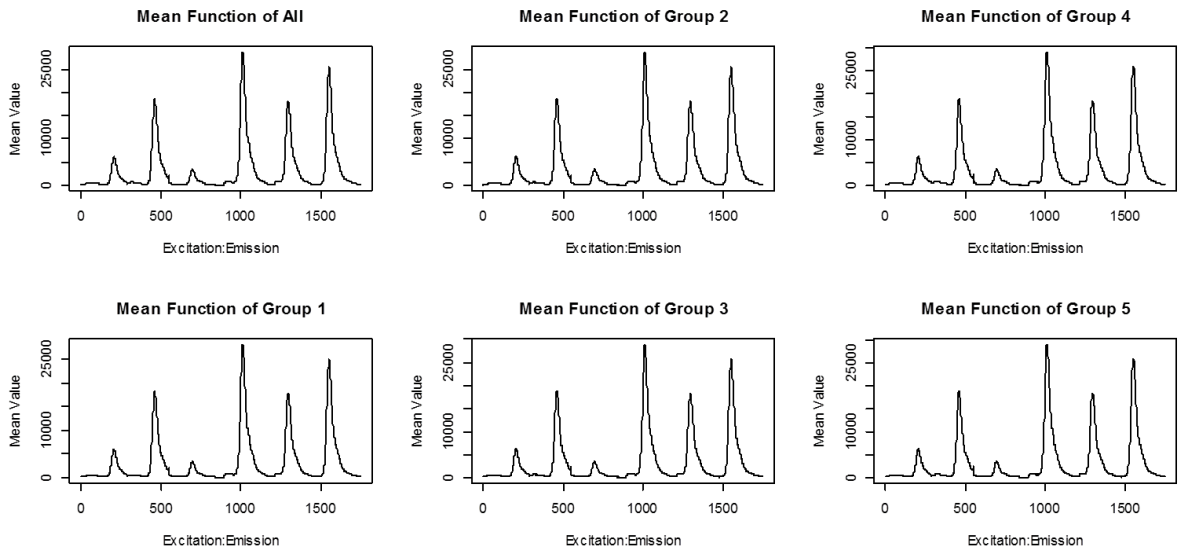
3.1 Exploratory Data Analysis

In our experiment, MDA, LBJ, and BCCA measurements combined with rhodamine and frosted cuvette standards are considered. For each of the measurement and standard combination, we have 5 different groups as mentioned above. The amount of sunlight in each group is different and measured by the number of LUX. The first group consists of 25 observations, while the other four groups have 15 observations each. It adds up to 85 observations for each measurement and standard combination.

An Exploratory Data Analysis (EDA) was performed on the initial data. For each of the observations, there are six distinct excitation wavelengths, ranging from 330nm to 480nm. Given the specified excitation wavelength, we have 292 different emission wavelengths which vary with the excitation wavelength. For instances, the observation, which has the value of the excitation wavelength equal to 330nm, consists of 292 different excitation wavelengths ranging from 375nm 660 nm. The one observation of the initial data contains 1752 distinct excitation-emission wavelength combinations in general. Figure 3.1 and Figure 3.2 show the values of

mean functions and standard deviation functions of the data with MDA FastEEM box and Rhodamine standard. Figure 3.3 and Figure 3.4 show the values of mean functions and standard deviation functions with the data of MDA FastEEM box and Frosted Cuvette standard. Other treatments share the same results so that we do not list them out.

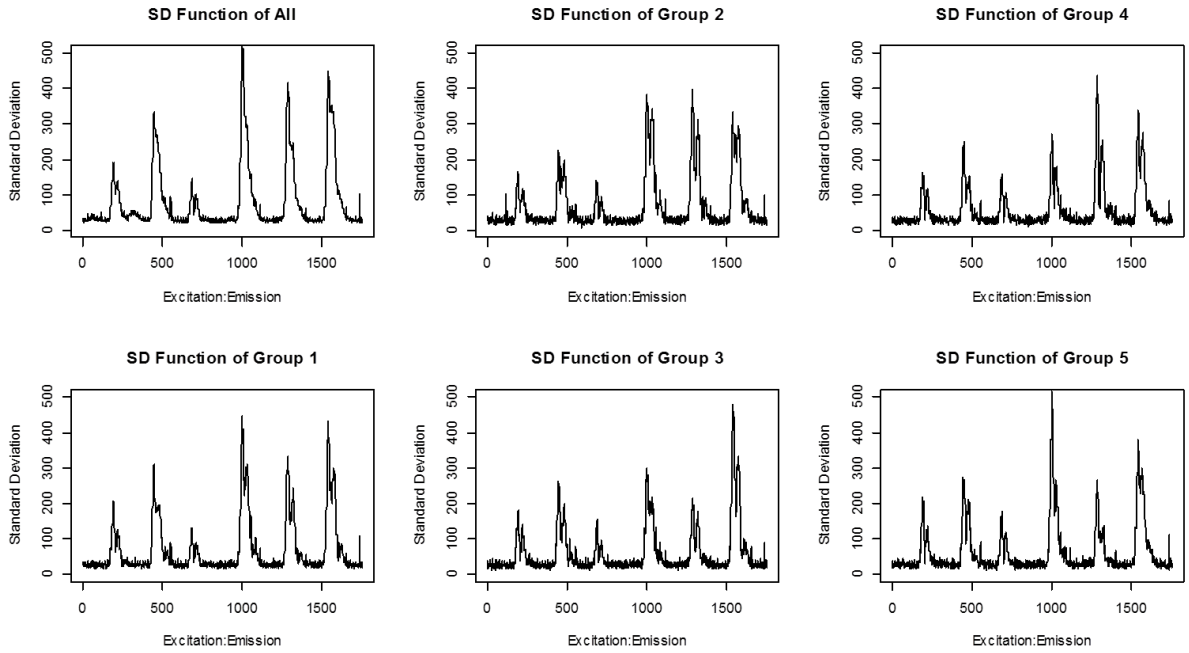
Mean



Note: Horizontal axis – combinations of excitation and emission, Vertical axis – value of the intensity

Figure 3.1: Mean Functions of MDA and Rhodamine

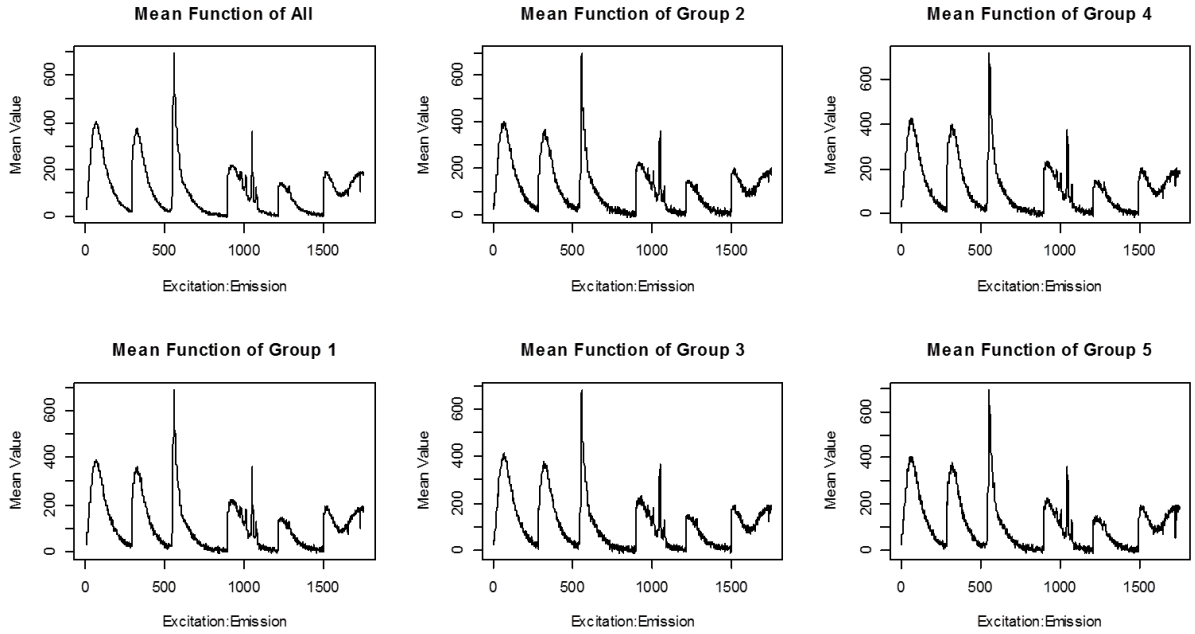
Standard Deviation



Note: Horizontal axis – combinations of excitation and emission, Vertical axis – value of the intensity

Figure 3.2: SD Functions of MDA and Rhodamine

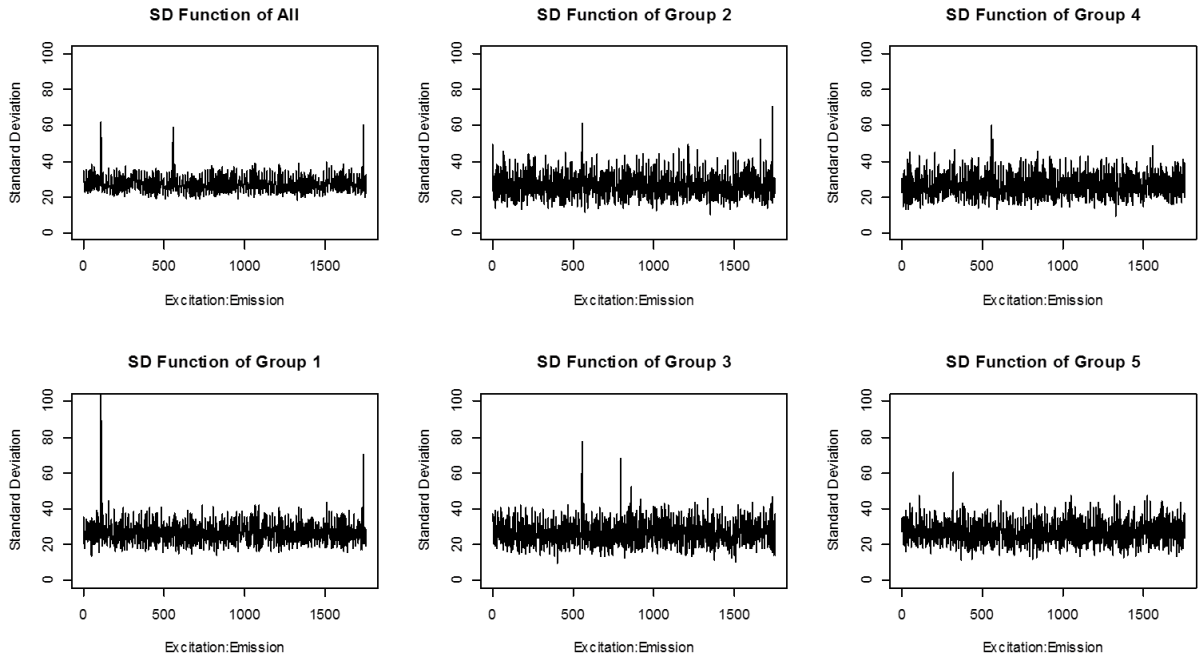
Mean



Note: Horizontal axis – combinations of excitation and emission, Vertical axis – value of the intensity

Figure 3.3: Mean Functions of MDA and Frosted Cuvette

Standard Deviation



Note: Horizontal axis – combinations of excitation and emission, Vertical axis – value of the intensity

Figure 3.4: SD Functions of MDA and Frosted Cuvette

In all the figures, the horizontal axis stands for the indications of excitation-emission combination. Even though there may be differences among groups, the mean functions shown in Figure 3.1 and Figure 3.3 are hard to distinguish with each other. The standard deviation functions are not all the same among groups. However, the differences are not significant, and it is also not sufficient to make statistical inferences using the standard deviation functions in case of covariance. Although all the figures do not show any statistical differences, it is clear that all the

mean functions have six peak values. The data which belongs to the frosted cuvette standard should have less variance because the standard deviation functions do not change dramatically among groups. The results of the other two FastEEM Boxes which are LBJ and BCCA are not shown here because the same results are as MDA FastEEM box.

Another way of Exploratory Data Analysis is to present the data graphically, which is the EEM plot of the means of each group. EEM plot is a colorific picture where different colors indicate different values of intensities. It is more visual using EEM plots because we divide the wavelength of emission and excitation into each axis. We draw the mean functions of each group of the initial data using EEM plots as shown in Figure 3.5-3.10. Like the results of the mean functions in EDA, we are not able to distinguish the groups using the EEM plots. However, we could clearly find out that the variations of groups may exist in a narrow band of wavelengths of the emission. If we reduce the dimension of the data by selecting the specified wavelength of the emission, most of the statistical methods are able to use because the number of dimension is much less than the number of the total observations. This is the primary idea of dimension reduction techniques.

EEM Plots

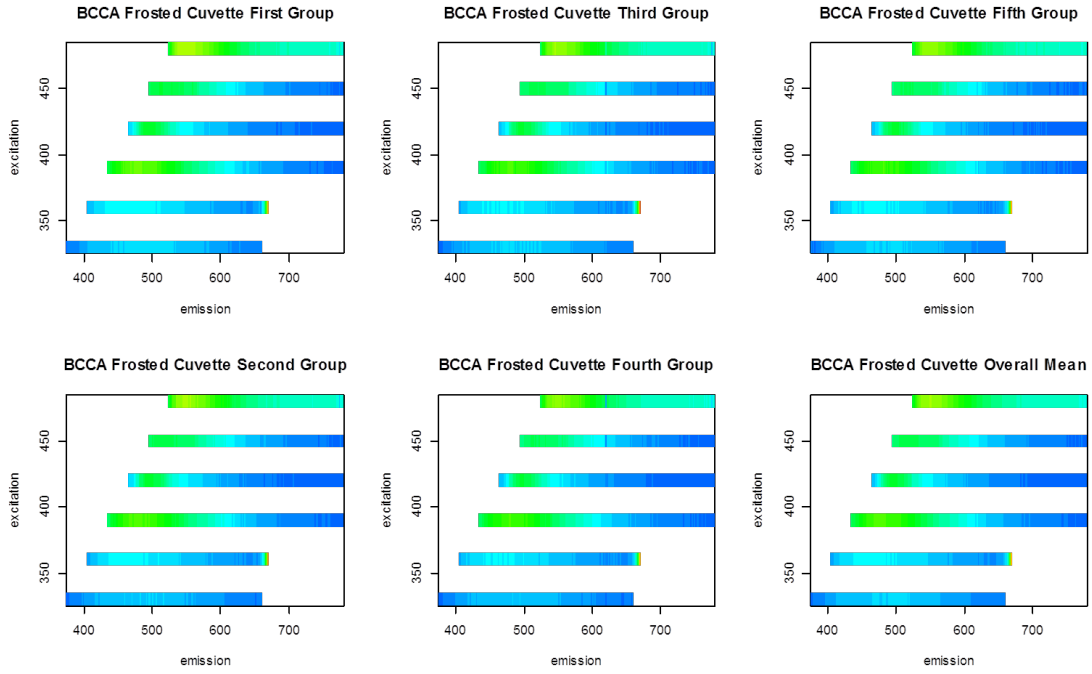


Figure 3.5: EEM Plots of BCCA and Frosted Cuvette

EEM Plots

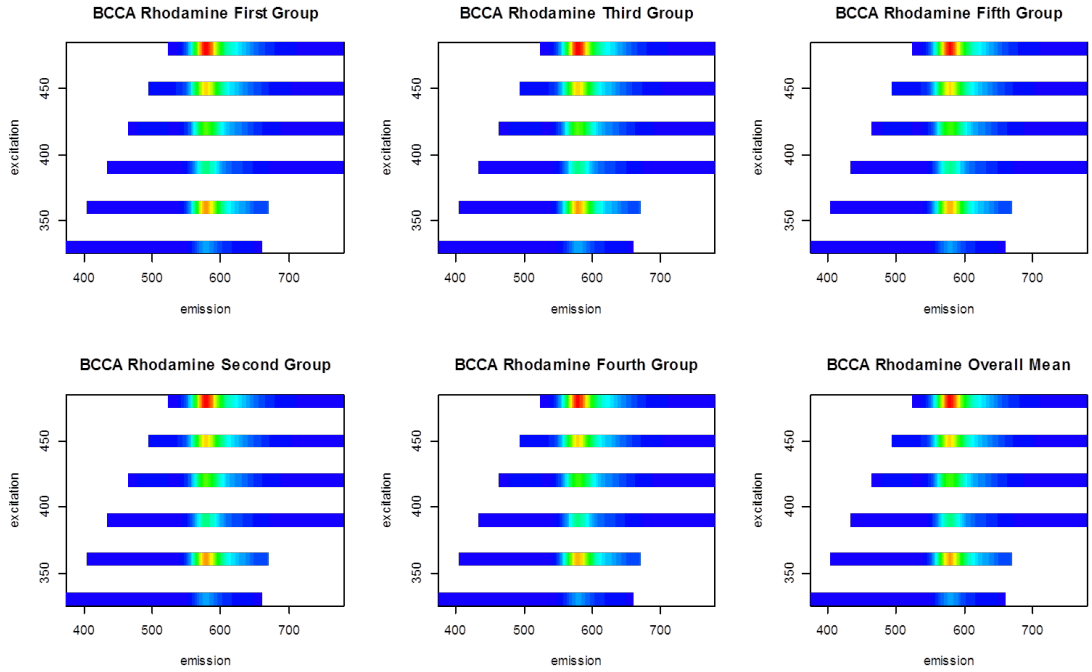


Figure 3.6: EEM Plots of BCCA and Rhodamine

EEM Plots

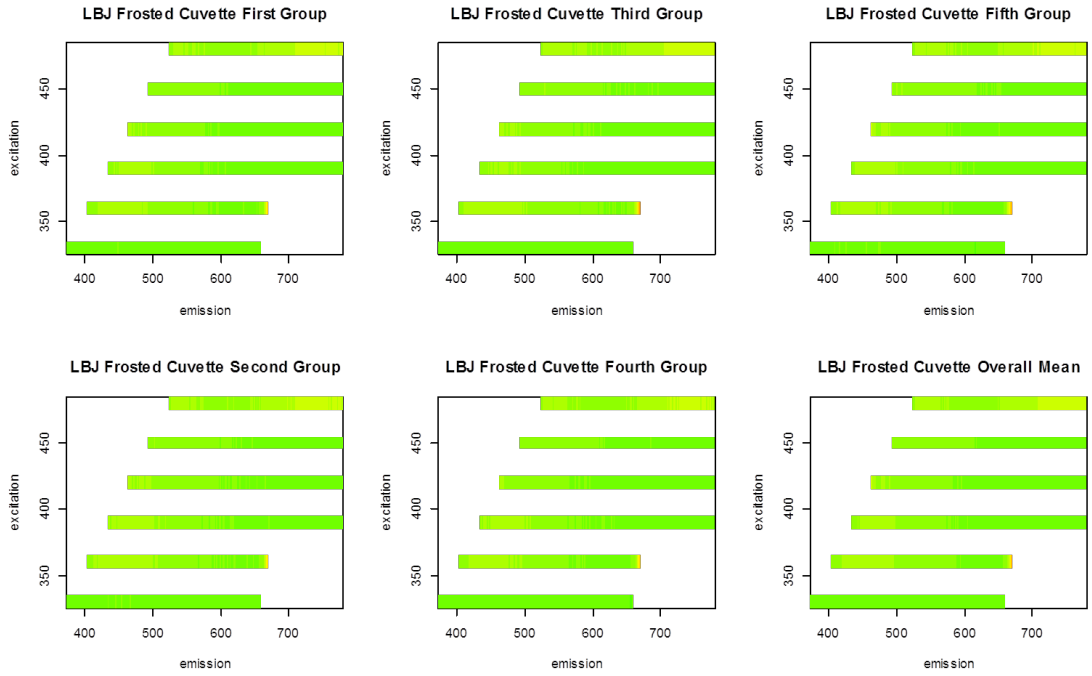


Figure 3.7: EEM Plots of LBJ and Frosted Cuvette

EEM Plot

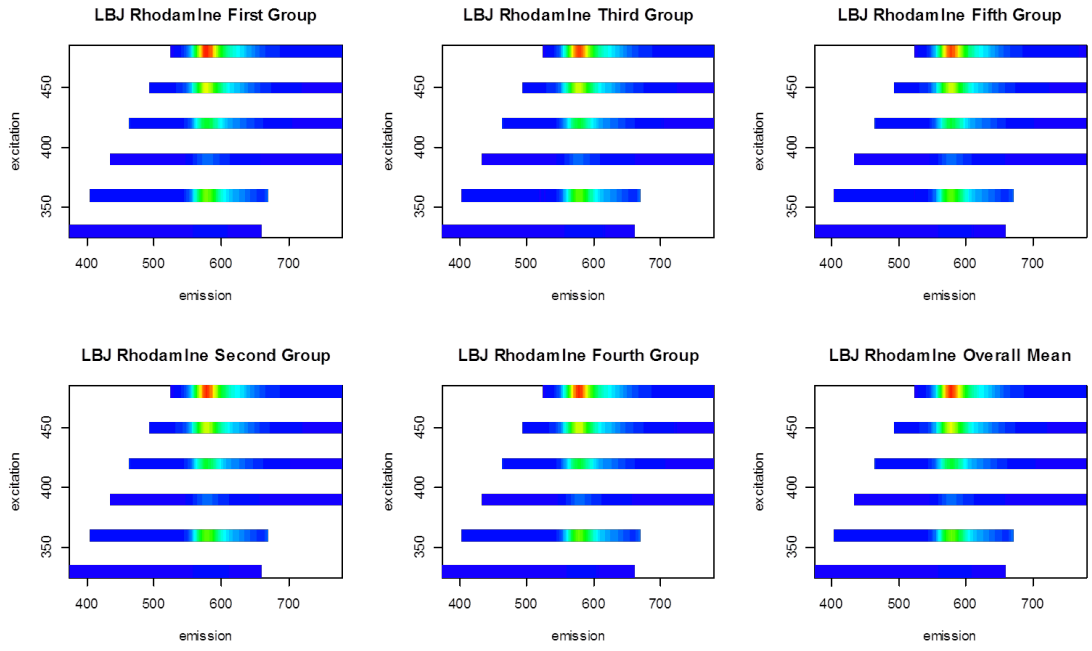


Figure 3.8: EEM Plots of LBJ and Rhodamine

EEM Plots

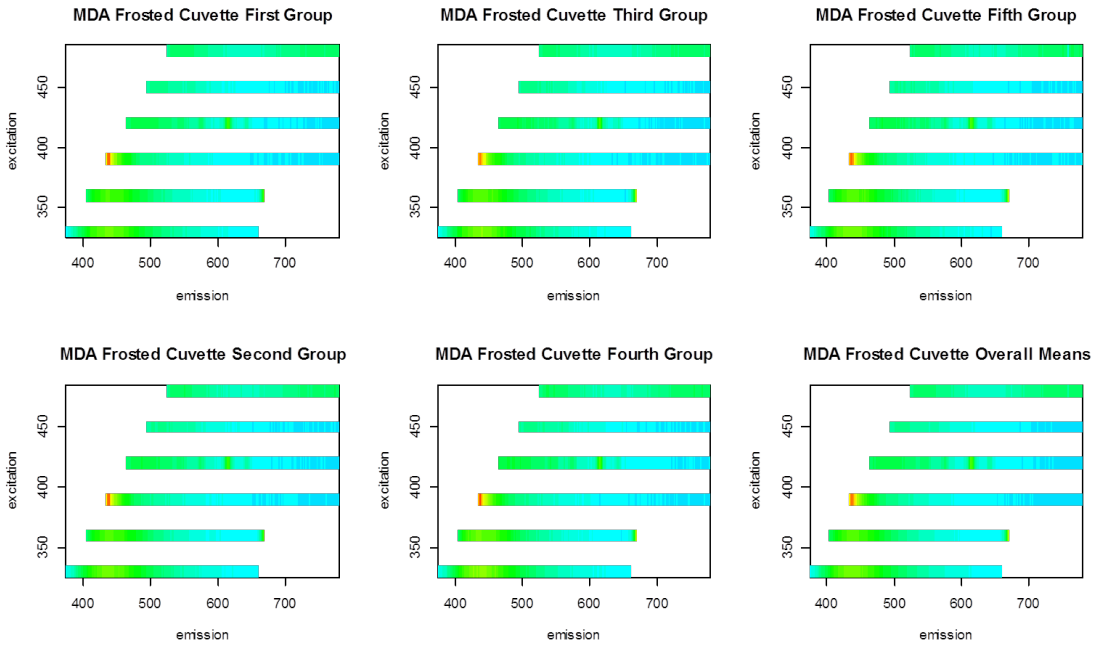


Figure 3.9: EEM Plots of MDA and Frosted Cuvette

EEM Plots

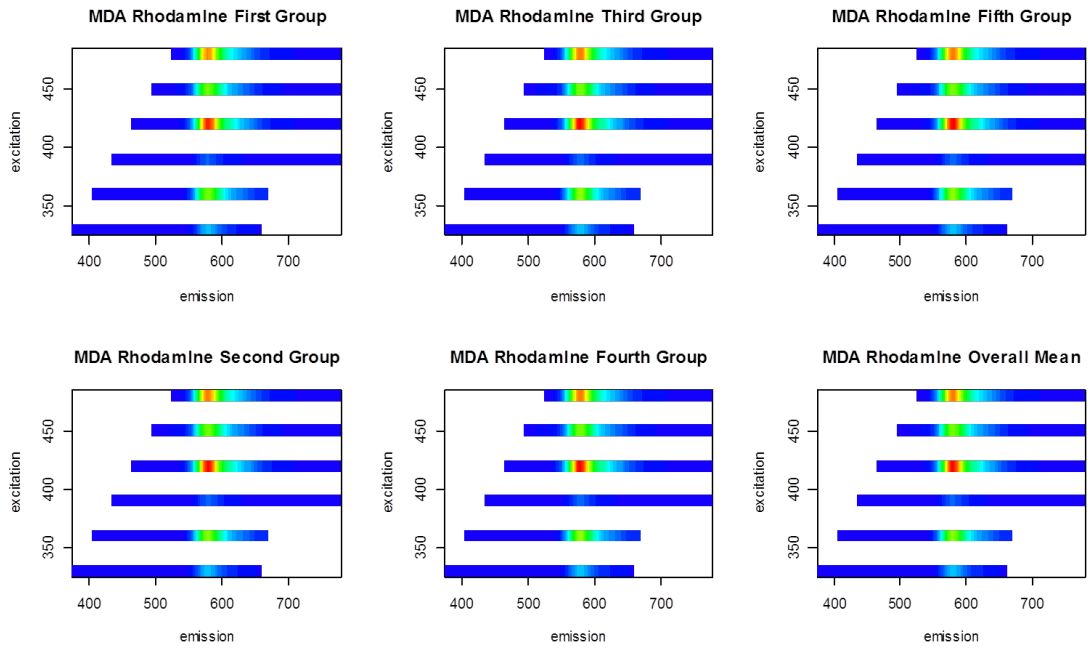


Figure 3.10: EEM Plots of MDA and Rhodamine

3.2 Result of Classical MANOVA

We often measure several dependent variables on each experimental unit instead of just one variable. In the case of one variable, we use ANOVA table as a statistical method for indicating the statistical significance of factors. Similar with ANOVA, MANOVA is a test with the null hypothesis that the mean vectors from more than two multivariate normal distributions are equal. For more information about MANOVA, see Chapter 3.5. Depending on different statistics of MANOVA which are

wilks, Roy, Lawley-Hotelling and Pillai, the results of MANOVA may vary. In our experiment, only one-way MANOVA is considered. What is more important, the traditional MANOVA assumes that the number of total observations is much greater than the number of dimensions in order to assure the correlation matrix be nonsingular. Because the initial data has only 85 observations in total and a dimension of 1752, we have to reduce the dimension for MANOVA. A useful and remarkable technique for dimension reduction is to retain the data with the six peak values from the mean functions as shown in Figure 3.1 and Figure 3.3. This six peak values share the same value of emission wavelength being 580nm with the data of rhodamine standard. For the data of frosted cuvette standard, 580nm is not a satisfied statistical truncated point. We suggest using six peak values of mean functions shown in Figure 3.3 for technical dimension reductions.

After reducing the dimension of the initial data to six, we could get the results of MANOVA in Table 3.1. The first horizontal line indicates the six measurement and standard combinations. The first vertical line stands for four different statistics and the max values of them. We could infer that the room light factor is not statistically significant at the five percent level for the data of MDA FastEEM box and LBJ FastEEM box combined with frosted cuvette standards (In general, the Roy's statistic is not recommended in any situation except the collinear case under standard assumptions). However, the room light factor for the other measurement and standard combinations is statistically significant at the five percent level. The fact that the standard deviation functions do not change a lot among groups

with frosted cuvette standard may be an explanation for the non-significance of the room light factor. Also, discriminant functions, canonical correlation vectors and the eigenvalues of the covariance matrix are provided in Table 3.2. Although we do not use any of the results in Table 3.2 in the article, it is useful for further analyses.

Table 3.1: Results of MANOVA

P-Value	MDA_rho	MDA_fc	BCCA_rho	BCCA_fc	LBJ_rho	LBJ_fc
Wilks	3.914e-17	0.295	1.134e-18	2.393e-5	1.348e-8	0.868
Roy	5.525e-24	0.020	3.838e-23	2.730e-7	1.230e-11	0.139
H-L	2.170e-30	0.301	4.078e-30	6.066e-6	7.855e-11	0.875
Pillai	4.407e-9	0.288	3.155e-11	8.223e-5	8.692e-7	0.862
Max P-value	4.407e-9	0.301	3.155e-11	8.223e-5	8.692e-7	0.875

Table 3.2: Related Statistics

Order	Eigenvalue	Dscriminant Function	Canonical Correlation
1	3.63	44.51	0.89
2	0.11	66.26	0.311
3	0.06	-21.26	0.23
4	0.01	-60.46	0.11
5	7.51e-16	-92.68	2.74e-8
6	7.51e-16	170.98	2.74e-8

3.3 Principal Components Analysis

Although MANOVA with the data of the emission wavelength being 580nm successfully indicates the significance of room light factor, it may not provide sufficient information we need. In another word, most parts of the initial data are removed. The best way for dimension reduction is to use the Principal Components Analysis (PCA). Each component is a linear combination of the depending variables. PCA is usually used to reducing the number of dimensions because few components take over large proportion of the total variance. There are several processes in the PCA. We begin the PCA by deciding the number of principal components. In this article, we assume that 85 percent of the total variance is considered. The results of the first step of PCA are shown in Figure 3.11. The purple line stands for the 85 percent cumulative variance proportion with the related number of components selected in the horizontal axis. For examples, for the data of LBJ FastEEM Box and Rhodamine standard, the first eight principal components account 85 percent of the total variance.

The Proportion of Variance

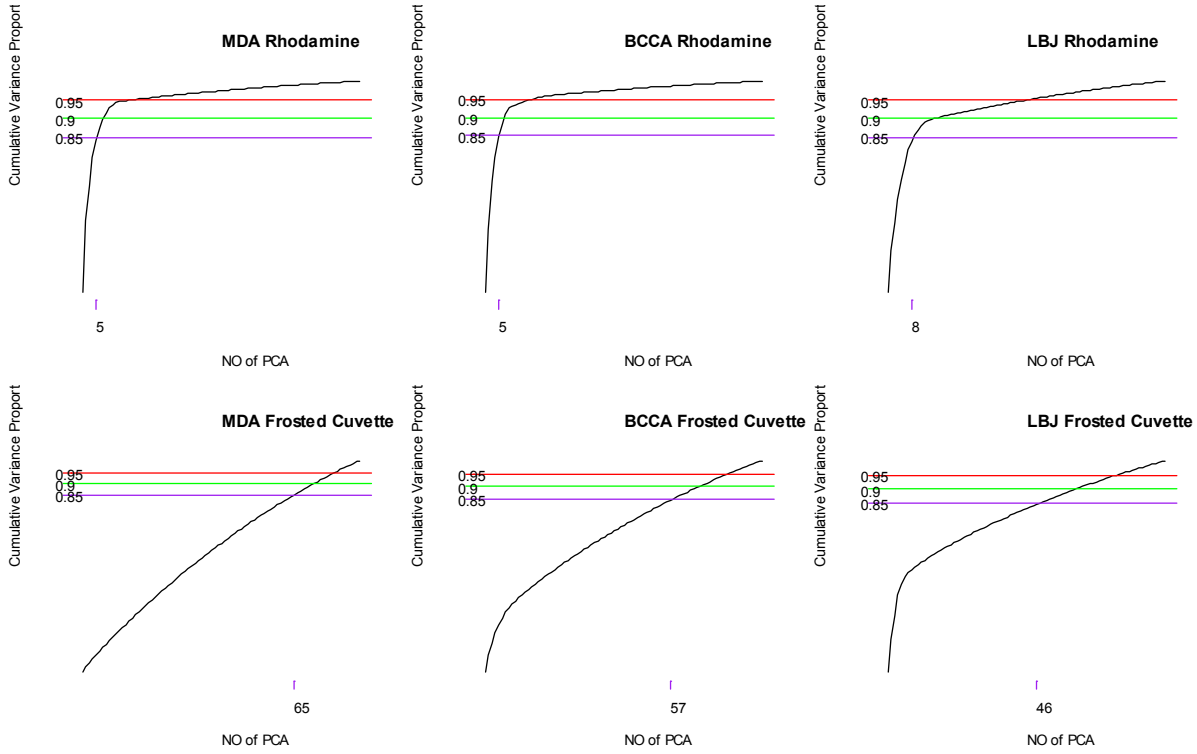


Figure 3.11: Cumulative Variance Proportion of PCA

Once we have decided the number of principal components that are sufficient for our analysis, we can transfer the initial data to a new one using the vectors obtained from the principal components. For examples, we have selected the first eight components from the data of LBJ measurement and Rhodamine standard. Each component is a linear combination of 1751 excitation-emission variables. The weight of each excitation-emission variable is the related number in the eigenvectors obtained from the correlation matrix. As a result, we successfully reduce the

dimension of the data from 1751 to 8 by using the first eight principal components. Unlike the rhodamine standards, the dimension reduction is not extreme obvious with the data of the frosted cuvette standards. The results are shown in Table 3.3.

Table 3.3: The Results of PCA

P-value	MDA_rho	MDA_fc	BCCA_rho	BCCA_fc	LBJ_rho	LBJ_fc
NO of PCA	5	65	5	57	8	46
Wilks	1.298e-32	3.322e-6	6.559e-10	2.383e-17	1.915e-5	0.009
Roy	4.642e-40	1.023e-7	2.506e-13	1.176e-13	2.038e-8	2.203e-5
H-L	5.136e-71	6.218e-8	5.913e-13	8.401e-23	1.599e-6	0.003
Pillai	4.918e-13	0.001e-1	1.543e-7	2.448e-11	0.001e-1	0.017
Max P-value	4.918e-13	0.001e-1	1.543e-7	2.448e-11	0.001e-1	0.017

Assuming that the results of the test do no change between the initial data and the transferred data by PCA, the room light factor is significant at the five percent level for all measurement and standard combinations. By PCA, we successfully reduce the dimension of the data and get the results as expected without removing necessary information of the initial data.

3.4 Adaptively Truncated Hotelling T-Square Test

Jong Soo Lee, Dennis D. Cox and Michele Follen (2010) first came up with the idea of Adaptively Truncated Hotelling T-Square Test (ATHTST). ATHTST is a new adjusted method based on the traditional Hotelling T-Squre Test. Their same goal is to test equality of mean functions from two samples of multivariate normal distributed data. The traditional Hotelling T-Square Test are conducted under the condition that the number of dimensions of the data is much less than the number of total observations. If the condition fails, we are not able to obtain the values of test statistics from the traditional Hotelling T-Square Test. The reason is that the variance-covariance matrix is singular. In the real world, especially for the functional data, the condition usually fails, which means the total number of observations is much less than the number of dimensions. However, ATHTST may be a solution if the condition fails.

Instead of calculating the inverse of the matrix included in the test statistics of the traditional Hotelling T-Square Test, ATHTST takes the decomposition of the matrix and adaptively selects the maximum value from functions of eigenvectors. As a result, we call the maximum value our new test statistics for ATHTST. Unlike the traditional test statistics of Hotelling T-Square Test which has an obvious distribution, the new test statistics does not have a common distribution. Under the enlarged null hypothesis that the distributions of the two populations are the same, randomization methods which we also call permutation are proposed to find a

null distribution. Permutation gives accurate significance levels irrespective of the common distribution under the null distribution.

The functions of eigenvectors and eigenvalues which are used in the test statistic of ATHTST are shown below:

$$T_k^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{k} \hat{\lambda}_j^{-1} ((\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \hat{\mathbf{v}}_j)^2 \qquad (3.1)$$

The equation 3.1 contains the jth eigenvector $\hat{\mathbf{v}}_j$, the jth eigenvalue $\hat{\lambda}_j$, the number of observations in the first group n_1, the number of observations in the second group n_2, the mean vector of the first group $\bar{\mathbf{y}}_1$, and the mean vector of the second group $\bar{\mathbf{y}}_2$. The number of k is decided by criteria. Then, we obtain the maximum value of T_k^2 with certain transformation in case of k as our test statistics.

In the article of Jong Soo Lee, Dennis D. Cox and Michele Follen (2010), they came up with two different transformations. Then, they compared it with the other five test statistics which are Wald-Wolfowitz version of ATHTST (Wald & Wolfowitz, 1944), Max-T statistic (Taylor, Worskey, & Gosselin, 2007), Fan and Lin's statistic (Fan & Lin, 1998) and Shen and Faraway's statistic (Shen & Faraway, 2004). What is more important, they performed simulations using the permutation methods. The simulations require the assumption of the alternative hypothesis. They tested six different alternative hypotheses. For each alternative hypothesis, the first group contains the data from multivariate normal distribution with mean vector being

zero. The second group adds a function to the mean vector. The six different functions are shown in Figure 3.12.

Functions

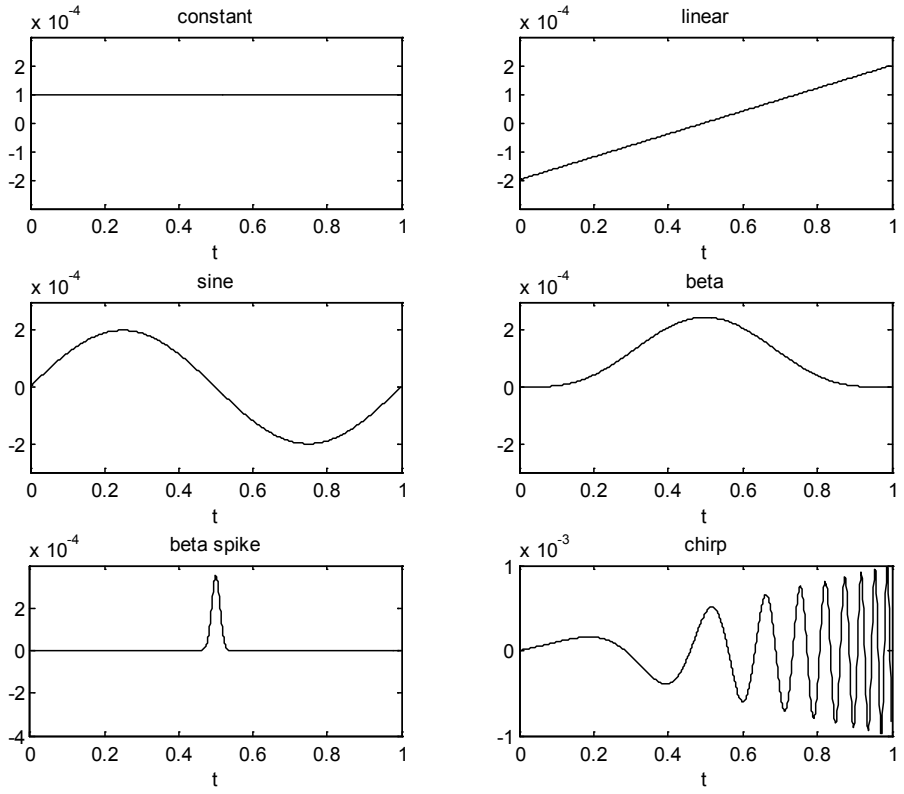


Figure 3.12: Plot of All Six Alternatives

The results of p-values obtained from 10,000 permutation iterations of the 7 tests and 6 alternatives are given in the Table 3.4. It is obvious that Wald-Wolfowitz modification is very computationally efficient compared with the other

tests. We could conclude that most method do well in test A through D except for Fan and Lin's statistic. The p-value of Shen and Faraway' statistic exceeds 0.05 on test E and F.

Table 3.4: P-values Obtained From 10,000 Permutation Iterations

Test	A	B	C	D	E	F	Time
1	0.0023	0.0509	0.0260	0.0019	0	0	35,589
2	0.0018	0.0532	0.0514	0.0006	0	0	25589
3	0.0024	0.0273	0.0193	0.0028	0	0	290
4	0.0002	0.0414	0.0468	0.0003	0	0	290
5	0.0076	0.0184	0.1014	0.0022	0	0.0061	766
6	0.1833	0.3142	0.3216	0.2104	0	0	771
7	0.0001	0.0091	0.0103	0	0.2683	0.1361	452

Note: Test as rows: 1: ATHTST F-transform 2:ATHTST X-transform 3:Wald-Wolfowitz of ATHTST F-transform 4: Wald-Wolfowitz of ATHTST X-transform 5:Max-T statistic 6:Fan and Lin's statistic 7:Shen and Faraway's Statistic. Alternatives as columns: A:constant B:linear C:sine D:beta E:beta spike F:chirp. The last column lists the CPU time for each test in seconds.

3.5 Adaptively Truncated MANOVA Test

Based on the thought of ATHTST (Lee, Cox, & Follen, 2010), we would like to make some adjustments on the traditional Multivariate Analysis of Variance (MANOVA). At first, we should have some concepts about traditional MANOVA. We assume that k independent random samples of size n are obtained from p-variable normal populations with equal covariance matrix. The equation for each observation vector is

$$\mathbf{y}_{ij} = \mathbf{u} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} \quad \text{i=1,2,....k; j=1,2,....n.} \tag{3.2}$$

We would like to show the two important matrices as below. One is called the "between" matrix **H**, the other is called the "within" matrix **E**. $\bar{\mathbf{y}}_{i.}$ means the sample mean vector of the ith group. $\bar{\mathbf{y}}_{..}$ means the sample mean vector of all groups and n is the total number of observations.

$$\mathbf{H} = n\sum_{i=1}^{k}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})'$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})'$$

For traditional MANOVA, we need to calculate the eigenvalues of the Matrix $\mathbf{E}^{-1}\mathbf{H}$. Once we have successfully obtained the eigenvalues from the Matrix $\mathbf{E}^{-1}\mathbf{H}$, there are four test statistics based on the eigenvalues which are called Wilks, Roy, Lawley-Hotelling and Pillai. Each test statistic has an approximated distribution and is able to transform to common F or χ^2 distribution. However, the matrix **E** is not

inverse able if the number of observations is much less than the number of dimensions. In another word, the matrix **E** is singular. By the functional data setting, like the data from our experiment, the number of dimensions is usually much larger than the number of observations, and the assumption of equal covariance matrix may not always hold. We try to develop a new test statistic based on the thoughts of ATHTST.

The goal of our modifications on MANOVA is to obtain approximated eigenvalues of the matrix $\mathbf{E}^{-1}\mathbf{H}$ under the condition that the matrix **E** is singular. At first, we would like to calculate the eigenvalues of the matrix **E** and **H** separately. Because the matrix **E** and **H** are symmetric, we have the following decomposition.

$$\mathbf{E} = \sum_{j=1}^{p} \lambda_j \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j'$$

$$\mathbf{H} = \sum_{j=1}^{p} \theta_j \hat{\mathbf{w}}_j \hat{\mathbf{w}}_j'$$

Because the number of p is quite larger, we have to retain the first k parts of them. The number of k is determined by a certain criteria. For example, we can choose k by $\lambda_k > 0.00001\lambda_1$ (the eigenvalues should be ordered with λ_1 being largest).

The first test statistic for Adaptively Truncated MANOVA is to divide the eigenvalues from two matrices. The equation 3.3 shows the k truncated values. We would like to use the negative mean of all the k truncated values as our test statistic T.

$$T_K = \sum_{i=1}^{k} \frac{\theta_i}{\lambda_i} \quad T = -\frac{T_1 + T_2 + \ldots + T_k}{k} \tag{3.3}$$

The second test statistic for Adaptively Truncated MANOVA is to ignore the eigenvalues of the matrix **H**. Even though it may lose information of the original dataset, it could free us more time for computation. When we perform simulations on these test statistics, we could find out that the second test statistic S costs much less computation time. The test statistic S is shown below.

$$S_K = \sum_{i=1}^{k} \frac{1}{\lambda_i} \quad S = -\frac{S_1 + S_2 + \ldots + S_k}{k} \tag{3.4}$$

The third test statistic P is more complex but accurate. The first one does not include the eigenvectors and assumes that the multiple of upper triangle matrix from the decomposition of two matrices is identity matrix. However, it should not always be the identity matrix. We would like to consider adding "weight" to the division of eigenvalues. The equation 3.5 and 3.6 outlines the test statistic P.

$$P_K = \sum_{i=1}^{k} \alpha_i \frac{\theta_i}{\lambda_i} \quad P = -\frac{P_1 + P_2 + \ldots + P_k}{k} \tag{3.5}$$

$$\alpha_i = < \hat{\mathbf{v}}_i, \widehat{\mathbf{w}}_i > \tag{3.6}$$

Just like what Jong Soo Lee, Dennis D. Cox and Michele Follen do in the ATHTST, we would like to perform simulations on the test statistics of Adaptively Truncated MANOVA. For convenience, we only conduct the simulations with the data of three groups. There exist three different circumstances of the groups: all the groups

are from same distribution; two groups are from the same distribution but the third group is not; none of the group is from the same distribution. Mathematically, we create the first population consists of independent Gaussian process with mean function being zero vector and a certain covariance. We mark the first population as "0". Then, we create other groups by adding a function to the mean vector of the first group but the same covariance. The functions are exactly the same functions used in the simulations of ATHTST except the last two. We list the functions by equations as shown in Table 3.5. As a result, we have 8 alternatives in total. In each group, we have 250 observations with 1000 dimensions.

Table 3.5: A List of 4 Functions

Alternative	Functions	Name
A	0.0001	Constant
B	0.0004(t-0.5)	Linear
C	$0.0002\sin(2\pi t)$	Sine
D	$0.0001\text{beta}_{5,5}(t)$	Beta

We would like to perform 10,000 permutation iterations on the three test statistics of Adaptively Truncated MANOVA in order to obtain p-values. In addition, we calculate the p-value when the null hypothesis is true taht all the three groups are from the same multivariate normal distribution. The results of the simulation are shown in Table3.6.

Table 3.6: P-values and CPU time

	Data of three Groups									Time
	000	00A	00B	00C	00D	ABC	ABD	BCD	ACD	
T	0.8488	0	0	0	0	0.0175	0.0107	0.0157	0.0166	5
S	0.1837	1	1	1	1	1	1	1	1	1
P	0.2609	0.0045	0.0148	0.0036	0.0210	0.0117	0	0.0127	0.0062	10

Note: T, S and P stands for the three test statistics. "00A" means the first and second group has same distribution "0", the third group comes from the distribution which adds mean function "A" on "0". Etc. Time means how long the system runs one iteration of the simulations in seconds.

From Table 3.6, we could conclude that test statistic S is a bad statistic because the p-value is constant at 1. Even though the CPU time is the lowest, the test statistic S is not appropriate for testing the equality of the mean vectors from more than two groups. Test statistic T reports that all the p-values are below five percent. However, it remains 0 for the data with the structure like "00A", "00B", "00C" or "00D". In another word, it is a bad statistic if the alternative contains a certain kind of structure. Generally, test statistic P is the best one that fits our goal of testing the equality of the mean vectors from more than two groups. No matter what the alternative is, the test statistic P shows different p-values which are all under five percent. The disadvantage of the test statistic P is that it costs much more CPU time than the other two statistics. It is mainly because that it has more computations than

the other two statistics. Finding new ways for decomposition of the matrices may reduce the CPU time. From what we have discussed above, we can safely draw the conclusion that the test statistic P is the most suitable statistic for Adaptively Truncated MANOVA compared with the other two test statistics.

3.6 Chapter Summary

In this chapter, we firstly show the results of Exploratory Data Analysis (EDA). There are two ways of displaying the mean vectors of the functional data from the experiment. The mean functions of each group are not able to distinguish from each other with EDA. However, EDA provides us the structures of the original data setting at least. In general, it is more visual to draw EEM plots. Next, we select six peak values from each mean function based on the figures of EDA. Then, the number of observations is much larger than the number of dimensions so that traditional MANOVA is able to apply. The result of MANOVA shows that the room light factor is not statistically significant at the five percent level when we use MDA and LBJ FastEEM Box combined with frosted cuvette standard. While, the others are statistically significant. Without removing much information of the original data, PCA is another good consideration for the techniques of dimension reduction. By PCA, we successfully reduce the dimension of the original data so that the condition of MANOVA satisfies. By what we have expected, the room light factor is statistically significant at the five percent level in all measurements combined with standards.

However, the relationship between the original hypothesis with the transformed hypothesis is not clear yet.

The second part of this chapter focuses on the modifications of existing traditional method for functional data. We firstly introduce the ATHTST. Based on the thoughts of ATHTST, we come up with the idea of adaptively truncated MANOVA. Like ATHTST, we conduct the similar simulations for adaptively truncated MANOVA. Even though the test statistics of adaptively truncated MANOVA do not work better compared to the test statistics of ATHTST, it may be a primary thought of developing traditional MANOVA. It costs too much time for obtaining p-values from simulations. Therefore, finding new ways for decomposition of the matrix may reduce the computation time. Meanwhile, the assumption of equal variance in both simulations of ATHTST and Adaptively Truncated MANOVA needs further discussion.

Chapter 4

CONCLUSION

The objective of this thesis is to determine if the factor of room light is significant using the functional data setting from our experiment. Because the conditions of traditional methods usually fail with functional data, we need to reduce the dimension of the original data without removing any important information. The thesis shows the background of cervical cancer in the first chapter. Then, we come up with the issues in the research project and how the experiment is designed. Proper statistical methods are given in Chapter 2. It is necessary to reduce the dimension of the data if we would like to apply the traditional statistical methods to the functional data. Two dimension reduction techniques are provided in Chapter 2.

The EDA suggests that the mean functions in each group are quite similar and difficult to distinguish with each other. However, the EDA provides us a useful idea of the techniques of dimension reduction. Having chosen the six peak values in the figures, the traditional MANOVA concludes that the room light factor is not statistically significant at the five percent level with the data of MDA and LBJ FastEEM Box combined with frosted cuvette standards. Another remarkable way is PCA. The PCA successfully shows the result as we have expected. The room light factor is statistically significant at the five percent level. Compared with traditional

MANOVA using the subset of data by the EDA, the PCA remains much more information of the original data setting. For this reason, the PCA should be our first consideration of techniques of dimension reduction.

Furthermore, the thesis extends the existence ATHTST to MANOVA. Depending on the thoughts of ATHTST, we come up with three test statistics for Adaptively Truncated MANOVA. Also, we perform simulations for the new three test statistics of Adaptively Truncated MANOVA. Before we can apply the test statistics to the real data, we need much more future work. It is quite useful to show the empirical cumulative distributed function of p-values under the null hypothesis for our next step. Meanwhile, the assumption that the data is from normal distribution with equal covariance needs further check. If we could develop new methods for the decomposition of matrix, it could save us much more time. Whether the new test statistics are stable and suitable for unbalanced large data is still in doubt. It is the next coming issue of our research project.

REFERENCES

Cox, D. D., & Lee, J. S. (2008). Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika Trust*, 7-10.

Fan, J., & Lin, S. (1998). Test of significance when data are curves. *Journal of the American Statistical Association*, 1007-1021.

Lee, J. S., & Cox, D. D. (2008). Sources of Variablity in FastEEm Mearsuments. *EVIES*, 42-45.

Lee, J. S., Cox, D. D., & Follen, M. (2010). Adatptively truncated Hotelling's T-Squre test for functional data with applications. 2-22.

Lee, J. S., Follen, M., MacAulay, C., Pikkula, B., Serachitopol, D., Price, R., et al. (2007). Sources of variability in fluorescence spectroscopic measurements in a Phase II clinical trial of 850 patients. *Gynecologic Oncology*, 4-8.

Lee, J. S., Shuhatovich, O., Price, R., Pikkula, B., Follen, M., Mckinnon, N., et al. (2005). Design and preliminary analysis of a study to assess intra-device and inter-device variability of fluorescence spectroscopy instruments for detecting cervical neoplasia. *Elsevier*, 98-111.

Miller, A. B. (1992). *Cervical Cancer Screening Programmes.* Toronto: World Heath Organization Geneva.

Pikkula, B. M., Shuhatovich, O., Price, R. L., & Serachitopol, D. M. (2007). Instrumentation as a source of variability in the application of fluorescence spectroscopic devices for detecting cervical neoplasia. *Biomedical Optics*, 5-6.

Rencher, A. C. (2002). *methods of Multivariate Analysis.* John Wiley & Sons Inc.

Shen, Q., & Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica*, 1239-1257.

Taylor, J., Worskey, k., & Gosselin, F. (2007). Maxima of discretely sampled random fields with an application to 'bubbles'. *Biometrika*, 1-18.

Wald, A., & Wolfowitz, J. (1944). Statistical tests based on permutation of the observations. *Annals of Mathermatical Statistics*, 358-372.

Yamal, J.-M., Zewdie, G. A., Cox, D. D., Atkinson, E. N., Cantor, S. B., MacAulay, C., et al. (2012). Accuracy of optical spectroscoy for the detection of cervical intraepithelial neoplasia without colposcopic tissue information; a step toward automation for low resource settings. *Biomedical Optics*, 4-7.